INTERPOLATION OF THE DAILY RAINFALL MODEL

by

L. McNEILL, A. BRANDÃO, W. ZUCCHINI and A. JOUBERT

Department of Statistical Sciences

University of Cape Town

December 1993

Report to the Water Research Commission on the Project

"Interpolation and Mapping of Daily Rainfall Model Parameters of South Africa"

Project Leaders : Prof W. Zucchini

: Ms L. McNeill

WRC Report No 305/1/94
ISBN 1 86845 074 0

EXECUTIVE SUMMARY

INTERPOLATION OF THE DAILY RAINFALL MODEL

by

L.McNEILL, A.BRANDÃO, W.ZUCCHINI and A.JOUBERT

Department of Statistical Sciences
University of Cape Town
December 1993

Report to the Water Research Commission on the Project
"Interpolation and Mapping of Daily Rainfall Model Parameters
for South Africa"

Project Leaders : Prof. W. Zucchini
: Ms. L. McNeill

Motivation

In southern Africa[1], rainfall is the element of climate most influential in determining the variety and abundance of flora and fauna, land use, economic development and practically all aspects of human activity. The major climatic and agricultural regions of southern Africa are based largely on the areal distribution and seasonality of rainfall. Most studies have focused on the simplest characteristic of the rainfall process such as annual and monthly means. However, as was pointed out by Tyson (1986):

> "... it is clear that rainfall over Africa is a highly variable quantity, particularly over the dry western parts of South Africa. Consequently the concept of mean annual rainfall at any one locality must be treated with caution."

The same comment holds for monthly means. Furthermore monthly means provide little or no information on many properties of the rainfall that are relevant to the wide variety of rainfall-related activities. For example, the risk and severity of storms, the risk, severity and duration of drought and the timing of rainfall within each year are all aspects of rainfall that are of importance to decision making.

It is of course possible to make a special study of any particular property of daily rainfall. For example, Adamson (1981) tabulated and mapped the risk and severity of n-day storm depths (for $n = 1, 2, 3, 7$) at 2200 sites in southern Africa. However the variety of statistics that might be of interest to different decision makers is effectively infinite, which renders that approach problematic.

An alternative and more flexible approach is to model the daily rainfall process itself and thereby encapsulate all the properties of daily rainfall by

[1]Throughout this report, 'southern Africa' is defined to include South Africa, Lesotho and Swaziland.

means of a small number (in our case 16) of model parameters. Until the advent of cheap fast computers this approach would have been fruitless because it is difficult or impossible to determine properties of interest purely analytically, based on such a model. For example it is doubtful that one could derive a formula for the probability of events such as 'there will be at least 50 mm rainfall at Pretoria in July but not more than 20mm on any one day'.

Computers have made it easy to evaluate the probability of any such event or sequence of events, regardless of complexity. Once calibrated, the model can be used to generate long artificial rainfall sequences (typically 1000-2000 years) which preserve all the statistical properties of rainfall; not merely the means and variances, but also the frequency of occurrence of any sequence of values.

The point of being able to generate sequences of artificial rainfall is that it enables one to estimate statistics relating to rainfall events. For example, suppose that we require an estimate of the probability that Stellenbosch will have less than 20 mm rainfall in February. This can be done by using the model to generate a 1000-year daily rainfall sequence at Stellenbosch and counting the number of years in which this event occurred. Suppose that in 689 out of the 1000 years the February rainfall total was less than 20 mm. Then an estimate of the required probability is $689/1000 = 0{,}689$.

In effect one estimates probabilities of this type by simply regarding the artificial rainfall sequence generated as a very long real rainfall record. One can do this because the model used to generate the sequences preserves the properties of real rainfall sequences, for example the averages, standard deviations and in fact the entire probability distribution of daily, monthly and annual rainfall totals, as well as the correlation between rainfall totals on consecutive days, the seasonal distribution of wet and dry runs, and so on.

One can use the artificial sequences generated to estimate a wide variety

of quantities that may be of interest, for example

- What is the probability of having no rain between two specified dates, e.g. between 15 July and 30 July ?
- What is the probability of having a run of 20 consecutive dry days starting sometime in November ?
- Which day (week, month, 50-day period, ...) of the year has the highest (or lowest) average amount of rainfall ?
- What is the distribution of monthly rainfall (mean, median, standard deviation, ...) for any given month of the year ?
- What is the probability that, between 15 October and 31 December, there will be at least 200 mm, and that there will be no 10-day run having less than 5 mm ?

One can answer any of these and similar questions by simply averaging over the generated sequence, that is treating the generated sequence as if it were a very long real rainfall record.

The Water Research Commission project by Zucchini and Adamson entitled 'The Occurrence and Severity of Drought in South Africa' (WRC Report No. 91/1/84 - 91/3/84) described a daily rainfall model for South Africa. The model, which was calibrated at 2550 sites across the country, captures all the probabilistic properties of the daily rainfall process at those sites. It can be used to quantify the daily, monthly and annual statistics of rainfall, its seasonality, the risk of storms and the probabilities of droughts of various durations and intensity. In fact it can be used to estimate the probability of any rainfall event or sequence of events with a resolution of one day or longer. Thus the model provides a versatile decision support tool enabling hydrologists, water resources managers, natural resource planners and other decision

makers to assess the probable consequences of decisions whose outcome depends on the amounts and timing of rainfall. Some applications of the model are described, for example, in Zucchini, Adamson and McNeill (1992). The model is now used routinely by various institutions in Forestry, Agriculture, Nature Conservation, Agricultural and Civil Engineering and Hydrology, as well as by researchers at a number of South African universities, by some farmers, and by a number of companies and financial institutions, such as the Standard Bank of South Africa. It is offered as one of the products of the Computing Centre for Water Research (CCWR), according to whose records it has been used over 2000 times, mainly to infill missing values of daily rainfall prior to the data being run through daily rainfall budgeting models.

Although the model was calibrated at a large number of sites, the sites having sufficiently long records to allow for accurate calibration are concentrated in and around urban centres. Many parts of the country, notably the north-western Cape, the north-eastern Transvaal and Lesotho, are poorly covered, due to the shortage of rainfall records. Consequently users of the model have been obliged to base their estimates and conclusions on the rainfall properties of calibrated sites, which are often quite distant from the location of interest. Thus, whereas the usefulness of the model has been established, its application has been limited to those sites for which it has been calibrated.

Direct estimation of the model parameters is possible using as few as five years of daily rainfall data, although the accuracy of estimates based on so little data would be questionable. However, to establish and service sufficiently many rain gauges to accumulate records of even such relatively modest length is obviously not practical. It is therefore necessary to make do with the data that are available.

Objectives

The main objective of this project has been to produce estimates of the parameters of the daily rainfall model of Zucchini and Adamson (1984) for sites throughout South Africa at which there is little or no rainfall data available, thereby making it possible to use the model to generate artificial rainfall sequences and study rainfall characteristics at any given location or over any given area in South Africa. Parameter estimates were to be made available in the form of:

1. Isoline maps.

2. Digitised values at a regular grid of points one minute of degree square throughout southern Africa, (that is, at a resolution of about 1,5 kilometres), to be made available on magnetic tape.

3. An algorithm for generating parameters at any point.

These constitute three different ways of presenting the same information. During the course of the project, the Project Steering Committee recognised that the maps stipulated under item 1 and the algorithm stipulated in item 3 above would be of limited use once the digitised values were available and recommended that the project team focus on item 2.

A second objective was to develop methodology and computer software for the type of interpolation problem investigated in the project with a view to its future use in the interpolation of other climate variables, such as temperature and relative humidity.

The Database

Rainfall data from a number of sources, including the South African Weather Bureau, the Department of Forestry, the Department of Agriculture, the South African Sugar Association, as well as data collected by farmers and other members of the public, are held by the Computing Centre for Water Research (CCWR), and this data set was used as the data base for this project. Dent *et al.* (1989) describe the data base and its quality in more detail.

In order to fit a reasonably accurate model of daily rainfall at any location, it is necessary to have a fairly long record of daily rainfall at that site. Zucchini and Adamson (1984) fitted their daily rainfall model to some 2550 stations throughout southern Africa, which, in 1981, had at least 30 years of daily data available.

In 1992, there were some 3397 stations with at least 30 years of data in southern Africa (including Lesotho and Swaziland). As the major objective of this project was to extend the geographical coverage of the model, it was decided to include also all stations with between 20 and 30 years of data. The first phase of the project was thus to re-fit the model at each of these stations. Figure 1 shows the location of these sites.

It is clear from this map that there are a number of areas with a very low density of data points, in particular the western, north-western and central Cape, Lesotho, and an area in the north-east of the country around the Kruger National Park. For these areas, it was decided to include those stations having at least five years of data, giving an additional 512 stations. While models fitted at such sites might not be very accurate in themselves, they would contribute useful information to the estimation process described in the report. The accuracy of the fitted model was incorporated into the final estimation process in such a way that stations where the fitted model had low accuracy would be appropriately down-weighted. In all, there were

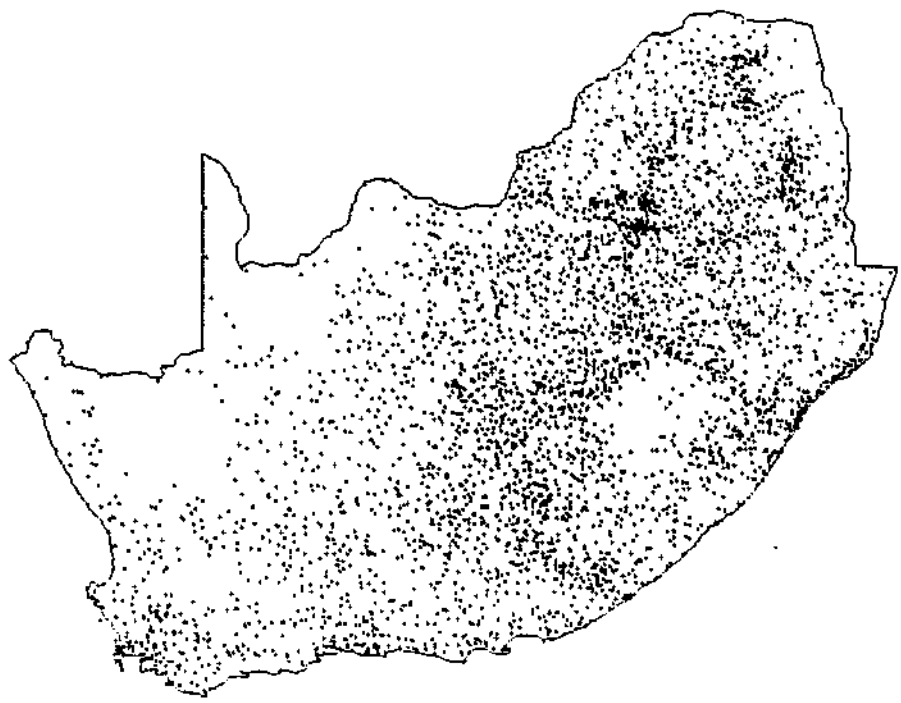


Figure 1: Stations with at least 20 years of data.

5070 stations finally selected. Their locations are shown in Figure 2. Despite the incorporation of the additional stations, some areas of the country are still poorly represented in the data set. In addition, the station locations tend to be clustered around areas of human habitation, so that in mountainous areas there may be a bias towards the lower altitudes, which could give rise to a corresponding downward bias in rainfall estimates for those areas.

The rainfall data-base was complete up to the end of February 1992, except for a few stations where record-keeping had been discontinued prior to this date. Thus the actual time period covered varies from one station to another; for example, a 10 year record covers the period 1982-1992 while a 20

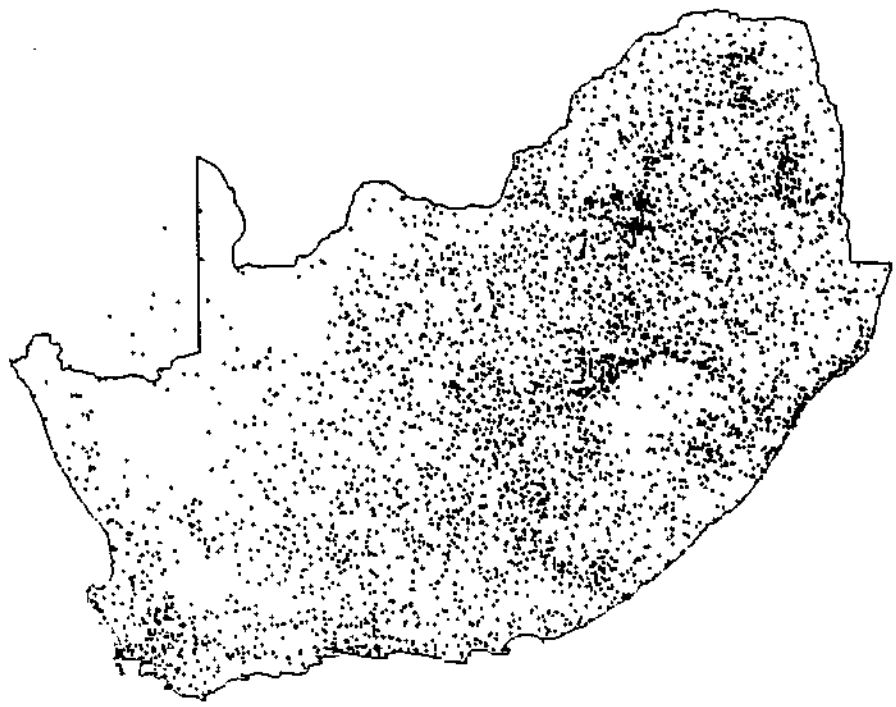


Figure 2: Stations used in this report.

year record covers the period 1972-1992. In analysing the data any possible long-term trends have been ignored; the magnitude of any such trends is in practice very small in comparison with the typical year-to-year variation in the rainfall values.

The data held by CCWR have been screened as far as possible for recording and coding errors. Missing or doubtful values are appropriately flagged in the data base, although there seem to be occasional inconsistencies in the coding of some of the older stations in that missing values are sometimes coded in the same way as zero rainfall. While the model fitting program is designed to deal with missing values in an appropriate way, it is difficult to

quantify the effect of coding and recording errors in the data on the fitted parameters.

Identifying suspect data values is not a trivial task since each value must be considered both in the light of the time of year and the geographical location; a value that is reasonable at one site at a given time of year might be most unlikely in another situation. Fortunately, the majority of rainfall records are unaffected by this problem and, furthermore, many of those that are, contain a only few such anomalies. With this in mind, a number of checks were performed at various stages of the project to identify suspect values, such values were re-coded as missing values

Apart from possible errors in the daily rainfall values another potential source of error is the station locations. Although the locations of a few stations are recorded to the nearest second of a degree of latitude and longitude, the majority are recorded to the nearest minute. This means that locations are accurate to within 1 to 2 km at best. In most parts of the country the pattern of daily rainfall will change very little over such a distance, however in coastal and mountainous areas the changes can be quite significant. This variability must be viewed as a limitation imposed by the resolution of the data; it cannot be removed but must be taken into account in the estimation process.

The Model

In the recent literature the process of daily rainfall is described by a model comprising two components; the first describes the occurrence of wet and dry days while the second describes the distribution of the amounts of rain on wet days, and the parameters of the model are allowed to vary seasonally. Woolhiser (1992) gives a recent review. In modelling the occurrence of wet and dry days a first order Markov chain was found to be appropriate. That is, the rainfall process exhibits a one day 'memory'. Thus the model estimates

the probability of a wet day given that the preceding day was also wet, and the probability of a wet day following a dry day. Clearly these probabilities also vary seasonally in a smooth way. The model incorporates the seasonal effect by fitting a 5-term Fourier series to the data at a given site.

The method of maximum likelihood was used to fit the Fourier parameters to the sequence of historical daily rainfall at a given site. Figure 3 illustrates the model fitted to the probability of a wet day following a dry day at Stellenbosch in the south-west Cape, together with the actual frequencies observed over a 104-year rainfall period.

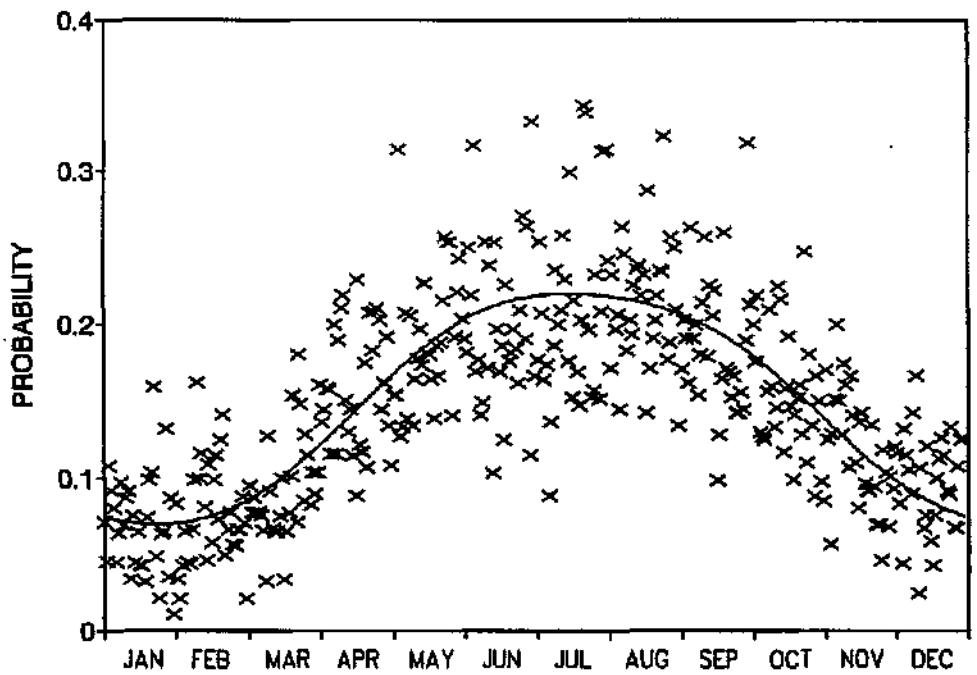


Figure 3: Empirical probabilities and Fourier series model for the probability of a wet day following a dry day at Stellenbosch.

The amount of rain on wet days was found by Zucchini and Adamson (1984) to show a seasonally varying mean but a constant coefficient of vari-

ation. A 5-term Fourier series was again used to model the mean amounts. Given the estimated mean, the method of moments can then be used to fit any two-parameter distribution to the data. The Weibull distribution was found to provide a good fit for stations throughout southern Africa.

The first stage of this project was therefore to re-fit the model outlined above to the 5070 stations shown in Figure 2.

As mentioned above, in order to improve the spatial coverage of the sites we had to make use of a number of sites with quite short rainfall records which can therefore be calibrated relatively imprecisely. Thus the accuracy of the parameter estimates at the calibrated sites varies substantially. In fact the accuracy depends not only on the length of the rainfall record, but also on various aspects of the timing and amount of rainfall at the site. For example, the model parameters for sites in arid areas with highly seasonal rainfall can be estimated less accurately than parameters in areas of high rainfall with less marked wet and dry seasons.

These discrepancies in the accuracy of the parameter estimates at the calibrated sites need to be taken into account in the interpolation process. More specifically it is necessary to have a reasonably accurate measure of the standard errors of the estimates in order to assign appropriate weights to each of the available data points. A substantial portion of the work done on this research project was focused on finding ways to quantify the accuracy of the parameter estimates at calibrated sites. Initially the standard theoretical approach to the problem was attempted, but this led to unacceptable levels of bias. The reasons why this approach fails are discussed in an appendix to the report. An alternative approach was based on the so-called *bootstrap* method, and this proved successful. This method requires an enormous amount of computation, and its implementation would not have been possible without the co-operation of the CCWR who made their computer facilities available to us and kindly assisted with software implementation.

Interpolation of the Parameters

Having re-calibrated the model at the 5070 sites, the major objective of the project was to interpolate the model parameters, (16 at each site), to a grid of 1 minute of a degree of latitude and longitude throughout southern Africa, or some 500 000 points in all.

The key theoretical issue in this project was to identify the most appropriate method of interpolating the calibrated parameter values. All existing methods of interpolation that we could find in the literature were considered; the main ones are briefly reviewed in the report. For a variety of reasons detailed in the report, we decided to make use of the method known as *kriging.* However, as outlined below, the standard kriging techniques (and software) are not directly applicable to our problem so it was necessary to develop new variations on the kriging methodology and to write the corresponding software.

The parameters of the daily rainfall model fall neatly into two types, the 'amplitude parameters' and the 'phase parameters'. Roughly speaking, the former encapsulate information relating to the amount of rainfall at a site and the latter provide information relating to the timing of the rainfall. The coefficient of variation, which is somewhat anomalous, being neither an amplitude nor a phase parameter, can be regarded as being of the first type. The amplitude parameters are scalar quantities (in our case either probabilities or millimetres) but the phase parameters are what are known as circular variables (in our case the days of the year). The magnitude of a scalar variable is determined on an ordinary linear scale but the magnitude of a circular variable is a somewhat subtler concept which needs to be measured as a direction on a circle. As an example, consider the fact that the time interval between day 364 of the year (30 December) and day 365 (31 December) is the same that between day 365 and day 1 (1 January). Even the 'mean' of two circular values has to be defined in a special way; it is

not the simple arithmetic average of the two values. The main consequence of this is that circular variables need to be modelled entirely differently to scalar variables. Furthermore kriging techniques for circular variables were not available and had to be derived; the theory for kriging circular variables which was developed in this project has recently been published in a scientific journal (McNeill, 1993).

The phase parameters of the model do have one property that is not enjoyed by the amplitude parameters, namely they do not depend to any significant extent on local topographic features. Thus one can find pairs of sites, only a few kilometres apart, which have substantially different mean rainfall (typically in mountainous areas), but the seasonality of the rainfall will be approximately the same (they will tend to receive rain at the same time of the year). This property allows one to interpolate the phase parameters directly, without taking local features into account.

The interpolation of the amplitude parameters, by contrast, has to take account of local topographic features. Altitude measurements were available to us on a grid of 1 minute of degree of latitude and longitude throughout southern Africa. In effect this determined the finest resolution that we could achieve for interpolating the model parameters. The question of how best to make use of this altitude information occupied much of our attention. We considered a variety of interpolation techniques which incorporate additional information. A brief review of the main techniques is given in the report. The literature on the interpolation of other aspects of rainfall, such as the mean annual precipitation, describes a variety of measures derived from altitude data, the main ones being gradient, aspect, roughness and exposure. The precise definition of each these measures is, of course, somewhat arbitrary so that there are many variations on how one might define, for example, exposure. One of the main advantages of the kriging technique that we finally adopted is that it is not required to specify such measures in advance

- the method can be used to determine which functions of altitude are most important for the interpolation.

Another of the problems that we had to consider was the magnitude of the data set with which we were dealing. Some techniques are not applicable to such large data sets with the currently available computers - they simply require too much computing. We also required a methodology which would take account of the varying accuracy of the data points. This was important in our application because, as mentioned above, some of the parameter estimates were based on very short rainfall records. The method finally selected was the so-called *kriging with external drift* ; the 'external drift' in this case being the functions of altitude. All computations were done on a local basis; that is, the parameters at each grid point were interpolated using only data values in the vicinity of the grid point; this relieves one of the necessity of first partitioning the country into homogeneous regions, interpolating each region separately and then dealing with the subsequent problem of patching together the estimates from the disjoint regions in a smooth way.

Validation

The rainfall model itself was extensively tested and validated by Zucchini and Adamson (1984). In the present report we focused on the validation of the interpolated parameter estimates. This was carried out by 'hiding' a number of the available data points, using the remaining data points to obtain interpolated estimates at the locations of the hidden points and then comparing the interpolates to the 'true' values. (It needs be kept in mind that the 'true' values are in fact also estimates.) The agreement was found to be within the limits of accuracy indicated by the bootstrap variance calculations.

Another way to validate the results is to to calculate derived characteristics, such as the mean annual precipitation, based on simulated data generated by the model; this enables us to test the model as a whole in the

form in which it will be used in practice, and also allows comparison with the same statistics derived from other sources. We therefore calculated a mean annual precipitation (MAP) at the location of each of 373 selected test sites using four different methods:

- Using a 100 year simulation based on the daily rainfall model parameters estimated for that station.
- Using a 100 year simulation based on the daily rainfall model parameters estimated by the kriging procedure at the grid point with the same latitude and longitude as the station.
- Using the MAP calculated directly from the daily rainfall data for that station held by CCWR.
- Taking the value of MAP from the CCWR data base of gridded MAP values, as estimated by Dent *et al.* (1989).

There are a number of reasons to expect differences between the four values; these are discussed in the report. In general, however, the agreement between the four sets of figures is very close (Figure 4), which helps to confirm that the interpolated model parameters produce realistic simulated rainfall sequences.

Summary

The main objective of the project described in this report was to produce estimates of the parameters of the daily rainfall model of Zucchini and Adamson (1984) for sites throughout southern Africa at which there is little or no rainfall data available, thereby making it possible to use the model to generate artificial rainfall sequences and study rainfall characteristics at any given location or over any given area in southern Africa.

The parameters of the daily rainfall model have been interpolated on a regular grid one minute of degree square throughout southern Africa, that

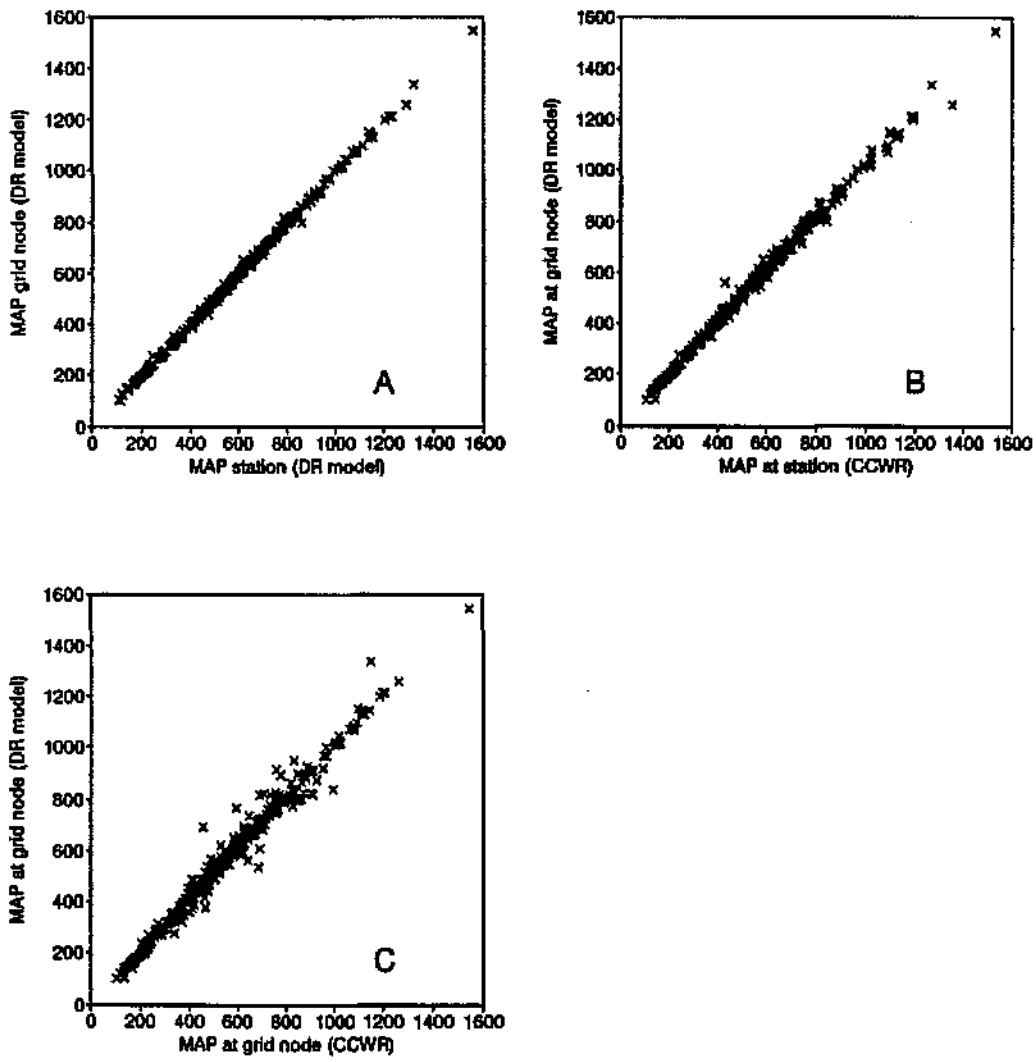


Figure 4: Comparison of MAP values (in mm).

is, at a resolution of about 1,5 kilometres, making the parameter estimates of the model available for approximately 500 000 sites.

The daily rainfall model is routinely used by researchers and decision makers in a wide variety of applications. It is hoped, now that the model is now applicable at practically any site in southern Africa, that it will find even wider application.

It needs to be emphasised that although the theory behind the model is rather technical, the model is easy to use by anybody who can operate a micro-computer. No statistical or other specialist knowledge is required to *apply* the model. The feedback that we have received, during the last eight or nine years, from users with very different mathematical backgrounds, has been encouraging; no-one has indicated that they found the model difficult to apply. We are not aware of any user who has misunderstood what it is that the model provides or who has misinterpreted the estimates derived from the model.

One of the by-products of the project has been the contribution to the theory of kriging, namely the development of a technique for the kriging of circular variables, described in McNeill (1993). The report also briefly reviews kriging and other interpolation techniques and comments on their suitability in the context of hydrological data. This provides a convenient starting point and an up-to-date list of references for researchers wishing to interpolate other values.

Recommendations

The daily rainfall model has 16 parameters. We have generated estimates of these parameters for approximately 500 000 grid points, covering southern Africa on a grid of 1 minute by 1 minute. This information is currently stored at the CCWR; the data file occupies 3 megabytes of computer disc space for each of the 16 parameters or almost 50 megabytes in total. As this quantity

of information is too large to be conveniently distributed in its entirety to individual researchers and other interested parties, we recommend that the CCWR be approached to:

- store the data file of estimated parameter values.
- extend their present service of supplying artificially generated rainfall sequences via the 'DRAINGEN' program to incorporate an option for using the grid point data. (They currently supply generated sequences for the 2550 stations covered in the Zucchini and Adamson (1984) report.)
- maintain an archive of the interpolation software which was used to estimate the grid point values so that it will be possible to re-run the programs at some future date to update the parameter estimates.

We also recommend that some consideration be given to finding appropriate means of publicising the existence of the model and its potential uses. We believe that the number of current users is much smaller than the number of potential users, who are either unaware of the model or who might be mistakenly under the impression that it is a complicated tool requiring specialist knowledge. With this in mind, a PC compatible diskette containing a small data set and sample programs will be made available on request. Further software development, aimed at providing application tools to make optimum utilisation of the generated data, would be a valuable addition.

Further research is required to develop methodology for generating simulated sequences of daily rainfall for an area rather than a single point, in such a way as to preserve the appropriate spatial correlation of individual rainfall occurrences.

References

ADAMSON, P.T. (1981). Southern African storm rainfall. *Technical Report 102.* Department of Water Affairs, Directorate of Scientific Services, Private Bag X313, Pretoria.

DENT, M.C., LYNCH, S.D. and SCHULZE, R.E. (1989). Mapping mean annual and other rainfall statistics over southern Africa. *Water Research Commission Report* 109/1/89, Water Research Commission, Pretoria.

McNEILL, L. (1993). Interpolation and smoothing of mapped circular data. *S. African Statistical Journal* **27**, 23–49.

TYSON, P.D. (1986). *Climatic Change and Variability in Southern Africa.* Oxford University Press, Cape Town.

WOOLHISER, D.A. (1992). Modelling daily precipitation - progress and problems. In *Statistics in the Environmental and Earth Sciences*, A.T. Walden and P. Guttorp, eds. Edward Arnold, New York, 71–89.

ZUCCHINI, W. and ADAMSON, P.T. (1984). The occurrence and severity of droughts in South Africa. *WRC Report* No. 91/1/84, Water Research Commission, Pretoria.

ZUCCHINI, W., ADAMSON, P. and McNEILL, L. (1992). A model of southern African rainfall. *S. Afr. Jnl. Sci.*, **88**, 103–109.

Acknowledgements

The Steering Committee appointed by the Water Research Commission for this project consisted of the following persons:

Dr G C Green	-	Water Research Commission (Chairman)
Mr F P Marais	-	Water Research Commission (Secretary)
Mr H Maaren	-	Water Research Commission
Mr S van Biljon	-	Department of Water Affairs
Dr A L du Pisani	-	Weather Bureau
Mr J F Erasmus	-	Department of Agricultural Development
Prof D Hughes	-	Rhodes University
Mr S D Lynch	-	University of Natal
Prof L G Underhill	-	University of Cape Town

The financing of the project by the Water Research Commission and the contribution and support of the members of the Steering Committee is acknowledged with thanks.

Extensive use was made of the databases and computing facilities of the Computing Centre for Water Research, and we extend our thanks to Mark Dent, Mike Horn, Rajesh Nundlall and Barbara Main for help and advice on the use of their facilities. Special thanks are due to Reinier de Vos and Arne Kure for their generous assistance in adapting programs to the CCWR computer system and linking to CCWR databases, and also to Maria Hill for help with data and other enquiries.

We would also like to thank Shirley Butcher of the Department of Surveying and Geodetic Engineering of the University of Cape Town for providing the digitized coordinates of the boundary of South Africa which were used to prepare the maps in this report.

List of Symbols: Model Parameters

Section 3.3 explains the meaning of each of the individual model parameters. Throughout the text the following abbreviations are used for the parameters:

WWA0	Zero'th amplitude:	Prob$(W_t \mid W_{t-1})$
WWA1	First amplitude:	Prob$(W_t \mid W_{t-1})$
WWA2	Second amplitude:	Prob$(W_t \mid W_{t-1})$
WWP1	First phase:	Prob$(W_t \mid W_{t-1})$
WWP2	Second phase:	Prob$(W_t \mid W_{t-1})$
DWA0	Zero'th amplitude:	Prob$(W_t \mid D_{t-1})$
DWA1	First amplitude:	Prob$(W_t \mid D_{t-1})$
DWA2	Second amplitude:	Prob$(W_t \mid D_{t-1})$
DWP1	First phase:	Prob$(W_t \mid D_{t-1})$
DWP2	Second phase:	Prob$(W_t \mid D_{t-1})$
DEPA0	Zero'th amplitude:	Mean depth on wet days $(mm \times 10^{-1})$
DEPA1	First amplitude:	Mean depth on wet days $(mm \times 10^{-1})$
DEPA2	Second amplitude:	Mean depth on wet days $(mm \times 10^{-1})$
DEPP1	First phase:	Mean depth on wet days $(mm \times 10^{-1})$
DEPP2	Second phase:	Mean depth on wet days $(mm \times 10^{-1})$
CV	Coefficient of Variation:	Depth on wet days

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In southern Africa[1], rainfall is the element of climate most influential in determining the variety and abundance of flora and fauna, land use, economic development and practically all aspects of human activity. The major climatic and agricultural regions of southern Africa are based largely on the areal distribution and seasonality of rainfall. (See for example, Dove (1988), Schumann and Thompson, (1934), Schumann and Hofmeyr (1938), Schulze (1947), (1958), Jackson (1951), Wellington (1955).) Most studies have focused on the simplest characteristic of the rainfall process such as annual and monthly means. However, as was pointed out by Tyson (1986):

> "... it is clear that rainfall over Africa is a highly variable quantity, particularly over the dry western parts of South Africa. Consequently the concept of mean annual rainfall at any one locality must be treated with caution."

The same comment holds for monthly means. Furthermore monthly means provide little or no information on many properties of the rainfall that are relevant to the wide variety of rainfall-related activities. For example, the risk and severity of storms, the risk, severity and duration of drought and

[1]Throughout this report, 'southern Africa' is defined to include South Africa, Lesotho and Swaziland.

the timing of rainfall within each year are all aspects of rainfall that are of importance to decision making.

It is of course possible to make a special study of any particular property of daily rainfall. For example, Adamson (1981) tabulated and mapped the risk and severity of n-day storm depths (for $n = 1, 2, 3, 7$) at 2200 sites in southern Africa. However the variety of statistics that might be of interest to different decision makers is effectively infinite, which renders that approach problematic.

An alternative and more flexible approach is to model the daily rainfall process itself and thereby encapsulate all the properties of daily rainfall by means of a small number (in our case 16) of model parameters. Until the advent of cheap fast computers this approach would have been fruitless because it is difficult or impossible to determine properties of interest purely analytically. For example it is doubtful that one could derive a formula for the probability of events such as "there will be at least 50 mm rainfall at Pretoria in July but not more than 20mm on any one day".

Computers have made it easy to evaluate the probability of any such event or sequence of events, regardless of complexity. Once calibrated, the model can be used to generate long artificial rainfall sequences (typically 1000-2000 years) which preserve all the statistical properties of rainfall; not merely the means and variances, but also the frequency of occurrence of any sequence of values.

The point of being able to generate sequences of artificial rainfall is that it enables one to estimate statistics relating to rainfall events. For example, suppose that we require an estimate of the probability that Stellenbosch will have less than 20 mm rainfall in February. This can be done by using the model to generate a 1000-year daily rainfall sequence at Stellenbosch and counting the number of years in which this event occurred. Suppose that in 689 out of the 1000 years the February rainfall total was less than 20 mm.

Then an estimate of the required probability is $689/1000 = 0{,}689$.

In effect one estimates probabilities of this type by simply regarding the artificial rainfall sequence generated as a very long real rainfall record. One can do this because the model used to generate the sequences preserves the properties of real rainfall sequences, for example the averages, standard deviations and in fact the entire probability distribution of daily, monthly and annual rainfall totals, as well as the correlation between rainfall totals on consecutive days, the seasonal distribution of wet and dry runs, and so on.

One can use the artificial sequences generated to estimate a wide variety of quantities that may be of interest, for example

- What is the probability of having no rain between two specified dates, e.g. between 15 July and 30 July ?
- What is the probability of having a run of 20 consecutive dry days starting sometime in November ?
- Which day (week, month, 50-day period,...) of the year has the highest (or lowest) probability of having non-zero rainfall ?
- Which day (week, month, 50-day period,...) of the year has the highest (or lowest) probability of having at least 25 mm of rainfall ?
- Which day (week, month, 50-day period,...) of the year has the highest (or lowest) average amount of rainfall ?
- What is the average rainfall for any given period of the year, e.g. between 29 February and 13 April ? What is the corresponding standard deviation, median, mode, 90% confidence interval ?
- What is the distribution of monthly rainfall for any given month of the year ?

- What is the distribution of annual rainfall ?
- What is the least amount of rainfall that can be reasonably expected (for example, with probability 0,9) between 15 December and 15 February ?
- What is the probability of having more than 200 mm in any 3 consecutive days between 1 September and 31 January ?
- What is the probability that, between 15 October and 31 December, there will be at least 200 mm, and that there will be no 10-day run having less than 5 mm ?

One can answer any of these and similar questions by simply averaging over the generated sequence, that is treating the generated sequence as if it were a very long real rainfall record.

Zucchini and Adamson (1984a) described a daily rainfall model for sites in southern Africa. The model, which was calibrated at 2550 sites across the region, captures all the probabilistic properties of the daily rainfall process at those sites. Some applications of the model are described, for example, in Zucchini, Adamson and McNeill (1992).

The model is now used routinely by various institutions in Forestry, Agriculture, Nature Conservation, Agricultural and Civil Engineering and Hydrology, as well as by researchers at a number of South African universities, by some farmers, and by a number of companies and financial institutions, such as the Standard Bank of South Africa. It is offered as one of the products of the Computing Centre for Water Research (CCWR), according to whose records it has been used over 2000 times, mainly to infill missing values of daily rainfall prior to the data being run through daily rainfall budgeting models.

Although the model was calibrated at a large number of sites, the sites having sufficiently long records to allow for accurate calibration are concentrated in and around urban centres. Many parts of the country, notably the north-western Cape, the north-eastern Transvaal and Lesotho, are poorly covered, due to the shortage of rainfall records. Consequently users of the model have been obliged to base their estimates and conclusions on the rainfall properties of calibrated sites, which are often quite distant from the location of interest. Thus, whereas the usefulness of the model has been established, its application has been limited to those sites for which it has been calibrated.

Direct estimation of the model parameters is possible using as few as five years of daily rainfall data, although the accuracy of estimates based on so little data would be questionable. However, to establish and service sufficiently many rain gauges to accumulate records of even such relatively modest length is obviously not practical. It is therefore necessary to make do with the data that are available.

The main objective of this project has been to produce estimates of the parameters of the daily rainfall model of Zucchini and Adamson (1984a) for sites throughout southern Africa at which there is little or no rainfall data available, thereby making it possible to use the model to generate artificial rainfall sequences and study rainfall characteristics at any given location or over any given area in southern Africa.

This report describes the theory and methods used to reach this objective, namely to obtain estimates of the model parameters on a regular grid one minute of degree square throughout southern Africa, that is, at a resolution of about 1,5 kilometres. Thus the parameter estimates of the model are now available for approximately 500 000 sites.

To achieve this objective, the model was first calibrated at a total of some 5070 sites for which data are available. (A brief description of the

data that were available to us is given in Chapter 2, details of the model are given in Chapter 3.). As many sites as possible were used in order to increase the density of coverage, especially in areas with a low density of data points, such as the western, north-western and central Cape, Lesotho and the north-western and north-eastern Transvaal. The 2550 sites covered in the Zucchini and Adamson (1984a) report were re-calibrated so as to take advantage of the additional data that have become available since the release of that report. The resulting estimates formed the raw material for the interpolation procedure.

In order to increase the number of sites we had to make use of a number of sites with quite short rainfall records which can therefore be calibrated relatively imprecisely. Thus the accuracy of the parameter estimates at the calibrated sites varies substantially. In fact the accuracy depends not only on the length of the rainfall record, but also on various aspects of the timing and amount of rainfall at the site. For example, the model parameters for sites in arid areas with highly seasonal rainfall can be estimated less accurately than parameters in areas of high rainfall with less marked wet and dry seasons.

These discrepancies in the accuracy of the parameter estimates at the calibrated sites need to be taken into account in the interpolation process. More specifically it is necessary to have a reasonably accurate measure of the standard errors of the estimates in order to assign appropriate weights to each of the available data points. A substantial portion of the work done on this research project was focused on finding ways to quantify the accuracy of the parameter estimates at calibrated sites. Initially the standard theoretical approach to the problem was attempted, but this led to unacceptable levels of bias. The reasons why this approach fails are discussed in Appendix A. An alternative approach (described in Chapter 4) was based on the so-called *bootstrap* method, and this proved successful. This method requires an enormous amount of computation, more than would have been possible 20 years

ago.

The re-fitting of the model parameters and the estimation of their standard errors has not been without problems. A number of stations were found to have isolated data values which were clearly questionable. Identifying suspect data values is not a trivial task since each value must be considered both in the light of the time of year and the geographical location; a value that is reasonable at one site at a given time of year might be most unlikely in another situation. Fortunately, the majority of rainfall records are unaffected by this problem and, furthermore, many of those that are, contain a only few such anomalies. As far as possible, suspect data values were identified and either corrected or re-coded as missing values.

The key theoretical issue in this project was to identify the most appropriate method of interpolating the calibrated parameter values. This is the subject of Chapter 5. All existing methods of interpolation that we could find in the literature were considered; the main ones are briefly reviewed in the report. For a variety of reasons detailed in the report, we decided to make use of the method known as *kriging*. However, as outlined below, the standard kriging techniques (and software) are not directly applicable to our problem so it was necessary to develop new variations on the kriging methodology and to write the corresponding software.

The parameters of the daily rainfall model fall neatly into two types, the 'amplitude parameters' and the 'phase parameters'. Roughly speaking, the former encapsulate information relating to the amount of rainfall at a site and the latter provide information relating to the timing of the rainfall. The coefficient of variation, which is somewhat anomalous, being neither an amplitude nor a phase parameter, can be regarded as being of the first type. The amplitude parameters are scalar quantities (in our case either probabilities or millimetres) but the phase parameters are what are known as circular variables (in our case the days of the year). The magnitude of a scalar vari-

able is determined on an ordinary linear scale but the magnitude of circular variable is a somewhat subtler concept which needs to be measured as a direction on a circle. As an example, consider the fact that the time interval between day 364 of the year (30 December) and day 365 (31 December) is the same that between day 365 and day 1 (1 January). Even the 'mean' of two circular values has to be defined in a special way; it is not the simple arithmetic average of the two values. The main consequence of this is that circular variables need to be modelled entirely differently to scalar variables. Furthermore kriging techniques for circular variables were not available and had to be derived; the theory for kriging circular variables which was developed in this project has recently been published in a scientific journal (McNeill, 1993).

The phase parameters of the model do have one property that is not enjoyed by the amplitude parameters, namely they do not depend to any significant extent on local topographic features. Thus one can find pairs of sites, only a few kilometres apart, which have substantially different mean rainfall (one site might be in a rain shadow area), but the seasonality of the rainfall will be approximately the same (they will tend to receive rain at the same time of the year). This property allows one to interpolate the phase parameters directly, without taking local features into account.

The interpolation of the amplitude parameters, by contrast, has to take account of local topographic features. Altitude measurements extracted by Dent *et al.* (1989) were available to us on a grid of 1 minute of a degree of latitude and longitude throughout southern Africa. In effect this determined the finest resolution that we could achieve for interpolating the model parameters. The question of how best to make use of this altitude information occupied much of our attention. We considered a variety of interpolation techniques which incorporate additional information. A brief review of the main techniques is given in the report. The literature on the interpolation

of other aspects of rainfall, such as the mean annual precipitation, describes a variety of measures derived from altitude data, the main ones being gradient, aspect, roughness and exposure. The precise definition of each these measures is, of course, somewhat arbitrary so that there are many variations on how one might define, for example, exposure. One of the main advantages of the kriging technique that we finally adopted is that it is not required to specify such measures in advance - the method can be used to determine which functions of altitude are most important for the interpolation.

Another of the problems that we had to consider was the magnitude of the data set with which we were dealing. Some techniques are not applicable to such large data sets with the computers currently available - they simply require too much computing. We also required a methodology which would take account of the varying accuracy of the data points. This was important in our application because, as mentioned above, some of the parameter estimates were based on very short rainfall records. The method finally selected was the so-called *kriging with external drift* ; the 'external drift' in this case being the functions of altitude. All computations were done on a local basis; that is, the parameters at each grid point were interpolated using only data values in the vicinity of the grid point; this relieves one of the necessity of first partitioning the country into homogeneous regions, interpolating each region separately and then dealing with the subsequent problem of patching together the estimates from the disjoint regions in a smooth way.

The rainfall model itself was extensively tested and validated by Zucchini and Adamson (1984a). In the present report we focused on the validation of the interpolated parameter estimates. This was carried out by 'hiding' a number of the available data points, using the remaining data points to obtain interpolated estimates at the locations of the hidden points and then comparing the interpolates to the 'true' values. (It needs be kept in mind that the 'true' values are in fact also estimates.) The agreement was found

to be quite close.

Chapter 6 gives an outline of the algorithm needed to generate artificial rainfall sequences using the model. Our conclusions and recommendations are given in Chapter 7.

Chapter 2

The Data

2.1 The Database

Rainfall data from a number of sources, including the South African Weather Bureau, the Department of Forestry, the Department of Agriculture, the South African Sugar Association, as well as data collected by farmers and other members of the public, are held by the Computing Centre for Water Research (CCWR), and this data set was used as the data base for this project. Dent *et al.* (1989) describe the data base and its quality in more detail.

2.2 Selection of Stations

In order to fit a model of daily rainfall at any location, it is necessary to have a fairly long record of daily rainfall at that site. Zucchini and Adamson (1984a) fitted their daily rainfall model to some 2550 stations throughout southern Africa, which, in 1981, had at least 30 years of daily data available.

In 1992, there were some 3397 stations with at least 30 years of data in southern Africa (including Lesotho and Swaziland). As the major objective of this project was to extend the geographical coverage of the model, it was

decided to include also all stations with between 20 and 30 years of data. The first phase of the project was thus to re-fit the model, as described in Chapter 3, at each of these stations. Figure 2.1 shows the location of the sites. It is clear from this map that there are a number of areas with a very low density of data points, in particular the western, north-western and central Cape, Lesotho, and an area in the north-east of the country around the Kruger National Park. For these areas, it was decided to include those

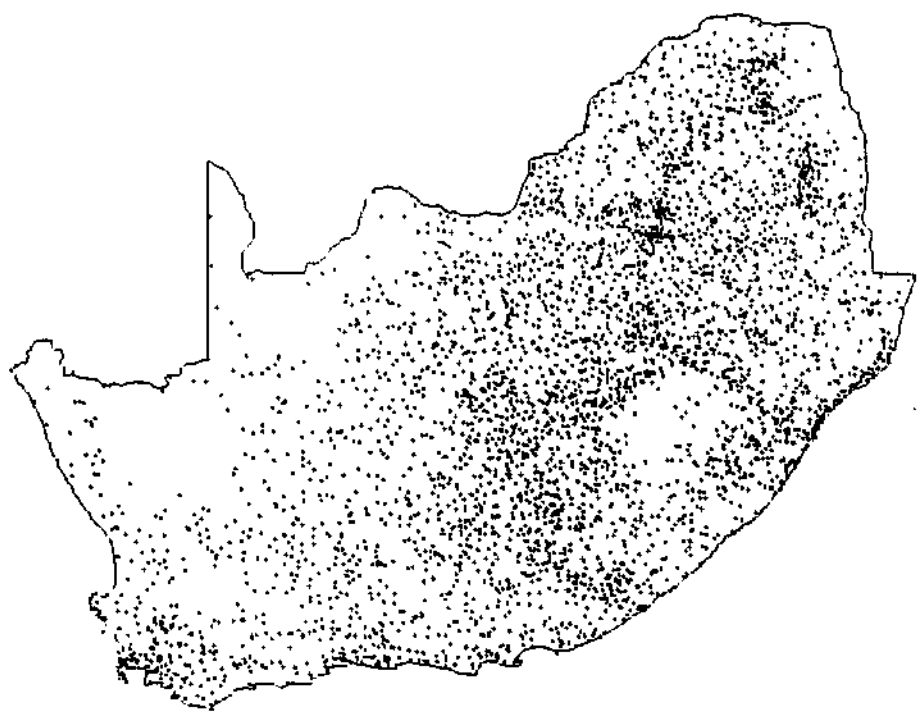


Figure 2.1: Stations with at least 20 years of data.

stations having at least five years of data, giving an additional 512 stations (Figure 2.2). While models fitted at such sites might not be very accurate in themselves, they would contribute useful information to the estimation

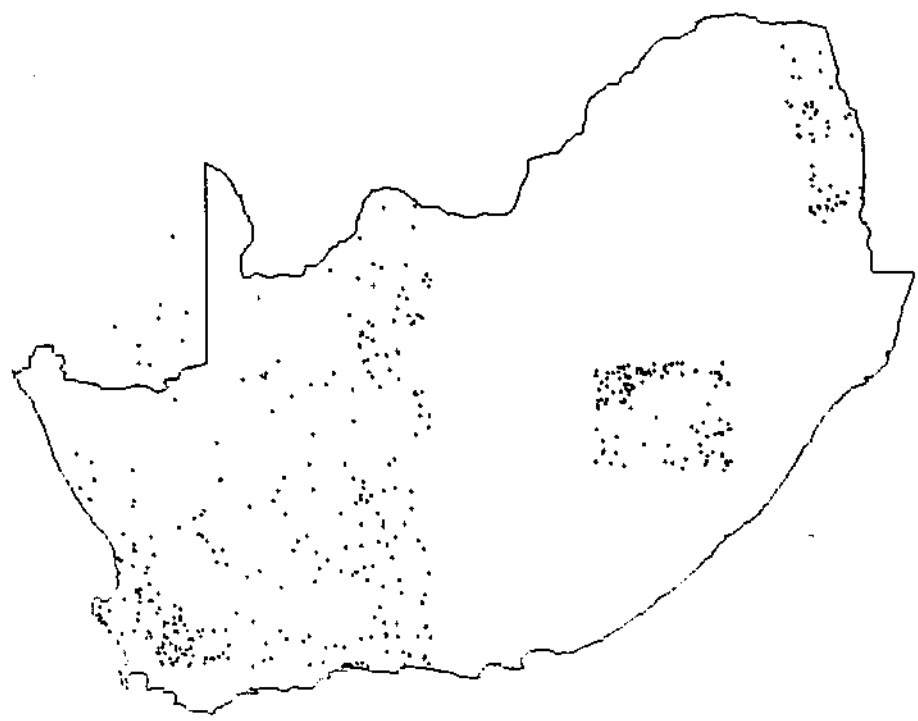


Figure 2.2: Stations with between 5 and 20 years of data.

process described in Chapter 5. The accuracy of the fitted model was incorporated into the final estimation process in such a way that stations where the fitted model had low accuracy would be appropriately down-weighted. In all, there were 5070 stations finally selected. Their locations are shown in Figure 2.3. Despite the incorporation of the additional stations, some areas of the country are still poorly represented in the data set. In addition, the station locations tend to be clustered around areas of human habitation, so that in mountainous areas there may be a bias towards the lower altitudes, which could give rise to a corresponding downward bias in rainfall estimates for those areas. This point will be addressed in Chapter 5.

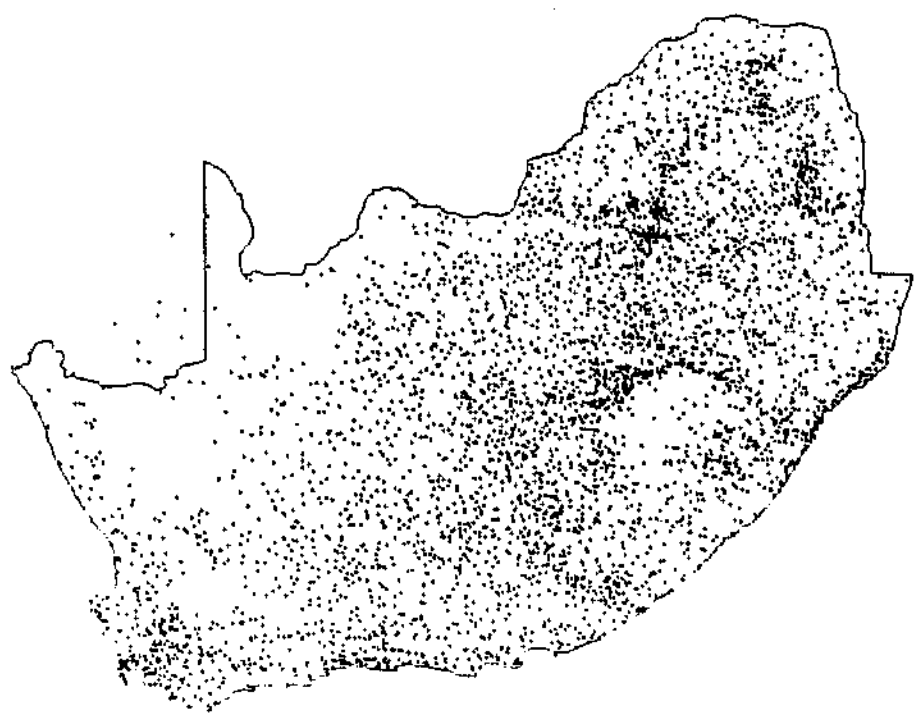


Figure 2.3: Stations used in this report.

Figure 2.4 shows the distribution of the number of years of available rainfall record at each of the 5070 stations. The data was complete up to the end of February 1992, except for a few stations where record-keeping had been discontinued prior to this date. Thus the actual time period covered varies from one station to another; for example, a 10 year record covers the period 1982-1992 while a 20 year record covers the period 1972-1992. In analyzing the data any possible long-term trends have been ignored; the magnitude of any such trends is in practice very small in comparison with the typical year-to-year variation in the rainfall values.

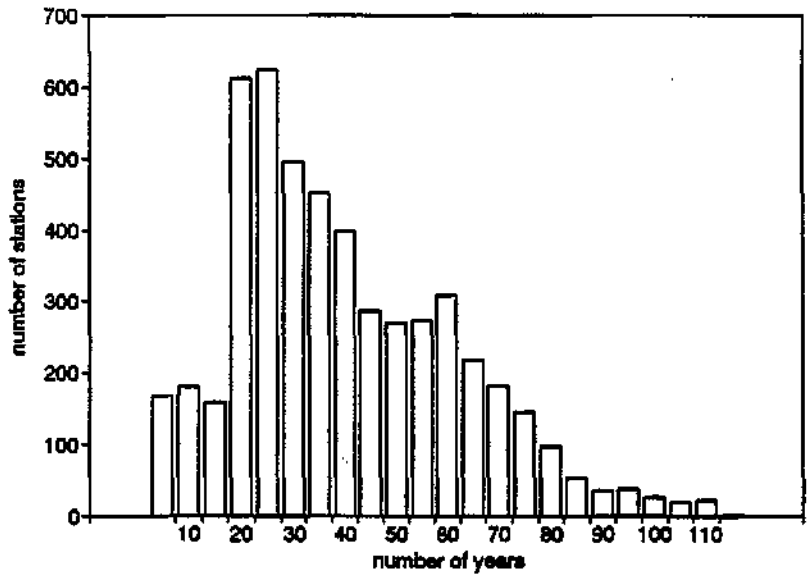


Figure 2.4: Distribution of length of rainfall record.

2.3 Accuracy of the Data

The data held by CCWR has been screened as far as possible for recording and coding errors. Missing or doubtful values are appropriately flagged in the data base, although there seem to be occasional inconsistencies in the coding of some of the older stations in that missing values are sometimes coded in the same way as zero rainfall. While the model fitting program is designed to deal with missing values in an appropriate way, it is difficult to quantify the effect of coding and recording errors in the data on the fitted parameters.

With this in mind, a number of checks were performed at various stages of the project to identify suspect values. One of the first checks was to construct histograms (see Figure 3.1) of each of the fitted model parameters and to

investigate any outliers. In addition, 200 year simulations (based on the fitted daily rainfall model) were carried out to estimate mean annual precipitation (MAP) values at each site, which were compared with those obtained directly from the CCWR data. Some 89 sites were found to be discrepant, apparently mainly as a result of the inconsistencies in coding of missing and zero data described above. The data for these sites were re-checked by the CCWR, and suspect values were re-coded as missing values where necessary. After bootstrapping the data to estimate the variances as described in Chapter 4, a further check was made by comparing the bootstrap means with the original estimates. This led to the exclusion of some additional stations, as described in Section 4.3.

Apart from possible errors in the daily rainfall values another potential source of error is the station locations. Although the locations of a few stations are recorded to the nearest second of a degree of latitude and longitude, the majority are recorded to the nearest minute. This means that locations are accurate to within 1 to 2 km at best. In most parts of the country the pattern of daily rainfall will change very little over such a distance, however in coastal and mountainous areas the changes can be quite significant. As an example, Table 2.1 lists the fitted model parameters at three stations on the slopes of Table Mountain in Cape Town which all have the same recorded location. It can be seen that for some parameters[1] the differences are quite considerable. This is also reflected in the 'nugget effect' apparent in the semi-variograms discussed in Chapter 5. This variability must be viewed as a limitation imposed by the resolution of the data; it cannot be removed but must be taken into account in the estimation process.

[1]See symbol list on page iv for an explanation of the parameter codes.

	Station Code		
	020719 W	020719AW	020719BW
WWA0	-0.847	-1.018	0.183
WWA1	0.622	0.576	0.634
WWA2	0.140	0.183	0.085
WWP1	195.40	204.56	191.87
WWP2	131.09	127.22	132.51
DWA0	-1.614	-1.646	-1.175
DWA1	0.292	0.258	0.395
DWA2	0.051	0.067	0.033
DWP1	216.67	216.34	211.50
DWP2	49.42	53.02	97.74
DEPA0	203.40	192.38	114.20
DEPA1	88.63	91.82	37.90
DEPA2	25.23	27.12	11.77
DEPP1	173.44	173.40	176.81
DEPP2	164.90	163.71	175.26
CV	1.265	1.278	1.233

Table 2.1: Fitted parameters: stations coded 020719.

Chapter 3

The Daily Rainfall Model

The sequences of rainfall values exhibit a number of distinctive features. In particular the distribution of daily precipitation depths varies seasonally, rainfall depths on consecutive days are not independently distributed, that is, the probability that a wet day will follow a wet day is higher than the probability that a wet day will follow a dry day, and finally the distribution of rainfall is partly discrete and partly continuous. Any useful model for the description of precipitation sequences must of course preserve all these properties.

Several models have been proposed for simulating daily precipitation. (Gabriel and Neumann, 1962; Richardson, 1981; Roldan and Woolhiser, 1982; Stern and Coe, 1984; Zucchini and Adamson, 1984a. For a recent review, see Woolhiser, 1992.) Most precipitation models are specified by a discrete occurrence process describing the sequence of wet and dry days, and a continuous distribution function for the amount of precipitation of days with rain. The parameters of the model are allowed to vary seasonally.

3.1 A Model to Describe the Occurrence of Wet and Dry Sequences of Days

A first-order Markov chain is used to describe the occurrence of wet and dry days. By this one assumes that the state of day t depends on the state of the previous day, $t-1$. This does not imply that the state at time t is independent of the state on day $t-2, t-3$, etc ..., but rather that the information given by $t-1$ is equivalent to all the information given by $t-1, t-2$, etc .... One also assumes that, except for the seasonality, the process is stationary.

A first-order Markov chain has been found to be an adequate model for precipitation occurrence in many different regions. (See, for example, Gabriel and Neumann, 1962; Caskey, 1963; Weiss, 1964; Hopkins and Robillard, 1964; Haan *et al.*, 1976; Smith and Schreiber, 1973; Woolhiser and Pegram, 1979; Richardson, 1981; Roldan and Woolhiser, 1982; Zucchini and Adamson, 1984a, Woolhiser,1992.) The order of the Markov chain may of course be increased, but this has to be done at the cost of increasing complexity and the number of parameters in the model. A further problem arises if one attempts to increase the order of the Markov chain in arid areas, namely the estimation of the probability that a rain day follows two or more consecutive rain days. In arid areas there are relatively few runs of three or more consecutive rain days and thus there is hardly any data on which to base estimates of this conditional probability. (Note that this has to be estimated for each day of the year.) Finally, it was demonstrated in Zucchini and Adamson (1984a) that a first order Markov chain provides an adequate description of the occurrence of wet and dry sequences of days in the complete range of southern African conditions.

3.1.1 Notation and Preliminaries

The day will be used as the time unit. That is, the year is divided into $NT(= 365)$ equal intervals, denoted by $t = 1, 2, \ldots, NT$. A day with total rainfall greater than 0 mm is considered as a wet day.

The following notation will be used:

R represents the occurrence of rain (i.e. wet day).

$\bar{R}$ represents the non-occurrence of rain (i.e. dry day).

For $t = 1, 2, \ldots, NT$

$NR(t)$ is the number of times it was wet in period t.

$N\bar{R}(t)$ is the number of times it was dry in period t.

$N\bar{R}R(t)$ is the number of times it was dry in period $t-1$ and wet in period t.

$NR\bar{R}(t)$ is the number of times it was wet in period $t-1$ and dry in period t.

$N\bar{R}\bar{R}(t)$ is the number of times it was dry in period $t-1$ and dry in period t.

$NRR(t)$ is the number of times it was wet in period $t-1$ and wet in period t.

$ND(t) = N\bar{R}R(t) + N\bar{R}\bar{R}(t)$ is the number of times that it was dry in period $t-1$ and there was an observation (wet or dry) in period t.

$NW(t) = NRR(t) + NR\bar{R}(t)$ is the number of times that it was wet in period $t-1$ and there was an observation (wet or dry) in period t.

$\pi_{R/R}(t)$ the probability that period t is wet given that period $t-1$ is wet.

$\pi_{\bar{R}/R}(t)$ the probability that period t is dry given that period $t-1$ is wet.

$\pi_{R/\bar{R}}(t)$ the probability that period t is wet given that period $t-1$ is dry.

$\pi_{\bar{R}/\bar{R}}(t)$ the probability that period t is dry given that period $t-1$ is dry.

Then $\pi_{R/R}(t) + \pi_{\bar{R}/R}(t) = 1$

$$\pi_{\bar{R}/\bar{R}}(t) + \pi_{R/\bar{R}}(t) = 1.$$

Therefore the transition probabilities are fully defined given $\pi_{R/R}(t), \pi_{\bar{R}/\bar{R}}(t)$ and the wet or dry state on day $t-1$, and one only needs to estimate these two probabilities.

From elementary probability theory we have

$$\begin{aligned} NRR(t) &\sim B(NW(t),\ \pi_{R/R}(t)) \\ N\bar{R}R(t) &\sim B(ND(t),\ \pi_{\bar{R}/\bar{R}}(t)), \qquad t = 1, 2, \ldots, NT \end{aligned}$$

where $B(N, \pi)$ denotes the binomial distribution with parameters N and π.

3.1.2 Estimation

The functions $\pi_{R/R}(t)$ and $\pi_{\bar{R}/\bar{R}}(t)$ are estimated using the same method but different data. To simplify the notation in what follows, one makes use of the following generic names:

$$\text{Let} \qquad M(t) \sim B(MM(t),\ \pi(t)), \qquad t = 1, 2, \ldots, NT.$$

First we note that the binomial distribution belongs to the exponential family. Therefore we have a set of independent random variables $M(t), t = 1, 2, \ldots, NT$, each with a distribution from the exponential family; each $M(t)$ depends on a single parameter $\pi(t)$ and the distributions of all $M(t)$, $t = 1, 2, \ldots, NT$, are of the same form (i.e. all binomial). Thus the properties of a generalized linear model are satisfied, and estimates of $\pi(t)$ may be obtained by using the theory for estimation for generalized linear models. (Dobson, 1983.)

The probabilities $\pi(t)$ are represented as functions of a linear combination of parameters $\gamma_1, \gamma_2, \ldots, \gamma_{NT}$. That is

$$g(\pi(t)) = \lambda(t)$$

where g is the link function and $\lambda(t)$ is a linear combination of the γ_i.

To ensure that the estimated values of $\pi(t)$ are restricted to the interval $[0,1]$, one uses the logit link function, given by

$$g(\pi(t)) = \log\left(\frac{\pi(t)}{1-\pi(t)}\right) = \lambda(t).$$

To obtain the linear combination of the γ_i, $\lambda(t)$, we look at some of the properties of $\pi(t)$, namely that it is a smooth, periodic and approximately sinusoidal shaped function. Transforming $\pi(t)$, using the logistic transformation, to a logit $\lambda(t)$ given by

$$\lambda(t) = \log\left(\frac{\pi(t)}{1-\pi(t)}\right) ,$$

one obtains a representation which has the similar properties to $\pi(t)$, and thus we can approximate $\lambda(t)$ by the first few terms of its Fourier representation. This approximation has been used by Stern and Coe (1984) and Zucchini and Adamson (1984a).

The exact Fourier representation of $\lambda(t)$ is given by

$$\lambda(t) = \sum_{i=1}^{NT} \gamma_i \varphi_i(t), \qquad t = 1, 2, \ldots, NT$$

where

$$\varphi_i(t) = \begin{cases} \cos(\omega(t-1)i/2) & \text{i} = 2,4,\ldots \\ \sin(\omega(t-1)(i-1)/2) & \text{i} = 3,5,\ldots \end{cases}$$

$$\varphi_1(t) = 1 \; ; \qquad t = 1, 2, \ldots, NT,$$

and

$$\omega = \frac{2\pi}{NT}.$$

Define the function $\lambda(t, L)$ by

$$\lambda(t, L) = \sum_{i=1}^{L} \gamma_i \varphi_i(t), \qquad t = 1, 2, \ldots, NT; \quad L \leq NT$$

where $\varphi_i(t)$ is defined as before and L is the order of the Fourier series approximation. One is thus making the following approximation:

For some $L < NT$

$$\lambda(t, L) \approx \lambda(t) , \qquad t = 1, 2, \ldots, NT.$$

A procedure to choose the order of the Fourier series approximation (i.e. the value of L) will be discussed later. Generally this approximation is accurate for small values of L. The number of parameters, L, is always chosen to be an odd number. This restriction is made partly for programming convenience and partly for the following reason:

If we rewrite the Fourier representation of $\lambda(t, L)$ by its polar form, we get

$$\lambda(t, L) = \begin{cases} \alpha_0 + \sum_{i=1}^{p} \alpha_i \cos\left(\frac{2\pi i}{NT}((t-1) - \beta_i)\right), & L \text{ odd} \\ \alpha_0 + \sum_{i=1}^{p} \alpha_i \cos\left(\frac{2\pi i}{NT}((t-1) - \beta_i)\right) + \alpha_p \cos\frac{2\pi p(t-1)}{NT}, & L \text{ even} \end{cases}$$

where

$$\begin{aligned} \alpha_0 &= \gamma_1 \\ \alpha_i &= \left(\gamma_{2i}^2 + \gamma_{2i+1}^2\right)^{\frac{1}{2}}, \qquad i = 1, 2, \ldots, p \\ \beta_i &= \frac{NT}{2\pi i} \arctan\left(\frac{\gamma_{2i+1}}{\gamma_{2i}}\right), \qquad i = 1, 2, \ldots, p \end{aligned}$$

and p is the integer part of $\frac{L-1}{2}$. The α_i is called the amplitude and β_i is called the phase of the ith harmonic.

If L is even, then the highest harmonic does not have a phase parameter. Thus the quality of the fit of the model depends on the time origin selected. If L is odd we obtain the same degree of approximation for all time origins.

We have used the Fourier representation of $\lambda(t)$ as the basis for obtaining approximations. Other representations are also feasible, e.g. polynomials or rational functions. There are several reasons for selecting the Fourier representation rather than other possibilities. Firstly, $\lambda(t)$ is known to be approximately sinusoidal in shape and consequently we can expect that even for small values of L, the approximation $\lambda(t, L) \approx \lambda(t)$ will be reasonably accurate. Secondly, $\lambda(t, L)$ is periodic, which is a property that $\lambda(t)$ is known to have. Thirdly, the individual components in the representation are orthogonal, which is a convenient mathematical property.

The log-likelihood function of the observed values as a function of the probabilities $\pi(t)$, is given by

$$\ell(\pi(t);\, M(t)) = \sum_{t=1}^{NT} \left[M(t) \log\left(\frac{\pi(t)}{1-\pi(t)}\right) + MM(t) \log(1-\pi(t)) + \log \binom{MM(t)}{M(t)} \right].$$

Therefore, the log-likelihood function of the observed values as a function of the parameters $\gamma_1, \gamma_2, \ldots, \gamma_L$ is given by

$$\ell(\gamma; M(t)) = \sum_{t=1}^{NT} \left[M(t)\lambda(t, L) - MM(t) \log(1 + e^{\lambda(t,L)}) + \log \binom{MM(t)}{M(t)} \right].$$

The score vector U with respect to $\gamma_1, \gamma_2, \ldots, \gamma_L$ has elements given by

$$\begin{aligned} U_j = \frac{\partial \ell(\gamma; M(t))}{\partial \gamma_j} &= \sum_{t=1}^{NT} \left[M(t) - MM(t)\, \frac{e^{\lambda(t,L)}}{1 + e^{\lambda(t,L)}} \right] \varphi_j(t) \\ &= \sum_{t=1}^{NT} [M(t) - MM(t)\pi(t)] \varphi_j(t) \end{aligned}$$

since $\mathrm{Var}(M(t)) = MM(t)\, \dfrac{e^{\lambda(t,L)}}{(1 + e^{\lambda(t,L)})^2}$ and

$$\begin{aligned} E(M(t)) &= \frac{MM(t) e^{\lambda(t,L)}}{(1 + e^{\lambda(t,L)})} \quad \text{and so} \\ \frac{\partial E(M(t))}{\partial \lambda(t, L)} &= \frac{MM(t) e^{\lambda(t,L)}}{(1 + e^{\lambda(t,L)})^2} = \mathrm{Var}(M(t)). \end{aligned}$$

Similarly, the information matrix $\mathbf{I}_{L\times L}$ has elements given by

$$\mathbf{I}_{jk} = \sum_{t=1}^{NT} \varphi_j(t)\varphi_k(t)MM(t)\,\frac{e^{\lambda(t,L)}}{(1+e^{\lambda(t,L)})^2}.$$

Since $\dfrac{e^{\lambda(t,L)}}{(1+e^{\lambda(t,L)})^2} = \pi(t)(1-\pi(t))$ it follows that

$$\mathbf{I}_{jk} = \sum_{t=1}^{NT} \varphi_j(t)\varphi_k(t)MM2(t)\pi(t)(1-\pi(t)).$$

The maximum likelihood estimates for $\gamma_1, \gamma_2, \ldots, \gamma_L$ are then obtained by solving the iterative equation

$$\mathbf{I}^{(m-1)}\widehat{\gamma}^{(m)} = \mathbf{I}^{(m-1)}\widehat{\gamma}^{(m-1)} + U^{(m-1)}$$

where m indicates the mth approximation and $\widehat{\gamma}$ is the vector of estimates.

Some initial approximation $\gamma^{(0)}$ is used to evaluate $\mathbf{I}^{(0)}$ and $U^{(0)}$, then the iterative equation is solved to give $\gamma^{(1)}$ which in turn is used to obtain better approximations for $\mathbf{I}$ and U, and so on until adequate convergence is achieved. When the difference between successive approximations $\gamma^{(m)}$ and $\gamma^{(m-1)}$ is sufficiently small, $\gamma^{(m)}$ is taken as the maximum likelihood estimate vector.

3.1.3 Model Selection

Whenever a model is fitted to observed data, two types of discrepancy arise. The discrepancy due to approximation (the fewer the number of parameters fitted, the higher the value of this discrepancy) and the discrepancy due to estimation (the more parameters fitted, the higher the value of this discrepancy). When choosing the number of parameters to be fitted, one attempts to minimize the combined effect arising from the two discrepancies.

Selection of the number of parameters, L, may be done by using the criterion of the Kullbach-Leibler measure of discrepancy (Linhart and Zucchini, 1986; Zucchini and Adamson, 1984a.)

Under the assumption that for some L_0, $\lambda(t)$ is exactly fitted by L_0 parameters, i.e.

$$\lambda(t) = \lambda(t, L_0), \quad L_0 < NT,$$

the above method leads to the Akaike Information Criterion where

$$AIC = -\ell(\gamma; M(t)) + L$$

where $\ell(\gamma; M(t))$ is the log-likelihood function given before.

Each value of L leads to a different approximating model. The criterion is computed for $L = 1, 3, 5, \ldots$ and the model which leads to the smallest value of the criterion is selected.

The AIC criterion is much easier to compute than the full Kullbach-Leibler discrepancy and leads to almost identical results if the discrepancy due to approximation is small, which it is in this application. (Linhart and Zucchini, 1986)

3.2 The Distribution of Rainfall on Days when Rain Occurs

Several models have been proposed for the distribution of precipitation amounts given the occurrence of a wet day. These include the exponential (Todorovic and Woolhiser, 1975; Richardson, 1981); gamma (Ison *et al.*, 1981; Buishand, 1977; Stern and Coe, 1984); two-parameter gamma (Buishand, 1978); three-parameter mixed exponential (Woolhiser and Pegram, 1979); kappa (Mielke, 1973); lognormal and Weibull (Zucchini and Adamson, 1984a).

Woolhiser and Roldan (1982b) found that out of the exponential, gamma and mixed exponential distributions, the latter fitted the model of precipitation amounts best. Zucchini and Adamson (1984a) found that for stations in southern Africa, the lognormal distribution did not fit some stations, while the Weibull seemed to provide better fits.

It is known that the distribution of precipitation depths when rain occurs is positively skewed (i.e. smaller amounts occurring more frequently than the larger amounts) and that it exhibits the same seasonal variability as found with the probabilities $\pi(t)$. To account for this seasonality, the simplest solution is to fit a family of distributions and then to allow the parameters to change over the year, where these parameters are expressed in terms of its Fourier series approximation.

The method of modelling precipitation amounts is based on Zucchini and Adamson (1984a). Here one does not fit any model initially, the first two moment functions of the distribution are fitted instead. These are then used to estimate the parameters (by the method of moments) to any desired two-parameter model. Different families can be fitted to a single record, e.g. one for the rainy season and a second for the dry season.

3.2.1 Notation

The year is divided into NT equal intervals denoted by $t = 1, 2 \ldots, NT$.

$M(t)$ represents the number of times that it rained in period t.

$R(i,t)$ represents the rainfall depth on the ith year that it rained in period t, where $i = 1, 2, \ldots, M(t)$.

C represents the coefficient of variation which we assume to be constant for all t (Zucchini and Adamson, 1984a).

$\mu(t)$ represents the mean rainfall per rainy day in period $t = 1, 2, \ldots, NT$.

3.2.2 Estimating the Mean and Coefficient of Variation

As observed before $\mu(t)$ can be approximated by its truncated Fourier series representation thus reducing the number of parameters to be estimated. That is, we make the approximation:

$$\mu(t, L) \approx \mu(t), \qquad t = 1, 2, \ldots, NT; \quad L < NT$$

where $\mu(t)$ is defined as

$$\mu(t) \;=\; \sum_{i=1}^{NT} \mu_i \varphi_i(t) \qquad t = 1, 2, \ldots, NT$$

and

$$\mu(t, L) \;=\; \sum_{i=1}^{L} \mu_i \varphi_i(t) \qquad t = 1, 2, \ldots, NT; \quad L \leq NT$$

and $\varphi_i(t)$ is defined as before.

Define $m(t)$ to be the observed means for each period, i.e.

$$m(t) = \frac{1}{M(t)} \sum_{i=1}^{M(t)} R(i, t), \qquad t = 1, 2, \ldots, NT; \quad i = 1, 2, \ldots, M(t); \quad M(t) > 0$$

where $m(t)$ is not defined when $M(t) = 0$, i.e. it never rained in period t.

We use the method of least squares on $m(t)$ to estimate $\mu_1, \mu_2, \ldots \mu_L$, that is, minimize

$$\sum_{t=1}^{NT} (m(t) - \mu(t, L))^2 \tag{3.1}$$

with respect to the μ_i, $i = 1, 2, \ldots, L$. Approximations to the least squares estimators when some of the $M(t) = 0$, something which occurs often in arid regions, are given by

$$\widehat{\mu}_i = K(i) \sum_{\substack{t=1 \\ M(t)>0}}^{NT} m(t) \varphi_i(t) \tag{3.2}$$

where

$$K(i) = \sum_{\substack{t=1 \\ M(t)>0}}^{NT} \varphi_i(t)^2 \qquad i = 1, 2, \ldots, L.$$

The $m(t)$ in (3.1) are given the same weight and so periods which had very little rainfall have a large influence in the estimates of $\mu(t)$. To overcome this difficulty, the following criterion is used instead:

Minimize

$$S(\mu) = \sum_{t=1}^{NT} \sum_{i=1}^{M(t)} (R(i,t) - \mu(t,L))^2 \tag{3.3}$$

with respect to μ_i, $i = 1, 2, \ldots, L$.

By adding and subtracting $m(t)$ inside the squared term of (3.3), $S(\mu)$ can be rewritten as

$$S(\mu) = S + \sum_{t=1}^{NT} M(t)(m(t) - \mu(t,L))^2 \tag{3.4}$$

where

$$S = \sum_{t=1}^{NT} \sum_{i=1}^{M(t)} (R(i,t) - m(t))^2$$

and $m(t)$ is defined as before if $M(t) \neq 0$ and $m(t) = 0$ if $M(t) = 0$.

To minimize (3.4) its first partial derivatives are set equal to zero:

$$\frac{\partial S(\mu)}{\partial \mu_i} = -2 \sum_{t=1}^{NT} M(t)(m(t) - \mu(t,L))\varphi_i(t), \qquad i = 1, 2, \ldots, L.$$

These L equations can be solved using the Newton-Raphson iteration method. For this, we need the second partial derivatives:

$$\frac{\partial^2 S(\mu)}{\partial \mu_i \partial \mu_j} = 2 \sum_{t=1}^{NT} M(t)\varphi_i(t)\varphi_j(t), \qquad i, j = 1, 2, \ldots, L.$$

Denote the ith element of the vector $f^{(k)}$ by

$$f_i^{(k)} = \sum_{t=1}^{NT} M(t)(m(t) - \mu^{(k)}(t,L))\varphi_i(t), \qquad i = 1, 2, \ldots, L \tag{3.5}$$

and the (i,j)th element of the matrix $F^{(K)}$ by

$$F_{ij}^{(k)} = \sum_{t=1}^{NT} M(t)\varphi_i(t)\varphi_j(t), \qquad i,j = 1,2,\ldots,L \tag{3.6}$$

where k denotes the kth iteration.

Then an algorithm to estimate μ_i, $i = 1,2,\ldots,L$ is given by:

Step 1: Obtain initial estimates $\mu_1^{(0)},\ldots,\mu_L^{(0)}$ using (2) and compute $\mu^{(0)}(t,L)$.

Step 2: Compute $f^{(k)}$ using (5) and $F^{(k)}$ using (6).

Step 3: Compute the vector $\delta^{(k)}$ which is the solution to the system of L linear equations given by

$$F^{(k)}\delta^{(k)} = f^{(k)}$$

Step 4: Set $\mu^{(k+1)} = \mu^{(k)} - \delta^{(k)}$.

Step 5: Test for convergence, e.g. if the elements of $f^{(k)}$ are sufficiently close to zero. If the convergence criterion is met, stop, otherwise increase k by 1 and go to Step 2.

Note that $F^{(k)}$ is symmetric. This fact can be used to reduce the number of computations performed.

An estimator of C is given by:

$$\widehat{C} = \left[\frac{\left[\sum_{t=1}^{NT}\sum_{i=1}^{M(t)}(R(i,t)-\widehat{\mu}(t))^2\right]}{\left[\sum_{t=1}^{NT} M(t)\widehat{\mu}(t)^2\right]}\right]^{\frac{1}{2}}. \tag{3.7}$$

3.2.3 Selecting the Number of Parameters

$$\Delta(L) = \sum_{t=1}^{NT}(\mu(t) - E(\widehat{\mu}(t,L)))^2, \qquad L = 1,3,5,\ldots \tag{3.8}$$

would be a suitable discrepancy on which to base the selection, except that some $M(t)$ are zero and so only approximately unbiased estimators are available. The reliability of this criterion is therefore difficult to determine.

If one is prepared to make distributional assumptions, then selection criteria are relatively easy to derive, for example based on the Kullbach-Leibler discrepancy.

A reasonable procedure is to select L for a parametric family of models and then use the same L in the estimation of $\mu(t)$.

3.2.4 Fitting the Weibull Family

Zucchini and Adamson (1984a) found the Weibull family to fit the rainfall depth models for stations in southern Africa and so this family was used to model the observed rainfall amounts on days that rain was recorded.

Having estimated the mean value function $\mu(t)$ and the coefficient of variation, C, one can apply the method of moments to estimate the parameter functions of the Weibull distribution.

Denote the scale parameter by $\alpha(t)$, $t = 1, 2, \ldots, NT$ and the shape parameter by β.

Now

$$C = \left\{ \frac{\Gamma(1 + 2/\beta)}{\Gamma(1 + 1/\beta)^2} - 1 \right\}^{\frac{1}{2}}. \tag{3.9}$$

To obtain β as a function of C a rational function approximation has to be derived as no closed expression of this function is available.

The following approximation has been obtained from Zucchini and Adamson (1984a):

$$\widehat{\beta} = \frac{339.5410 + 148.445\widehat{C} + 192.7492\widehat{C}^2 + 22.4401\widehat{C}^3}{1 + 257.1162\widehat{C} + 287.8362\widehat{C}^2 + 157.2230\widehat{C}^3}. \tag{3.10}$$

Using the relationship

$$\mu(t) = \alpha(t)\Gamma(1 + 1/\beta) \qquad t = 1, 2, \ldots, NT$$

we obtain the estimator

$$\widehat{\alpha}(t) = \frac{\widehat{\mu}(t)}{\Gamma(1 + 1/\widehat{\beta})} \qquad t = 1, 2, \ldots, NT.$$

3.3 The Amplitude-Phase Representation

Up until now we have used the representation

$$\lambda(t,L) = \sum_{i=1}^{L} \gamma_i \varphi_{i(t)}, \qquad t = 1,2,\ldots,NT; \qquad L \leq NT$$

where L is an odd integer, and where

$$\varphi_i(t) = \begin{cases} \cos(\omega(t-1)i/2) & \text{i = 2,4,...} \\ \sin(\omega(t-1)(i-1)/2) & \text{i = 3,5,...} \end{cases}$$
$$\varphi_1(t) = 1 \; ; \qquad t = 1,2,\ldots,NT,$$

and

$$\omega = \frac{2\pi}{NT}.$$

Although this representation is convenient for computational purposes since the terms $\varphi_i(t)$ need only be computed once, it is not a very convenient representation for purposes of interpretation of the parameters and for comparing the parameters of different stations. For this the amplitude-phase representation is more appropriate. For example the first phase parameter represents the time of year of maximum probability of rain, or of maximum rain depths, while the zero'th amplitude represents the average rainfall depth, or the average probability of rain throughout the year and the first amplitude describes the range of rainfall depth, or of the probability of rainfall. The phase parameter has the further advantage in interpolating rainfall parameters in that they are not affected by altitude. The amplitude-phase representation is given by

$$\lambda(t,L) = \alpha_0 + \sum_{i=1}^{P} \alpha_i cos\left(\frac{2\pi i}{NT}(t-1-\phi_i)\right)$$

where $\alpha_0 = \gamma_1$ and

$$\begin{aligned} \alpha_i &= \left(\gamma_{2i}^2 + \gamma_{2i+1}^2\right)^{\frac{1}{2}}, \qquad i = 1,2,\ldots,P, \\ \phi_i &= \frac{NT}{2\pi i} \arctan\left(\frac{\gamma_{2i+1}}{\gamma_{2i}}\right), \qquad i = 1,2,\ldots,P, \end{aligned}$$

where $P = \frac{L-1}{2}$.

Obtaining maximum likelihood estimates for the amplitude-phase representation is equivalent to obtaining the maximum likelihood estimates for the parameters γ_i, $i = 1, 2, \ldots, L$ and then transforming them as above.

In order to obtain phases that are always between 0 and NT we use the following convention to compute the ϕ_i, $i = 1, 2, \ldots, P$

$$\begin{array}{llllll}
\text{If} & \gamma_{2i} > 0 & \text{then} \begin{cases} \text{if} & \gamma_{2i+1} < 0 & \text{then} & \phi_i = C[A + 2\pi] \\ \text{if} & \gamma_{2i+1} \geq 0 & \text{then} & \phi_i = CA \end{cases} \\
\text{If} & \gamma_{2i} = 0 & \text{then} \begin{cases} \text{if} & \gamma_{2i+1} < 0 & \text{then} & \phi_i = C[\frac{3\pi}{2}] \\ \text{if} & \gamma_{2i+1} \geq 0 & \text{then} & \phi_i = C[\frac{\pi}{2}] \end{cases} \\
\text{If} & \gamma_{2i} < 0 & \text{then} & & & \phi_i = C[A + \pi],
\end{array}$$

where $C = \frac{NT}{2\pi i}$ and $A = \arctan(\frac{\gamma_{2i+1}}{\gamma_{2i}})$ and the range of arctan is defined to be in the interval $(-\pi/2, \pi/2]$.

With this convention we in fact have that the phases $\phi_i \in (0, NT/i]$, $i = 1, 2, \ldots, L$.

The model described above was fitted to the daily rainfall data at each of the stations selected as discussed in Chapter 2. Thus, for each station we estimated the sixteen model parameters as listed on page iv. Histograms of each of the parameters are shown in Figure 3.1, while Figure 3.2 maps the mean value of each parameter, averaged over all rainfall stations within each Weather Bureau block.

3.4 Rainfall Model Validation

In order to ensure that simulated sequences of daily rainfall data generated by the model preserve those properties of the process which are of interest to the user, Zucchini and Adamson (1984a) tested the model at six stations which

broadly represent the various rainfall/climate regions of southern Africa. The properties which they tested included

- The annual mean and variance and the distribution of annual totals and sums of annual totals.
- The monthly means and variances.
- The expected number of wet days, and its seasonal variation.
- The runs characteristics of daily rainfalls and their seasonal variation, for example the 5 or 10 day rainfall total at various times of the year.
- The distribution of n-day extreme rainfalls.

Their tests showed that the relevant properties were faithfully reproduced by the model at each of the test sites. In view of the fact that the model used here was identical to that used by Zucchini and Adamson (1984a) the model validation process was not repeated.

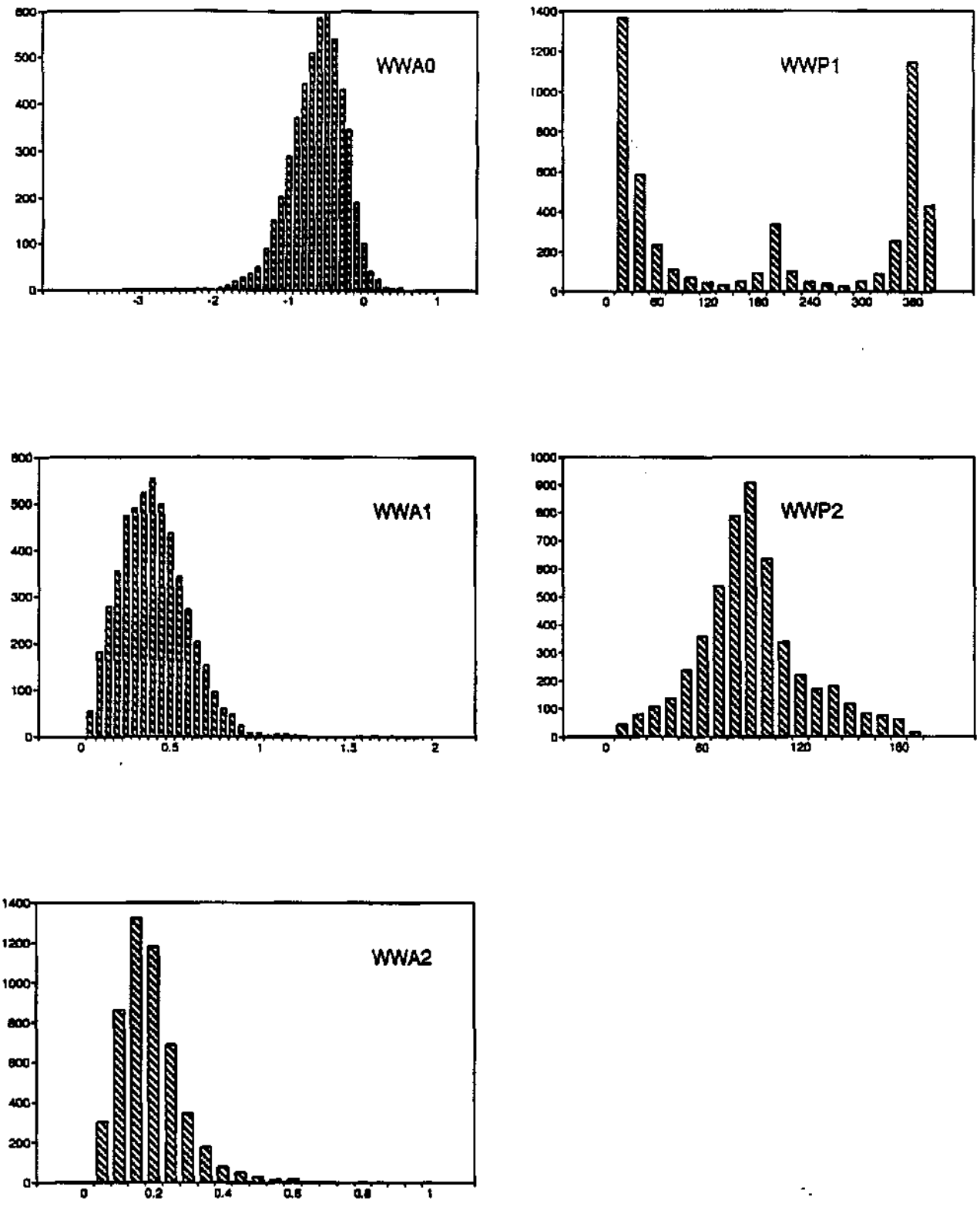
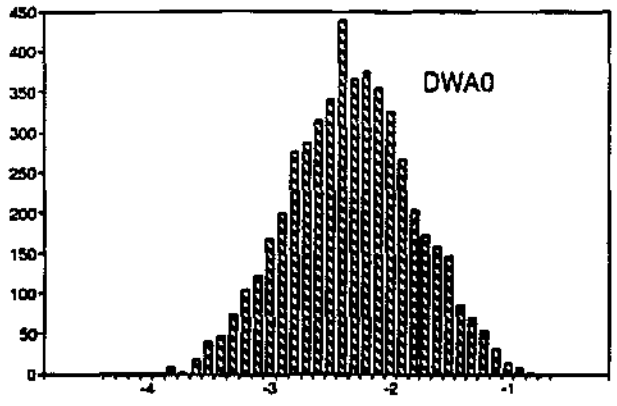
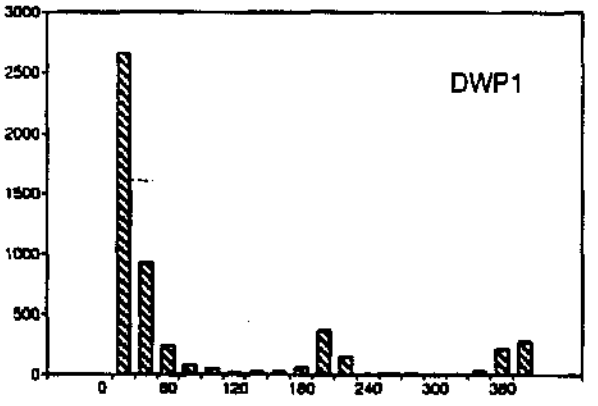
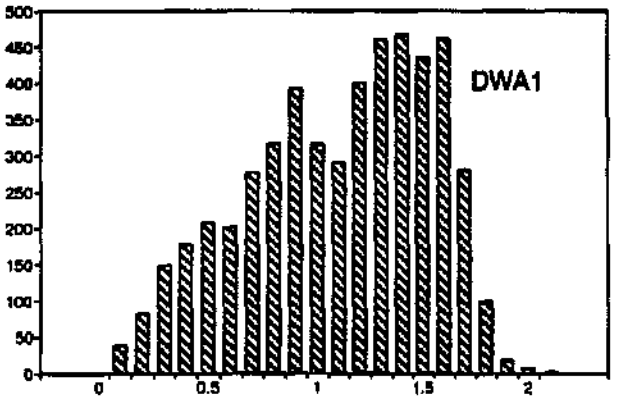
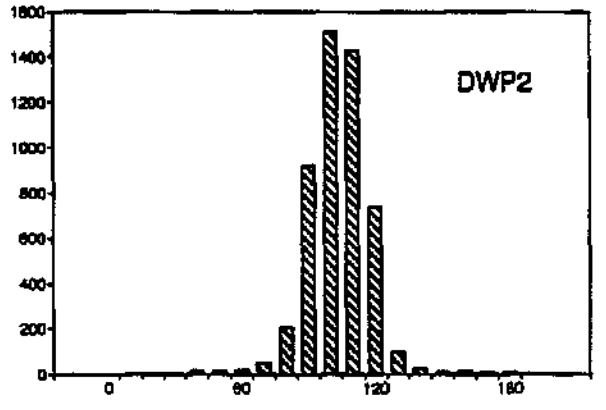


Figure 3.1: Histogram of parameter values.









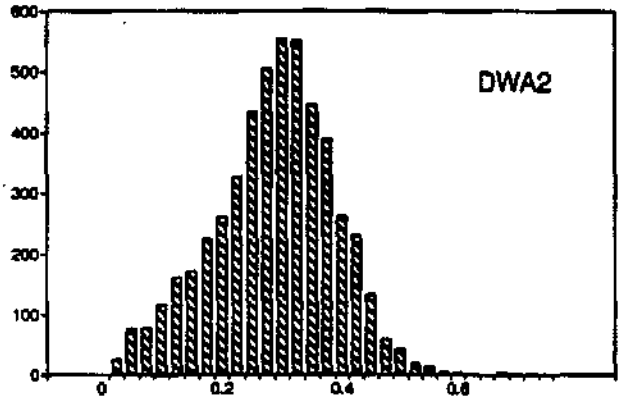


Figure 3.1: Histogram of parameter values (contd.).

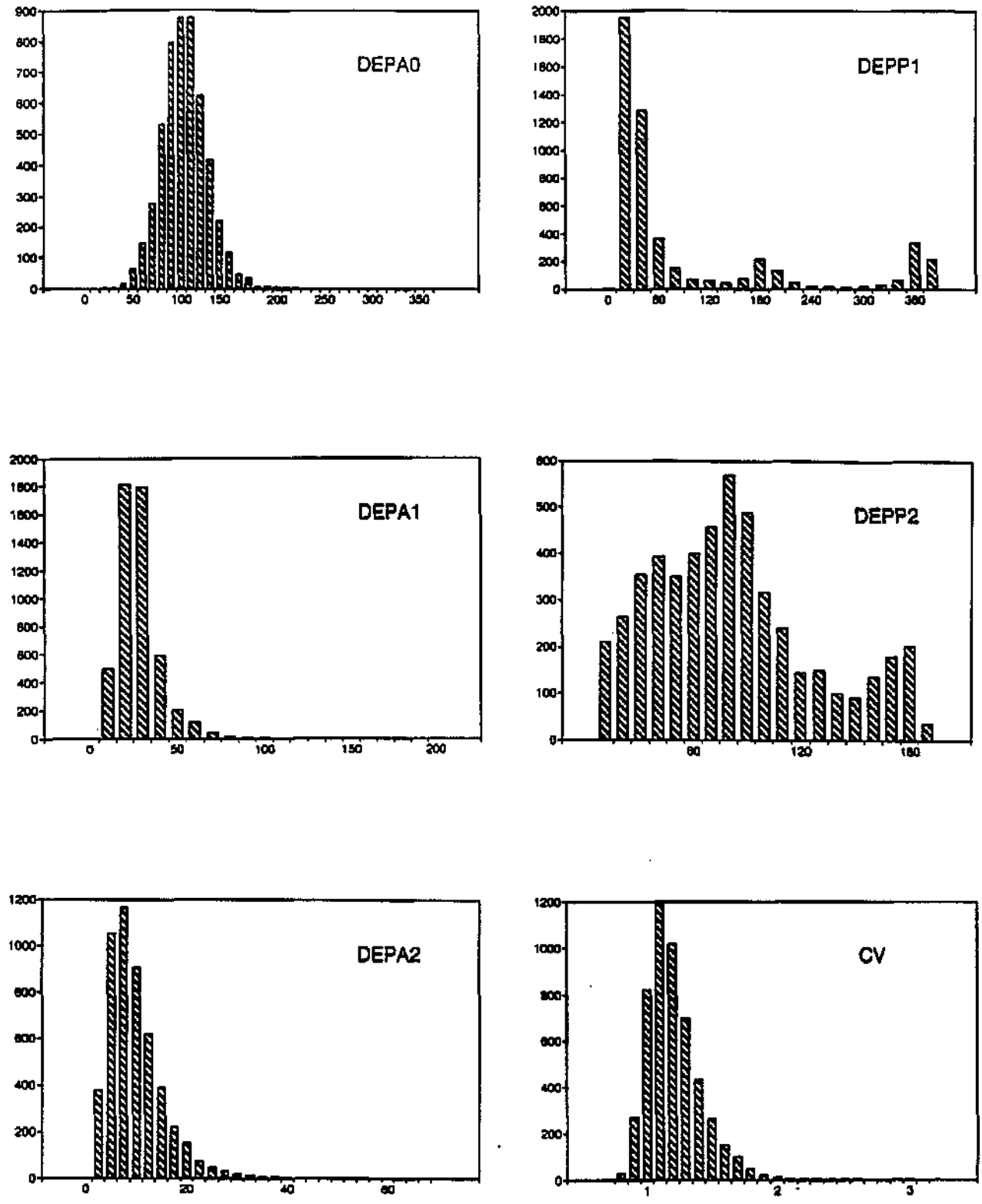


Figure 3.1: Histogram of parameter values (contd.).

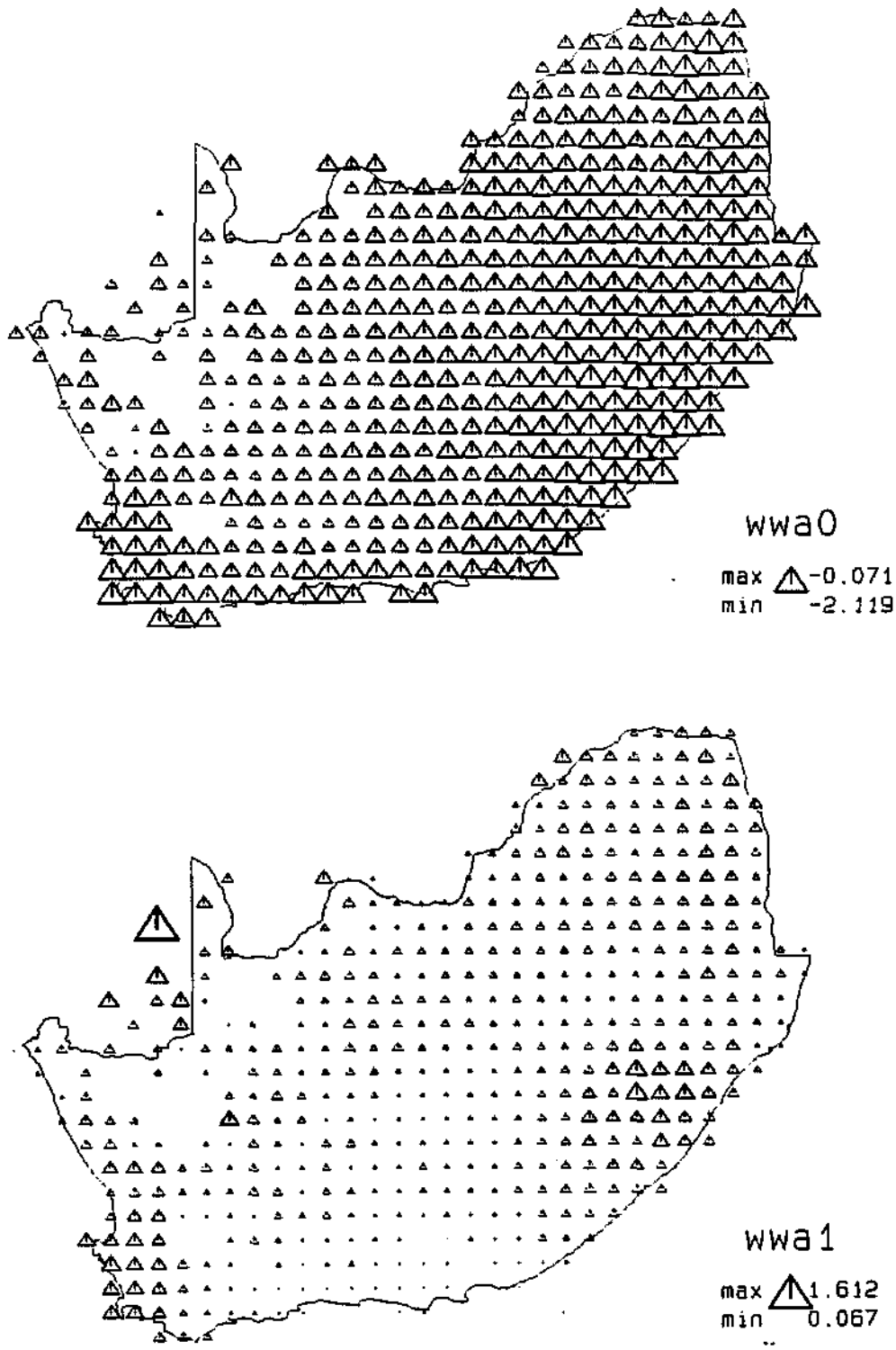


Figure 3.2: Mean parameter values for each Weather Bureau block.

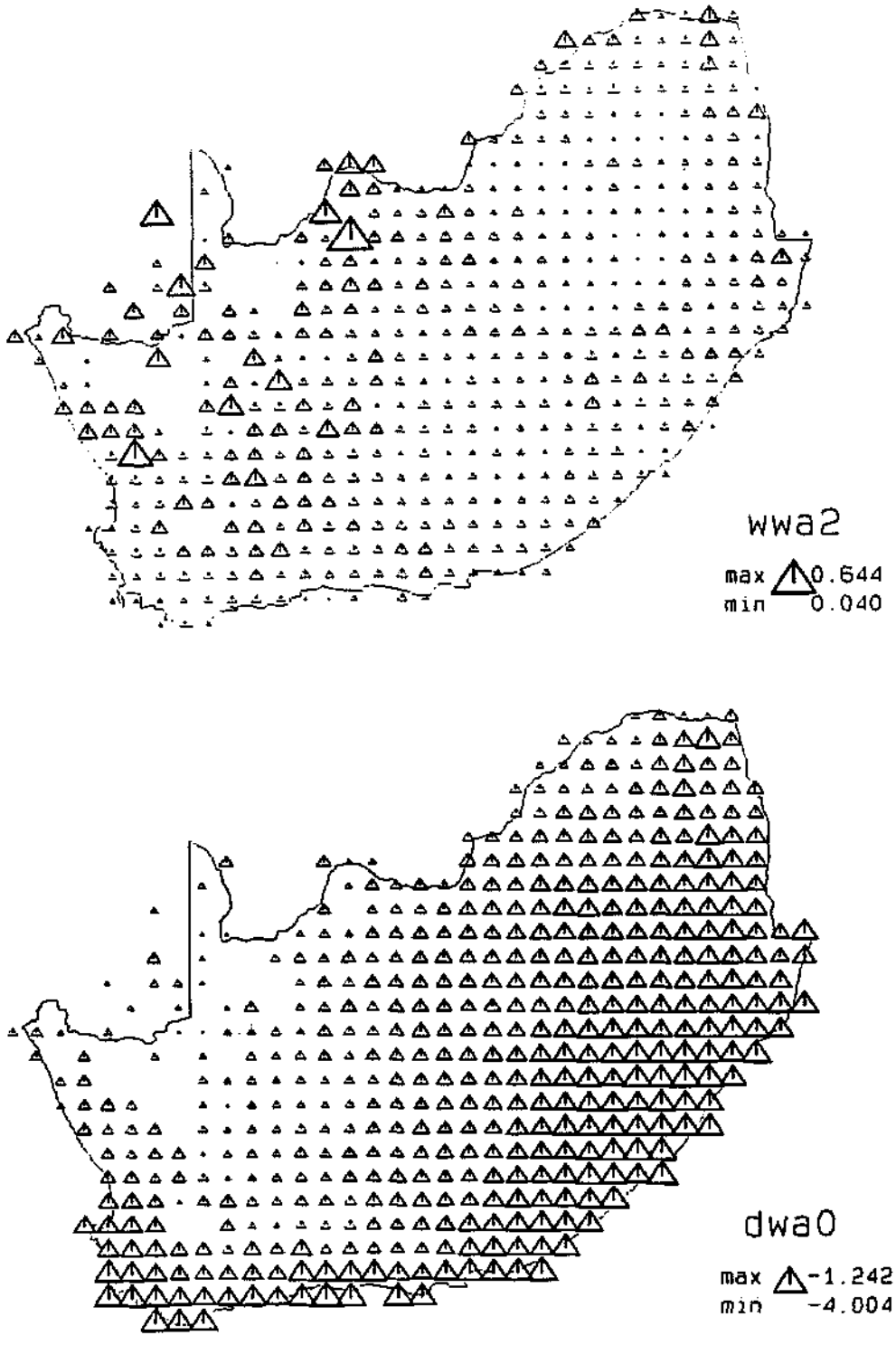


Figure 3.2: Mean parameter values for each Weather Bureau block (contd.).

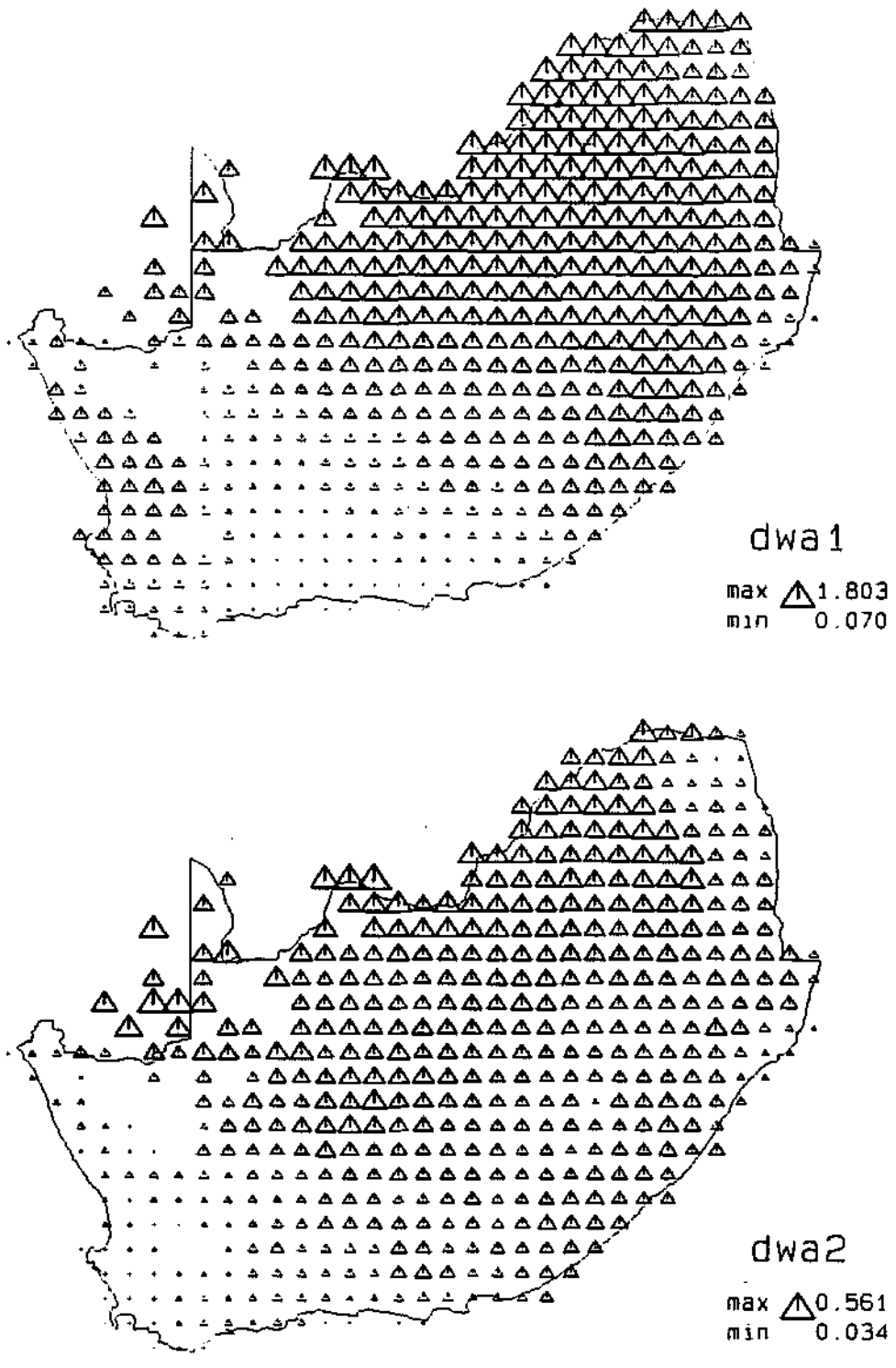


Figure 3.2: Mean parameter values for each Weather Bureau block (contd.).

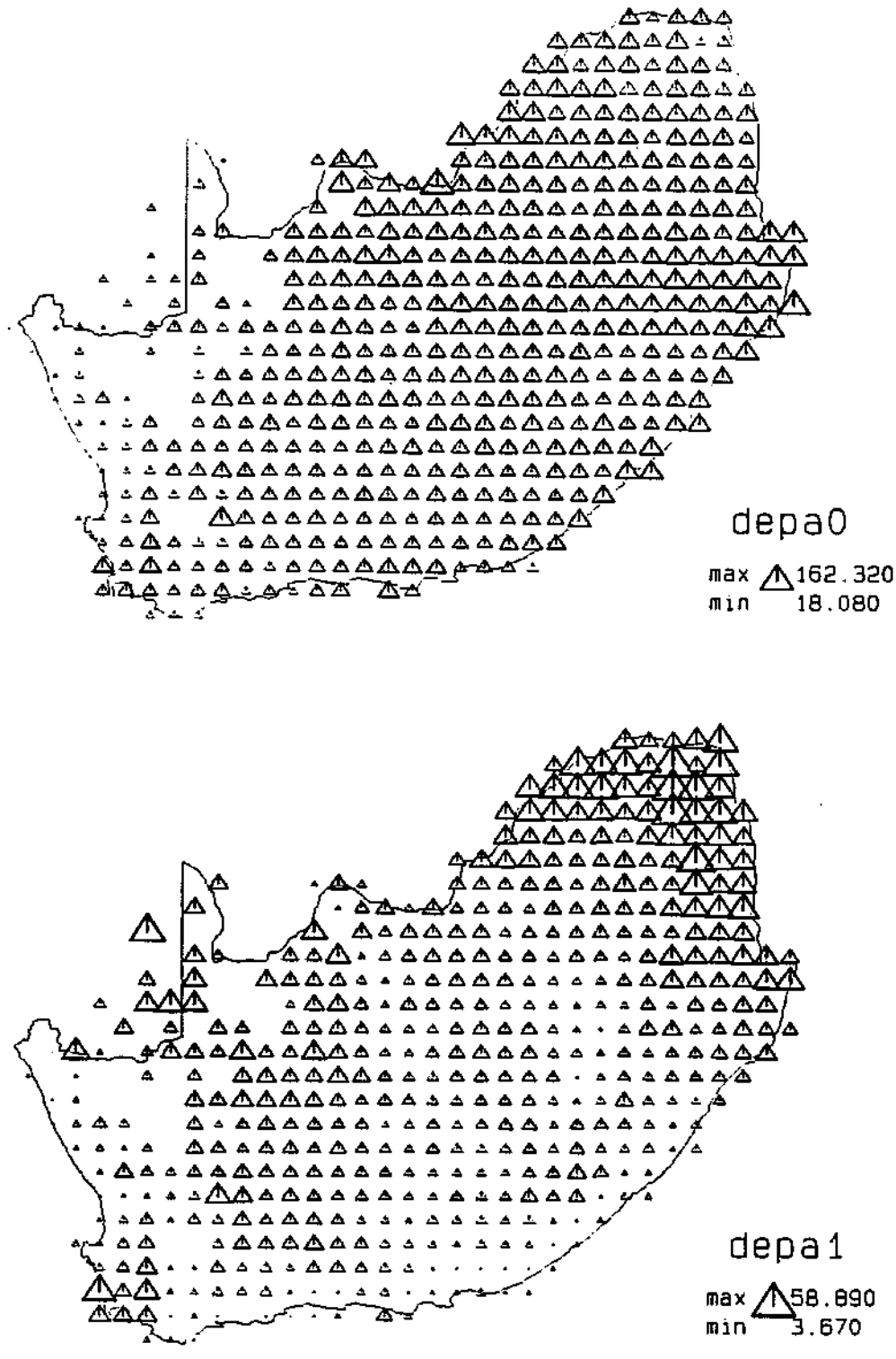


Figure 3.2: Mean parameter values for each Weather Bureau block (contd.).

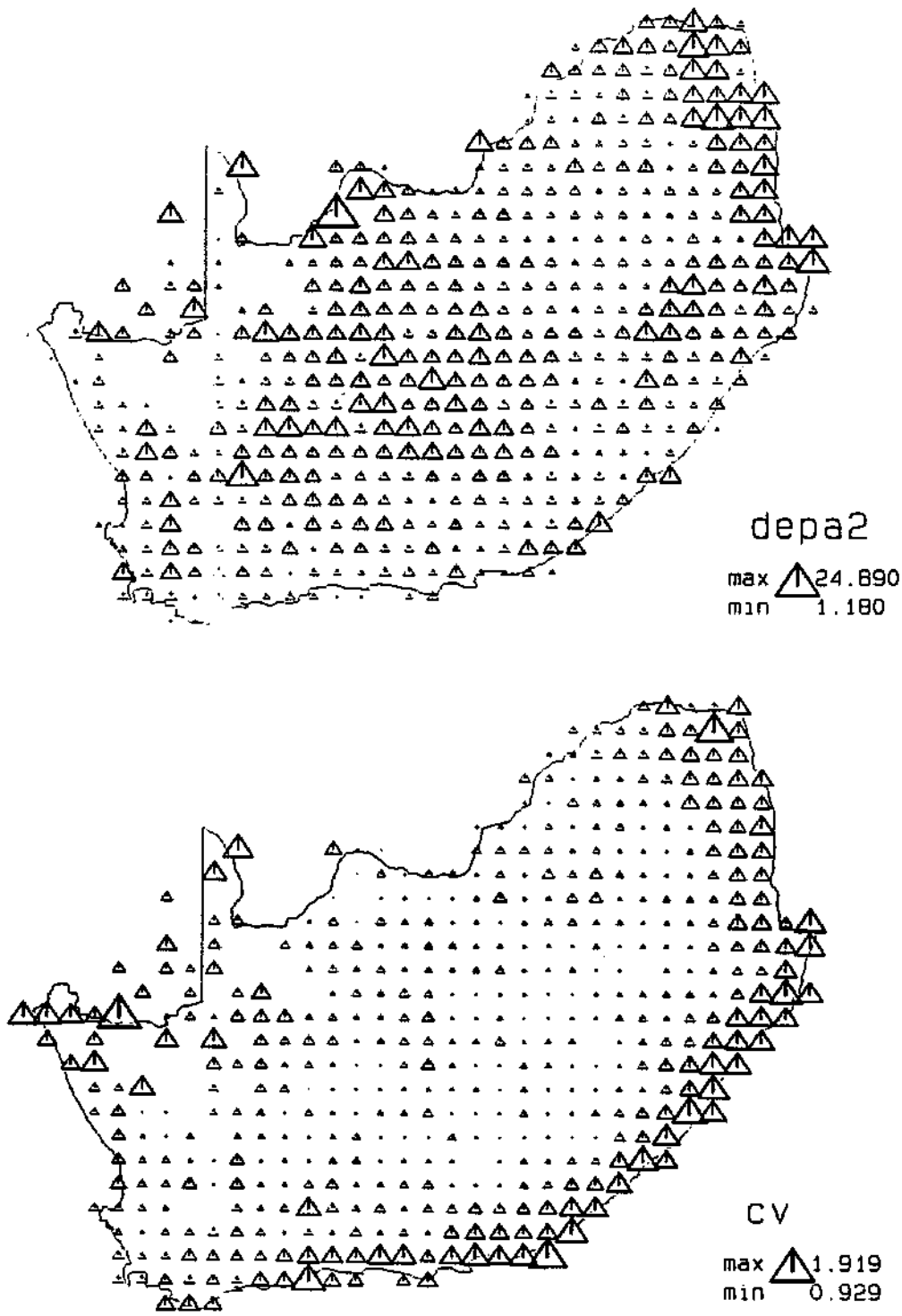


Figure 3.2: Mean parameter values for each Weather Bureau block (contd.).

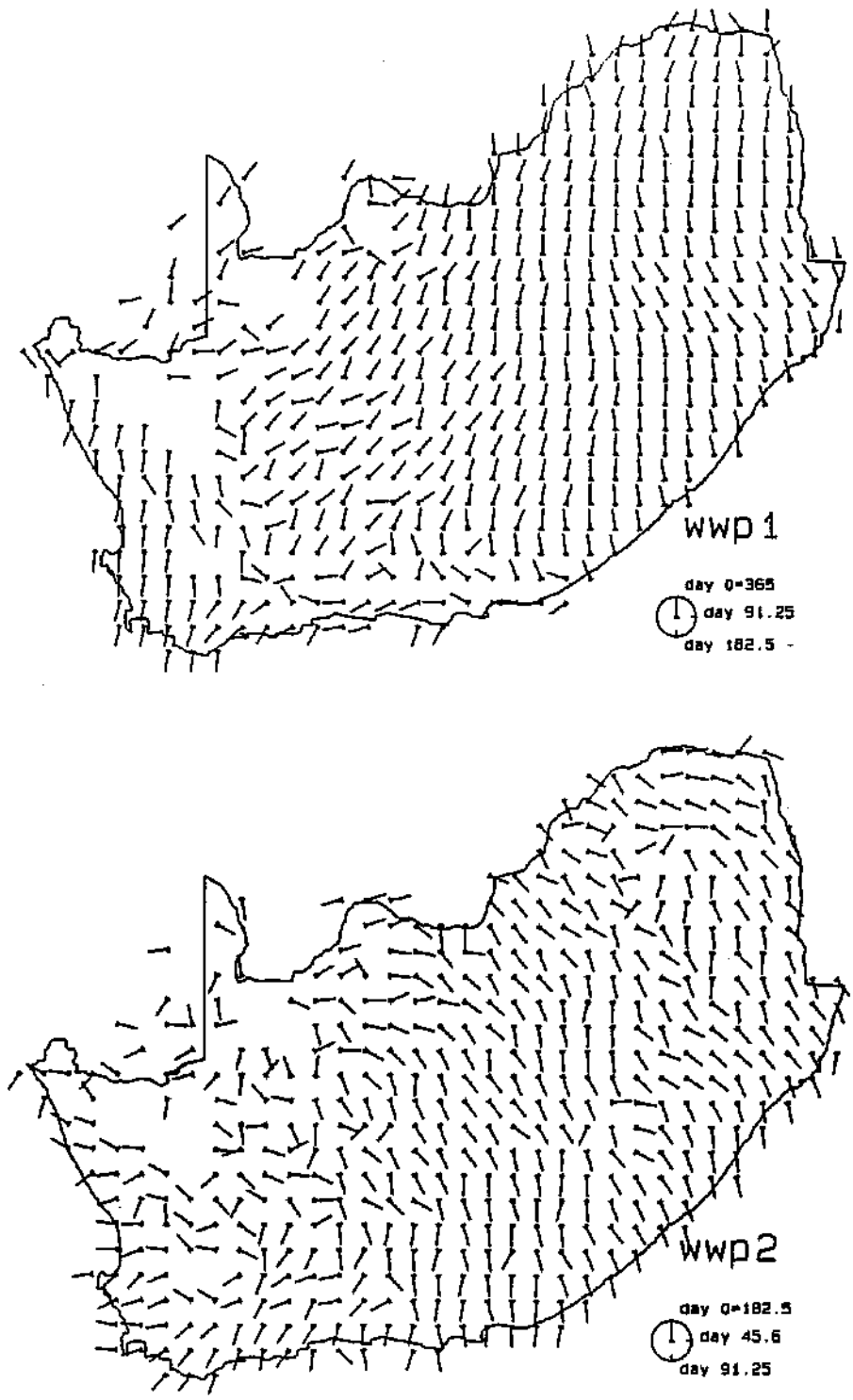


Figure 3.2: Mean parameter values for each Weather Bureau block (contd.).

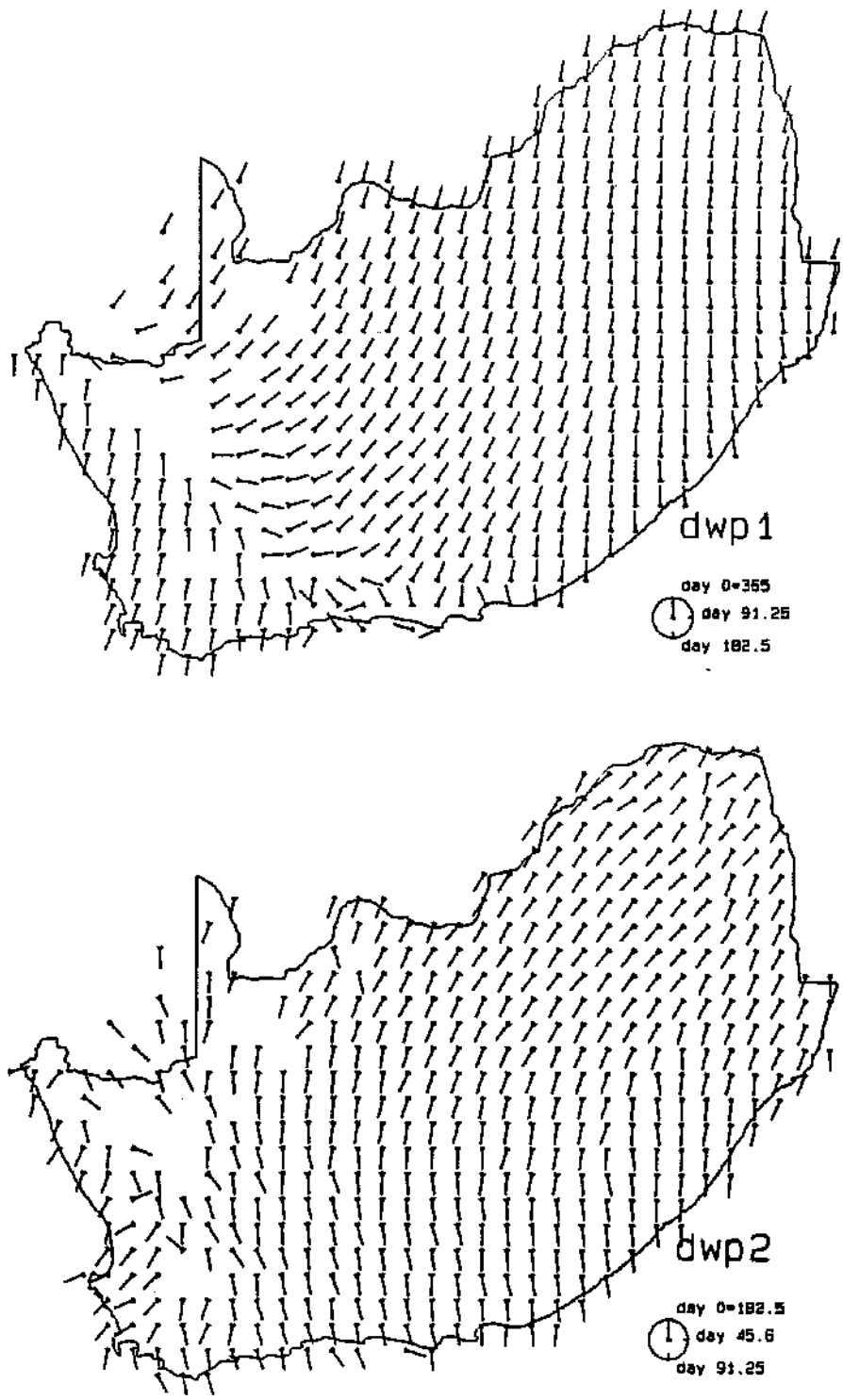


Figure 3.2: Mean parameter values for each Weather Bureau block (contd.).

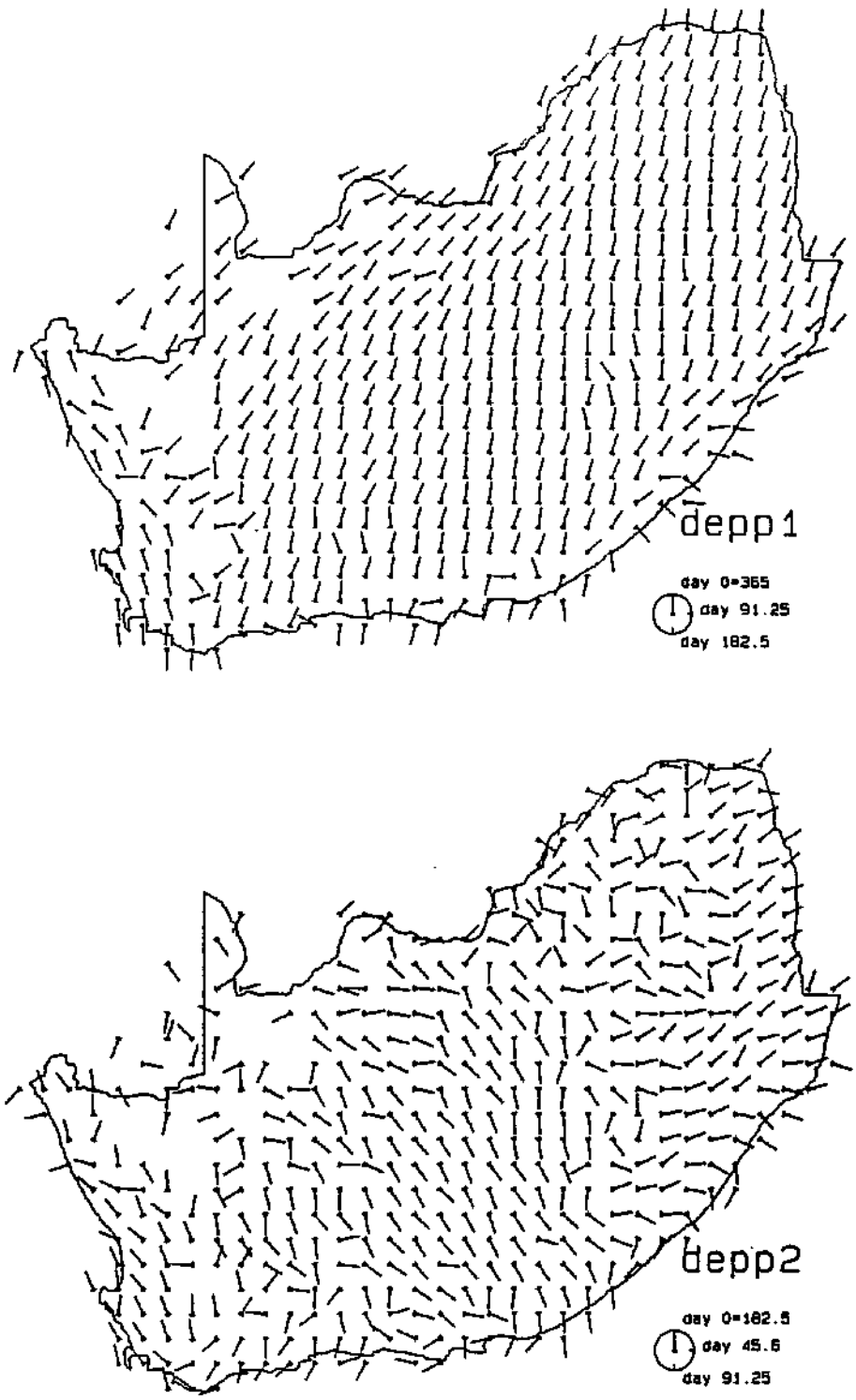


Figure 3.2: Mean parameter values for each Weather Bureau block (contd.).

Chapter 4

Variances of Model Parameters

To apply the kriging method for the interpolation of the rainfall model parameters, it is necessary to estimate the standard errors of the parameter estimates. Initially a classical methodology was applied to solve this problem. Although we were successful in deriving suitable formulae for the parameters of the Markov chain module of the model, maximum likelihood estimation of the Weibull distribution that relates to rainfall depths on rainy days leads to difficulties. The theory of the maximum likelihood estimation method and the reasons why we had to abandon this approach are discussed in Appendix A.

4.1 The Parametric Bootstrap Method

Efron (1979) proposed a methodology called the *bootstrap* by which for a random sample $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ from an unknown probability distribution F, one can estimate the sample distribution of a specific random variable V, on the basis of the observed realizations of $\mathbf{X}$, $\mathrm{x} = (x_1, x_2, \ldots, x_n)$, where V possibly depends on both $\mathbf{X}$ and F. The bootstrap method is explained here as it has been applied to our situation.

In Chapter 3 we discussed fitting models to the occurrence and non-

occurrence of rainfall (the discrete part of the model) as well as to rainfall depths on days when rain occurs (the continuous part). Parameter estimates for the probability that a wet day follows a wet day and that a wet day follows a dry day, for mean rainfall and for the coefficient of variation are then obtained. The problem to be solved is to get some measure of accuracy of these estimates, namely their standard errors. The bootstrap method gives us a way to do this. We have a special case of the bootstrap method in that we know that the probability distribution of the number of times a wet day follows a wet day is the binomial as is the probability distribution that a wet day follows a dry day and that rainfall depths follow a Weibull distribution. We therefore apply a parametric bootstrap procedure.

Let $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_{16})$ denote all the parameter estimates of the rainfall model, that is the parameter estimates for the probability that a wet day follows a wet day, the probability that a wet day follows a dry day, for the mean rainfall and for the coefficient of variation. Denote the probability distributions that describe the rainfall model by $\mathbf{F} = (F_1, F_2)$, that is, F_1 is the binomial distribution and F_2 is the Weibull distribution. Then the bootstrap algorithm is given by:

Algorithm

Step 1: Generate NY years of daily rainfall observations given the parameter estimates $\widehat{\boldsymbol{\beta}}$ and probability distribution $\widehat{\mathbf{F}}$. Denote this by $X^*(i,t)$, $i = 1, 2, \ldots, NY$ and $t = 1, 2, \ldots, NT$. This is called the bootstrap sample.

Step 2: Estimate the model parameters for the bootstrap sample in the same manner as $\widehat{\boldsymbol{\beta}}$ was obtained. Denote these parameter estimates by $\widehat{\boldsymbol{\beta}^*}$.

Step 3: Repeat Step 1 and Step 2 NB times.

Step 4: From the repeated Monte Carlo sampling in Step 3 we obtain a random

sample of parameter estimates

$$\widehat{\beta}^{*1}, \widehat{\beta}^{*2}, \ldots, \widehat{\beta}^{*NB}$$

which can be used to estimate the bootstrap distribution of $\widehat{\beta}^{*}$.

Step 5: Approximate the sampling distribution of $\widehat{\beta}$ by the bootstrap distribution of $\widehat{\beta}^{*}$. In our case we are estimating the standard error of the parameter estimates by the standard deviation of $\widehat{\beta^{*}}$.

4.2 Implementing the Bootstrap Method

The first step in implementing the bootstrap procedure is to choose the number of years (NY) of daily rainfall sequences to generate in each bootstrap sample and how many bootstrap repetitions to perform (NB). For this project it is appropriate to set NY equal to the number of years of the daily rainfall record at any given station so as to reflect the variance of parameter estimates based on a sequence of this length. NB is usually chosen to be a large number, say 1000, so as to obtain an accurate estimate of the variance. However, having to perform 1000 bootstraps for every single rainfall station would be an immense task. That is, for every one of the ±5000 rainfall stations one would have to generate NY (in the region of 60) years of daily rainfall sequences 1000 times, and for each of these sequences, compute 16 parameter estimates as well as their mean and variance. On the other hand, too few bootstrap replicates will not give the accurate results. Thus, a decrease in the number of bootstrap samples generated must not be at the cost of accuracy of the final results. We used the following strategies.

For a subset of rainfall stations, referred to as test stations, standard errors were obtained with NB set to 50, 100, 200, 300, ...1000. The resulting standard errors were compared and a decision was taken to perform

100 bootstraps in all subsequent runs as the standard errors did not differ significantly for values of NB between 100 and 1000.

Secondly, we investigated the possibility that the standard error at each station could be related to the number of years of rainfall data available. If such a relationship existed it would indicate that the bootstrap standard errors for parameter estimates could be derived as a function of the number of years of rainfall data; that is, it would not be necessary to perform the bootstrap procedure for every rainfall station in southern Africa. For the test stations, the bootstrap variance was plotted against the number of years in the historical rainfall record. No clear pattern was found; it appeared that other factors such as the geographical location, which in turn determines the variability of rainfall, also have a major effect on the variance of the parameter estimates. It was therefore decided that the bootstrap procedure had to be performed for all rainfall sites. Figure 4.1 shows the plot of the bootstrap variances versus the number of years for all model parameters, at all sites. The plots show that the variance of the parameter estimates does decrease as the number of years of data increase. It is also interesting to note that there is levelling off, that is, beyond about 60 years there is relatively little decrease in the variance for most parameters.

4.2.1 Checking the Bootstrap Method

As already mentioned, we were successful in obtaining standard errors for the Markov chain part of the rainfall model from approximations to large sample theory, that is, the inverse of the negative matrix of second derivatives of the log-likelihood function provides information relating to the accuracy of the parameter estimates. This provides us with a way of testing the ability of the bootstrap method to give satisfactory standard errors of the rainfall model parameter estimates. The standard errors obtained from the bootstrap

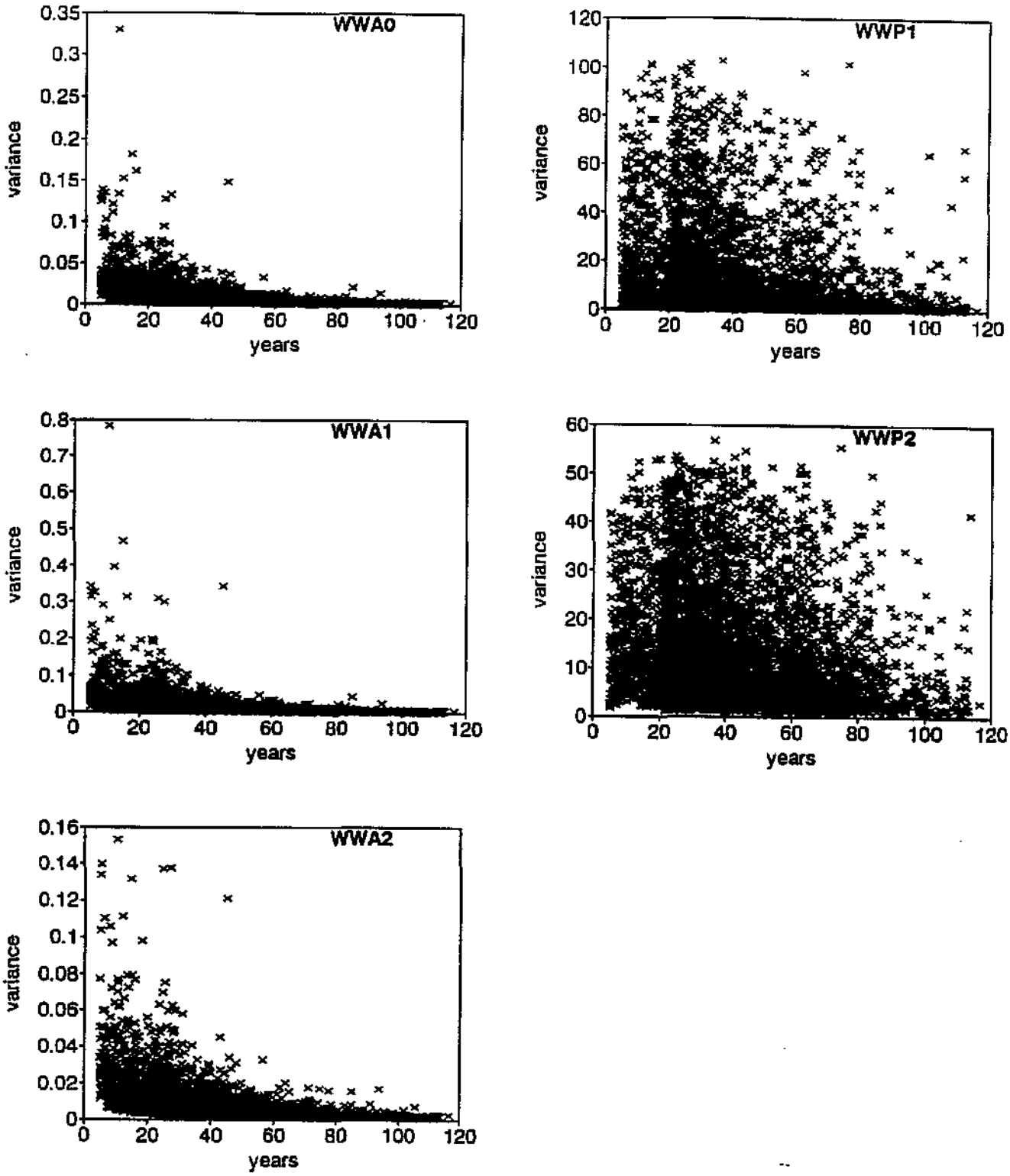


Figure 4.1: Bootstrap variances versus number of years.

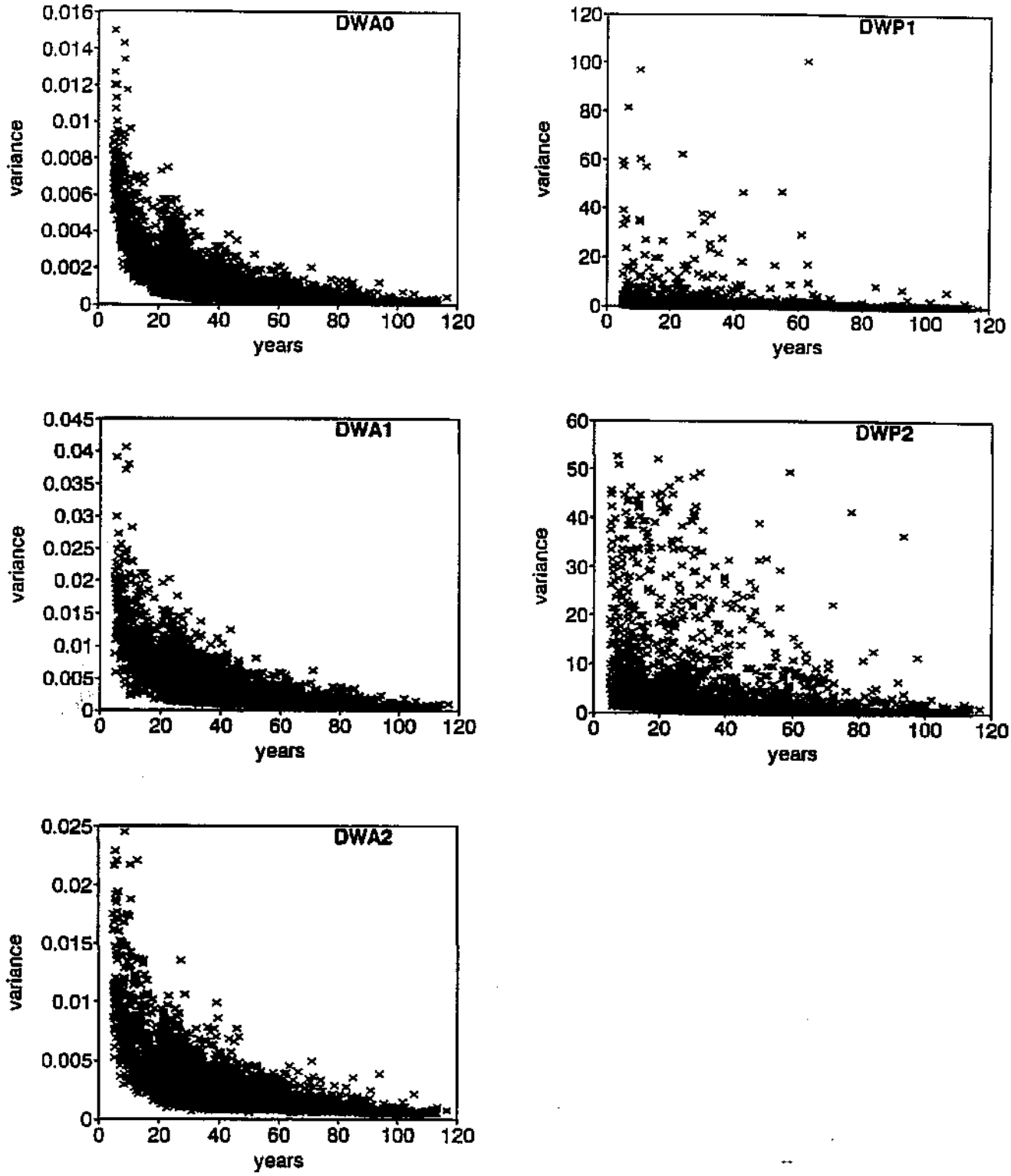


Figure 4.1: Bootstrap variances versus number of years (contd.).

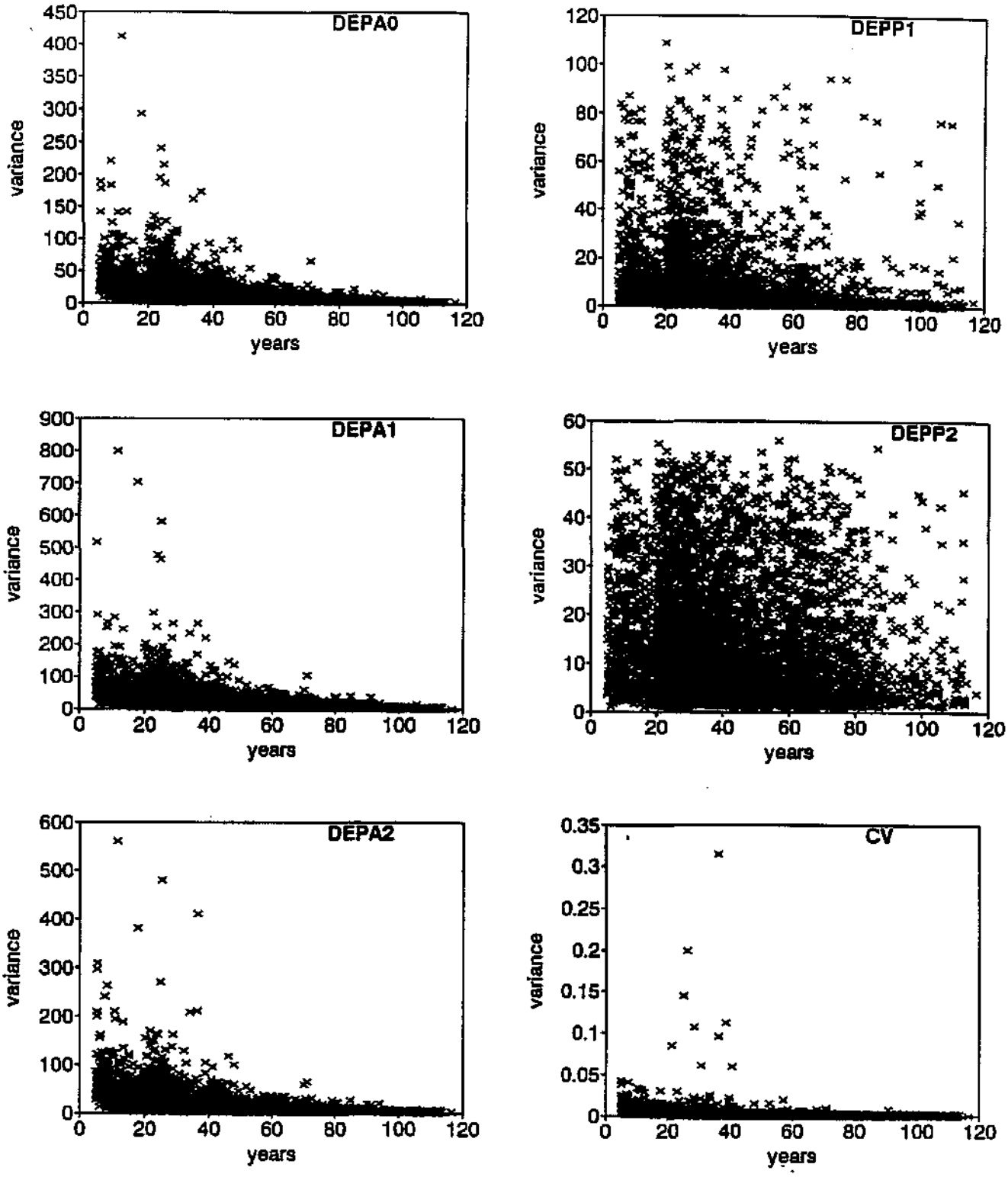


Figure 4.1: Bootstrap variances versus number of years (contd.).

samples for the Markov chain model were compared to those obtained by the classical approach. These compared very favourably and since, at least for the model for the occurrence of wet and dry days, the bootstrap method is an acceptable method to obtain standard errors for the parameter estimates, we assumed that it will also give appropriate estimates for the standard errors for the rainfall depth parameters. In order to have standard errors of all the parameter estimates computed in a uniform way, bootstrap standard errors were used for the Markov chain model as well as for the rainfall depth model in the subsequent interpolation of model parameters.

4.3 Conclusion

The bootstrap method was examined as a possible way to obtain standard errors for the rainfall model parameter estimates. The procedure was found to give satisfactory results and therefore bootstrap variances and means for all parameter estimates were computed for all the selected stations mentioned in Chapter 2. The bootstrap method not only produced variances for the parameter estimates, but it also provided us with a further check on how well the rainfall model behaves. For each station, 100 bootstrap samples were generated and the model parameters estimated. One would expect the mean of the parameters from the bootstrap samples to be very close to the parameter estimates obtained from the historical record. There were a few stations that showed a significant difference between the bootstrap means and the original parameter estimates. Cut-off points were established for the maximum permissible difference between the bootstrap mean and the original parameter estimates with the aid of histograms and the scatterplots for the various parameter estimates. The few stations that exceeded this maximum difference were considered as outliers and were removed from the data set. The final number of rainfall stations included in the remaining

analyses was 5070. Figure 4.2 shows the mean of the parameters from the bootstrap samples plotted against the original parameter estimates. As can be seen from this plot, once outlying stations were removed, the bootstrap means compare favourably with the original parameters.

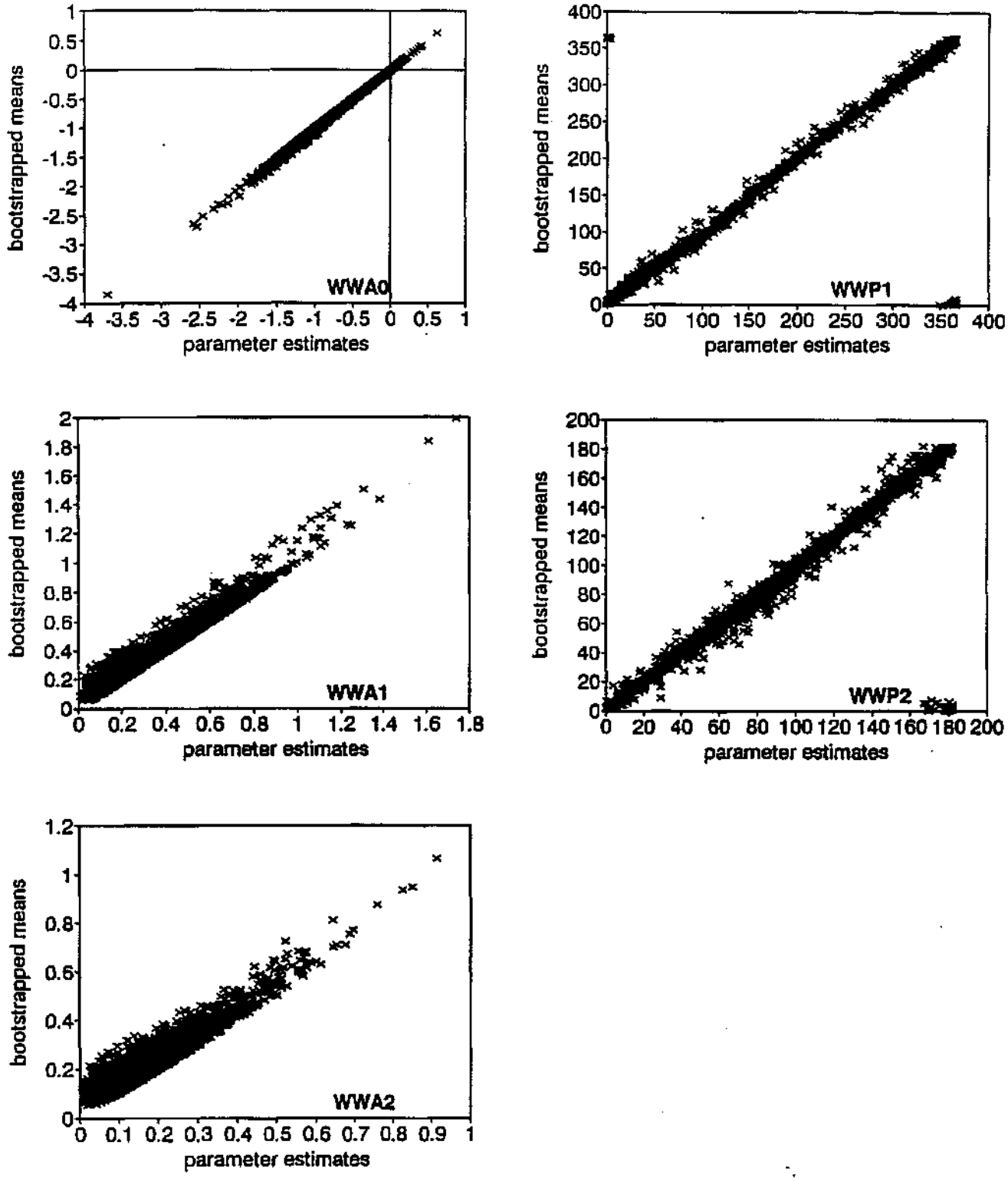


Figure 4.2: Bootstrap means versus original parameter value.

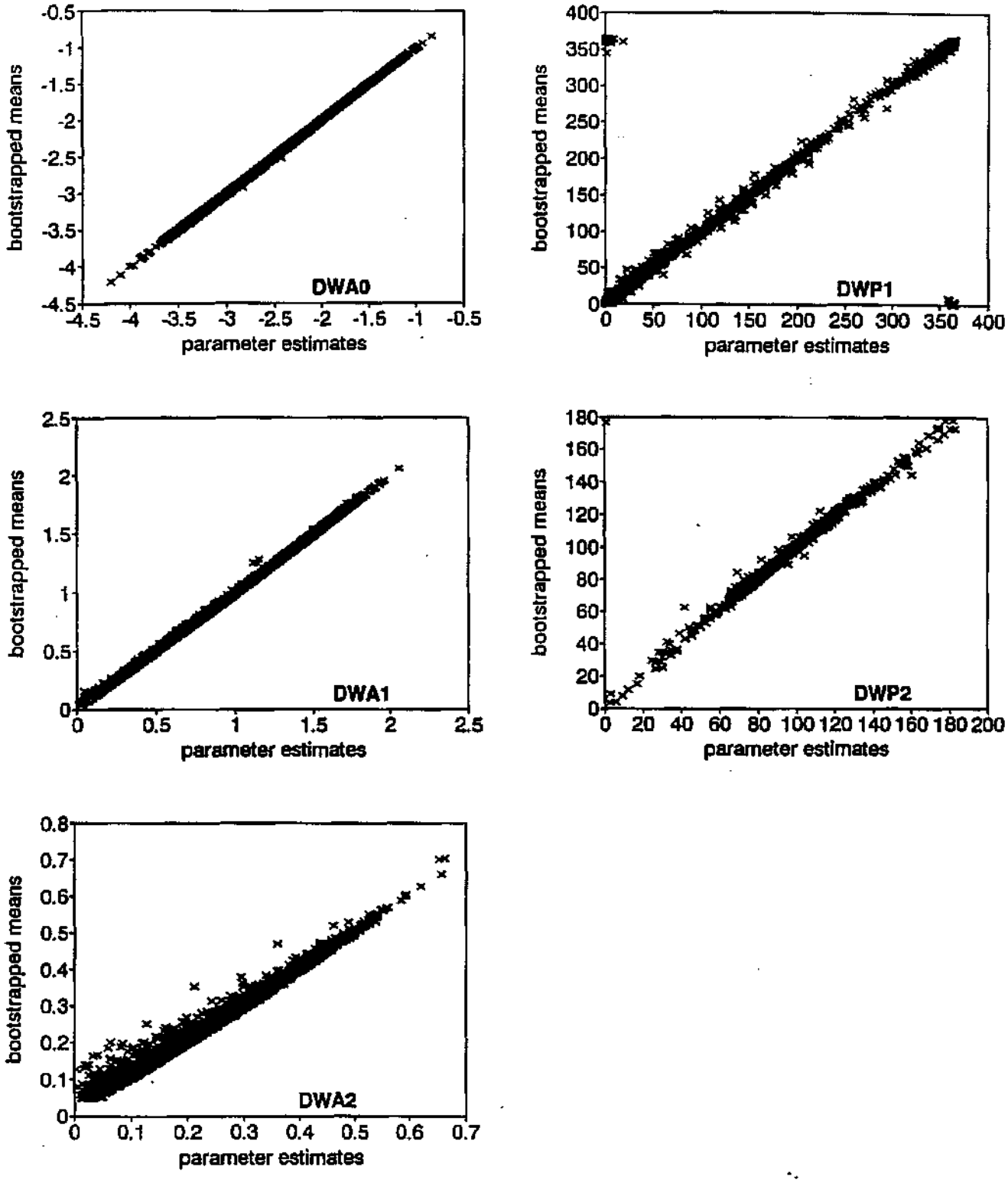


Figure 4.2: Bootstrap means versus original parameter value (contd.).

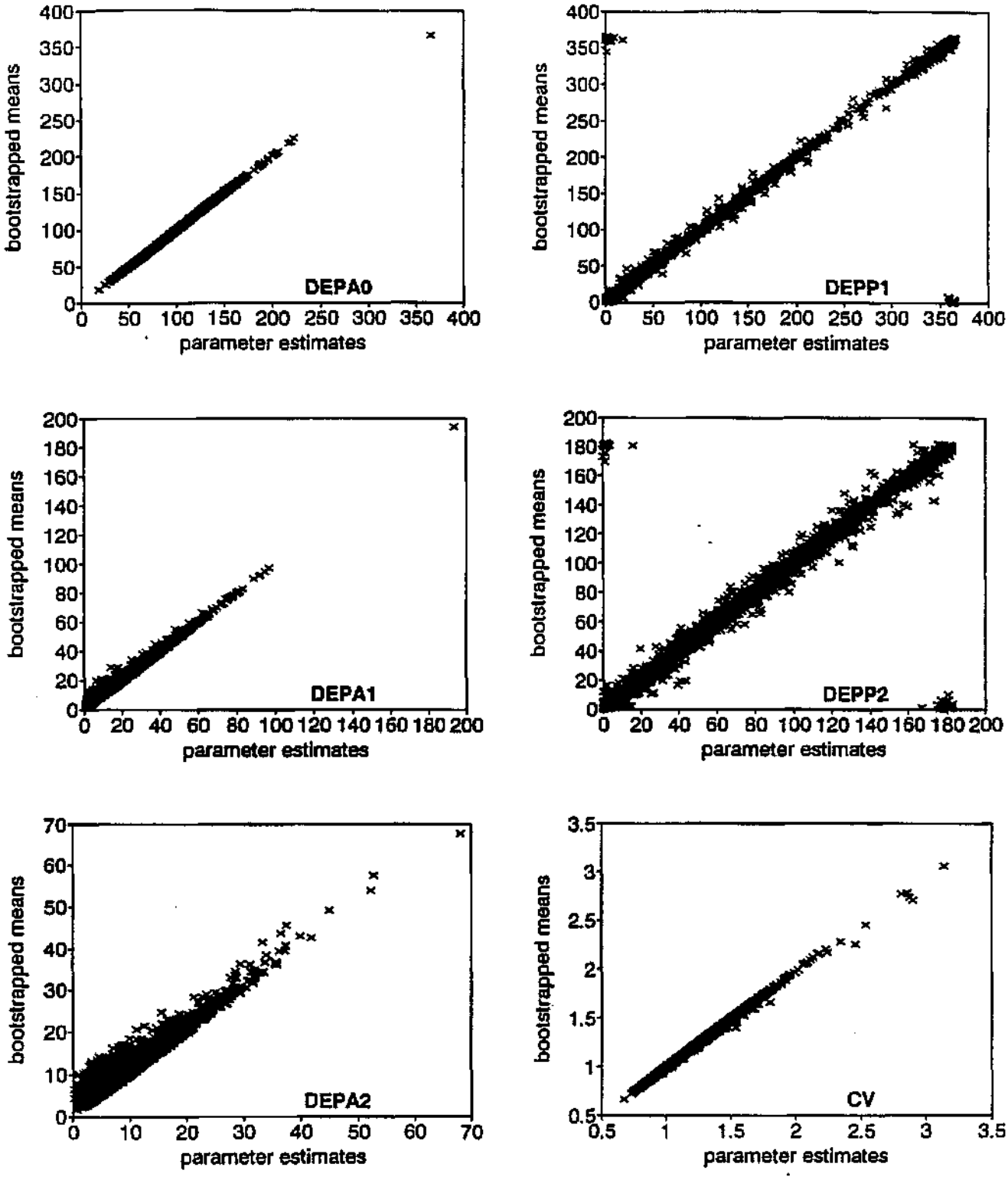


Figure 4.2: Bootstrap means versus original parameter value (contd.).

Chapter 5

Estimating the Model Parameters

Having fitted the daily rainfall model at all stations at which sufficient daily data are available it is necessary to turn to the problem of estimation of the parameters at points where no data, or too little data, are available. Specifically, our objective is to estimate the parameters on a grid of 1 minute of a degree of latitude and longitude throughout South Africa, Lesotho and Swaziland.

The map of selected rainfall stations (Figure 2.3) shows that in some areas of the country there is a high density of stations while in others, notably the north-western Cape, the data is very sparse. Available data tends to be clustered around areas of human habitation. One consequence of this is that, in mountainous regions of the country, the higher lying areas tend to be less well covered by rain gauges, so that to ignore this in the analysis would tend to give rise to under-estimation of rainfall.

Large-scale spatial patterns are clearly observable in most of the model parameters (Figure 3.2). These large scale trends may be attributed to general circulation patterns affecting the climate of southern Africa and involving the movements of large masses of air, giving rise to *frontal* rainfall. On a

smaller scale rainfall patterns are affected by the local topography and other physical features; in particular *orographic* rainfall is induced by the forced ascent of air on the windward side of mountain barriers, while *convectional* rainfall is due to updraughts caused by localized heating and can thus be affected by ground cover and land use. In all types of rainfall, rising air is cooled so that it approaches saturation; a further factor in the formation of actual rain droplets is the presence of suitable nuclei; these may be provided by ice crystals in the clouds or by other particles such as occur in dust or man-made air pollution so that, for example, large cities may have higher rainfall than the surrounding rural areas. It is clear that local anomalies can be accurately estimated only if the rainfall data is sufficiently dense in a given locality or if information on local explanatory variables is incorporated into the estimation process.

Elevation data is available on a grid of 1 minute by 1 minute throughout southern Africa (Dent *et al.*, 1989), and one would expect that local estimation of model parameters could be improved by incorporating this information. In addition, by making use of elevation data we would hope to overcome the bias in the station locations towards the lower-lying parts of each region.

One might expect that the amplitude parameters, which relate to rainfall *amounts*, would be more susceptible to topographic effects than the phase parameters which relate to *seasonality* of rainfall. This is exemplified by a comparison between the models for Tamboerskloof in Cape Town (station code 020716 W, elevation 100m) and the station at Woodhead Dam on the slopes of Table Mountain (station code 020719BW, elevation 747m) as shown in Figure 5.1. These two sites are only about five kilometres apart but show a large difference in the average level of the mean depth of rain on wet days (DEPA0), while the other parameters are very similar at the two sites. It was therefore decided to make use of the altitude information only in the

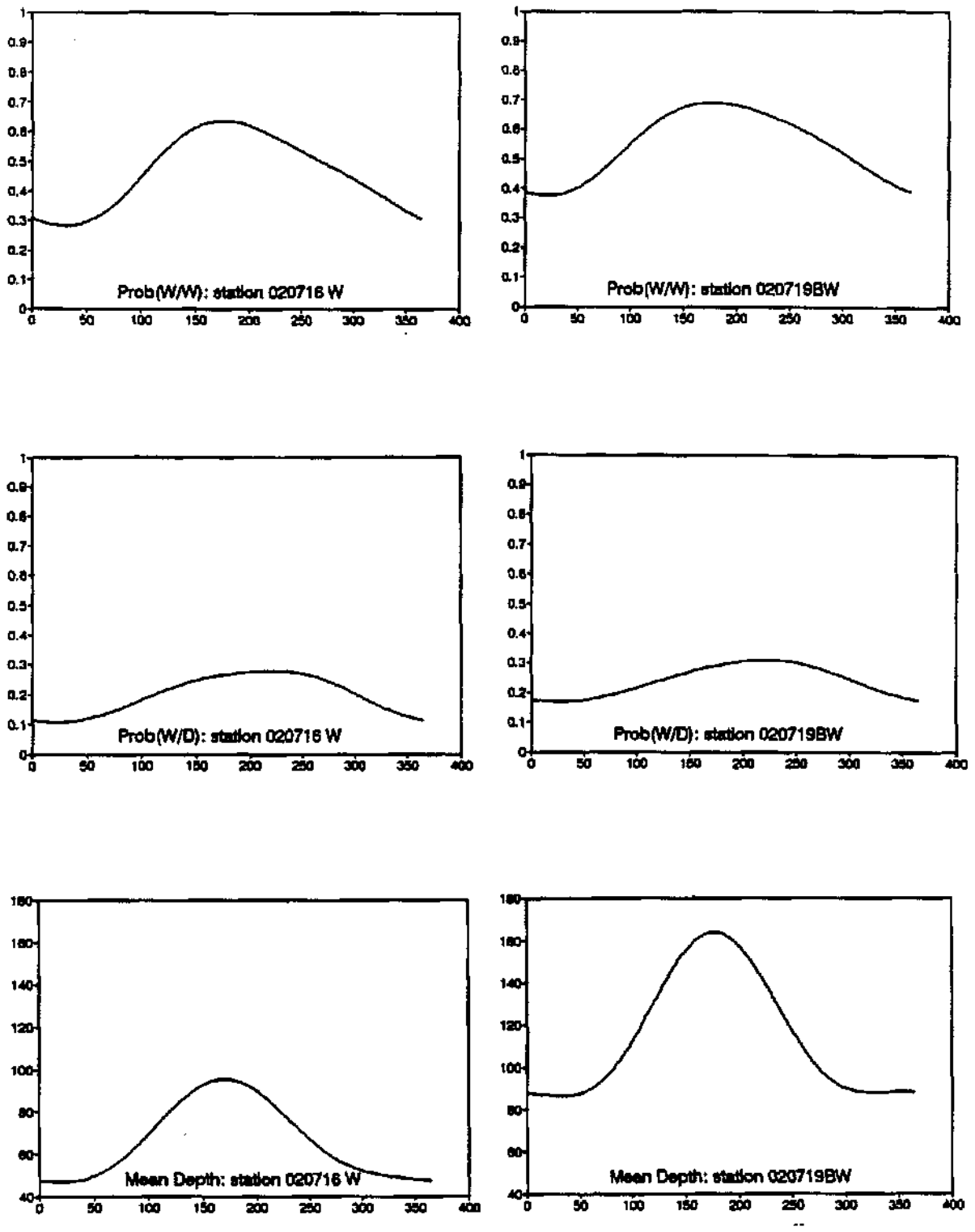


Figure 5.1: Comparison of two stations on Table Mountain.

estimation of the amplitude parameters. Further justification for this decision is provided by the semi-variogram models of the parameters, described in later sections of this chapter.

In the following section we review some of the approaches taken by previous researchers in the field of rainfall modelling to the incorporation of the effects of topography into the modelling process. We then review a number of methods for the interpolation and smoothing of spatial data and motivate the selection of *kriging* for the estimation of the rain model parameters.

5.1 Rainfall and Topography: A Review

As mentioned in the previous section, orographic rain results when air rises over mountains, so that one may expect the highest rain to occur on the windward slopes; for narrow mountain ranges the tops of the mountains and leeward slopes may also experience relatively high rain, however for more extensive mountain ranges the leeward slopes may be in a rain shadow area. This suggests that using only the altitude at a given point to predict the rain anomaly at that point will in general not be very successful, and this has been found to be the case by a number of researchers, for example, Armstrong (1992) and Creutin and Obled (1982). Thus a considerable body of research has been directed at deriving functions of the altitude at surrounding points which will be more suitable for predicting local rainfall patterns.

An early study is that of Spreen (1947) who investigated the relationship between elevation, slope, orientation (aspect) and exposure (defined below). Using a graphical regression technique Spreen found that 88% of the variation in mean winter precipitation in western Colorado could be explained by these four variables, compared with 30% for elevation alone. Other studies, using the same measures, plus a number of others such as 'roughness', have been carried out in other parts of the world, for example in New Zealand by

Hutchinson (1968), in Israel by Wolfson (1975), and in South Africa by Whitmore (1968), Schulze (1976), Hughes (1982) and Dent, Lynch and Schulze (1989).

Most of these authors have used multiple regression techniques to incorporate the topographic variables into the rainfall modelling process; a difficulty of this approach is that if the area under study is large it may first need to be segmented into homogenous sub-regions within each of which the relationship between rainfall and the topographic variables is approximately constant. Dent *et al.* (1989) initially delineated some 712 regions in their study of mean annual and monthly rainfall in southern Africa, but experienced considerable difficulty in patching together the resultant estimates at the sub-region boundaries.

All the topographic variables used by these authors are based on gridded altitude data, using a local grid centered on a given point to calculate the relevant variates at that point. Definitions of the most commonly used measures are given below.

Gradient and Aspect Given a tangent plane to the surface at any point, the gradient is the maximum rate of change in altitude on this plane and the aspect is the compass direction of this maximum (decreasing) rate of change (Skidmore, 1989). The estimate of these values will depend on the grid size and limits used, as well as the algorithm used; Skidmore compares six possible algorithms. Some authors (Spreen (1947), Hutchinson (1968)) define aspect as the direction in which the exposure (defined below) is a maximum. Aspect is a circular variable, and thus cannot be used directly in a standard regression model.

Roughness In view of the fact that the roughness of the terrain may cause updraughts and turbulence which may in turn influence the occurrence and longevity of storms (London and Emmitt, 1986) a number of re-

searchers have included a measure of roughness. Hobson (1972) gives three methods for the calculation of roughness: one based on 'bump frequency', one based on comparing estimated surface area with the corresponding planar area and a third based on the variation in the direction of normals to planar surfaces defined by adjacent groups of three elevation readings.

Exposure A number of authors have attempted to define directly a function of topography which encapsulates the fact that windward slopes tend to get higher rain due to their higher 'exposure' to the rain bearing winds. Dent *et al.* (1989) used the definition of exposure suggested by Seed (1987) which involves counting the number of points in a 5 minute by 5 minute mask which have a lower elevation than the point at the centre. Spreen (1947) used as his definition of exposure the number of one-degree sectors of a 20 mile radius circle centered on the station in which there is no land higher than 1000 feet above the station. Hutchinson (1968) used a similar definition but with a five mile radius. Hughes (1982) used an index based on the (weighted) sum of areas of grid squares with elevation higher than the gauge, taken over all squares of area 0,25 km^2 lying in a 45 degree sector oriented south-west and of radius 10 km. The weighting used was the logarithm of the excess elevation. The south-west orientation was chosen to coincide with the main rain-bearing wind direction in the area, the other aspects of the measure were chosen after a number of trials with exposure indices of varying complexity, and Hughes comments that *'the choice of a measure of exposure proved to be very difficult'*. It is clear from these different definitions that, apart from the difficulty of finding a satisfactory definition of exposure, there is almost certainly a need to 'customize' the measure for different geographical regions.

In practice all these measures are calculated as a function of the a_i where the a_i are the local values of altitude, usually available at a grid of points, and are thus influenced by the grid spacing and also by the extent of the local area or mask used in the calculation. Many of the measures can be expressed in the form $\sum w_i a_i$, that is, a linear function of the local altitudes. In view of the fact that researchers have noted the difficulty in finding a suitable measure of 'exposure' based on a priori considerations, it is appropriate to ask whether it may not be possible to use the data itself to determine, on a local basis, that function of the a_i which best explains the rainfall anomalies, and let this function provide a local definition of 'exposure' which can then be calculated at ungauged locations to predict the anomalies there. By defining a single measure in this way we could also avoid the difficulty that arises when a number of correlated measures are used as the explanatory variables in a multiple regression and also the need to consider the possible interacting effects of such variables. This approach is discussed further in Section 5.3.3.

5.2 Methods of Interpolation and Smoothing

In this section we outline the commonly used methods for fitting a surface to data available at points in two dimensions. In the case of exact interpolation, the fitted surface is required to coincide with the original values at the data points. This can be viewed as a limiting case of the more general smoothing problem, in which the fitted surface need not match the original values. For the rainfall model data, which consist of estimated model parameters, we know that there is error in the data values, as measured by the bootstrap variance, and thus exact interpolation is not appropriate.

Throughout this section we use the notation that the variable v_i (in our

case v_i represents one of the rainfall parameters) is measured at locations $\mathbf{z}_i = (x_i, y_i)'$ where $i = 1, 2, \ldots n$, and x and y represent appropriate coordinates such as longitude and latitude. The location of the point to be estimated is given by $\mathbf{z}_0 = (x_0, y_0)'$, and d_{ij} represents the distance between $\mathbf{z}_i$ and $\mathbf{z}_j$, while d_{i0} represents the distance between $\mathbf{z}_i$ and $\mathbf{z}_0$.

5.2.1 Trend Surface Analysis

In trend surface analysis a simple polynomial function such as a plane or quadratic surface is fitted to the data using ordinary least squares (Grant (1957), Krumbein (1959), Watson(1971, 1972)). For example, if the fitted function is quadratic in the x and y coordinates of the data locations, then the fitted surface has the form:

$$f(x, y) = \beta_0 + \beta_1 x + \beta_2 y + \beta_3 x^2 + \beta_4 xy + \beta_5 y^2$$

While this method may be appropriate when the trend has a relatively simple functional form, this is rarely the case in practice in the earth and atmospheric sciences, except perhaps over fairly small areas. The degree of the polynomial must be selected by the user, and in fact this is the only way in which the user can control the degree of smoothing; interpolation is possible for most data sets only by allowing the number of terms in the model to equal the number of data points. When the residuals from the trend, or local 'anomalies', are spatially correlated, as they generally are in spatial applications, use of the usual F-tests will often lead to the fitting of a surface of too high an order which is perceived by the user as 'too wavy'. Ripley (1981, Chapter 4) illustrates this effect. In addition, clustering of the data points tends to give excessive weight to the fit of the surface in the vicinity of the clusters. In the presence of spatial correlation of the residuals it would be more appropriate to use *generalized* (weighted) least squares (Draper and Smith, 1981).

The fitting of polynomial models lends itself readily to the inclusion of information on covariates, and thus the method of trend surface analysis has been popular for interpolating rainfall values, incorporating information on continentality, altitude, and other topographic features as described in the previous section. However, a problem often encountered is that the relationship with the covariates may change across the study area, and this may necessitate partitioning the area, which is in itself a major problem in that homogeneous areas must first be delineated, and also leads to the subsequent problem of patching together the various fitted equations in a smooth way, as described by Dent *et al.* (1989).

5.2.2 Smoothing Splines

The idea of fitting *local* polynomial functions leads naturally to the concept of smoothing splines. There are a number of generalizations of spline smoothing to two dimensions, but the most commonly used is the thin-plate smoothing spline (two-dimensional Laplacian smoothing spline) which can be viewed as the function f which minimizes the penalized least-squares expression

$$n^{-1} \sum_{i=1}^{n} [v_i - f(x_i, y_i)]^2 + \lambda J_2(f)$$

where

$$J_2(f) = \int\int [f_{xx}^2 + 2f_{xy}^2 + f_{yy}^2]\, dx\, dy$$

(Duchon, 1976; Wahba and Wendelberger, 1980). The degree of smoothing is controlled by the smoothing parameter λ; if λ is set to zero, the solution will interpolate the data points. If there is measurement error in the data the smoothing parameter is usually selected by generalized cross-validation; software for this is available in GCVPACK (Bates *et al.*, 1987).

Unequal error variance in the data points could be accommodated into the spline smoothing method by weighting the fit differently at individual

points. It is less clear how spline smoothing applied to the rainfall parameter values could incorporate concomitant information on topography.

5.2.3 Kriging and Optimal Interpolation

The technique of kriging was developed by Matheron (1963); the almost identical but less well known method of *optimal interpolation* was developed at about the same time by Gandin (1963). In these methods the data is modelled as a realization of a stochastic process with a covariance function which is assumed stationary, that is, dependent only on distance, at least locally, and the kriging estimate is derived as the minimum variance unbiased linear predictor. By explicitly modelling the covariance of the data points, the method is especially suited to clustered data exhibiting spatial autocorrelation.

If we use the general model

$$v_i = \tau_i + \eta_i + \epsilon_i \tag{5.1}$$

where τ represents large-scale trend, η represents the local spatially correlated component, and ϵ represents measurement error, then kriging provides an estimator of v_0 of the form

$$\sum_{i=1}^{n} w_i v_i$$

where the weights w_i are chosen to minimize the expected squared error of estimation of the measurement-error free values, that is, to minimize

$$\mathrm{E}\left[(\sum_{i=1}^{n} w_i v_i - (\tau_0 + \eta_0))^2\right]$$

In the case of so-called *simple* kriging, the data are assumed to be detrended so that the τ terms may be assumed to be zero. In this case the solution is given (Cressie, 1991) by

$$\mathbf{Kw} = \mathbf{c} \tag{5.2}$$

where the matrix $\mathbf{K}$ has elements $k_{ij} = \text{cov}(v_i, v_j)$ and $c_i = \text{cov}(v_i, \eta_0) = \text{cov}(\eta_i, \eta_0)$.

More generally, the trend term is either assumed to be a constant ('ordinary kriging'), or else modelled as a simple deterministic trend function, for example, a quadratic function of the x and y coordinates ('universal kriging'). Thus we have, at any location $\mathbf{z}$,

$$v_{\mathbf{z}} = \sum_{l=1}^{p} f_l(\mathbf{z})\beta_l + \eta_{\mathbf{z}} + \epsilon_{\mathbf{z}} \tag{5.3}$$

where the $f_l(\mathbf{z})$ are functions of x and y and the β_l are coefficients to be determined. The problem is then equivalent to a generalized least squares prediction problem, and we can write, in the usual regression notation,

$$\mathbf{v} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\eta} + \boldsymbol{\epsilon}$$

By using the Lagrange multiplier technique to introduce a constraint to ensure unbiasedness, the solution for the $\mathbf{w}$ can be shown to be (Cressie, 1991)

$$\begin{pmatrix} \mathbf{K} & \mathbf{X} \\ \mathbf{X}' & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{w} \\ -\boldsymbol{\lambda} \end{pmatrix} = \begin{pmatrix} \mathbf{c} \\ \mathbf{x}_0 \end{pmatrix} \tag{5.4}$$

where $\boldsymbol{\lambda}$ is a $p \times 1$ vector of Lagrange multipliers and $\mathbf{0}$ is a $p \times p$ matrix of zeros, the matrix $\mathbf{K}$ has elements $k_{ij} = \text{cov}(\eta_i + \epsilon_i, \eta_j + \epsilon_j)$ and $c_i = \text{cov}(\eta_i + \epsilon_i, \eta_0) = \text{cov}(\eta_i, \eta_0)$. The elements of the vector $\mathbf{x}_0$ are the values $f_l(\mathbf{z}_0)$. By partitioning the left-hand matrix in equation 5.4 as shown the solution for $\hat{v}_0$ can also be written (Goldberger, 1962 or Stein and Corsten, 1991) as

$$\hat{v}_0 = \mathbf{x}_0'\hat{\boldsymbol{\beta}} + \mathbf{c}'\mathbf{K}^{-1}(\mathbf{v} - \mathbf{X}\hat{\boldsymbol{\beta}}) \tag{5.5}$$

where $\hat{\boldsymbol{\beta}}$ is the generalized least squares estimate of $\boldsymbol{\beta}$. We see that this is equivalent to generalized least squares regression estimate of τ_0 combined with a simple kriging prediction of the value of the local component η_0, based on the regression residuals.

If the trend is assumed to be constant, then the matrix $\mathbf{X}$ above reduces to a vector of 1's and β is simply an estimate of the mean. Although it is unlikely to be reasonable to assume a constant mean throughout a large study area, it is common practice to use *local* or *moving-window* kriging, in which only the data points in the vicinity of the point to be estimated are used in the estimation process. In this case the constant mean assumption is more likely to be realistic, and Journel and Rossi (1989) have shown that local kriging with a constant mean model gives results essentially the same as those given by a more complex trend model, while avoiding some of the difficulties associated with the latter.

In many applications of kriging the measurement error term ϵ in equation 5.1 is ignored, that is, the data are assumed to be error-free. In this case kriging acts as an interpolator, so that if $\mathbf{z}_0 = \mathbf{z}_i$ then $\hat{v}_0 = v_i$.

The method of kriging can be extended to the situation where data on covariates is also available, using either *co-kriging*, in which the estimate is given by $\hat{v}_0 = \sum w_i v_i + \sum \tilde{w}_j u_j$ where the u_j are the covariate values, or alternatively by incorporating the covariates as part of the trend function. These two options are discussed further in Section 5.3

More detailed accounts of the theory and practice of kriging are given by Clark (1979), Cressie (1991) and Isaaks and Srivastava (1989).

5.2.4 Moving Average Methods (Kernel Smoothing)

These methods use a simple weighted average of the neighbouring data points, with the weights being chosen as some (inverse) function of distance or *kernel* function. Thus to estimate the value v_0 at the location $\mathbf{z}_0$ based on values v_i at locations $\mathbf{z}_i$ we have

$$\hat{v}_0 = \sum_{i=1}^{n} w_i v_i$$

where the weights w_i are given by

$$w_i = \frac{c_0}{\lambda} K\left(\frac{d_{i0}}{\lambda}\right)$$

where K is the chosen kernel function (a decreasing function of distance) and λ is known as the *bandwidth*. The constant c_0 is usually included to ensure that the weights sum to unity (Hastie and Tibshirani, 1990). Interpolation of the data points is achieved by choosing a weighting function which tends to infinity as the distance tends to zero. In general, the degree of smoothing is determined by the bandwidth and the rate of decay of the kernel function.

Kernel smoothing methods are computationally simple and do not require the assumption of any functional form for the underlying trend. They are however inappropriate for clustered data which exhibit short-scale spatial correlation, since in this case the clusters tend to dominate the smooth in their vicinity, leading to bias.

Another disadvantage of using moving average methods for estimating the parameters of the rainfall model is that there is no obvious way to incorporate covariate information. It is also not clear how one would incorporate information on heterogeneity of the measurement error into these methods, although a possible approach might be to multiply the weight of each data point by some inverse function of the variance.

5.2.5 Multiquadric Surfaces

A predictor for two-dimensional data based on the fitting of multi-quadric surfaces was proposed by Hardy (1971). The surface to be interpolated is represented by the summation of the heights of a series of n quadric surfaces, where the i'th surface has its vertex at the i'th data point. The parameters of the individual surfaces, which may be circular hyperboloids of two sheets, paraboloids or cones, are determined in such a way as to ensure that the final surface interpolates the data points. Lee *et al.* (1974) tested several types

of quadric surface for the estimation of areal rainfall and concluded that the cone was the most appropriate choice of surface, giving good estimates and being simple to compute. For the circular cone with vertex at (x_i, y_i) given by $v^2 = c_i^2((x-x_i)^2+(y-y_i)^2)$, the height at a point with coordinates (x_j, y_j) is given by

$$v_j = c_i d_{ij}$$

where d_{ij} is the distance between (x_i, y_i) and (x_j, y_j). Thus if the sum of the heights of the cones is to interpolate the data points, the constants c_i must be determined by the equations

$$\mathbf{v} = \mathbf{D}\mathbf{c}$$

where the elements of the matrix $\mathbf{D}$ are the inter-point distances, and the elements of the vector $\mathbf{v}$ are the data values. From this we see that $\mathbf{c} = \mathbf{D}^{-1}\mathbf{v}$ and hence

$$\hat{v}_0 = \mathbf{d}'\mathbf{D}^{-1}\mathbf{v}$$

where $\mathbf{d}$ is the vector of distances d_{i0}.

The method is specifically designed to interpolate the data points exactly and is thus not a general purpose smoother. In fact, the equation above shows that the solution is in fact a special case of simple kriging, with a linear function of distance used for the covariance function.

5.2.6 Selecting a Smoothing Method

A number of researchers have compared some or all of these methods on real and simulated data. Creutin and Obled (1982) tested splines, optimal interpolation and kriging, amongst other methods, to estimate rainfall amounts in southern France, while Tabios and Salas (1985) used trend surface analysis, kriging, optimal interpolation, multi-quadric interpolation and inverse-distance averaging to estimate annual precipitation in Nebraska and Kansas.

Both studies used a number of known sites as test sites to evaluate the various methods being tested. Generally the more sophisticated methods gave better results. In their conclusions, Creutin and Obled recommended optimal interpolation while Tabios and Salas recommended optimal interpolation or kriging.

Despite the apparent differences between these approaches, there are numerous connections between them. Some of these have already been mentioned in the discussion above but some further points are worth noting. Firstly, kriging can be seen as an extension of trend-surface analysis which uses generalized least squares in place of ordinary least squares to take account of the spatial correlation in the residuals, and also uses a local smoothing of the residuals to extract further predictive value from them.

There is also a formal equivalence between kriging and splines; as shown by Kimeldorf and Wahba (1970) and Watson (1984). More practical aspects of the relationship between the two methods are also discussed in some detail by Wahba (1990) and Cressie (1990).

Silverman (1985) discusses the relationship between splines and kernel smoothers and shows that the one-dimensional cubic smoothing spline is (approximately) equivalent to a kernel smoother with a bandwidth which is varied according to the local density at each data point used in the estimation, so that more clustered points are thus down-weighted.

In view of the discussions above the choice of an appropriate methodology for smoothing the rain model parameters would appear to be essentially between a spline-based approach or a geostatistical approach. While the computational techniques used in splines have perhaps been better developed than those in kriging, the geostatistical approach was selected for the following reasons:

- The model based formulation of kriging makes it suitable for the ac-

commodation of a varying error variance.

- A technique for including information on covariates via 'co-kriging' or 'kriging with external drift' already exists. In fact, as discussed in Section 5.3, these approaches make it possible to relate the rainfall at one point to the altitude at a number of neighbouring points, thus largely obviating the need to try to pre-define functions such as 'exposure' which have previously been used for this purpose.

5.3 Estimation of the Amplitude Parameters

In this section we consider the estimation of the nine amplitude parameters of the daily rainfall model and the coefficient of variation. Estimation of the phase parameters, which are circular in nature, and thus need to be treated rather differently, is discussed in Section 5.4

The approach taken is univariate; that is, each of the parameters is estimated independently of the others. Although there will probably be some spatial correlations between the parameters, which might suggest some advantage to be gained from a multivariate approach, possibly using co-kriging, it has been found in practice that co-kriging is generally only beneficial when the covariates are sampled more densely than the variable of interest. For the rainfall model, the data locations are the same for all parameters, so that a multivariate kriging is unlikely to give much advantage, and would be considerably more complex.

In order to use the kriging approach it is necessary first to model the spatial covariance function of each parameter; if co-kriging is to be used to incorporate the topographical information then the relevant cross-covariances are also required. Alternatively, if one is using kriging with external drift,

then one must select appropriate topographical variables to be included in the drift function. We consider each of these aspects in turn.

5.3.1 Estimation of the Spatial Covariance Function

The spatial covariance defines the covariance of two points as a function of the distance between them. That is,

$$\sigma(\mathbf{h}) = E[(v_{\mathbf{z}} - \mu_{\mathbf{z}})(v_{\mathbf{z}+\mathbf{h}} - \mu_{\mathbf{z}+\mathbf{h}})]$$

where μ denotes the mean value at a given location. In kriging applications it is more common to work with the *semi-variogram* function, defined as

$$\gamma(\mathbf{h}) = 1/2E[(v_{\mathbf{z}} - v_{\mathbf{z}+\mathbf{h}})^2] \tag{5.6}$$

The term semi-variogram is due to Matheron although its use had been recommended earlier in a time series context by Jowett (1952). There are a number of advantages in working with the semi-variogram, using the estimator

$$\hat{\gamma}(\mathbf{h}) = 1/(2N_{\mathbf{h}}) \sum (v_{\mathbf{z}_i} - v_{\mathbf{z}_j})^2$$

where the summation is over all $N_{\mathbf{h}}$ pairs which are a vector distance $\mathbf{h}$ apart. In practice, for non-gridded data, the summation is calculated over all pairs belonging to a number of distance *intervals*, for example, 0-1 km, 1-2 km, 2-3 km etc. If the spatial continuity is more marked in some directions than others, then it is necessary to calculate separate semi-variograms for each direction, but often there is no directional effect so that we need only consider the distance $h = \| \mathbf{h} \|$.

One of the advantages of working with the semi-variogram is that its estimation does not require any prior estimate of the mean; in addition, the estimate is relatively little affected by trend for small values of h, these being the more critical values for the kriging process, since in practice it is

usual to obtain the kriging estimate at a given location using only the data points which are in the vicinity of that location. Other reasons for preferring to use the semi-variogram rather than the covariance function are given by Srivastava (1988) and Cressie (1991).

In the case where the mean is constant there is a simple relationship between the semi-variogram and the covariance, given by

$$\gamma(h) = \sigma(0) - \sigma(h)$$

so that, having estimated the semi-variogram, one may readily obtain the corresponding covariance required for the solution of the kriging equations.

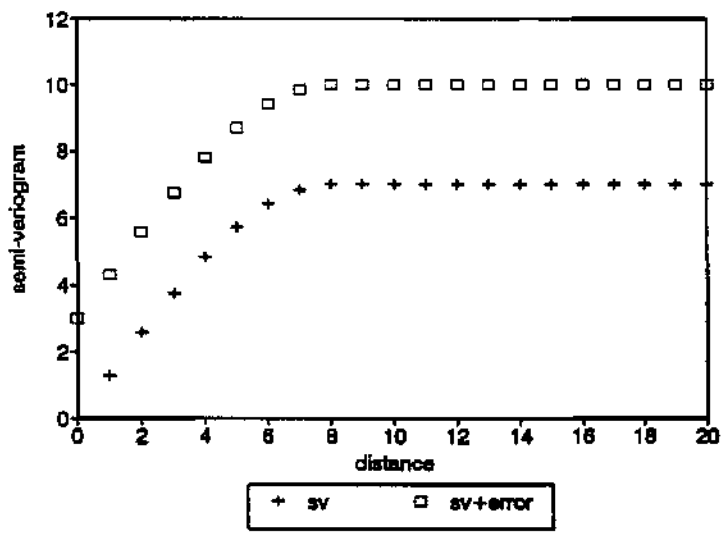


Figure 5.2: Effect of error on the semi-variogram.

When there is measurement error in the data it is necessary to break down the semi-variogram into components corresponding to the original model components given in equation 5.1 In the case where $E(\epsilon_i) = 0$, $\text{var}(\epsilon_i) = \sigma_\epsilon^2$ and the ϵ_i are uncorrelated with one another and with the η_i then it is easy

to show that

$$E[((\eta_i + \epsilon_i) - (\eta_j + \epsilon_j))^2] = E[(\eta_i - \eta_j)^2] + 2\sigma_\epsilon^2$$

and thus we have

$$\gamma_{\eta+\epsilon}(h) = \gamma_\eta(h) + \sigma_\epsilon^2 \tag{5.7}$$

so that the error term increases the semi-variogram by a constant amount equal to σ_ϵ^2 (Figure 5.2).

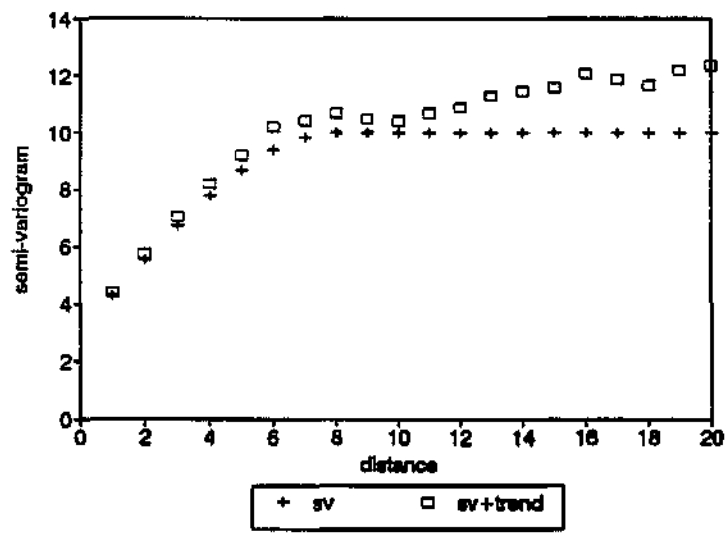


Figure 5.3: Effect of trend on the semi-variogram.

The effect of trend on the semi-variogram is dependent on the exact form of the trend. If the trend is constant then it is clear from equation 5.6 that it will cancel out of $\gamma(h)$. In general the trend will vary little over small distances, so that the effect of the trend will be relatively minor for small lag distances. Figure 5.3 shows a typical semi-variogram for a data set containing trend.

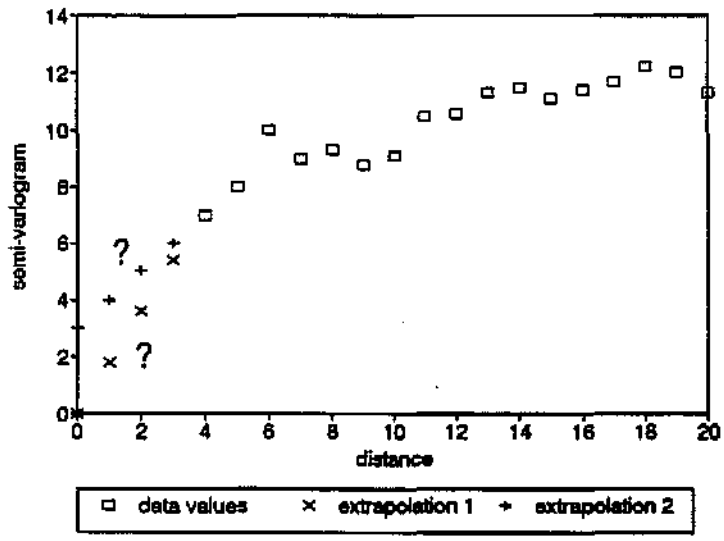


Figure 5.4: Estimating the nugget effect.

Unless there are repeated measurements at some locations, or some other way of independently estimating σ_ϵ^2, the error variance can only be estimated from the empirical semi-variogram by extrapolating the fitted model to the point $h = 0$. As it is quite possible that there is also significant short-scale variation in η, commonly referred to as the *nugget effect*, such extrapolation may be quite inaccurate (Figure 5.4).

For the rain model parameters there are three sources of apparent small-scale variation. Firstly there is the measurement error term of equation 5.1, whose variance was estimated by the bootstrap procedure. Secondly we have the inaccuracy in the measurement of the station locations, and thirdly the small-scale variation in the local component, η, of equation 5.1 which will typically be the result of topographic variability. Because the value of σ_ϵ^2 is not constant across all data points we cannot use equation 5.7 directly in this case, but since the bootstrap procedure provides individual measurement

error variances $\sigma^2_{\epsilon_i}$ corresponding to each datum v_i, $i = 1, 2, \ldots n$ we can use the fact that

$$E[((\eta_i + \epsilon_i) - (\eta_j + \epsilon_j))^2] = E[(\eta_i - \eta_j)^2] + \sigma^2_{\epsilon_i} + \sigma^2_{\epsilon_j}$$

to calculate an *adjusted* semi-variogram estimator by using

$$\hat{\gamma}_\eta(h) = 1/(2N_h)\left\{\sum (v_i - v_j)^2 - \hat{\sigma}^2_{\epsilon_i} - \hat{\sigma}^2_{\epsilon_j}\right\}$$

where the summation is, as usual, over all N_h pairs which are a distance h apart.

Figure 5.5 shows the unadjusted and adjusted semi-variograms for the nine amplitude parameters and the coefficient of variation of the daily rainfall model. For most of the parameters, the unadjusted values suggest a definite nugget effect, while in many of the graphs the adjusted values appear to pass approximately through the origin, that is, $\gamma_\eta(0) = 0$. Those which still show a nugget effect after adjustment (notably DEPA0, DEPA1, CV, WWA0, and DWA0) suggest that the corresponding parameters are sensitive to local topographical changes, or possibly other sources of small-scale variation. For one parameter, WWA2, the adjustment seems to have over-corrected, resulting in negative values throughout the empirical semi-variogram. This is probably due to the somewhat skew distribution of this parameter, and also the fact that the error variance of this parameter is relatively high compared to the actual parameter values, which are typically quite small.

Models were fitted to the semi-variograms of each of the amplitude parameters and to the coefficient of variation. The method of fitting, based on the weighted least squares method of Cressie (1985), gives more weight to those points on the semi-variogram which are based on a larger number of data pairs, and also gives more weight to those points corresponding to smaller values of h. In each case the model fitted was the sum of a spherical model for the local component η plus a linear model for the trend component τ.

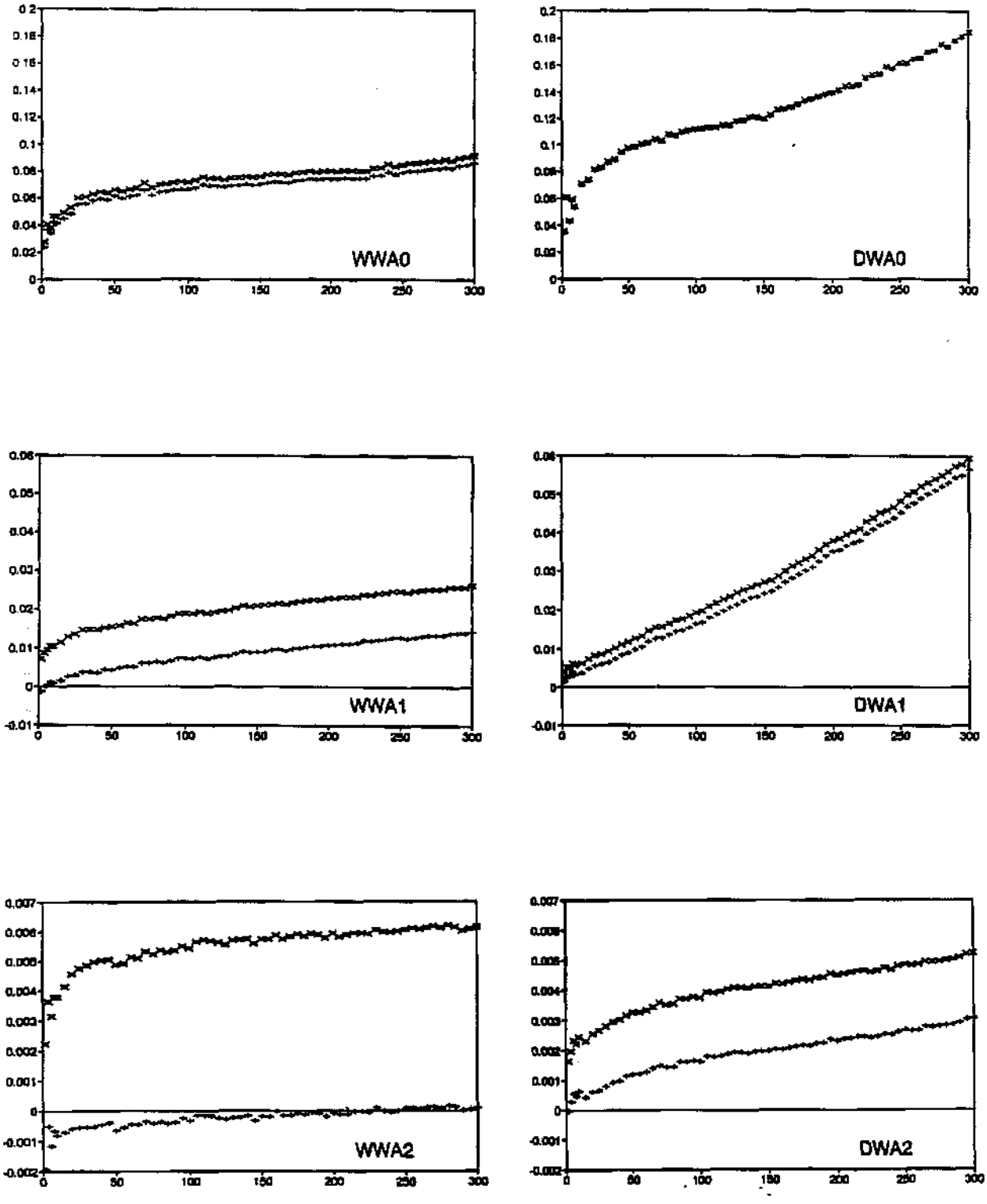


Figure 5.5: Semi-variograms: amplitude parameters.

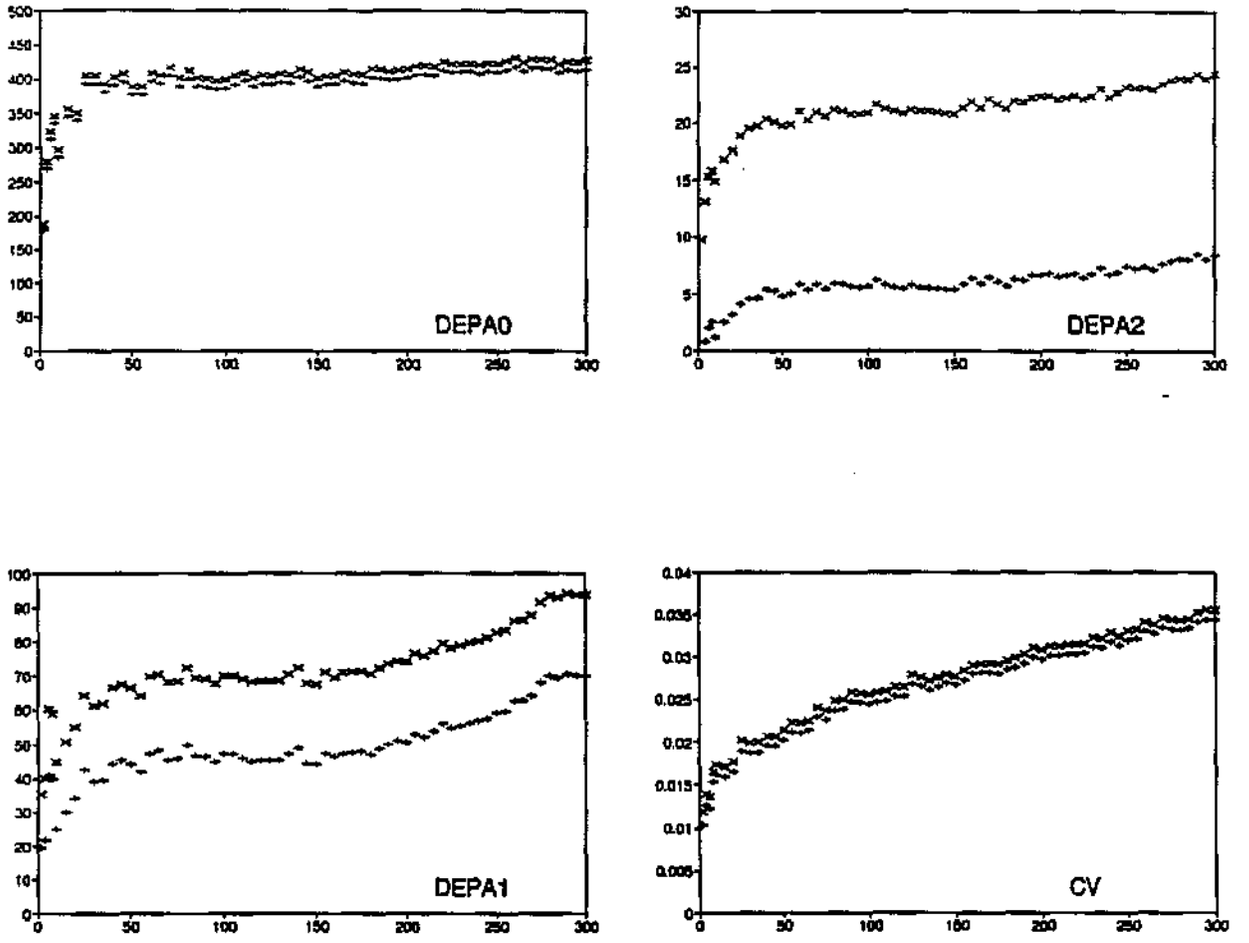


Figure 5.5: Semi-variograms: amplitude parameters (contd.).

The spherical model is given by

$$\gamma(h) = \begin{cases} a[(3/2)(h/r) - (1/2)(h/r)^3] & 0 \leq h \leq r \\ a & h > r \end{cases}$$

where a, the asymptote, is commonly known as the *sill*, while r, the *range*, indicates the maximum extent of the spatial correlation.

The linear model is given by

$$\gamma(h) = sh$$

where s is a slope parameter.

The fitted values a, r, and s, for each rainfall model parameter, are listed in Table 5.1. The ranges of the local component vary between 10 and 40 kilometres. All the models fit well up to a distance of at least 240 km, which corresponds to the maximum distance used in the local kriging calculations. These fitted models are used to provide the covariances required for the kriging calculations.

parameter	sill	range	slope
WWA0	0.0530	15	0.000140
WWA1	0.0035	36	0.000036
WWA2	0.0027	19	0.000007
DWA0	0.0790	10	0.000300
DWA1	0.0030	10	0.000140
DWA2	0.0007	12	0.000010
DEPA0	389	10	0.0400
DEPA1	47	30	0.0002
DEPA2	5	40	0.0020
CV	0.0180	10	0.000064

Table 5.1: Fitted semi-variogram models: amplitude parameters.

5.3.2 Cokriging of Rain and Altitude

One possible approach to the incorporation of altitude into the kriging process is to use co-kriging (Matheron, 1971) in which the covariates are essentially treated as an extension of the data vector so that the solution (for a single covariate) is a weighted sum of values of the variate to be interpolated and the values of the covariate. Thus we have

$$\hat{v}_0 = \sum_{i=1}^{n} w_i v_i + \sum_{j=1}^{m} \tilde{w}_j u_j$$

with constraints $\sum w_i = 1$ and $\sum \tilde{w}_j = 0$.

Assuming a no-trend model for the data, the co-kriging weights are given by

$$\begin{pmatrix} \mathbf{C_{vv}} & \mathbf{C_{vu}} & \mathbf{1} & \mathbf{0} \\ \mathbf{C_{uv}} & \mathbf{C_{uu}} & \mathbf{0} & \mathbf{1} \\ \mathbf{1}' & \mathbf{0}' & 0 & 0 \\ \mathbf{0}' & \mathbf{1}' & 0 & 0 \end{pmatrix} \begin{pmatrix} \mathbf{w} \\ \tilde{\mathbf{w}} \\ -\lambda_v \\ -\lambda_u \end{pmatrix} = \begin{pmatrix} \mathbf{c_{v0}} \\ \mathbf{c_{u0}} \\ 1 \\ 0 \end{pmatrix}$$

where the matrix $\mathbf{C}_{vu}$ is now the $n \times m$ matrix of cross-covariances, $\mathbf{C}_{uv} = \mathbf{C}'_{vu}$ and the vector $\mathbf{c}_{u0}$ also contains cross-covariances (of the u_i with v_0), while λ_v and λ_u are Lagrangian parameters. More generally, a polynomial trend model may be included, as in universal kriging. Stein and Corsten (1991) show how co-kriging with a polynomial trend function may be expressed as a generalized least squares predictor.

In using co-kriging, the covariates need not be available at the same points as the variate of interest, nor at the sites to be estimated, although the locations of the covariate information do affect the method of estimation of the cross-covariance function (see below). Co-kriging is generally most valuable when the covariates are sampled more intensely than the predictand. An application of co-kriging to the estimation of rainfall data is described by Krajewski (1987).

In order to use the co-kriging approach it is necessary to model the *cross-covariance* of u and v in addition to their respective covariances. The spatial cross-covariance of two variate v and u is defined as

$$\sigma_{vu}(\mathbf{h}) = E[(v_{\mathbf{z}} - \mu^v_{\mathbf{z}})(u_{\mathbf{z}+\mathbf{h}} - \mu^u_{\mathbf{z}+\mathbf{h}})]$$

where μ^v and μ^u denote the mean values, of v and u respectively, at the relevant location.

While the cross-covariance function may be estimated directly it is natural, in view of the advantages of the semi-variogram over the ordinary co-

variance function to look for an appropriate way of defining a *cross semi-variogram.* The traditional definition (see for example Journel and Huijbregts, 1978) is

$$\gamma_{vu}(\mathbf{h}) = \frac{1}{2}E[(v_{\mathbf{z}} - v_{\mathbf{z}+\mathbf{h}})(u_{\mathbf{z}} - u_{\mathbf{z}+\mathbf{h}})]$$

In order to estimate this function the variables v and u must be available at a number of common locations. The function is also symmetric in v and u. This implies that $\gamma_{vu}(\mathbf{h}) = \gamma_{vu}(-\mathbf{h})$ and this is not always appropriate; for example, in studying the relationship between rain and altitude, it is generally the windward slopes of mountains which receive higher rain, so that the direction of a mountain in relation to a point at which rainfall is to be estimated cannot be ignored.

Myers (1982) proposes an alternative definition which involves modelling the semi-variograms of v, u and $v + u$, and then using these to estimate the corresponding covariances, from which the cross-covariance of u and v may be obtained. This method also requires a number of common data locations and is also symmetric in v and u, and thus suffers from the same problem as the previous definition in that it does not cater for directional effects.

Clark *et al.* (1987) have suggested that a better definition of the cross semi-variogram is

$$\gamma_{vu}(\mathbf{h}) = \frac{1}{2}E[(v_{\mathbf{z}} - u_{\mathbf{z}+\mathbf{h}})^2]$$

Use of this definition does not require common data locations; furthermore the definition is *not* symmetric in v and u. It is recommended that the two variables first be standardized to zero mean and unit variance so that values are commensurate, since gross differences in scale could adversely affect the precision of computations.

For our application, where the cross-covariance will certainly show directional effects, it was decided to try modelling the cross-covariance directly. In exploring this approach the cross-covariance of the parameter DEPA0 with

altitude was studied for the south western Cape region (approximately west of Mossel Bay and south of Calvinia). Calculations were done for eight directions, and the results are shown in Figure 5.6 (in which the distance units are minutes of a degree of latitude or longitude). Figure 5.7 displays the same information in the form of a contour map. It is clear from the figures that the cross-covariance is generally positive and decreases with distance, but there is a distinct group of negative values at a lag distance of approximately 20 units (about 35 km) in a north-west direction. This corresponds with the knowledge that the main rain-bearing wind direction in this area is approximately north-west, so that it is likely that rain gauges which are sited so as to have points of high elevation to the north-west will be in the rain shadow of that higher ground and thus have reduced rain.

It is clear, however, that in order to take account, for example, of locally varying directional effects, the cross-covariance models would have to be re-calculated and fitted regionally, or perhaps, to avoid discontinuities at regional boundaries, re-computed in a neighbourhood of each point being estimated as suggested by Haas (1990). This would necessitate an enormous amount of computation. Further, there is then a need to parameterize appropriate cross-covariance models which could be used as part of an automatic fitting procedure, since it would be impractical for the user to interactively model cross-covariances at the half a million or so locations being estimated in this project. Some further research was done to investigate the feasibility of such automatic modelling, but the results were generally disappointing (Sedupane, 1992).

5.3.3 Kriging with External Drift

An alternative to the co-kriging approach is to include the covariates as part of a trend function, which is essentially similar to the kriging formulation

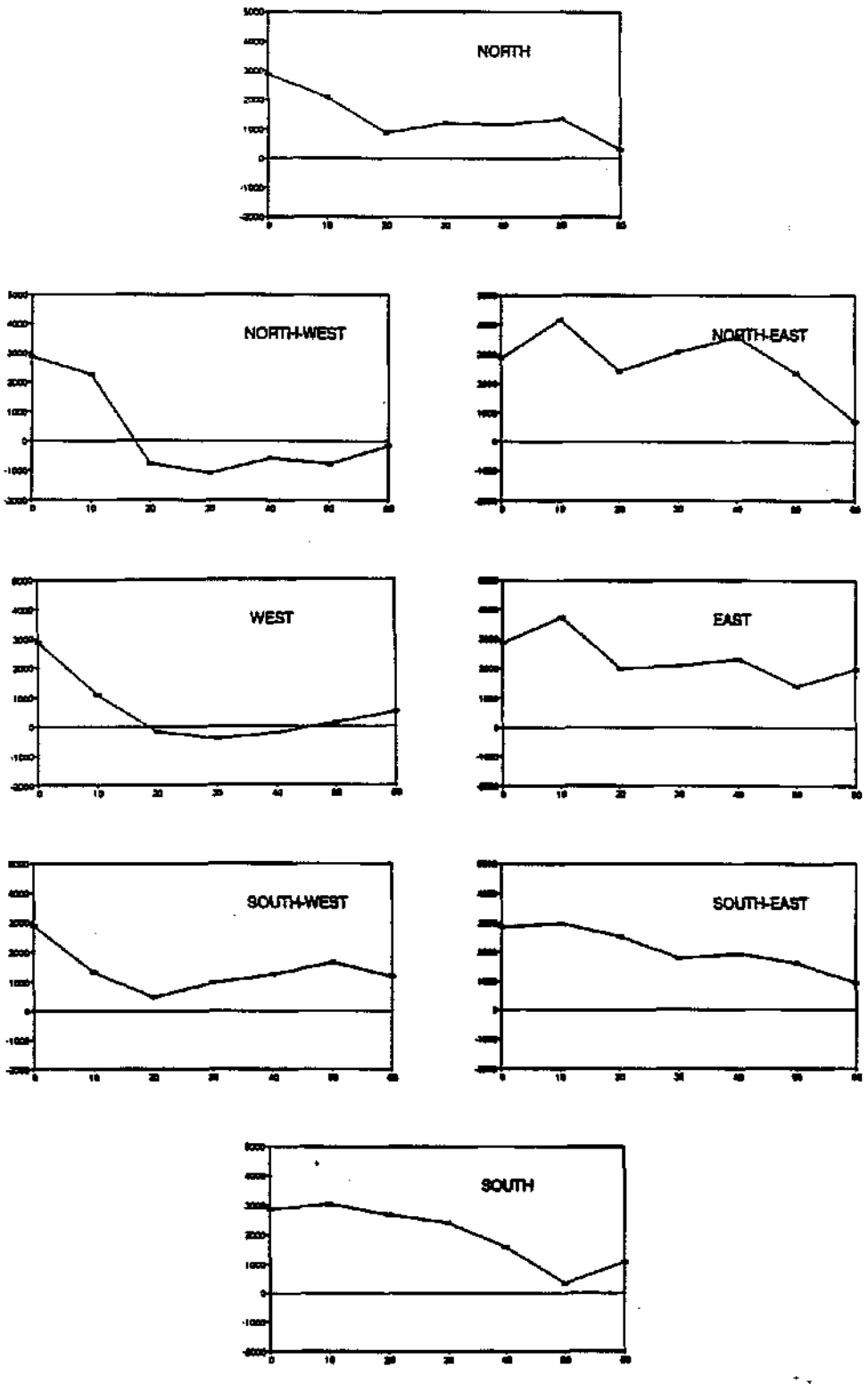


Figure 5.6: Cross-covariance of rain and altitude: SW Cape.

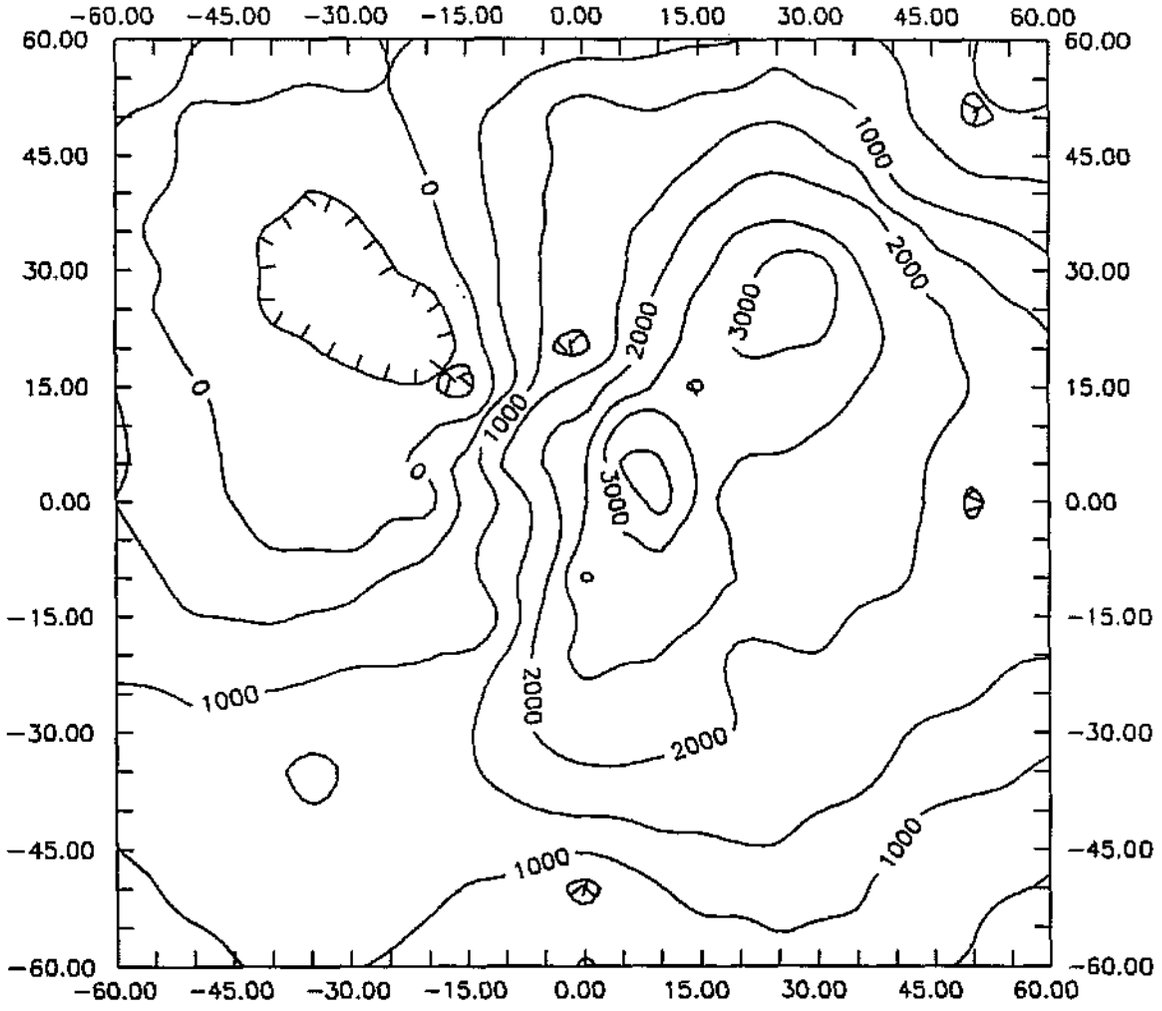


Figure 5.7: Contoured cross-covariance: SW Cape.

given in Section 5.2.3 so that, in equation 5.3, some of the functions f_l may be functions of the covariates. Thus, for example, the functions f_l might be functions of altitude as well as latitude and longitude. This approach requires firstly that the form of the relationship between predictand (rainfall parameter) and covariate (altitude or function of altitude) is known or can be approximated by a simple function such as a polynomial, and also that the covariate information is available at all the sites at which the predictand is known and also at all those points at which an estimate of the predictand is required. This method, which Matheron has described as *kriging with external drift*, was applied by Ahmed and De Marsily (1987) to the estimation of aquifer transmissivity, assuming a linear dependence on specific capacity. It has also been used by Armstrong (1992) to estimate monthly rainfall in Lesotho, using (sometimes estimated) annual rainfall as the covariate or 'drift', and by Hudson (1992) to estimate monthly temperatures in Scotland using elevation as the covariate.

The use of this approach means, in effect, that we would be using a generalized least squares multiple regression of rain on some function of altitude, together with ordinary kriging of the residuals. This would appear to bring us back to the problems previously mentioned for the multiple regression approach, namely the need for restricting the regression calculations to homogenous sub-regions and also the need to pre-define the appropriate functions of topography to be included in the regression model. The use of a moving-window (local) kriging approach as discussed in Section 5.2.3 avoids the first of these problems by re-calculating the estimates at each point using only data points within a limited neighbourhood, thus effectively re-fitting the regression at each point. While computationally intensive, the process is computationally stable, and does not produce the sharp discontinuities in the output that can occur at regional boundaries when regional regression models are used.

To avoid the need to pre-define functions of topography it was decided to first calculate a number of orthogonal functions of elevation at each gridded altitude point, which together would account for all possible patterns up to a third degree surface. This would effectively incorporate a number of the functions defined in Section 5.1; for example, both slope and aspect can be measured in terms of first degree functions, while some of the definitions of roughness and exposure could also be expressed as low-order polynomials of the gridded altitude values. It should be emphasised, however, that what is proposed here is more general than the use of a pre-defined function of altitude such as slope, in that no particular polynomial is chosen *a priori*, but rather, a set of functions is used which effectively encompasses all possible patterns that can be described by third degree functions, while the moving-window kriging process estimates appropriate weightings to give to the component functions in the neighbourhood of each point being estimated.

An advantage of defining orthogonal functions is that they are by definition uncorrelated and thus we avoid the multicollinearity problems commonly associated with multiple regression.

For gridded data the calculation of the orthogonal functions is a simple matter. If we write the altitude values at a grid of points in an $q \times q$ matrix, D, and then calculate the matrix $M'DM$ where M is the $q \times 4$ matrix whose columns give the coefficients for orthogonal polynomials of degree 0,1,2,3 respectively, then the resultant matrix has as its elements the required orthogonal functions.

We illustrate the procedure using a 5×5 grid of altitude points. Let the

altitude values at a grid of points be given by:

$$D = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25} \\ a_{31} & a_{32} & a_{33} & a_{34} & a_{35} \\ a_{41} & a_{42} & a_{43} & a_{44} & a_{45} \\ a_{51} & a_{52} & a_{53} & a_{54} & a_{55} \end{bmatrix}$$

Then the matrix M of orthogonal polynomial coefficients is given by:

$$M = \begin{bmatrix} 1 & -2 & 2 & -1 \\ 1 & -1 & -1 & 2 \\ 1 & 0 & -2 & 0 \\ 1 & 1 & -1 & -2 \\ 1 & 2 & 2 & 1 \end{bmatrix}$$

and $M'DM$ can be written as

$$M'DM = \begin{bmatrix} \xi_{00} & \xi_{01} & \xi_{02} & \xi_{03} \\ \xi_{10} & \xi_{11} & \xi_{12} & \xi_{13} \\ \xi_{20} & \xi_{21} & \xi_{22} & \xi_{23} \\ \xi_{30} & \xi_{31} & \xi_{32} & \xi_{33} \end{bmatrix}$$

where ξ_{00} is simply the sum of the elements of D, while ξ_{01} is the linear contrast of the columns of D, which corresponds to a plane with E-W slope, and ξ_{10} corresponds to a plane sloping N-S. By including both ξ_{10} and ξ_{01} in the external drift function we allow for a plane of any inclination to form the 'drift' function. By including also ξ_{02}, ξ_{11} and ξ_{20} we allow for an arbitrary second degree surface and so on. We decided to include all terms up to third degree, thus allowing for a cubic surface, and using 10 orthogonal functions in all. This allows for reasonably complex topographical patterns.

Tables of orthogonal polynomial coefficients for various values of q, together with formulae for their computation in the general case, are given in

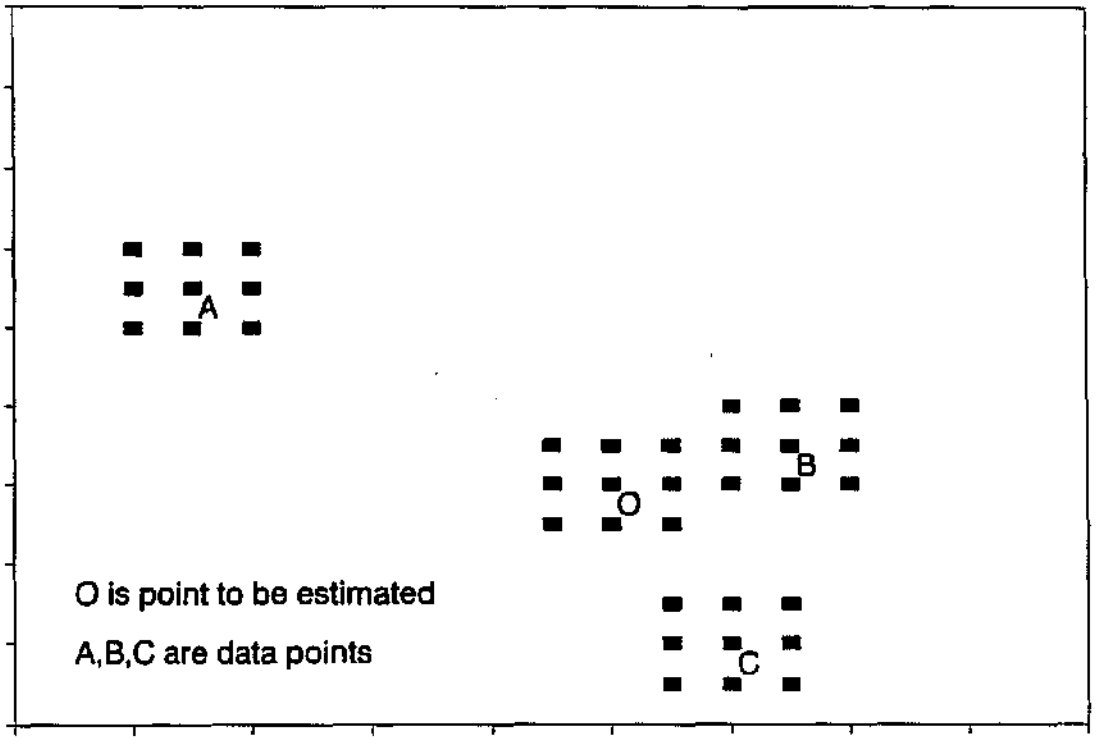


Figure 5.8: Calculating the functions of topography.

a number of books of statistical tables, for example, Pearson and Hartley (1962).

Thus the full external drift function has the form:

$$\beta_{00}\xi_{00}+\beta_{10}\xi_{10}+\beta_{01}\xi_{01}+\beta_{20}\xi_{20}+\beta_{11}\xi_{11}+\beta_{02}\xi_{02}+\beta_{30}\xi_{30}+\beta_{21}\xi_{21}+\beta_{12}\xi_{12}+\beta_{03}\xi_{03}$$

where the β coefficients will be selected optimally by the kriging program to model the relationship between the rainfall model parameter and the components of the pattern of topography in the neighbourhood of the point being estimated.

The values $\xi_{00},\ldots\xi_{03}$ are calculated at each data point and at each point to be estimated. Figure 5.8 illustrates this, using a 3×3 grid. In practice, a

3×3 grid is too small to allow the calculation of independent functions up to third degree; a grid of at least 5×5 is required, but it is not immediately obvious how to choose the optimal size of the grid. In addition, the optimal choice may depend on whether the data has been de-trended. If one is working with de-trended residuals, then they will probably only reflect relatively local topographic effects, while the effects of larger mountain features will have been absorbed by the trend. On the other hand if no prior smoothing has been used then the data will possibly reflect the effects of mountain features some distance away, so that a larger mask should probably be used for the altitude function calculations. A disadvantage of using a larger mask would be that as the number of points in the mask increases, so does the potential complexity of possible topographic patterns, so that it might be necessary to use orthogonal functions of higher degree to obtain a realistic approximation to the surface. Another problem with using a large mask is that the gridded altitude data is not currently available for some of the areas north of the South African border, so that as the grid size increases, there are an increasing number of data sites for which we cannot calculate the necessary functions without some additional estimation.

5.3.4 Cross-Validation

In order to decide on the optimal grid size and also on the optimal degree of the ξ functions to be used in the external drift kriging procedure, a number of test sites were selected as described below, and the values at these sites were estimated from the remaining sites using a range of grid sizes and functions. The sum of the squared estimation errors at the test sites was then used to compare the various options.

Thus, for a given grid size, the corresponding orthogonal altitude functions were first calculated and stored for each point within southern Africa.

Then the kriging estimation process was carried out firstly using no topographical information, then using only the term ξ_{00}, that is, the average altitude, then using only first degree functions, that is, including ξ_{10} and ξ_{01}, then using first and second degree functions, and finally using the full set of third degree functions. For any sites where the external drift matrix $\mathbf{X}$ in equation 5.4 was singular, (this could happen for example if all the altitude values in the mask were identical) a drift of lower order was substituted until a non-singular matrix was obtained.

The whole process was repeated twice; once using the de-trended parameters, (the trend estimation procedure is described in Appendix B), and once with the original parameters. In all cases, a moving window version of kriging was used, such that the closest 33 points (within a maximum search distance of 120 km) were selected. In those parts of the country where the stations are fairly dense the closest 33 points were generally all within a radius of not more than 60 km. If the number of points found within the maximum search distance of 120 km was insufficient for estimation with the chosen degree of orthogonal functions then a drift of lower order was used at that site. For example, if a cubic drift was selected, but less than 30 data points were found within 120 km of a given location to be estimated, then a quadratic drift was used, if less than 20 data points were found, then linear drift was used, while if less than 10 data points were available, then only ξ_{00} was used.

In selecting a set of data points as test sites to cross-validate the various options it must be remembered that the data, that is, the rain parameter values, are themselves estimated values subject to error, so that we do not have 'true' data values with which to compare our estimates. It was therefore decided to select from each Weather Bureau block the rainfall station at which the variance of the estimate of DEPA0 was a minimum, and to use these points as the test points. No test point was selected from blocks having less than five data points, as this could make the resultant data set rather sparse

	degree of orthogonal function							
	with no de-trending				with prior de-trending			
grid size	0	1	2	3	0	1	2	3
5 × 5	363	378	437	568	373	387	443	558
15 × 15	366	373	410	466	377	383	417	466
25 × 25	374	385	391	437	not calculated			
35 × 35	378	400	413	465	not calculated			
no altitude	381	not applicable			386	not applicable		

Table 5.2: Mean squared estimation error: DEPA0.

around that point. Apart from these omitted blocks the 373 test points are thus roughly on a grid across the country, with one in each Weather Bureau block. The decision to use the variance of the parameter DEPA0 as the selection criterion was based on the fact that this parameter is probably the one most sensitive to topography.

The levels of the various factors used in the cross-validation exercise were:

- grid size: (5 × 5, 15 × 15, 25 × 25, 35 × 35 minutes of a degree)
- degree of orthogonal functions: (0,1,2,3)
- prior de-trending / no prior de-trending

For each rain model parameter the mean squared estimation error, averaged over the 373 test points, was calculated for various possible combinations of the factors shown above. The optimal factor combination could vary depending on the specific rain parameter under consideration. In practice, results were very similar for all parameters and thus only results for the parameter DEPA0 are given here (Table 5.2).

It is clear from the results that prior de-trending of the data gives no improvement. The spurious correlations at short lags which were induced by

the smoothing process (see Appendix B) may well be responsible for this. Also, the fact that local kriging was used throughout means that the effects of trend are likely to be small, and thus the de-trending only introduces an extra level of complexity into the kriging process, apparently without any compensating gain in accuracy.

A rather surprising feature of the results is that fitting the more complex topographical models produces poorer results, although the deterioration is less marked for the larger grid sizes. The estimation using the average altitude ξ_{00} does however give a small improvement over estimation ignoring altitude.

These results are made clearer if we graph the absolute values of the errors at individual test sites, as in Figure 5.9. We see from these graphs, which are based on a grid size of 9×9, that, although the models using average, linear or quadratic altitude generally give rise to smaller absolute error than the model ignoring altitude, the quadratic model gives rise to one very poor estimate, and the cubic model gives two. Bearing in mind that we may be extrapolating the altitude functions; that is, the values of the ξ at the point being estimated could be outside the range of the ξ values at the neighbouring data locations, it is not so surprising that we occasionally get rather poor estimates. Thus the more complex models are less robust. The fact that the models which do not include altitude at all do almost as well as the models with altitude is probably due to the fact that the test points, which were chosen for their low variance, are typically stations with many years of data and not necessarily at high altitude, so that a simple interpolation from the neighbouring data points gives fairly accurate results.

In order to show more clearly the effect of including altitude in the models, parameter estimates were calculated at one minute intervals along two transects; Figure 5.10 shows the estimates of DEPA0 along the two transects, together with the altitude values. In the first, running west to east in

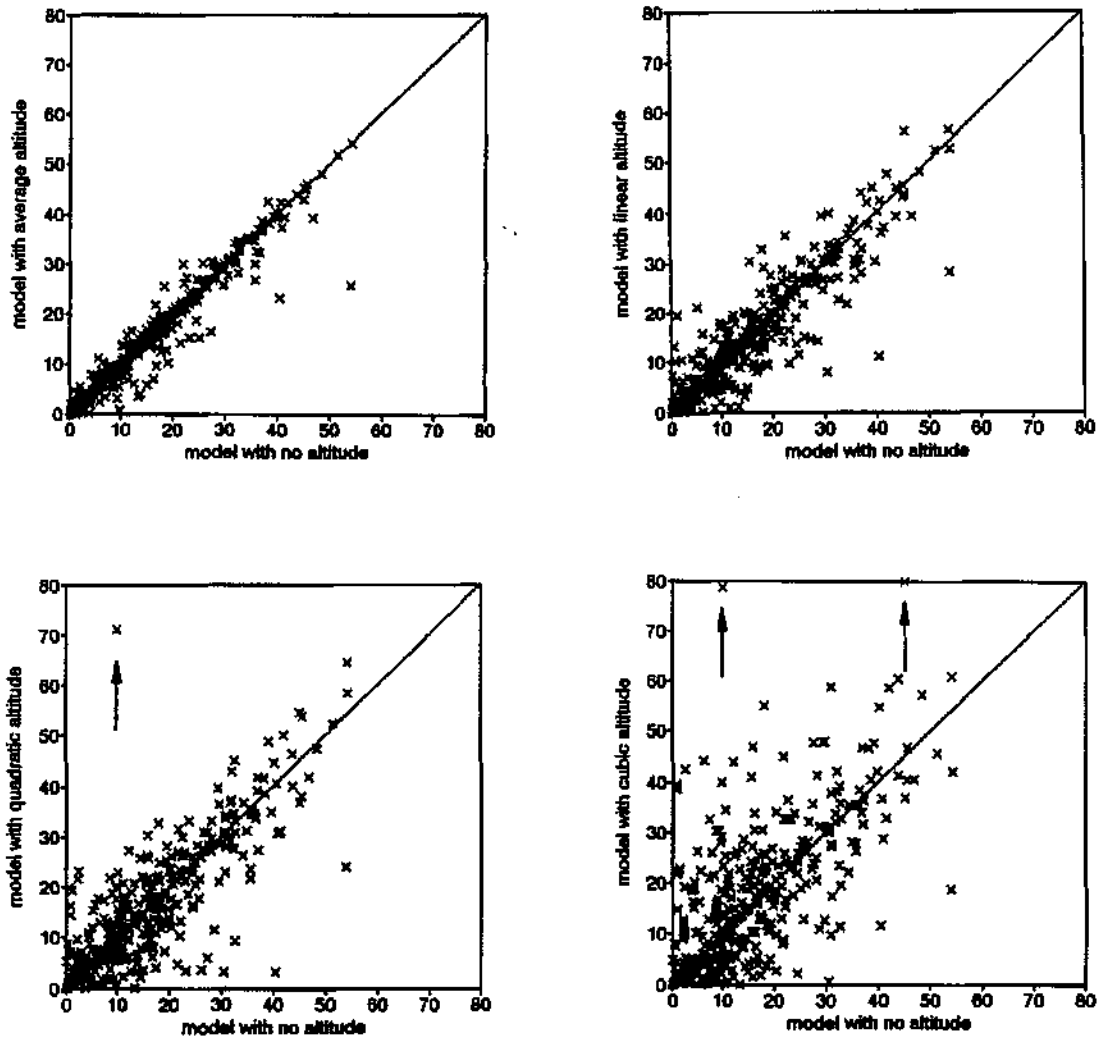


Figure 5.9: Estimation errors at individual test sites.

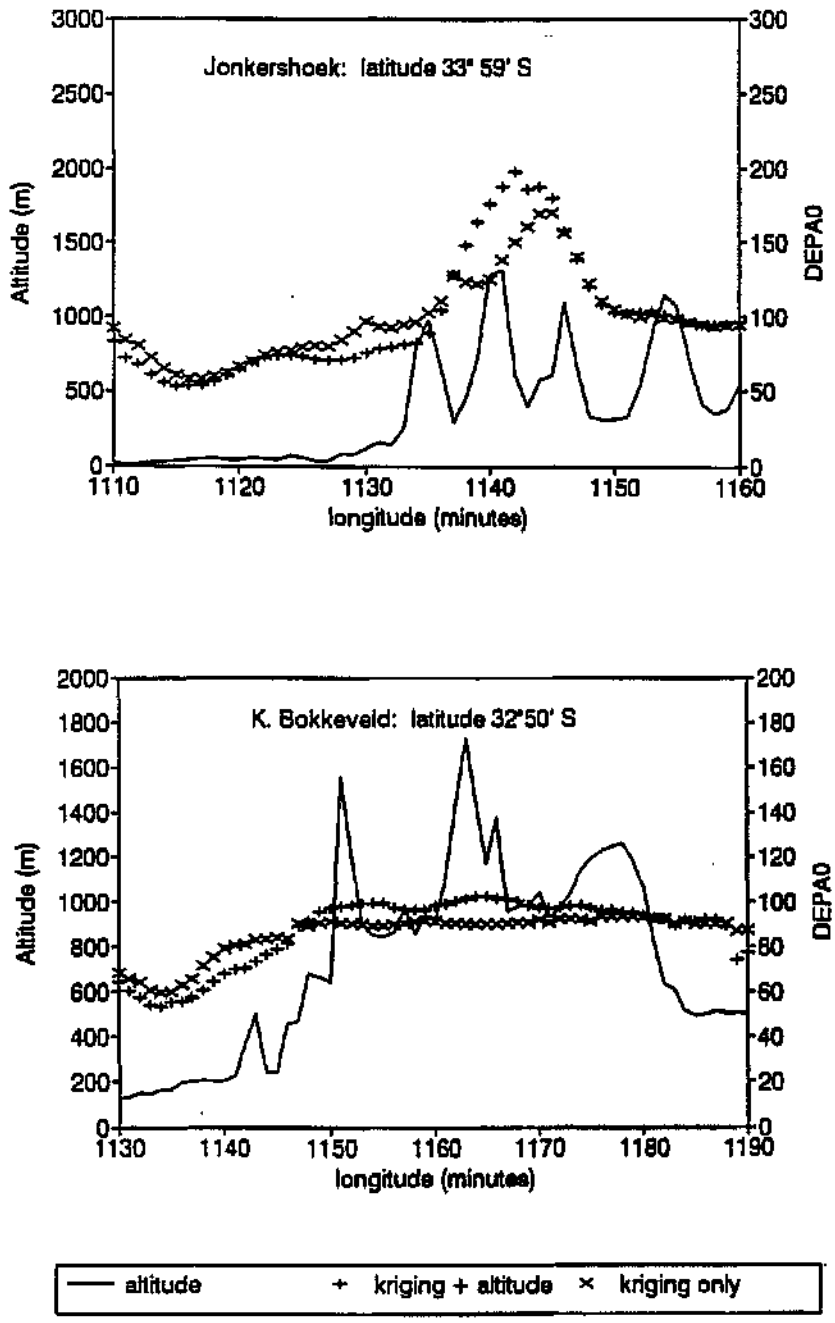


Figure 5.10: Predicted DEPA0 and altitude.

the Jonkershoek mountains near Stellenbosch, there are a number of rainfall stations within the mountains, so that the ordinary kriging model without any altitude functions follows the shape of the mountains quite well. In the second transect, running west to east across the mountain ranges just north of Porterville, the model without altitude does not pick up the individual mountain peaks at all as there are few rainfall stations in the area, while the model including ξ_{00} shows a small rise in DEPA0 as each peak is crossed. By contrast, the values of DWA0, which is a measure of the *probability* of rain, taken along the same transect (Figure 5.11), show almost no response to altitude; the values decrease steadily as one moves west, away from the coast.

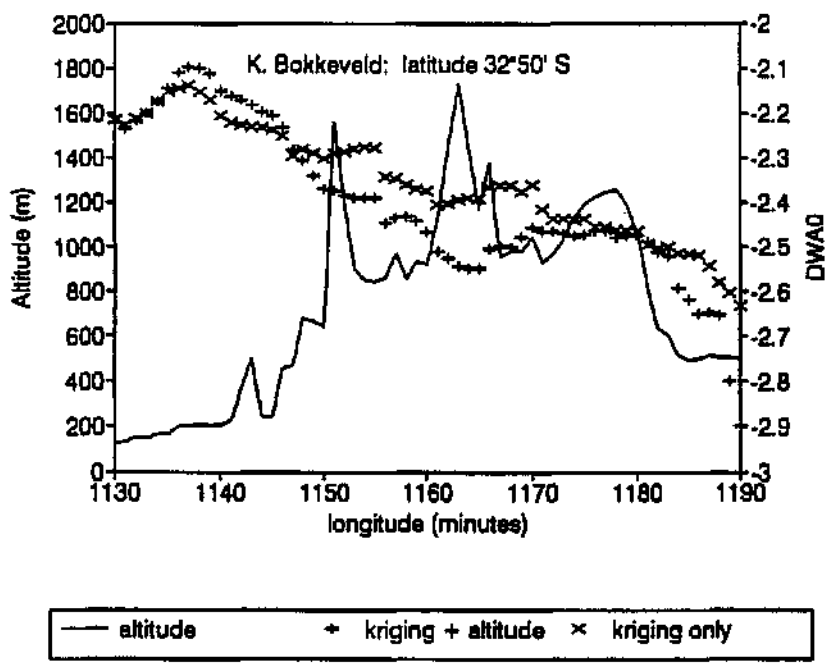


Figure 5.11: Predicted DWA0 and altitude.

On the basis of these results it was decided that the final estimates should

be done without prior de-trending and including only the term ξ_{00} in the drift function. Further checking of grid sizes suggested that a grid of 9×9 would be optimal, and this was used for the final estimation at a grid of sites covering the country at 1 minute by 1 minute intervals. Maps of the estimates at intervals of 30 minutes by 30 minutes, that is, at the centre of each Weather Bureau block, are shown in Figure 5.12.

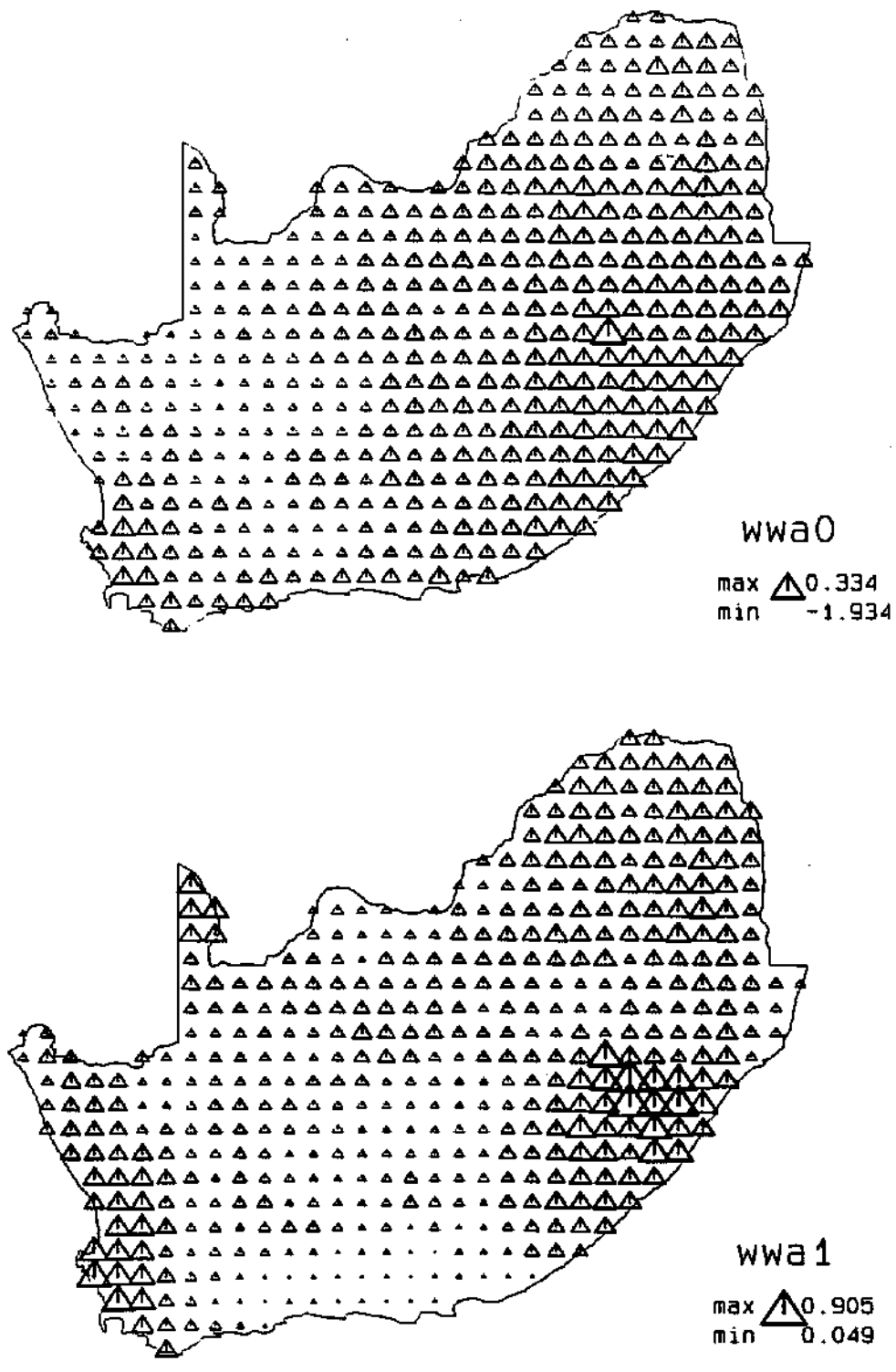


Figure 5.12: Estimated parameter values at centres of Weather Bureau blocks (amplitude parameters).

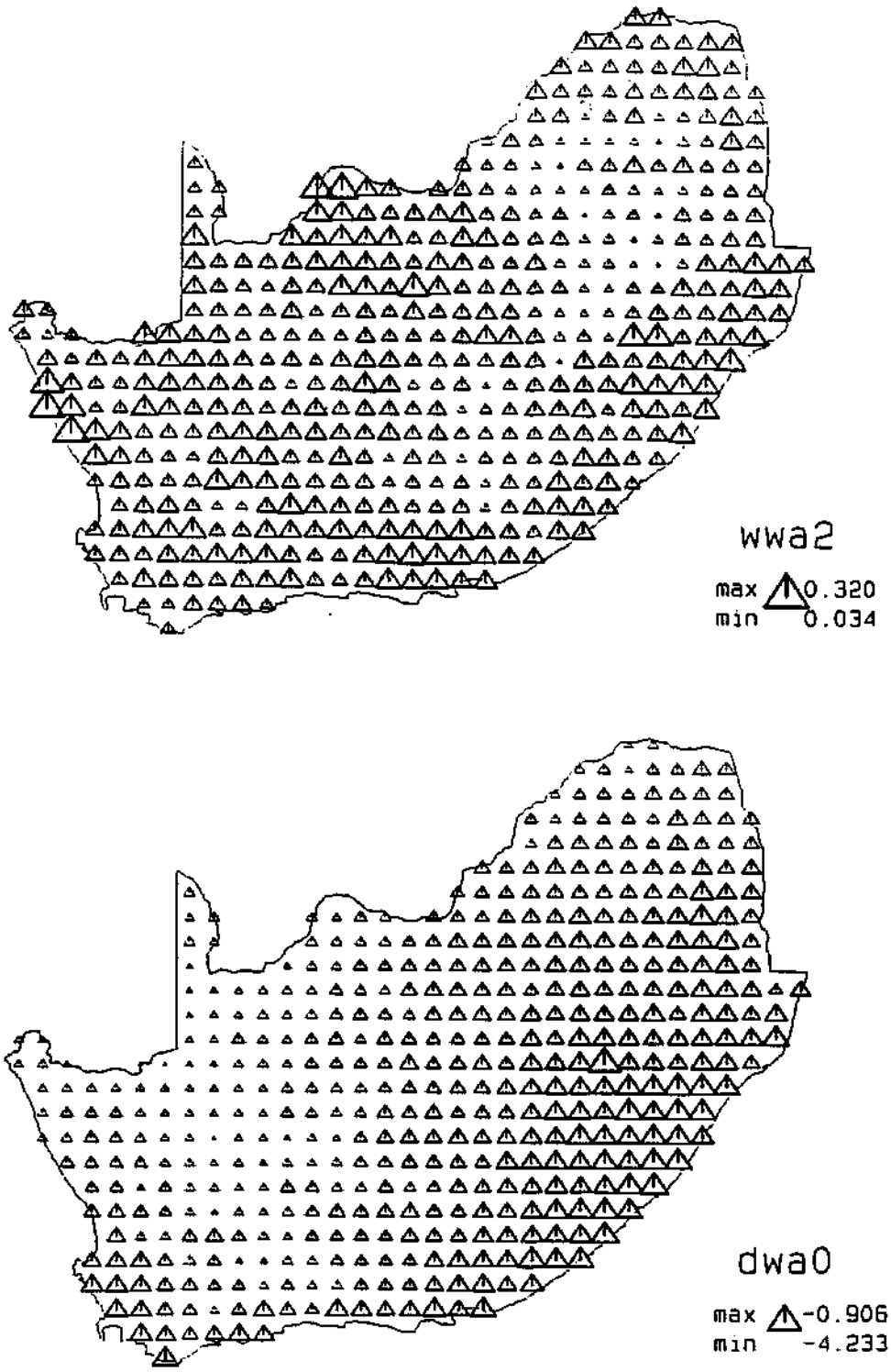


Figure 5.12: Estimated parameter values at centres of Weather Bureau blocks (amplitude parameters) (contd.).

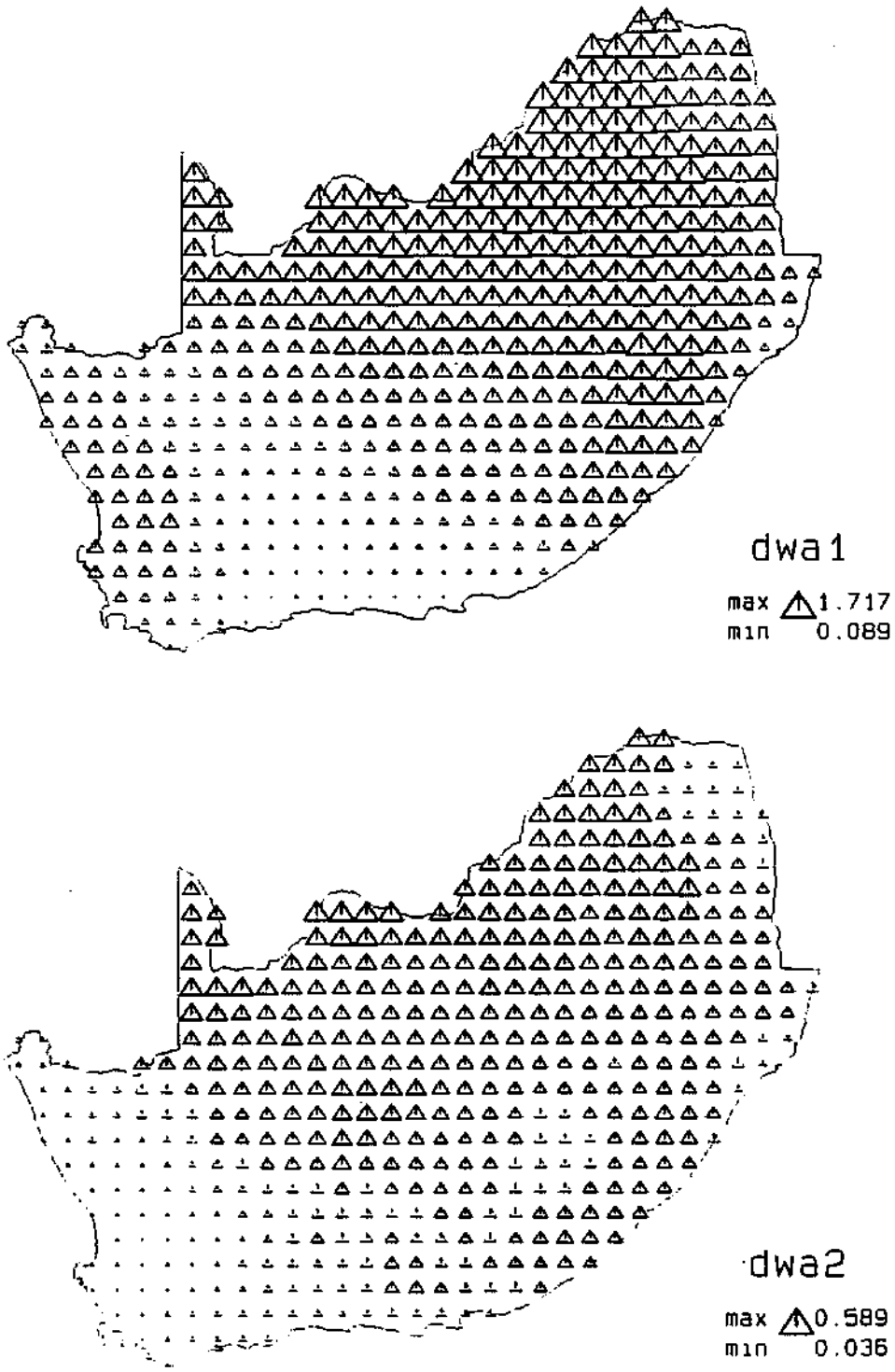


Figure 5.12: Estimated parameter values at centres of Weather Bureau blocks (amplitude parameters) (contd.).

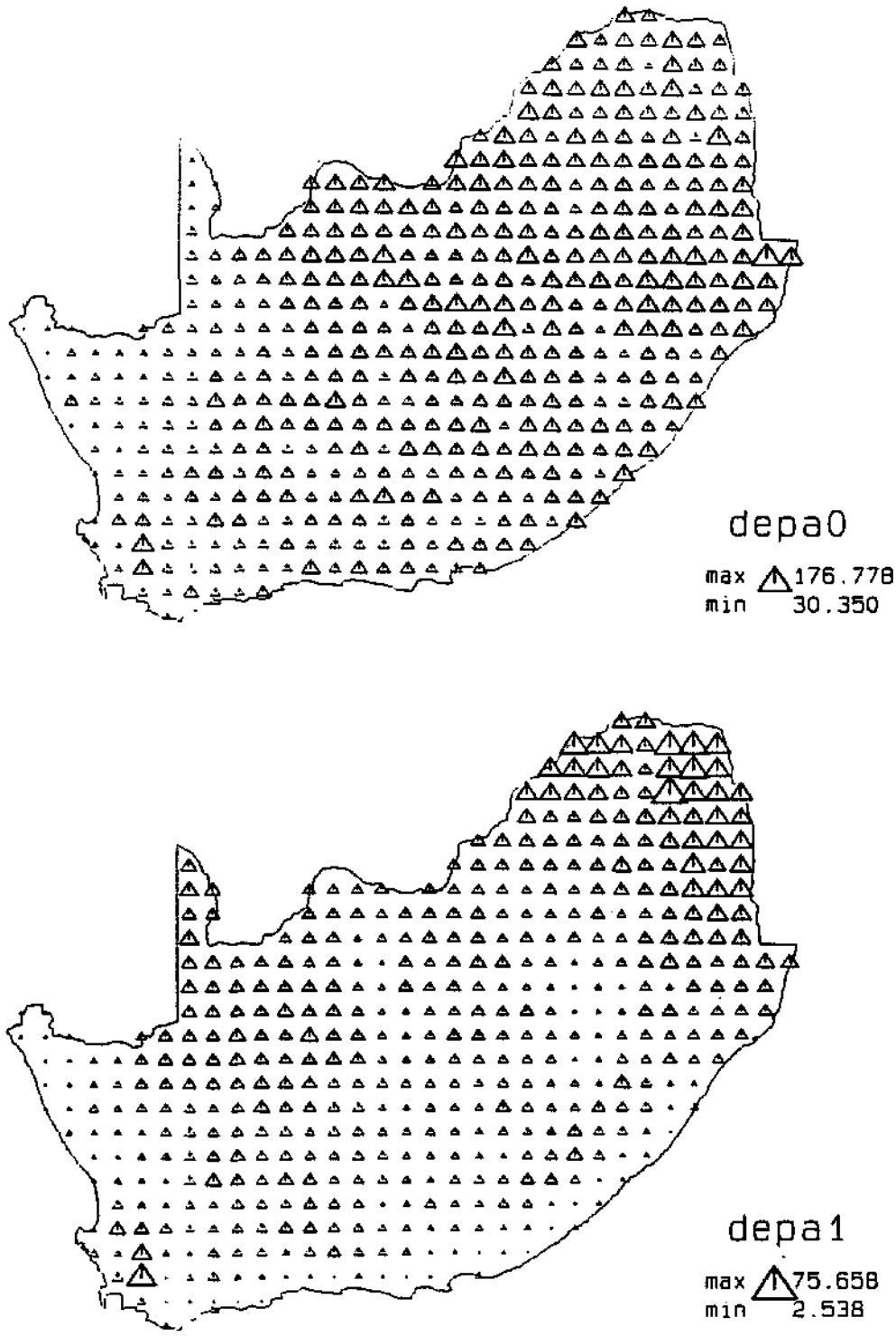


Figure 5.12: Estimated parameter values at centres of Weather Bureau blocks (amplitude parameters) (contd.).

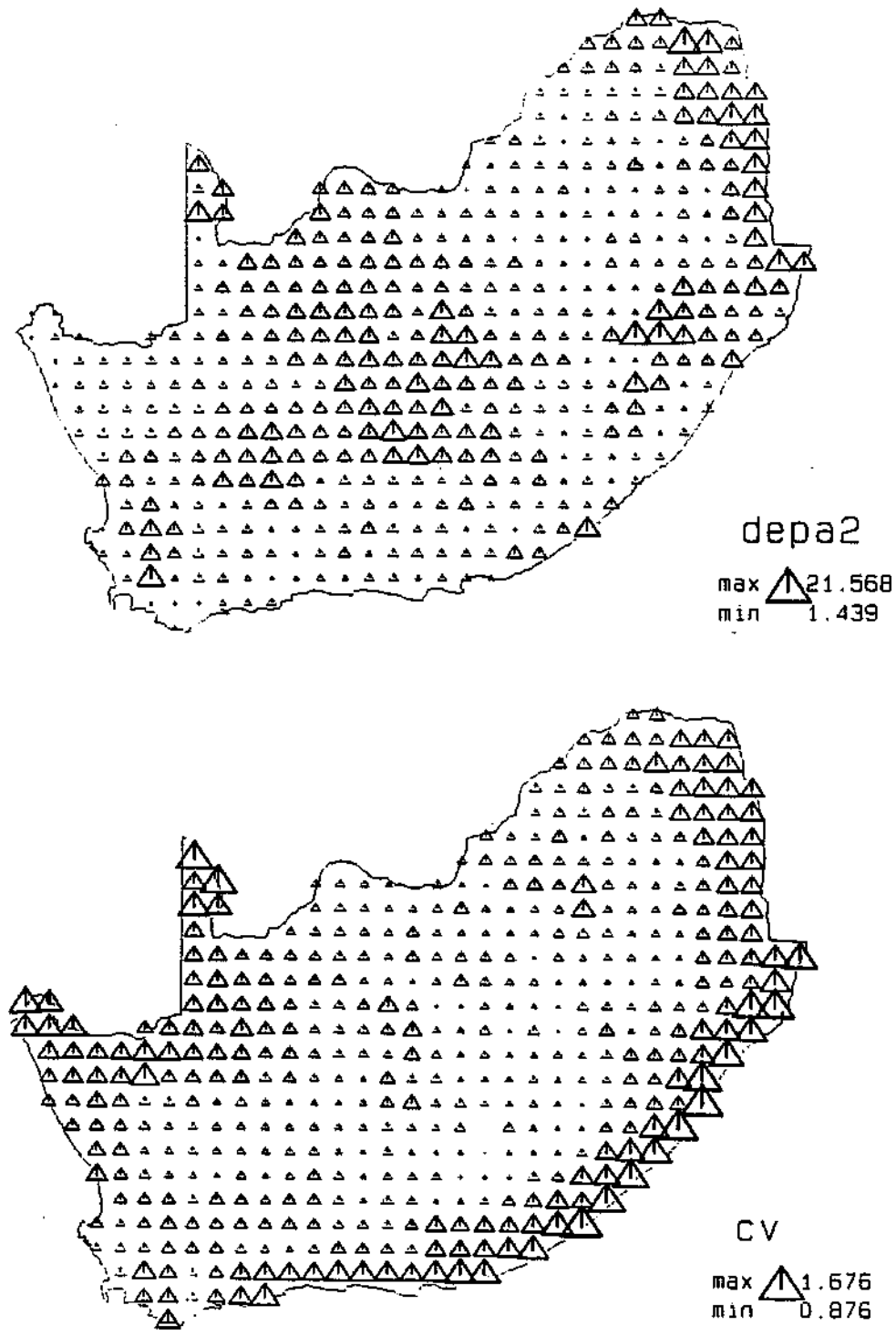


Figure 5.12: Estimated parameter values at centres of Weather Bureau blocks (amplitude parameters) (contd.).

5.4 Estimation of the Phase Parameters

The phase parameters of the daily rainfall model are *circular* in nature. In particular the first phase parameters take values between 0 and 365 while the second phase parameters take values between 0 and 365/2. Given two sites, one with $\phi_1 = 364$ (December 30) and the other with $\phi_1 = 3$ (January 3), say, an obvious estimate of the value of this phase parameter at a site situated midway between the two sites would be given, *not* by the arithmetic average $(364 + 3)/2 = 183.5$, but by the value $\phi_1 = 1$ (January 1). From this simple example it is clear that normal methods of calculation are inappropriate for circular data. Such data arise in a number of fields. The most common examples arise either from directional data in two-dimensional space, such as in studies of wind direction or direction of magnetization of rock specimens, or else from periodic phenomena, such as the time of day or the time of year of the occurrence of certain events. The phase parameters of the rainfall model are of the latter type. Many other examples are given in the texts by Mardia (1972), Batschelet (1981) and Upton and Fingleton (1989).

Statistical techniques for circular data tend to be more computationally complex than their counterparts for data taking values on the real line. Notions of correlation, regression and bias are still the subject of discussion (Mardia (1975), Jupp and Mardia (1989)). In particular, the development of smoothing techniques for spatially distributed circular data is very much in its infancy.

5.4.1 Smoothing Methods for Circular Data

Watson (1985) is perhaps the first to discuss the problems of interpolating and smoothing circular data available at a number of spatial locations, and he outlines a couple of possible approaches. For the interpolation problem he suggests the use of a weighted average of the data points, with the

weights chosen proportional to the inverse of the square of the distance from the point to be estimated. He goes on to outline a possible method for smoothing spatial directional data, represented by angles $\theta_1, \ldots \theta_n$, based on calculating estimated values $f(\mathbf{z}_1)$ to $f(\mathbf{z}_n)$ at the given spatial locations $\mathbf{z}_i$, $(i = 1, 2, \ldots n)$ to maximize

$$\sum_{i=1}^{n} k_i \cos(\theta_i - f(\mathbf{z}_i)) + \tau \sum_{1 \le i < j \le n} \cos(f(\mathbf{z}_i) - f(\mathbf{z}_j))/d_{ij}$$

where k_i is some inverse function of the measurement error variance of the i'th data point, d_{ij} is some monotonic increasing function of distance between the i'th and j'th data points and τ is a smoothing parameter. Watson does not give full details of the method; in particular, the discussion does not explain how the method generates estimates at points other than the original data locations $\mathbf{z}_i$. Also, if the original data locations are clustered in space, then excessive weight will be given to 'high density' areas. This criticism applies also to the simple weighted average method.

Mendoza (1986) has subsequently implemented a method of smoothing circular data available at locations on a plane and illustrates its use to smooth data on the cross-bedding directions of sandstone. His method, which is rather similar to Watson's except that it uses a spline-based measure of smoothness, finds $f(\mathbf{z})$ to minimize

$$\sum_{i=1}^{n} w_i(1 - \cos(\theta_i - f(\mathbf{z}_i)) + \lambda R_2(f)$$

where θ_i is the observed angle and $f(\mathbf{z}_i)$ the smoothed angular value at the i'th location, w_i is a weighting factor for the i'th data point and $R_2(f)$ is a measure of roughness of the function f. The roughness criterion is similar to that used in spline-smoothing and is given by

$$R_2(f) = \int\int (\frac{\partial^2}{\partial x^2} f)^2 + 2(\frac{\partial^2}{\partial x \partial y} f)^2 + (\frac{\partial^2}{\partial y^2} f)^2 \, dx \, dy$$

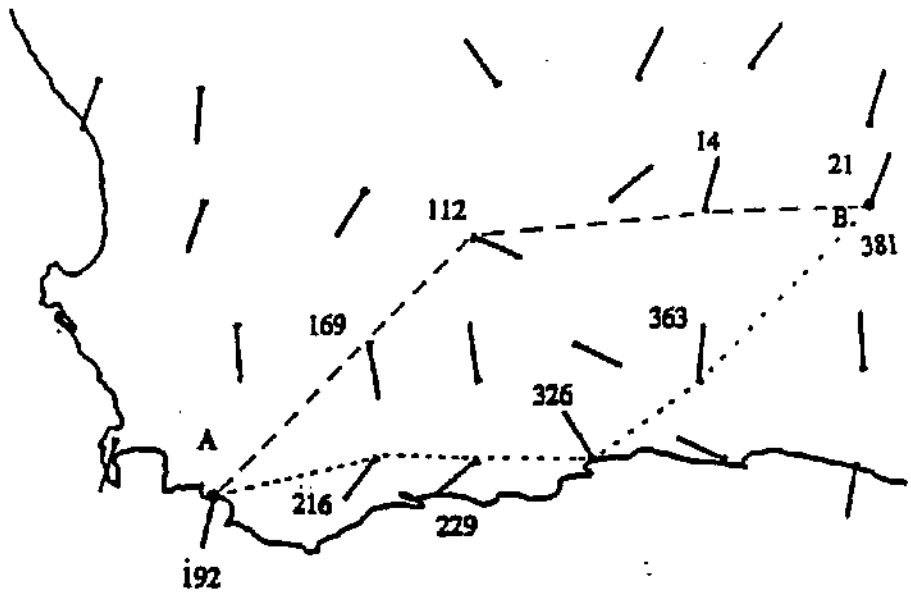


Figure 5.13: Re-labelling of circular values.

Whereas for most calculations involving directional data only trigonometric functions of the data values are used, so that θ_i and $\theta_i + 360$ (in degrees) are equivalent, in the calculation of $R_2(f)$ such values are *not* equivalent, so that it is first necessary to choose the value θ_i to represent each data point. Mendoza suggests that these values should be chosen '*so that observations are not rougher than they should be*'. Thus, for example, a sequence of adjacent values (in degrees) of 240 300 350 30 is re-expressed as 240 300 350 390. This may not be easy to achieve consistently for spatial data. For example, Figure 5.13 shows some data values in the south-western Cape, taken from the values of WWA0 (converted to degrees), in which smoothing the data along the path indicated by the dotted line from the point A (192 degrees) to the point B leads to a labelling of 381 degrees for point B whereas smoothing the data along the path indicated by the dashed line leads to a value of 21 degrees for the same point; the lack of a natural ordering of points in

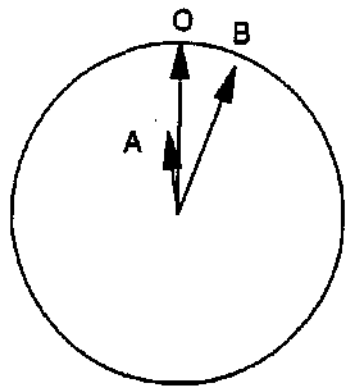


Figure 5.14: Vector distance and angular distance.

two dimensional space means that appropriate values may not be uniquely defined.

Young (1987) extends the technique of kriging to *vector* data in a natural way, using the Euclidean norm as a measure of distance, and suggests that the resulting technique will be appropriate for directional or circular data. This is not the case however, since there is no guarantee that the vector estimate resulting from the vector kriging process will in fact be of unit length, and thus it may minimize the vector distance but not the angular distance. Thus, for example, in Figure 5.14 the vector B is closer (as measured by the Euclidean norm distance) to vector O than is vector A. However, in terms of angular distance, A is closer to O than B is. To get an estimate which minimizes the angular distance it is necessary either to constrain the solution to lie on the unit circle, or else to try to minimize angular distance directly.

In the next section we explore the feasibility of extending kriging in this way.

5.4.2 Kriging for Circular Data

We start by reviewing some basic notation and summary statistics for circular data.

Means and Variances for Circular Data

Circular data can be represented by points on a circle of unit radius. Where the data are not directions, as in the present application, the range of values can easily be mapped on to the circle; for example, in the case of the first phase parameters of the rainfall model , which take values between 0 and 365, we can multiply the values by $2\pi/365$ to get an equivalent value in radians. The mean of the data is then defined to be the direction of the resultant vector. That is, if we represent the data points by the unit vectors $\mathbf{e}_1$,$\mathbf{e}_2$, $\ldots \mathbf{e}_n$, with polar coordinates $(1,\theta_1),(1;\theta_2),\ldots(1,\theta_n)$ then the mean vector of the sample is given by

$$\mathbf{m} = \sum_{i=1}^{n} \mathbf{e}_i/n$$

If we assign a unit mass to each data point in Figure 5.15, then $\mathbf{m}$ represents the centre of gravity of the data. The cartesian coordinates of $\mathbf{m}$ are $\overline{x} = \sum \cos\theta_i/n$ and $\overline{y} = \sum \sin\theta_i/n$ and the polar coordinates are $(\overline{R},\overline{\theta})$ where

$$\overline{R}^2 = \overline{x}^2 + \overline{y}^2$$

and

$$\tan\overline{\theta} = \sum \sin\theta_i / \sum \cos\theta_i$$

This has a singularity if $\sum \sin\theta_i = \sum \cos\theta_i = 0$, so that the centre of gravity is at the origin, and thus the resultant direction is not uniquely defined.

It can be shown that

$$\overline{R} = \sum_{i=1}^{n} \cos(\theta_i - \overline{\theta})/n$$

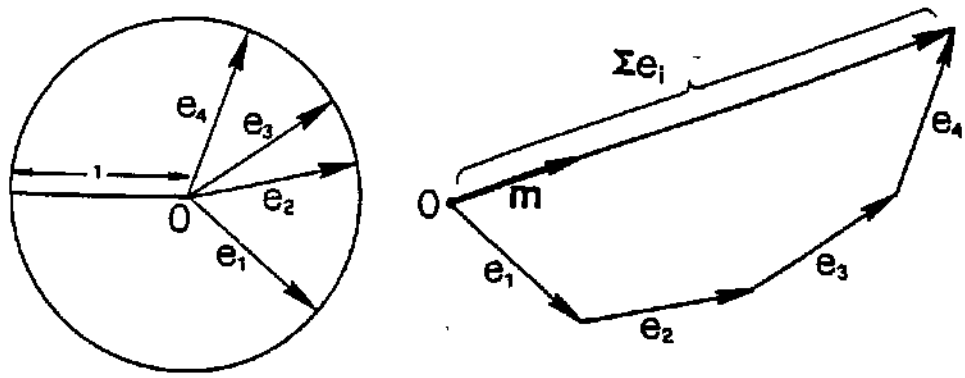


Figure 5.15: Mean of circular data.

and thus the measure $[1 - \overline{R}]$ provides a measure of the variance, with properties which are in many ways analogous to those of the variance for non-circular data (Mardia, 1972)

To obtain a weighted mean with weights $w_1, \ldots w_n$ it is natural to define this as the direction corresponding to the point with coordinates $\overline{x} = \sum w_i \cos\theta_i/n$ and $\overline{y} = \sum w_i \sin\theta_i/n$, which is equivalent to assigning the weights as masses to the data points and finding the centre of gravity as before. Note that multiplying all the weights by a non-zero constant, l, does not affect the direction of the weighted mean, but changes the length of the mean vector $\overline{R}$ by a factor l.

The Kriging Equations

We consider a model equivalent to that used in *ordinary* kriging, that is, we assume the data are a realization of a stochastic process with common mean and variance, and that the covariance, to be defined below, is a function of distance only.

Given unit vectors $\mathbf{e}_1, \mathbf{e}_2, \ldots \mathbf{e}_n$, at spatial locations $\mathbf{z}_1, \mathbf{z}_2, \ldots \mathbf{z}_n$, we seek an estimator, based on the weighted sum $\sum w_i \mathbf{e}_i$, of the value at some location $\mathbf{z}_0$. Specifically, if $\sum w_i \mathbf{e}_i$ is written in polar form as $(R_0, \hat{\theta}_0)$ then we seek $\mathbf{w}$ to minimize

$$E[1 - \cos(\hat{\theta}_0 - \theta_0)] \tag{5.8}$$

where θ_0 is the unknown angular value at location $\mathbf{z}_0$. The use of the function $1 - \cos(\hat{\theta} - \theta)$ as a measure of estimation error is common in circular statistics, and is analogous to the usual least squares criterion (Fisher and Lewis, 1985).

We show in Appendix B that an approximate solution is given by:

$$\mathbf{w} = \frac{K^{-1}\mathbf{s}}{\sqrt{\mathbf{s}' K^{-1} \mathbf{s}}} = \frac{K^{-1}\mathbf{s}}{r} \tag{5.9}$$

where $k_{ij} = E[\cos(\theta_i - \theta_j)]$, $s_i = E[\cos(\theta_i - \theta_0)]$ and $r = \sqrt{\mathbf{s}' K^{-1} \mathbf{s}}$ is a scalar normalizing constant.

For the case where the data include measurement error, as is the case with the rainfall model parameters, so that we observe $\vartheta_i = \theta_i + \epsilon_i$ instead of θ_i, a similar argument shows that the approximate solution to obtaining an optimal estimate of θ_0 is given by the same expression but with $k_{ij} = E[\cos(\vartheta_i - \vartheta_j)]$ and $s_i = E[\cos(\vartheta_i - \theta_0)]$.

The form of the solution given by equation 5.9 is, apart from the normalizing constant r, exactly analogous to the solution of the usual (non-circular) problem of *simple* kriging. The form of this solution is intuitively appealing, in that it gives more weight to those data values which are close (in space) to the point to be estimated (via $\mathbf{s}$) and gives less weight to points which are clustered with other data points (via K^{-1}). Thus, although equation 5.9 gives only an approximate solution to the minimization of equation 5.8, it can be justified in its own right as a form of weighted average which caters specifically for clustered data, and can also cater for varying error variances. It is thus of interest to see how well such a method performs in practice.

The performance of the method was therefore compared with the simple weighted average method described in Section 5.4.1, using a number of test sites. Before carrying out the estimation it is first necessary to model the spatial covariance so as to have values for $E[\cos(\vartheta_i - \vartheta_j)]$ and $E[\cos(\vartheta_i - \theta_0)]$.

Modelling the Spatial Covariance

A number of measures of association have been proposed for circular data, and reviews can be found in Jupp and Mardia (1989) and Breckling (1989). For our purposes here it suffices to find a measure of the relationship between two circular variables with the same distribution, and, in particular, with the same mean, since, in using local kriging, trends can be ignored. In view of the form of the kriging equations an obvious choice is

$$\sigma_{ij} = E[\cos(\theta_i - \theta_j)]$$

This is in fact equivalent to the measure proposed by Breckling (1989) in the case where the means of θ_i and θ_j are the same.

If we assume that the covariance is a function of distance only then we can estimate the covariance function from the data by plotting $\cos(\theta_i - \theta_j)$ as a function of d_{ij}. Alternatively we may prefer to define a circular semi-variogram as

$$\gamma(h) = E_{(d_{ij}=h)}[\frac{1}{2}(1 - \cos(\theta_i - \theta_j))]$$

as a circular analogue of the usual semi-variogram, where the expectation is over all locations i and j with separation distance h. The circular semi-variogram as thus defined takes values between 0 and 1.

In order to study the empirical circular semi-variogram as a function of separation distance we can plot the average of the values $(1 - \cos(\theta_i - \theta_j))/2$ for all pairs of points with a given separation distance as a function of the separation distance. When the data is measured with error, this empirical semi-variogram will be increased by an amount depending on the error

variance. Specifically, suppose that we observe angular values ϑ_i such that

$$\vartheta_i = \theta_i + \epsilon_i$$

where the angles ϵ_i represent measurement error, so that we may assume that the values of ϵ_i at different sites are independent of one another and also of the values θ_i. We also assume that the distribution of ϵ_i is symmetric with mean zero so that $E[\sin \epsilon_i] = 0$. Then we show in Appendix B that

$$E[\cos(\vartheta_i - \vartheta_j)] = E[\cos(\theta_i - \theta_j)]E[\cos \epsilon_i]E[\cos \epsilon_j] \tag{5.10}$$

Similarly

$$E[\cos(\vartheta_i - \theta_0)] = E[\cos(\theta_i - \theta_0)]E[\cos \epsilon_i]$$

Estimates of the terms $E[\cos \epsilon_i]$ are available for each parameter at each rainfall station from the bootstrap variance calculations described in Chapter 4, and thus it is possible to estimate the covariance of the underlying θ values using the estimator

$$\sigma_\theta(h) = 1/N_h \left\{ \sum \left(\frac{\cos(\vartheta_i - \vartheta_j)}{E[\cos(\epsilon_i)]E[\cos(\epsilon_j)]} \right) \right\}$$

where the summation is over all N_h pairs of points a distance h apart. A similar adjustment may be made to the semi-variogram calculated from the observed ϑ values to get an estimate of the semi-variogram of the θ values.

Figure 5.16 shows the unadjusted and adjusted semi-variogram for each of the phase parameters of the daily rainfall model. The effect of the adjustment is less marked for the first phase parameters than for the second phase parameters, indicating that the former are subject to relatively less estimation error. After adjustment, the graphs show little evidence of any residual nugget effect, which means that the phase parameters do not change significantly over short distances. This confirms our earlier suggestion that while the amplitude parameters would be sensitive to local topography, the phase parameters would not. The few negative values in some of the adjusted

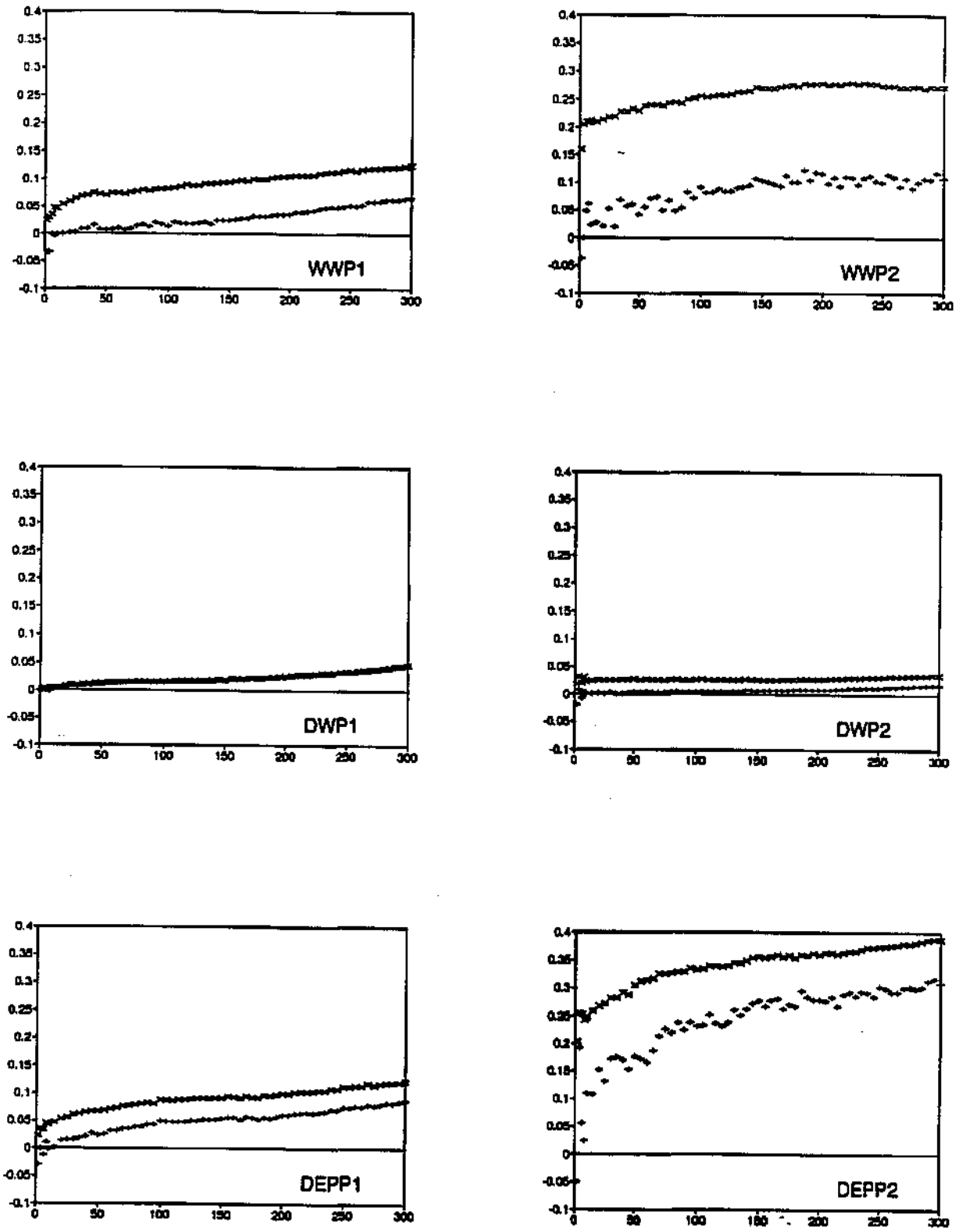


Figure 5.16: Semi-variograms: phase parameters.

graphs result from the measurement error adjustment; clearly, in fitting a model to such data one would require all the fitted values to be positive. Using the weighted least squares method discussed in Section 5.3.1, an exponential model, given by the equation

$$\gamma(h) = s(1 - \exp(-3h/r)) \tag{5.11}$$

was fitted to each of the adjusted semi-variograms. In this equation s is the sill and r is the *effective range*, that is, the range at which the value of γ reaches 95% of the sill. Table 5.3 gives the sill and effective range for each rain model parameter.

parameter	sill	r
WWP1	0.0112	10
WWP2	0.1110	22
DWP1	0.0157	11
DWP2	0.0078	28
DEPP1	0.0659	29
DEPP2	0.2790	13

Table 5.3: Fitted semi-variogram models: phase parameters.

5.4.3 Validation and Discussion

To test the circular kriging method proposed in the previous section, the method was compared with the simpler method of inverse distance weighting using a test set of 101 rainfall stations and a data set of 325 rainfall stations selected from the full data set. The test sites selected lie approximately on a regular grid, with one station having been selected at random from every alternate Weather Bureau block, while the test sites were selected at random from the remaining stations in such a way as to have a similar spatial distribution to the full data set (Figure 5.17).

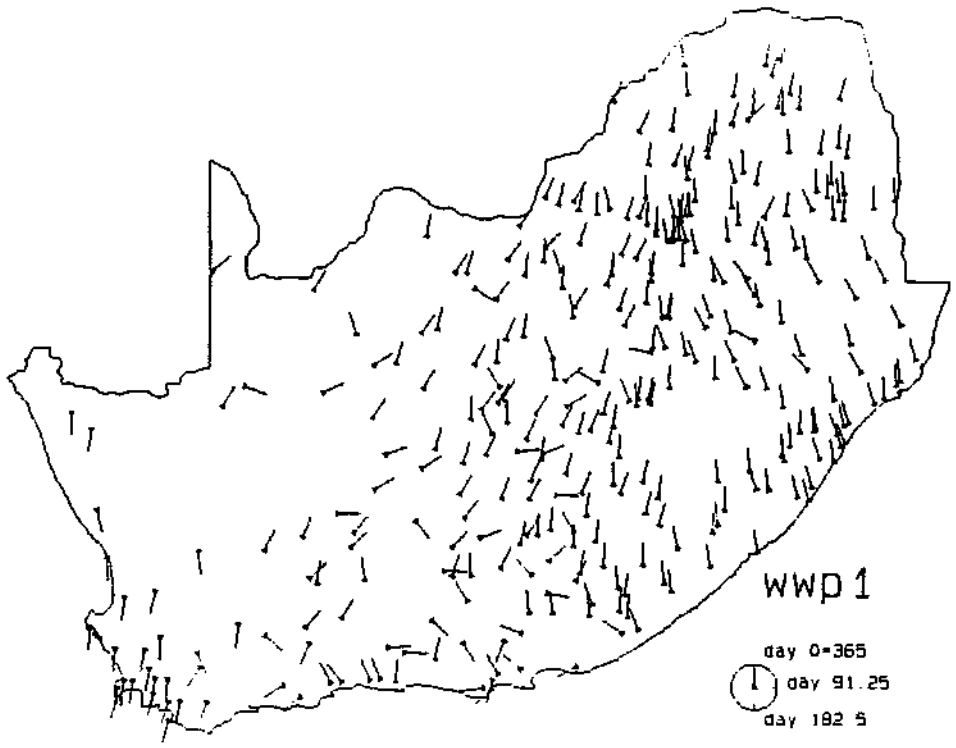


Figure 5.17: Map of data sites used in circular kriging validation.

The data sites are rather sparse; while this can be expected to result in rather poor estimates, it should also help to highlight the difference between the two methods, since, if too dense a data set is used, almost all smoothing methods will give good results. In both methods a search radius of 300 kilometres was used, that is, only points within 300 kilometres of the point to be estimated were included in the calculation.

The average of the error terms, $[1 - \cos(\vartheta_i - \hat{\theta}_i)]$ (averaged over the 101 test data points), was compared for the two methods, and the results for the parameter WWP1 are shown in Table 5.4 below, from which it is clear that the kriging method has resulted in considerably lower errors on average.

There are several reasons why the kriging method may give better results

method	av. error
inverse distance	0.1318
kriging	0.0988

Table 5.4: Comparison of prediction errors: WWP1.

than the inverse distance method. One is, of course, that in using the inverse distance method we have made no attempt to optimize the particular inverse distance function used; it would be possible to use a parametric function of distance, with the parameter value controlling the effective bandwidth selected, for example, by cross-validation, but this to some extent reduces the main advantage of the inverse distance method, namely its simplicity. Another possible reason for the superiority of the kriging method is that the inverse distance method does not take account of clustering in the data; however, a study of Figure 5.17 suggests that this is probably not of great consequence for this particular data set, as the clustered data points generally have similar values to the more isolated points around them. A third reason for the superiority of the kriging method is its explicit use of the error variance of the data; the inverse distance method will give relatively high weight to the few points which are closest to the point to be estimated even if those data points have high measurement error, whereas the kriging method will adjust for this; this is quite important in the present application where the error variance, as measured via the bootstrap procedure, was relatively high at some sites.

In comparing individual estimated values with the original values in the test data set one must bear in mind that even the values in the test data set are not entirely accurate but are subject to the parameter estimation errors as estimated by the bootstrap procedure. Therefore, in plotting a map of the values estimated by the kriging method (Figure 5.18) we have included for comparison, not the original data values, but a range of values

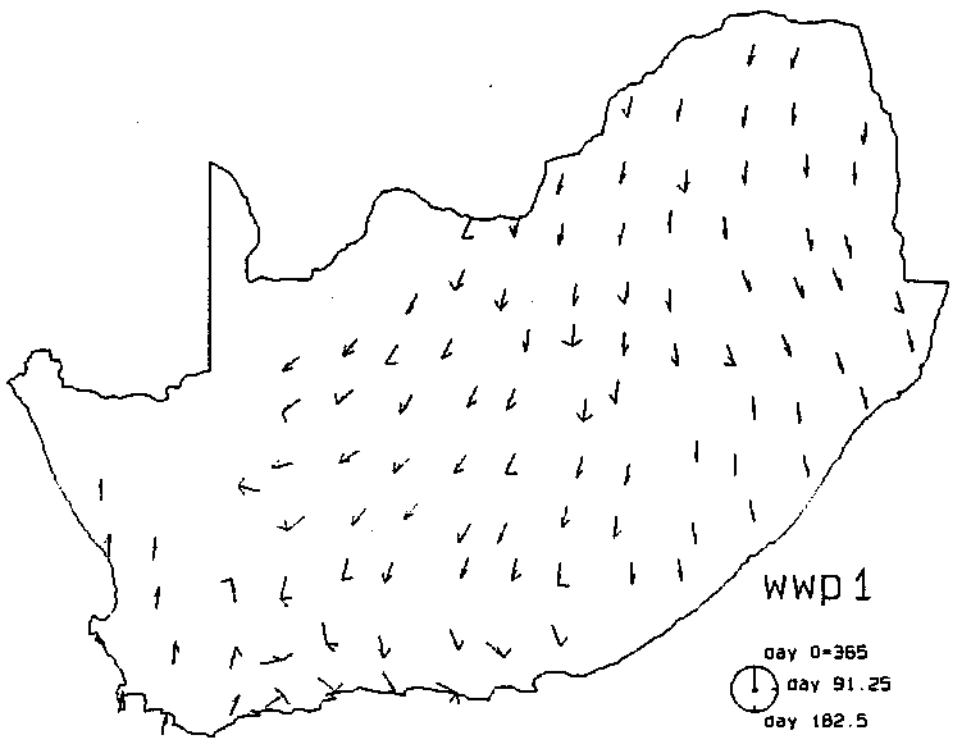


Figure 5.18: Map of kriging estimates at test sites.

(indicated by the two short lines emanating from each site in the figure) which represent a range of $\vartheta_i \pm \varepsilon_i$, where ϑ_i is the original data value at that site and $\varepsilon_i = \arccos(\sum_{j=1}^{100} \cos(\vartheta_{ij} - \overline{\vartheta_{ij}}))$, where the ϑ_{ij} are the individual bootstrap estimates described in Chapter 4. For our data set this range corresponds roughly to an approximate 95% confidence interval for data having a von Mises distribution, based on the formulae given in Section 9.6 of Upton and Fingleton (1989).

It can be seen in Figure 5.18 that the fit of the estimated values is generally good, except for five sites which lie in the area of change-over between the winter rainfall area in the south west and the summer rainfall area further to the north and east. The test data set is relatively sparse in this area; clearly more data points are needed for accurate estimation in this region. In practice, of course, the full data set has over 5000 points compared with the 325 used here, which should give much more accurate results throughout the country.

In the comparison described above the semi-variogram parameters were estimated directly from the empirical semi-variogram. However, since the solution given by equation 5.9 is only approximate there is no reason why these parameters should be optimal and it is likely that better results would be obtained if the parameters a and r were estimated via cross-validation. However, the cross-validation approach is more computationally intensive and it is thus of interest to test the sensitivity of the kriging method to the semi-variogram parameters. The estimation process was therefore repeated with a range of values of these parameters, but the results suggested that average estimation error was fairly insensitive to variation in the sill and range parameters (McNeill, 1993). Thus it would seem that, for this data set at least, using cross-validation to estimate optimal parameters is not likely to give much improvement over the computationally quicker method used here, based on modelling the empirical semi-variogram.

Another possible method of improving the accuracy of estimation would be by re-estimating the semi-variogram locally, as suggested by Haas (1990). For example, it is likely that the effective range of the spatial covariance would be smaller in the change-over region between the winter and summer rainfall area than it is in the middle of the summer rainfall area. However, as mentioned in Section 5.3.2, such a moving-window approach is excessively computationally intensive and thus effectively impractical in a project such as this. In addition, the advantage of a locally-calibrated semi-variogram model must be offset against the fact that relatively few data points will be used to estimate each local model and thus the model-fitting procedure will be less robust.

For the final estimation of the phase parameters throughout southern Africa, the semi-variograms for all parameters were estimated using the full data set, with the fitted semi-variogram models given in Table 5.3. Figure 5.19 shows maps of the resultant estimates at the centre of each Weather Bureau block.

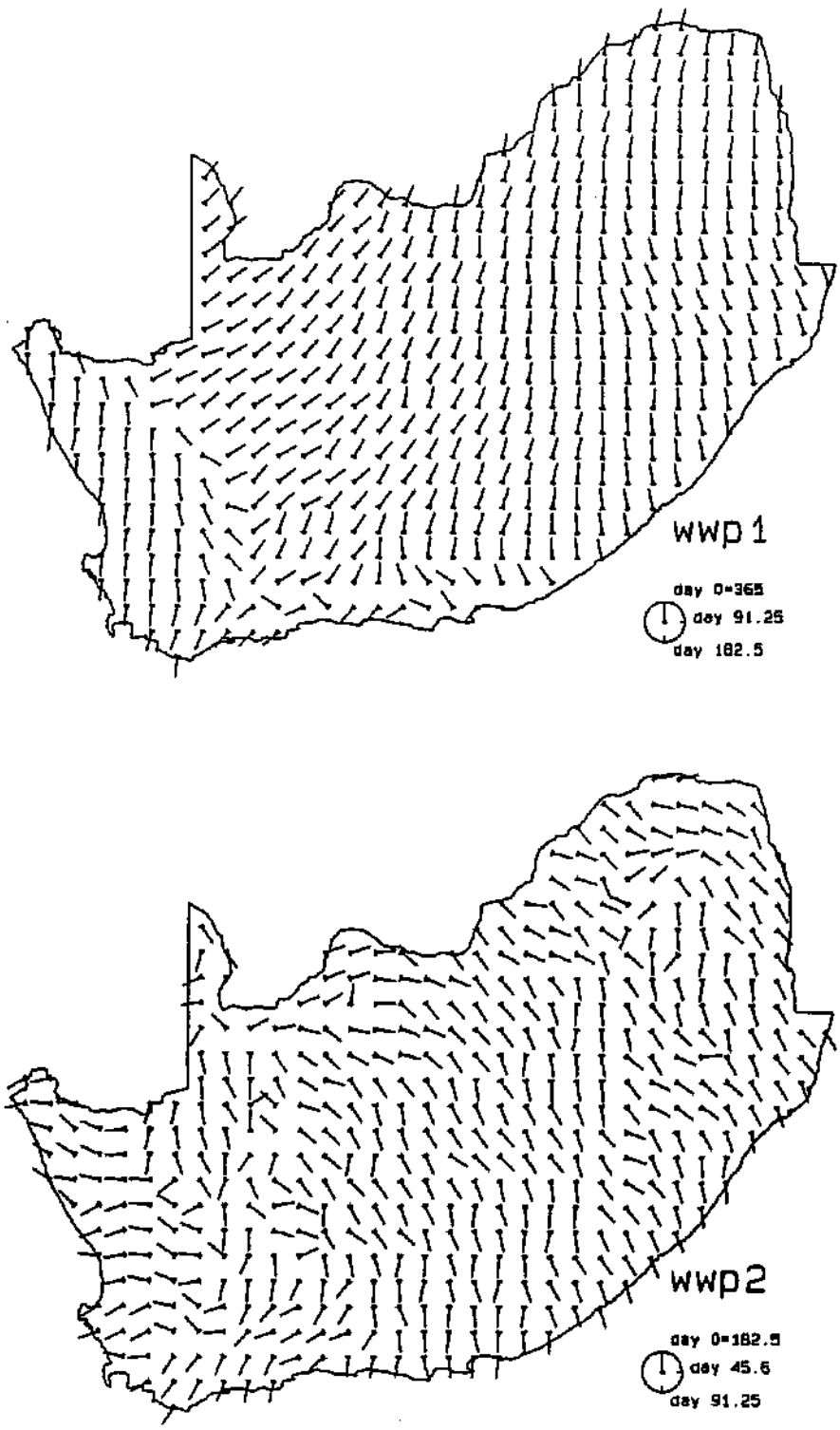


Figure 5.19: Estimated parameter values at centres of Weather Bureau blocks (phase parameters).

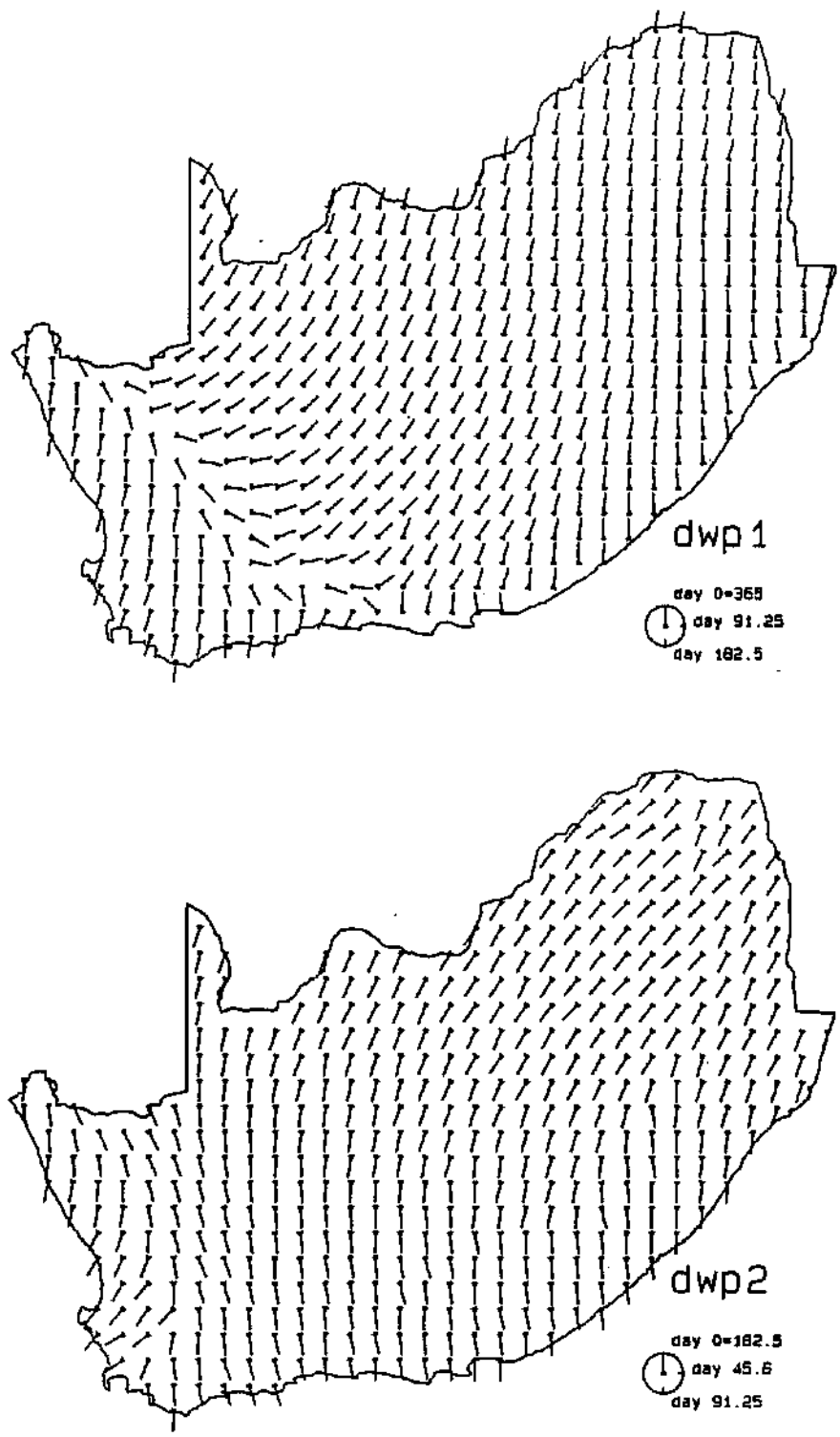


Figure 5.19: Estimated parameter values at centres of Weather Bureau blocks (phase parameters) (contd.).

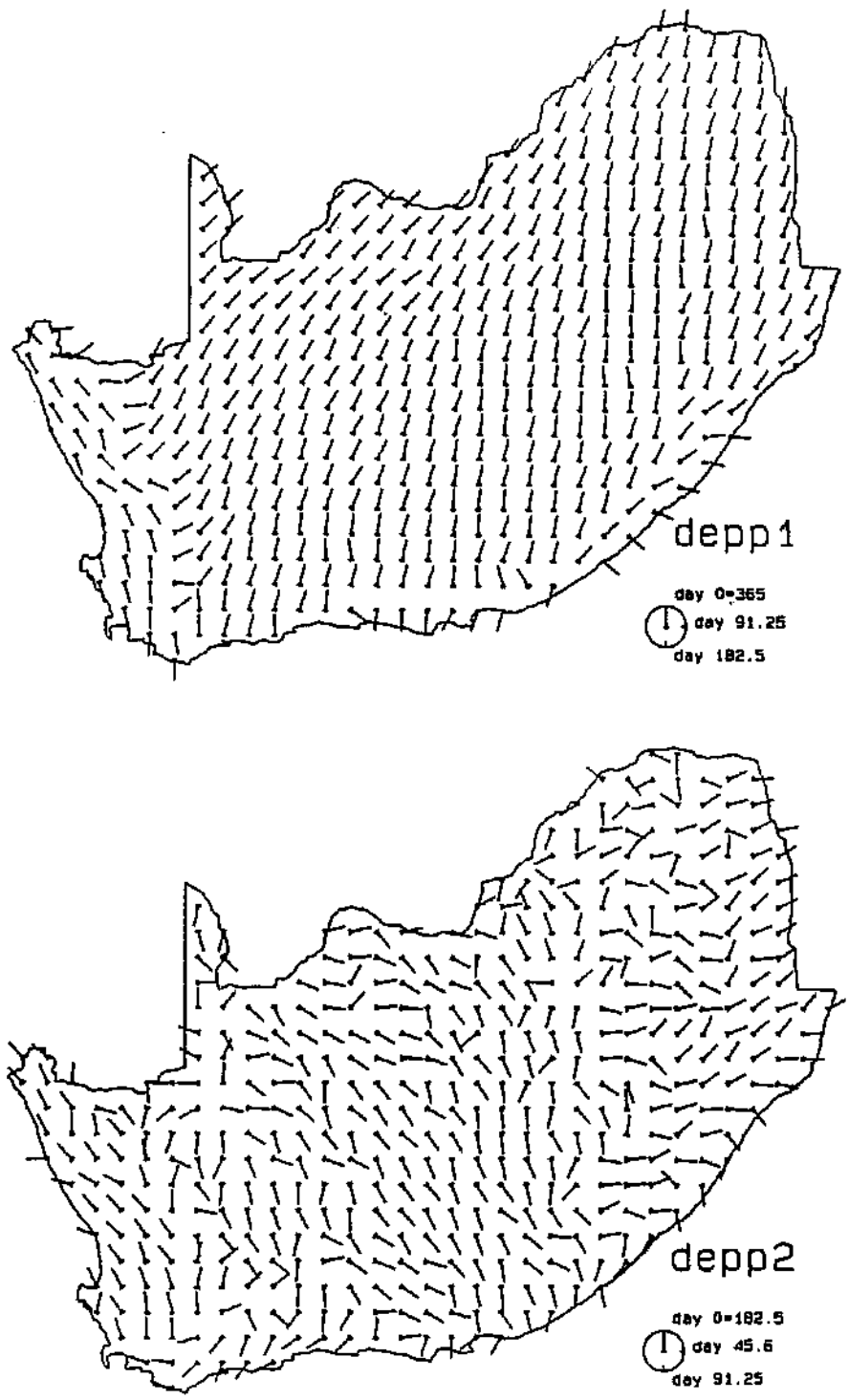


Figure 5.19: Estimated parameter values at centres of Weather Bureau blocks (phase parameters) (contd.).

5.5 Validation

The validation of the original daily rainfall model by Zucchini and Adamson (1984a) was discussed in Section 3.4. Briefly, the relevant characteristics of simulated daily rainfall data generated by the model at individual rainfall stations were compared with the corresponding values deduced directly from the original data. The kriging process described in this chapter extends the original 5000 stations at which the model is available to some 500000 points throughout southern Africa. Since most of these points do not coincide with the location of rain gauges, it is not possible to validate them in the same way. In addition, at those grid points which do coincide with rainfall stations we do not expect the estimated model parameters for the grid point to be equal to the values fitted to the original data at the station, since the kriging process takes into account the estimation error in the fitting of the original parameters and also the error introduced by the limited accuracy of the station locations. However, a comparison of grid point and station values will give some indication of the validity of the kriging process.

Rather than comparing individual model parameter values, it is more meaningful to compare derived characteristics, such as the mean annual precipitation, based on simulated data generated by the model; this enables us to test the model as a whole in the form in which it will be used in practice, and also allows comparison with the same statistics derived from other sources. We therefore calculated a mean annual precipitation (MAP) at the location of each of the 373 test sites described in Section 5.3.4 using four different methods:

- Using a 100 year simulation based on the daily rainfall model parameters estimated for that station.

- Using a 100 year simulation based on the daily rainfall model parame-

ters estimated by the kriging procedure at the grid point with the same latitude and longitude as the station.

- Using the MAP calculated directly from the daily rainfall data for that station held by CCWR.
- Taking the value of MAP from the CCWR data base of gridded MAP values, as estimated by Dent *et al.* (1989).

The results are shown in Table 5.5. There are a number of reasons for the differences between the four values; in particular, sampling variability introduced by the simulation process, uncertainty in the exact station location relative to the grid point, estimation error in the daily rainfall model parameters, estimation error in the kriging procedure, estimation error in the CCWR gridded MAP value calculations, outliers in the daily rainfall data, and also the use of data for a different time period (the grid values estimated by Dent *et al.* (1989) include data up to May 1987 and thus exclude the most recent rainfall data. In general, however, the agreement between the four sets of figures is quite close.

The MAP estimates based on the kriged values are also compared with the other three sets of values in Figure 5.20. As might be expected, the agreement with the values based on the daily rainfall model fitted to the station data (diagram A) is the closest; any discrepancy is due to the allowance for model-fitting error and the influence of neighbouring rainfall stations and of the altitude data in the kriging process. In diagram B, where the kriged values are compared with those calculated directly from the CCWR rainfall data at the station, the discrepancies incorporate also any inherent 'lack of fit' of the seasonal Markov chain model described in Chapter 3. In diagram C, the discrepancies are generally greater, as they now include also the effects of estimation errors inherent in the regression procedure used by Dent *et al.* (1989).

Station Code	Latitude		Longitude		Years of data	DRmodel (stn)	DRmodel (grid)	CCWR (stn)	CCWR (grid)
2885 W	-34	45	20	0	77	464	468	471	483
3032 W	-34	32	20	2	112	473	478	467	466
4891 W	-34	21	18	30	30	370	373	368	363
5605 A	-34	5	18	51	94	654	648	650	642
6733 W	-34	13	19	25	112	534	538	529	473
7698 A	-34	8	19	54	53	438	436	431	430
8136 A	-34	16	20	5	60	404	399	402	393
9815 W	-34	5	20	58	90	408	407	412	401
11065 W	-34	5	21	33	42	450	452	451	428
12215 W	-34	5	22	8	24	451	454	466	391
17582 A	-34	12	24	50	94	671	657	657	657
21055 W	-33	55	18	32	85	481	483	472	484
22038 W	-33	38	19	2	87	766	763	751	767
23678 W	-33	48	19	53	106	333	317	322	320
24197 W	-33	47	20	7	79	315	312	321	269
25599 W	-33	59	20	50	93	1026	1016	1019	1015
26510 W	-34	0	21	17	55	654	644	642	645
27302 W	-33	32	21	41	112	200	196	194	198
28838 W	-33	58	22	28	107	884	884	878	778
29805 W	-33	55	22	57	99	838	839	828	830
30090 W	-34	0	23	3	80	916	920	900	953
31237 W	-33	57	23	38	105	1011	1006	994	1003
32209 W	-33	59	24	7	96	1145	1145	1140	1144
33384 W	-33	54	24	43	43	566	564	550	642
34767 W	-33	47	25	26	28	435	425	449	400
35179 A	-33	59	25	36	39	615	634	604	611
36729 W	-33	39	26	25	105	649	653	641	637
37696 W	-33	36	26	54	89	666	667	660	672
40653 W	-33	23	18	22	98	479	479	475	461
41417 W	-33	27	18	44	112	462	456	453	452
42227 W	-33	17	19	6	113	487	485	473	415
43109 W	-33	19	19	34	33	589	594	605	577
44050 W	-33	20	20	2	85	227	226	222	223
45134 W	-33	14	20	35	98	167	166	158	165
46479 W	-33	29	21	16	112	320	321	316	368
47716 W	-33	26	21	54	63	415	414	424	415
48043 W	-33	13	22	2	98	166	166	169	169
49060 W	-33	30	22	32	76	321	321	333	324
50887 W	-33	17	23	30	78	269	272	262	233
51430 W	-33	10	23	45	63	261	260	263	255
52590 W	-33	20	24	20	98	232	233	236	206
53432 W	-33	12	24	45	60	238	238	247	231
54805 W	-33	25	25	27	29	385	385	384	356
55300 W	-33	30	25	40	31	336	334	332	328
56709 W	-33	19	26	24	101	607	610	605	693
57046AW	-33	16	26	32	110	706	695	704	696
58192 W	-33	12	27	7	110	527	527	510	523
59722 W	-33	2	27	55	72	844	826	832	757
60620 W	-32	50	17	51	30	248	242	236	235
61298 W	-32	58	18	10	17	284	290	276	263
62444 W	-32	54	18	45	111	483	487	464	410
63538 A	-32	58	19	18	38	622	623	614	529
68857 W	-32	47	21	59	106	124	116	120	117
69559 W	-32	49	22	19	67	163	164	167	169
70033 W	-32	33	22	32	64	191	189	195	193
71264 W	-32	54	23	9	72	180	181	186	184
72712 W	-32	52	23	54	47	193	199	202	214
73871 W	-32	31	24	30	98	281	270	276	284
74296 W	-32	56	24	40	80	273	273	271	268
75215 W	-32	35	25	8	90	320	324	323	327
76884 W	-32	44	26	0	86	477	480	465	469
77522 W	-32	42	26	18	97	439	452	428	439

Table 5.5: Comparison of MAP values (in mm).

Station Code	Latitude	Longitude	Years of data	DRmodel (stn)	DRmodel (grid)	CCWR (stn)	CCWR (grid)
78227 A	-32 47	26 38	112	522	524	514	513
79632 W	-32 32	27 22	104	977	970	952	956
80694 W	-32 34	27 54	105	738	738	724	647
81007 W	-32 37	28 1	72	821	819	799	693
83572 A	-32 2	18 20	14	232	221	227	223
84159 W	-32 9	18 36	55	240	234	230	240
85112 W	-32 22	19 4	82	704	694	692	457
87186 W	-32 6	20 7	52	366	355	371	369
88293 W	-32 23	20 40	61	246	276	240	339
89385 W	-32 25	21 13	57	295	294	296	293
90196 W	-32 16	21 37	71	173	173	172	176
91835 W	-32 25	22 26	67	189	188	192	194
92141 W	-32 21	22 35	96	237	238	239	238
93314 W	-32 14	23 11	101	213	213	218	219
94730 W	-32 10	23 55	59	406	416	392	406
95119 W	-32 29	24 4	100	286	276	282	285
96101 W	-32 11	24 34	104	284	285	279	281
97239 W	-32 29	25 8	68	362	363	346	348
98190 W	-32 10	25 37	86	321	323	315	312
99811 W	-32 1	26 26	84	446	447	438	431
100329 W	-32 29	26 41	106	1033	1018	997	1026
101604 W	-32 24	27 27	102	766	779	765	733
102762 W	-32 12	27 56	101	716	720	694	702
103516 W	-32 6	28 18	60	613	615	597	605
104762 W	-32 12	28 56	69	1145	1146	1133	1116
106850 W	-31 40	18 29	23	147	143	130	136
107396 W	-31 36	18 44	97	146	146	149	147
109215 W	-31 35	19 38	55	215	214	214	208
110385 W	-31 55	20 13	76	143	143	144	165
111373 W	-31 43	20 43	65	123	127	126	131
112346 W	-31 46	21 12	53	167	167	168	169
113025 W	-31 55	21 31	83	181	178	176	181
114747 W	-31 57	22 25	76	205	204	207	212
116083 W	-31 53	23 3	95	225	224	229	229
117447 W	-31 57	23 45	87	280	281	282	257
118395 W	-31 35	24 14	81	327	331	318	323
119209 W	-31 59	24 37	95	336	335	330	345
120338 W	-31 38	25 12	97	350	349	348	323
121518 W	-31 38	25 48	78	358	359	361	332
122480 W	-32 0	26 16	111	446	443	444	445
123304 W	-31 34	26 41	90	545	537	526	496
125150 W	-32 0	27 35	93	668	664	653	649
127485 A	-31 35	28 47	61	620	625	604	595
128032 W	-31 32	29 2	72	784	817	775	812
134478 A	-31 28	19 46	100	222	213	218	210
137337 W	-31 7	21 12	61	171	169	168	165
138041 W	-31 11	21 32	77	162	162	157	160
139658 W	-31 28	22 22	67	233	235	237	234
140616 W	-31 18	22 51	76	245	245	243	243
141329 W	-31 29	23 11	78	226	228	226	224
142805 W	-31 25	23 57	113	325	326	323	323
143579 W	-31 9	24 20	76	296	299	291	285
144900 W	-31 30	25 0	67	325	321	322	322
145029 A	-31 29	25 1	75	368	371	361	356
146588 W	-31 18	25 50	110	433	434	426	422
147409 W	-31 19	26 14	69	502	500	505	483
148352 A	-31 22	26 42	101	561	561	550	542
149082 A	-31 22	27 3	99	622	620	598	592
150085 W	-31 25	27 33	100	722	725	695	710
151604 W	-31 4	28 21	93	772	785	748	757
152190 W	-31 10	28 37	67	1115	1099	1091	1091
157035 W	-30 35	17 32	30	110	103	145	133

Table 5.5: Comparison of MAP values (contd.).

Station Code	Latitude		Longitude		Years of data	DRmodel (stn)	DRmodel (grid)	CCWR (stn)	CCWR (grid)
165898 A	-30	58	22	0	60	204	205	209	204
166238 W	-30	58	22	8	45	198	195	188	201
167665 W	-30	35	22	53	87	216	216	221	216
168250 W	-30	40	23	9	61	230	238	234	221
169090 W	-31	0	23	33	71	289	292	302	286
170009 A	-30	39	24	1	94	298	294	287	303
171758 W	-30	36	24	56	81	348	351	344	324
172163 W	-30	43	25	6	112	400	400	393	388
173497 W	-30	47	25	47	84	435	436	419	406
174550 W	-30	40	26	19	106	483	486	473	451
175371 W	-30	41	26	43	72	522	509	507	524
176631AW	-30	31	27	22	73	678	695	640	622
177178 A	-30	58	27	36	106	621	618	590	588
178689 W	-30	59	28	23	29	820	808	804	813
179790 W	-30	40	28	57	67	912	911	897	912
180032 W	-30	32	29	2	64	817	820	818	773
181073 W	-30	43	29	33	77	954	951	923	832
182379 A	-30	49	30	13	62	1231	1214	1195	1206
185023 W	-30	23	17	31	28	113	99	105	102
193561 A	-30	21	21	49	65	171	177	179	175
196375 W	-30	15	23	13	72	213	203	215	210
198838 W	-30	26	24	28	112	332	332	327	335
199107 W	-30	17	24	34	68	315	316	313	307
200486 W	-30	16	25	16	97	401	406	390	395
201381 W	-30	1	25	43	76	437	436	421	426
202575 W	-30	5	26	20	33	523	531	489	487
203657 W	-30	27	26	52	60	636	641	623	587
204138 W	-30	16	27	5	81	715	714	686	685
205385 W	-30	25	27	43	42	830	824	834	707
206843 W	-30	3	28	29	43	617	621	609	585
207560 W	-30	20	28	49	76	707	711	673	682
208406 W	-30	16	29	14	75	750	754	733	743
209039 W	-30	9	29	32	68	1156	1133	1130	1108
210002 W	-30	2	30	1	75	881	873	810	925
211661 A	-30	1	30	53	39	1042	1047	1019	1019
214670 W	-29	40	17	53	114	219	210	216	216
223344 W	-29	44	22	12	24	235	238	235	202
224430 W	-29	40	22	45	57	250	243	250	228
225679 W	-29	49	23	23	79	267	269	264	271
226327 W	-29	57	23	41	85	244	248	248	243
227127 W	-29	37	24	5	86	321	324	311	302
228567 W	-29	57	24	49	82	383	390	387	369
229556 A	-29	46	25	19	41	438	458	428	422
230816 W	-29	36	25	58	46	515	506	509	489
231279 W	-29	39	26	10	85	500	496	483	479
232823 W	-29	43	26	58	77	616	619	585	582
233044 W	-29	44	27	2	70	537	559	430	503
236677 W	-29	47	28	53	18	619	654	586	582
237471 W	-29	51	29	16	53	1207	1199	1192	1184
238837 A	-29	57	29	58	49	881	866	864	862
239482 A	-29	32	30	17	73	900	889	875	876
240891 W	-29	51	31	0	111	1032	1025	1020	1020
241019 W	-29	49	31	1	56	995	1000	966	967
251261 W	-29	21	21	9	78	138	144	144	141
252894 W	-29	24	22	0	63	181	184	191	196
253648 W	-29	18	22	22	69	216	222	227	223
255202 W	-29	22	23	7	89	223	221	229	235
256453 W	-29	3	23	46	103	340	334	331	330
257845 W	-29	5	24	29	77	380	386	364	366
258458 W	-29	8	24	46	97	395	400	376	381
259727 W	-29	7	25	25	83	443	441	426	420
260878 W	-29	18	25	53	61	496	500	467	478

Table 5.5: Comparison of MAP values (contd.).

Station Code	Latitude		Longitude		Years of data	DAmodel (stn)	DAmodel (grid)	CCWR (stn)	CCWR (grid)
261722 W	-29	2	26	25	86	566	568	540	487
262129 W	-29	9	26	35	66	564	556	554	516
263859 A	-29	19	27	29	47	755	745	717	714
264022 W	-29	22	27	31	50	820	819	776	735
268640 A	-29	10	29	52	80	894	901	877	877
269532 A	-29	22	30	18	68	1221	1215	1184	1195
270544 W	-29	4	30	49	83	1085	1072	1092	1081
271099 W	-29	9	31	4	66	1089	1086	1083	1080
272121 W	-29	1	31	35	78	1077	1073	1025	1063
282823 W	-28	43	20	58	47	169	163	165	159
283098 W	-28	38	21	4	70	149	151	151	155
286824 W	-28	44	22	58	18	308	313	297	310
287441 W	-28	51	23	15	89	294	299	296	293
288528 W	-28	48	23	48	71	340	343	323	329
289102 W	-28	42	24	4	55	334	332	328	333
290463 W	-28	43	24	46	51	419	418	406	410
291899 A	-28	59	25	30	77	431	422	410	410
292461 W	-28	41	25	46	65	436	444	433	447
293597 A	-28	57	26	20	64	575	581	542	529
294847 W	-28	37	26	59	62	590	588	570	563
295408 W	-28	48	27	14	66	654	652	635	631
296378 W	-28	49	27	43	78	710	706	691	711
297694 W	-28	34	28	24	42	777	780	752	761
298301 W	-28	31	28	41	68	838	836	825	822
299419 W	-28	59	29	14	26	1292	1260	1356	1262
300567 A	-28	57	29	49	88	761	761	724	723
301692 A	-28	32	30	24	62	804	803	754	772
302687 W	-28	57	30	53	43	938	917	883	758
303633 W	-28	33	31	22	28	827	822	833	911
304822 W	-28	42	31	58	70	1145	1142	1102	1109
305037 W	-28	37	32	2	71	1026	1022	994	1002
317447 A	-28	27	21	15	86	163	162	151	174
320348 W	-28	18	22	42	92	339	347	326	334
321110 W	-28	20	23	4	74	332	336	327	328
322071 W	-28	11	23	33	70	386	390	384	371
323649 W	-28	19	24	22	95	412	414	407	383
324607 W	-28	7	24	51	79	443	448	424	426
325870 W	-28	30	25	29	39	367	376	368	408
326668 W	-28	8	25	53	60	522	510	496	470
327883 W	-28	13	26	30	73	503	504	491	482
328628 W	-28	28	26	51	45	602	598	580	545
329215 W	-28	5	27	8	60	549	550	545	548
330750 W	-28	30	27	55	65	666	663	656	643
331058 W	-28	28	28	2	61	810	805	790	781
332663 W	-28	3	28	53	53	685	680	663	680
333226 W	-28	16	29	8	76	646	643	618	618
334825 W	-28	15	29	58	72	922	915	909	907
335550 A	-28	10	30	19	60	808	798	778	778
336283 W	-28	13	30	40	61	804	796	782	853
337148 W	-28	28	31	5	53	848	839	826	994
339354 A	-28	24	32	12	66	935	924	886	884
356285 W	-27	45	22	40	70	368	358	368	339
358049 W	-27	49	23	32	40	509	516	490	463
359606 W	-27	58	24	27	88	466	459	450	415
360597 A	-27	57	24	50	51	462	462	441	430
361354 W	-27	54	25	12	65	447	449	414	436
362862 W	-27	52	25	59	53	551	544	548	532
363651 W	-27	51	26	22	66	476	473	450	451
364322 W	-27	52	26	41	83	512	514	502	497
365400 W	-27	40	27	14	65	607	606	586	580
366710 W	-27	50	27	54	47	659	653	626	626
367766 W	-27	48	28	26	96	715	718	692	696

Table 5.5: Comparison of MAP values (contd.).

Station Code	Latitude	Longitude	Years of data	DRmodel (stn)	DRmodel (grid)	CCWR (stn)	CCWR (grid)
368634 W	-27 34	28 52	60	758	750	721	760
369238 W	-27 58	29 8	84	717	717	691	691
370486 W	-27 36	29 47	67	808	804	785	863
371579 W	-27 39	30 20	73	762	766	736	743
372652 W	-27 42	30 59	75	799	801	781	764
373680 W	-27 50	31 23	63	1559	1544	1531	1547
374264 W	-27 54	31 39	71	933	929	883	887
375366 W	-27 36	32 13	41	680	680	642	622
392148 W	-27 28	22 35	65	327	320	332	314
393778 W	-27 25	23 26	34	475	443	436	480
394874 W	-27 4	24 0	14	359	362	348	400
395855 W	-27 15	24 29	36	329	352	325	332
396813 W	-27 3	24 58	74	474	474	466	427
397086 W	-27 26	25 3	21	419	418	413	460
398556 W	-27 16	25 49	84	526	526	511	512
399667 W	-27 7	26 23	42	575	559	568	557
400203 W	-27 23	26 37	80	558	554	526	541
401798 W	-27 18	27 27	75	608	609	586	575
402061 W	-27 21	27 33	74	644	654	613	625
403886 W	-27 16	28 30	51	658	672	623	647
404316 W	-27 16	28 41	79	566	570	546	536
405753 W	-27 3	29 26	56	727	728	731	705
406607 W	-27 7	29 51	84	776	768	768	760
407639 W	-27 9	30 22	63	795	804	770	796
408798 W	-27 18	30 57	66	855	864	820	820
409375 W	-27 15	31 13	84	805	808	768	783
410133 W	-27 13	31 35	40	934	914	894	896
411175 W	-27 25	32 6	45	645	633	603	603
430354 W	-26 54	23 42	37	349	349	345	343
431896 W	-26 56	24 30	42	462	473	452	463
432237 A	-26 57	24 38	71	466	465	445	437
433858 W	-26 48	25 29	67	535	532	509	504
434020 W	-26 50	25 31	65	541	535	529	526
435400 W	-26 40	26 14	61	627	633	622	614
436747 W	-26 57	26 55	69	607	598	600	597
437134 A	-26 44	27 5	77	644	631	622	618
438315 W	-26 45	27 41	75	699	691	694	657
439764 W	-26 44	28 26	76	720	718	695	684
440157 W	-26 37	28 36	78	737	727	718	721
441777 W	-26 57	29 26	68	729	723	697	692
442781 A	-26 31	29 57	49	755	764	724	723
443451 W	-26 31	30 16	83	813	808	797	786
444748 W	-26 56	30 55	49	782	776	742	825
446741 S	-26 51	31 55	67	580	580	553	553
468318 W	-26 18	24 11	78	449	451	437	407
471490 W	-26 10	25 47	54	573	573	571	565
472175 W	-26 25	26 6	62	607	611	595	590
473586 W	-26 26	26 53	69	620	608	609	574
474198 W	-26 16	27 7	78	659	662	638	603
475861 W	-26 11	26 0	91	817	807	814	807
476072 W	-26 12	26 3	96	860	798	844	839
477309 W	-26 9	26 41	83	702	706	692	678
478360 W	-26 30	29 12	82	750	739	740	731
479545 W	-26 5	29 49	66	690	686	645	663
480689 W	-26 19	30 30	38	838	831	839	837
481167 W	-26 17	30 36	78	900	882	885	881
482357 W	-26 27	31 12	65	1137	1129	1126	1121
483053 W	-26 23	31 32	52	687	664	658	703
504836 W	-25 58	23 58	20	407	390	394	421
505834 W	-25 54	24 26	34	407	397	397	368
508649 W	-25 49	25 52	74	590	585	589	547
509759 W	-25 39	26 26	79	599	605	596	599

Table 5.5: Comparison of MAP values (contd.).

Station Code	Latitude	Longitude	Years of data	DRmodel (stn)	DRmodel (grid)	CCWR (stn)	CCWR (grid)
510712 W	-25 52	26 54	77	667	672	667	648
511469 W	-25 49	27 16	70	692	684	668	673
512613 W	-25 43	27 51	69	696	689	667	675
513382 W	-25 52	28 13	74	706	702	685	688
514618 W	-25 48	28 51	83	724	723	694	684
515826 W	-25 46	29 28	52	709	695	686	689
516285 A	-25 45	29 40	82	784	777	764	757
517430 W	-25 40	30 15	83	798	797	781	790
518859 W	-25 49	30 59	81	819	828	786	872
519017 W	-25 47	31 1	28	715	713	744	682
520450 W	-26 0	31 45	35	764	768	739	594
545626 W	-25 26	25 51	70	616	605	583	588
546082 W	-25 22	26 3	66	605	610	596	587
549354 W	-25 24	27 42	62	594	594	579	604
550567 W	-25 27	28 19	52	601	615	570	612
551120 W	-25 30	28 34	74	681	671	660	644
552610 W	-25 10	29 21	39	623	631	604	625
553851 W	-25 21	29 52	66	739	725	718	692
554788 W	-25 6	30 27	72	695	691	677	678
555487 W	-25 7	30 47	49	1136	1151	1098	1095
556110 W	-25 20	31 4	54	915	900	901	848
557029 W	-25 29	31 31	35	699	673	670	652
585526 W	-24 48	26 18	53	583	591	568	585
586441 W	-24 51	26 45	67	580	587	551	572
587350 W	-24 50	27 12	30	697	695	669	632
588406 W	-24 46	27 44	83	617	622	601	614
589594 W	-24 54	28 20	30	654	631	632	628
590307 W	-24 37	28 41	72	635	632	623	615
591538 W	-24 58	29 18	50	520	522	500	515
593015 W	-24 45	30 1	64	562	558	559	552
594141 W	-24 51	30 35	67	598	602	570	551
595202 W	-24 52	31 7	36	1053	1046	1022	1019
630511 W	-24 1	27 18	34	512	493	504	506
631011 W	-24 11	27 31	58	525	523	513	505
632465 W	-24 15	28 16	48	599	586	563	610
633503 W	-24 23	28 47	49	673	668	664	666
634011 W	-24 11	29 1	27	595	592	597	624
635076 W	-24 16	29 33	54	563	561	534	537
636308 W	-24 8	30 11	61	959	971	948	968
637720 W	-24 30	30 54	45	834	827	790	903
638748 W	-24 28	31 25	35	572	561	561	581
639504 W	-24 24	31 47	41	597	589	556	562
673264 W	-23 44	27 10	36	473	463	442	447
675117 W	-23 57	28 4	63	556	564	540	553
676523 W	-23 43	28 48	37	475	488	474	499
677834 W	-23 54	29 28	82	493	486	482	485
678725 W	-23 35	29 55	44	640	629	618	589
679268 W	-23 58	30 9	51	1320	1338	1268	1147
680354 W	-23 54	30 42	41	537	541	521	537
681069 W	-23 39	31 3	33	486	497	471	501
718674 W	-23 4	28 0	51	416	416	420	403
719370 A	-23 10	28 13	28	418	434	402	391
720727 W	-23 7	28 55	60	555	536	557	686
721772 W	-23 22	29 26	53	415	416	405	402
722497 W	-23 17	29 47	52	428	430	405	411
723060 W	-23 20	30 3	58	787	796	749	841
724790 W	-23 10	30 57	33	550	546	575	566
762532 W	-22 52	28 18	49	369	378	368	485
763313 W	-22 43	28 41	29	448	456	417	410
764161 W	-22 41	29 6	61	364	347	374	377
765869 W	-22 59	29 59	39	783	786	748	794
766863 A	-22 53	30 29	37	1072	1080	1026	1077

Table 5.5: Comparison of MAP values (contd.).

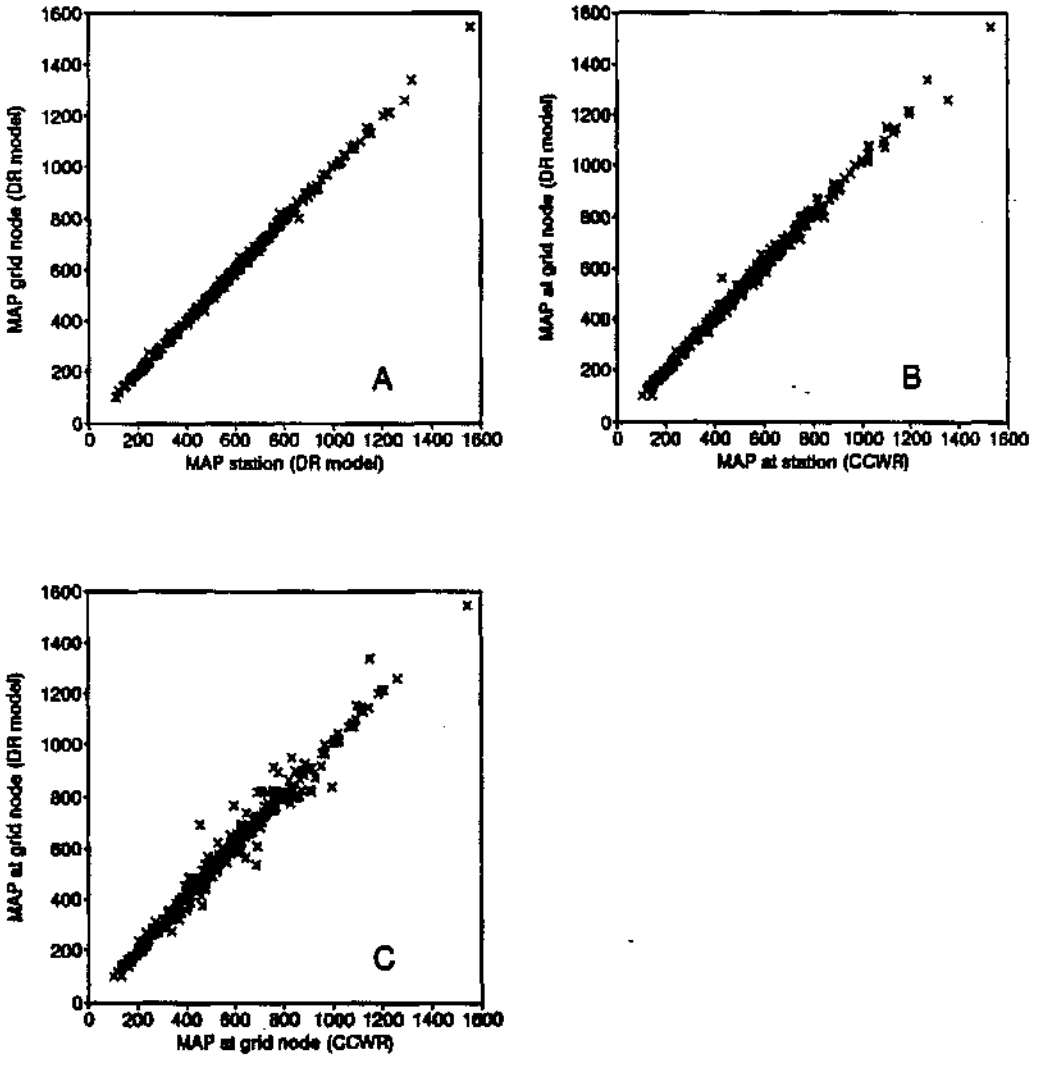


Figure 5.20: Comparison of MAP values (in mm).

Chapter 6

Implementing the Model

The application of generated daily rainfall sequences to estimate statistics of interest has been discussed in Zucchini and Adamson (1984a) and Zucchini *et al.* (1992). This chapter describes the algorithm required to generate the rainfall sequences.

To generate an artificial rainfall sequence at a particular site one first needs to know the parameters of the model for that site. If there is a rainfall station at the site whose model parameters have been calibrated, then it is only necessary to know the Weather Bureau station number. If there is no rainfall station at the site, then one has to use the interpolated parameter values. To obtain these one needs to give the longitude and latitude of the site.

The user also specifies the length (in years) of the required generated sequence. The output is in the form of daily values given in *tenths* of a *mm* so that, for example, a rainfall depth of $10,2$ *mm* is represented by the integer 102.

Algorithm for generating artificial rainfall sequences

Step 1: Input

- number of years of daily rainfall to be generated

- Read NG

– either station number of interest

- Read $STNNO$

– or grid points (i.e. longitude and latitude) for the site.

- Read $LONG, LAT$

Step 2: Extract model parameter estimates for the site of interest from the data base.

- Get $WWA0, WWA1, WWA2, WWP1, WWP2, DWA0, DWA1, DWA2, DWP1, DWP2, DEPA0, DEPA1, DEPA2, DEPP1, DEPP2, CV$

Step 3: Set initial state of day to be dry.

- $STATE = 0$

Step 4: Compute

– Probability that day t is wet given that day $t-1$ is wet, $t = 1, 2, \ldots, 365$

- $W = 0.01721421$
- $LOGIT = WWA0 + WWA1 * \cos(W * (t - 1 - WWP1)) + WWA2 * \cos(2 * W * (t - 1 - WWP2))$
- $PWW(t) = \exp(LOGIT)/(1 + \exp(LOGIT))$

– Probability that day t is wet given that day $t-1$ is dry, $t = 1, 2, \ldots, 365$

- $LOGIT = DWA0 + DWA1 * \cos(W * (t - 1 - DWP1)) + DWA2 * \cos(2 * W * (t - 1 - DWP2))$
- $PDW(t) = \exp(LOGIT)/(1 + \exp(LOGIT))$

– The shape and scale parameters of the Weibull distribution.

- $BI = \frac{1}{B}$.

  The shape parameter, B, is given by equation (3.10).

- $GI = 1/\Gamma(1 + BI)$
- $M(t) = (DEPA0 + DEPA1 * \cos(W * (t - 1 - DEPP1)) + DEPA2 * \cos(2 * W * (t - 1 - DEPP2))) * GI$

Step 5: Loop over years NY and over days t.

- $NY = 1, 2, \ldots, NG$ and $t = 1, 2, \ldots, 365$

Step 6: Generate uniform random number between 0 and 1, inclusive ($U(0, 1)$).

- Generate RND

Step 7: If $U(0, 1)$ random number is less than the probability of a wet day following a day with the status of the previous time period then

- the status of the present time period is wet.

Otherwise

- the status of the present time period is dry.
  - If $RND < PWW(t)$ given $STATE = 1$
  - or $RND < PDW(t)$ given $STATE = 0$ then
    - $STATE = 0$
  - Else
    - $STATE = 1$

Step 8: If present state is wet than determine the rainfall depth, (else set rain = 0).

- If $STATE = 1$ then
  - $GR(NY, t) = M(t) * (-\log(RND))^{BI}$

- Otherwise
  - $GR(NY,t) = 0$

Step 9: Repeat loop from Step 5 until enough years of rainfall have been generated.

- End of t loop and NY loop.

Step 10: Output the generated daily rainfall sequence.

- Write $GR(NY,t), NY = 1,2,\ldots,NG, \quad t = 1,2,\ldots,365.$

To generate and store 1000 years of daily rainfall on a 386 micro-computer (with math co-processor) takes less than 2 minutes. A FORTRAN version of this algorithm, which makes use of the parameter values at calibrated stations and the interpolated grid point values described in this report, is available from CCWR (see Chapter 7 and Appendix C).

Chapter 7

Summary and Recommendations

7.1 Summary

The main objective of the project described in this report was to produce estimates of the parameters of the daily rainfall model of Zucchini and Adamson (1984a) for sites throughout southern Africa at which there is little or no rainfall data available, thereby making it possible to use the model to generate artificial rainfall sequences and study rainfall characteristics at any given location or over any given area in southern Africa. Examples of the type of questions that the model can be used to answer are given in Chapter 1.

The parameters of the daily rainfall model have been interpolated on a regular grid one minute of degree square throughout southern Africa, that is, at a resolution of about 1,5 kilometres, making the parameter estimates of the model available for approximately 500 000 sites.

As was pointed out in the introduction, the daily rainfall model is routinely used by researchers and decision makers in a wide variety of applications. It is hoped, now that the model is now applicable at practically any

site in southern Africa, that it will find even wider application.

It needs to be emphasised that although the theory behind the model is rather technical, the model is easy to use by anybody who can operate a micro-computer. No statistical or other specialist knowledge is required to *apply* the model. The feedback that we have received, during the last eight or nine years, from users with very different mathematical backgrounds, has been encouraging; no-one has indicated that they found the model difficult to apply. We are not aware of any user who has misunderstood what it is that the model provides or who has misinterpreted the estimates derived from the model.

One of the by-products of the project has been the contribution to the theory of kriging, namely the development of a technique for the kriging of circular variables, described in McNeill (1993). The report also briefly reviews kriging and other interpolation techniques and comments on their suitability in the context of hydrological data. This provides a convenient starting point and an up-to-date list of references for researchers wishing to interpolate other values.

7.2 Recommendations

The daily rainfall model has 16 parameters. We have generated estimates of these parameters for approximately 500 000 grid points, covering southern Africa on a grid of 1' by 1'. This information is currently stored at the CCWR; the data file occupies 3 megabytes of computer disc space for each of the 16 parameters or almost 50 megabytes in total. As this quantity of information is too large to be conveniently distributed in its entirety to individual researchers and other interested parties, we recommend that the CCWR be approached to:

- store the data.

- extend their present service of supplying artificially generated rainfall sequences via the 'DRAINGEN' program to incorporate an option for using the grid point data. (They currently supply generated sequences for the 2550 stations covered in the Zucchini and Adamson (1984a) report.)

- maintain an archive of the interpolation software which was used to estimate the grid point values (see Appendix C) so that it will be possible to re-run the programs at some future date to update the parameter estimates.

We also recommend that some consideration be given to finding appropriate means of publicising the existence of the model and its potential uses. We believe that the number of current users is much smaller than the number of potential users, who are either unaware of the model or who might be mistakenly under the impression that it is a complicated tool requiring specialist knowledge. With this in mind, a summary version of the current report, together with a PC compatible diskette containing a small data set and sample programs, is currently in preparation. Further software development, aimed at providing application tools to make optimum utilization of the generated data, would be a valuable addition.

Further research is required to develop methodology for generating simulated sequences of daily rainfall for an area rather than a single point.

References

ADAMSON, P.T. (1981). Southern African storm rainfall. *Technical Report 102.* Department of Water Affairs, Directorate of Scientific Services, Private Bag X313, Pretoria.

AHMED, S. and DE MARSILY, G. (1987). Comparison of geostatistical methods for estimating transmissivity using data on transmissivity and specific capacity. *Water Resources Research*, **23**, 1717–1737.

ARMSTRONG, M., CHETBOUN, G. and HUBERT, P. (1992). Kriging the rainfall in Lesotho. In *Proceedings of the Fourth International Geostatistics Congress.* Lisbon. (to be published).

BATES, D.M., LINDSTROM, M.J., WAHBA, G. and YANDELL, B.S. (1987). GCVPACK - routines for generalized cross-validation. *Commun. Statist. Simula.*, **16**, 263–297.

BATSCHELET, E. (1981). *Circular Statistics in Biology.* London, Academic Press.

BOX, G.E.P. and JENKINS, G.M. (1970). *Time Series Analysis: Forecasting and Control.* Holden Day, New York.

BRECKLING, J. (1989). *The Analysis of Directional Time Series: Applications to Wind Speed and Direction.* Springer-Verlag, Berlin.

BUISHAND, T.A. (1977). *Stochastic modelling of daily rainfall sequences.* Meded. Landbouwhagesch, Wageningen, 77–83.

BUISHAND, T.A. (1978). Some remarks on the use of daily rainfall models. *Journal of Hydrology*, **36**, 295–308.

CASKEY, J.E. (1963). A Markov chain model for the probability of precipitation occurrence in intervals of various length. *Monthly Weather Review*, **91**, 298–301.

CHATFIELD, C. (1980). *The Analysis of Time Series: An Introduction.* (2nd ed.) Chapman and Hall, New York.

CLARK, I. (1982). *Practical Geostatistics.* Elsevier, London

CLARK, I., BASINGER, K.L. and HARPER, W.V. (1989). MUCK: A novel approach to co-kriging. In *Proceedings of the Conference on Geostatistical, Sensitivity, and Uncertainty Methods for Ground-Water Flow and Radionuclide Transport Modelling*, B.E.Buxton, ed. Battelle Press, Columbus, Ohio, 473–493.

CRESSIE, N. (1985). Fitting variogram models by weighted least squares. *Math. Geol.*, **17**, 563–586.

CRESSIE, N. (1991). *Statistics for Spatial Data.* Wiley, New York.

CRESSIE, N. (1990). Reply to Wahba's letter to the editor. *American Statistician*, **44**, 256–258.

CREUTIN, J.D. and OBLED, C. (1982). Objective analyses and mapping techniques for rainfall fields: an objective comparison. *Water Resources Research*, **18**, 413–431.

DENT, M.C., LYNCH, S.D. and SCHULZE, R.E. (1989). Mapping mean annual and other rainfall statistics over southern Africa. *Water Research Commission Report* 109/1/89, Water Research Commission, Pretoria.

DIGGLE, P.J. (1990). *Time Series: A Biostatistical Introduction.* Clarendon Press, Oxford.

DOBSON, A.J. (1983). *An Introduction to Statistical Modelling.* Chapman and Hall, New York.

DOVE, K. (1888). *Das Klima des Aussertropischen Südafrika.* Van den Hoek und Ruprech, Göttingen.

DRAPER, N.R.and SMITH, H. (1981). *Applied Regression Analysis.* Wiley, New York.

DUCHON, J. (1976). Interpolation des fonctions de deux variables suivant le principe de la flexion des plaques minces. *R.A.I.R.O. Analyse Numérique*, **10**, 5–12.

EFRON, B. (1979). Bootstrap methods: another look at the jackknife. *Annals of Statistics*, **7**, 1–26.

FISHER, N.I. and LEWIS, T. (1985). A note on spherical splines. *J. R. Statist. Soc. B,* **47**, 482–488.

GABRIEL, K.R. and NEUMANN, J. (1962). A Markov chain model for daily rainfall occurrence at Tel Aviv. *Quarterly Journal of the Royal Meteorological Society*, **88**, 90–95.

GANDIN, L.S. (1963). *Objective Analysis of Meteorological Fields.* Leningrad: GIMIZ. (translated from the Russian in 1965 by R. Hardin, Israel Program for Scientific Translations: Jerusalem).

GOLDBERGER, A.S. (1962). Best linear unbiased prediction in the generalized linear regression model. *J. Amer. Statist. Assoc.*, **57**, 369–375.

GRANT, F. (1957). A problem in the analysis of geophysical data. *Geophysics*, **22**, 309–344.

HAAN, C.T., ALLEN, D.M. and STREET, J.O. (1976). A Markov chain model of daily rainfall. *Water Resources Research*, **12**, 443–449.

HAAS, T.C. (1990). Lognormal and moving window methods of estimating acid deposition. *J. Amer. Statist. Assoc.*, **85**, 950–963.

HARDY, R.L. (1971). Multiquadric equations of topography and other irregular surfaces. *Journal of Geophysical Research*, **76**, 1905–1915.

HARTER, H.L. and MOORE, A.H. (1967). Asymptotic variances and covariances of maximum-likelihood estimators, from censored samples, of the parameters of Weibull and gamma populations. *Annals of Mathematical Statistics*, **38**, 557–570.

HASTIE, T.J. and TIBSHIRANI, R.J. (1990). *Generalized Additive Models.* Chapman and Hall, London.

HOBSON, R.D. (1972). Surface roughness in topography: quantitative approach. In *Spatial Analysis in Geomorphology*, R.J.Chorley, ed. Methuen, London, 221–245.

HOPKINS, J.W. and ROBILLARD, P. (1964). Some statistics of daily rainfall occurrences for the Canadian prairie provinces. *Journal of Applied Meteorology*, **3**, 600–602.

HUDSON, G. (1992). Kriging temperature in Scotland using the external drift method. In *Proceedings of the Fourth International Geostatistics Congress*. Lisbon. (to be published).

HUGHES, D.A. (1982). The relationship between mean annual rainfall and physiographic variables applied to a coastal region of southern Africa. *S. Afr. Geog. Jour.*, **64**, 41–50.

HUTCHINSON, P. (1968) An analysis of the effect of topography on rainfall in the Taieri catchment area, Otago. *Earth Science Journal*, **2**, 51–68.

ISAAKS, E.H. and SRIVASTAVA, R.M. (1989). *Applied Geostatistics.* Oxford University Press, New York.

ISON, N.T., FEYERHERM, A.M. and BARK, L.D. (1971). Wet period precipitation and the gamma distribution. *Journal of Applied Meteorology*, **10**, 658–665.

JACKSON S.P. (1951). Climates of Southern Africa. *S. Afr. Geogr. J.*, **38**, 17–37.

JOHNSON, N.L. and KOTZ, S. (1970). *Distributions in Statistics: Continuous Univariate Distributions–1.* Wiley, New York.

JOURNEL, A.G. and HUIJBREGTS, C.J. (1978). *Mining Geostatistics.* Academic Press, London.

JOURNEL, A.G. and ROSSI, M.E. (1989). When do we need a trend model in kriging? *Math. Geol.*, **21**, 715–739.

JOWETT, G.H. (1955). Sampling properties of local statistics in stationary stochastic series. *Biometrika*, **42**, 160–169.

JUPP, P.E. and MARDIA, K.V. (1989). A unified view of the theory of directional statistics, 1975-1988. *International Statistical Review*, **57**, 261–294.

KIMELDORF, G.and WAHBA, G. (1970). A correspondence between Bayesian estimation of stochastic processes and smoothing by splines. *Annals of Mathematical Statistics*, **41**, 495–502.

KRAJEWSKI, W.F. (1987). Cokriging radar-rainfall and rain gage data. *Journal of Geophysical Research*, **92**, 9571–9580.

KRUMBEIN, W.C. (1959). Trend-surface analysis of contour-type maps with irregular control-point spacing *Journal of Geophysical Research*, **64**, 823–834.

LEE, P.S., LYNN, P.P. and SHAW, E.M. (1974). Comparison of multiquadric surfaces for the estimation of areal rainfall. *Hydrological Sciences Bulletin*, **19**, 303–317.

LINHART, H. and ZUCCHINI, W. (1986). *Model Selection*. Wiley, New York.

LONDON, W. and EMMITT, G.D. (1986). Topographical influences on radar echo properties - implications to weather modification projects in mountainous terrain. 2nd Conference on Planned and Inadvertent Weather Modification.

MARDIA, K.V. (1972). *Statistics of Directional Data*. Academic Press, London.

MARDIA, K.V. (1975). Statistics of directional data (with discussion). *J. R. Statist. Soc. B*, **37**, 349–393.

MATHERON, G. (1963). Principles of geostatistics. *Econ. Geol.*, **58**, 1246–1266.

No.5. Fontainebleau, France.

MATHERON, G. (1982). Pour une analyse krigeante de données régionalisées. Note interne, N-732, Centre de Géostatistique, Fontainebleau, France.

McNEILL, L. (1993). Interpolation and smoothing of mapped circular data. *S. African Statistical Journal* **27**, 23–49.

MENDOZA, C.E. (1986). Smoothing unit vector fields. *Math. Geol.*, **18**, 307–322.

MIELKE, P.W. (1973). Another family of distributions for describing and analyzing precipitation data. *Journal of Applied Meteorology*, **10**, 275–280.

MYERS, D.E. (1982). Matrix formulation of co-kriging. *Mathematical Geology*, **14**, 249–257.

PEARSON, E.S. and HARTLEY, H.O. (1962). *Biometrika Tables for Statisticians: vol. 1*. Cambridge University Press, Cambridge.

RICHARDSON, C.W. (1981). Stochastic simulation of daily precipitation, temperature and solar radiation. *Water Resources Research*, **17**, 182–190.

RIPLEY, B.D. (1981). *Spatial Statistics*. Wiley, New York.

ROLDAN, J. and WOOLHISER, D.A. (1982). Stochastic daily precipitation models: (1) A comparison of occurrence processes. *Water Resources Research*, **18**, 1451–1459.

SCHULZE B.R. (1947). The climates of South Africa according to the classification of Köppen and Thornthwaite. *S. Afr. Geogr. J.*, **29**, 32–42.

SCHULZE B.R. (1958). The climate of South Africa according to Thornthwaite's rational classification. *S. Afr. Geogr. J.*, **40**, 31–53.

SCHULZE, R.E. (1976). On the application of trend surfaces of precipitation to mountainous areas. *Water SA*, **2**, 110–118.

SCHUMANN, T.E.W. and HOFMEYR W.R. (1938). The partition of a region into rainfall districts with special reference to South Africa. *Jl. R. Met. Soc.*, **64**, 482–488.

SCHUMANN, T.E.W. and THOMPSON W.R. (1934). A study of South African secular variations and agricultural aspects. *Pretoria University Occasional Series* **1**.

SEDUPANE, S.M. (1992). Modelling cross-covariance between rainfall and altitude in the western Cape, Transvaal and Natal. Unpublished Honours project report. Dept. Statistical Sciences, University of Cape Town.

SEED, A.W. (1987). Techniques for mapping rainfall. Unpublished M.Sc.Eng. thesis. Dept. Agricultural Engineering, University of Natal, Pietermaritzburg.

SILVERMAN, B.W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *J.R. Statist. Soc. B*, **47**, 1–52.

SKIDMORE, A.K. (1989). A comparison of techniques for calculating gradient and aspect from a gridded digital elevation model. *Int. J. Geographical Information Systems*, **3**, 323–334.

SMITH, R.E. and SCHREIBER, H.A. (1973). Point processes of seasonal thunderstorm rainfall, I, Distributions of rainfall events. *Water Resources Research*, **9**, 871–884.

SPREEN, W.C. (1947). A determination of the effect of topography upon precipitation. *Trans. Amer. Geophys. Union,* **28**, 285–290.

SRIVASTAVA, R.M. (1988). A non-ergodic framework for variograms and covariance functions. Technical Report no. 114, Dept. Statistics and Dept. Applied Earth Sciences, Stanford University, California. 113p.

STEIN, A. and CORSTEN, L.C.A. (1991). Universal kriging and cokriging as a regression procedure. *Biometrics*, **47**, 575–587.

STERN, R.D. and COE, R. (1983). A model fitting analysis of daily rainfall data. *Journal of the Royal Statistical Society*, A **147**, 1–34.

TABIOS, G.Q. and SALAS, J.D. (1985). A comparative analysis of techniques for spatial interpolation of precipitation. *Water Resources Bulletin,* **21**, 365–380.

TODOROVIC, P. and WOOLHISER, D.A. (1975). A stochastic model of n-day precipitation. *Journal of Applied Meteorology*, **14**, 17–24.

TYSON, P.D. (1986). *Climatic Change and Variability in Southern Africa.* Oxford University Press, Cape Town.

UPTON, G.J.G. and FINGLETON, B. (1989). *Spatial Data Analysis by Example, vol.2: Categorical and Directional Data.* Wiley, New York.

WAHBA, G. (1990). Letter to the editor. *American Statistician,* **44**, 255–256.

WAHBA, G. and WENDELBERGER, J. (1980). Some new mathematical methods for variational objective analysis using splines and cross-validation. *Monthly Weather Review*, **108**, 36–57.

WATSON, G.S. (1971). Trend-surface analysis. *Journal of the International Association for Mathematical Geology*, **3**, 215–226.

WATSON, G.S. (1972). Trend surface analysis and spatial correlation. *Geological Society of America, Special Paper*, **146**, 39–46.

WATSON, G.S. (1984) Smoothing and interpolation by kriging and with splines. *Math. Geol.*, **16**, 601–615.

WATSON, G.S. (1985). Interpolation and smoothing of directed and undirected line data. In *Multivariate Analysis - VI*. P.R. Krishnaiah, ed. Elsevier Science Publishers B.V. 613–625.

WEISS, L.L. (1964). Sequences of wet or dry days described by a Markov chain probability model. *Monthly Water Review*, **92**, 169–176.

WELLINGTON J.H. (1955). *Southern Africa: a Geographical Study. vol. 1. Physical Geography.* Cambridge University Press.

WHITMORE, J.S. (1968). The relationship between mean annual rainfall and locality and site factors. *S. Afr. Jour. Sci.*, **64**, 423–427.

WOLFSON, N. (1975). Topographical effects on standard normals of rainfall over Israel. *Weather*, **30**, 138–143.

WOOLHISER, D.A. (1992). Modelling daily precipitation - progress and problems. In *Statistics in the Environmental and Earth Sciences*, A.T. Walden and P. Guttorp, eds. Edward Arnold, New York, 71–89.

WOOLHISER, D.A. and PEGRAM, G.G.S. (1979). Maximum likelihood estimation of Fourier coefficients to describe seasonal variations of parameters in stochastic daily precipitation models. *Journal of Applied Meteorology*, **18**, 34–42.

WOOLHISER, D.A. and ROLDAN, J. (1982). Stochastic daily precipitation models: (2) A comparison of distributions of amounts. *Water Resources Research*, **18**, 1461–1468.

YOUNG, D.S. (1987). Random vectors and spatial analysis by geostatistics for geotechnical applications. *Math. Geol.*, **19**, 467–479.

ZIMMERMAN, D.L and ZIMMERMAN, M.B. (1991). A comparison of spatial semivariogram estimators and corresponding ordinary kriging predictors. *Technometrics,* **33**, 77–91.

ZUCCHINI, W. and ADAMSON, P.T. (1984a). The occurrence and severity of droughts in South Africa. *WRC Report* No. 91/1/84, Water Research Commission, Pretoria.

ZUCCHINI, W. and ADAMSON, P.T. (1984b). The occurrence and severity of droughts in South Africa : Appendix 6. *WRC Report* No. 91/1/84(A), Water Research Commission, Pretoria.

ZUCCHINI, W., ADAMSON, P. and McNEILL, L. (1992). A model of southern African rainfall. *S. Afr. Jnl. Sci.*, **88**, 103–109.

ZUCCHINI, W. and SCHMIDT, M. (1990). Die wichtigsten stetigen verteilungen. *Herausgegeben vom Institut für Statistik und Ökonometrie der Universität Göttingen*, Einband: Wolfram Bach.

Appendix A

Maximum Likelihood Estimates of the Weibull Distribution

The probability density function of the ordinary non-seasonal Weibull distribution is given by

$$f(x) = \left(\frac{\beta}{\alpha}\right)\left(\frac{x}{\alpha}\right)^{\beta-1} e^{-(x/\alpha)^{\beta}}, \qquad x \geq 0,$$

with mean

$$\mu = \alpha\Gamma(1 + 1/\beta),$$

variance

$$\sigma^2 = \alpha^2\Gamma(1 + 2/\beta) - \Gamma(1 + 1/\beta)^2$$

and coefficient of variation

$$\gamma = \frac{\sigma}{\mu} = \left(\frac{\Gamma(1 + 2/\beta)}{\Gamma(1 + 1/\beta)^2 - 1}\right)^{1/2}.$$

If we allow the mean of the Weibull distribution to vary seasonally and model this seasonal mean by its truncated Fourier series representation, that is we define

$$\mu(t, L) = \sum_{i=1}^{L} \theta_i\varphi_i(t), \qquad t = 1, 2, \ldots, NT, \qquad L \leq NT,$$

where $\varphi_i(t)$ is defined as in Chapter 3, then the seasonal probability density function of the Weibull distribution is given by

$$f(x) = \left(\frac{\beta}{\alpha(t)}\right)\left(\frac{x_t}{\alpha(t)}\right)^{\beta-1} \exp(-(x_t/\alpha(t))^{\beta}), \qquad x_t \geq 0,$$

where

$$\begin{aligned} \mu(t,L) &= \sum_{i=1}^{L} \theta_i \varphi_i(t), \\ \beta &= \beta(\gamma) \qquad \text{is independent of t} \\ \alpha(t) &= \frac{\sum_{i=1}^{L} \theta_i \varphi_i(t)}{\Gamma(1+1/\beta)}. \end{aligned}$$

The likelihood function of observation x_t is then

$$\begin{aligned} L(\psi) &= L(\theta_i, \beta; x) = \prod_{t=1}^{T} f(x_t) = \left[\beta\Gamma(1+1/\beta)^{\beta}\right]^T \prod_{t=1}^{T} \frac{x_t^{\beta-1}}{\sum_{i=1}^{L} \theta_i \varphi_i(t)^{\beta}} \\ &\quad \exp\left[-\sum_{t=1}^{T}\left(\frac{\Gamma(1+1/\beta)x_t}{\sum_{i=1}^{L}\theta_i\varphi_i(t)}\right)^{\beta}\right] \end{aligned}$$

and the log-likelihood is given by

$$\begin{aligned} \ell(\psi) &= T(\log\beta + \beta\log(\Gamma(1+1/\beta))) + \sum_{t=1}^{T}\left\{(\beta-1)\log x_t - \beta\log(\sum_{i=1}^{L}\theta_i\varphi_i(t))\right\} \\ &\quad - \Gamma(1+1/\beta)^{\beta}\sum_{t=1}^{T}\left(\frac{x_t}{\sum_{i=1}^{L}\theta_i\varphi_i(t)}\right)^{\beta} \end{aligned}$$

Maximum likelihood estimates can be obtained by minimising $\ell(\psi)$ and this is achieved by setting its first partial derivatives with respect to the parameters, θ_i, and β, equal to zero. The first partial derivatives are given by

$$\frac{\partial\ell(\psi)}{\partial\theta_j} = -\beta\sum_{t=1}^{T}\frac{\varphi_j(t)}{\sum_{i=1}^{L}\theta_i\varphi_i(t)} + \Gamma(1+1/\beta)^{\beta}\beta\sum_{t=1}^{T}\frac{x_t^{\beta}\varphi_j(t)}{(\sum_{i=1}^{L}\theta_i\varphi_i(t))^{\beta+1}}$$

$$\frac{\partial \ell(\psi)}{\partial \beta} = \frac{T}{\beta} + T\left[\log(\Gamma(1+1/\beta)) - \frac{\Psi(1+1/\beta)}{\beta}\right] + \sum_{t=1}^{T}\{\log x_t - \log(\sum_{i=1}^{L}\theta_i\varphi_i(t))\} - \Gamma(1+1/\beta)^{\beta}\sum_{t=1}^{T}\left(\frac{x_t}{\sum_{i=1}^{L}\theta_i\varphi_i(t)}\right)^{\beta} \left\{\log\left(\frac{x_t}{\sum_{i=1}^{L}\theta_i\varphi_i(t)}\right) - \frac{\Psi(1+1/\beta)}{\beta} + \log(\Gamma(1+1/\beta))\right\}$$

These equations cannot be solved explicitly and therefore the Newton-Raphson iterative method is used to solve them. For this we require the second partial derivatives. These are

$$\frac{\partial^2 \ell(\psi)}{\partial \theta_j \partial \theta_k} = \beta\sum_{t=1}^{T}\frac{\varphi_j(t)\varphi_k(t)}{\left(\sum_{i=1}^{L}\theta_i\varphi_i(t)\right)^2} - \beta(\beta+1)\Gamma(1+1/\beta)^{\beta} \sum_{t=1}^{T}\frac{x_t^{\beta}\varphi_j(t)\varphi_k(t)}{\left(\sum_{i=1}^{L}\theta_i\varphi_i(t)\right)^{\beta+2}}$$

$$\frac{\partial^2 \ell(\psi)}{\partial \beta \partial \beta} = -\frac{T}{\beta^2} - \frac{T\Psi'(1+1/\beta)}{\beta^3} - \Gamma(1+1/\beta)^{\beta}\sum_{t=1}^{T}\left(\frac{x_t}{\sum_{i=1}^{L}\theta_i\varphi_i(t)}\right)^{\beta} \left\{\left[\log\left(\frac{x_t}{\sum_{i=1}^{L}\theta_i\varphi_i(t)}\right) - \frac{\Psi(1+1/\beta)}{\beta} + \log(\Gamma(1+1/\beta))\right]^2 + \frac{\Psi'(1+1/\beta)}{\beta^3}\right\}$$

$$\frac{\partial^2 \ell(\psi)}{\partial \theta_j \partial \beta} = -\sum_{t=1}^{T}\frac{\varphi_j(t)}{\sum_{i=1}^{L}\theta_i\varphi_i(t)} + \Gamma(1+1/\beta)^{\beta}\sum_{t=1}^{T}\frac{x_t^{\beta}\varphi_j(t)}{\left(\sum_{i=1}^{L}\theta_i\varphi_i(t)\right)^{\beta+1}} \{1 - \Psi(1+1/\beta) + \beta[\log(\Gamma(1+1/\beta)) + \log x_t - \log\left(\sum_{i=1}^{L}\theta_i\varphi_i(t)\right)]\}.$$

An algorithm for parameter estimation as well as algorithms to compute the gamma function $\Gamma(\alpha)$, the digamma function, $\Psi(\alpha)$, and the trigamma

function, $\Psi'(\alpha)$ are given at the end of this appendix.

A.1 Properties of MLE of the Weibull distribution

The Weibull distribution was fitted to model rainfall depth and maximum likelihood estimates of the parameters were obtained for two test stations. It was seen that, although the parameter estimates for the mean rainfall depth were close to those obtained by the method of moments, the coefficient of variation differed significantly (Table A.1).

	Durban	Elsenburg
Moment estimate	1.633	1.266
Maximum likelihood	1.013	0.8549

Table A.1: Estimates of coefficient of variation

According to Johnson and Kotz (1970):

1. It is not generally true that maximum likelihood estimates are unbiased, and in particular the maximum likelihood estimate of the shape parameter (β) of the Weibull distribution is a biased estimate. The coefficient of variation, CV, is given by

$$CV = \frac{\sqrt{\Gamma(2/\beta+1) - \Gamma(1/\beta+1)^2}}{\Gamma(1/\beta+1)}.$$

   That is, the coefficient of variation of the Weibull distribution is a function of the shape parameter alone and therefore if the estimate of β is biased, so is the estimate of CV.

2. If the maximum likelihood estimates are 'regular', in the sense of having the usual asymptotic distribution, then the asymptotic variance-

covariance matrix for the estimators is given by the inverse of the matrix with entries

$$\left\{-\frac{\partial^2 \ell(\psi)}{\partial \xi_i \partial \xi_j}\right\}$$

where ξ_i represents the model parameters. Maximum likelihood estimators are 'regular' only for $\beta > 2$. In our case, for several rainfall stations we have that $\beta < 2$ therefore we do not obtain correct measures of the standard errors of the estimates.

Alternative distributions, such as the gamma distribution used by Stern and Coe (1984), were considered for modelling rainfall depth. The coefficient of variation for the gamma distribution is also dependent only on a shape parameter whose maximum likelihood estimate is biased. The bias can be estimated if the mean rainfall depth is assumed to be constant. In our situation, the mean rainfall depth is allowed to vary seasonally so this assumption is violated and the extent of the bias is unknown, therefore one cannot objectively correct for bias. Also, the coefficient of variation essentially determines the variability of rainfall, which is a property that it is important for the model to preserve, especially in southern Africa where the variability in the rainfall is a major feature of our climate. It was thus decided to abandon the classical approach to solving this problem and to develop alternative methods of estimating the required standard errors discussed in the Chapter 4.

A.2 Algorithms

A.2.1 Algorithm to compute parameter estimates

Step 1: Estimate initial $\widehat{\theta}_i$, $\quad i = 1, 2, \ldots, L$ by

$$\widehat{\theta}_i = \begin{cases} \bar{x} & \text{if} \quad \text{i} = 0 \\ 0 & \text{if} \quad \text{i} = 2,3, \ldots, \text{L}, \end{cases} \quad \text{and}$$

$$\widehat{\beta} = s_x/\bar{x},$$

where $\bar{x}$ and s_x is the mean and the standard deviation of the observations x_t, $t = 1, 2, \ldots, T$, respectively.

Step 2: Compute $f^{(k)}$ and $F^{(k)}$, where $f^{(k)}$ is the vector of first derivatives and $F^{(k)}$ is the matrix of second partial derivatives, computed at the kth iteration.

Step 3: Compute the vector $\delta^{(k)}$ which is the solution to the system of NP linear equations

$$F^{(k)}\delta^{(k)} = f^{(k)},$$

where NP represents the number of parameters.

Step 4: Set $\beta^{(k+1)} = \beta^{(k)} - \delta^{(k)}$, where $\beta^{(k)}$ contains the parameter estimates at the kth iteration.

Step 5: Test for convergence, for example, if the elements of $f^{(k)}$ are sufficiently close to zero. If the convergence criterion is met then stop, otherwise increase k by 1 and return to step 4.

A.2.2 Algorithm to compute $\Gamma(\alpha)$

This and all following algorithms were obtained from Zucchini and Schmidt (1990). The gamma function is given by:

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}e^{-t}dt, \qquad \alpha \neq 0, -1, -2, \ldots$$

If $\alpha < 10$ the following recurrence relationship is applied in order to increase the argument of the gamma function to a number greater than or equal to 10:

$$\Gamma(\alpha + 1) = \alpha\Gamma(\alpha).$$

Step 1:	Input	α
Step 2:	Set	$A = \alpha$
		$G = 1$
Step 3:	Test if	$A \geq 10$ then go to Step 5
Step 4:	Set	$G = G * A$
		$A = A + 1$
	Go to	Step 3
Step 5:	Set	$T = (1 + (0.0833333 + 0.00347222 - 0.002681327/A)/A)/A$
		$\text{Gamma} = \exp(-A + (A - 0.5) * \log(A) + 0.918939) * T *$
		A/G
Step 6:	Output	Gamma

A.2.3 Algorithm to compute $\Psi(\alpha)$

The digamma function is given by:

$$\Psi(\alpha) = \frac{\mathrm{d} \ln \Gamma(\alpha)}{\mathrm{d}\alpha}, \qquad \alpha \neq 0, -1, -2, \ldots$$

If $\alpha < 4$ the following recurrence relationship is applied in order to increase the argument of the gamma function to a number greater than or equal to 4:

$$\Psi(\alpha + 1) = \Psi(\alpha) + 1/\alpha.$$

Step 1:	Input	α
Step 2:	Set	$A = \alpha$
		$P = 0$
Step 3:	Test if	$A \geq 4$ then go to Step 5
Step 4:	Set	$P = P - 1/A$
		$A = A + 1$
	Go to	Step 3
Step 5:	Set	$T = 1/(A * A)$
		$U = T * (0.08333333 - T * (0.008333333 - T * 0.003968254))$
		Digamma $= P + \log(A) - 0.5/A - U$
Step 6:	Output	Digamma

A.2.4 Algorithm to compute $\Psi'(\alpha)$

The trigamma function is given by:

$$\Psi'(\alpha) = \frac{\mathrm{d}^2 \ln \Gamma(\alpha)}{\mathrm{d}\alpha^2}, \qquad \alpha \neq 0, -1, -2, \ldots$$

If $\alpha < 4$ the following recurrence relationship is applied in order to increase the argument of the gamma function to a number greater than or equal to 4:

$$\Psi'(\alpha + 1) = \Psi'(\alpha) + 1/\alpha^2.$$

Step 1:	Input	α
Step 2:	Set	$A = \alpha$
		$P = 0$
Step 3:	Test if	$A \geq 4$ then go to Step 5
Step 4:	Set	$P = P + 1/A * A$
		$A = A + 1$
	Go to	Step 3
Step 5:	Set	$T = 1/(A * A)$
		$U = T * (0.1666667 - T * (0.03333333 - T * 0.02380953))$
		$\text{Trigamma} = P + 1/A + 0.5 * T + U/A$
Step 6:	Output	Trigamma

Appendix B

Kriging

B.1 Trend Removal by Kriging

The usual polynomial models of trend are not suitable for modelling topography or rainfall except over fairly small areas due to the complexity of the surfaces typically encountered. Also, methods of smoothing based on simple moving averages, such as are commonly used in time series analysis, are unsuitable for irregularly-spaced data. An alternative possibility is to re-write the general kriging model

$$v_i = \tau_i + \eta_i + \epsilon_i$$

as

$$v_i = \mu + \tau_i + \eta_i + \epsilon_i$$

where μ is the overall mean, and τ represents trend, considered now as a stochastic component with zero mean similar to η, but on a larger scale. One can then use kriging as a filter to separate the high and low frequency components τ and η. Thus we estimate the trend as

$$\widehat{\mu + \tau} = \sum_{i=1}^{n} w_i v_i$$

where the w_i are given by:

$$\begin{pmatrix} \mathbf{K} & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \mathbf{w} \\ -\lambda \end{pmatrix} = \begin{pmatrix} \mathbf{c} \\ 1 \end{pmatrix}$$

where $k_{ij} = \mathrm{cov}(v_i, v_j)$, and $c_i = \mathrm{cov}(v_i, \tau_0) = \mathrm{cov}(\tau_i, \tau_0)$. This is commonly known as *factorial* kriging (Matheron, 1982), and is similar to ordinary kriging except that the covariance must be decomposed into components corresponding to the high and low frequency components. While the separation of τ and η is to some extent subjective since the terms small-scale and large-scale are relative, there is often a natural distinction apparent in the empirical semi-variogram or covariance function. The semi-variograms of the amplitude parameters discussed in Section 5.3 show such a separation, with a levelling-off at a range somewhere between 10 and 40 kilometres, and this was used as a basis for the models described in table 5.1. Having estimated the trend component at each data point, using the equations above, we can then subtract the trend from the original values to get de-trended data. The semi-variograms of the de-trended data showed that the long-range trend effects had indeed been eliminated, but also showed a spike at a lag distance of approximately four kilometres suggesting a spurious negative correlation induced by the de-trending process. This phenomenon is well known in the time-series field (see for example Diggle, 1990, section 2.6).

B.2 Circular Kriging Equations

Proof of Equation 5.9

We wish to find weights w_i to minimize $E[1 - \cos(\hat{\theta}_0 - \theta_0)]$ where $(R_0, \hat{\theta}_0)$ is the vector $\sum_{i=1}^{n} w_i \mathbf{e}_i$ written in polar form.

If we write $\mathbf{e}_i = (x_i, y_i)'$ so that $\theta_i = \arccos(x_i) = \arcsin(y_i)$ and $\mathbf{e}_0 = (x_0, y_0)'$

with $\theta_0 = \arccos(x_0) = \arcsin(y_0)$, then we have

$$\sin\hat{\theta}_0 = \sum_{i=1}^{n} w_i y_i / R_0 = \sum_{i=1}^{n} w_i \sin\theta_i / R_0$$

and

$$\cos\hat{\theta}_0 = \sum_{i=1}^{n} w_i x_i / R_0 = \sum_{i=1}^{n} w_i \cos\theta_i / R_0$$

where R_0 is the length of the vector $\sum w_i \mathbf{e}_i$ so that

$$\begin{aligned} R_0^2 &= (\sum_{i=1}^{n} w_i x_i)^2 + (\sum_{i=1}^{n} w_i y_i)^2 \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j \cos\theta_i \cos\theta_j + \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j \sin\theta_i \sin\theta_j \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j \cos(\theta_i - \theta_j) \\ &= \mathbf{w}'Q\mathbf{w} \end{aligned}$$

where $q_{ij} = \cos(\theta_i - \theta_j)$.

Now

$$\begin{aligned} \cos(\hat{\theta}_0 - \theta_0) &= \cos\hat{\theta}_0 \cos\theta_0 + \sin\hat{\theta}_0 \sin\theta_0 \\ &= 1/R_0 \left(\sum_{i=1}^{n} w_i(\cos\theta_i \cos\theta_0 + \sin\theta_i \sin\theta_0) \right) \\ &= \mathbf{w}'\mathbf{c}/\sqrt{(\mathbf{w}'Q\mathbf{w})} \end{aligned}$$

where $c_i = \cos(\theta_i - \theta_0)$

Thus in order to minimize $E[1 - \cos(\hat{\theta}_0 - \theta_0)]$ we need to find $w_1, ...w_n$ to maximize

$$E[\mathbf{w}'\mathbf{c}/\sqrt{\mathbf{w}'Q\mathbf{w}}]$$

It is clear from the formula above that the solution will be unique only up to a constant multiplier; that is, if $\mathbf{w}$ is a solution, then so is $l\mathbf{w}$ for any non-zero constant l.

If we use a first order Taylor series expansion of $E[\mathbf{w}'\mathbf{c}/\sqrt{\mathbf{w}'Q\mathbf{w}}]$, so that we approximate it by $\mathbf{w}'\mathbf{s}/\sqrt{\mathbf{w}'K\mathbf{w}}$, where $k_{ij} = E[q_{ij}]$ and $s_i = E[c_i]$, and

use the uniqueness constraint $E[\mathbf{w}'Q\mathbf{w}] = \mathbf{w}'K\mathbf{w} = 1$ then we can use the Lagrange multiplier approach to find the optimal values of the w_i. With the chosen uniqueness constraint, the function to be maximized becomes simply $\mathbf{w}'\mathbf{s}$, so if we set

$$G = \mathbf{w}'\mathbf{s} - \lambda(\mathbf{w}'K\mathbf{w} - 1)$$

then

$$\frac{\partial G}{\partial \mathbf{w}} = \mathbf{s} - 2\lambda K\mathbf{w}$$

and

$$\frac{\partial G}{\partial \lambda} = \mathbf{w}'K\mathbf{w} - 1$$

which leads to the solution

$$\mathbf{w} = \frac{K^{-1}\mathbf{s}}{\sqrt{\mathbf{s}'K^{-1}\mathbf{s}}} = \frac{K^{-1}\mathbf{s}}{r}$$

where $r = \sqrt{\mathbf{s}'K^{-1}\mathbf{s}}$ is a scalar normalizing constant.

Proof of Equation 5.10

$$\begin{aligned}
E[\cos(\vartheta_i - \vartheta_j)] &= E[\cos(\theta_i + \epsilon_i - \theta_j - \epsilon_j)] \\
&= E[\cos(\theta_i - \theta_j)\cos(\epsilon_i - \epsilon_j) - \sin(\theta_i - \theta_j)\sin(\epsilon_i - \epsilon_j)] \\
&= E[\cos(\theta_i - \theta_j)(\cos\epsilon_i \cos\epsilon_j + \sin\epsilon_i \sin\epsilon_j) \\
&\quad - \sin(\theta_i - \theta_j)(\sin\epsilon_i \cos\epsilon_j - \cos\epsilon_i \sin\epsilon_j) \\
&= E[\cos(\theta_i - \theta_j)\cos\epsilon_i \cos\epsilon_j] \\
&= E[\cos(\theta_i - \theta_j)]E[\cos\epsilon_i]E[\cos\epsilon_j]
\end{aligned}$$

Appendix C

Programs

The list below gives brief details of the main programs used in this project. The programs have been written in ANSI 77 FORTRAN and conform to the full ANSI standard.

DRMODEL Fits model parameters at selected sites. See Chapter 3.

DRBOOT Generates 100 parametric bootstrap samples, using the fitted model parameters at each site, and uses these to estimate the variances of the parameters. See Chapter 4.

SVGMAMP Calculates the unadjusted and adjusted semi-variograms for all the amplitude parameters and the coefficient of variation. See Chapter 5.

SVGMCIR Calculates the unadjusted and adjusted semi-variograms for the phase parameters. See Chapter 5.

ORTHOALT Calculates the orthogonal functions of altitude. See Chapter 5.

KGXDRIFT Carries out the kriging estimation of the amplitude parameters (and CV) using an 'external drift' model which incorporates the orthogonal functions of altitude. See Chapter 5.

KRIGCIRC Carries out the kriging estimation of the phase parameters. See Chapter 5.

DRGEN Generates an n-year sequence of simulated daily data for a given station or grid point. See Chapter 6.

These programs are available from the Computing Centre For Water Research at the following address:

Computing Centre For Water Research

c/o University of Natal

P O Box 375

Pietermaritzburg

3200

Tel. (0331) 63320 ext. 177/178

Fax (0331) 61896.