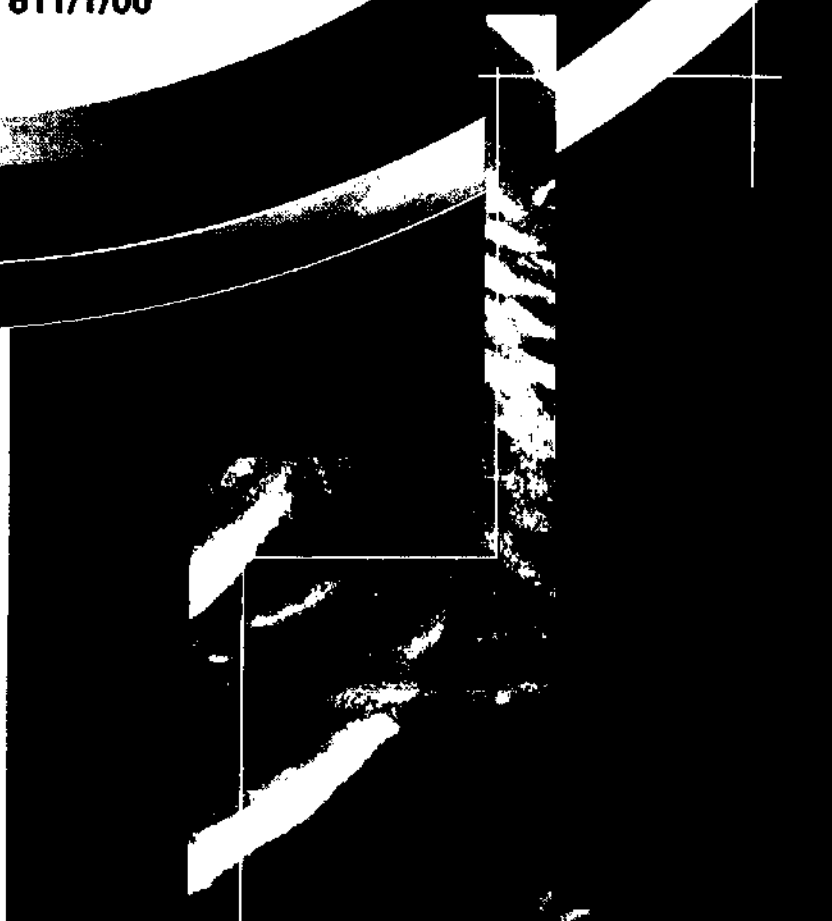
LONG DURATION DESIGN RAINFALL ESTIMATES FOR SOUTH AFRICA

JC Smithers • RE Schulze

WRC Report No. 811/1/00



Water Research Commission

LONG DURATION DESIGN RAINFALL ESTIMATES FOR SOUTH AFRICA

By

J C Smithers and R E Schulze

School of Bioresources Engineering and Environmental Hydrology
University of Natal
Pietermaritzburg
South Africa

Final Report to the
Water Research Commission

WRC Report No: 811/1/00
ISBN No: 1 86845 650 1

EXECUTIVE SUMMARY

Design rainfall depths for various durations are required for the many engineering and conservation design decisions made annually in South Africa and which result in millions of Rands of construction. For example, 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. Hence 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).

Design rainfall depths for durations of one day and longer were last estimated on a national scale at approximately 2400 stations in South Africa by Adamson (1981). One day design rainfall depths are computed using rainfall data measured at 08:00 every day for the preceding 24 h period by standard, non-recording raingauges. Since the study by Adamson (1981) a longer period of data is now available for analysis. Moreover, new techniques for estimating design values using a regional approach have now become accepted practice internationally, as regional approaches have been found to generally result in more reliable design values than traditional single site approaches.

The major objective of this study was the revision of medium to long duration (i.e. 1 - 7 day) rainfall Depth-Duration-Frequency (DDF) relationships for South Africa. In addition, the development of a processing system to enable future updates of medium to long duration design rainfall values to be performed relatively easily and quickly was envisaged.

One of the requirements of frequency analyses is a collection of long periods of records. A good distribution of daily rainfall data with relatively long records is available in South Africa. For example, nearly 4 000 stations have record lengths of 20 years and longer while more than 1 800 raingauges have more than 40 years of record. Of concern to future hydrological studies that require long rainfall records is the decrease in the number of operational raingauges maintained by the South African Weather Bureau (SAWB). Based on the daily rainfall database housed by

the Computing Centre for Water Research, 2480 stations were operational in the SAWB daily raingauge network for the period 1976 - 1985, while the number of raingauges decreased to 1786 for the period 1986 - 1995. Based on this trend, it is expected that the number and spatial density of stations with long records will decrease even further in the future.

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 rainfall, or flood, frequency curve that is applicable anywhere in that 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).

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 slight heterogeneity exists within a region, 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 the 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 rainfall or flood 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 combined 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 and hence a regional approach to the estimation of 1 to 7 day design rainfall for South Africa was adopted in this study.

Regional approaches are 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. This approach, in conjunction with other techniques, has been successfully used by Smithers and Schulze (1998) to estimate short duration design rainfall in South Africa.

In this study a regionalised, index storm based frequency analysis using L-moments was adopted for design rainfall estimation. Homogeneous rainfall regions in South Africa were identified using daily rainfall data from 1 789 stations which have at least 40 years of record. Regionalisation was performed using site characteristics and tested independently using at-site data. The General Extreme Value (GEV) probability distribution was found to be the most suitable function to estimate 1 day design rainfall values in South Africa. For each of the homogeneous regions and for durations of 1 to 7 days quantile growth curves, which relate the ratio between design rainfall depth and an index storm to return period, have been developed. These regionalised quantile growth curves , in conjunction with index values derived from at-site data, were used to estimate design rainfall values at 3 945 rainfall stations in South Africa which have at least 20 years of daily record.

This report consists of seven chapters plus appendices. Chapter 2 contains a review of design rainfall estimation and in particular summarises the Regional L-Moment Algorithm (RLMA), as proposed by Hosking and Wallis (1993; Hosking and Wallis, 1997), and also reports on 1 day

and longer design rainfall studies conducted in South Africa. The spatial distribution, record lengths and missing data in the daily rainfall database are examined in Chapter 3. The techniques used to infill missing daily rainfall data are described in Chapter 4. The results of the application of the RLMA are contained in Chapters 5 (identification of homogeneous regions) and 6 (estimation of design rainfall). The results produced by the study are discussed and conclusions are drawn in Chapter 7, which also contains recommendations for future research in this field. Appendix A contains examples of design rainfall values and 90% error bounds for return periods ranging from 2 to 200 years and for durations of 1, 2, 3 and 7 days at selected sites in South Africa. Design rainfall values for all 3 945 stations are contained in Portable Document Format (PDF) on the two diskettes which accompanying this report.

DAILY RAINFALL DATABASE

The data used in this study were limited to those in the daily rainfall database maintained by the Computing Centre for Water Research (CCWR). Of the 1 1171 stations available in the database, data from 78.9 % of the stations were contributed by the South African Weather Bureau (SAWB), 7.7 % by the Department of Agriculture's Institute for Soil, Climate and Water (ISCW), 3.3 % of the stations are joint SAWB and ISCW stations, 1.4 % of the stations by the South African Sugar Association Experiment Station (SASEX) and the remainder (8.8 %) by private individuals. The data used in this study were thus only as up to date as the data contained in the daily rainfall database maintained by the CCWR as of January 1999. One shortcoming of this database is that data from the ISCW were last updated in approximately 1985 and this study would have benefited with more recent data from this source.

The reliability of design rainfall values increases with longer records and records lengths less than 10 years are generally not suitable for design rainfall estimation. Hence an assessment of the number of daily rainfall stations, the available record lengths and the amount of missing data was made. In contrast to the findings of Smithers and Schulze (1998) with the short duration rainfall database for South Africa, both the number of stations with relatively long periods of record and the spatial distribution of these stations in South Africa is good. The distribution of record lengths for all stations in the database is shown in Figure 1.

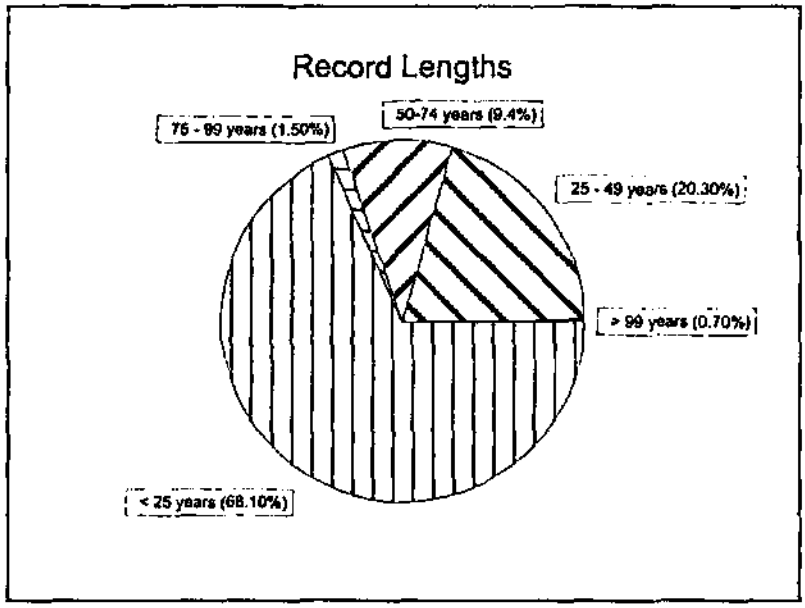


Figure 1 Distribution of 11 171 daily rainfall record lengths in southern Africa

An assessment of the amount of missing data in the daily rainfall database for stations with record lengths longer than 20 years was made. The results of this analysis indicate that more than 20 % of daily rainfall stations in South Africa, which have record lengths longer than 20 years, have more than 10 % of their data missing in the rainfall season. These missing data could be crucial to the estimation of design rainfalls and therefore the data need to be synthesised.

INFILLING MISSING DAILY RAINFALL DATA

The Expectation Maximisation Algorithm (EMA), formalised by Dempster *et al.* (1977), was adopted by Makhuvha *et al.* (1997a; 1997b) to infill missing data in monthly rainfall records. The EMA recursively substitutes missing data and then re-estimates the regression. Makhuvha *et al.* (1997a) treated all the records simultaneously and Makhuvha *et al.* (1997b) showed that this approach outperformed other regression based methods in terms of accuracy, variance preservation and speed of infilling.

Based on initial results from a report currently under preparation by Smithers *et al.*(1999), which evaluates the performance of inverse distance weighting, driver station, stochastic and EMA

techniques for infilling missing data, the EMA approach to infill missing daily rainfall was adopted in this study.

Prior to infilling missing rainfall data, outliers need to be identified and the sites grouped (Pegram, 1997b). Pegram (1997a) developed a set of routines (CLASSR) to enable a user to detect outliers and select suitable groupings of stations for the infilling of missing monthly rainfall totals. In addition, a modified version of the EMA used by Makhuvha *et al.* (1997a; 1997b) was utilised by Pegram (1997a) to create the PATCHR routines which are used to infill missing monthly rainfall totals. The suite of programs, developed by Pegram (1997a) to infill monthly rainfall totals, were modified as part of this project to operate on a daily time step.

For each of 3 945 daily rainfall stations in southern Africa, extracted from the daily rainfall database housed by the CCWR and which have 20 or more years of continuous records, 9 initial control stations were identified using the Euclidean Distance (*ED*) between each target station and all other potential control stations. The characteristics which were normalised, then weighted and used in the calculation of *ED* were the distances between, and differences in, mean annual precipitation and altitude of the target and potential control stations and an index of the overlapping years of record between the target and control stations.

One of the problems associated with the infilling of missing daily rainfall data is that the daily rainfall total for the same event may be incorrectly recorded by some observers and appear in the records as occurring on different days at adjacent or near by stations, i.e. some observers record the rainfall measured at 08:00 as occurring on the previous day whilst other observers may record the total for the 24 h period ending at 08:00 against the date for the current day. Hence a phasing problem is introduced into the data. This phasing problem has previously been identified in South African daily rainfall data by Schulze (1980) and Meier (1997) and its impact becomes important when modelling runoff at a daily time step from a distributed catchment set up using rainfall data from a number of different daily raingauges. In addition, the phasing problem could lead to erroneous relationships between stations being developed by the EMA and thus influence the infilled values. For these reasons the phasing problem of daily rainfall data was addressed in this study. Although it may be argued that the phasing problem is not always the result of an error by an observer and it could be attributed to the random nature of daily

rainfall, the persistent and systematic nature of this phasing error confirms that in the majority of the cases the data have been incorrectly recorded by the one of the observers. In an attempt to automate the correction of the phasing error for the purposes of infilling missing values using the EMA, rainfall events were identified in each of the 9 initial control stations and the entire event was lagged or advanced by a single day if the shift of the event improved the phasing of the rainfall data between the control and target station.

When collecting hydrometeorological data it is inevitable that errors will occur. In daily rainfall data, in addition to the phasing errors discussed in the previous paragraph, errors in recorded rainfall amounts may be due to incorrect recording of the rainfall depth by the observer or due to errors introduced when the data are transcribed into an electronic form. An example of such an error may the incorrect placement of the decimal point for the rainfall on a particular day, as has been illustrated for extreme events in South Africa by Schulze (1984). One method of attempting to identify such errors is to investigate inconsistencies between the data from stations which are relatively close to each other and in this regard the concept of the covariance biplot is useful in identifying hydrologically similar raingauges and for identifying outliers (Basson *et al.*, 1994; Pegram, 1997a; Pegram, 1997b). Thus, a hierarchical procedure was implemented such that when monthly totals of daily rainfall were identified using the covariance biplot as potential outliers, then outliers in the monthly rainfall totals were computed. If a single high (or low) outlier was identified in the monthly rainfall totals, then outliers in the total rainfall for a four day moving window were identified in the outlier month, and rainfall for a each day which was identified as an outlier in all the windows in which it appeared, was considered an outlier and excluded from the infilling process.

The cluster analysis output from the CLASSR program (Pegram, 1997b) was used to automate the identification of suitable control stations from the initial 9 control stations selected on the basis of *ED* as described above. Using the EMA algorithm (Makhuvha *et al.*, 1997a), as implemented by Pegram (1997a), missing data in the target and control stations were then infilled simultaneously. In this implementation only the infilled values from the target station were retained as it was postulated that missing data in the control stations may be infilled better, possibly by using more suitable control stations, when each of the control stations was considered as the target station. However, in the event that for a particular target station one or

more of the control stations had already been infilled (i.e. they had already been considered as target stations), then the infilled values were used to infill the current target station under consideration.

REGIONALISATION OF DAILY RAINFALL

A procedure similar to that used by Smithers and Schulze (1998) was adopted for the regionalisation of the daily rainfall stations into relatively homogenous regions for the estimation of design rainfalls. This approach was based on the Regional L-Moment Algorithm (RLMA) developed by Hosking and Wallis (1993; 1997), which identifies potentially homogeneous regions by a cluster analysis of site characteristics and then tests the homogeneity of the region using the statistics of the sites in the region. Subdivision of South Africa was achieved by a cluster analysis of site characteristics using Ward's minimum variance hierarchical algorithm (SAS, 1989), which tends to form clusters of roughly equal size (Hosking and Wallis, 1997). The site characteristics, which were normalised, consisted of latitude (°), longitude (°), altitude (m), concentration of precipitation (%), mean annual precipitation (mm), rainfall seasonality (category) and distance from sea (m). The cluster analysis is the most subjective aspect of the RLMA and it may be necessary to subjectively relocate sites or create new clusters, but based on geographical and physical considerations (Hosking and Wallis, 1997). The measures of discordancy (D) and heterogeneity (H) developed by Hosking and Wallis (1993; 1997) were used to identify anomalies in the data and test for homogeneous regions respectively.

The distribution of rainfall stations in South Africa with at least 40 years of record was considered adequate for regionalisation. Ten of the 1 806 rainfall stations which met this criteria were excluded from the regionalisation procedure and used to independently evaluate the performance of the RLMA. In addition to the 10 stations hidden from the regionalisation for the purposes of assessing the performance of the RLMA, a further 8 stations were excluded which displayed significant trends in the annual rainfall totals and which were discordant from the surrounding stations. After limited subjective relocation of stations, the remaining 1 789 stations were used to identify 78 relatively homogeneous clusters. The number of stations per cluster ranged from 3 and 66 stations with an average of 23, and it was found that the degree of

heterogeneity was not related to the number of stations per cluster. The spatial distribution of the 78 relatively homogeneous clusters in South Africa is shown in Figure 2.

ESTIMATION OF DESIGN RAINFALL

Once relatively homogeneous rainfall regions have been identified, the next step in the RLMA is the selection of an appropriate probability distribution to be used in the frequency analysis. Given a homogeneous region, a goodness-of-fit 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 goodness-of-fit 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. This approach resulted in the General Extreme Value (GEV) being adopted as the most appropriate distribution to use in all clusters.

Uncertainty is inherent in any statistical analysis and hence it is necessary to assess the magnitude of this uncertainty. Conventionally 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 within regions, the use of an inappropriate frequency distribution and dependence between observed data at different sites.

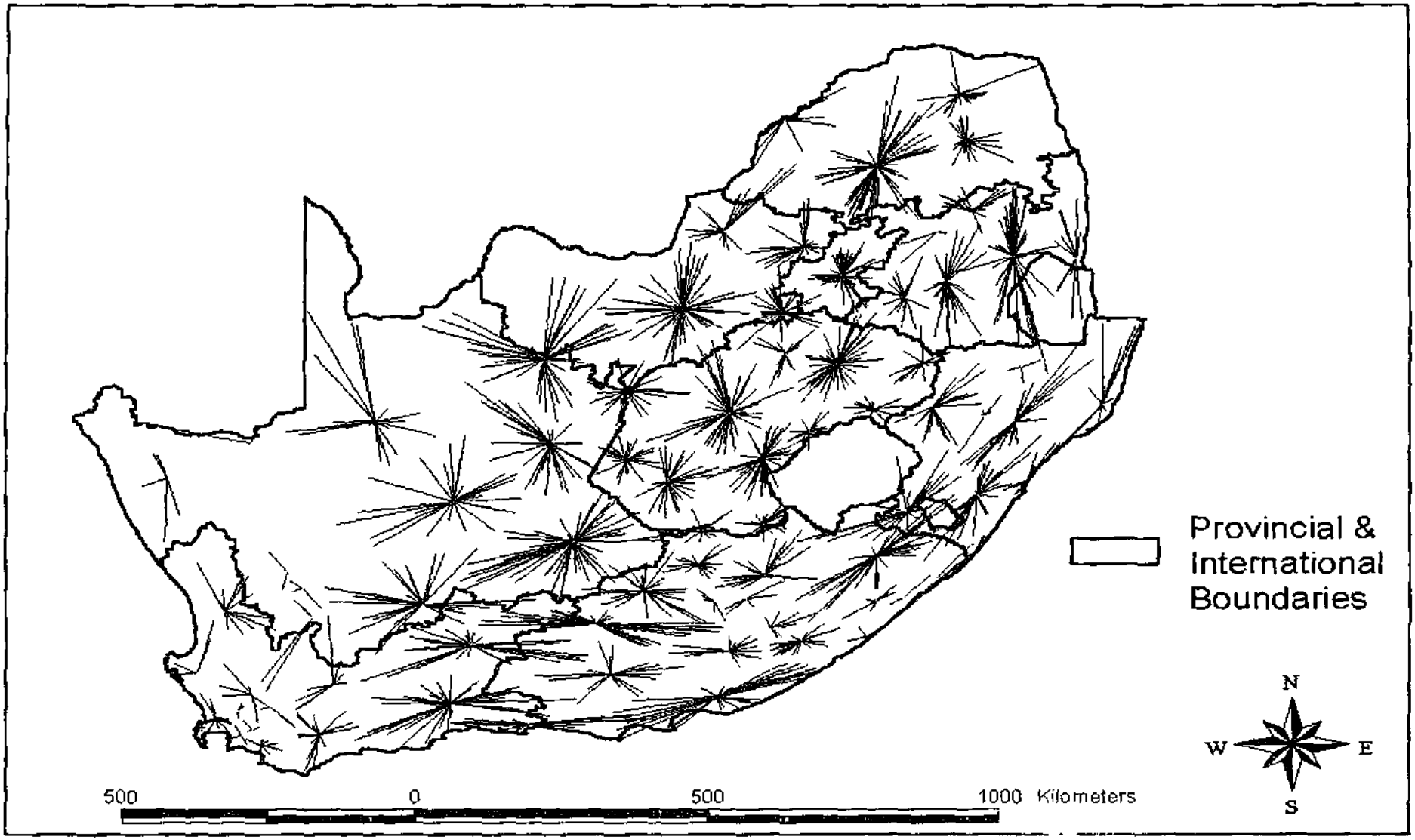


Figure 2 Distribution of 78 relatively homogeneous clusters in South Africa

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. Using this approach, 90% error bounds were estimated for each of the regional quantile growth curves which, relates the ratio between the 1 day design rainfall and an index value to return period. The index value used in the RLMA was the mean of the 1 day Annual Maximum Series (AMS).

In order to assess the performance of the RLMA, 10 daily rainfall stations which cover a range of climatic regions in South Africa were excluded from the regionalisation process. Each of these stations was allocated to the cluster with the closest Euclidean distance between the site characteristics of the station and the mean of the site characteristics of all sites within a cluster. The locations of the hidden stations are shown in Figure 3 and cluster numbers determined for each of the hidden stations are listed in Table 1.

Table 1 Hidden stations and cluster numbers

Station	Name	Cluster
0021055 W	Cape Town Maitland	51
0059572 A	East London	4
0144899 W	Middleburg	6
0239482 A	Cedara	15
0261368 W	Bloemfontein	10
0299357 W	Cathedral Peak Hotel	17
0317447AW	Upington	35
0442811 W	Nooitegedacht	24
0513404 W	Pretoria	16
0677834 W	Pietersburg	28

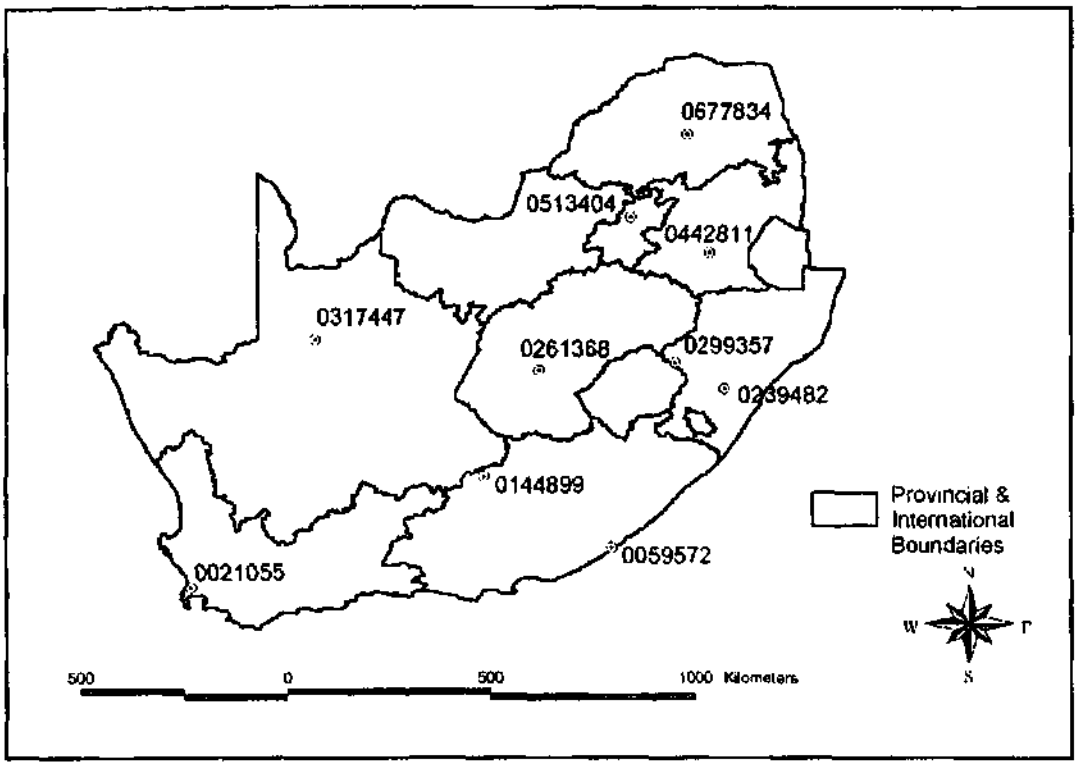


Figure 3 Location of the 10 hidden daily rainfall stations in South Africa

A comparison between the design rainfall estimated using the at-site data and and estimated from the regional quantile curve is shown in Figure 4 for the 10 hidden stations which were not used in the regionalisation procedure. Included in Figure 4 are the 90% error bounds of the design values estimated from the error bounds of the quantile growth curve.

As shown in Figure 4, the 1 day design rainfall depths estimated from the observed data and from the regional growth curve are similar for return periods up to 20 years and, with the exception of three stations (0021055 W, 0239482 A and 0513404 W), the values estimated from the regional growth curve generally exceed the values estimated from the at-site data for return periods greater than 20 years. The regional growth curve pools information from stations within a relatively homogeneous region and is thus considered to result in more reliable estimates of design rainfall than values estimated directly from the at-site data. Thus the recommended design values estimated using the regional growth curve are generally more conservative for longer return periods than those estimated directly from the at-site data.

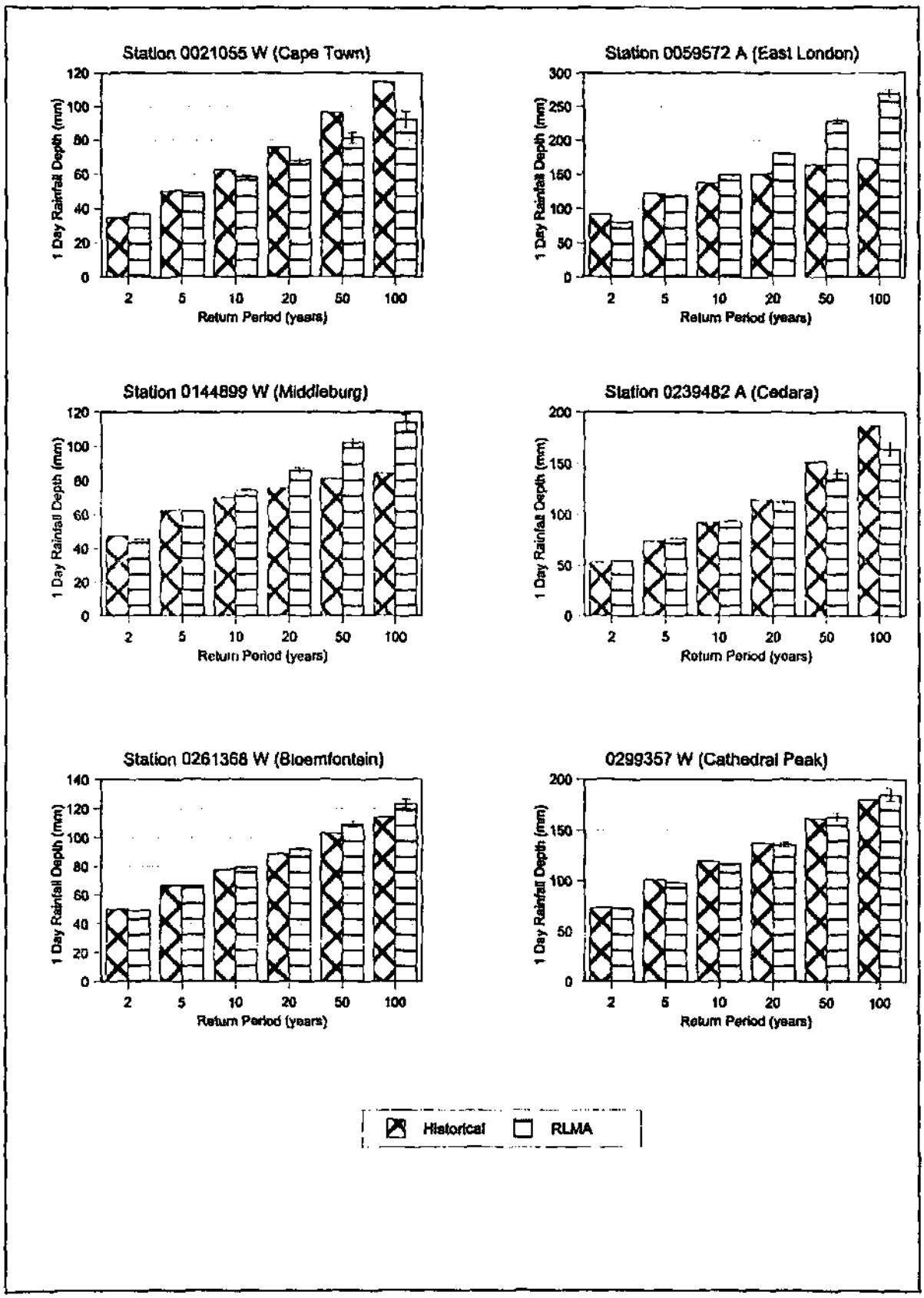


Figure 4 Comparison of design rainfall depths computed from at-site data and from regional growth curves at 10 stations not used in the regional process (I-beams indicate 90% error bounds)

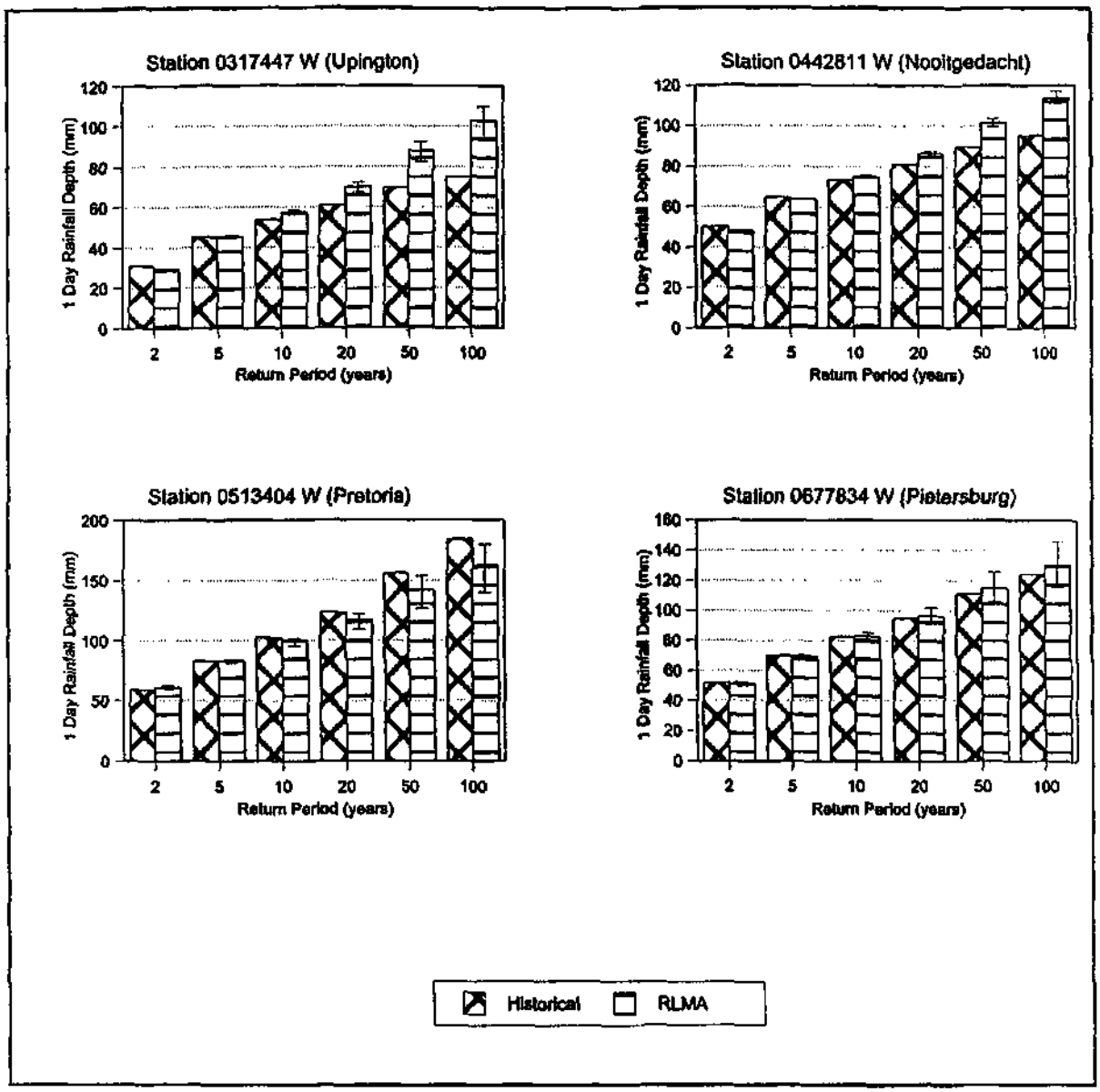


Figure 4 (cont) Comparison of design rainfall depths computed from at-site data and from regional growth curves at 10 stations not used in the regional process (I-beams indicate 90% error bounds)

A comparison was performed between the 1 day design rainfall estimated in this study using a regional approach and those estimated by Adamson (1981). This analysis indicated that for return periods less than 50 years the differences between the design rainfall estimated in this study and by Adamson (1981) are less than 20 % at the majority of the stations. As expected, the differences are bigger for longer return periods and for return periods ≥50 years a trend is discernible with the Adamson design values exceeding the values computed in this study.

Apart from the longer record lengths and the stringent data quality control procedures used in this study, some of the differences in design values estimated may be attributed to the different approaches taken in the two studies. Adamson (1981) used a single site approach with a censored LN distribution whereas this study adopted a regional approach and used the GEV distribution. In addition, the use of L-moments in this study to fit the GEV distribution results in less influence by outliers in the data. As shown in Figure 4, design rainfall depths computed using the regional approach generally exceed the values computed directly from the at-site data. In addition, the regional approach has been shown to result in more reliable and robust estimates compared to single site point estimates. Thus it is postulated that the design values computed in this study may be used with confidence.

Design rainfalls were computed at 3 945 sites in South Africa using the quantile growth curves for the 78 clusters and the mean of the AMS computed at the 3 945 stations, each of which has at least 20 years of record. These point design values are included as PDF files on the diskettes which accompanying this report.

ACKNOWLEDGEMENTS

The research in this report emanated from a project funded by the Water Research Commission and entitled "Long Duration Design Rainfall Estimates for South Africa".

The steering committee responsible for the project consisted of the following members:

Mr H Maaren	Water Research Commission (Chairman)
Dr GC Green	Water Research Commission
Dr J du Preez	Water Research Commission (Secretary)
Dr MC Dent	Computing Centre for Water Research
Prof RE Schulze	University of Natal (Project leader)
Prof PWL Lyne	University of Natal
Prof GGS Pegram	University of Natal
Mr EJ Schmidt	South African Sugar Association
Mr CB Schultz	*Gibb Africa*

The financing of the project by the Water Research Commission and the contribution of the members of the Steering Committee is acknowledged gratefully. This project was only possible with the co-operation of many individuals and institutions. The lead author therefore wishes to record his sincere thanks and to acknowledge the following contributions to this project:

- Professor Roland Schulze, project leader, School of Bioresources Engineering and Environmental Hydrology (BEEH), University of Natal, for his support, encouragement, guidance and opportunity to undertake this project,
- Professor Geoffrey Pegram, School of Civil Engineering, Surveying and Construction, University of Natal, for his guidance, encouragement and for making available the code for CLASSR and PATCHR software,
- Department of Water Affairs and Forestry who sponsored the development of the CLASSR and PATCHR software through BKS (Pty) Ltd,
- Professor Peter Lyne, head of the School of BEEH, University of Natal, for supporting this project,

- the Computing Centre for Water Research (CCWR), for computing facilities, assistance and access to daily rainfall data,
- Miss Kershani Chetty and Miss Marilyn Royappen of the School of BEEH for assisting with data analysis,
- the University of Natal Research Fund for contributing towards the funding for this study, and
- to my wife, Kary, and children, Jonathan and Bronwen, for their love, understanding, support and encouragement.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

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 design rainfall depths for various durations 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).

The required duration of design rainfall which is used in design flood estimation may range from as short as 5 minutes for small urban catchments, which have a rapid hydrological response, to a few days for large regional flood studies. Techniques have been developed and evaluated for the estimation of short duration ($\leq$ 24 h) design rainfall in South Africa by Smithers and Schulze (1998). Design rainfall depths for durations of one day and longer were last estimated on a national scale at approximately 2400 stations in South Africa by Adamson (1981). Since the study by Adamson (1981) a longer period of data is now available for analysis. Moreover, new techniques for estimating design values using a regional approach have now become accepted practice internationally as regional approaches generally result in more reliable design values than traditional single site approaches.

The major objective of this study was the revision of medium to long duration (i.e. 1 - 7 day) rainfall Depth-Duration-Frequency (DDF) relationships for South Africa. In addition, the development of a processing system to enable future updates of medium to long duration design rainfall values to be performed relatively easily and quickly was envisaged. A secondary objective was the revision of point to area rainfall depth relationships for medium to long durations in South Africa.

A distinction is drawn between 1 day and 24 h design rainfall values. When automatically recorded short duration rainfall data are available, for example as recorded autographically or by data loggers, a sliding 24 h window is used to extract the maximum 24 h duration event irrespective of the time of the start of the event. Using the sliding window approach, the true maximum for any 24 h period may be estimated and hence the 24 h design rainfall value will exceed the 1 day value computed from rainfall measured at fixed daily increments. The 1 day design rainfall values can be converted to 24 h values using fixed ratios, which may vary regionally.

The 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. A good distribution of daily rainfall data with relatively long records is available in South Africa. For example, nearly 4000 stations have record lengths of 20 years and longer while more than 1800 raingauges have more than 40 years of record.

Of concern to future hydrological studies that require long rainfall records is the decrease in the number of operational raingauges maintained by the South African Weather Bureau (SAWB). Based on the daily rainfall database housed by the Computing Centre for Water Research, 2480 stations were operational in the SAWB daily raingauge network for the period 1976 - 1985, while the number of raingauges decreased to 1786 for the period 1986 - 1995. Based on this trend, it is expected that the number and spatial density of stations with long records will decrease further in the future.

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).

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 slight heterogeneity exists within a region, 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 the 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 rainfall or flood 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 quantiles computed using both at-site data and regional information 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 and hence a regional approach to the estimation of 1 to 7 day design rainfall values was adopted in this study.

Regional approaches are 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. This approach in conjunction with other techniques has been successfully used by Smithers and Schulze (1998) to estimate short duration design rainfall in South Africa.

In this study a regionalised, index storm based frequency analysis using L-moments was adopted for design rainfall estimation. Homogeneous rainfall regions in South Africa were identified using daily rainfall data from 1789 stations which have at least 40 years of record. Regionalisation was performed using site characteristics and tested independently using at-site data. The General Extreme Value (GEV) probability distribution was determined to be the most suitable function to estimate 1 day design rainfall values in South Africa. For each of the homogeneous regions and for durations of 1 to 7 days quantile growth curves which relate the ratio between design rainfall depth and an index storm to return period have been developed. These regionalised quantile growth curves, in conjunction with index values derived from at-site data, were used to estimate design rainfall values at 3 945 daily rainfall stations in South Africa which have at least 20 years of record.

This document consists of seven chapters and appendices. Chapter 2 contains a review of design rainfall estimation and in particular summarises the Regional L-Moment Algorithm (RLMA), as proposed by Hosking and Wallis (1993; Hosking and Wallis, 1997), and also reports on 1 day and longer design rainfall studies conducted in South Africa. The spatial distribution, record lengths and missing data in the daily rainfall database are examined in Chapter 3. The techniques used to infill missing daily rainfall data are described in Chapter 4. The results of the application of the RLMA are contained in Chapters 5 (identification of homogeneous regions) and 6 (estimation of design rainfall). The results produced by this study are discussed and some conclusions are drawn in Chapter 7. The appendices contain examples of the design rainfall values and confidence intervals for return periods ranging from 2 to 200 years and for durations of 1 to 7 days at selected sites in South Africa. The design rainfall values and their 90% error bounds computed at all 3 945 stations for durations of 1 to 7 days are contained in Portable Document Format (PDF) on the diskettes which accompany this report.

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 design rainfall is a key concept in the design of hydraulic structures where a return period is selected according to the cost and significance of the structure.

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 design storm estimation since 1981 there still appears to be little consensus in the literature on preferred approaches to adopt. However, the relatively recent developments in regional approaches to the estimation of extreme events at a single site hold much promise for more general acceptance. Thus the objective of this chapter is to review the Regional L-moment Algorithm (RLMA), which is a regionalised index-value approach to frequency analysis based on L-moments. Hence L-moments are summarised in Section 2.1 and the RLMA reviewed in Section 2.2. This is followed by a review in Section 2.3 of one day design storm studies in South Africa.

Depending on the size of the catchment and its hydrological response time, the storm duration of interest may be from as little as 5 minutes for small urban catchments to a storm duration of a number of days for large catchments. Generally the procedures for estimating design storms are the same irrespective of the duration. However, for short duration storms ($\leq$ 24 h) the rainfall data are usually recorded continuously by data loggers or on autographic charts, while for longer durations (1 - 7 days) the source of the data is usually standard, non-recording raingauges from which observations are made in South Africa at 08:00 every day for the preceding 24 h period.

In a short duration design rainfall study for South Africa, Smithers and Schulze (1998) reviewed the literature of both single site and regional techniques for design storm estimation and concluded the following:

- Substantial benefits of using a regional approach have been reported in the literature, assuming that relatively homogeneous regions can be identified.
- L-moments are subject to less bias than ordinary product moments and should be used to fit probability distributions to the data.
- The relatively recently developed RLMA, developed by Hosking and Wallis (1993; 1997), appears to be a robust procedure and has been applied successfully in a number of studies.

Readers are referred to Smithers and Schulze (1998) for a detailed review leading to these conclusions. In addition, the RLMA was successfully applied by Smithers and Schulze (1998) for the estimation of short duration design storms and hence this approach was also adopted in this study for the estimation of design storms for durations of 1 day and longer. L-moments are defined and the RLMA is described in the following two sections.

2.1 L-moments

L-moments, as defined by Hosking (1990), are linear combinations of Probability Weighted Moments (PWMs). Greenwood *et al.* (1979) summarise the theory of PWMs. Unbiased sample estimates for the first four PWMs can be computed from the set of relationships making up Equation 1 (Stedinger *et al.*, 1993; Vogel and Fennessy, 1993).

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

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

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

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

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

$$\lambda_1 = b_0 \equiv \text{L-location (mean)} \qquad \text{...2a}$$

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

$$\lambda_3 = 6b_2 - 6b_1 + b_0 \qquad \text{...2c}$$

$$\lambda_4 = 20b_3 - 30b_2 + 12b_1 - b_0 \qquad \text{...2d}$$

where

λ_r = r-th L-moment.

Hosking (1990) defines the L-moment ratios (τ, τ_3 τ_4) as:

$$\tau = \frac{\lambda_2}{\lambda_1} \equiv \text{L-CV (coefficient of L-variation)} \qquad \text{...3a}$$

$$\tau_3 = \frac{\lambda_3}{\lambda_2} \equiv \text{L-skewness} \qquad \text{...3b}$$

$$\tau_4 = \frac{\lambda_4}{\lambda_2} \equiv \text{L} - \text{kurtosis} \qquad \ldots 3c$$

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.

The RLMA is a regional index value based procedure which is robust, uses simulation modelling to assess frequency distributions, utilises L-moments as summary statistics, allows a range of distributions to be evaluated and also pools regional information. The RLMA 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 The Regional L-moment Algorithm

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 are 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 from 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

$$Q_i(F) = \mu_i q(F), \quad i = 1,.....N \qquad ...4$$

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

$$\hat{\theta}_k^R = \sum_{i=1}^{N} n_i \hat{\theta}_k^{(i)} \Big/ \sum_{i=1}^{N} n_i \ . \qquad ...5$$

This is a record length weighted average, with the estimate at site i given a 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

$$\hat{Q}_i(F) = \hat{\mu}_i \hat{q}(F) \ . \qquad ...6$$

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 improvement in the accuracy of the regional growth curves for return periods shorter than 1000 years with more than 20 stations per cluster.

  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 RLMA, Hosking and Wallis (1993; 1997) proposed several stages in a regional frequency analysis and developed statistics, based on L-moments, that provide objective support in the procedures. These stages are discussed next.

2.2.1 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 3, and T denotes the transposition of a matrix. Hosking and Wallis (1997) define the discordancy index for site i as

$$D_i = \tfrac{1}{3} N(\mathbf{u}_i - \bar{\mathbf{u}})^T \mathbf{A}^{-1}(\mathbf{u}_i - \bar{\mathbf{u}}) \qquad ...7$$

where

$$\bar{\mathbf{u}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{u}_i \text{ , and} \qquad ...8$$

$$\mathbf{A} = \sum_{i=1}^{N} (\mathbf{u}_i - \bar{\mathbf{u}})(\mathbf{u}_i - \bar{\mathbf{u}})^T \qquad ...9$$

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.

2.2.2 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). The aim of this step is to form groups of sites that approximate the condition of homogeneity, i.e. the site's frequency distributions are highly similar 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, Mean Annual

Precipitation (MAP) and other physical properties). Hosking and Wallis (1997) recommend that the site characteristics, and not the site statistics, be used as the basis 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 $t^{(i)}$, $t_3^{(i)}$ and $t_4^{(i)}$. The regional average L-CV, L-skewness and L-kurtosis, denoted by t^R, t_3^R and t_4^R respectively, are weighted proportionally to the sites record length (n_i). For example

$$t^R = \sum_{i=1}^{N} n_i t^{(i)} \Big/ \sum_{i=1}^{N} n_i \qquad \ldots 10$$

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

$$V = \sqrt{\sum_{i=1}^{N} n_i (t^{(i)} - t^R)^2 / \sum_{i=1}^{N} n_i} \qquad \ldots 11$$

The 4-parameter 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

$$H = \frac{(V - \mu_v)}{\sigma_v} \qquad ...12$$

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 Choice of regional frequency distribution

After initial regionalisation has been performed, regions may still be slightly heterogeneous (i.e. $1 < H < 2$). 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 the upper bound of the distribution cannot be estimated with sufficient accuracy over the range of return periods of interest an unbounded distribution would better approximate the true distribution than a bounded 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 since 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 goodness-of-fit 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 goodness-of-fit 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 6.1.

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 site's 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 then 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

$$B_4 = \frac{1}{N_{sim}} \sum_{m=1}^{N_{sim}} (t_4^m - t_4^R) \qquad ...13$$

and the standard deviation of t_4^R as

$$\sigma_4 = \sqrt{\frac{1}{N_{sim} - 1} \times \left[\sum_{m=1}^{N_{sim}} (t_4^m - t_4^R)^2 - N_{sim} B_4^2 \right]}. \qquad ...14$$

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

$$Z^{DIST} = \frac{(\tau_4^{DIST} - t_4^R + B_4)}{\sigma_4} \qquad ...15$$

where

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

The fit is adequate if Z is close to zero and it is suggested that $|Z| \leq 1.64$ is a reasonable criterion to indicate that the fit of the assumed distribution is adequate (Hosking and Wallis, 1993; 1997).

2.2.4 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 2), and sample L-moment ratios $t^{(i)}$, $t_3^{(i)}$ and $t_4^{(i)}$, analogous to the population τ, τ_3 and τ_4 in Equation 3, then the regional average L-moment ratios t^R, t_3^R and t_4^R, which are weighted proportionally to the sites' record length, are computed as:

$$t^R = \sum_{i=1}^{N} n_i t^{(i)} / \sum_{i=1}^{N} n_i \qquad \ldots 16$$

$$t_r^R = \sum_{i=1}^{N} n_i t_r^{(i)} / \sum_{i=1}^{N} n_i \ , \ r = 3,4,.... \qquad \ldots 17$$

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 and t_4^R calculated in Equations 16 and 17. As shown in Equation 18, 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.

$$\hat{Q}_i(F) = l_1^{(i)} \hat{q}(F) \qquad \ldots 18$$

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 regional 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.

2.2.5 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 model's assumptions are

satisfied. Such confidence intervals are of limited use as all the assumptions regarding the data as rarely valid and uncertainty concerning the "correct" model selection is generally present (Hosking and Wallis, 1997). In order to obtain realistic assessments of the accuracy of the quantiles estimated using 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 into account in a way which is consistent with the data.

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 as follows:

(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, the procedure is as follows:

- 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:

$$R_i(F) = \sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(\frac{\hat{Q}_i^{m}(F) - Q_i(F)}{Q_i(F)}\right)^2} \qquad ...19$$

where

$R_i(F)$ = Root Mean Square Error (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, $R^R(F)$, where

$$R^R(F) = \frac{1}{N}\sum_{i=1}^{N} R_i(F) \qquad ...20$$

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 One Day Design Storm Estimation Studies in South Africa

A review of studies for the estimation of short duration (≤ 24 h) design rainfalls in South Africa was performed by Smithers and Schulze (1998) and is not repeated here. Relatively few studies in South Africa have looked specifically at rainfall durations of 1 day and longer.

The SAWB (1956) used the Extreme Value Type 1 (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. 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 to 100 years. Adamson (1981) used data from approximately 2400 stations in southern Africa and computed the 1, 2, 3 and 7 day design rainfalls for return periods up to 200 years. A censored log-Normal model of a partial duration series was used in this analysis of design rainfalls. More recently Pegram and Adamson (1988) used the Two Component Extreme Value (TCEV) distribution to estimate catchment based long duration design storms for selected catchments in Kwazulu-Natal.

All of the above studies in South Africa estimated point design rainfall values using at-site data only. Some regional smoothing was done in some of the studies (e.g. SAWB, 1956; Schulze, 1980) as the results are presented as isolines, interpolated from the point estimates, of design rainfalls for a specified return period. Thus no previous study has attempted to pool regional information and thus increase the reliability of the design values.

The daily rainfall database housed by the Computing Centre for Water Research (CCWR) was utilised to implement the RLMA and thus to estimate medium to long term design rainfall values for South Africa. Chapter 3 investigates the availability and quality of the records contained within this database.

CHAPTER 3

DAILY RAINFALL DATABASE

The reliability of design rainfall values increases with longer records and records lengths less than 10 years are generally not suitable for design rainfall estimation. Hence an assessment of the number of daily rainfall stations, the available record lengths and the amount of missing data is made in this chapter.

3.1 Station Distribution and Record Lengths

With the assistance of the CCWR, direct access to the daily rainfall database housed on the CCWR's mainframe computer was established. This enabled easy extraction of the daily records and prevented duplication of the database on the CCWR computing system. Of the 11 171 stations available on the database, 78.9 % of the stations have been contributed by the South African Weather Bureau (SAWB), 7.7 % by the Agriculture Research Council's Institute for Soil, Climate and Water (ISCW), 3.3 % of the stations are joint SAWB and ISCW stations, 1.4 % of the stations by the South African Sugar Association Experiment Station (SASEX) and the remainder (8.8 %) by private individuals.

The data used in this study were those contained in the daily rainfall database maintained by the CCWR as of January 1999. A limitation of this database is that data from the ISCW were last updated in approximately 1985 and this study would have benefited with more recent data from this source.

A database of site information was established which is required for the identification of relatively homogeneous regions in the index-storm based regional L-moments approach to design storm estimation. The site characteristics included in database are:

- latitude (°),
- *longitude* (°),

- altitude (m),
- seasonality (category), and
- concentration of precipitation (%), as defined by Markham (1970),
- mean annual precipitation (mm),
- distance from sea (m).

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

$$P_{\%,i} = 0.25 \times \frac{(P_{m,i-1} + 2P_{m,i} + P_{m,i+1})}{MAP} \times 100 \qquad \ldots 21$$

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 where 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.

In contrast to the findings of Smithers and Schulze (1998) with the short duration rainfall database for South Africa, the number of stations with relatively long periods of record have a good spatial distribution in South Africa. The distribution of record lengths for all stations in the database is shown in Figure 1. The spatial distribution in South Africa of the stations which have records lengths longer than 30 and 50 years are shown in Figures 2 and 3 respectively.

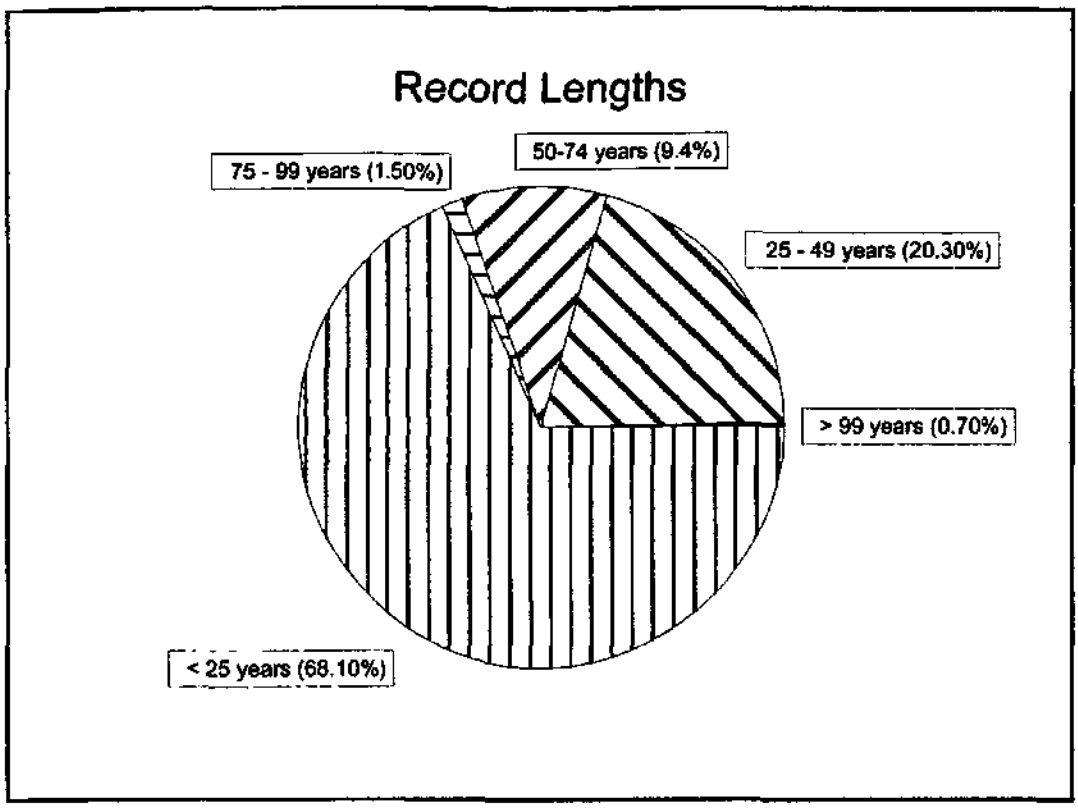


Figure 1 Distribution of daily rainfall record lengths in southern Africa

3.2 Missing Data

An assessment of the amount of missing data in the daily rainfall database for stations with record lengths longer than 20 years is shown in Figure 4. From Figure 4 it is evident that more than 20 % of daily rainfall stations in South Africa, which have record lengths longer than 20 years, have more than 10 % of their data missing in the rainfall season. These missing data could be crucial to the estimation of design rainfalls and therefore the data need to be repaired. Thus the infilling of missing data is addressed in the following chapter.

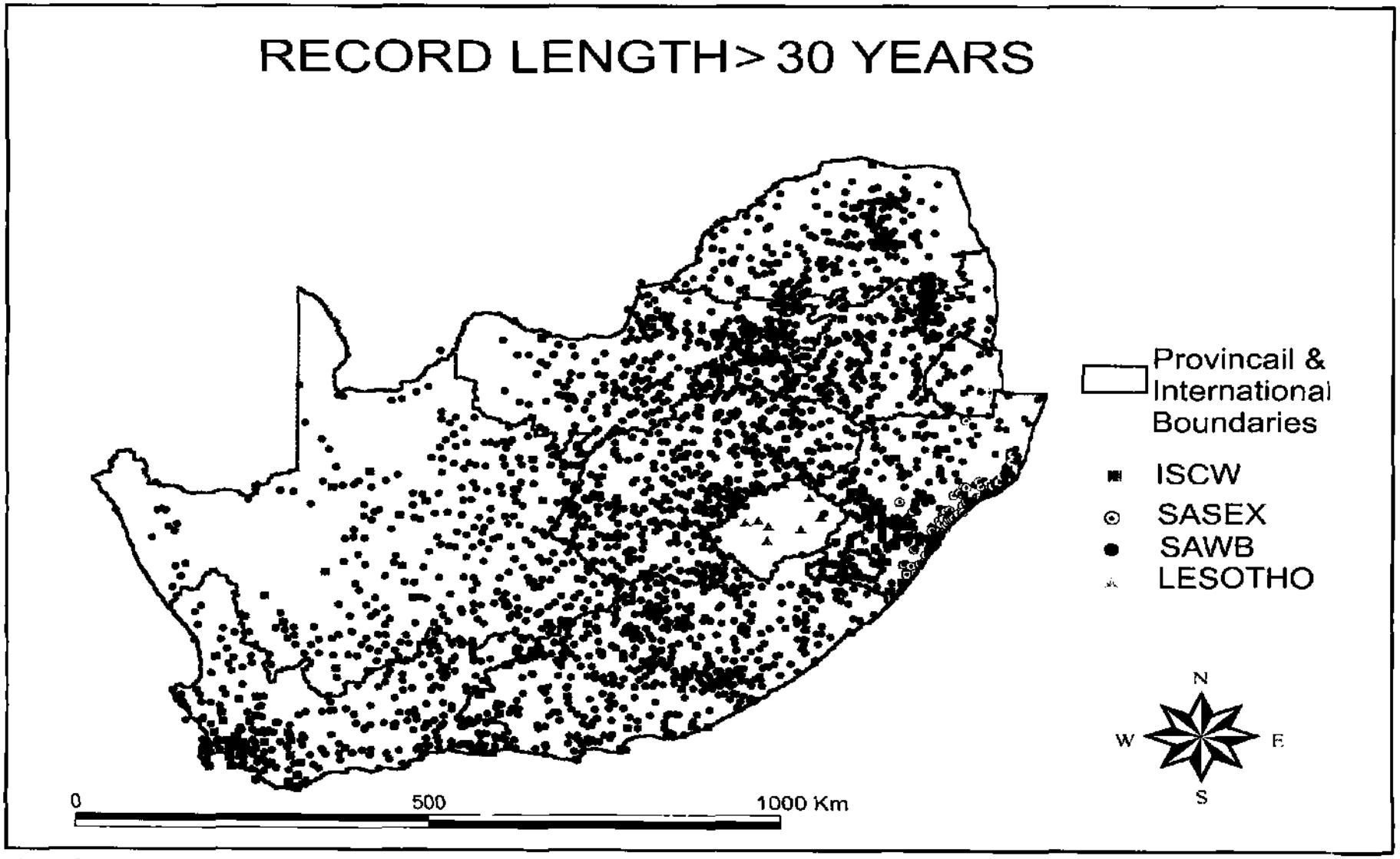


Figure 2 Location of daily raingauges with record lengths > 30 years

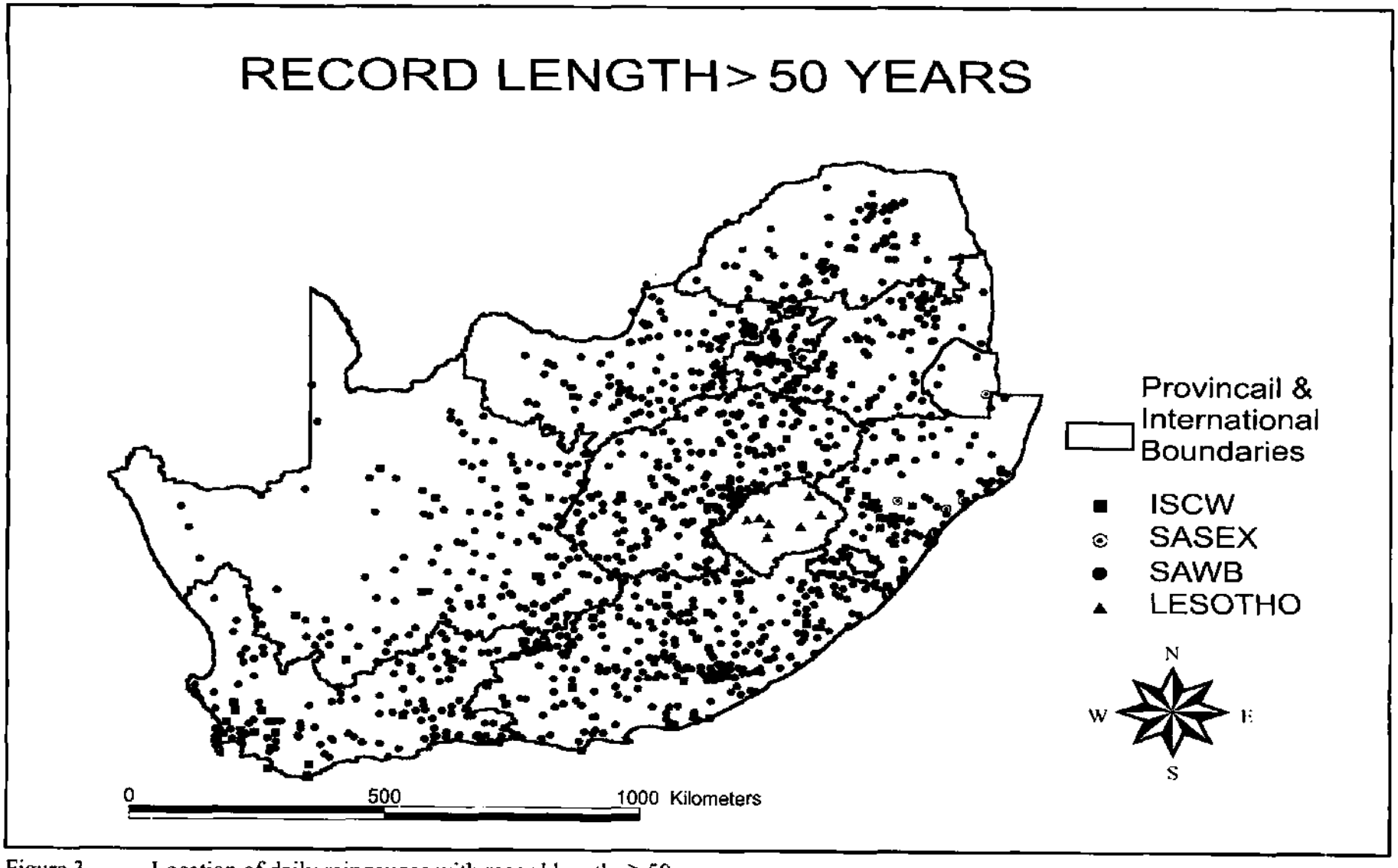


Figure 3 Location of daily raingauges with record lengths > 50 years

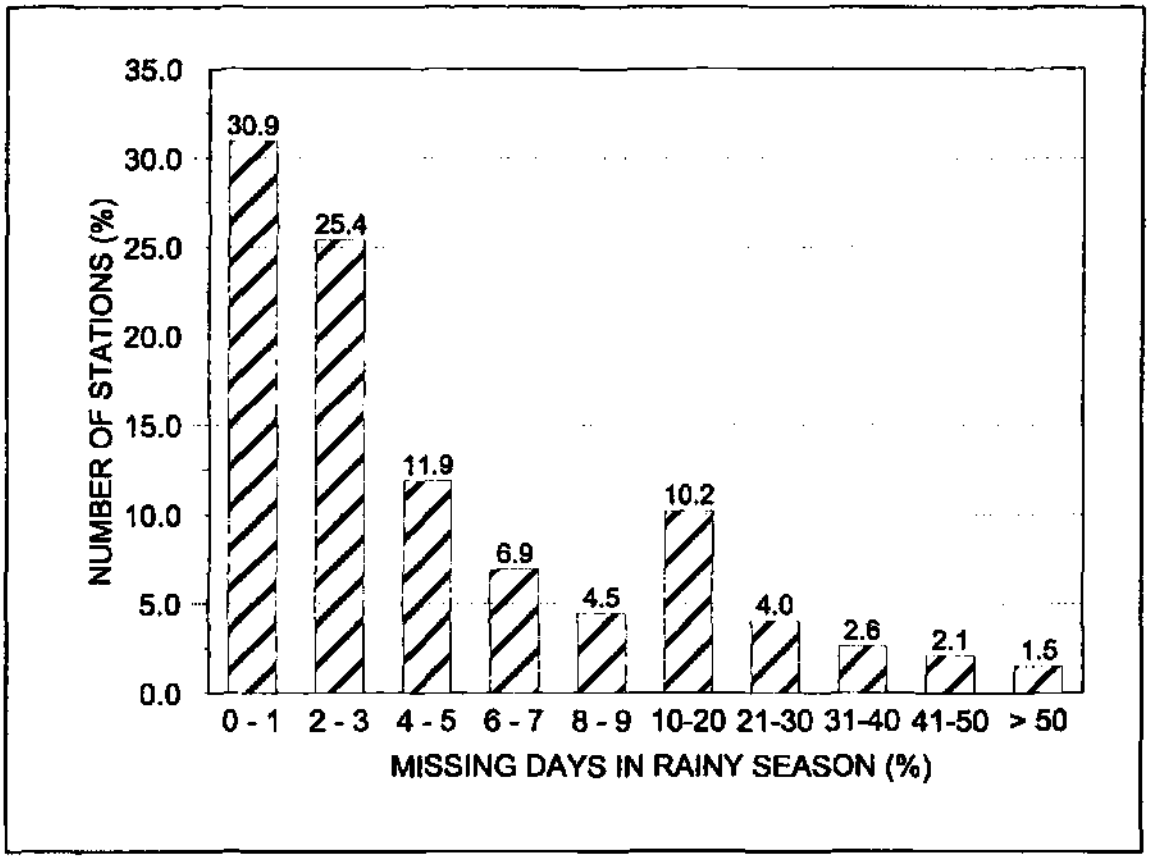


Figure 4 Analysis of missing daily rainfall data in South Africa at stations which have more than 20 years of record

CHAPTER 4

INFILLING MISSING DAILY RAINFALL DATA

Missing records in hydrological data limit the applications in which the data can be used. For example, missing records in daily rainfall data need to be infilled using an appropriate technique before a daily simulation model can use the data. Similarly, in the context of this study, missing days of data may contain extreme events which could be crucial to the estimation of design rainfall depths.

4.1 Selected Techniques for Infilling Missing Rainfall Data

A number of techniques for infilling and extending daily rainfall records have been developed and applied in South Africa. These include stochastic, inverse distance weighting, driver station and expectation maximisation techniques.

4.1.1 Stochastic

Zucchini and Adamson (1984) used a first order Markov chain to model the occurrences of daily rainfall in South Africa. The seasonal distribution of rainfall depths on rain days was modelled using a Weibull distribution with the mean estimated using a truncated Fourier series as shown in Equation 22. The shape parameter of the Weibull distribution was treated as a constant.

$$\mu(T) = \gamma_0 + \sum_{i=1}^{q} \gamma_i \cos\left[\frac{2\pi i}{365}(T - 1 - \delta_i)\right] \qquad ...22$$

where

$\mu(T)$ = mean of Weibull distribution of daily rainfall depth for day = T, T=1, 2, ..., 365,

γ_i = amplitude parameters, and

δ_i = phase parameters.

The wet:wet and wet:dry probabilities $\pi(T)$ were modelled as

$$\pi(T) = e^{\lambda(T)} / (1 + e^{\lambda(T)}) \qquad \ldots 23$$

where

$$\lambda(T) = \alpha_0 + \sum_{i=1}^{p} \alpha_i \cos\left[\frac{2\pi i}{365}(T - 1 - \beta_i)\right],$$

α_i = amplitude parameters, and

β_i = phase parameters.

The stochastic infilling technique makes use of a 12-parameter (p=2 and q=2) synthetic rainfall generator developed by Zucchini and Adamson (1984), with parameters available for 2550 daily rainfall stations in southern Africa. The daily rainfall data at the target station are scanned for missing data. When a period of missing data is found, a year of stochastic rainfall series is generated, and the missing data are infilled from the corresponding period in the stochastic series. This process is repeated for each period of missing data encountered, with a new year of synthetic series generated for each period of missing data. This method of infilling periods of missing daily rainfall has been automated by the CCWR for users in southern Africa.

A limitation of this method when applied, for example in rainfall-runoff modelling, is the synthetically generated values of daily rainfall takes no cognisance of the rainfall on the day in question at surrounding raingauges. Hence, the synthesised value may not reflect the general regional rainfall trend for the day in question and may not be synchronised with observed streamflow values.

4.1.2 Inverse distance weighting

The Inverse Distance Weighting (IDW) procedure weights the rainfall from selected surrounding control stations in relation to their individual distances from the target station. Hence, the closer a control station is situated to the target station, the higher the weighting that is assigned to the

control station. A procedure based on the inverse of the distance squared was developed by Meier (1997) where control rainfall stations within one degree latitude and longitude of the target station are identified. The area surrounding the target station is divided into four quadrants and the closest 10 rainfall stations in each quadrant to the target station are identified. Daily rainfall files for these 10 stations are retrieved from the daily rainfall database housed by the CCWR. When a period of missing data is encountered in the rainfall data at the target station, the closest station in each quadrant with an observed, non-missing value of rainfall occurring on that day is identified. Using the 1' x 1' MAP grid generated by Dent *et al.* (1987), this value is adjusted by the ratio of the MAP at the grid point of the closest station and the MAP at the grid point of the target station and the adjusted values are used to synthesize the missing value, as shown in Equation 24. If a single quadrant had no non-missing rainfall value for the day required, a single value was used from the remaining 3 quadrants. When 2 quadrants had no non-missing rainfall for the required day, 2 stations were selected from each of the remaining 2 quadrants. Similarly, if only a single quadrant had non-missing values on the required day, then 3 stations were selected from that quadrant.

$$r_t = \frac{\sum_{i=1}^{4} \frac{r_i \times MAP_{(t)}}{d_i^2 \times MAP_{(c,i)}}}{\sum_{i=1}^{4} \frac{1}{d_i^2}} \qquad \ldots 24$$

where

r_t	=	synthesised rainfall at target station,
r_i	=	observed rainfall at closest station with non-missing data in quadrant i,
$MAP(t)$	=	mean annual precipitation at the target station,
$MAP(c,i)$	=	mean annual precipitation at the control station in quadrant i, and
d_i^2	=	distance from the control station in quadrant i to the target station.

4.1.3 Driver station

In the driver station approach, periods of missing data in the target raingauge data are infilled from surrounding control raingauges. On days of missing data in the target data set, the data are infilled, after adjustment by the ratio of the MAP of the target and control raingauges as shown in Equation 25, from the raingauge deemed to be the most suitable. In the event of the most suitable raingauge also having missing data for the period in question, data from the next most suitable raingauge are used for infilling after appropriate scaling by the respective raingauge MAPs. This technique has been automated for users by the CCWR.

$$r_t = \frac{r_i \times MAP_{(t)}}{MAP_{(c)}} \qquad \ldots 25$$

where

r_t	=	synthesised rainfall at target station,
r_i	=	observed rainfall at selected control station with non-missing data on required day,
$MAP(t)$	=	mean annual precipitation at the target station, and
$MAP(c)$	=	mean annual precipitation at the control station.

4.1.4 Expectation maximisation algorithm

The Expectation Maximisation Algorithm (EMA), formalised by Dempster *et al.* (1977), was adopted by Makhuvha *et al.* (1997a; 1997b) to infill missing data in monthly rainfall records. The EMA recursively substitutes missing data and then re-estimates the multiple linear regression relationship between the data at the target station and the data from the selected nearby control stations. Makhuvha *et al.* (1997a) treated all the records simultaneously and Makhuvha *et al.* (1997b) showed that this approach outperformed other regression based methods in terms of accuracy, variance preservation and speed of infilling. The modifications made to the EMA are detailed in Makhuva *et al.* (1997a) and the EMA as summarised by Makhuva *et al.* (1997b) consists of 2 steps, which for a reasonable initial guess $\varphi^{(0)}$ of the parameters, which consist of μ and $\sum$, and for the $(r + 1)$-th interation are:

- **E Step**

Individual elements:

$$z_{ij}^{(r+1)} = z_{ij}, \text{ if } z_{ij} \text{ is observed} \qquad \text{...26a}$$

$$z_{ij}^{(r+1)} = \mu_j^{(r)} + \left[z_i^* - \mu_i^*\right]^T \beta_i^{*(r)}, \text{ if } z_{ij} \text{ is missing} \qquad \text{...26b}$$

Product elements:

$$\left[Z_{ij} Z_{ik}\right]^{(r+1)} = Z_{ij}^{(r)} Z_{ik}^{(r)}, \text{ if either is, or both are, observed} \qquad \text{...27a}$$

$$\left[Z_{ij} Z_{ik}\right]^{(r+1)} = Z_{ij}^{(r)} Z_{ik}^{(r)} + \sigma_{jk}^{(r)} - \beta_i^{*(r)T} \sigma_{ik}^{(r)}, \text{ if both are missing} \qquad \text{...27b}$$

where

Z = represents the matrix of rainfall depths at more than 2 sites,

z_{ij} = target site's data,

z_{ik} = any one of the control site's data,

z_i^* = is the vector of complete observations in row i of Z,

$\sigma_{ij}^{*(r)}$ = $\text{cov}(z_{ij} z_{ik})^{(r)}$,

$\mu_i^{*(r)}$ = is the subset of $\mu^{(r)}$ corresponding to z_i^*,

$\sum_{ii}^{*(r)}$ = is the covariance matrix of z_i^*, a subset of $\sum^{(r)}$,

$\beta_i^{*(r)}$ = $[\sum_{ii}^{*(r)}]^{-1} \sigma_{ij}^{*(r)}$, and

$\sigma_{jk}^{*(r)}$ = (j,k)-th element of $\sum^{(r)}$.

- **M Step**:

$$\mu_j^{(r+1)} = \sum_i^n z_{ij}^{(r+1)} / n \qquad \text{...28}$$

$$\sigma_{jk}^{(r+1)} = \sum_i^n \left[z_{ij} z_{ik}\right]^{(r+1)} / n - \mu_j^{(r+1)} \mu_k^{(r+1)} \qquad \text{...29}$$

Further details of the EMA can be obtained from Makhuva *et al.* (1997a; 1997b).

Prior to infilling missing rainfall data, outliers need to be identified and the sites grouped (Pegram, 1997b). Pegram (1997a) developed a set of routines (CLASSR) to enable a user to detect outliers and select suitable groupings of stations for the infilling of missing monthly rainfall totals. In addition, a modified version of the EMA used by Makhuvha *et al.* (1997a; 1997b) was utilised by Pegram (1997a) to create the PATCHR routines which are used to infill missing monthly rainfall totals. Currently version 5 of CLASSR and PATCHR are available.

Both the CLASSR and PATCHR programs operate on monthly time step data (Pegram, 1997b). In this study, station selection was therefore performed using data at monthly time intervals and hence the CLASSR5 program was modified to create CLASSR5A which enables data input from different data formats. Similar to the approach used by Pegram and Pegram (1993), the PATCHR5 program was modified in this study to operate on a daily time interval and the modified program has been termed PATCHR6.

The EMA technique requires the selection of suitable control stations. For the EMA procedure a classification is performed using the CLASSR program to ascertain the suitability of using the selected target and control stations for the simultaneous infilling of missing data. A procedure was thus developed to select potential control stations for each target station and is discussed in the following section.

4.2 Selection of Initial Control Stations

For all 3 945 daily rainfall stations in southern Africa which were extracted from the daily rainfall database housed by the CCWR and which have 20 or more years of continuous records, the Euclidean Distance (*ED*) between each target station and all other potential control stations was calculated. As shown in Equation 30, the characteristics used in the calculation of *ED* were the distances between the target and all potential control stations as well as the differences in mean annual precipitation and altitude of the target and all potential control stations. An index of the overlapping years of record computed between the target and control stations, as shown in Equation 31, is also included in *ED*. These characteristics were normalized such that the range of each characteristic lay in the range 0 to 1.

$$ED = \sum_{i=1}^{4} W_i Y_i \qquad \ldots 30$$

where

ED = Euclidean Distance,

W_i = weight assigned to i-th characteristic, distance (i = 1), MAP (i=2), altitude (i=3) and overlapping record (i=4), and

Y_i = i-th normalized characteristic.

$$OR = 1 - \frac{MAX(T_s, C_s) - MIN(T_e, C_e)}{T_e - T_s} \qquad \ldots 31$$

where

OR = index of overlap of records at target and control stations,

T_s = start year of record for target station,

T_e = end year of record for target station,

C_s = start year of record for control station, and

C_e = end year of record for control station.

The performance of the EMA has been shown by Smithers *et al.* (1999) to be sensitive to the weights assigned to the characteristics. Based on the initial results by Smithers *et al.* (1999) all W_i were set to 1, while an additional requirement was that the distance between the target and control stations had to be less than 50 km. Thus, for each target station an initial set of control stations, with a maximum of 9 control stations, was selected. In cases where the maximum physical distance between any of the 9 control stations and the target stations exceeded 50 km, the number of control stations was reduced accordingly.

4.3 Phasing of Daily Rainfall Data

One of the problems associated with the infilling of missing daily rainfall data, and which is not applicable when infilling missing monthly rainfall totals, is that the daily rainfall total for the same event may be incorrectly recorded by some observers and appear in the records as occurring on different days at adjacent or nearby stations, i.e. some observers record the rainfall measured at 08:00 as occurring on the previous day whilst other observers may record the total for the 24 h period ending at 08:00 against the date for the current day. Hence a so-called "phase" problem is introduced into the data. This phasing problem has previously been identified in South African daily rainfall data by Schulze (1980) and Meier (1997) as it becomes important when modelling runoff from a distributed catchment configuration using data from a number of different daily raingauges. In addition, the phasing problem could lead to erroneous relationships between stations being developed by the EMA and could thus influence the infilled values. It is for the above reasons that the phasing problem was addressed in this study.

From daily rainfall observations at adjacent stations, it was noted that the phase shift in daily rainfall data was not always consistent when viewed over long periods of time, i.e. adjacent or nearby stations may have certain periods in their records where the data are out of phase and other periods where the data are in phase. The inconsistency over time may be due to a change in one of the observers or a change in the observation procedure by one of the observers at some point in the record. Although it may be argued that the phasing problem is not an error and may be explained by the random nature of rainfall in space and time, the systematic nature of this phasing error confirms that the data were, in the majority of cases, recorded on the incorrect day by one of the observers. In an attempt to automate the correction of the phasing error for the purposes of infilling missing values using the EMA, the following procedure was implemented:

- Daily rainfall data from the target and 9 control stations, selected as described above, were aligned for each calendar day.
- Beginning at the first record and working sequentially to the last record of the observed series, rainfall events, which had at least one day with no rain at the start and end of the event, were identified at each of the control stations.

- In the case of missing rainfall data at the target station for the identified event, the first control station which did not have missing data was used as the "target" station for that event for the purposes of phasing the data at the remaining control stations.
- For each event and at each control station, the sum of the differences between the control and target stations were computed for three options. These were for:
  - no shift in the data,
  - lagging the control data for the event by one day or
  - moving the control data for the event forward by a day.
- The summed differences for the each of the three possibilities were compared and the option with the lowest difference between the control and target station for the event was implemented.

An example of this automated procedure to correct for phase errors in daily rainfall data is shown for Station 0239482 A (Cedara) in Table 1.

4.4 Outlier Detection

When collecting hydrometeorological data it is inevitable that errors will occur in the data sets. In daily rainfall data, in addition to the phasing errors discussed in the previous section, errors in recorded rainfall amounts may be due to incorrect recording of the rainfall depth by the observer or due to errors introduced when the data are entered into an electronic form. An example of such an error may be the incorrect placement of the decimal point for the rainfall on a particular day, as has been illustrated for extreme events in South Africa by Schulze (1984). One method of attempting to identify such errors is to investigate inconsistencies between the data from stations which are relatively close to each other.

Table 1 Example of automated phasing correction of daily rainfall at Station 0239482 A

Year	Month	Day	Rainfall (mm*10)				
			0239482 A	0239421 W		0269477 A	
			Original	Original	Phased	Original	Phased
1919	10	15	0	0	0	0	0
1919	10	16	0	0	0	0	0
1919	10	17	229	0	254	356	356
1919	10	18	0	254	15	69	69
1919	10	19	15	15	0	0	0
1919	10	20	18	0	0	20	20
1919	10	21	0	0	0	0	0
1919	10	22	160	0	0	10	10
1919	10	23	0	0	0	0	0
1919	10	24	0	0	0	0	0
1919	10	25	38	0	43	86	86
1919	10	26	3	43	0	0	0
1919	10	27	15	0	18	84	84
1919	10	28	0	18	0	8	8
1919	10	29	0	0	0	0	0
1919	10	30	3	0	0	25	25
1919	10	31	3	43	43	0	0
1919	11	1	0	0	0	0	0
1919	11	2	38	0	33	69	69
1919	11	3	109	33	64	117	117
1919	11	4	5	64	0	10	10
1919	11	5	5	0	0	0	0
1919	11	6	5	0	13	15	15
1919	11	7	79	13	84	91	91
1919	11	8	3	84	0	0	0
1919	11	9	0	0	0	0	0
1919	12	1	0	8	8	0	0
1919	12	2	0	8	8	76	76
1919	12	3	0	0	0	0	0
1919	12	4	8	0	15	0	0
1919	12	5	25	15	23	38	38
1919	12	6	13	23	10	30	30
1919	12	7	401	10	320	165	165
1919	12	8	0	320	0	0	0
1919	12	9	3	0	0	0	0
1919	12	10	46	0	25	66	66
1919	12	11	15	25	15	30	30
1919	12	12	66	15	20	114	114
1919	12	13	145	20	140	241	241
1919	12	14	25	140	23	91	91
1919	12	15	89	23	66	41	41
1919	12	16	20	66	0	178	178
1919	12	17	8	0	15	84	84
1919	12	18	0	15	0	0	0
1919	12	19	0	0	0	0	0

The concept of the covariance biplot is useful in identifying rainfall data which are strongly correlated and for identifying outliers (Basson et al., 1994; Pegram, 1997a; Pegram, 1997b). An example of the station year biplot, produced using the CLASSR routines developed by Pegram (1997a) for the months of January and December for Station 0239482 A and nine control stations is shown in Figure 5. Stations with apparent outlier data are far removed in the biplot from the remainder of the stations.

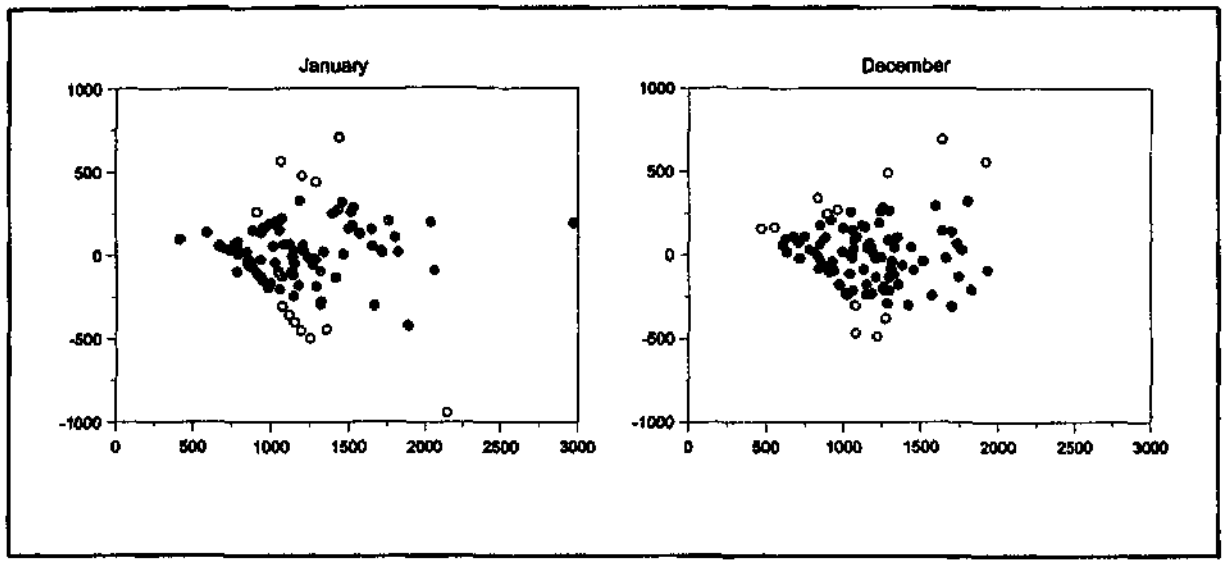


Figure 5 Examples of station year biplots at Station 0239482 A, where circles indicate potential outlier points

The routines developed by Pegram (1997a) are interactive and require the user to manually identify outliers in the data using the output produced by the program (e.g. as shown in Figure 5). In order to automate the detection of outliers, the following procedure was implemented using monthly rainfall totals:

- For each year in the station year biplot the angle α relative to the origin (0,0) of the line drawn through the point, as shown in Figure 6, was calculated.

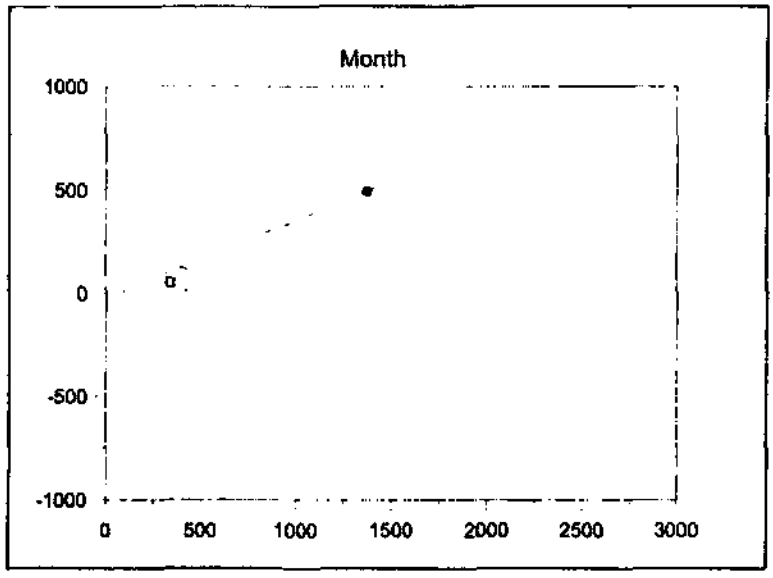


Figure 6 Schematic diagram of angle calculations in covariance biplot

- Outliers in the X and Y directions and angle values of the station year biplot were identified by computing upper and lower cutoff limits (C_U and C_L respectively) as presented by Basson *et al.* (1994) and shown in Equation 32.

$$C_U = F_{75} + 1.5\,(F_{75} - F_{25}) \qquad \text{...32}$$
$$C_L = F_{25} - 1.5\,(F_{75} - F_{25})$$

where

F_x = x-th non-exceedence percentile of the data, i.e. $F_{75} - F_{25}$ is the interquartile range either side of the median.

- An additional cutoff limit for the angles (C_A) was calculated as

$$C_A = M_A + 1.5\,SD_A \qquad \text{...33}$$

where

M_A = mean of the angles of all points in the biplot, and

SD_A = standard deviation of angles in the biplot.

- Outliers in monthly rainfall data were identified using the station year biplot when:
  - the point was both an outlier in the X-direction and the angle of the point was an outlier, or

- the point was both an outlier in the Y-direction and the angle of the point was an outlier, or
- if the absolute value of the angle exceeded C_A.

- When the data for a particular month and year at a station were identified as an outlier using the station year biplot, then outliers in monthly totals of rainfall were established.
- If outliers in the monthly totals of rainfall were found, then a further analysis on the daily rainfall data was pursued.
- If more than a single high (or low) outlier was identified within the monthly rainfall totals, then identification and exclusion of high (or low) outliers in the daily rainfall data was not performed.
- In order to identify outliers in the daily rainfall data for a given year and month, which had been identified as an outlier using the station year biplot and which had a single high (or low) monthly rainfall total, rainfall totals were computed at each station for a moving window, which increments by one day at a time and which is four days wide. Outlier values (i.e. stations) in the rainfall total for the 4 day window period were then identified. If a day at a particular station was identified as a high (or low) outlier in all the windows in which it appears, and the monthly rainfall total was also identified as a high (or low) monthly rainfall total, then the data for that day and station were flagged as an outlier and excluded in subsequent infilling procedures.

4.5 Infilling Procedure

After an initial rough infilling of the missing monthly rainfall totals had been performed to enable further analysis of the data, the CLASSR program (Pegram, 1997a) provides output to assist in identifying stations which have similar characteristics and which can be reasonably used to jointly infill missing data at the stations. This output consists of station vs months and stations vs years biplots, for both months and stations, as well as a cluster analysis of similar stations for each month. Generally it was found that the grouping of stations identified by the biplots and cluster analysis corresponded reasonably well.

In order to automate the identification of suitable control stations to infill the target station, the following procedure was implemented:

- Nine initial potential control stations were selected for each target station, as described in Section 4.2.
- A cluster analysis of the target and 9 control stations for annual average and each individual month is output from the CLASSR programme (Pegram, 1997a). Using this output, the number of times that the potential control stations had the same cluster membership as the target station was counted, and the stations were ranked according to this total.
- All potential control stations which were identified as having the same cluster membership as the target station for all 13 cluster analyses (annual average and 12 individual months) were adopted as control stations, irrespective of the final number of control stations.
- If fewer than four control stations were identified using the above procedure, all stations having a rank ≤ 4 were adopted as control stations (i.e. at least 9 of the 13 cluster analyses had the same cluster membership as the control station).
- Using the EMA algorithm (Makhuvha et al., 1997a), as implemented by Pegram (1997a), missing data in the target and control stations were then infilled simultaneously. In this implementation only the infilled values from the target station were retained, as it was postulated that missing data in the control station may be infilled better, possibly by using more suitable control stations, when the control station was considered as the target station. However, in the event that for a particular target station one of more of the control stations had already been infilled (i.e. they had previously in the analysis already been considered as target stations), then the infilled values were used to infill the current target station under consideration.

Although some cognisance is taken of overlapping records in the selection of initial potential control stations, some missing daily rainfall data were still found after the above procedure had been implemented. In such cases, the remaining missing data were infilled using the first non-missing data encountered in the initial control stations (9), and adjusting the daily infilled value using the ratio of the median monthly rainfall values for the two stations. Median monthly

rainfall values were computed from the observed daily rainfall data, with months which had missing daily rainfall data excluded from the analysis.

Once the missing daily rainfall data had been infilled using the EMA, the RLMA was implemented. The first step in the RLMA is to identify homogeneous regions and this aspect is addressed in Chapter 5.

CHAPTER 5

REGIONALISATION OF DAILY RAINFALL

A procedure similar to that used by Smithers and Schulze (1998) was adopted for the regionalisation of the daily rainfall stations into relatively homogenous regions for the estimation of design rainfalls. This approach was based on the Regional L-Moment Algorithm (RLMA) developed by Hosking and Wallis (1993; 1997), which identifies potentially homogeneous regions by a cluster analysis of site characteristics and then tests the homogeneity of the region using the statistics of the sites in the region.

5.1 Identification of Homogeneous Daily Rainfall Regions

The measures of discordancy (D) and heterogeneity (H) developed by Hosking and Wallis (1993; 1997) and described in Sections 2.2.1 and 2.2.2, were used to identify anomalies in the data and test for homogeneous regions respectively. These tests have been used successfully in South Africa by Smithers and Schulze (1998) in the regionalisation of short duration rainfall frequency distributions. A station is considered to be discordant with the rest of the group if $D > 3$ and a cluster of stations is "acceptably" heterogeneous if $H < 2$. In the selection of stations to be used in the regionalisation procedure, a compromise between record length and distribution of stations resulted in the selection of stations which have at least 40 years of record. The distribution of the 1806 daily rainfall stations in South Africa which have at least 40 years of record is shown in Figure 7. The L-CV and L-skewness of the 1806 daily rainfall stations are shown in Figure 8. Stations which have $D > 3$ (i.e. are discordant) are circled in Figure 8. The data for all 1806 stations considered together are clearly very heterogeneous with $H = 39.3$, thus indicating that further subdivision is necessary to achieve relatively homogeneous rainfall regions.

Ten of the 1806 rainfall stations with record lengths of at least 40 years were excluded from the regionalisation procedure. These 10 hidden stations were then used to independently evaluate the performance of the RLMA.

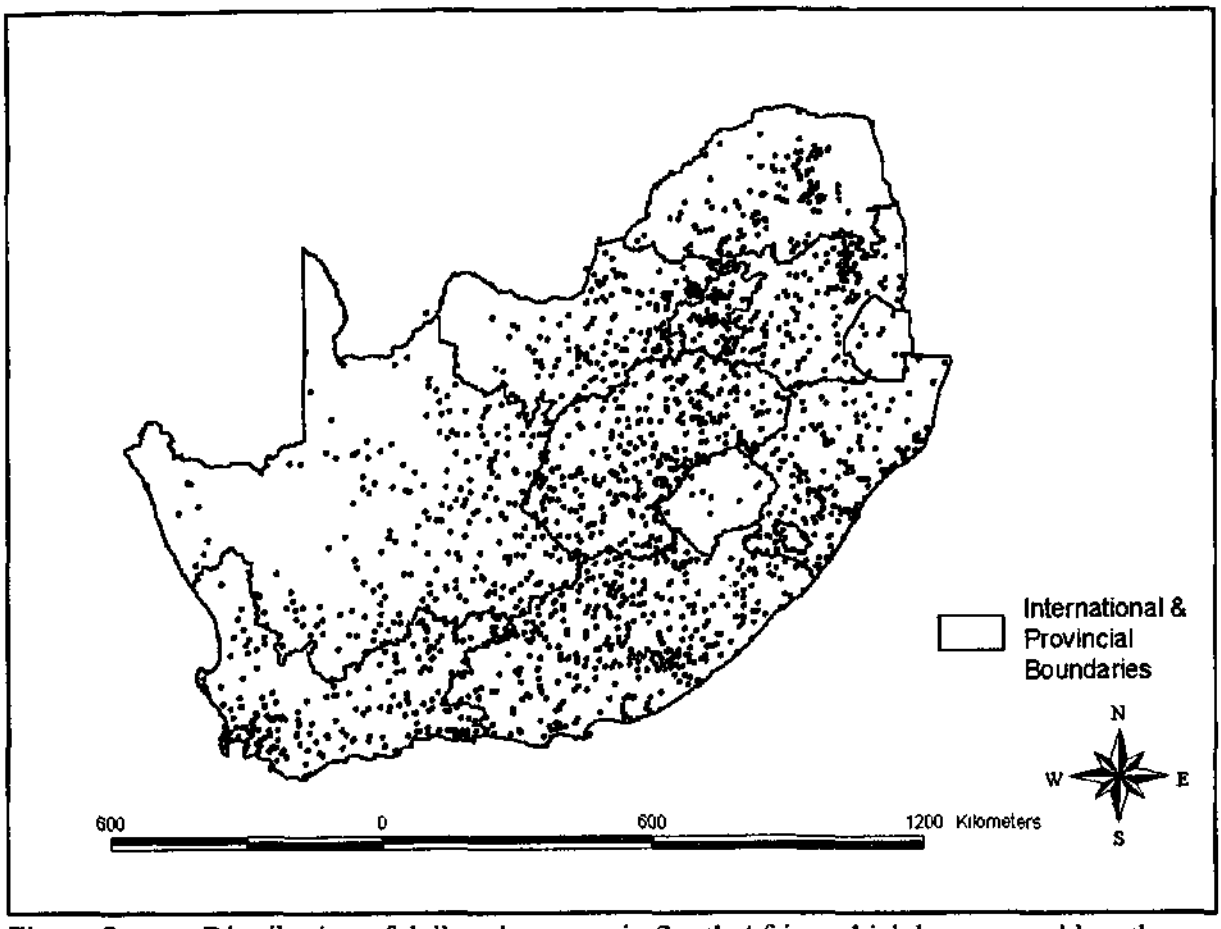


Figure 7 Distribution of daily raingauges in South Africa which have record lengths ≥40 years

The random distribution in South Africa of the stations (71) which have a discordancy index > 3, as shown in Figure 9, does not reveal any regional bias in the discordancy, i.e. all the discordant stations do not occur in similar climatic or geographic regions. In addition, a similar analysis performed on each of theses discordant stations, using nearby stations, indicated that these stations were consistent with the surrounding stations and were therefore included in subsequent analyses.

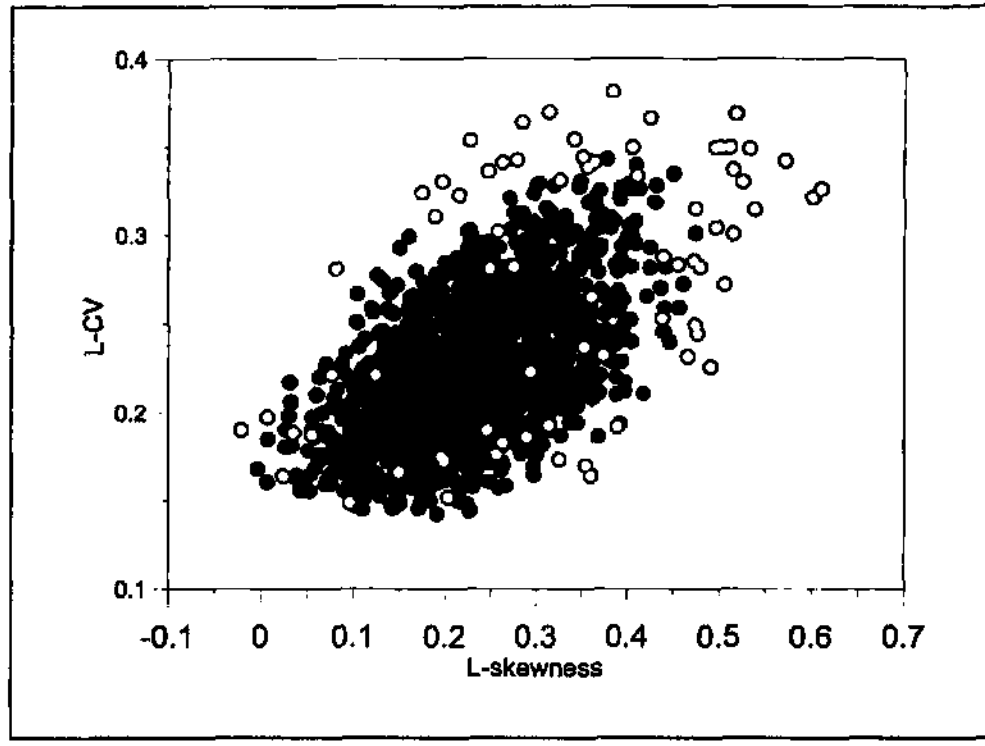


Figure 8 L-moment ratios for 1806 daily rainfall stations in South Africa which have at least 40 years of record (circles indicate discordant stations)

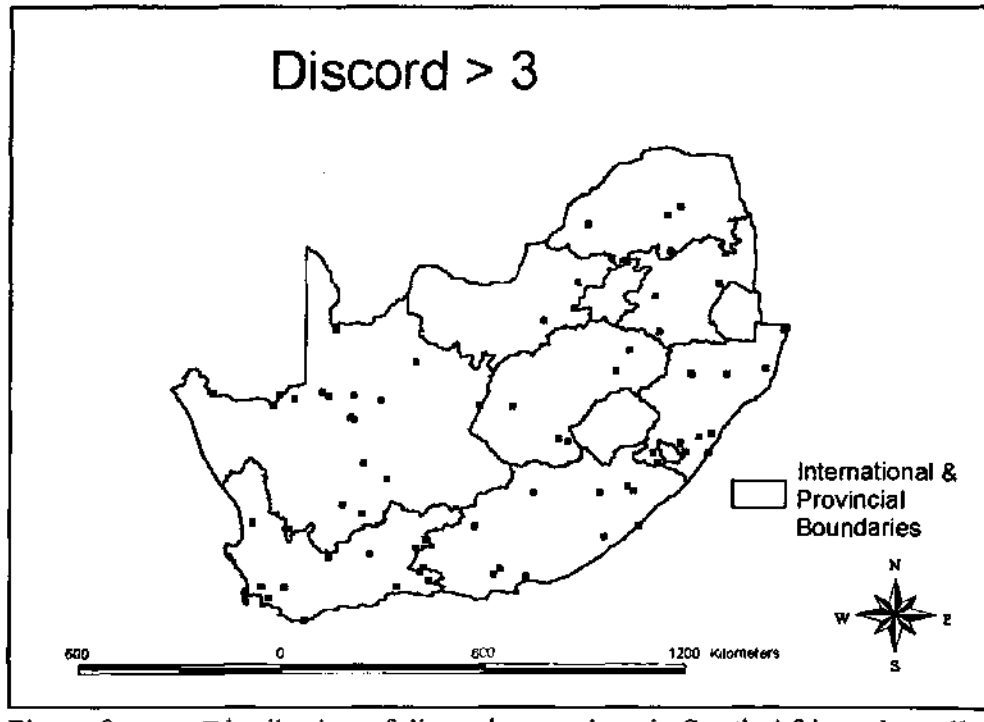


Figure 9 Distribution of discordant stations in South Africa when all stations which have record lengths ≥40 years are considered as a single region

5.2 Results of Cluster Analyses

Subdivision of South Africa was achieved by a cluster analysis of site characteristics, using Ward's minimum variance hierarchical algorithm (SAS, 1989), which tends to form clusters of roughly equal size (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 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).

The site characteristics used in the cluster analysis were:

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

All site characteristics were transformed to lie in the range between 0 and 100, as the cluster analysis is very sensitive to the Euclidian distance or scale (Hosking and Wallis, 1997).

The number of clusters to create is a subjective decision. Simulation results by Hosking and Wallis (1997) indicate that very little improvement in the accuracy of the regional growth curves for return periods < 1000 years is achieved with more than 20 stations per cluster. Using the 1806 daily rainfall stations in South Africa which have at least 40 years of daily rainfall record, the mean and Standard Deviation (SD) of the heterogeneity measure (H), computed for each of the clusters, and for the number of clusters ranging from 15 to 150 is shown in Figure 10. Smithers and Schulze (1998), in a regionalisation of extreme short duration rainfall data, identified 15 homogeneous clusters in South Africa and hence 15 clusters was used as a starting

point. From Figure 10 it is evident that when fewer than 60 clusters were formed, the mean of the H values was greater than 2, which is the upper threshold for acceptably heterogeneous clusters. In addition, a local "minimum" in the mean and SD of the H values is apparent for 60 clusters. Hence, initially 60 clusters were formed, of which 24 clusters were still unacceptably heterogeneous ($H > 2$) and thus further clustering and subjective adjustments were necessary. The number of stations per cluster in the 60 clusters ranged from 4 to 85 and, as shown in Figure 11, the H value was not affected by the number of stations in the cluster.

In each of the clusters which were unacceptably heterogeneous ($H>2$), the stations were further divided into two or more clusters using the clustering analysis procedure of site characteristics, and this process was continued until all the newly created clusters were relatively homogeneous ($H < 2$). This process resulted in 113 clusters, all of which were acceptably heterogeneous ($H < 2$), and the number of stations per cluster ranged from 2 to 74.

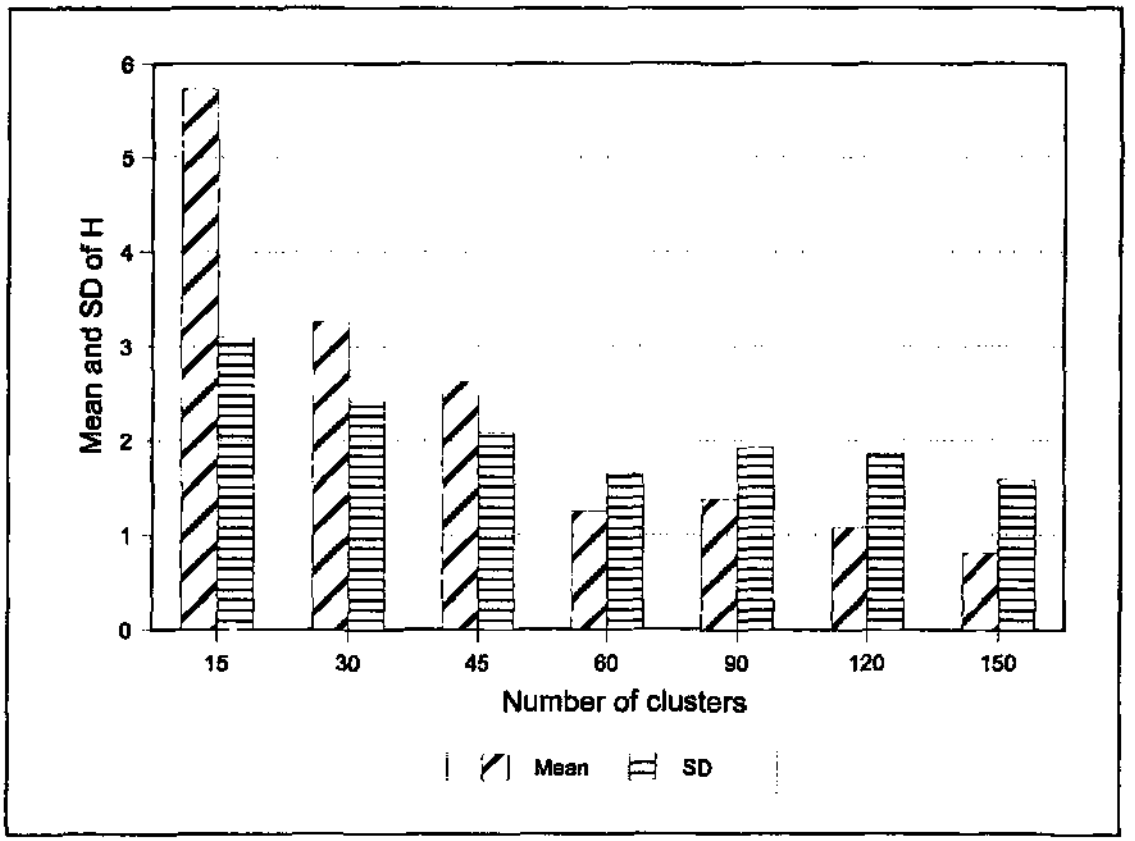


Figure 10 Mean and standard deviation of heterogeneity measure (H) for different number of clusters

In addition to the 10 stations hidden from the regionalisation for the purposes of assessing the performance of the RLMA, a further 8 stations were excluded which displayed significant trends in the annual rainfall totals and which were discordant from the surrounding stations. Thus 1 789 stations were used to create the 113 clusters. Fourteen clusters contained 3 or fewer stations. Thirteen of these clusters were joined to adjacent clusters to form larger clusters, which were still acceptably heterogeneous, resulting in102 relatively homogeneous clusters. The spatial distribution of the stations making up each of the102 relatively homogenous clusters indicated some overlap of stations at adjacent clusters. While this spatial overlap is acceptable since the clustering is based on 7 characteristics and the clusters are relatively homogenous, further clustering of stations was performed to make the clusters more physically coherent. Hence, clusters with overlapping stations were joined and a cluster analysis was performed using the pooled stations to create 2 or more clusters. Thus stations were not moved subjectively between clusters but the selection of clusters to pool together was based on the physical coherence of the initial clusters. This re-clustering process was continued until reasonable physical coherence of clusters was attained. The number of clusters at this point was 78 and the number of stations per cluster ranged from 3 and 66 stations with an average of 23, as shown in Figure 12. The distribution of the 78 relatively homogeneous clusters in South Africa is shown in Figure 13.

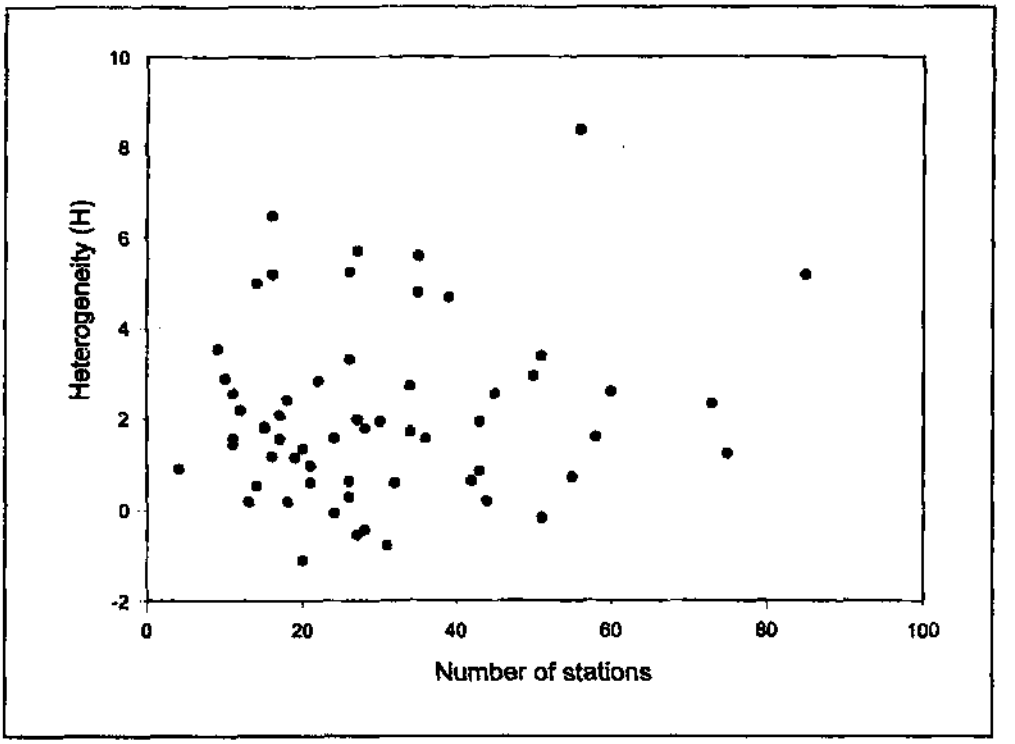


Figure 11 Heterogeneity (H) vs number of stations for 60 clusters

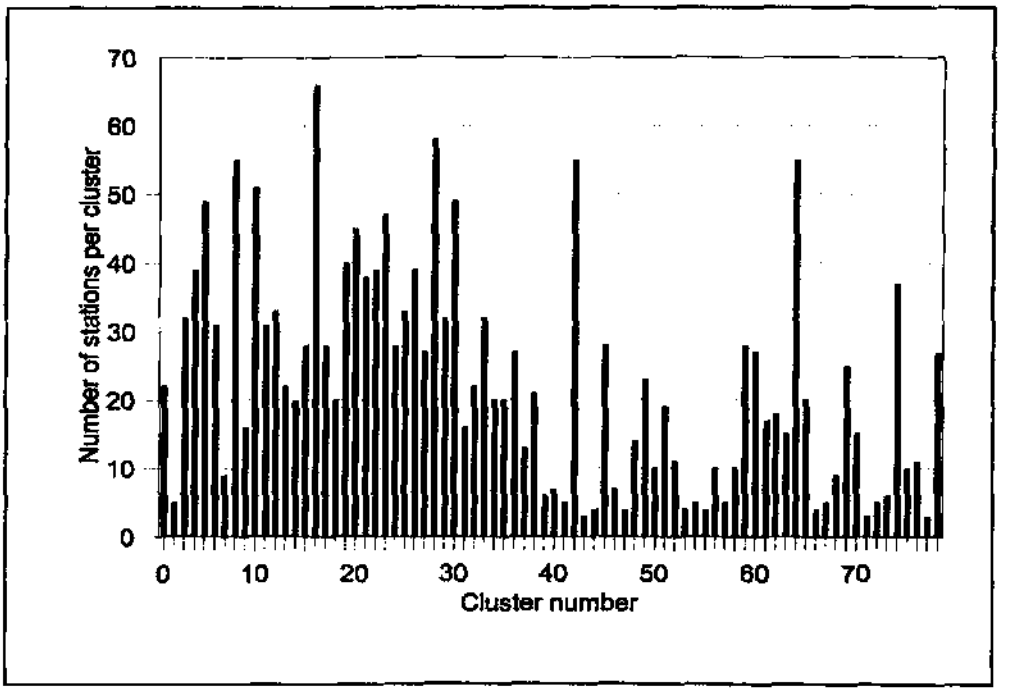


Figure 12 Number of stations per cluster in 78 relatively homogeneous clusters

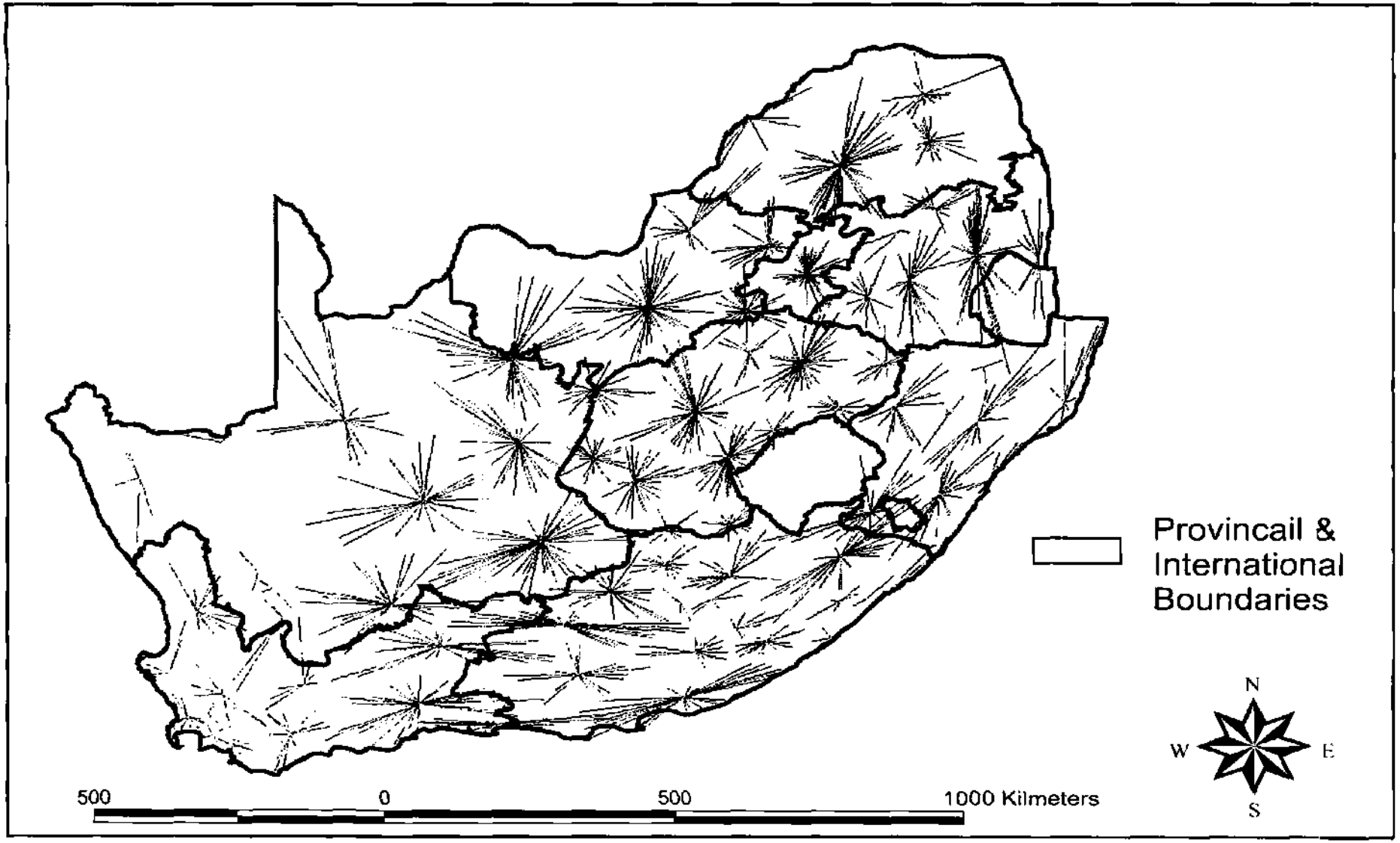


Figure 13 Distribution of 78 relatively homogeneous daily rainfall clusters in South Africa

CHAPTER 6

ESTIMATION OF DESIGN RAINFALL

Once relatively homogeneous rainfall regions have been identified, the next step in the RLMA is the selection of an appropriate probability distribution to be used in the frequency analysis. The selected distribution was then used to estimate regional quantile growth curves for each cluster and hence at-site design rainfall values were estimated. The accuracy of these design values were then assessed and comparisons performed with design values previously used in South Africa.

6.1 Choice of Frequency Distribution

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. This procedure is described in Section 2.2.3.

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 Z statistic is computed as shown in Equation 34. According to Hosking and Wallis (1997) the fit is adequate if Z is "sufficiently close to zero" and they suggest that $|Z| \leq 1.64$ is a reasonable criterion to indicate that the fit of the assumed distribution is adequate. A formal definition of the statistic is presented in Section 2.2.3.

$$Z = \frac{\left(t_4^R - t_4^{PD}\right)}{\sigma_4} \qquad \ldots 34$$

Five probability distributions were evaluated as potential candidate distributions in the frequency analysis. These were the Generalised Logistic (GLO), General Extreme Value (GEV), 3-parameter log-Normal (LN3), Pearson-III (P3) and General Pareto distributions (GPA). One option was to determine the most appropriate probability distribution in each of the 78 relatively homogeneous clusters. However, from a practical point of view it was decided to determine, for the one day duration, an appropriate distribution which is applicable to all clusters and which is then assumed to apply to longer durations as well. This approach of a single appropriate distribution for all clusters is supported by Wallis (1997). Using the *Z*-test statistic, the number of clusters in which the candidate distributions were acceptable and the best candidate distribution for each cluster was computed, with the results summarised in Table 2. Both the GLO and GEV had similar performances in terms of the number of times each distribution was selected as the best distribution for each of the clusters. However, the GEV distribution was acceptable in substantially more clusters and hence was selected as the most appropriate distribution to use in all the clusters.

Table 2 Performance of candidate probability distributions in 78 clusters

Criterion	GLO	GEV	LN3	P3	GPA
Number of clusters in which distribution is acceptable	35	52	30	11	2
Number of clusters in which the distribution is the best	33	30	11	4	0

6.2 Assessment of Regional Quantile Growth Curves

Uncertainty is inherent in any 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.

Regional growth curves for each duration were developed for each cluster and relate the ratio between design rainfall and an index value to return period. Examples of growth curves for selected clusters and design rainfall values estimated using the growth curves are shown in this section. The GEV distribution, which is shown in Section 6.1 to be an appropriate distribution for extreme 1 day rainfall in South Africa, was used to estimate the design storms.

6.2.1 Accuracy of estimates

The accuracy of quantile estimates were assessed by their bias and RMSE which were computed by a Monte Carlo simulation procedure as described in Section 2.2.5. For each site in each cluster and for all durations considered a random sample was generated which had 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 and duration, a region was simulated having the same number of stations, record lengths, heterogeneity 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 (1996) were used to implement the procedure. For each of the 100 repetitions, the errors in the simulated growth curve and quantiles were calculated and then accumulated and averaged to estimate the bias and Root Mean Square Error (RMSE).. Thus, 90 % error bounds 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 % error bounds for the regional quantile growth curve for Clusters 25 and 77 are shown in Figure 14. In order to estimate an at-site design rainfall depth, the regional quantile growth curve was re-scaled by the at-site mean of the Annual Maximum Series (AMS), which

is equivalent to the first L-moment (L_1). Thus, by using the error bounds of the quantile growth curve, error bounds in the design estimate may be obtained.

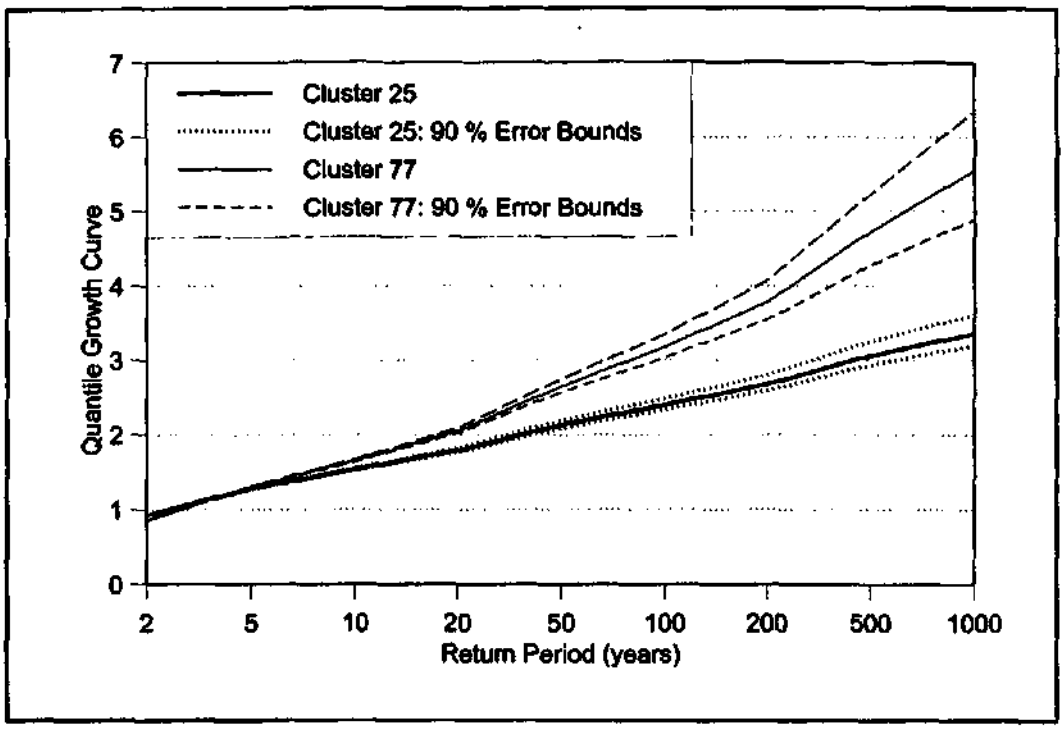


Figure 14 Examples of estimated regional growth curves and their 90 % error bounds

6.2.2 *At-site vs regional quantiles*

In order to assess the performance of the RLMA, 10 daily rainfall stations which cover a range of climatic regions in South Africa were excluded from the regionalisation. Each of these stations was allocated to the cluster with the closest Euclidean distance between the site characteristics of the station and the mean of the site characteristics of all sites within a cluster. The location of the hidden stations is shown in Figure 15 and cluster numbers determined for each of the hidden stations are listed in Table 3.

Table 3 Hidden stations and cluster numbers

Station	Name	Cluster
0021055_W	Cape Town Maitland	51
0059572 A	East London	4
0144899 W	Middleburg	6
0239482 A	Cedara	15
0261368 W	Bloemfontein	10
0299357 W	Cathedral Peak Hotel	17
0317447AW	Upington	35
0442811 W	Nooitegedacht	24
0513404 W	Pretoria	16
0677834 W	Pietersburg	28

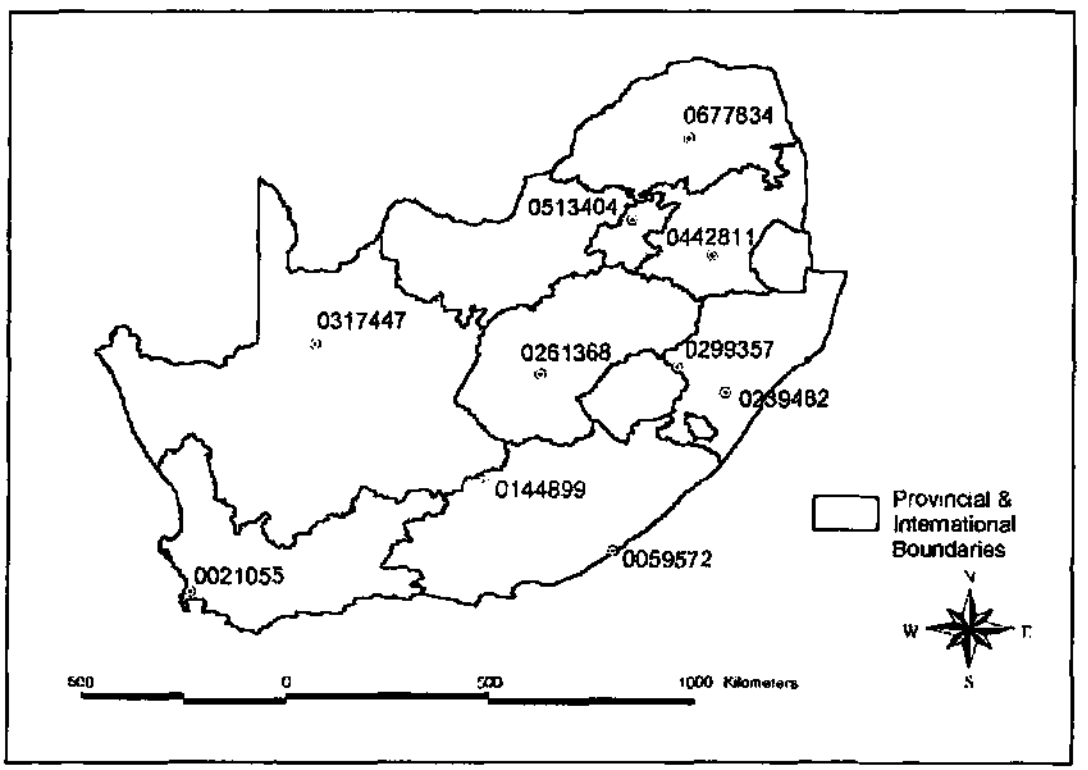


Figure 15 Location of the 10 hidden daily rainfall stations in South Africa

A comparison between the design rainfall estimated using the at-site data and and estimated from the regional quantile curve is shown in Figure 16 for the 10 hidden stations which were not

used in the regionalisation procedure. Included in Figure 16 are the 90% error bounds of the design values estimated from the error bounds of the quantile growth curve.

As shown in Figure 16, the 1 day design rainfall depths estimated from the observed data and from the regional growth curve are similar for return periods up to 20 years and, with the exception of three stations (0021055 W, 0239482 A and 0513404 W), the values estimated from the regional growth curve generally exceed the values estimated from the at-site data for return periods greater than 20 years. The regional growth curve pools information from stations within a relatively homogeneous region and is thus considered to result in more reliable estimates of design rainfall than values estimated directly from the at-site data. Hence the recommended design values estimated using the regional growth curve are generally more conservative for longer return periods than those estimated directly from the at-site data.

6.3 Estimation of One Day Design Rainfall Depths for South Africa

Ninety per cent error bounds in the quantile growth curves were generated for all 78 relatively homogeneous clusters in South Africa. Each of 3 945 daily rainfall stations which have more than 20 years of record were assigned to a cluster based on the minimum Euclidian distance between the site and the mean of that cluster's site characteristics. In addition, quantile growth curves for each of the 78 clusters were estimated. Thus, design rainfalls and error bounds for the design values were estimated at each of the 3 945 sites using the appropriate regional quantile growth curve re-scaled for each site using the at-site mean of the AMS (L_1) estimated from the at-site data. Examples of these results for a few stations are contained in Appendix A for durations of 1 to 7 days. The design rainfalls for all 3 945 sites are contained in Portable Document Format (PDF) on the diskettes which accompany this report.

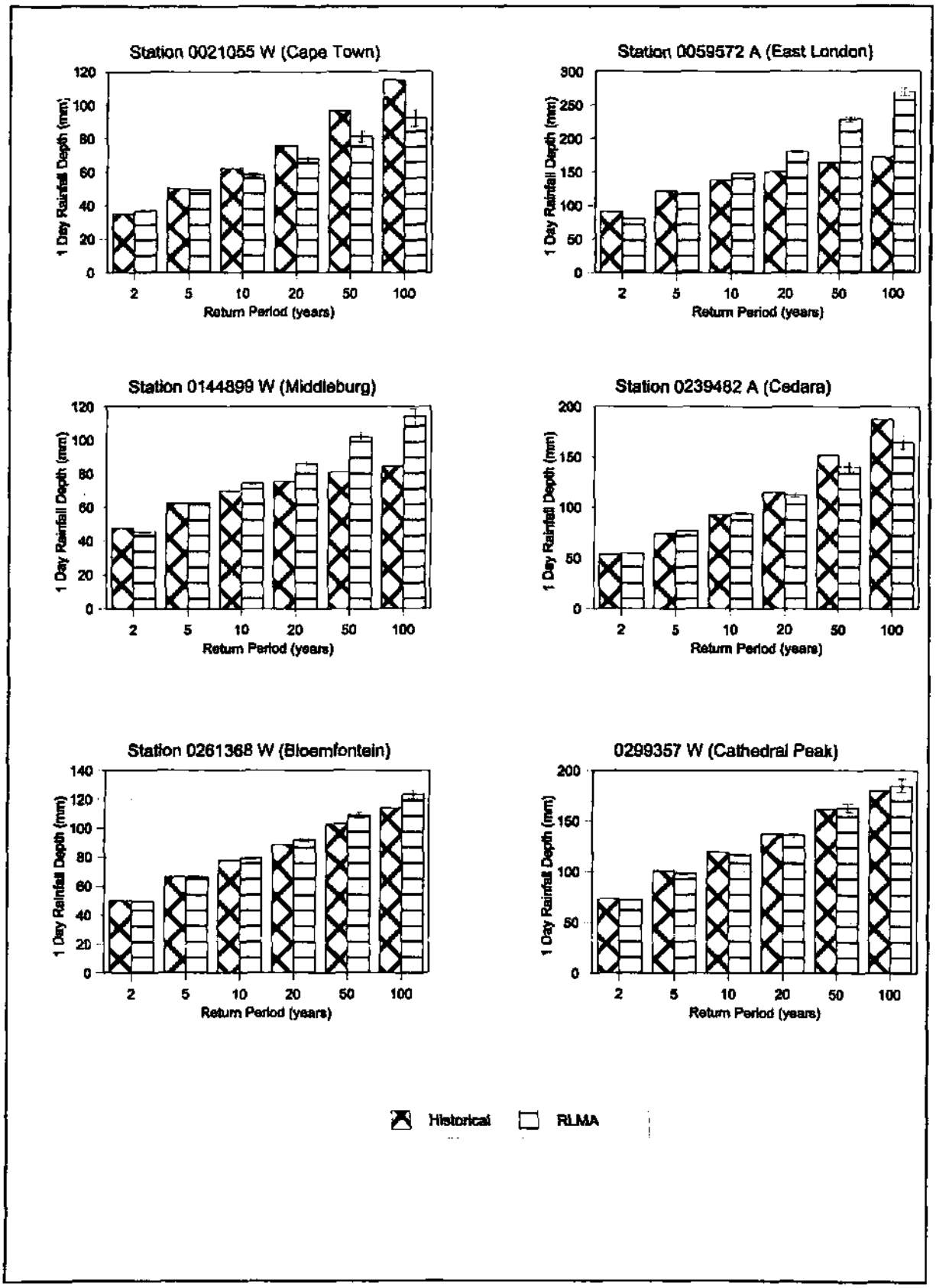


Figure 16 Comparison of design rainfall depths computed from at-site data and from regional growth curves at 10 stations not used in the regionalisation process (I-beams indicate 90% error bounds)

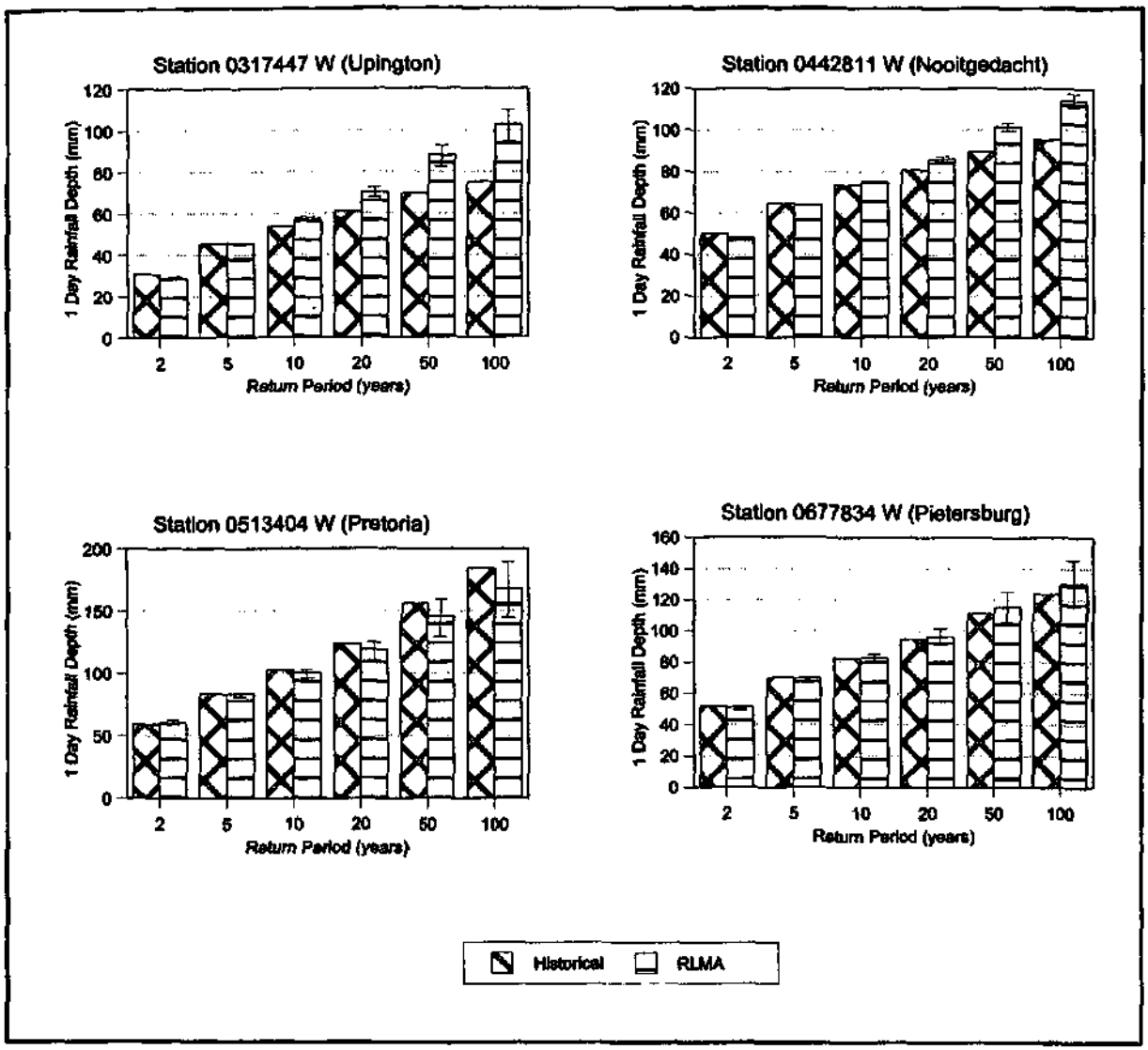


Figure 16 (cont) Comparison of design rainfall depths computed from at-site data and from regional growth curves at 10 stations not used in the regionalisation process (I-beams indicate 90% error bounds)

6.4 Comparison of Design Values with Previous Estimates

A comparison was performed between the 1 day design rainfall estimated in this study using a regional approach and those estimated by Adamson (1981). The Relative Difference (*RD*) was computed, as shown in Equation 35, between 1 day design rainfall estimated in this study and

those estimated by Adamson (1981) at 2 105 stations in South Africa and for return periods of 2 to 200 years.

$$RD_T = \frac{P_{RLMA,T} - P_{ADAM,T}}{P_{RLMA,T}} \qquad ...35$$

where

RD_T = Relative Difference for return period = T years,

$P_{RLMA,T}$ = T year return period design rainfall estimated using the RLMA and GEV distribution in this study,

$P_{ADAM,T}$ = T year return period design rainfall estimated by Adamson (1981), who used a single site approach and a censored LN distribution.

A frequency analysis was performed for the RD_T values computed at the 2 105 stations and the results are summarised in Figure 17. From Figure 17 it is evident that for return periods less than 50 years the differences between the design rainfall estimated in this study and by Adamson (1981) are less than 20 % at the majority of the stations. As expected the differences are bigger for longer return periods and for return periods $\geq$ 50 years there is a definite trend with the Adamson design values exceeding the values computed in this study. The differences in the design rainfall values estimated in the two studies may be attributed to the following factors:

- The longer record lengths used in this study.
- The stringent data quality control procedures used in this study.
- The different approaches to design rainfall estimation used in the two studies:
  - Adamson (1981) used a single site approach with a censored LN distribution.
  - This study used a regional approach and adopted the GEV distribution.
- L-moments used in this study to fit the GEV distribution are less influenced by outliers in the data.

As shown in Figure 16, design rainfall depths computed using the regional approach generally exceed the values computed directly from the at-site data. In addition, the regional approach has

been shown in many international studies to result in more reliable and robust estimates compared to design values computed using only single at-site data. Thus, it is postulated that the design values computed in this study may be used with confidence.

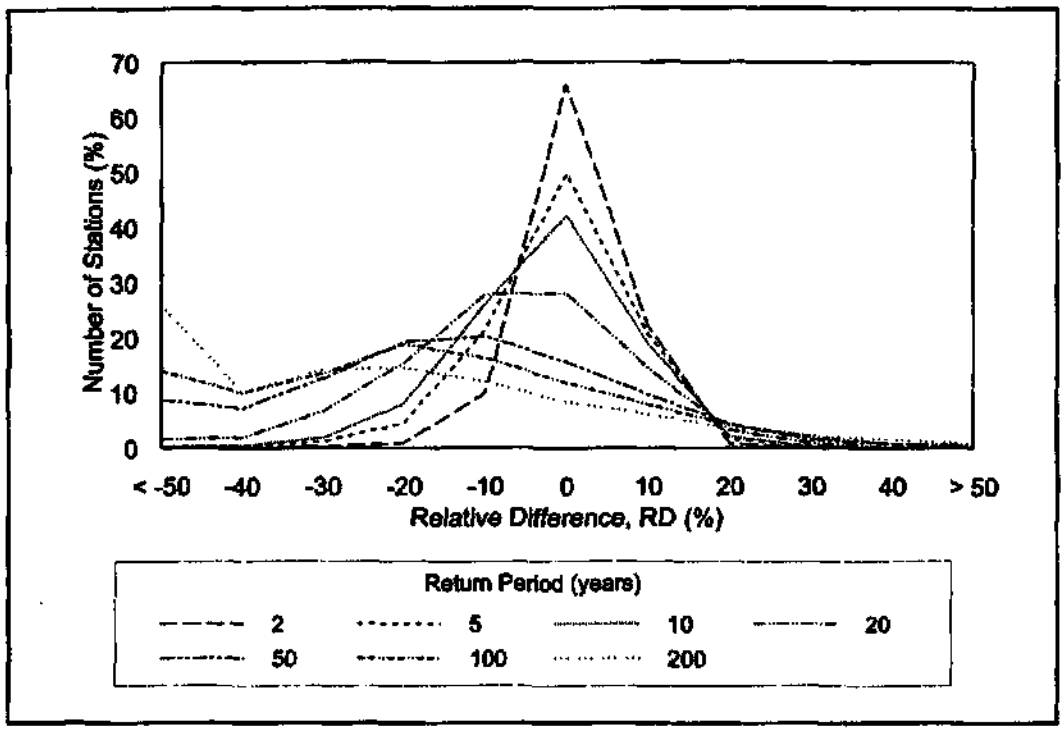


Figure 17 Comparison between 1 day design rainfall estimated in this study and values estimated by Adamson (1981)

CHAPTER 7

CONCLUSIONS AND RECOMMENDATIONS

Design rainfall depths for durations of 1 day and longer were last estimated for South Africa in the early 1980s using data up to the late 1970s. Thus nearly 20 additional years of record are currently available for analysis at many stations and many stations which did not qualify for inclusion in previous studies are now included. In addition, the growing acceptance that regionalised approaches to frequency analyses results in more reliable and robust design values necessitated the revision of design rainfall depths for durations of 1 day and longer.

A relatively dense network of daily raingauges with long records are available in South Africa. The daily rainfall database maintained by the Computing Centre for Water Research was used in this study and no additional data were acquired during this project. One shortcoming of this database is that data from the Institute of Soils, Climate and Water (ISCW) were last updated in approximately 1985 and this study would have benefitted with up to date data from the ISCW. The available record lengths were deemed to be sufficient for a study of this nature with more than 1800 raingauges in South Africa having record lengths longer than 40 years.

It was noted that, for the approximately 4 000 daily raingauges which have record lengths longer than 20 years, more than 20% of the raingauges had at least 10% of the data missing in the rainy season. Hence, a considerable effort went into developing and adapting techniques to infill the missing data.

The procedures developed by Pegram (1997b), which operate on monthly rainfall totals and utilise the EMA algorithm to infill missing data, were modified to operate on a daily time step. The routines were, furthermore, automated and numerous data checks were built into the programs. For example, outliers in the daily rainfall data were only discarded if the monthly total was also considered an outlier, using both the covariance biplot and conventional threshold limits. A further outlier test on daily values was then performed (i.e. after the total for the month had been identified as an outlier) and daily values were only discarded if the value was identified as an outlier in each of the 4 days of a moving 4 day window. The selection of control

stations was automated in a two stage process where the initial control stations were selected based on distance, MAP, altitude and overlapping years of record and the second stage discarded any of the initial control stations if they did not appear to fit the character of the target and remaining control stations.

The regional index value approach based on L-moments adopted in this study to estimate design rainfall depths was successfully implemented. The clustering of stations using site characteristics enables the independent testing of the cluster of sites for homogeneity. Data from approximately 1800 sites with ≥ 40 years of record were used to identify 78 relatively homogeneous daily rainfall clusters in South Africa. This is substantially more than the 15 clusters identified by Smithers and Schulze (1998) in a study of short duration design rainfalls. Many more daily rainfall sites were available for this study with a far denser distribution in South Africa than the sites available for the short duration design rainfall study. Hence, the larger number of stations used in this study enabled a far more detailed investigation of rainfall regionalisation and used data that are more reliable than those used in the short duration design rainfall study.

The allocation of stations to a particular cluster is not unique. Further localised pooling and re-clustering of stations from adjacent clusters may result in a different number and configuration of relatively homogenous clusters. However, the 78 clusters identified are relatively homogenous and are spatially coherent and thus further clustering was not attempted.

The GEV distribution was adopted for use in all the clusters in this study and is consistent with findings in South Africa by Smithers (1996) and Smithers and Schulze (1998) and with other international design rainfall studies summarised by Smithers and Schulze (1998).

The accuracy of the quantile growth curves for each cluster was assessed using a Monte Carlo simulation procedure and 90 % error bounds were computed. The comparison at 10 sites, which were not used in the regionalisation procedure, between 1 day design rainfall depths estimated using the regionalised and at-site approaches indicated that for return periods up to 50 years the two approaches resulted in similar design values, but that design values estimated using the regionalised approach generally exceeded the at-site values for longer return periods (≥ 50 years).

Using the quantile growth curves and the mean of the annual maximum series (L_1) computed from the observed data, the 1 to 7 day design rainfall depths and their 90 % error bounds were computed for approximately 4 000 sites in South Africa. The amount of printout required to reflect all these design values is substantial and thus the design rainfall values for each station are contained in Portable Document Format (PDF) on the diskettes which accompany this report.

A comparison between the 1 day rainfall depths estimated in this study and by Adamson (1981) indicated that the design values were similar between the two studies for return periods < 50 years. However, for longer return periods the design depths estimated by Adamson (1981) were generally larger than the values generated in this study. Some of the differences in design values estimated may be attributed to the different approaches taken in the two studies. Adamson (1981) used a single site approach with a censored LN distribution whereas this study adopted a regional approach with the GEV distribution. In addition, the effect of outliers in the data was reduced in this study by the use of L-moments to fit the probability distributions.

It was found in this study that design rainfall depths computed using the regional approach generally exceed the values computed directly from the at-site data. The regional approach has also been shown in numerous other studies to result in more reliable and robust estimates compared to single site point estimates. Thus, it is concluded that the design rainfall depths computed in this study may be used with confidence.

It is recommended that a user friendly front end computer program be developed to enable users to identify appropriate stations and to retrieve the information. One option would be to develop a program which could be distributed and installed on a user's computer. Another option which should be considered is the development of an interactive WWW based system where users can access the results. The WWW based system approach to the dissemination of the results enables users to access the most up to date information and does not require the re-distribution of the program to users when any updates or changes are made, which would be necessary for a distributed computer version of the program.

In order to estimate design rainfall depths at ungauged sites, it is necessary to estimate the L_1 value at the ungauged sites. Hence it is recommended that further research should investigate

estimating L_1 as a function of site characteristics, which are available at ungauged sites, hence enabling design rainfall depths to be estimated at the ungauged site.

For users without access to a computer to enable rapid access to the results, it is recommended that the point design rainfall values be used to produce an isoline design rainfall map of South Africa for each return period. The design rainfall isoline maps may be generated using the values from approximately 4 000 stations computed in this study. Another alternative would be to estimate the index value (L_1) and hence compute design rainfall values on a minute-by-minute latitude and longitude grid over South Africa and then use these gridded points to draw the isolines.

In order to apply the point design rainfall depths to estimate design floods over large catchments, it is necessary to adjust the point design rainfall values, as estimated in this study, to account for the decrease in average storm depth with increase in catchment size. The areal reduction factors currently used in South Africa are largely based on results from international studies. It is recommended that the databases and information compiled during this study and during the study on short duration design rainfall values should be used to update the storm areal reduction factors for South Africa.

CHAPTER 8

REFERENCES

Adamson, P.T., 1981. Southern African storm rainfall. Technical Report No. TR 102, Department of Water Affairs, Pretoria, RSA.

Basson, M.S., Allen, R.B., Pegram, G.G.S. and van Rooyen, J.A., 1994. Chapter 3: Hydrological Data Preparation. Probabilistic Management of Water Resources and Hydropower Systems. Water Resources Publications, Highlands Ranch, Colorado, USA, 424 pp.

Cunnane, C., 1989. Statistical distributions for flood frequency analysis. WMO Report No. 718, World Meteorological Organization, Geneva, Switzerland.

Dempster, A.P., Laird, N.M. and Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm (with discussion). Journal of Royal Statistics Society, B 39: 1-38.

Dent, M.C., Lynch, S.D. and Schulze, R.E., 1987. Mapping mean annual and other rainfall statistics over southern Africa. Report 109/1/89, Water Research Commission, Pretoria, RSA.

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.

Gabriele, S. and Arnell, N., 1991. A hierarchical approach to regional flood frequency analysis. Water Resources Research, 27(6): 1281-1289.

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.

Guttman, N.B., 1993. The Use of L-moments in the Determination of Regional Precipitation Climates. National Climate Centre, Asheville, North Carolina, USA.

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., 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., 1997. Regional Frequency Analysis: An Approach Based on L-Moments. Cambridge University Press, Cambridge, UK, 224 pp.

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.

Makhuvha, T., Pegram, G., Sparks, R. and Zucchini, W., 1997a. Patching rainfall data using regression methods. 1. Best subset selection, EM and pseudo-EM methods: Theory. Journal of Hydrology, 198: 289-307.

Makhuvha, T., Pegram, G., Sparks, R. and Zucchini, W., 1997b. Patching rainfall data using regression methods. 2. Comparisons of accuracy, bias and efficiency. Journal of Hydrology, 198: 308-318.

Markham, C.G., 1970. Seasonality of precipitation in the United States. Annals of Association of American Geographers, 60: 593-597.

Meier, K.B., 1997. Development of a spatial database for agrohydrological model application in southern Africa, Unpublished MSc dissertation. Department of Agricultural Engineering, University of Natal, Pietermaritzburg, RSA, 141 pp.

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.

Pegram, G.C. and Pegram, G.G.S., 1993. Intergration of rainfall via multiquadric surfaces over polygons. Journal of Hydraulic Engineering, 119(2): 151-163.

Pegram, G.G.S., 1997a. Patching Rainfall Data (A Guide). Report H 6/6/0194, Department of Water Affairs and Forestry, Directorate of Project Planning, Pretoria, RSA.

Pegram, G.G.S., 1997b. Patching rainfall data using regression methods. 3. Grouping, patching and outlier detection. Journal of Hydrology, 198: 319-334.

Pegram, G.G.S. and Adamson, P.T., 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.

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.

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.

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.

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, 276 pp.

Smithers, J.C., 1996. Short-duration rainfall frequency model selection in Southern Africa. Water SA, 22(3): 211-217.

Smithers, J.C., Chetty, K., Royappen, M. and Schulze, R.E., 1999. A comparison of selected techniques for infilling missing daily rainfall data. *ACRU*cons Report 29, School of Bioresources Engineering and Environmental Hydrology, University of Natal, South Africa.

Smithers, J.C. and Schulze, R.E., 1998. An evaluation of techniques for estimating short duration design rainfalls in South Africa, Department of Agricultural Engineering, University of Natal, Pietermaritzburg, RSA. Report to Water Research Commission, Pretoria, RSA. 356 pp.

Stedinger, J.R., Vogel, R.M. and Foufoula-Georgiou, E., 1993. Frequency analysis of extreme events. Handbook of Hydrology. McGraw-Hill, New York, USA.

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.

Vogel, R.M. and Fennessy, N.M., 1993. L-Moments diagrams should replace product moment diagrams. Water Resources Research, 29(6): 1746-1752.

Wallis, J.R., 1997. Personal communication. IBM Research Division, T.J. Watson Research Centre, New York, USA.

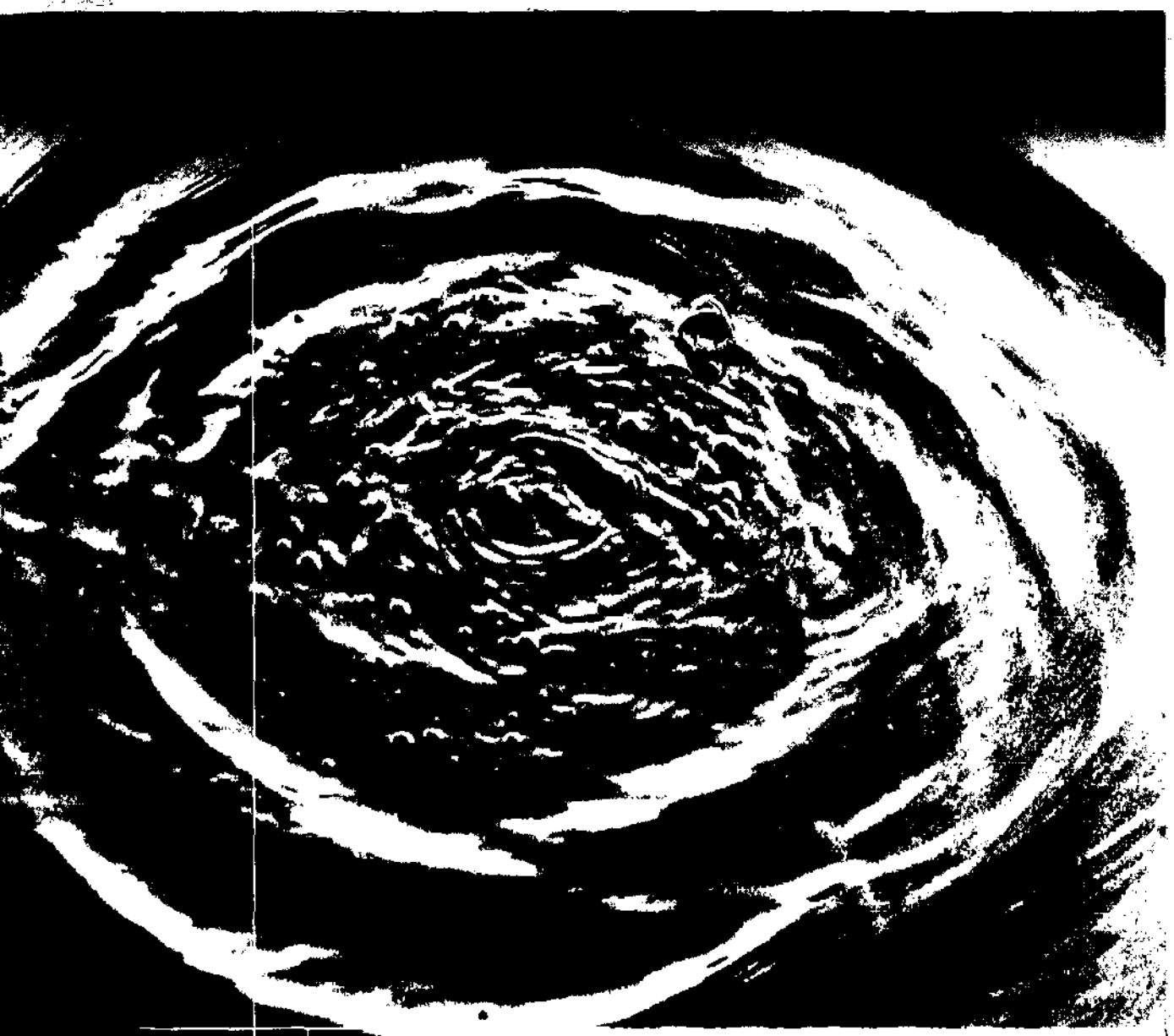
Zucchini, W. and Adamson, P.T., 1984. The occurence and severity of droughts in South Africa. WRC Report 91/1/84, Water Research Commission, Pretoria, RSA.

APPENDIX A

EXAMPLES OF DESIGN RAINFALL DEPTHS FOR ONE TO SEVEN DAY DURATIONS

L = lower 90 % error bound (mm)
D = design rainfall depth (mm)
U = upper 90 % error bound (mm)

SAWB No.	Station Name	Latitude		Longitude		M.A.P	Altitude	Years	Duration	Return Period (years)																	
										2			5			10			20			50			100			200		
		(°)	(')	(°)	(')	(mm)	(m)		(days)	L	D	U	L	D	U	L	D	U	L	D	U	L	D	U	L	D	U	L	D	U
0001517 W	DANGER POINT (VRT)	34	37	19	18	447.3	46	93	1	37	38	39	52	53	54	62	65	67	73	78	82	88	97	106	99	113	127	110	131	152
									2	48	49	50	67	68	69	79	82	85	91	97	101	107	117	126	118	133	147	130	150	170
									3	54	56	57	76	77	79	90	93	96	102	109	114	119	131	141	132	148	164	145	166	189
									4	58	60	61	80	82	84	94	98	100	107	114	118	122	135	144	133	152	165	143	169	187
									5	61	62	64	83	86	87	97	102	104	110	118	122	126	139	148	137	156	169	147	173	192
									6	63	65	66	87	89	90	101	105	107	113	122	126	129	143	152	140	160	173	149	177	195
									7	66	68	69	90	92	93	104	108	111	116	125	129	132	146	154	142	162	175	151	179	196
0001605 W	GANSBAAI	34	35	19	21	526.7	17	72	1	37	38	39	53	54	55	63	66	68	74	79	83	89	98	107	100	115	129	112	133	154
									2	48	50	51	68	69	70	80	83	86	92	98	102	108	118	127	120	134	148	131	152	172
									3	55	56	58	77	78	79	91	94	97	104	110	115	121	132	143	134	150	166	146	168	191
									4	59	60	62	82	84	85	96	100	101	108	115	120	124	137	146	135	154	167	145	171	190
									5	61	63	65	84	87	88	98	103	105	111	119	123	128	141	150	138	158	171	149	175	194
									6	64	66	67	88	90	91	102	107	109	115	123	127	131	145	154	142	162	175	151	179	197
									7	66	68	69	90	92	93	104	109	111	116	125	129	132	146	154	142	163	175	151	179	196
0001726 W	UILENKRAAL (BOS)	34	36	19	25	530	9	32	1	37	38	39	52	53	54	62	65	67	72	78	82	88	97	105	99	113	127	110	131	152
									2	46	47	48	65	66	67	77	80	82	88	93	98	103	113	122	114	128	141	125	145	164
									3	53	54	55	74	76	77	87	91	94	100	106	111	117	127	138	129	144	160	141	162	184
									4	58	60	62	81	83	84	95	99	101	107	114	118	123	136	144	134	152	165	144	169	188
									5	61	63	65	84	86	87	98	103	105	111	119	123	127	141	149	138	158	171	148	175	194
									6	64	66	67	88	90	91	102	107	109	115	123	127	131	145	154	142	162	175	151	179	197
									7	67	68	70	91	93	94	105	110	112	118	126	130	134	148	156	144	164	177	153	181	199



Water Research Commission

PO Box 824, Pretoria, 0001, South Africa

Tel: +27 12 330 0340, Fax: +27 12 331 2565

Web: http://www.wrc.org.za