Report No. 1

THE OCCURRENCE AND SEVERITY OF DROUGHTS IN SOUTH AFRICA

by

W. ZUCCHINI and P.T. ADAMSON

Report describing a research project carried out by the Department of Civil Engineering, University of Stellenbosch, under contract to the Water Research Commission.

Department of Civil Engineering
University of Stellenbosch
December 1984

PREFACE

This report is part of a larger study on the occurrence and severity of drought in South Africa. As various aspects of this study may be of interest to researchers and practitioners who are not specifically concerned with drought, it was decided to separate the results of the research into three self-contained reports, this being the main one. Naturally this has led to some repetition but it is hoped that this disadvantage is outweighed by making the methods and results more accessible to a wider audience.

ACKNOWLEDGMENTS

We particularly wish to thank Professor L. Hiemstra of the Department of Civil Engineering, University of Stellenbosch who motivated this research project, organised funding and provided guidance and support throughout its course.

We owe thanks to Professor H. Linhart of the Institut für Statistik und Ökonometrie, University of Göttingen who collaborated on the problem relating to model selection, and Dr. D. Martin and Dr. T. Stewart of the C.S.I.R., who helped us overcome some of the mathematical difficulties.

We wish to thank the Division of Hydrology of the Department of Environment Affairs, and in particular Mr. W. Alexander and Mr. S. van Biljon for their willing cooperation at all stages of this project and for allowing us to use the Division's facilities.

Thanks are also due to the Weather Bureau of the Department of Transport for their help in supplying us with the data and allowing us to use their computer to rearrange these into a form which was more suitable for our purposes.

We wish to thank the Steering Committee for their guidance and support.

Finally we wish to thank Mr. J.P. Trichard of the Computer Division, Department of Hydrology for his unstinting co-operation in many aspects of the computing, Miss G. Maschile of the Division of Hydrology and Mrs. C. Adamson for their expert drafting, and last but by no means least Mrs. I. Cousins for her amazingly patient and immaculate typing.

PREFACE TO REPRINT

A number of corrections have been made to the original report. The page numbering and structure of the original report remain unchanged.

March 1987

CONTENTS

FIGURES

1. INTRODUCTION

Drought and mismanagement of available water resources are major factors which reduce production of many water related enterprises to below potential. It is not surprising therefore that an abundance of drought studies are reported in the literature. Palmer and Denny (1971) list no less than 3150 selected references on the subject. These investigations vary from the purely descriptive to mathematical modelling and simulation. Definitions of drought and classifications of drought severity abound, as do attempts to forecast the occurrence of future droughts. On the other hand, methodology to tackle the problem of assessing drought risk in a manner which is sufficiently general to be universally applicable and simultaneously simple enough to be suitable for implementation on a large scale is not available. It was with the object of developing such methodology that the research described in this report was initiated. During the course of the project, and as the complexity of the problem came to be appreciated, it became increasingly evident that in order to make progress we would have to restrict our attention to some selected aspects of the problem. Much more research needs to be carried out on the subject.

Droughts are usually classified as either meteorological, agricultural or hydrological, depending on the variables under investigation. The important variables in meteorological drought are rainfall, snowfall, wind speed, wind direction, humidity and temperature. In agricultural drought soil moisture content and evapotranspiration are the major variables. Hydrological drought analyses are mostly concerned with water in rivers, lakes, reservoirs and underground water storage spaces.

The variables which indicate the presence and severity of agricultural and hydrological drought derive from those associated with meteorological drought, or at least are directly influenced by them. The single most important of these, particularly in South Africa, is rainfall. It often correlates quite well with humidity and temperature as well as wind speed and direction, but the availability of rainfall data which have been collected in more places and for longer periods than data for any of the other variables is a more important factor. It is therefore not surprising that rainfall forms the basis of most drought investigatio

Underlying our notion of drought is the assumption that the water-related activities in a region should be in harmony with the amount of water which is "normally" available for those specific activities. Any significant deviation from normal conditions is usually harmful, and should the deviation be to the deficit side then a drought occurs. In many situations "normal" is taken to be the mean amount of water available. This notion of drought is, however, rather inadequate. In reality some activities require less water than the mean amount available, and some require more. It is the occurrence of negative deviations from the required levels rather than from the mean which constitute droughts. In other words a drought occurs when there is less water available than is needed, and not when there is less than is expected.

In any specific drought investigation it is therefore important to identify the water-related variable (or variables) which is relevant to the activity under consideration and to establish the water requirements, which we will call the desired level, for the activity to function effectively. Deviations of the available variables from the desired level can then be traced. The deviations below the desired level are characterised by their

duration, depth (or intensity) and time of origin. In addition the drought-affected region usually needs to be identified.

Ideally all four properties of the drought, i.e. its timing, duration, intensity and areal extent should be studied simultaneously. Furthermore one should take account of the fact that the drought process and the desired level of water are both dynamic processes which usually follow the seasonal cycle and are stochastic rather than deterministic in nature. To adequately describe all these facets of drought in a single model is extremely difficult. For example no applicable model to describe the process of daily rainfall at several locations simultaneously has yet been proposed. It is therefore necessary to make simplifying assumptions or alternatively to study the different properties separately. The first simplification is usually to limit the description of drought to a single measure, for example the drought duration at a point. As there is a strong correlation between drought duration and severity (Hully 1980) this scheme is less restrictive than it may appear superficially. In other studies the complications due to seasonality are avoided by selecting as time unit either the year or the growing season of the crop under investigation.

The most widely applied classification system is the Palmer drought index which is a function of accumulated weighted differences between actual and required precipitation; the latter being determined by evapotranspiration, moisture recharge, runoff and antecedent rainfall conditions. It is intended primarily as a measure of wetness/dryness in agriculture and is a good illustration of the large number of variables and of the complexity of their interrelationships which need to be taken into account even if one

wishes to describe just agricultural drought in detail. Simpler classification systems have been proposed, but unless adequately long historical records of the variables which make up the index are available, it is not possible to meaningfully assess the risk of future drought events as measured by the index in question. Consequently most drought studies stop after classifying droughts into severity classes at various points in a region and then drawing contour lines of equal severity to assess the areal extent of a particular drought.

Herbst et al (1966) define a general drought model based on rainfall. This method, which may provide a viable alternative to the Palmer drought index for identifying and comparing droughts, is attractive because only rainfall data are required to classify a drought. It takes into consideration the average requirements which must be met each month before a drought condition is deemed to exist. However this methodology appears too general to have operational value. (The results of this report do not exclude the application of such drought models; in fact these results can be used to construct such models quite easily.

We have emphasised that drought investigations should be specific to the water-related activity. Once we know the water requirements associated with a particular activity there is no difficulty in defining a drought. If sufficient historical data are available at the point where the activity takes place then we can also begin to construct models to assess the risk of drought. In any case it is clear the search for any single all-embracing definition of drought is futile. Different water users have different needs and what may be a drought for one user need not be a drought for another. On the other hand it is obviously unrealistic in a single study to develop a separate

methodology for each conceivable application and therefore some compromise is necessary.

At the initial stages of this project we decided that a reasonable way to solve this dilemma was to study families of drought indices rather than a single drought index.

The family of drought models was to be based on daily rainfall (weekly, monthly or even annual rainfall can be used) because this is the only relevant variable for which historical records of sufficient length and at sufficiently many places are available for a large-scale study of drought based on statistical models.

A suitable family was identified and it was our intention that each user would make use of that index within the family which came closest to meeting his specific requirements. The family we chose is quite simple and is characterised by a single parameter, namely the half-life of an exponentially decaying function. This function is then used to describe the decay in the "benefit" associated with a unit quantity of rainfall as the time from the rainfall event increases. For example the "benefit" associated with 20 mm of rainfall does not vanish the moment that it stops raining. One would wait at least a day or two before declaring a state of drought. In other words the effects of rainfall persist after the event and our family of models assume that these effects decrease *exponentially* with time. The rate at which this decrease takes place will depend on the effect in question and we therefore allow the user to decide which rate may be most appropriate for his purpose. In most agricultural applications (with the possible exception of sugar cultivation) it is important that rain should occur quite frequently, particularly at the flowering stage of the plant. Here the benefits associated with

any single rainfall event decay quite rapidly. On the other hand, for those who are concerned with reservoir levels, the precise timing of rainfall is less important and the benefits associated with a rainfall event persist for a longer period. For such applications a model with a long half-life (slow rate of decay) would be suitable.

Having decided on a family of drought indices we then set about looking for suitable statistical models to describe them. This task turned out to be more difficult than we had hoped. It is easy enough to construct a model for a fixed member of the family of drought indices, but we were unable to find any single model which could satisfactorily cover an adequate range of drought indices even for a given rainfall record. A further complication is that different models are sometimes required for different regions. It would therefore be necessary to fit a number of models, some of which are necessarily complex, to each historical record individually. The final product of this research would then consist of a list of models and their estimated parameters for each rainfall station. Since it was one of the main objectives of the project to provide methods which were simple enough to be attractive to practitioners (and not only to statisticians), such an approach was considered unsatisfactory. On the other hand any further simplification in the definition of the family of drought indices was out of the question. The only reasonable solution to this problem was to model the rainfall process itself and thereby indirectly provide a means of modelling the complete family of drought indices.

With the advantage of hindsight it is now clear to us that this is what we should have done in the first place rather than concentrating on definitions of drought. Had we persisted with our original approach users would have been obliged to use *our* drought indices, even if they were free

to select the half-life. By providing a model for the rainfall process, users are now entirely free to select *any* rainfall-based index whatsoever. The model provides a means to assess the risk of drought as defined by any such index, including the family discussed above.

There is a good deal of useful literature on the construction of daily rainfall models, but two problems required further research. The first concerned the question of model selection, in particular selecting the number of parameters which should be used. The second, which apparently only arises in arid and semi-arid regions, involved the search for alternative estimation procedures for one of the two components of the basic model. After a number of false starts the solution to this problem turned out to be both simple and satisfying - one must estimate the logits (transformed probabilities) rather than the probabilities themselves. In a recent paper "A Model fitting Analysis of Daily Rainfall Data" read before the Royal Statistical Society and published in the Society Journal, Stern and Coe (1984) proposed the same approach, together with one extension. This paper is followed by a comprehensive discussion by 14 prominent British statisticians and was favourably received.

The model was validated using six rainfall stations selected from different climatic regions of the country and was found to fit remarkably well. It was then fitted to 2550 stations selected on the basis of the length and quality of their historical records and also in such a way as to provide an adequate coverage of the country as a whole. The estimated parameters for these stations are given in the report.

An entire spectrum of properties of the daily rainfall process at any of these stations is condensed into a

relatively small number of parameters and with the aid of a micro-computer one can unlock a wealth of information relating to occurrence of rainfall. For example, all the monthly and annual properties of rainfall including means, variances and in fact the complete probability distributions, can be computed. The probability of dry spells of any particular length and starting at any particular time are easy to compute. The distribution of wet and dry days for any period, the probabilities of getting any desired amount of rainfall over any desired period, the time of the year when most rainy days occur, and so on, can all be computed using the model. A number of maps which illustrate the variation of some of the characteristics over the country are given.

As mentioned above to apply the model one needs to use a computer - chiefly because most of the properties outlined above cannot be usefully derived analytically from the parameters. In fact one uses the computer to "generate" artificial rainfall sequences and keeps a record of how many times a condition under investigation is successfully met. As the length of the artificial record increases (and this can be increased to any desired length) so the proportion of successes converges to the probability of the condition being met. It would naturally be preferable to derive such results by more direct means but this does not seem feasible except in a few rather special cases. For example the few "analytic results" relating to application of the model which are discussed in Stern and Coe (1984) appear to require no less computation, and are much more complex and therefore prone to errors. Furthermore each new application would require the derivation of new results and the development of corresponding computer programs. In this respect, and therefore in terms of general applicability, the simulation approach is vastly superior. One

uses the same rainfall "generator" in all applications and simply keeps a record of whatever aspect of the artificial record is of interest, no matter how complex this aspect may be. In this way any properties of any rainfall-based drought model can be established for any of the 2550 stations considered.

A problem which we have not been able to solve is the areal description of drought. No satisfactory methodology exists to tackle this problem on a seasonal basis. A study of streamflow deficits, which provide a measure of the integrated effect of drought over a catchment, was (at least in part) carried out with the object of providing a means of assessing drought risk over a region rather than at a point. The results of this research, which was also intended to cover a second important variable associated with drought, namely streamflow, will be discussed in a separate report: "Assessing the Risk of Streamflow Deficiencies".

Another aspect of areal droughts investigated was the distribution of past annual rainfall deficits. A simple model was fitted to the annual totals of 500 selected stations and the percentile points associated with each of the Weather Bureau water years (October-September) from 1920/21 to 1979/80 were represented on maps, one for each year. This sequence of 60 maps provides a history of areal droughts for the Republic as a whole.

In the course of this project a number of other topics relating to drought were also examined. In particular we investigated the possibility of using tree-ring indices to significantly augment the length of rainfall records. The available rainfall and streamflow data are too short to allow one to accurately assess the risk of exceptionally

severe droughts, such as that recently experienced. Tree-ring indices may be suitable for this purpose - they are correlated with annual rainfall and have much longer records, sometimes going back several hundred years. There is presently only one site in South Africa for which suitable tree-ring indices have been compiled, namely Die Bos in the Cedar Mountains. These records extend back to 1564. We were able to find a significant correlation between tree-ring indices and annual rainfall at a neighbouring rainfall station but the relationship was not close enough to meet the required degree of accuracy. This research was published in Water S A , the Water Research Commission Journal, and is therefore not repeated here. As we were only able to examine one site it is difficult for us to come to any conclusion as to whether such a study would be more fruitful elsewhere, but we believe that this matter is worth further investigation.

Methods for estimating missing values in rainfall records were also developed. In order to apply the results relating to the family of drought models mentioned earlier it was necessary to have complete data records. Most of the South African Weather Bureau records, however, have gaps and it was therefore necessary to find a systematic way of filling these. When it was subsequently decided to fit the rainfall process directly it was no longer necessary to fill these gaps because the relevant estimators could equally well be applied using incomplete data. Although this research on filling in gaps is now something of a by-product of the project it may nevertheless be of use in other contexts and so we have presented it in a separate report : "Estimating the Missing Values in Rainfall Records".

A second by-product is the theory which is also described in a separate report : "Augmenting Hydrological Records".

This describes methods to significantly extend hydrological records (as opposed to estimating relatively few missing values) using related records. The problem here is that standard regression techniques introduce a systematic bias in the variance of the augmented record. Alternative methods were developed.

This particular report is set out as follows:

Chapters 2 to 6 are about the rainfall model. The theory behind the model is derived in Chapters 2 and 3 and algorithms to implement the theory is given in Chapter 4. Chapter 5 contains material on the validation of the model at six test stations which are representative of most of the major climatic regions of the country. Chapter 6 contains some selected examples of application of the model.

Chapter 7 discusses the family of drought indices characterised by an exponential response function to rainfall events. Examples of application are then considered in Chapter 8.

The more technical aspects of the report are discussed in appendices. Appendix 6 contains the estimates of the rainfall model parameters for the 2550 stations.

2. A MODEL TO DESCRIBE THE OCCURRENCE OF WET AND DRY SEQUENCES OF DAYS

By the term "process of daily precipitation" we will mean the sequence of random quantities comprising the precipitation depths on consecutive days. This process exhibits a number of distinctive features:

(i) The distribution of daily precipitation is partly discrete and partly continuous.

On any given day of the year there is a positive probability that there will be zero precipitation (discrete part). On the other hand when precipitation does occur it is convenient to consider its depth as having a continuous distribution.

Naturally it is only possible to measure precipitation depths to a certain degree of accuracy and therefore the measured depths are in fact discrete, e.g. there is a positive probability that the measured depth will be exactly 10 mm. However the probability that the "true" precipitation depth will be exactly 10 mm is zero. In other words the distribution of precipitation depths on wet days is continuous.

These considerations aside, it is simply more convenient to model the precipitation depths on wet days by means of continuous distributions.

(ii) The distribution of daily precipitation depths is seasonal.

It is common knowledge that the process of daily precipitation is not stationary but follows a cyclical pattern with a period of one year.

(iii) The precipitation depths on consecutive days are not independently distributed.

In most regions 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. Consequently the conditional distribution of precipitation depth on a given day depends on the state of precipitation on the previous day.

The above features must be reflected in any reasonable model for the process of precipitation. In the recent literature this process is described by means of a model comprising two components: The first, a first-order Markov chain, describes the occurrence of wet or dry days. The second, some univariate distribution, describes the amount of precipitation on wet days. The parameters of the model are allowed to vary seasonally. Particular models of this type have been discussed by Gabriel and Neumann(1962), a number of subsequent authors (see Richardson (1981) for references) and more recently by Roldan and Woolhiser (1982). This chapter is about the first of these two components.

We will firstly state the assumptions which are implied when one uses a first-order Markov chain to describe the process of wet or dry sequences. It is then pointed out that a naive description of such a model leads to estimation difficulties because it contains too many parameters. The method of fitting a truncated Fourier series as applied for example by Woolhiser and Pegram (1979) is outlined and a new criterion for model selection is derived for this type of model. We then propose a new method of fitting first-order Markov chains to such data. This method overcomes the problem which is frequently encountered in arid and semi-arid regions, namely that of obtaining inadmissible estimates

of the parameters. A model selection procedure for this new method of fitting is also derived. Finally, an example of application is given.

2.1 FIRST-ORDER MARKOV CHAIN ASSUMPTION

The seasonal nature of the precipitation process is such as to tend to cluster wet days (and dry days). But in many regions there is also a short-term persistence in the sequence of wet days which operates over and above that due to seasonality. In other words whether it is wet or dry on day t depends not only on t, the day of the year, but also on the state on previous days, t-1, t-2, etc ... . The number of previous days which are relevant in this respect is often referred to as the "memory" of the process. In our context the memory is certainly finite and, for practical purposes, of short duration.

If, apart from the seasonal fluctuations, the precipitation process is stationary, i.e. exhibits no systematic changes, then it can be described in terms of a (seasonal) Markov chain. In using a *first-order* chain one uses the approximation that the memory of the process has a duration of 1 day. In other words one assumes that, for the purposes of predicting whether day t will be wet or dry, knowing the state on day t-1 is equivalent to knowing the state on all days preceding t.

This does *not* imply that one is assuming the state on day t is distributed independently of that on say day t-2 or any other day. A first-order Markov chain has the property that the state on day t is not distributed independently of that on day t-2, t-3, etc ... .

Whether or not the sequence of wet or dry days really does conform to a first-order Markov chain is not possible to

establish with certainty. But, to our knowledge, there have been no reports of situations where such a model was found to be inadequate. In fact the model has proved itself to be a good approximation for the purpose in a wide variety of regions. (Gabriel and Neumann 1962 , Caskey 1963 , Weiss 1964 , Hopkins and Robillard 1964 , Haan et al 1976 , Woolhiser and Pegram 1979 , Richardson 1981 , Roldan and Woolhiser 1982 .)

In cases of doubt it is of course possible to increase the order of the Markov chain, but this has to be done at the cost of increasing the complexity and number of parameters in the model. A method on which to base the decision of whether to increase the order or not is given in Tong (1975). However this method would have to be extended to apply to seasonal Markov chains.

2.2 NOTATION AND PRELIMINARIES

In the above discussion we have used the day as the basic time unit. The methods given below can be applied if pentads, weeks, months or some other time unit is used. Suppose in fact that the year is divided into NT equal intervals or "times" which we denote by T = 1,2,...,NT.

We will use the following notation. For T = 1,2,...,NT

N(T) is the number of observations made in period T,
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,
NDW(T) is the number of times it was dry in period T-1 and wet in period T,
NDD(T) is the number of times it was dry in period T-1 and dry in period T,
NWW(T) is the number of times it was wet in period T-1 and wet in period T,

NWD(T) is the number of times it was wet in period T-1 and dry in period T,

$$ND(T) = NDW(T) + NDD(T),$$
$$NW(T) = NWW(T) + NWD(T).$$

From the above it can be seen that ND(T) is the number of times that it was dry in period T-1 *and* there was an observation (either wet or dry) in period T. Similarly NW(t) is the number of times that it was wet in period T-1 *and* there was an observation (either wet or dry) in period T.

Note that in the above for T = 1 the period T-1 is NT. For example the day preceding day T = 1 (1 January) is day T = 365 (31 December of the previous year).

Our object is to estimate the following probabilities which specify the Markov chain model:

$\pi_R(T)$ the probability that period T is wet,
$\pi_{\bar{R}}(T)$ the probability that period T is dry,

$\pi_{W/W}(T)$ the probability that period T is wet *given* that period T-1 is wet,
$\pi_{D/W}(T)$ the probability that period T is dry *given* that period T-1 is wet,

$\pi_{W/D}(T)$ the probability that period T is wet *given* that period T-1 is dry,
$\pi_{D/D}(T)$ the probability that period T is dry *given* that period T-1 is dry.

These probabilities need to be estimated for each T = 1,2,...,NT.

The above functions satisfy the following relationships (which are obvious if one reflects on the above definitions):

$$\pi_R(T) + \pi_{\bar{R}}(T) = 1 \quad ,$$
$$\pi_{W/W}(T) + \pi_{D/W}(T) = 1 \quad ,$$
$$\pi_{W/D}(T) + \pi_{D/D}(T) = 1 \quad , \qquad T = 1,2,\ldots,NT \ .$$

So in fact one really only needs to estimate $\pi_R(T)$, $\pi_{W/W}(T)$ and $\pi_{W/D}(T)$ - the estimates for $\pi_{\bar{R}}(T)$, $\pi_{D/W}(T)$ and $\pi_{D/D}(T)$ are then automatically available from the above relationships.

It follows from elementary probability theory that for a given number of observations $N(T)$ and a probability $\pi_R(T)$ of period T being wet, the number of wet days $NR(T)$ is a random variable having a binomial distribution. Using the standard notation this is written as:

$$NR(T) \sim B(N(T), \pi_R(T))$$
$$NWW(T) \sim B(NW(T), \pi_{W/W}(T))$$
$$NDW(T) \sim B(ND(T), \pi_{W/D}(T)) \quad , \qquad T = 1,2,\ldots,NT \ .$$

So the problem of fitting a model to the occurrence of wet and dry sequences is reduced to that of estimating the parameters $\pi_R(T)$, $\pi_{W/W}(T)$ and $\pi_{W/D}(T)$ from the given observations.

The remaining sections concern methods of estimating the above three functions.

2.3 NAIVE ESTIMATORS

The obvious estimator for $\pi_R(T)$, the probability that period T is wet is

$$\hat{\pi}_R(T) = NR(T)/N(T) \quad , \qquad T = 1,2,\ldots,NT$$

that is the proportion of times it was wet in period T in the historical record. This is in fact the estimator one

obtains using the method of maximum likelihood for the binomial distribution. Similarly the maximum likelihood estimators of $\pi_{W/W}(T)$ and $\pi_{W/D}(T)$ are given by

$$\hat{\pi}_{W/W}(T) = NWW(T)/NW(T) \quad ,$$
$$\hat{\pi}_{W/D}(T) = NDW(T)/ND(T) \quad , \qquad T = 1,2,\ldots,NT \ .$$

If these estimators were satisfactory then the problem of fitting a model to the wet and dry sequences would be solved.

The above estimators are not suitable for the historical records which are available.

Unless a very long historical record (in the order of hundreds to thousands of years is available) these estimators yield very poor estimates of the required probabilities. In particular if $\hat{\pi}_R(T)$ is plotted against T (period) one finds that the estimates are highly scattered. One may find for example that the estimated probabilities of wet days on 1, 2 and 3 January are respectively 0,4 0,0 and 0,6 which obviously does not make sense. There are good reasons to believe that the probability that 1 January is a wet day should be very close to the probabilities for 2 and 3 January. In other words we know that $\pi_R(T)$ is a *smooth function* of T whereas the estimates we obtain are not.

The same difficulties arise in the estimation of $\pi_{W/W}(T)$ and $\pi_{W/D}(T)$.

Furthermore if some of the N(T), NW(T) or ND(T) are zero then the corresponding probabilities for these periods cannot be estimated by this method. (For most records it is unlikely that N(T) will ever be zero, but as a rule several of the NW(T) and ND(T) *are* zero for a number of periods.)

In statistical terms the above difficulties arise because one is attempting to estimate too many parameters. In the case of daily records one is attempting to estimate 365 parameters for *each* of the functions $\pi_R(T)$, $\pi_{W/W}(T)$ and $\pi_{W/D}(T)$, i.e. 1095 of them!

None of the available historical records is sufficiently long to justify the estimation of so many parameters and it is therefore necessary to somehow find a way of reducing this number. This can be done in several different ways and essentially they all involve making use of *a priori* information (or assumptions) about the behaviour of the functions $\pi_R(T)$, $\pi_{W/W}(T)$ and $\pi_{W/D}(T)$. As already mentioned we know (or at least believe) that these are *smooth* functions of T, i.e. we expect the properties of precipitation on consecutive days to be very similar. Secondly it is well-known that these functions should be periodic (with a period NT, i.e. one year) and that they are approximately sinusoidal in shape. This information is used in the construction of the remaining methods of estimation which we consider.

2.4 APPROXIMATIONS BASED ON THE FOURIER SERIES REPRESENTATION

The functions $\pi_R(T)$, $\pi_{W/W}(T)$ and $\pi_{W/D}(T)$ are all estimated using the *same method* but with *different data*. To simplify the notation we will use $\pi(T)$ as the generic name representing any one of these three functions. We will also use generic names for the other quantities involved as follows:

Let $M(T) \sim B(MM(T), \pi(T))$, $\quad T = 1,2,\ldots,NT.$

In the case where we are dealing with

(i) $\pi_R(T)$ we have that M(T) = NR(T) and MM(T) = N(T),

(ii) $\pi_{W/W}(T)$ we have that M(T) = NWW(T) and MM(T) = NW(T),

(iii) $\pi_{W/D}(T)$ we have that M(T) = NDW(T) and MM(T) = ND(T).

We repeat that the methods given below can be applied in each of the above three cases. One simply uses the appropriate M(T) and MM(T), T = 1,2,...,NT.

The properties which we would expect $\pi(T)$ to have were discussed at the end of the previous section. These properties (smoothness, periodicity and approximately sinusoidal shape) make it reasonable for us to suppose that $\pi(T)$ can be quite accurately approximated by the *first few terms of its Fourier representation*. This approximation has been used by a number of authors, e.g. Woolhiser and Pegram (1979).

The exact Fourier representation of $\pi(T)$ is of the form

$$\pi(T) = \sum_{i=1}^{NT} \theta_i \, \phi_i(T) \qquad T = 1,2,\ldots,NT$$

where

$$\phi_i(T) = \begin{cases} 1 & i = 1 \\ \cos\left(\frac{i}{2} \cdot \frac{2\pi(T-1)}{NT}\right) & i = 2,4,6,\ldots \\ \sin\left(\frac{i-1}{2} \cdot \frac{2\pi(T-1)}{NT}\right) & i = 3,5,7,\ldots \end{cases}$$

$$\theta_i = \sum_{T=1}^{NT} \pi(T) \, \phi_i(T) \qquad i = 1,2,\ldots,NT \, .$$

Now define $\pi(T,L)$ to be the function which is given by the sum of the first L terms of the Fourier representation of $\pi(T)$, i.e.

$$\pi(T,L) = \sum_{i=1}^{L} \theta_i \, \phi_i(T) \quad , \qquad \begin{array}{l} T = 1,2,\ldots,NT, \\ L < NT \, . \end{array}$$

Note that $\pi(T,NT) = \pi(T)$. The approximation which we make is that for some $L < NT$

$$\pi(T,L) \approx \pi(T) \quad , \quad T = 1,2,\ldots,NT \ .$$

Recall that $\pi(T)$ is in fact unknown, i.e. there are NT quantities which we need to estimate, viz $\pi(1), \pi(2),\ldots,\pi(NT)$. Using the full Fourier representation there are still NT quantities which are unknown, viz $\theta_1, \theta_2,\ldots,\theta_{NT}$. However under the above approximation we need use only L quantities to represent $\pi(T)$, viz $\theta_1, \theta_2,\ldots,\theta_L$. It turns out that for nearly all situations the approximation is sufficiently accurate for *small* values of L (usually $L < 11$ for daily rainfall sequences). Consequently the number of parameters which need to be estimated is greatly reduced.

In statistical terms the above considerations can be described as follows: Whatever method is used to fit a statistical model to data there will in general be a discrepancy between the "true" or operating model and the model which one actually fits to the data. This discrepancy stems from two sources. The first, called the *discrepancy due to approximation*, occurs when one approximates the function of interest using some other function. In our case we want to approximate $\pi(T)$ using $\pi(T,L)$. This discrepancy can be reduced by increasing the number of terms in the approximation, i.e. the number of parameters, L. In fact by setting $L = NT$ this discrepancy is reduced to zero.

In contrast the *discrepancy due to estimation* arises because we do not know the exact values of the parameters - they have to be estimated from the historical record. This discrepancy tends to increase if the number of parameters, L, is increased.

In other words one has two sources of discrepancy which act in opposition. By adjusting the number of parameters to

decrease the one type of discrepancy one necessarily increases the other type of discrepancy. By using the naive estimators discussed in the previous section we were implicitly selecting a model with zero discrepancy due to approximation but the *largest* possible discrepancy due to estimation. It turns out that by decreasing the number of parameters one can achieve a *substantial reduction in the discrepancy due to estimation* for a relatively small increase in the discrepancy due to approximation.

Objective methods to select L in such a way as to minimise the estimated overall discrepancy are discussed in the next section. The remainder of this section describes a method of estimating the parameters $\theta_1, \theta_2, \ldots, \theta_L$ for a given L.

The number of parameters L is always taken to be an odd integer. This restriction is made partly for programming convenience and partly for the following reason: The truncated Fourier representation of $\pi(T)$ which we called $\pi(T,L)$ can be rewritten as a sum of shifted cosine terms:

$$\pi(T,L) = \begin{cases} \alpha_0 + \sum_{i=1}^{P} \alpha_i \cos\left(\frac{2\pi i}{NT}(T-1-\beta_i)\right) & \text{, if L is odd} \\ \alpha_0 + \sum_{i=1}^{P-1} \alpha_i \cos\left(\frac{2\pi i}{NT}(T-1-\beta_i)\right) + \alpha_P \cos\frac{2\pi P}{NT}(T-1), & \text{if L is even} \end{cases}$$

where the relationship between the parameters θ_i and the new parameters α_i, β_i is given by

$\alpha_0 = \theta_1$

$\alpha_i = (\theta_{2i}^2 + \theta_{2i+1}^2)^{\frac{1}{2}}$ for $i = 1,2,\ldots,P$

$\beta_i = \frac{NT}{2\pi i}$ Arctan $(\theta_{2i+1}/\theta_{2i})$ for $i = 1,2,\ldots,P$

and P is the integer part of (L-1)/2.

In this representation the parameters α_i and β_i are called respectively the amplitude and phase of the ith harmonic. This representation is equivalent to our previous representation of $\pi(T,L)$ but is expressed in terms of different parameters.

Note that if L is even then the highest harmonic (the last cosine term) does not contain a phase parameter. This leads to the undesirable property that a shift in the time origin (e.g. from 1 January to 1 October) results in changes in the parameters α_i *and* β_i. It also results in a change in the discrepancy due to approximation (unless $L = NT$). In other words if L is taken to be an even integer then the quality of the fit which we obtain after estimating the parameters *will depend on the time origin selected.* If, on the other hand, L is taken to be an odd integer then a shift in the time origin leads to a simple translation in the phase parameters β_i but *no change in the amplitude parameters,* α_i. Consequently by restricting L to be an odd integer we obtain the same degree of approximation for all time origins.

We have used the Fourier representation of $\pi(T)$ 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 $\pi(T)$ is known to be approximately sinusoidal in shape and consequently we can expect that even for small values of L the approximation $\pi(T,L) \approx \pi(T)$ will be reasonably accurate. Secondly $\pi(T,L)$ is periodic, which is a property that $\pi(T)$ is known to have. Thirdly the individual components in the representation are orthogonal, which is a convenient mathematical property.

2.5 ESTIMATION

We now consider the problem of estimating the L coefficients $\theta_1, \theta_2, \ldots, \theta_L$ which will give us estimates

$$\hat{\pi}(T) = \sum_{i=1}^{L} \hat{\theta}_i \, \phi_i(T) \qquad T = 1,2,\ldots,NT$$

This problem can be formulated as follows: Suppose that, for $T = 1,2,\ldots,NT$, $MM(T)$ independent Bernoulli trials are performed and that at each trial there is a probability $\pi(T,L)$ of a "success", where

$$\pi(T,L) = \sum_{i=1}^{L} \theta_i \, \phi_i(T)$$

and suppose that $M(T)$ "successes" were observed. How can one estimate the parameters $\theta_1, \theta_2, \ldots, \theta_L$?
(In the case of $\pi(T) = \pi_R(T)$, this independence assumption is not met because of the first-order dependence structure of wet and dry days. Therefore the estimates of $\hat{\pi}_R(T)$ (i.e. When $MM(T) = N(T)$, $M(T) = NR(T)$) obtained using the estimation technique which follows, will only be approximately correct.)

The likelihood of the observed values as a function of these parameters is given by

$$L(\theta_1, \theta_2, \ldots, \theta_L \; ; \; M(T), \; T = 1,2,\ldots,NT)$$

$$= \prod_{T=1}^{NT} \binom{MM(T)}{M(T)} \pi(T,L)^{M(T)} \, (1-\pi(T,L))^{MM(T)-M(T)} .$$

For simplicity we denote the likelihood by $L(\theta)$. Now

$$\log L(\theta) = \sum_{T=1}^{NT} \log\binom{MM(T)}{M(T)} + \sum_{T=1}^{NT} M(T) \log \pi(T,L)$$

$$+ \sum_{T=1}^{NT} (MM(T)-M(T)) \log (1-\pi(T,L)) .$$

The maximum likelihood estimators of $\theta_1, \theta_2, \ldots, \theta_L$ are those values of these parameters which maximise the likelihood, or equivalently the log-likelihood. They are therefore the solutions to the system of equations given by

$$\frac{\partial \log L(\theta)}{\partial \theta_i} = 0 \qquad , \qquad i = 1,2,\ldots,L$$

It is straightforward to show that this system of equations is given by

$$\sum_{T=1}^{NT} \frac{M(T)-\pi(T,L)MM(T)}{\pi(T,L)(1-\pi(T,L))} \phi_i(T) = 0 \qquad , \qquad i = 1,2,\ldots,L$$

where $\pi(T,L)$ is the function expressed in terms of the parameters $\theta_1, \theta_2,\ldots,\theta_L$ given above.

This system of equations cannot be solved for the θ_i analytically; it has to be solved using an iterative method. As the log-likelihood is a concave function of the parameters, and as good starting values for the iteration can be given (see below), the Newton-Raphson method can be expected to perform well. This was indeed found to be the case and convergence is rapid.

To apply the Newton-Raphson method the matrix of second derivatives is required. This is given by

$$\frac{d^2 \log L(\theta)}{\partial\theta_i \partial\theta_j} = - \sum_{T=1}^{NT} \frac{M(T)[1-\pi(T,L)]^2+[MM(T)-M(T)]\pi(T,L)^2}{\pi(T,L)^2[1-\pi(T,L)]^2} \phi_i(T)\phi_j(T)$$

$$i,j = 1,2,\ldots,L.$$

This L×L matrix, when evaluated at the solutions to the system of maximum likelihood equations, also provides an estimator for the variance-covariance matrix of the maximum likelihood estimates.

To start the iteration one needs suitable starting values $\theta_1^{(o)}, \theta_2^{(o)},\ldots,\theta_L^{(o)}$ of the parameters. The ordinary least squares estimates of the parameters provide excellent starting values. These are given by:

$$\theta_i^{(o)} = K(i) \sum_{\substack{T=1 \\ MM(T)\neq 0}}^{NT} \left(\frac{M(T)}{MM(T)}\right)\phi_i(T) \quad , \qquad i = 1,2,\ldots,L$$

where

$$K(i) = 1/ \sum_{\substack{T=1 \\ MM(T)\neq 0}}^{NT} \phi_i(T)^2 \quad , \qquad i = 1,2,\ldots,L$$

In cases where only a few of the MM(T) are equal to zero the approximation K(i) = NT is adequate.

We now give an outline of the Newton-Raphson algorithm to estimate the parameters. We use the notation:

$f(\theta^{(k)})$ to denote the (column) vector of the L functions $\partial \log L(\theta)/\partial\theta_i$ evaluated at the point $\theta^{(k)}$, and $F(\theta^{(k)})$ the L×L matrix $\partial^2 \log L(\theta)/\partial\theta_i\partial\theta_j$ evaluated at $\theta^{(k)}$, where $\theta^{(k)}$ is the vector of estimates of the L parameters obtained after k iterations. $\delta^{(k)}$ is a vector of L entries defined in the algorithm below.

ALGORITHM

Step 1 Obtain an initial estimate, $\theta^{(o)}$, and set k = 0.

Step 2 Compute $f(\theta^{(k)})$ and $F(\theta^{(k)})$.

Step 3 Compute $\delta^{(k)}$, the solution to the linear system of equations given by
$F(\theta^{(k)})\ \delta^{(k)} = f(\theta^{(k)})$

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

Step 5 Check for convergence, i.e. if the entries of $f(\theta^{(k)})$ are sufficiently close to zero. If convergence has occurred then stop, otherwise increment k by 1 and return to step 2.

The vector f and the matrix F above enjoy several special features which can be used to reduce the computational effort. Although we will not discuss these here the interested reader is referred to the algorithm which is discussed in section 2.9, for which a detailed algorithm is given in chapter 4. It is a fairly simple matter to modify the detailed algorithm to deal with the case described above.

2.6 MODEL SELECTION

We now discuss the question of how to select L, the number of parameters to be fitted. The general theory on which this selection is derived in Linhart and Zucchini (1982a, 1982b, 1986). In this section we will only give an outline of this theory and simply state the results without giving details.

Recall that there are two sources of discrepancy involved when a model is fitted to data, viz the discrepancy due to approximation and the discrepancy due to estimation, and that reducing either one of these one necessarily increases the other. Recall also that the levels of these two component discrepancies are controlled (for a given historical record) by the number of parameters, L.

The idea behind the method of model selection which we will describe is to select L in such a way as to minimise the combined effect arising from these two discrepancies, viz the overall discrepancy.

The natural measure of discrepancy when one uses the method of maximum likelihood for estimation is the Kullback-Leibler discrepancy. In our application this is given by

$$\Delta(\theta) = \text{Constant} - \sum_{T=1}^{NT} \pi(T)MM(T) \log \pi(T,L)$$
$$- \sum_{T=1}^{NT} (1-\pi(T))MM(T) \log (1-\pi(T,L)) .$$

A constant estimator of this discrepancy, called an empirical discrepancy, is given by

$$\Delta_n(\theta) = - \log L(\theta)$$

where $L(\theta)$ is the likelihood function given in section 2.5.

It can be shown that if one is trying to select the number of parameters, L, which leads to the smallest expected discrepancy (for a given historical record) then a criterion for selection is

$$\Delta_n(\theta) + \text{tr } \hat{\Omega}^{-1} \hat{\Sigma}$$

where $\hat{\Sigma}$ is an L×L matrix having entry (i,j) given by

$$\sum_{\substack{T=1 \\ MM(T)\neq 0}}^{NT} \frac{M(T)[1-M(T)/MM(T)]}{\hat{\pi}(T,L)^2[1-\hat{\pi}(T,L)]^2} \phi_i(T) \phi_j(T) \quad , \quad i,j = 1,2,\ldots,L$$

and $\hat{\Omega}$ is an L×L matrix having entries (i,j) given by

$$\sum_{T=1}^{NT} \frac{M(T)[1-\hat{\pi}(T,L)]^2 + [MM(T)-M(T)]\hat{\pi}(T,L)^2}{\hat{\pi}(T,L)^2[1-\hat{\pi}(T,L)]^2}$$

where

$$\hat{\pi}(T,L) = \sum_{i=1}^{L} \hat{\theta}_i \phi_i(T)$$

and the $\hat{\theta}_i$ are the maximum likelihood estimators of the parameters θ_i when L parameters are fitted.

Note that $\hat{\Omega}$ is simply minus the matrix of second derivatives required in the Newton-Raphson algorithm evaluated at the maximum likelihood estimates. Consequently $\hat{\Omega}$ is available at no extra computational cost. Similarly $\hat{\Sigma}$ is available at only marginal additional computational cost because it is closely related to the vector of first derivatives evaluated in the estimation algorithm.

To implement the model selection method one estimates the parameters of the model for increasing values of L and in each case computes the value of the above criterion. Initially the value of the criterion will decrease (as L is increased) but after a certain point will begin to increase. The number of parameters which is estimated to be optimal is that which leads to the *smallest* value of the criterion.

It can be shown that, under the assumption that for some L_0 one has that $\pi(T) = \pi(T,L_0)$, i.e. that $\pi(T)$ can be *exactly* represented by $L_0 < NT$ parameters, then the theory on which the above method of selection is based leads to the well-known Akaike Information Criterion (AIC):

$$AIC = \Delta_n(\theta) + L.$$

In practice it turns out that the AIC leads to very similar results (unless L is very small) to those obtained using the method of discrepancies, even if the mentioned assumption is only approximately true. The Akaike criterion involves less computation and is therefore to be recommended except perhaps in cases where very little data are available and consequently a small value of L is likely to be selected.

Other methods of model selection are available; in particular methods based on statistical tests of hypotheses based on the likelihood ratio are widely employed. Using such a method an increase in L is only made if there is strong evidence that the current L is not large enough. However it is our opinion that methods based on such tests are inappropriate in this context. Here it is not the object to *prove* whether $\pi(T)$ may or may not be exactly represented by a certain number of parameters. Rather we are trying to find an appropriately simple representation which approximates $\pi(T)$ quite well but which does not contain more parameters than can be reasonably estimated.

2.7 INADMISSIBLE ESTIMATES

A problem which often arises in applying the above methods in regions with a marked dry season is that one obtains inadmissible estimates for $\pi(T)$, i.e. one often obtains $\hat{\pi}(T) < 0$. (The other type of inadmissible estimate $\hat{\pi}(T) > 1$ does not occur in South Africa.) If this problem only occurred in isolated cases, or for only a few time points, T, then it would not be unreasonable to simply replace the offending estimates by zero, or some suitably small quantity. Unfortunately in South Africa this phenomenon occurs for a good many stations and furthermore the estimates can be negative for a period of several months.

There are ways to deal with this problem. Woolhiser and Pegram (1979) employ what amounts to constrained maximum likelihood estimation of the parameters, i.e. they maximise the likelihood subject to the constraint $0 < \hat{\pi}(T,L) < 1$ for each L.

This method involves a substantial computing effort and fairly sophisticated optimisation software. Although the method could be implemented on some of the larger micro-computers, the

task of programing it would daunt many potential users. A further objection to the method is that the statistical theory which normally provides the properties of the estimators, e.g. the variance-covariance matrix of the estimates, applies to maximum likelihood estimation - not to constrained maximum likelihood estimators. It is not known to what extent the properties of maximum likelihood estimators hold for the constrained case. The same objection can be made when model selection using constrained maximum likelihood is used - again the theory has not been derived for this case.

The effort involved in computing constrained maximum likelihood estimates can be considerably reduced by using the theory of semi-infinite programing, see e.g. Flachs and Martin (1982). This method does however suffer from the disadvantage (in this context) that it is sensitive to certain starting values required in the iteration. Unless starting values are given which are very close to the solution then the iteration often does not converge. It may well be possible to overcome this difficulty by refining the method of iteration but the objections to constrained maximum likelihood still remain.

A quite different approach, suggested by Dr. T. Stewart, which entirely circumvents the problem of having to deal with constraints is to use a different representation for the probabilities $\pi(T)$. This method is the subject of the remainder of the chapter.

2.8 APPROXIMATIONS BASED ON THE FOURIER REPRESENTATION OF THE LOGITS

The problem which we discussed in section 2.7 arises because we are estimating probabilities which necessarily must lie in the interval [0,1] and the estimates we obtained sometimes fell outside this interval. Clearly this problem does

not arise when we are estimating quantities which do not have to lie in any bounded interval. Now a probability π can be transformed, using the logistic transformation, to a so-called *logit*, λ, which is given by

$\lambda = \log(\pi/(1-\pi))$, i.e.

$\pi = e^{\lambda}/(1+e^{\lambda})$.

From the above relationships it can be seen that there is a one-to-one correspondence between probabilities and logits, e.g. a probability of $\frac{1}{2}$ corresponds to a logit of zero, probabilities less than $\frac{1}{2}$ correspond to negative logits and those greater than $\frac{1}{2}$ to positive logits. For our purposes the attractive feature of logits is that unlike probabilities they are entirely *unconstrained*. This property can be used to circumvent the problem of obtaining inadmissible estimates for $\pi(T)$.

Instead of approximating $\pi(T)$ by a truncated form of its Fourier representation, we make this type of approximation for the corresponding logits, $\lambda(T)$. The Fourier representation of $\lambda(T)$ is given by

$$\lambda(T) = \sum_{i=1}^{NT} \gamma_i \phi_i(T)$$

where, as before,

$$\phi_i(T) = 1 \qquad i = 1$$

$$= \cos\left(\frac{i}{2} \cdot \frac{2\pi(T-1)}{NT}\right) \qquad i = 2,4,6,\ldots$$

$$\sin\left(\frac{i-1}{2} \cdot \frac{2\pi(T-1)}{NT}\right) \qquad i = 3,5,7,\ldots$$

and $\gamma_1, \gamma_2, \ldots$, are the Fourier coefficients.

We now define

$$\lambda(T,L) = \sum_{i=1}^{L} \gamma_i \phi_i(T) \quad , \quad T = 1,2,\ldots,NT; \quad L < NT.$$

We use an approximation which is entirely analogous to that in section 2.4:

$$\lambda(T,L) \approx \lambda(T) \quad T = 1,2,\ldots,NT .$$

The justification for making this approximation is the same as that for the original model. Again L is taken to be *an odd integer*.

We note that this representation has all the desirable properties (smoothness, periodicity and approximate sinusoidal shape), of the previous representation. It has the additional desirable property that the parameters γ_i *are unconstrained*.

It was also found that in all the cases where both this logit representation and the probability representation were fitted the former had a better fit to the data. For this reason alone the logit representation is preferable to the original probability representation.

2.9 ESTIMATION

We now discuss a method of estimating the L coefficients $\gamma_1, \gamma_2, \ldots, \gamma_L$ which will give us estimates

$$\hat{\lambda}(T) = \sum_{i=1}^{L} \hat{\gamma}_i \phi_i(T) \quad , \quad T = 1,2,\ldots,NT$$

and hence estimates

$$\hat{\pi}(T) = e^{\hat{\lambda}(T)}/(1+e^{\hat{\lambda}(T)}) \quad , \quad T = 1,2,\ldots,NT.$$

Here the problem is formulated analogously to that in section 2,5, i.e.: Suppose that, for T = 1,2,...,NT, MM(T) independent Bernoulli trials are performed and that at each trial

there is a logit $\lambda(T,L)$ of a "success" where

$$\lambda(T,L) = \sum_{i=1}^{L} \gamma_i \phi_i(T) \quad , \qquad T = 1,2,\ldots,NT; \quad L < NT$$

and suppose that $M(T)$ "successes" were observed. How can one estimate the parameters $\gamma_1, \gamma_2, \ldots, \gamma_L$?

The likelihood of the observed values as a function of these parameters is given by

$$L(\gamma_1,\gamma_2,\ldots,\gamma_L;\ M(T),\ T = 1,2,\ldots,NT)$$

$$= \prod_{T=1}^{NT} \binom{MM(T)}{M(T)} \left[e^{\lambda(T,L)} / \left(1+e^{\lambda(T,L)}\right)\right]^{M(T)} \left[1-e^{\lambda(T,L)} / \left(1+e^{\lambda(T,L)}\right)\right]^{MM(T)-M(T)} .$$

For simplicity we denote this likelihood by $L(\gamma)$. Now it follows that

$$\log L(\gamma) = \sum_{T=1}^{NT} \log \binom{MM(T)}{M(T)} + \sum_{T=1}^{NT} M(T)\, \lambda(T,L)$$

$$- \sum_{T=1}^{NT} MM(T) \log \left(1+e^{\lambda(T,L)}\right) .$$

The maximum likelihood estimators of $\gamma_1, \gamma_2, \ldots, \gamma_L$ are those values of these parameters which maximise the log-likelihood. They are given by the solutions to the system of equations:

$$\frac{\partial \log L(\gamma)}{\partial \gamma_i} = 0 \qquad , \qquad i = 1,2,\ldots,L .$$

Differentiating the log-likelihood with respect to γ_i, $i = 1,2,\ldots,L$, it is straightforward to show that this system of equations is given by

$$\sum_{T=1}^{NT} \left\{ M(T) - \frac{MM(T)e^{\lambda(T,L)}}{1+e^{\lambda(T,L)}} \right\} \phi_i(T) = 0 \quad , \qquad i = 1,2,\ldots,L ,$$

where $\lambda(T,L)$ is the function expressed in terms of the parameters $\gamma_1, \gamma_2, \ldots, \gamma_L$, given above.

As in the previous case the maximum likelihood equations cannot be solved analytically - they have to be solved using an iterative method. Also as in the previous case the log-likelihood is a concave function of the parameters and again good starting values can be given (see below). Consequently the Newton-Raphson iteration technique performs well and convergence is rapid.

To apply the Newton-Raphson method one requires the matrix of second derivatives which is given by

$$\frac{\partial^2 \log L(\gamma)}{\partial \theta_i \partial \theta_j} = -\sum_{T=1}^{NT} \frac{MM(T)e^{\lambda(T,L)}}{[1+e^{\lambda(T,L)}]^2} \phi_i(T)\, \phi_j(T) \quad , \quad i,j = 1,2,\ldots,L.$$

The L×L matrix when evaluated at the solutions to the system of maximum likelihood equations, also provides an estimate of the variance-covariance matrix of the maximum likelihood parameter estimates.

The following starting values, based on ordinary least-squares estimation, can be used to begin the iteration:

$$\gamma_i^{(o)} = K(i) \sum_{\substack{T=1 \\ MM(T)\neq 0}}^{NT} \log\left\{\frac{M(T)/MM(T)}{1-M(T)/MM(T)}\right\} \phi_i(T) \quad , \quad i = 1,2,\ldots,L \ ,$$

where

$$K(i) = 1/\sum_{\substack{T=1 \\ MM(T)\neq 0}}^{NT} \phi_i(T)^2 \quad , \quad i = 1,2,\ldots,L \ .$$

We now give an outline of the Newton-Raphson algorithm used to estimate the parameters. We use the notation $g(\gamma^{(k)})$ to denote the column vector of the L functions $\partial \log L(\gamma)/\partial \gamma_i$ evaluated at the point $\gamma^{(k)}$, and $G(\gamma^{(k)})$ to denote the L×L matrix $\partial^2 \log L(\gamma)/\partial \gamma_i \partial \gamma_j$ evaluated at $\gamma^{(k)}$, where $\gamma^{(k)}$ is the vector of estimates of the L parameters obtained after k iterations. $\delta^{(k)}$ is a vector of L entries defined in the algorithm below.

ALGORITHM

Step 1 Obtain an initial estimate, $\gamma^{(o)}$, and set k = 0.

Step 2 Compute $g(\gamma^{(k)})$ and $G(\gamma^{(k)})$

Step 3 Compute $\gamma^{(k)}$, the solution to the linear system of equations given by
$G(\gamma^{(k)})\ \delta^{(k)} = g(\gamma^{(k)})$

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

Step 5 Check for convergence, i.e. if the entries of $g(\gamma^{(k)})$ are sufficiently close to zero. If convergence has occurred then stop, otherwise increment k by 1 and return to step 2.

Complete details of the above algorithm are given in section 4.3.

2.10 MODEL SELECTION

The selection of L for the logit representation model can be carried out along similar lines as that for the probability representation model. The Kullback-Leibler discrepancy is given by:

$$\Delta(\gamma) = \text{Constant}$$
$$- \sum_{T=1}^{NT} \pi(T)MM(T) \log \left[e^{\lambda(T,L)} / \left(1+e^{\lambda(T,L)}\right)\right]$$
$$- \sum_{T=1}^{NT} (1-\pi(T))MM(T) \log \left[1 / \left(1+e^{\lambda(T,L)}\right)\right]$$

An empirical discrepancy (i.e. a consistent estimator of the discrepancy) is given by

$$\Delta_n(\gamma) = - \log L(\gamma)$$

where $\log L(\gamma)$ is the log-likelihood defined in section 2.9.

It can be shown that if one is trying to select the number of parameters, L, which leads to the smallest expected discrepancy (for a given historical record) then a criterion for selection is

$$\Delta_n(\gamma) + \text{tr } \hat{\Omega}^{-1} \hat{\Sigma}$$

where here $\hat{\Sigma}$ is an L×L matrix whose entry (i,j) is given by

$$\{\hat{\Sigma}\}_{ij} = \sum_{\substack{T=1 \\ MM(T)\neq 0}}^{NT} \frac{M(T)(1-M(T)/MM(T))(1+e^{\hat{\lambda}(T,L)})^4}{e^{2\hat{\lambda}(T,L)}} \phi_i(T)\phi_j(T), \quad i,j=1,2,\ldots,L$$

and Ω is also an L×L matrix whose entry (i,j) is given by

$$\{\hat{\Omega}\}_{ij} = \sum_{T=1}^{NT} \frac{MM(T)e^{\hat{\lambda}(T,L)}}{[1+e^{\hat{\lambda}(T,L)}]^2} \phi_i(T)\ \phi_j(T) \quad , \quad i,j = 1,2,\ldots,L$$

where

$$\hat{\lambda}(T,L) = \sum_{i=1}^{L} \hat{\gamma}_i\ \phi_i(T) \quad , \quad T = 1,2,\ldots,NT \ ,$$

and the $\hat{\gamma}_i$ are the maximum likelihood estimates of the parameters γ_i when L parameters are fitted.

In order to select L using this criterion one has to fit the model for different values of L and then choose that L which leads to the smallest value of the criterion. It usually turns out that the optimal L is quite small (less than 11 for daily data) and so it is recommended that one begin with L = 1 and then increase L in steps of 2 - because L should always be odd - until the criterion begins to increase in value. The criterion can sometimes increase and then decrease again, i.e. it can have a number of local minima. However this seldom happens in the vicinity of the global minimum which is almost without exception the first local minimum. In other words, for practical purposes, it is sufficient to increase L until the first minimum of the criterion is found.

It can be shown that under the assumption that if for some integer L_0: $\lambda(T) = \lambda(T,L_0)$, i.e. that $\lambda(T)$ can be *exactly* represented using $L_0 < NT$ parameters, then the above method leads to the Akaike Information Criterion rather than the criterion given above where

$$AIC = \Delta_n(\gamma) + L \ .$$

It again turns out that unless L is small (i.e. if $L < 5$) then the AIC criterion leads to almost identical results to those obtained using the method of discrepancies. The AIC criterion is simpler to compute and is therefore preferable in most cases. The exception to this is if only very little data is available and consequently a small value of L is likely to be selected.

2.11 THE AMPLITUDE-PHASE REPRESENTATION

In the above we have used the representation

$$\lambda(T,L) = \sum_{i=1}^{L} \gamma_i \, \phi_i(T) \qquad T = 1,2,\ldots,NT; \quad L < NT$$

where L is an odd integer.

This representation is particularly convenient for computation because the values $\phi_i(T)$, $i = 1,2,\ldots,L$; $T = 1,2,\ldots,NT$, need only be computed *once*. We have that

$$\phi_i(T) = \begin{cases} 1 & i = 1 \\ \cos\left(\frac{i}{2} \cdot \frac{2\pi(T-1)}{NT}\right) & i = 2,4,6,\ldots \\ \sin\left(\frac{i-1}{2} \cdot \frac{2\pi(T-1)}{NT}\right) & i = 3,5,7,\ldots \end{cases}$$

The computation of sine and cosine functions is relatively slow and as the terms $\phi_i(T)$ are required very frequently in the computation of the estimators it is particularly advantageous to compute the $\phi_i(T)$ once only at the beginning of the program and to store the values in an array. (e.g. For daily data this array would be 21×365, if 21 is regarded as the largest probable value of L.) An efficient algorithm to compute this array is given in section 4.3.

Although the above representation is convenient for computing, it is less convenient for interpreting the parameters and comparing the parameters for different stations. An *amplitude-phase representation* is much easier to interpret and to use for interpolation on a map. Using this representation we have:

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

where $\alpha_0 = \gamma_1$ and

$$\alpha_i = (\gamma_{2i}^2 + \gamma_{2i+1}^2)^{\frac{1}{2}} \quad , \qquad i = 1,2,\ldots,P \, ,$$

$$\beta_i = \frac{NT}{2\pi i} \operatorname{Arctan}(\theta_{2i+1}/\theta_{2i}) \quad , \qquad i = 1,2,\ldots,P \, ,$$

This representation of the logits is entirely equivalent to that involving the parameters γ_i, $i = 1,2,\ldots,L$. Furthermore the estimates of the $\alpha_0, \alpha_1,\ldots,\alpha_L, \beta_1, \beta_2,\ldots,\beta_L$ which are obtained by first estimating $\gamma_1, \gamma_2,\ldots,\gamma_L$ and then transforming them as above are the maximum likelihood estimates, i.e. the values we would have obtained if we had used the amplitude-phase representation in the first place and then estimated the parameters using the method of maximum likelihood. This is particularly convenient because it is computationally easier to estimate the γ_i, $i = 1,2,\ldots,L$.

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

$$\text{If } \gamma_{2i} > 0 \text{ then } \begin{cases} \text{if } \gamma_{2i+1} < 0 \text{ then } \beta_i = C[A + 2\pi] \\ \text{if } \gamma_{2i+1} > 0 \text{ then } \beta_i = CA \end{cases},$$

$$\text{If } \gamma_{2i} = 0 \text{ then } \begin{cases} \text{if } \gamma_{2i+1} < 0 \text{ then } \beta_i = C[3\pi/2] \\ \text{if } \gamma_{2i+1} > 0 \text{ then } \beta_i = C[\pi/2] \end{cases},$$

$$\text{If } \gamma_{2i} < 0 \text{ then } \beta_i = C[A + \pi],$$

where $C = NT/(2\pi*i)$ and $A = \text{Arctan}(\gamma_{2i+1}/\gamma_{2i})$ 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 $\beta_i \in (0, NT/i]$, $i = 1,2,\ldots,L$. This makes comparison between stations particularly convenient.

2.12 SUMMARY

We have described a model for the occurrence of wet and dry sequences of days using a first order Markov chain which has seasonal parameters. It was argued that the naive estimators

of these parameters yield unsatisfactory estimates and an alternative method was described based on truncated Fourier representations of the model parameter functions. This alternative, which has been considered in the literature, was found to be unsuitable for stations in arid and semi-arid regions because it leads to inadmissible estimates of the parameters. An alternative method was then presented which is based on truncated Fourier representations of the logit functions in the model. Methods of estimation were derived and the question of model selection was discussed. We derived an objective criterion to decide which model is the most appropriate for a given station having a given length of record.

To select the best model requires roughly seven times as much computation on average than is needed to fit a model with a fixed number of parameters. The computing cost of fitting the 2550 selected data records is considerable but that of carrying out individual selections for each record is prohibitive. It was therefore necessary for us to find a cheaper method for deciding how many parameters should be used for each record. In order to investigate the variation in the optimum number of parameters, L^*, the full model selection procedure was applied to 100 test stations; the results are illustrated in Figures 2.11.1 to 2.11.3.

The values of L^* for the probability of a wet day (Figure 2.11.1) ranged between 1 and 11, the average was 7,6 and the mode, which accounted for almost 50% of the cases, was $L^* = 7$. This was one of the reasons why we eventually elected to use 7 parameters for the complete set of 2550 records. A detailed examination of the computed values of the criterion, $C(L)$, as a function of the number of parameters, L, revealed that $C(L)$ is very close to $C(L^*)$ for $L = 7$ and $L = 9$. In other words, for those cases where $L^* \neq 7$, very little accuracy is lost by using 7 parameters.

The same is true if one uses 9 parameters. Keeping in mind that the method of model selection employed here is less stringent than methods based on conventional tests of hypotheses (and which therefore generally lead to a smaller number of parameters being selected), we decided that 7 parameters would be preferable to 9. An additional factor to be considered is that it requires 30% more computing time to fit the 9 parameter model. Finally it is simply more convenient to have fewer parameters because there are then fewer numbers which must be entered into the subsequent programs required to implement the model.

The length of the historical record plays a role in determining L^* and it would not have been unreasonable to allow the final number of parameters used for each station to depend on the length of its record, e.g. to use 9 parameters whenever at least 50 years of data are available. The potential gain in accuracy resulting from such a procedure is rather small and in some cases it even leads to slightly lower accuracy, viz. whenever $L^* < 9$. This refinement was therefore not adopted.

The values of L^* for the probability of a wet day given that the preceding day was wet (Figure 2.11.2) and given that the preceding day was dry (Figure 2.11.3) ranged between 1 and 9 and between 3 and 9 respectively. The averages were 4,0 and 6,1 and the modes 3 and 7 although 5 came very near to being the mode in both cases. Following arguments similar to those which led us to choose 7 parameters for the probability of a wet day we decided that 5 parameters should be used for each of two cases considered here. We note that it is not inconsistent to have selected 7 parameters for the one model and 5 each for the other two because the effective sample size is smaller in the latter two cases.

FIGURE 2.11.1

Optimum number of parameters to estimate the probability of a wet day for each of 100 test stations.

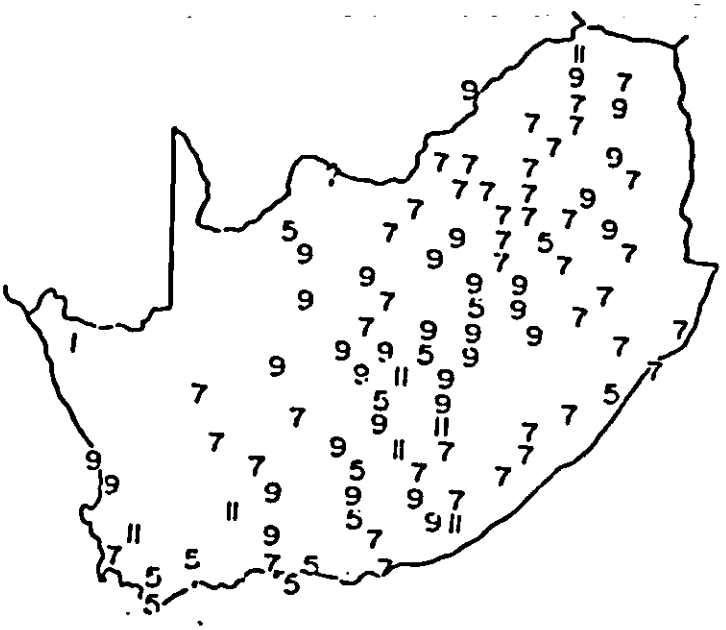


FIGURE 2.11.2

Optimum number of parameters to estimate the probability of a wet day given that the preceding day was wet for each of 100 test stations.

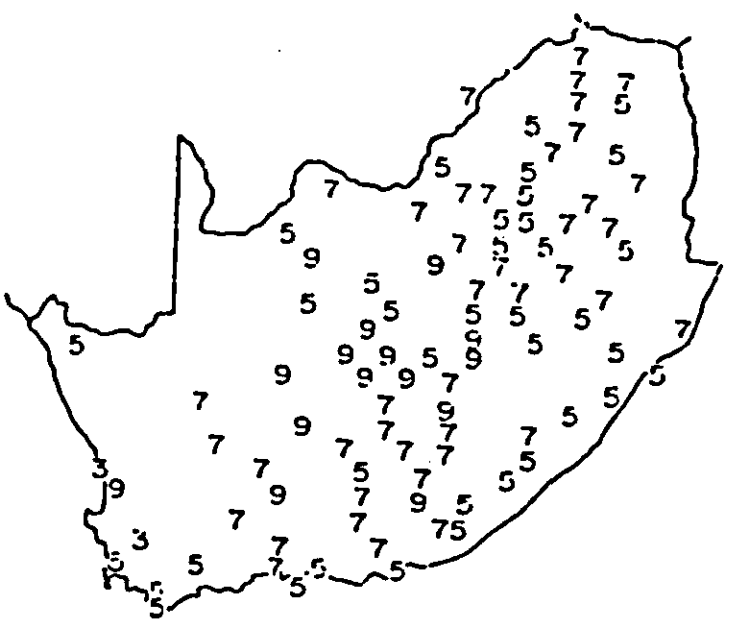
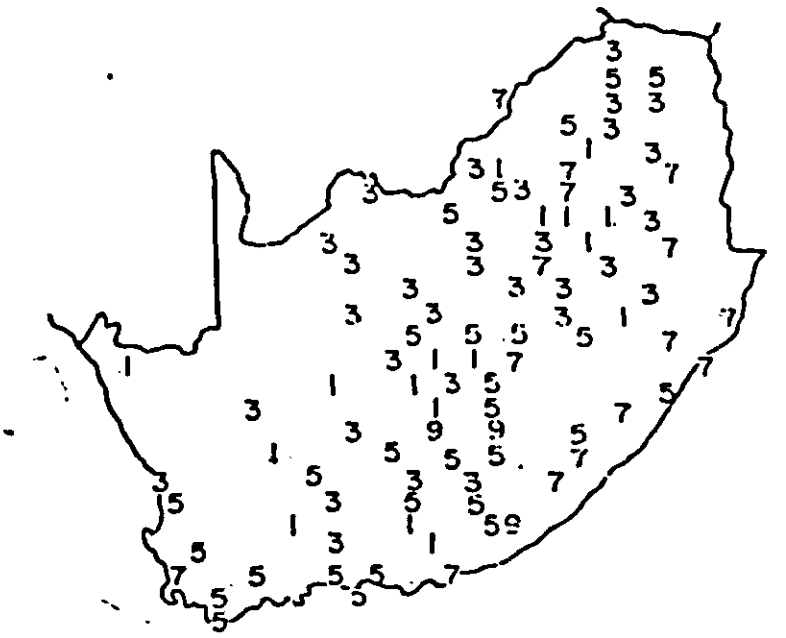


FIGURE 2.11.3

Optimum number of parameters to estimate the probability of a wet day given that the preceding day was dry for each of 100 test stations.



3. THE DISTRIBUTION OF RAINFALL ON DAYS WHEN RAIN OCCURS

This chapter concerns the second component of the rainfall model, namely the distribution of rainfall depths on those days when rain occurs. Rainfall depths are generally recorded to the nearest tenth of a millimeter in South Africa and so the smallest non-zero reading is given as 0,1 mm. This provides a convenient cut-off to distinguish wet and dry days and in what follows we will define a wet day as one in which at least 0,1 mm was recorded. We note that this definition affects both components of the rainfall model and it is therefore necessary to be consistent should one wish to change this boundary. Furthermore the models which are usually fitted to rainfall depths (when rain occurs) have a lower bound of zero and so if the selected boundary is much larger than 0,1 mm, say 2 mm or more, it will be necessary to model the differences between the observed depths and the boundary value rather than the observed depths themselves.

The distribution of rainfall depths on days when rain occurs exhibits the same type of seasonal behaviour as say the probability of having a wet day! For example the average rainfall amount (taken over wet days only) on 19 February is different, in general, to the average on 13 April. The same holds for the variance and many other aspects of this distribution except perhaps the coefficient of variation which seems to be approximately constant over the year. One has to use a different distribution for each day of the year and the simplest way of doing this is by fitting a single family of distributions and then allowing the parameters to change over the year. It is known that the distribution for each day is positively skewed but there is otherwise very little known to help one decide which particular family would be appropriate. Obvious candidates

are the lognormal, gamma, extreme (type 1) and Weibull. An interesting possibility (Woolhiser and Pegram 1979) is to use a mixture of two exponential distributions.

The selection of a suitable family presents a number of extra difficulties when one is dealing with a large number of records, as we were attempting to do. We began by fitting the lognormal family because the maximum likelihood estimators of its (seasonal) parameters were the easiest to derive and, more significantly, require much less computation to implement. This model appeared to fit a number of rainfall records reasonably well and so was applied to a further 100 test stations, partly in order to decide on how many parameters should be used in the model when it was fitted to the full set of 2550 records. On validating the model for these 100 stations it was found that in some cases it simply did not fit the data and so we were forced to discard it. The Weibull was later found to provide better fits, but as a result of our experience with the lognormal we decided that it would be unwise to settle for any one particular model for all 2550 records because if it later were to turn out that the model did not fit a particular station then our estimates would be useless for the purpose of fitting an alternative model. On the other hand model selection and validation are time consuming and costly exercises, particularly in this application, and we could not afford to handle each station separately. One possible solution is to deal with the problem on a regional basis, i.e. identify the regions where each of the models is expected to fit, but with this approach too a large number of records would have to be analysed in order to determine the regional boundaries accurately. The simplest solution (and probably the safest in the long run) is to initially not fit any model at all, but rather to fit the first two moment functions of the distribution. These can then be

used to estimate the parameters (by the method of moments) to any desired two-parameter model. As mentioned above, we have found that the Weibull distribution provides a good fit for the records which we tested. Should it happen, however, that some other two-parameter family is found to fit the data better, perhaps for some particular cluster of stations, then the results given in this report can equally easily be used to estimate the parameters of the alternative family. It is also possible to fit different families to a single record, say one for the rainy season and a second for the dry season.

As we eventually did not make use of the lognormal distribution the material relating to it has been relegated to an appendix. We have not left this out of the report altogether because it provides an illustration of how to go about fitting a seasonal model by the method of maximum likelihood, and a discussion on the question of how many terms should be used to fit the parameter functions of the model. Corresponding results for the gamma, Weibull, extreme (type I), normal and mixture of two exponential distributions can be derived along similar lines and will not be given here. We note that estimation based on maximum likelihood requires more computation than that based on the method of moments. For the lognormal family the difference is not large but it is very large for the Weibull family because no suitable sufficient statistics can be given (cf. the sufficient statistics m(T) and s(T) for the lognormal case in Appendix 1). Consequently even if we had been certain that the Weibull family was suitable for all 2550 stations we would nevertheless have been forced to estimate its parameters by the method of moments because the alternative of using maximum likelihood is too expensive.

3.1 ESTIMATING THE MEAN AND COEFFICIENT OF VARIATION

Suppose that the year is divided into NT intervals (e.g. 52 weeks, 365 days, etc...) denoted by T = 1,2,...,NT. Let M(T) represent the number of times that it rained in period T and R(I,T), I = 1,2,...,M(T), the rainfall depth on the Ith year that it rained in period T. (To be consistent with the notation in Chapter 2 we should really use NR(T) instead of M(T), but the latter is briefer and is unlikely to lead to any confusion.) Let $\mu(T)$ represent the mean rainfall per rainy day in period T = 1,2,...,M(T) and let C denote the coefficient of variation which we assume to be constant for all T (see section 3.2 for justifications for this assumption).

It is undesirable to estimate $\mu(T)$ separately for each T for the reasons that were outlined in Chapter 2 when we were discussing estimators for $\pi(T)$. Instead we will again make use of the Fourier Series representation:

$$\mu(T) = \sum_{i=1}^{NT} \mu_i \, \phi_i(T) \quad , \qquad T = 1,2,\ldots,NT \ ,$$

where $\phi_i(T)$ is defined in section 2.4 and $\mu_1, \mu_2, \ldots, \mu_{NT}$, are the Fourier coefficients of $\mu(T)$. Truncating the series to L terms we define

$$\mu(T,L) = \sum_{i=1}^{L} \mu_i \, \phi_i(T) \quad , \qquad T = 1,2,\ldots,NT \ ,$$
$$L < NT \ .$$

The approximation we then make is

$$\mu(T,L) \approx \mu(T) \quad , \qquad T = 1,2,\ldots,NT \ .$$

The effects of varying L are analogous to those given in section 2.4 and so will not be discussed in detail here.

Briefly, L must be large enough for the above approximation to be accurate but as small as possible in order to minimise the uncertainties associated with sampling variation. We suppose for the moment that L is fixed.

The simplest way to estimate $\mu_1, \mu_2, \ldots, \mu_L$ is to apply the method of ordinary least squares on the observed means for each period:

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

where m(T) is not defined if M(T) = 0, i.e. it never rained in period T. If none of the M(T) are zero then the minimum of

$$\sum_{T=1}^{NT} (m(T) - \mu(T,L))^2 \qquad (2)$$

is achieved using

$$\hat{\mu}_i = K(i) \sum_{T=1}^{NT} m(T)\, \phi_i(T) \quad , \qquad i = 1,2,\ldots,L \ ,$$

where $K(1) = 1/NT$ and $K(i) = 2/NT$ for $i = 2,3,\ldots,L$.

The solution is less simple if some of the M(T) are equal to zero, a situation that frequently occurs in arid regions. Approximations to the least squares estimators are given by

$$\tilde{\mu}_i = K(i) \sum_{\substack{T=1 \\ M(T)>0}}^{NT} m(T)\, \phi_i(T) \qquad (3)$$

where

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

This method of estimating $\mu(T)$ is attractive in so far as it requires relatively little computation, but unfortunately (1) is not a satisfactory criterion on which to base estimation. Each $m(T)$, $T = 1,2,\ldots,NT$, in (1) is given equal weight irrespective as to whether it represents the average of 1 or 100 observations and consequently those periods T which experience relatively little rainfall have a disproportionately large influence in final estimates of $\mu(T)$. The estimators given in (3) do however supply useful initial estimates for the two methods which follow.

To overcome the difficulty associated with (2) one can consider the following criterion:

$$\sum_{T=1}^{NT} \sum_{I=1}^{M(T)} (R(I,T) - \mu(T,L))^2 \qquad (4)$$

This must be minimised with respect to the μ_i, $i = 1,2,\ldots,L$.

Here the sum of the squares of all the individual deviations is considered and not only the deviations from the sample means. A further refinement is to use the method of weighted least squares, i.e. base estimation on the criterion given by

$$\sum_{T=1}^{NT} \sum_{I=1}^{M(T)} \left\{\frac{R(I,T)-\mu(T,L)}{C\mu(T,L)}\right\}^2 \qquad (5)$$

where C is the (constant) coefficient of variation. It does not require much more computation to minimise (5) with respect to $\mu_1, \mu_2,\ldots,\mu_L$ than it does to minimise (4) and iterative methods have to be used in both cases. Although (5) may be preferable to (4), in theory we would hesitate to recommend it in our application because it is more sensitive to outliers in the observations when $\mu(T)$ is small. A single storm event in the dry season, for example, can substantially influence the estimates which are based on (4), but even more so those based on (5). It is well-known that outliers

("unusual" observations) are especially problematic when least squares estimators are used, and although so-called "robust" techniques have been proposed (see e.g. Huber 1977) they involve rather more computation, particularly in our application. Estimation based on (4) is more robust than that based on (5) and consequently we recommend the former criterion.

We now show how one can compute the values of $\mu_1,\mu_2,\ldots,\mu_L$ which minimise (4). The problem of minimising (5) can be solved along similar lines and will not be discussed further. We will denote the sum of squares given in (4) by $S(\mu)$, where $\mu = (\mu_1,\mu_2,\ldots,\mu_L)^T$ denotes the vector of parameters. It is straightforward to show that

$$S(\mu) = s + \sum_{T=1}^{NT} M(T)(m(T) - \mu(T,L))^2 \qquad (6)$$

where $m(T)$ is defined in (1) for $M(T) \neq 0$, and in what follows we define it to be zero if $M(T) = 0$; and

$$s = \sum_{T=1}^{NT} \sum_{I=1}^{M(T)} (R(I,T) - m(T))^2 .$$

To minimise (6) we set its partial derivatives with respect to the parameters equal to zero:

$$\frac{\partial S(\mu)}{\partial \mu_i} = -2\sum_{T=1}^{NT} M(T)(m(T) - \mu(T,L))\,\phi_i(T) = 0, \quad i = 1,2,\ldots,L.$$

These L equations can be solved using the Newton-Raphson iteration method for which the second partial derivatives are required. These can also be used to estimate the standard error of the estimates. One has

$$\frac{\partial^2 S(\mu)}{\partial \mu_i \partial \mu_j} = 2\sum_{T=1}^{NT} M(T)\,\phi_i(T)\,\phi_j(T) \quad , \quad i,j = 1,2,\ldots,L .$$

We now give an outline of an algorithm to carry out the estimation. Let the ith element of the (column) vector $f^{(k)}$

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

$$f_i^{(k)} = -\sum_{T=1}^{NT} M(T)(m(T) - \mu^{(k)}(T,L))\phi_i(T) \quad , \quad i = 1,2,\ldots,L \quad (7)$$

$$F_{ij}^{(k)} = \sum_{T=1}^{NT} M(T)\,\phi_i(T)\phi_j(T) \quad , \quad i,j = 1,2,\ldots,L \quad (8)$$

where

$$\mu^{(k)}(T,L) = \sum_{i=1}^{L} \mu_i^{(k)}\,\phi_i(T) \quad , \quad T = 1,2,\ldots,NT \quad (9)$$

and $\mu_1^{(k)},\mu_2^{(k)},\ldots,\mu_L^{(k)}$ are the estimates of the parameters at the kth iteration.

ALGORITHM

STEP 1 Obtain initial estimates $\mu_1^{(o)},\ldots,\mu_L^{(o)}$ using (3) and compute $\mu^{(o)}(T,L)$ using (9). Set k = 0.

STEP 2 Compute $f^{(k)}$ using (7) and $F^{(k)}$ using (8).

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, for example test 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 go to Step 2.

To speed up the algorithm one should make use of the fact that the matrix $F^{(k)}$ is symmetric, i.e. it is only necessary to compute the entries of the upper triangle of the matrix. Subroutines to solve linear equations directly

are generally more efficient than those which compute the inverse of a matrix, and it is therefore recommended that the equations in Step 3 be solved directly rather than by pre-multiplying $f^{(k)}$ by the inverse of $F^{(k)}$.

Having estimated $\mu(T)$ it is quite easy to estimate the coefficient of variation, C. We note that

$$E \sum_{T=1}^{NT} \sum_{I=1}^{M(T)} \{R(I,T) - \mu(T)\}^2 = C^2 \sum_{T=1}^{NT} M(T)\mu(T)^2$$

An estimator of C is obtained by replacing $\mu(T)$ by $\hat{\mu}(T)$ and omitting the expectation, i.e.

$$\hat{C} = \left[\left[\sum_{T=1}^{NT} \sum_{I=1}^{M(T)} \{R(I,T) - \hat{\mu}(T)\}^2 \right] \Big/ \left[\sum_{T=1}^{NT} M(T)\hat{\mu}(T)^2 \right] \right]^{\frac{1}{2}}.$$

The variance of $\hat{C}$ is a function of up to the 4th order moment functions of the R(I,T) and is rather complicated. To estimate the standard error of $\hat{C}$ using such an expression would require one to decide how many Fourier terms should be fitted to the 3rd and 4th order moment functions. Bootstrap methods (cf. "Assessing the Risk of Deficiencies in Streamflow") would seem to be the only viable (though costly) alternative.

3.2 SELECTING THE NUMBER OF PARAMETERS

In theory the methods described in Linhart and Zucchini (1986) could be used to
select L, the number of terms in the approximation of $\mu(T)$. For example

$$\Delta(L) = \sum_{T=1}^{NT} (\mu(T) - E\hat{\mu}(T,L))^2 \quad , \quad L = 1,3,5,\ldots,$$

would be a suitable discrepancy on which to base selection. A complicating feature of our application is that M(T) can be zero for some of the periods and so in practice only approximately unbiased estimators are available to construct the corresponding criterion. For most stations in the arid and semi-arid regions there are several days with M(T) = 0, i.e. days of the year which have been dry over the whole period of observation. We cannot determine how reliable the criterion would be in such situations. On the other hand it is rather difficult to derive a criterion which takes this complicating feature into account and, even if this could be done, it is likely that the result would be cumbersome and not easy to compute.

If one is prepared to make distributional assumptions then selection criteria are relatively easy to derive, for example those based on the Kullback-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 $\mu(T)$. We fitted 100 test stations using the lognormal distribution and found L = 5 to be a suitable value for the mean of the logs and L = 1 for the standard deviation of the logs (cf. Appendix 1). These results together with an analysis of the values of criterion (4) for different L led us to decide that a 5-term approximation for $\mu(T)$ would be the most appropriate.

The fact that L = 1 turned out to be the best choice for estimating the standard deviation of the logs of the observations is strong evidence in favour of the assumption that the coefficient of variation is constant. This point is discussed in Appendix 1 and supports the findings of Stern and Coe (1984) and Yevjevich and Dyer (1983).

3.3 FITTING THE WEIBULL FAMILY

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. We denote the scale parameter by $\alpha(T)$, $T = 1,2,\ldots,NT$ and the shape parameter by β. The latter does not depend on T because it is a function of C but not of $\mu(T)$:

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

We require β as a function of C and as no closed expression of this function is available we have derived a rational function approximation (cf. Appendix 2). To estimate β given $\bar{C}$ one uses

$$\hat{\beta} = \frac{339,5410 + 148,4445\bar{C} + 192,7492\bar{C}^2 + 22,4401\bar{C}^3}{1 + 257,1162\bar{C} + 287,8362\bar{C}^2 + 157,2230\bar{C}^3} \quad .$$

Having estimated β one makes use of the relationship

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

to obtain the estimator

$$\hat{\alpha}(T) = \hat{\mu}(T)/\Gamma(1+1/\hat{\beta}) \quad , \qquad T = 1,2,\ldots,NT \; .$$

An algorithm to compute an approximation to the gamma function is given in Appendix 1 of the separate report "Assessing the Risk of Deficiencies in Streamflow". Clearly one should only compute the above gamma function once and not for each T.

It is quite easy to use the same approach in order to estimate the parameter functions of any other 2-parameter

family, e.g. lognormal, gamma, etc ... and so we will not discuss these here.

4. ALGORITHMS

This chapter describes the algorithms to implement the theory discussed in the previous two chapters. Except for one or two references to the algorithms in Appendix 2 and Appendix 3, the chapter is designed to be self-contained, i.e. it should be unnecessary to have to refer back to the theory in order to code the required computer programs. The algorithms described here are well within the capabilities of a typical desk-top microcomputer. They have all been implemented on an IBM PC microcomputer.

This chapter does not follow the standard prose style but has been written in "note form" for the convenience of the user.

There are four groups of algorithms which are given in sections 4.1 to 4.4. They relate to

4.1 the generating of artificial rainfall sequences for a given set of model parameters,

4.2 preparing historical records for parameter estimation,

4.3 estimating the parameters of the probability of the wet/dry sequences model,

4.4 estimating the mean value function of the rainfall depths on wet days and the coefficient of variation.

The parameters of 2550 stations have already been estimated (cf. Appendix 6) so for any of these stations it is only necessary to carry out the algorithm given in 4.1. Note that the parameter estimates given in Appendix 6 are for rainfall measured in units of one-tenth of a millimetre.

4.1 GENERATING ARTIFICIAL RAINFALL SEQUENCES

This section describes the algorithm to generate arbitrarily long artificial daily rainfall sequences for a station whose model parameters have been estimated. For stations not covered in Appendix 6 it is necessary to first estimate the model parameters using the algorithms described in the following three sections.

We will give the notation and then refer to an example to illustrate it:

AMWW(I) is the Ith amplitude for the probability that a wet day follows a wet day, I = 0,1,2,
PHWW(I) is the Ith phase for this probability, I = 1,2,

AMDW(I) is the Ith amplitude for the probability that a wet day follows a dry day, I = 0,1,2,
PHDW(I) is the Ith phase for this probability, I = 1,2,

AMM(I) is the Ith amplitude for the mean rainfall on wet days, I = 0,1,2,
PHM(I) is the Ith phase for this mean, I = 1,2,

C is the coefficient of variation.

For example the first station given in Appendix 6 is PETERS GATE (2069) and has

AMWW(0) = -0,5516	AMWW(1) = 0,4532	AMWW(2) = 0,1241
	PHWW(1) = 194,88	PHWW(2) = 133,80
AMDW(0) = -1,6836	AMDW(1) = 0,3345	AMDW(2) = 0,1050
	PHDW(1) = 184,03	PHDW(2) = 82,03
AMM(0) = 68,18	AMM(1) = 23,98	AMM(2) = 4,51
	PHM(1) = 198,20	PHM(2) = 132,57

C = 1,2533

Notes:

(a) The first seven numbers given for each station in Appendix 6 (which are estimates of the parameters of the probability of a wet day) are *not* required here and should be ignored.

(b) Should one decide to use more than the above number of estimates for the amplitudes and phases, e.g. to also use AMWW(3) and PHWW(3) then the algorithms given below will have to be modified in the obvious way. In order to increase the number of amplitudes and phases one would of course have to estimate them first.

The following arrays are required:

PWW(T) contains the probability that day T is wet given that it is wet on day T-1, T = 1,2,...,365,

PDW(T) contains the probability that day T is wet given that it is dry on day T-1, T = 1,2,...,365,

M(T) contains the scale parameter of the Weibull distribution, T = 1,2,...,365,

GR(N,T) contains the generated rainfall depths for year N, day T ; T = 1,2,...,365, N = 1,2,...,NG in units of 1/10 mm.

where NG is the required number of years of generated record.

The lower case letters in brackets on the right margin of the algorithm refer to notes following the algorithm.

ALGORITHM 4.1

```
STEP 1   INPUT     AMWW(I), AMDW(I), AMM(I), I = 0,1,2
                   PHWW(I), PHDW(I), PHM(I), I = 1,2
                   and C.
STEP 2   COMPUTE   B                                              (a)

STEP 3   COMPUTE   G                                              (b)

STEP 4   SET       GI = 1/G
                   BI = 1/B

STEP 5   COMPUTE   PWW(T), PDW(T), M(T), T = 1,2,...,365          (c)

STEP 6   SET       P = PDW(365)/(1-PWW(365) + PDW(365))

STEP 7   IF        RND < P  THEN SET IND = 1                      (d)
                            ELSE SET IND = 0

STEP 8   LOOP OVER YEARS : N = 1,2,...,NG

STEP 9   LOOP OVER DAYS  : T = 1,2,...,365

STEP 10  IF        IND = 1 THEN GO TO  STEP 11
                           ELSE GO TO  STEP 12

STEP 11  IF        RND < PWW(T)  THEN SET IND = 1                 (d)
                                 ELSE SET IND = 0
         GO TO  STEP 13

STEP 12  IF        RND < PDW(T)  THEN SET  IND = 1
                                 ELSE SET  IND = 0

STEP 13  IF        IND = 1  THEN SET  GR(N,T) = M(T)*
                                      (-LOG(RND))**BI           (d)
                            ELSE SET  GR(N,T) = 0
```

STEP 14 END OF T LOOP

STEP 15 END OF N LOOP

STEP 16 OUTPUT ARRAY GR.

Notes:

(a) the shape parameter of the Weibull distribution, B, is computed from the coefficient of variation, C, using the algorithm given in Appendix 2.

(b) G = Γ(1+1/B) where Γ denotes the gamma function and is available as a standard subprogram on most large computers. An algorithm to compute it is given in Appendix 1 of the report "Assessing the Risk of Deficiencies in Streamflow".

(c) The following steps are required to compute PWW(T), T = 1,2,...,365:

STEP 5.1 SET W = 0,01721421
STEP 5.2 *LOOP OVER* DAYS : T = 1,2,...,365
STEP 5.3 SET LOGIT= AMWW(0)
+ AMWW(1)*COS(W*(T-1-PHWW(1)))
+ AMWW(2)*COS(2*W*(T-1-PHWW(1))
STEP 5.4 SET PWW(T) = EXP(LOGIT)/(1+EXP(LOGIT))
STEP 5.5 *END OF T LOOP*

By using its corresponding phases and amplitudes the array PDW(T), T = 1,2,...,365 is computed in the same way.

To compute M(T) one replaces Step 5.3 and 5.4 by

M(T) = (AMM(0))
+ AMM(1)*COS(W*(T-1-PHM(1)))
+ AMM(2)*COS(2*W*(T-1-PHM(2)))*GI

(d) RND denotes a uniformly distributed random deviate in the interval (0,1) and is available on practically all computers. Note that at each stage in algorithm 4.1 where RND appears a fresh random deviate should be generated.

4.2 PREPARING THE DATA FOR PARAMETER ESTIMATION

In this section we describe an algorithm to extract from historical rainfall data the information required by the parameter estimation algorithms. It is assumed that the rainfall record is available in an array:

DEPTH(J,T), where T = 1,2,...,NT represents the period (e.g. day) in the year and J = 1,2,...,NY represents the year.

The daily rainfall records as supplied by the Department of Transport Weather Bureau are not in this form and so if data is obtained from this source then it will be necessary to reorganise it into an array as specified above.

Remarks about NT and NY

(a) For daily data NT = 365,
pentad data NT = 73,
weekly data NT = 52,
monthly data NT = 12.

(b) To overcome the irregularity arising during leap years one can add the precipitation for 29 February to that for 1 March, or some other day.

(c) For weekly data one of the "weeks" will have to consist of eight days and a second "week" will also have to consist of eight days on leap years. The table below

gives the recommended dates for the start and end of the 52 "weeks" of the year. This arrangement has the advantage that both 1 January and 1 October (the first day of the "water year") occur at the start of a week.

DATES FOR THE RECOMMENDED "WEEK" BEGINNINGS

WEEK	BEGIN	WEEK	BEGIN
1	1 Jan	27	2 Jul
2	8 Jan	28	9 Jul
3	15 Jan	29	16 Jul
4	22 Jan	30	23 Jul
5	29 Jan	31	30 Jul
6	5 Feb	32	6 Aug
7	12 Feb	33	13 Aug
8	19 Feb	34	20 Aug
9[1]	26 Feb	35	27 Aug
10	5 Mar	36	3 Sep
11	12 Mar	37	10 Sep
12	19 Mar	38	17 Sep
13	26 Mar	39	24 Sep
14	2 Apr	40[2]	1 Oct
15	9 Apr	41	9 Oct
16	16 Apr	42	16 Oct
17	23 Apr	43	23 Oct
18	30 Apr	44	30 Oct
19	7 May	45	6 Nov
20	14 May	46	13 Nov
21	21 May	47	20 Nov
22	28 May	48	27 Nov
23	4 Jun	49	4 Dec
24	11 Jun	50	11 Dec
25	18 Jun	51	18 Dec
26	25 Jun	52	25 Dec

[1] 8 days on leap years

[2] 8 days

(d) For monthly data the number of days in each month varies and this variation often results in an unnecessary increase in the number of parameters fitted to the model. It is therefore recommended that instead of using monthly totals the average daily precipitation is used for each month. That is, the January totals for each divided by 31, the February totals by 28 (or 29 on leap years) etc.

Gaps in the historical record

GAPS IN THE RECORD MUST BE INDICATED BY "-1" OR SOME OTHER NEGATIVE NUMBER.

The majority of rainfall records in South Africa contain gaps. The methods described here have been designed to automatically deal with incomplete data (within limits). If the historical record is very short (less than 5 years) and a high proportion (50% or more) of the data is missing then it is not unlikely that the algorithms for parameter estimation will not converge. This will however depend on where the gaps occur in the data record, e.g. if they always occur over one part of the year then the estimation algorithm will usually not converge. But except for such extreme cases gaps will not lead to any problems in the estimation algorithm nor to any systematic bias in the estimates.

It is also possible to fill in gaps in the historical record by making use of records from neighbouring stations. Methods to do this are discussed in Part II of Report 3: "Estimating the missing values in rainfall records".

The start and end of the historical record

RECORDS SHOULD BEGIN IN PERIOD 1 AND END IN PERIOD NT.

For example if the calendar year is used then daily records should begin on 1 January and end on 31 December, and monthly records should begin in January and end in December. This restriction simplifies the algorithm and the program. It is NOT necessary to waste data in order to meet this requirement. For example if the original available daily record starts on 1.10.1920 and ends on 31.3.1964 then one should not discard the 3 months of 1920 record and 3 months of 1964 record, but instead code the days 1.1.1920-30.9.1920 and 1.4.1964-31.12.1964 as missing, i.e. set the values to "-1". The record is then regarded as starting on 1.1.1920 and ending on 31.12.1964.

Arrays required for parameter estimation

The following information is required for the parameter estimation programs and must be computed from the historical record:

NT — the number of periods in the year (e.g. 365 for daily data),

NY — the number of years of data (including the missing values),

For each T = 1,2,...,NT:

N(T) — the number of observations made in period T, (missing values are not counted),

NR(T) — the number of times it was wet (non-zero rain) in period T.

NW(T) — the number of times it was wet in period T-1 AND there was an observation in period T (i.e. there was not a gap on period T).

NWW(I) the number of times it was wet in period T-1 AND wet in period T.

ND(T) the number of times it was dry (zero rain) in period T-1 AND there was an observation in period T.

NDW(T) the number of times it was dry in period T-1 AND wet in period T.

R(I,T) the Ith non-zero rainfall depth in period T, I = 1,2,...,NR(T); T = 1,2,...,NT.

Notes:

(a) The period which precedes period 1 is NT. So for example with daily data NWW(1) is the number of times it was wet on day 365 of the preceding year and wet on day 1 of the current year.

(b) Note that NR(T) and NW(T+1) can be different particularly (but not exclusively) when there are gaps in the record. Clearly $NW(T+1) < NR(T)$. Also $ND(T+1) + NW(T+1) < N(T)$.

The above arrays are required by the estimation algorithms as follows:

(i) NW( ) and NWW( ) are required to estimate the parameters for the probability that a wet period follows a wet period.

(ii) ND( ) and NDW( ) are required to estimate the parameters for the probability that a wet period follows a dry period.

(iii) N( ) and DEPTH(,) are required to fit the parameters of the mean rainfall depth in a wet period and the coefficient of variation.

Although this is not required in order to generate artificial rainfall sequences one may also wish to compute the probability that period T is wet. For this one needs

(iv) N( ) and NR( ).

Computing the required arrays

Computation of the arrays N( ), NR( ), NW( ), NWW( ), ND( ), NDW( ) and R(,) is straightforward particularly if one is prepared to compute them one at a time. However such a procedure requires one to pass over the record several times and is therefore computationally inefficient. The following algorithm requires only one pass over the data.

An indicator, IND, is used to indicate the state on the previous period:

$$IND = \begin{cases} -1 & \text{indicates that the previous observation is missing} \\ 0 & \text{indicates that the previous period was dry} \\ 1 & \text{indicates that the previous period was wet.} \end{cases}$$

ALGORITHM 4.2

STEP 1 INPUT NT, NY, DEPTH (I,T), I = 1,2,...,NY: T = 1,2,..,NT

STEP 2 SET N( ), NR( ), NW( ), NWW( ), ND( ), NDW( ) to zero
IND = -1

STEP 3 LOOP OVER YEARS: I = 1,2,...,NY

STEP 4 LOOP OVER PERIODS: T = 1,2,...,NT

STEP 5 IF DEPTH(I,T) { = 0 GO TO STEP 6
{ > 0 GO TO STEP 7
{ < 0 GO TO STEP 8

STEP 6 SET N(T) = N(T)+1
IF IND = { 0 THEN SET ND(T) = ND(T)+1
{ 1 THEN SET NW(T) = NW(T)+1 AND SET IND = 0
{ -1 THEN SET IND = 0
GO TO STEP 9

STEP 7 SET NR(T) = NR(T)+1
R(NR(T),T) = DEPTH(I,T)
IF IND = { 0 THEN SET NDW(T) = NDW(T)+1 AND SET IND = 1
{ 1 THEN SET NWW(T) = NWW(T)+1
{ -1 THEN SET IND = 1
GO TO STEP 9

STEP 8 SET IND = -1
GO TO STEP 9

STEP 9 END OF T LOOP

STEP 10 END OF I LOOP

STEP 11 *LOOP OVER PERIODS:* T = 1,2,...,NT

STEP 12 *SET* N(T) = N(T) + NR(T)
ND(T) = ND(T) + NDW(T)
NW(T) = NW(T) + NWW(T)

STEP 13 *END OF T LOOP*

STEP 14 *OUTPUT ARRAYS* N, NR, NW, ND, NWW, NDW, R

Notes:

(a) When deciding how the output to this algorithm should be stored we recommend that (i) to (iv) above be kept in mind.

(b) For efficient use of computer storage and to reduce computing time the arrays in this algorithm (including DEPTH and R) should be defined as INTEGER rather than REAL arrays in the program.

4.3 ESTIMATING THE PROBABILITIES OF WET AND DRY SEQUENCES

The algorithm described here can be used to estimate the probabilities associated with (i), (ii) and (iv) in section 4.2 One uses the same algorithm but with different input data. We will use the generic notation MM(T) and M(T), T = 1,2,...,NT to represent the relevant arrays as follows:

(i)* When we are estimating the probability that a wet period follows a wet period then
MM(T) = NW(T) and M(T) = NWW(T), T = 1,2,...,NT.

(ii)* When we are estimating the probability that a wet period follows a dry period then
MM(T) = ND(T) and M(T) = NDW(T), T = 1,2,...,NT.

(iv)* When we are estimating the probability that period T is wet then
MM(T) = N(T) and M(T) = NR(T), T = 1,2,...,NT.

Algorithm 4.3 deals with the problem of estimating a fixed number of parameters and Algorithm 4.3A with that of selecting the best number of parameters to use.

The following variables and arrays are used in Algorithm 4.3:

L	number of parameters to be fitted (odd integer)
PAR(I), I = 1,2,...,L;	vector of parameters
AM(I), I = 0,1,...,K;	corresponding amplitudes, K = (L-1)/2
PH(I), I = 1,2,...,K;	corresponding phases
P(T) , T = 1,2,...,NT;	current estimates of probabilities
L(T), T = 1,2,...,NT;	current estimates of logits

*c.f. (i), (ii) and (iv), pages 66-67.

B(I), I = 1,2,...,L: vector of first partial derivatives
A(I,J), I,J = 1,2,...,L; matrix of second partial derivatives
D(I), I = 1,2,...,L; vector of solutions to linear equations

ITER current iteration
DELTA convergence criterion

T0,T1,T2,TA,TB temporary variables

PHI(I,T), I = 1,2,...,L, T = 1,2,...,NT matrix of Fourier terms.

ALGORITHM 4.3

STEP 1 *INPUT* L, NT, MM(T), M(T); T = 1,2,...,NT

STEP 2 *COMPUTE* PHI(I,T); I = 1,2,...,L; T = 1,2,...,NT

STEP 3 *COMPUTE* P(T), L(T); T = 1,2,...,NT

STEP 4 *COMPUTE* PAR(I), I = 1,2,...,L

STEP 5 *SET* ITER = 0

STEP 6 *COMPUTE* A(I,J), B(I): I,J = 1,2,...,L, DELTA

STEP 7 *SOLVE* AD = B (for D)

STEP 8 *SET* PAR(I) = PAR(I)-D(I), I = 1,2,...,L

STEP 9 *SET* ITER = ITER+1

STEP 10 *IF* DELTA > 0,00005 *THEN GO TO* STEP 6

STEP 11 *COMPUTE* AM(I); I = 0,1,...,K, PH(I): I = 1,2,...,K

STEP 12 *OUTPUT* AM(I); I = 0,1,2,...,K, PH(I); I = 1,2,...,K

Details

STEP 2 An algorithm to compute PHI is given in Appendix 3.

STEP 3 Here we need to compute initial estimates of the probabilities and logits:

STEP 3.1 LOOP OVER T = 1,2,...,NT
STEP 3.2 IF MM(T) > 0 THEN SET P(T) = M(T)/MM(T)
ELSE SET P(T) = -1 GO TO STEP 3.4
STEP 3.3 IF M(T) { = 0 THEN SET L(T) = -5
{ = MM(T) THEN SET L(T) = 5
{ ≠ 0, ≠ MM(T) THEN SET L(T)=LOG(P(T)/(1-P(T)))
STEP 3.4 END OF T LOOP

STEP 4 The initial estimates of the parameter vector are computed here:

STEP 4.1 LOOP OVER I = 1,2,...,L
STEP 4.2 SET T0 = 0, T1 = 0
STEP 4.3 LOOP OVER T = 1,2,...,NT
STEP 4.4 IF MM(T) = 0 THEN GO TO STEP 4.7
STEP 4.5 SET T0 = T0 + L(T)*PHI(I,T)
STEP 4.6 SET T1 = T1 + PHI(I,T)*PHI(I,T)
STEP 4.7 END OF T LOOP
STEP 4.8 SET PAR(I) = T0/T1
STEP 4.9 END OF I LOOP

STEP 6 We compute the first and second partial derivatives, and the sum of squares of absolute differences between the current and preceding values of the probabilities:

STEP 6.1 SET B(I), A(I,J) = 0; I = 1,2,...,L, J = 1,2,...,I
DELTA = 0
STEP 6.2 LOOP OVER T = 1,2,...,365

```
STEP 6.3 SET LOGIT = PAR(1)
STEP 6.4 LOOP OVER I = 2,3,...,L
STEP 6.5 SET LOGIT = LOGIT + PAR(I)*PHI(I,T)
STEP 6.6 END OF I LOOP
STEP 6.7 SET TO = EXP(LOGIT)
              PROB = TO/(1+TO)
              T1 = M(T)-MM(T)*PROB
              T2 = MM(T)*PROB/(1+TO)
              DELTA = DELTA + ABS(P(T)-PROB)
              P(T) = PROB
STEP 6.8 LOOP OVER I = 1,2,...,L
STEP 6.9 SET B(I) = B(I) + T1*PHI(I,T)
STEP 6.10 LOOP OVER J = 1,2,...,I
STEP 6.11 SET A(I,J) = T2*PHI(I,T)*PHI(J,T)
STEP 6.12 END OF J LOOP
STEP 6.13 END OF I LOOP
STEP 6.14 END OF T LOOP
STEP 6.15 SET A(I,J) = A(J,I); I = 1,2,...,L; J = I+1,I+2,...,L
```

STEP 7 Unless a subprogram to solve a system of L×L linear equations is available, it will be necessary to write one. The Gauss reduction method is suitable here.

STEP 9 It is recommended that the number of iterations (ITER) be limited to 50. Normally convergence occurs within about 7 iterations.

STEP 11 The notation ATN( ) is used to represent the principal value of the arctangent function (in the interval $-\pi/2$ to $\pi/2$). SQR( ) denotes the square root. To transform the parameters to their amplitude and phase representations one proceeds as follows:

```
STEP 11.1 SET PI = 3,141593
                   AM(0) = PAR(1)
                   K = (L-1)/2
STEP 11.2 LOOP OVER I = 1,2,...,K
STEP 11.3 SET TA = PAR(2*I)
              TB = PAR(2*I+1)
              AM(I) = SQR(TA*TA+TB*TB)
STEP 11.4 IF TA < 0 THEN SET PH(I) = ATN(TB/TA) + PI
STEP 11.5 GO TO STEP 11.10
STEP 11.6 IF TA = 0 THEN GO TO STEP 11.9
STEP 11.7 IF TB > 0 THEN SET PH(I) = ATN(TB/TA)
                    ELSE SET PH(I) = ATN(TB/TA) + 2*PI
STEP 11.8 GO TO STEP 11.10
STEP 11.9 IF TB > 0 THEN SET PH(I) = 0,5*PI
                    ELSE SET PH(I) = 1,5*PI
STEP 11.10 SET PH(I) = PH(I)*NT/(2*PI*K)
STEP 11.11 END OF I LOOP
```

Model Selection

To decide on how many parameters should be used in the model one can proceed as follows:

ALGORITHM 4.3A

STEP 1 INPUT LMAX

STEP 2 SET CRITO = 10**50

STEP 3 LOOP OVER L = 1,3,...,LMAX

STEP 4 APPLY ALGORITHM 4.3 TO COMPUTE P(T), T = 1,2,...,NT

STEP 5 COMPUTE CRIT = -LLK+L

STEP 6 IF CRIT < CRITO THEN SET LO = L
AND SET CRITO = CRIT

STEP 7 END OF L LOOP

STEP 8 OUTPUT LO

Details

STEP 1 LMAX must be an odd integer and need be no greater than 25 unless the historical record is exceptionally long.

STEP 3 Note that the values of L are incremented by 2.

STEP 4 The final values of P(T), T = 1,2,...,NT are required, i.e. those which were computed on the last iteration in Algorithm 4.3

STEP 5 The following steps are required to compute LLK:

STEP 5.1 SET LLK = 0
STEP 5.2 LOOP OVER T = 1,2,...,NT
STEP 5.3 IF MM(T) = 0 THEN GO TO STEP 5.5
STEP 5.4 SET LLK = LLK + M(T)*LOG(P(T)) + (MM(T)-M(T))*LOG(1-P(T))
STEP 5.5 END OF T LOOP

STEP 6 Normally the criterion decreases with L, reaches a minimum and then increases again. It is therefore not necessary to continue with the L loop once the criterion has begun to increase.

To avoid unnecessary computation the values of AM(I), I = 0,1,2,...,K and PH(I), I = 1,2,...,K which correspond to the current minimum criterion CRITO, should be stored. In this way one avoids having to re-estimate them once LO has been found.

4.4 ESTIMATING THE MEAN RAINFALL IN WET PERIODS

In this section we describe an algorithm to compute the mean rainfall in wet periods, where NT represents the number of periods in the year, e.g. 365 days, 52 weeks etc ... .

The following information is required in order to carry out the estimation:

NR(T); T = 1,2,...,NT
R(I,T); I = 1,2,...,NR(T); T = 1,2,...,NT
and the number of parameters to be fitted, L, which must be an odd integer. See section 4.2 for the definition of NR(T) and R(I,T).

Most of the arrays and variables used in Algorithm 4.4 are analogous to those used in Algorithm 4.3 We will only list the additional ones used here:

Q(T); T = 1,2,...,NT Average observed rainfall in each period
F(T); T = 1,2,...,NT Current estimate of the mean

To compute the parameters one proceeds as follows:

ALGORITHM 4.4

STEP 1 INPUT L,NT,NR(T),R(I,T); I = 1,2,..,NR(T),T = 1,2,...,NT

STEP 2 COMPUTE PHI(I,T); I = 1,2,...,L, T = 1,2,...,NT

STEP 3 COMPUTE Q(T); T = 1,2,...,NT

STEP 4 COMPUTE PAR(I); I = 1,2,...,L

STEP 5 SET ITER = 0

STEP 6 COMPUTE B(I), A(I,J); I,J = 1,2,...,L

STEP 7 SOLVE AD = B (for D)

STEP 8 SET PAR(I) = PAR(I)-D(I); I = 1,2,...,L

STEP 9 SET ITER = ITER+1

STEP 10 COMPUTE DELTA

STEP 11 IF DELTA > 0,0001 THEN GO TO STEP 6

STEP 12 COMPUTE AM(I); I = 0,1,...,K, PH(I); I = 1,2,...,K

STEP 13 OUTPUT AM(I); I = 0,1,...,K, PH(I); I = 1,2,...,K

Details

STEP 2 This is the same as STEP 2 in Algorithm 4.3

STEP 3 In this step the initial estimates for the mean are computed:

STEP 3.1 *LOOP OVER* T = 1,2,...,NT
STEP 3.2 *IF* NR(T) < 0 *THEN GO TO* STEP 3.8
STEP 3.3 *SET* TO = 0
STEP 3.4 *LOOP OVER* I = 1,2,...,NR(T)
STEP 3.5 *SET* TO = TO + R(I,T)
STEP 3.6 *END OF* I *LOOP*
STEP 3.7 *SET* Q(T) = TO/NR(T)
STEP 3.8 *END OF* T *LOOP*

STEP 4 This is the same as STEP 4 of Algorithm 4.3 with NR(T) replacing MM(T) and Q(T) replacing L(T)

STEP 6 The first and second partial derivatives are computed as follows:

STEP 6.1 *SET* B(I),A(I,J) = 0; I = 1,2,...,L; J = 1,2,...,I
STEP 6.2 *LOOP OVER* T = 1,2,...,NT
STEP 6.3 *SET* TO = PAR(1)
STEP 6.4 *LOOP OVER* I = 1,2,...,L
STEP 6.5 *SET* TO = TO + PAR(I)*PHI(I,T)
STEP 6.6 *END OF* I *LOOP*
STEP 6.7 *SET* F(T) = TO
STEP 6.8 *END OF* T *LOOP*
STEP 6.9 *LOOP OVER* T = 1,2,...,NT

STEP 6.10 IF NR(T) < 0 *THEN GO TO* STEP 6.17
STEP 6.11 LOOP OVER I = 1,2,...,L
STEP 6.12 SET B(I) = B(I) - NR(T)*(Q(T)-F(T))*PHI(I,T)
STEP 6.13 LOOP OVER J = 1,2,...,I
STEP 6.14 SET A(I,J) = A(I,J) + NR(T)*PHI(I,T)*PHI(J,T)
STEP 6.15 END OF J *LOOP*
STEP 6.16 END OF I *LOOP*
STEP 6.17 END OF T *LOOP*
STEP 6.18 SET A(I,J) = A(J,I); I = 1,2,...,L; J=I+1,I+2,...,L

STEP 7 See the remark made for STEP 7 of Algorithm 4.3.

STEP 10 The convergence criterion here is computed using
DELTA = ABS(B(1)) + ABS(B(2))+...+ABS(B(L))

STEP 12 This is computed in the same way as STEP 11 of Algorithm 4.3.

To compute the coefficient of variation, C, one simply uses

$$C = \left[\left[\sum_{T=1}^{NT} \sum_{I=1}^{NR(T)} \{R(I,T) - F(T)\}^2\right] \Big/ \left[\sum_{T=1}^{NT} NR(T)F(T)^2\right]\right]^{\frac{1}{2}}$$

where F(T) is the final value of this array after executing Algorithm 4.4.

No methodology to decide which is the best value of L is available. In general we would recommend that L should be set to 5 or 7. By examining the reduction in the average residual variance one can usually detect when nothing more is to be gained by increasing L. This variance is given by

$$V = \frac{1}{N} \sum_{T=1}^{NT} \sum_{I=1}^{NR(T)} (R(I,T)-F(T))^2$$

where $N = \sum_{T=1}^{NT} NR(T)$ is the total number of rainy days. Usually the value of V decreases sharply as L goes from 1 to 3 and then to 5. After that the decrease becomes slower. One must then decide after which L the decrease becomes small enough.

5. RAINFALL MODEL VALIDATION

Model validation is a retrospective view of the model where we evaluate its performance in order to assess its accuracy. In fact we are assessing to what extent the rationale of model structure and parameter estimation is successful in preserving the properties of the process in which we are interested. These properties of the historical process have a sampling variance of their own and our aim is to establish that the model preserves these within reasonable limits such that any simulated sequence is representative of an alternative but equally likely historical sequence.

For a daily rainfall model we need to establish that the properties of the daily rainfalls *and their sums* are preserved. In order to be viewed successfully the model must *simultaneously* preserve:

(a) The annual mean and variance and the distribution of annual totals and sums of annual totals.

(b) The monthly means and variances.

(c) The expected number of wet days as it varies seasonally.

(d) The runs characteristics of daily rainfalls as they vary seasonally.

(e) The distribution of n-day extreme rainfalls.

For a two-tier model such as we have here, with a Markovian structure for the probability of wetness and a univariate model for the distribution of wet day rainfall totals, each of the examinations proposed above emphasises a different aspect of the model. For example, the runs characteristics

of daily rainfalls assesses the validity of the model's Markovian structure, whilst the distribution of n-day extremes emphasises the performance of the univariate model chosen for rainfall depths and in particular the estimators of mean and variance. Six rainfall gauges which broadly represent the various ranifall/climate regions of South Africa were chosen for study. Table 5.1 lists them with their official Weather Bureau number, and their positions within the country are shown in Figure 5.1. The position of all 2550 stations to which the model is eventually fitted is shown in Figure A6.1.

Table 5.1

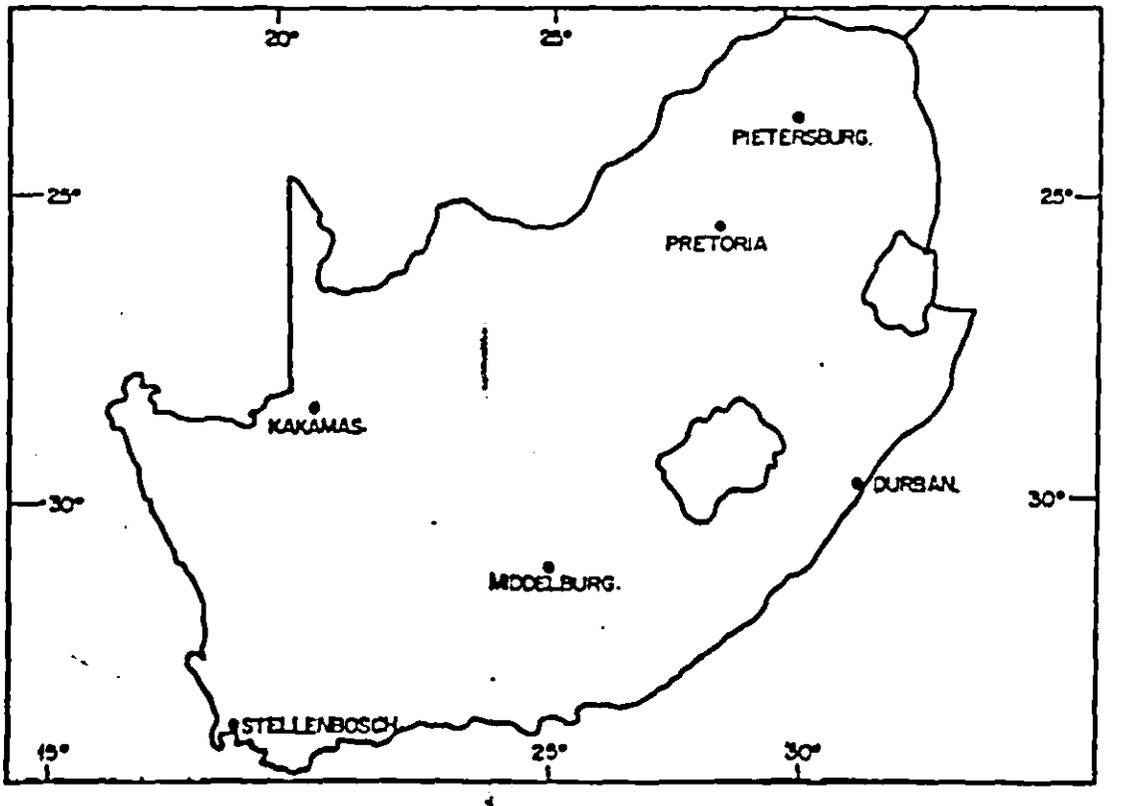
Station	Weather Bureau Index Number
Pretoria	513/404
Durban	240/891
Kakamas	282/166
Pietersburg	677/839
Stellenbosch	21/655
Middelburg (CP)	144/900

In each case the historical rainfall statistics of interest were compared with those estimated from 1000 years of simulated daily data.

Validation of annual properties

Table 5.2 shows a comparison of historical and simulated annual means and standard deviations. Both statistics are adequately preserved by the model. There would appear to be a slight underestimation of the annual standard deviation. Upon examination this was revealed to be due, when it occurred, to the model's inability to preserve the frequency of extreme n-day storm rainfalls in cases where these are

FIGURE 5.1 Locations of the six stations used for model validation.



associated with weather-generating processes that are distinct from those that generate the bulk of the rainfall. This point is dealt with when we consider model performance with regard to extremes.

TABLE 5.2 Comparison of historical and simulated annual means and standard deviations (mm)

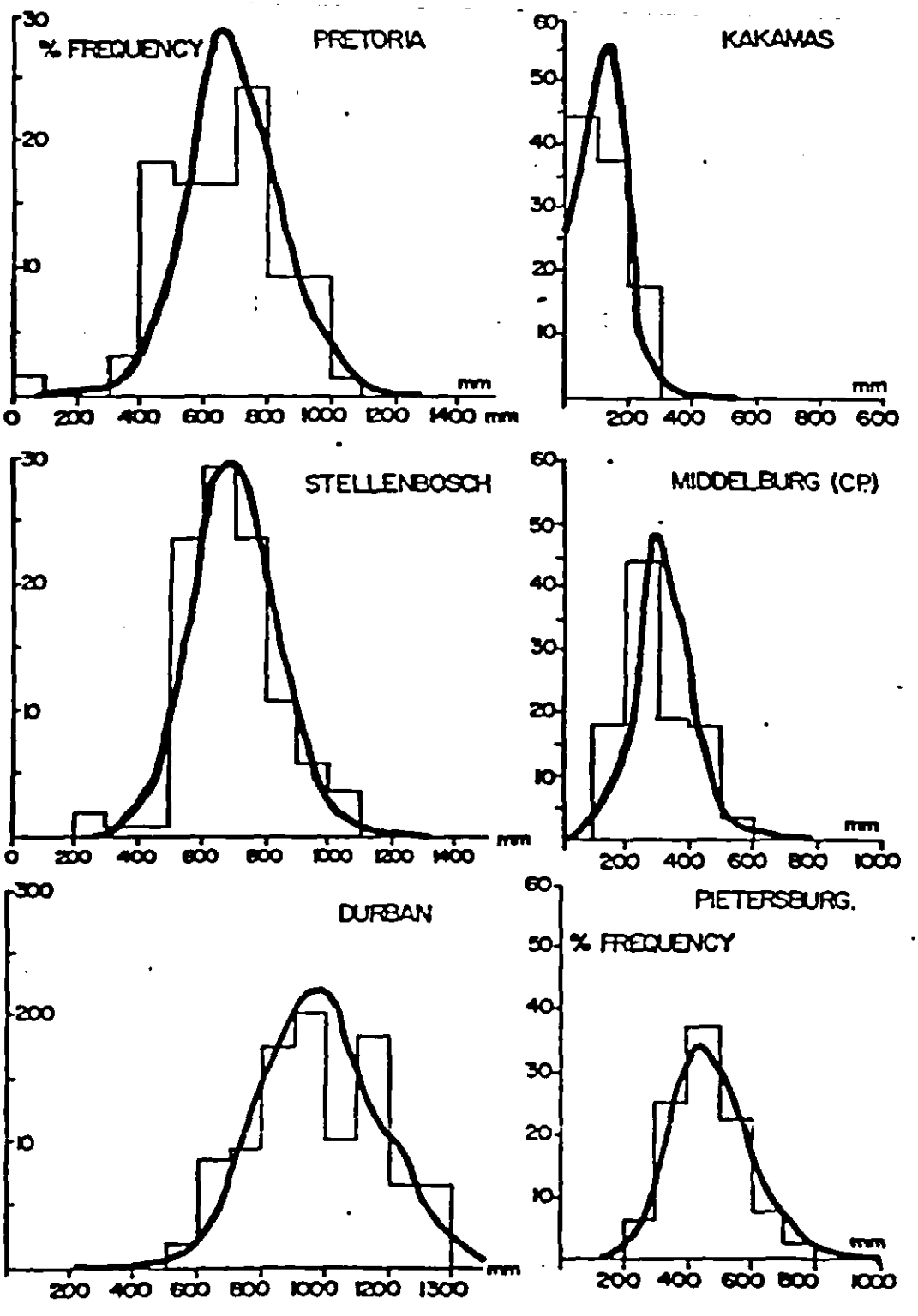
Station	Period of historical record	Mean	Standard Deviation	Simulation Mean	Simulation Standard Deviation
Pretoria	1905-1980	714	168	728	157
Durban	1871-1980	997	230	994	209
Stellenbosch	1895-1980	716	147	710	133
Kakamas	1936-1980	138	76	122	67
Middelburg (CP)	1917-1972	302	105	317	91
Pietersburg	1905-1952	461	112	475	124

Figure 5.2 shows the histograms of historical annual rainfall totals and their simulated distribution functions. As can be seen the overall shape of the distribution of annual totals is well preserved by the model.

In order to more closely examine the performance of the model in preserving the behaviour of low annual totals (i.e. the left tail of the distribution function) and in preserving runs of deficient annual rainfalls, the following strategy was adopted:

(a) Fit the gamma distribution with scale parameter β and shape parameter α by maximum likelihood to the historical annual totals. A univariate model selection criterion (cf. Appendix 2 of the "Assessing the risk of deficiencies in streamflow") showed the gamma model to be appropriate.

FIGURE 5.2 Histograms and simulated density functions of annual rainfall data.



(b) The reproductive property of the gamma model implies that the sum of n independently gamma distributed random variables with parameters α and β is also distributed as gamma and has parameters $n\alpha$ and β. Thus the two-year sums are distributed as $2\alpha,\beta$; the three-year sums as $3\alpha,\beta$ etc.

(c) Compute the distribution of the annual totals and their sums in this way and compare these with the distributions of 1000 simulated replicates of n-year totals, as obtained from the model. Such a comparison is given in Table 5.3. The results show a remarkable similarity even at the 1% level and imply that the model is at least as good as directly fitting an appropriate univariate model to the annual totals in order to investigate the distribution of annual run sums.

Validation of monthly properties

In many practical applications monthly rainfall sequences are required and it is important that the basic statistics of monthly rainfall are preserved by the model, in particular the means and standard deviations. Figures 5.3 and 5.4 show that these statistics are well preserved and Figure 5.5 shows that the mean number of wet days to be expected in each month is also maintained.

Validation of daily properties

A visual assessment of the fit of the truncated Fourier series to the various components of the model is available from Figures 5.6.1 to 5.10.2. The fits are generally excellent although the seasonal standard deviation for Stellenbosch, which was found to be an exception, is somewhat low.

TABLE 5.3 Distribution of n-year deficient rainfall totals estimated by fitting a gamma model to the historical data and by simulation

		PRETORIA		STELLENBOSCH		DURBAN		KAKAMAS		MIDDELBURG (CP)		PIETERSBURG	
	P(X>χ)	GAMMA	SIM	GAMMA	SIM	GAMMA	SIM	GAMMA	SIM	GAMMA	SIM	GAMMA	SIM
	.99	342	388	425	420	601	642	24	29	100	110	242	240
	.98	373	412	452	466	641	688	31	34	116	126	264	270
1	.95	423	477	496	507	705	740	42	43	143	156	299	318
YEAR	.90	471	526	537	529	764	786	55	51	171	170	332	363
	.80	534	571	589	606	841	815	73	62	210	202	375	397
	.50	669	675	698	722	1003	986	119	93	299	279	468	483
	.99	856	921	999	1021	1420	1486	91	84	296	318	604	642
	.98	907	969	1042	1089	1483	1522	104	98	326	349	639	671
2	.95	987	1026	1108	1145	1581	1616	126	122	374	390	694	720
YEAR	.90	1061	1146	1170	1221	1672	1702	149	141	420	434	745	779
	.80	1156	1192	1247	1293	1786	1799	179	168	481	489	811	832
	.50	1353	1377	1405	1460	2019	2202	248	238	613	596	946	996
	.99	1411	1450	1603	1588	2285	2224	171	166	520	517	994	1002
	.98	1476	1569	1658	1711	2365	2366	190	195	560	530	1039	1060
3	.95	1579	1684	1743	1797	2490	2511	221	219	624	660	1109	1100
YEARS	.90	1673	1772	1820	1874	2604	2592	251	242	684	713	1174	1196
	.80	1793	1861	1917	2006	2748	2723	291	279	762	788	1257	1280
	.50	2038	2069	2112	2309	3036	3020	378	351	927	954	1425	1445
	.99	1985	2116	2224	2306	3173	3156	259	266	759	819	1397	1311
	.98	2064	2182	2289	2367	3269	3227	283	279	807	880	1451	1396
4	.95	2185	2311	2389	2424	3416	3424	321	319	884	943	1534	1492
YEARS	.90	2297	2406	2480	2519	3550	3620	357	350	956	1007	1611	1570
	.80	2437	2523	2593	2634	3718	3681	405	391	1048	1111	1708	1682
	.50	2722	2769	2819	2900	4052	4191	508	487	1241	1292	1903	1999
	.99	2573	2612	2855	2962	4077	4266	353	360	1006	1144	1809	1800
	.98	2662	2809	2929	3016	4186	4319	380	388	1062	1207	1870	1899
5	.95	2801	2957	3042	3100	4353	4401	425	434	1150	1269	1966	2006
YEARS	.90	2928	3056	3145	3227	4505	4500	467	475	1233	1326	2053	2111
	.80	3087	3221	3273	3331	4693	4627	521	515	1338	1419	2162	2216
	.50	3406	3462	3526	3595	5069	4954	637	606	1556	1612	2381	2464

FIGURE 5.3 Simulated (.....) and historical (__) monthly mean rainfall.

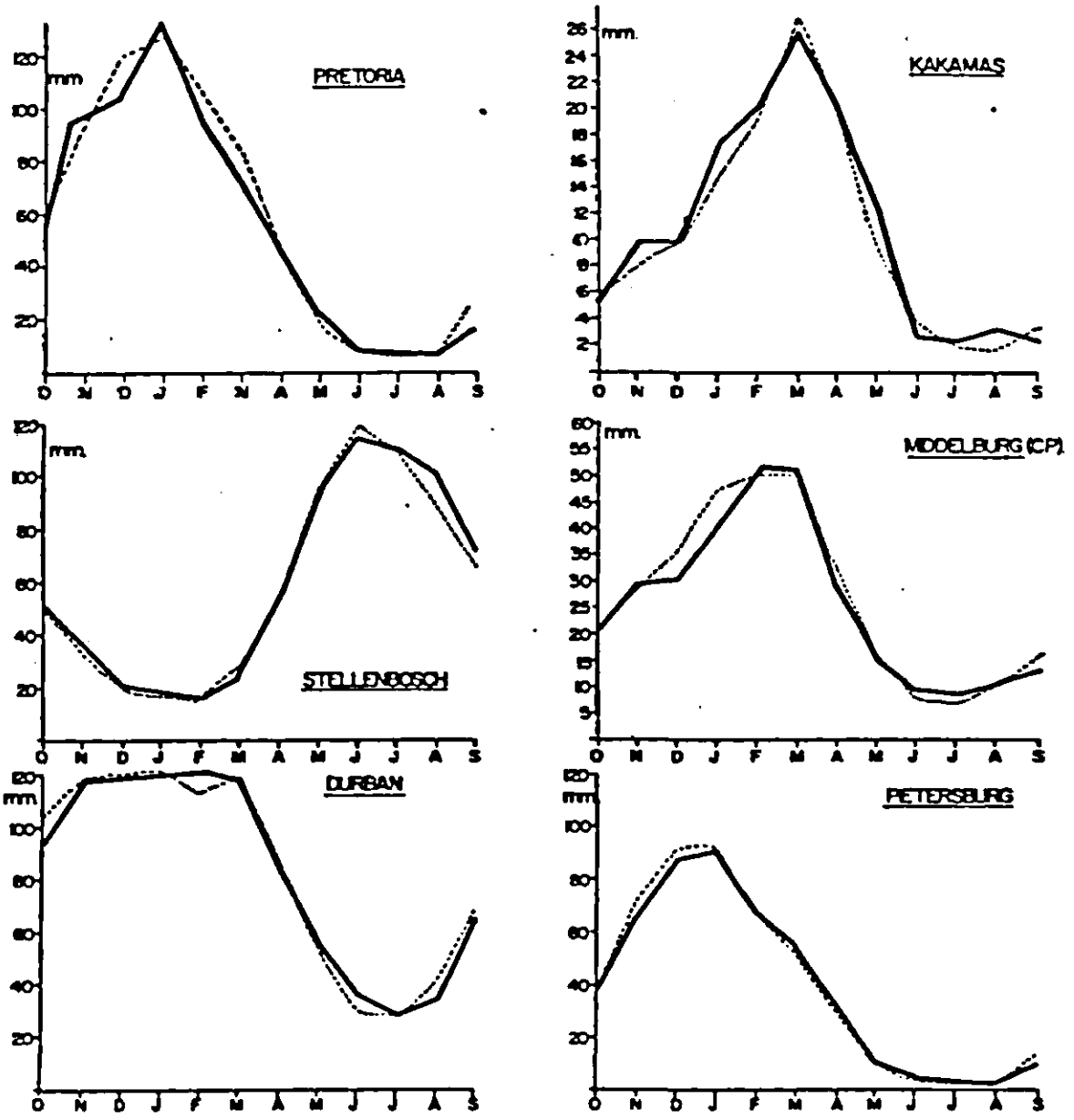


FIGURE 5.4 Simulated(-----) and historical (———) monthly standard deviations.

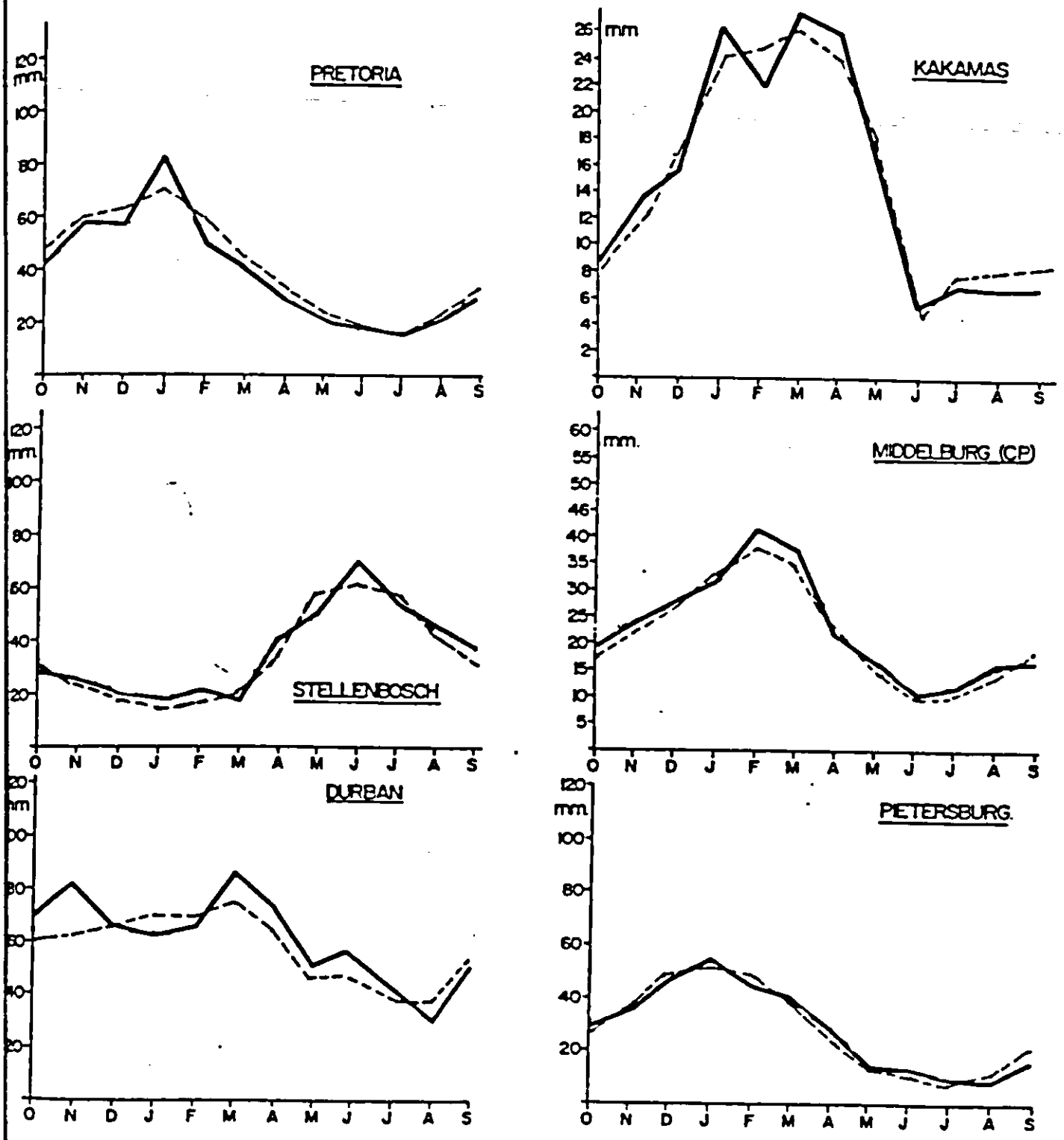
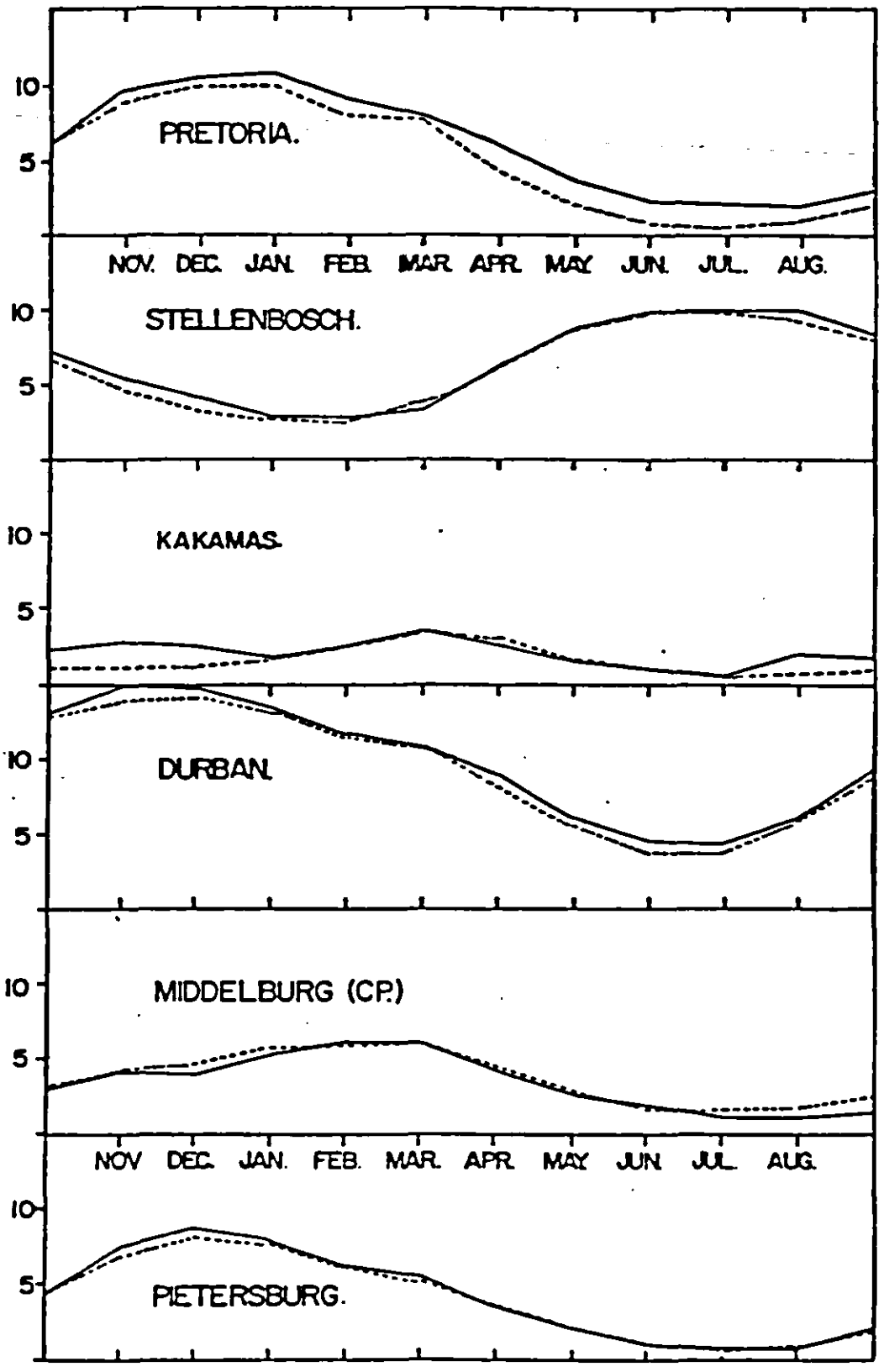
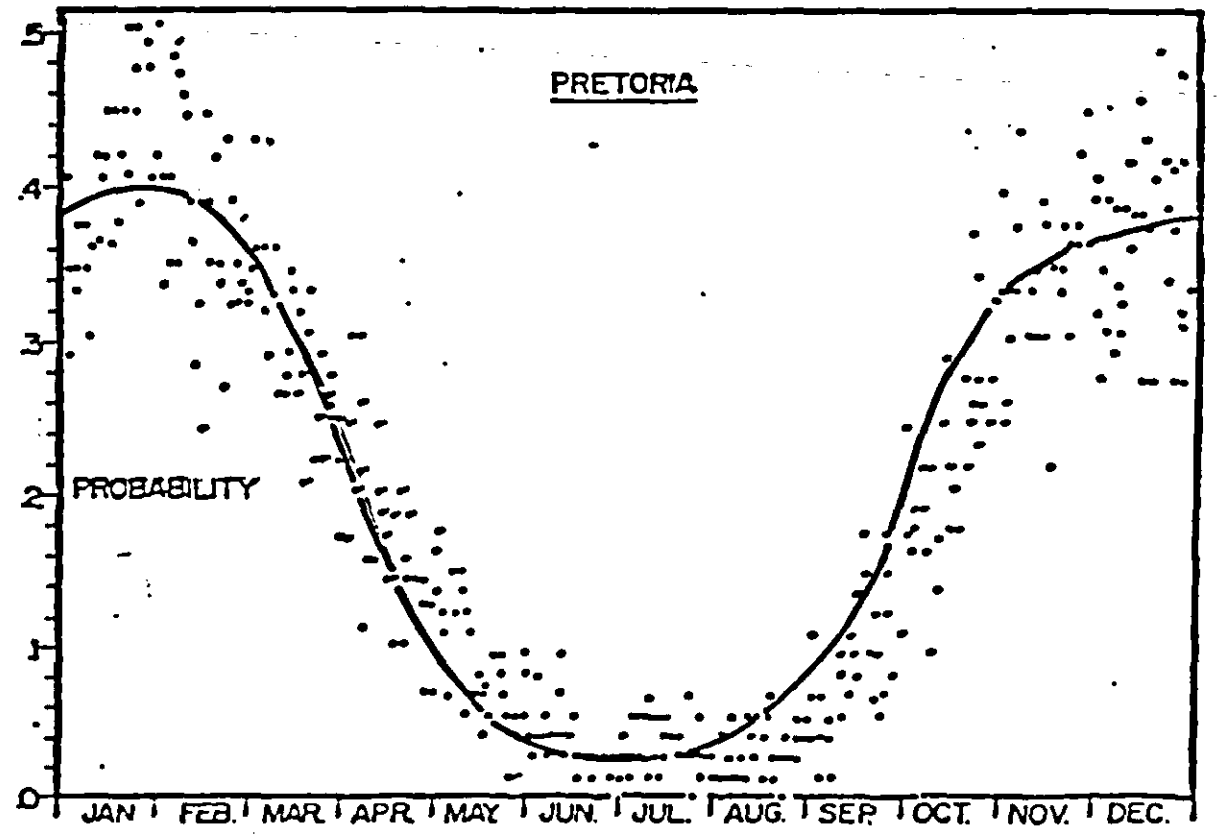


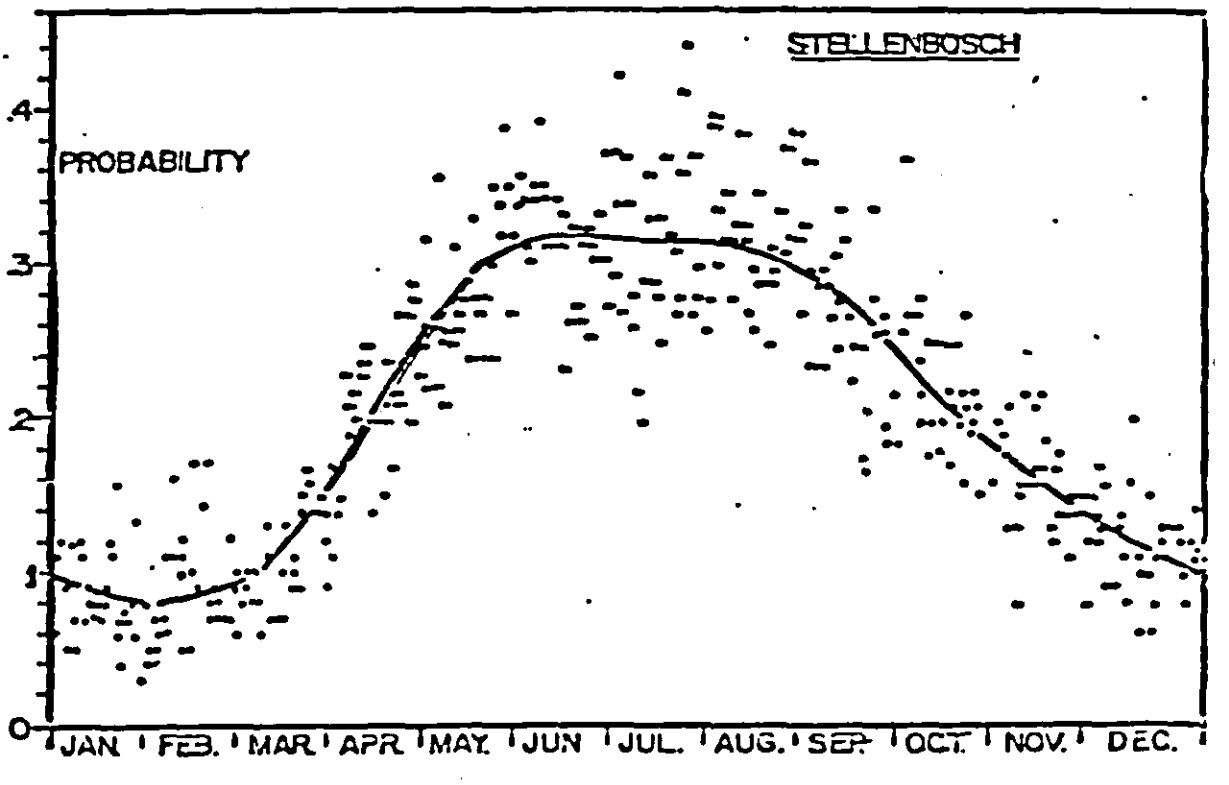
FIGURE 5.5 Simulated (- - - -) and historical (___) mean number of wet days per month.



FIGURES 5.6.1 and 5.6.2

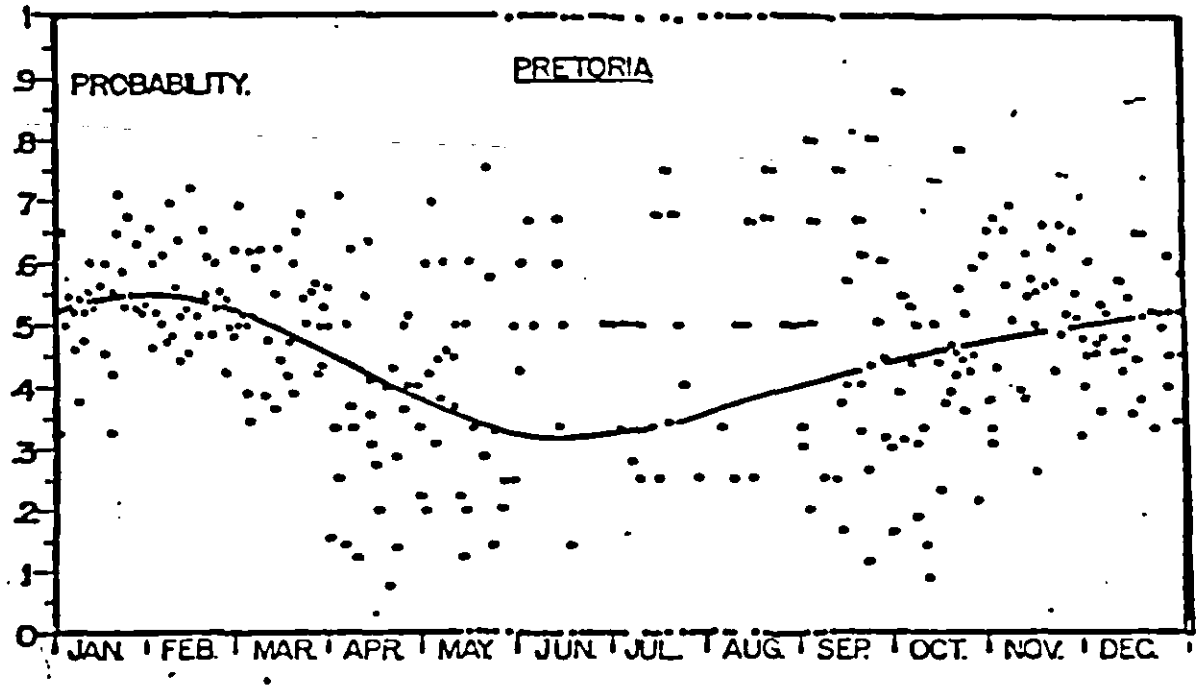
Empirical probabilities and estimates based on a 7-parameter model for the probability of having a wet day in Pretoria and Stellenbosch

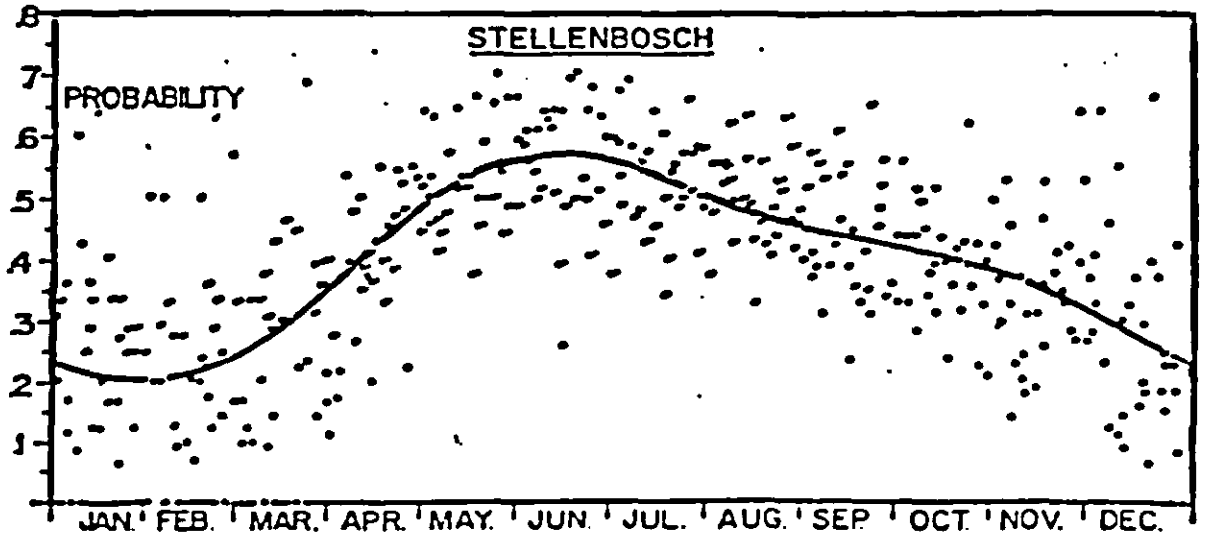




FIGURES 5.7.1 and 5.7.2

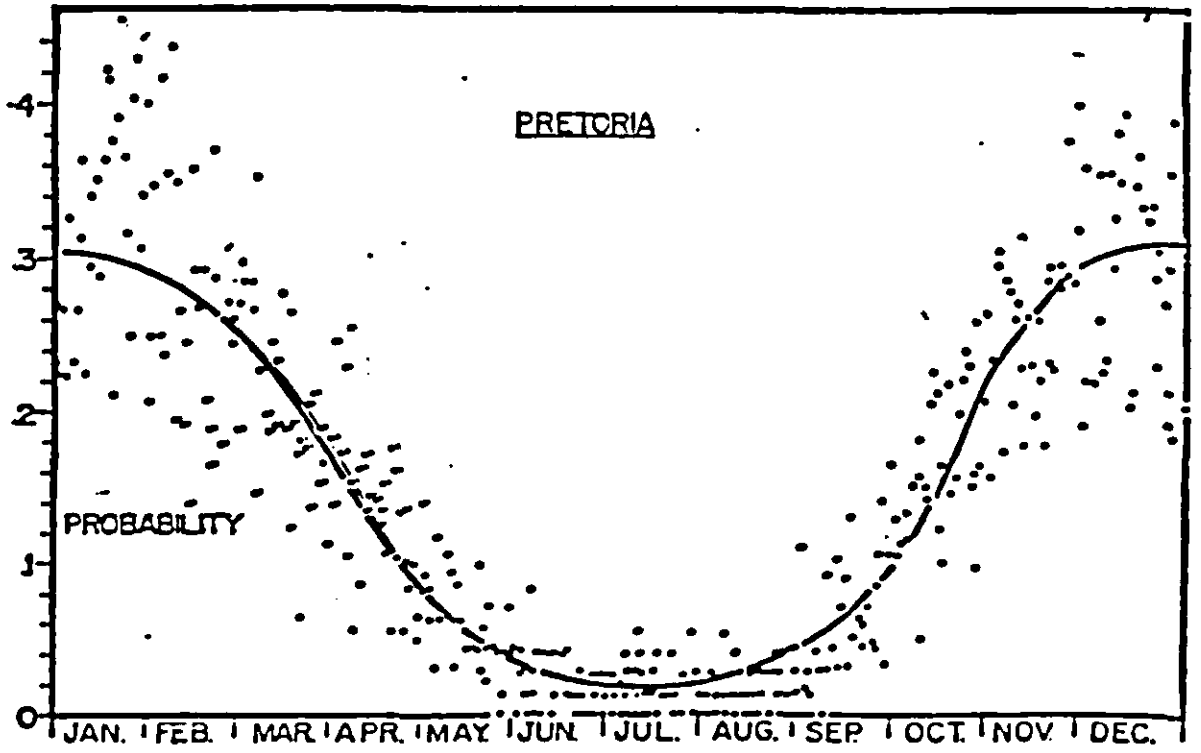
Empirical probabilities and estimates based on a 5-parameter model for the probability of a wet day given a wet preceding day for Pretoria and Stellenbosch

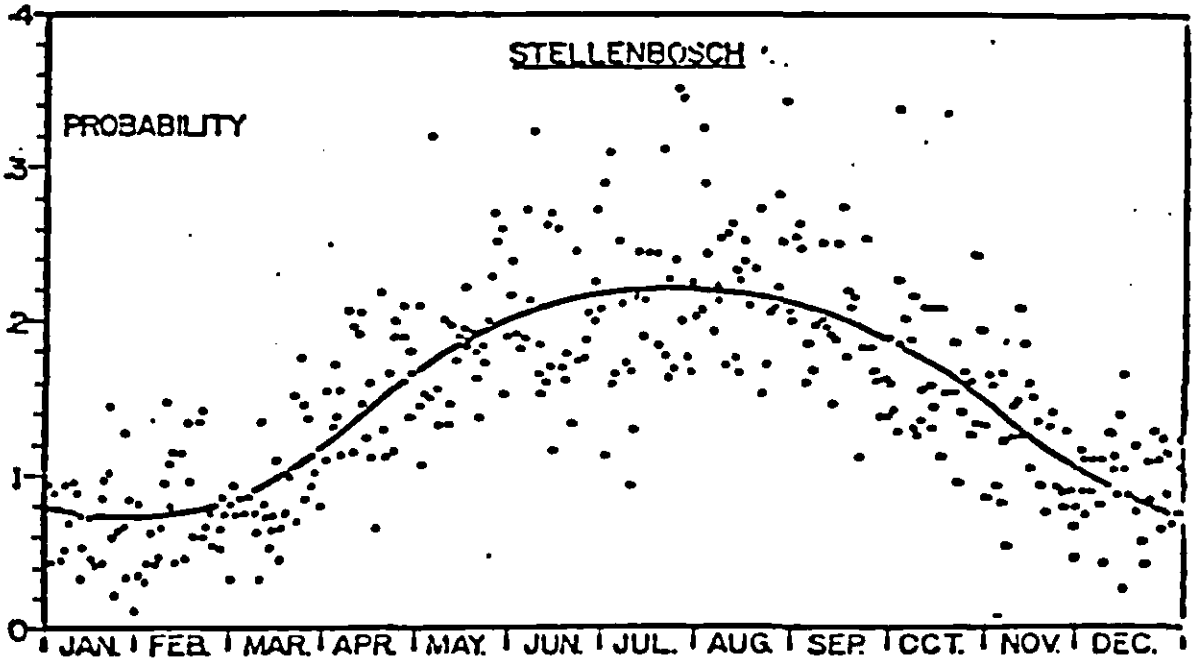




FIGURES 5.8.1 and 5.8.2

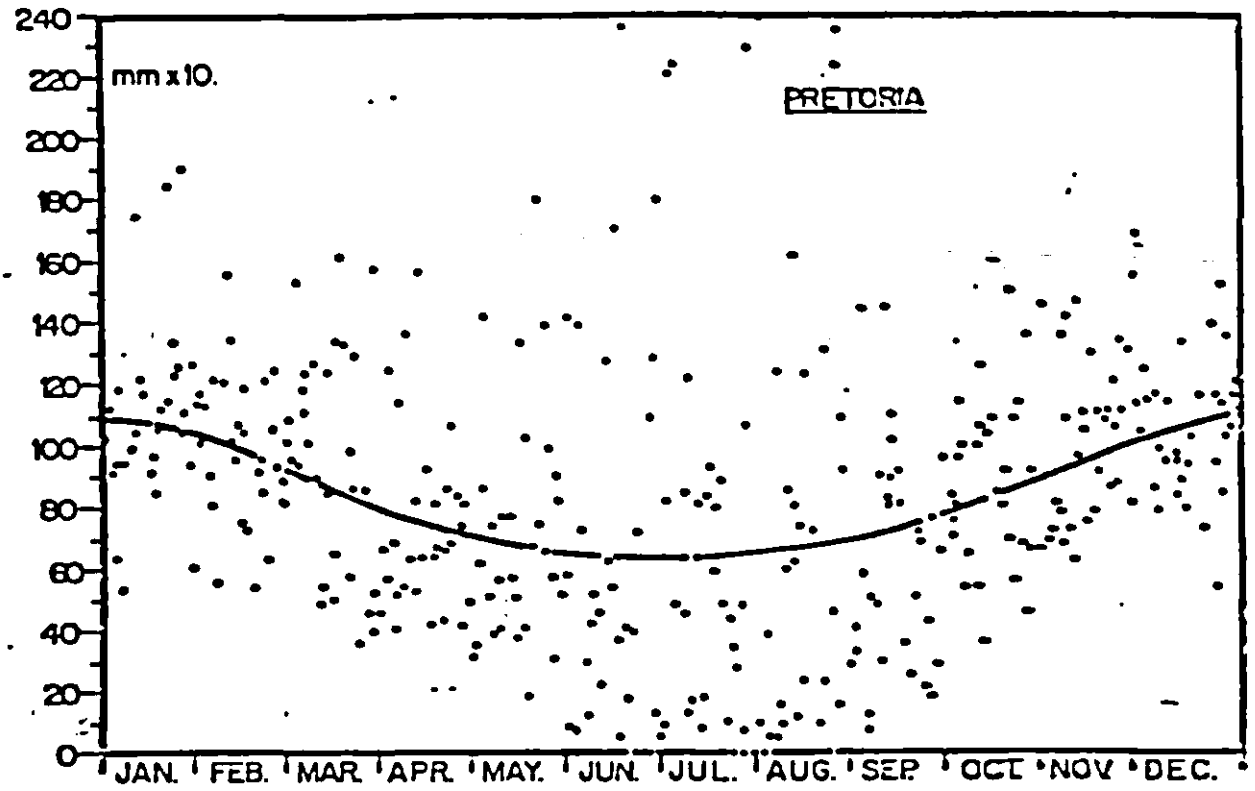
Empirical probabilities and estimates based on a 5-parameter model for the probability of a wet day given a dry preceding day for Pretoria and Stellenbosch

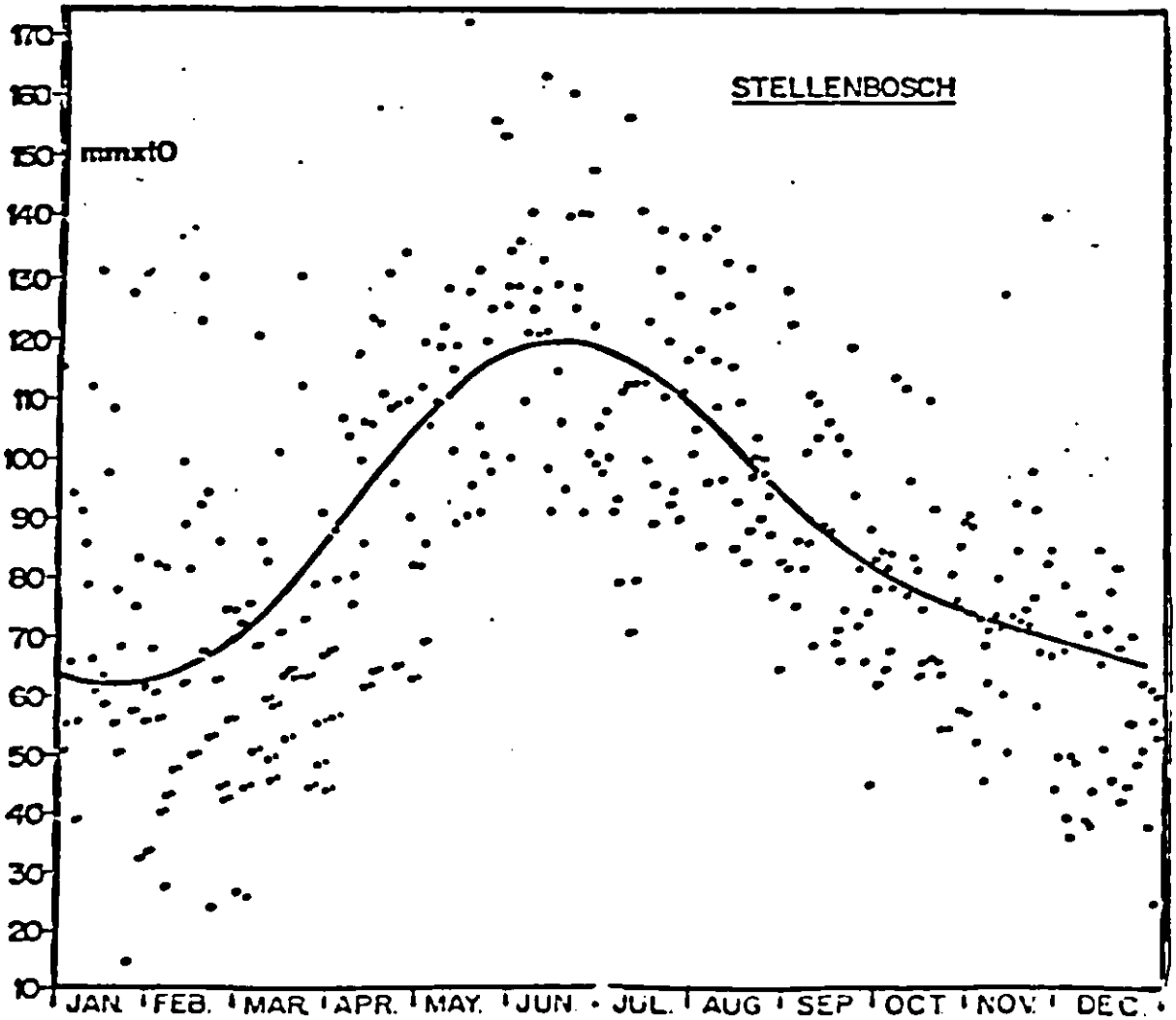




FIGURES 5.9.1 and 5.9.2

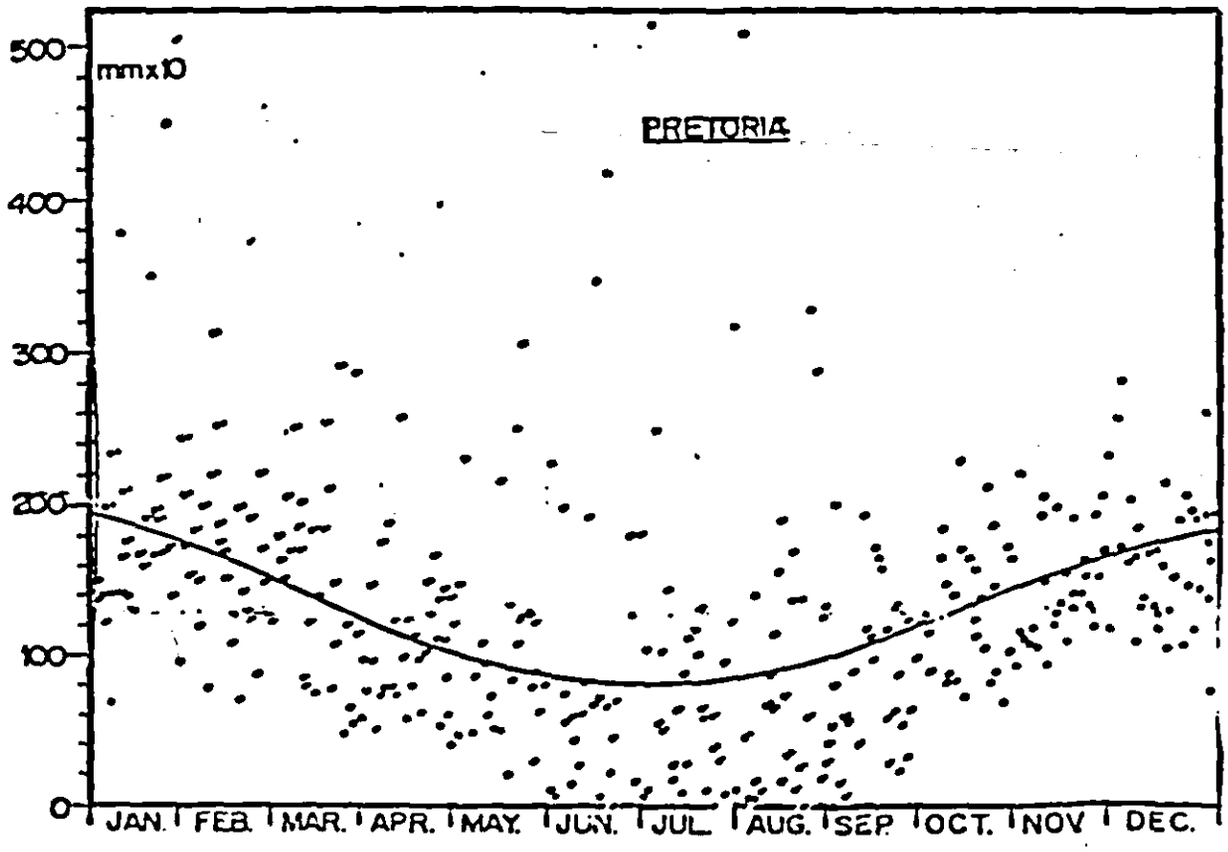
Daily averages and mean fitted by a 5-term Fourier series for Pretoria and Stellenbosch

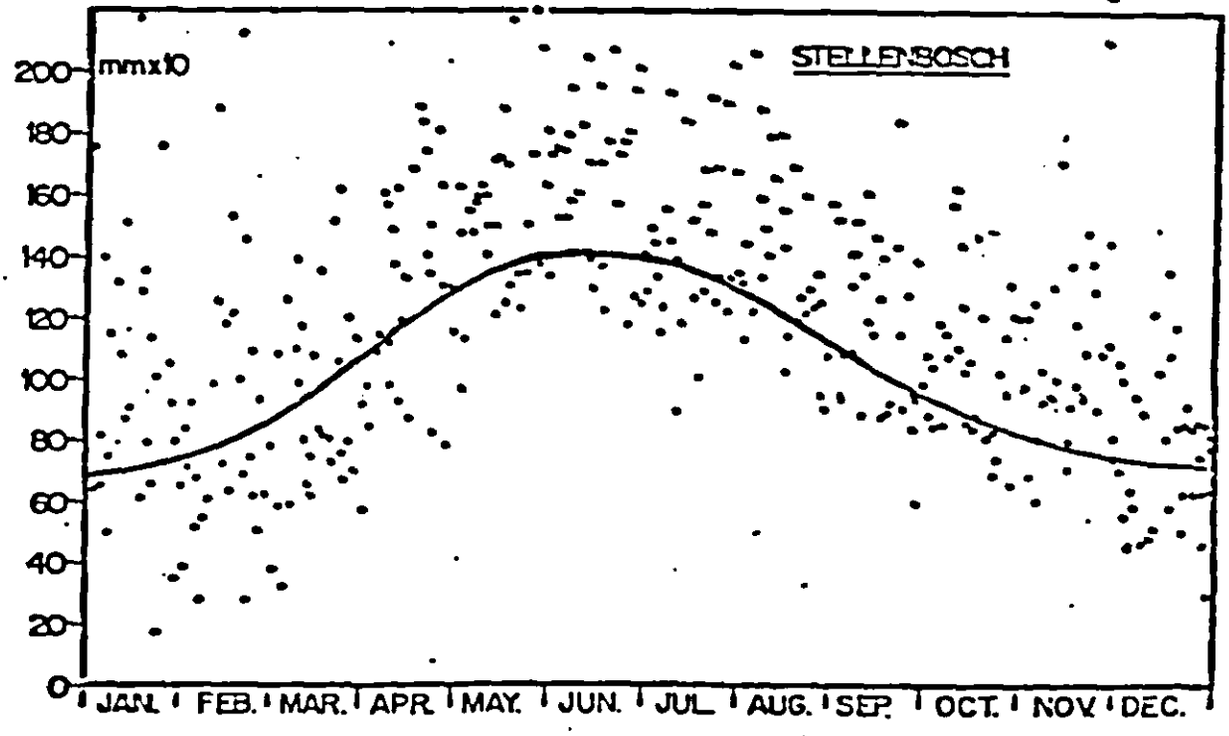




FIGURES 5.10.1 and 5.10.2

Standard deviations computed on a daily basis and those computed using a constant coefficient of variation and a 5-term Fourier series for the mean; Pretoria and Stellenbosch





Two aspects of daily rainfall require investigation with respect to model performance. Firstly that the distribution of daily totals is adequately represented under the assumption that they are distributed as Weibull; and secondly that the Markovian structure of the model preserves the frequency and seasonal variability of runs of wet and runs of dry days.

In order to establish that the Weibull model adequately preserves the distribution of daily rainfalls we proceed as follows:

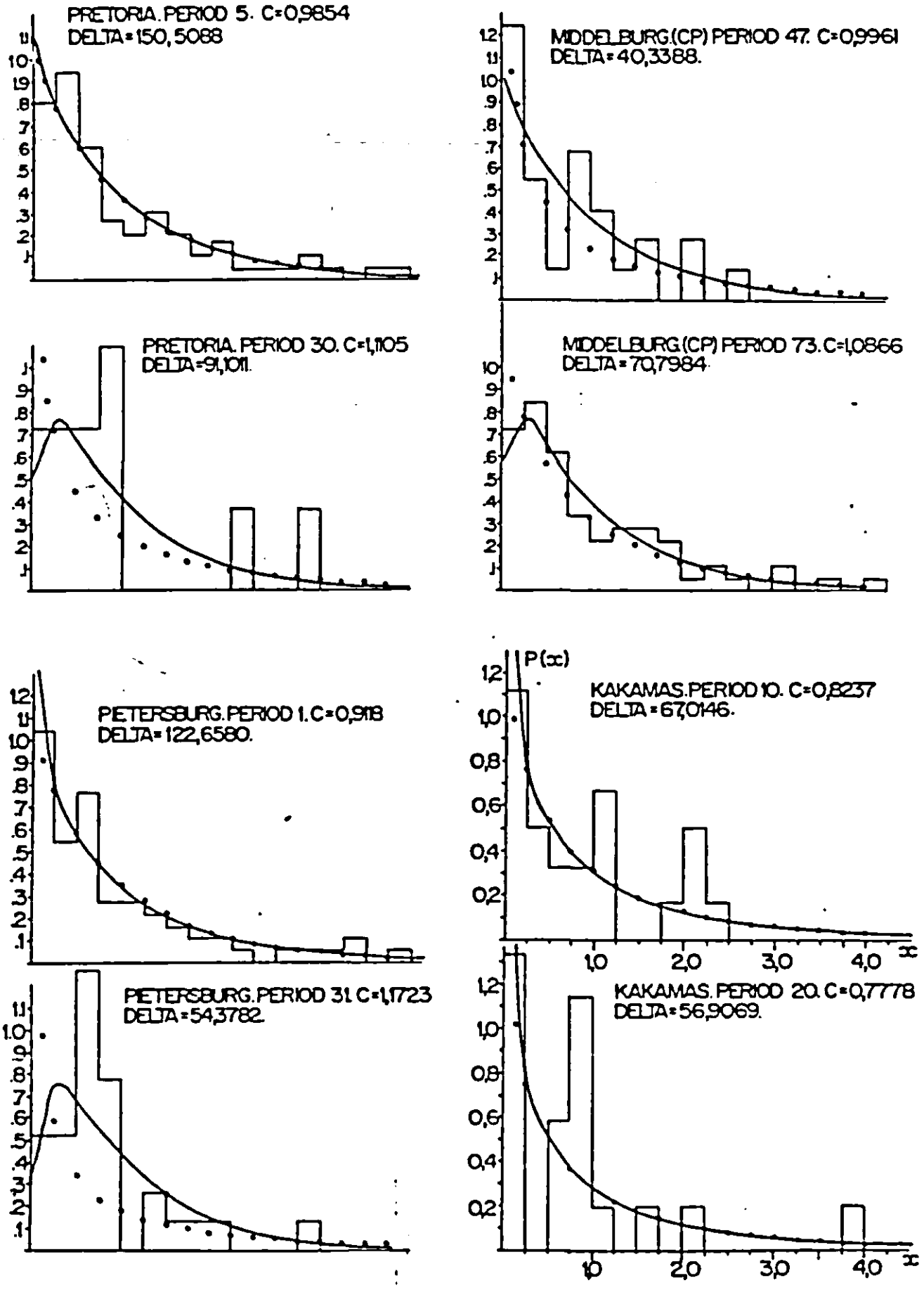
(a) Take a 5-day period at random during each of the wet and dry seasons and draw the frequency histogram of the events that occurred during this particular pentad over the period of historical record. A 5-day period was chosen for this exercise because during the "dry" periods in particular not enough observations would have been forthcoming from any shorter period for the assessment to be made. Thus period 1 would reflect the rainfall depths over the period from 1 to 5 January, period 2 for 6 to 10 January etc ... .

(b) For these historical samples compute the maximum likelihood estimates of the Weibull model and draw the density function on the histogram in standard form.

(c) Using the constant estimate of the coefficient of variation as given by the model compute the shape parameter of the Weibull model using the functional approximation given earlier. Compute the scale parameter using the periodic estimator of the mean and draw the computed standard density function on the histogram.

Figure 5.11 illustrates the results of this exercise and shows that this aspect of the model works quite well. For

FIGURE 5.11 Histogram of historical rainfall depths over pentads with maximum likelihood (—) and model (....) estimators of their density (standardised).



Pretoria (period 30) and Pietersburg (period 31), which represent the dry season, the model results may at first sight appear poor but during these periods we have few observations in the first place and in the second the fit of the truncated Fourier series to the means will give less weight to such periods individually.

Two properties of the run characteristics of daily rainfall require investigation. Firstly that the seasonal distribution of clusters of wet days is preserved by the model and secondly that the seasonal characteristics of runs of dry days is preserved. Figure 5.12 shows the distribution by month of clusters of wet days and may be interpreted as follows: in Pretoria in October 60% of wet days are isolated, 22% occur in clusters of two, 9% in clusters of three and 4% in clusters of 4. We note that during the dry months the relative distribution of the clustering changes with a higher incidence of isolated wet days, as one would expect.

We note that the Markovian structure proposed for the model is particularly successful with regard to this particular aspect of daily rainfalls.

Figures 5.13.1 to 5.13.4 show the seasonal distribution of dry day run lengths for Durban, Pretoria, Kakamas and Stellenbosch. As expected the relative proportion of shorter runs is high in the wet season and as the dry season becomes established the distribution changes to provide a greater proportion of longer runs. There exist cases where the probability that the whole month is dry is higher than that of, for example, a run of exactly 25 dry days (e.g. Pretoria, June). This is obvious since the frequency that the whole month is dry is higher than a dry day run of 25 days as the latter case would imply a wet day at the very beginning or end of the dry season month. Such an occurrence

FIGURE 5.12 Seasonal distribution of clusters of wet days: (—) historical; (----) simulated.

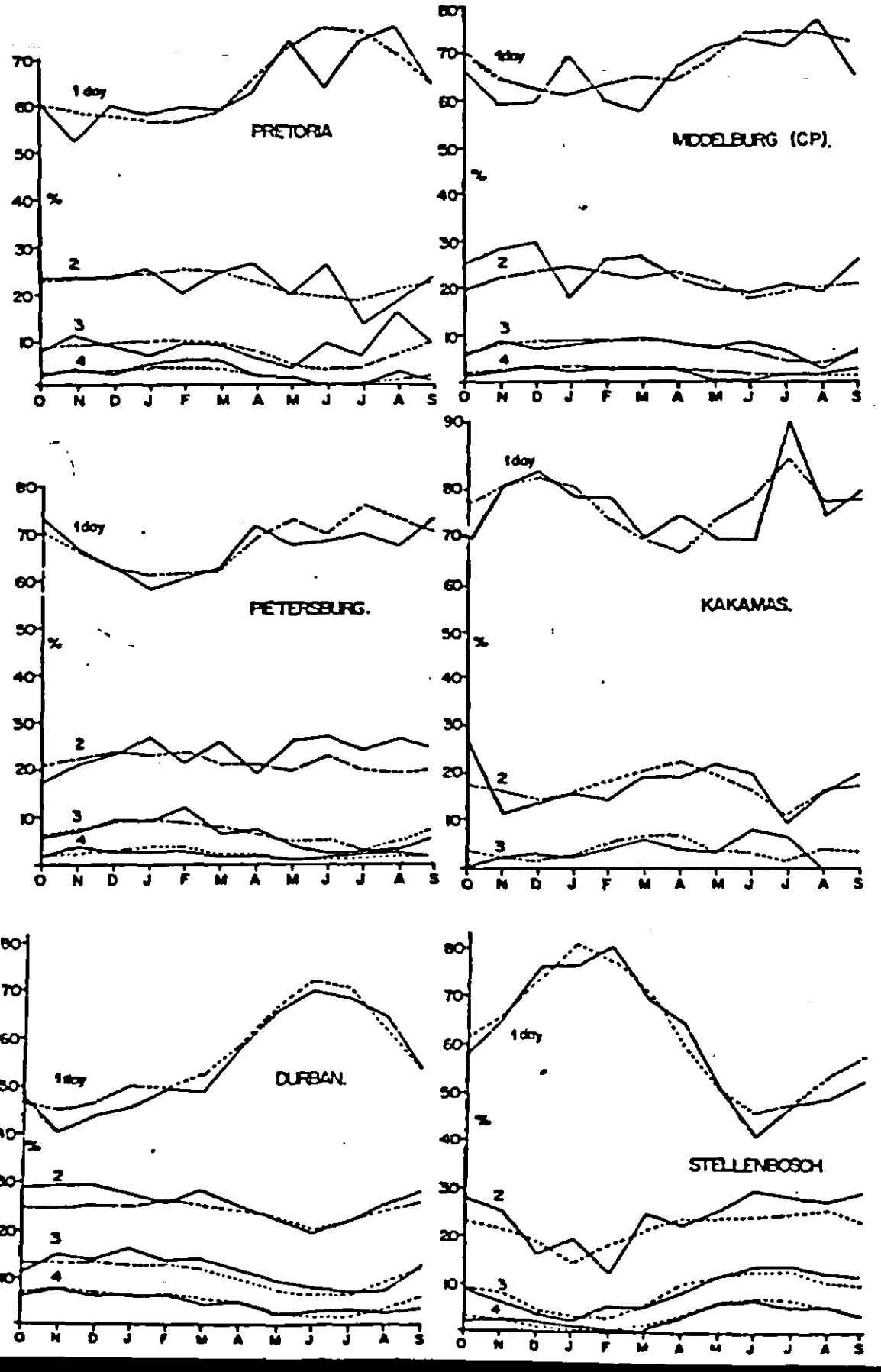


FIGURE 5.13.1 Histograms and simulated frequency distributions of run lengths of dry days by month : Durban.

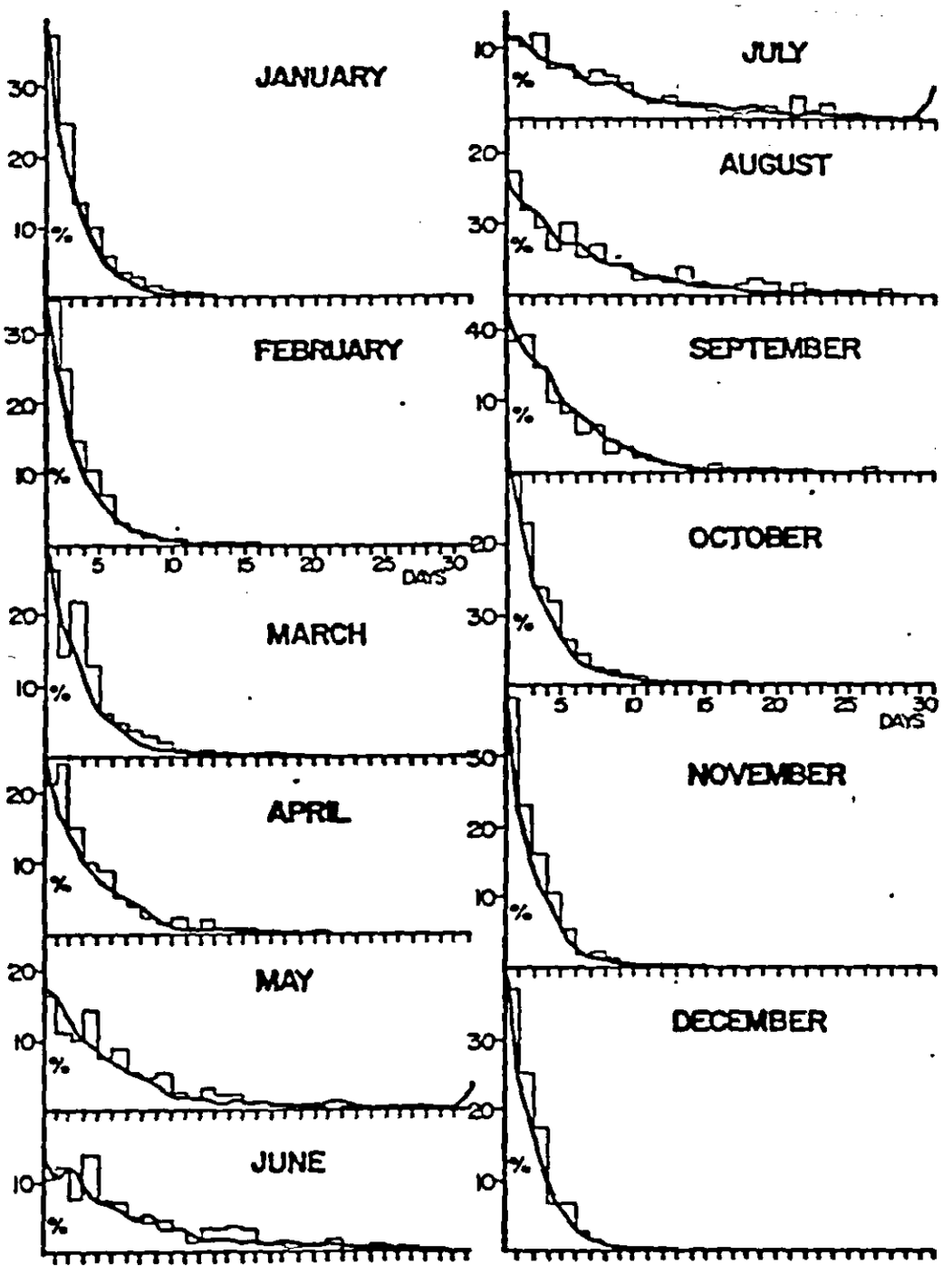


FIGURE 5.13.2 Histograms and simulated frequency distributions of run lengths of dry days by month : Pretoria

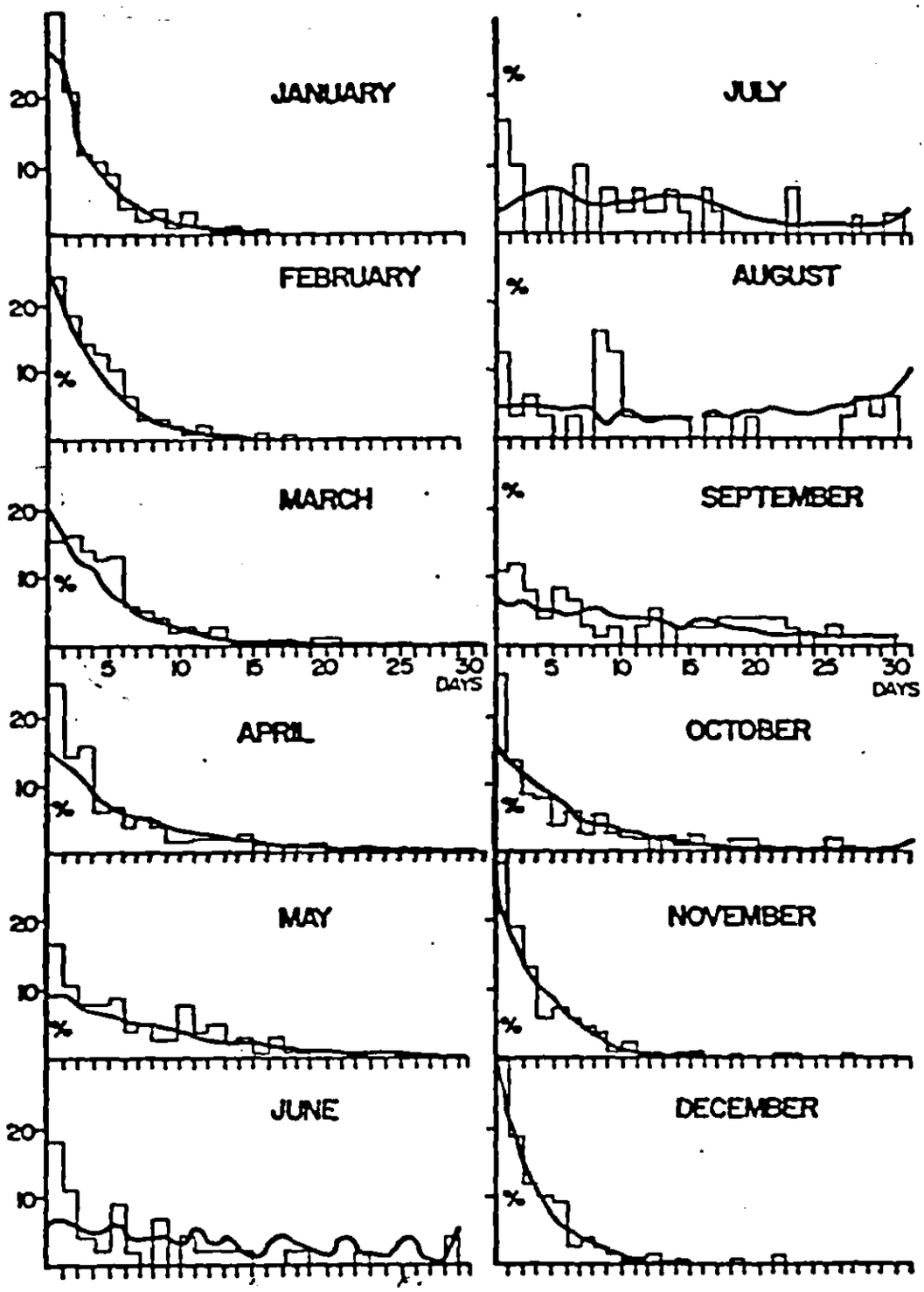


FIGURE 5.13.3 Histograms and simulated frequency distributions of run lengths of dry days by month : Kakamas

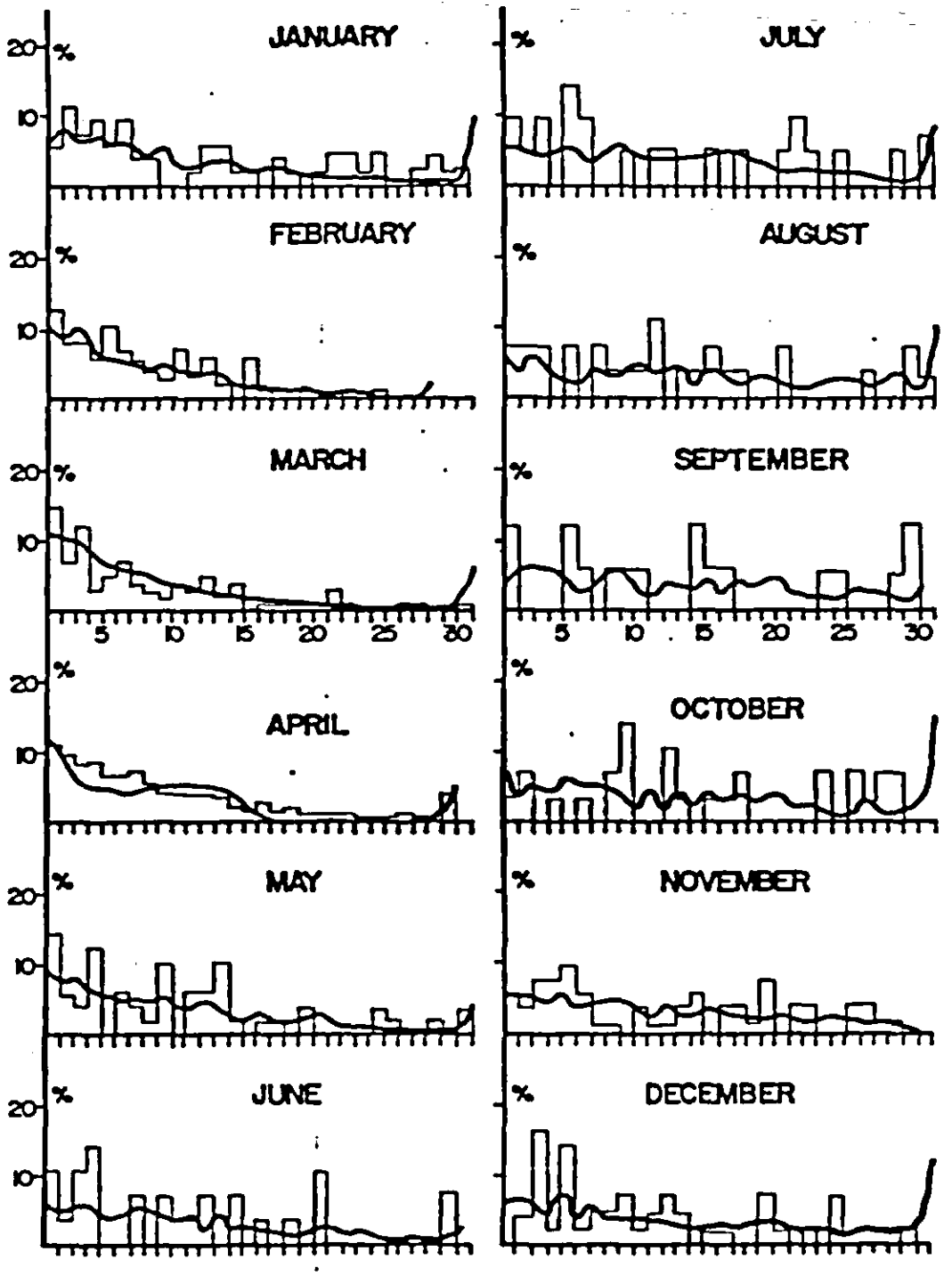
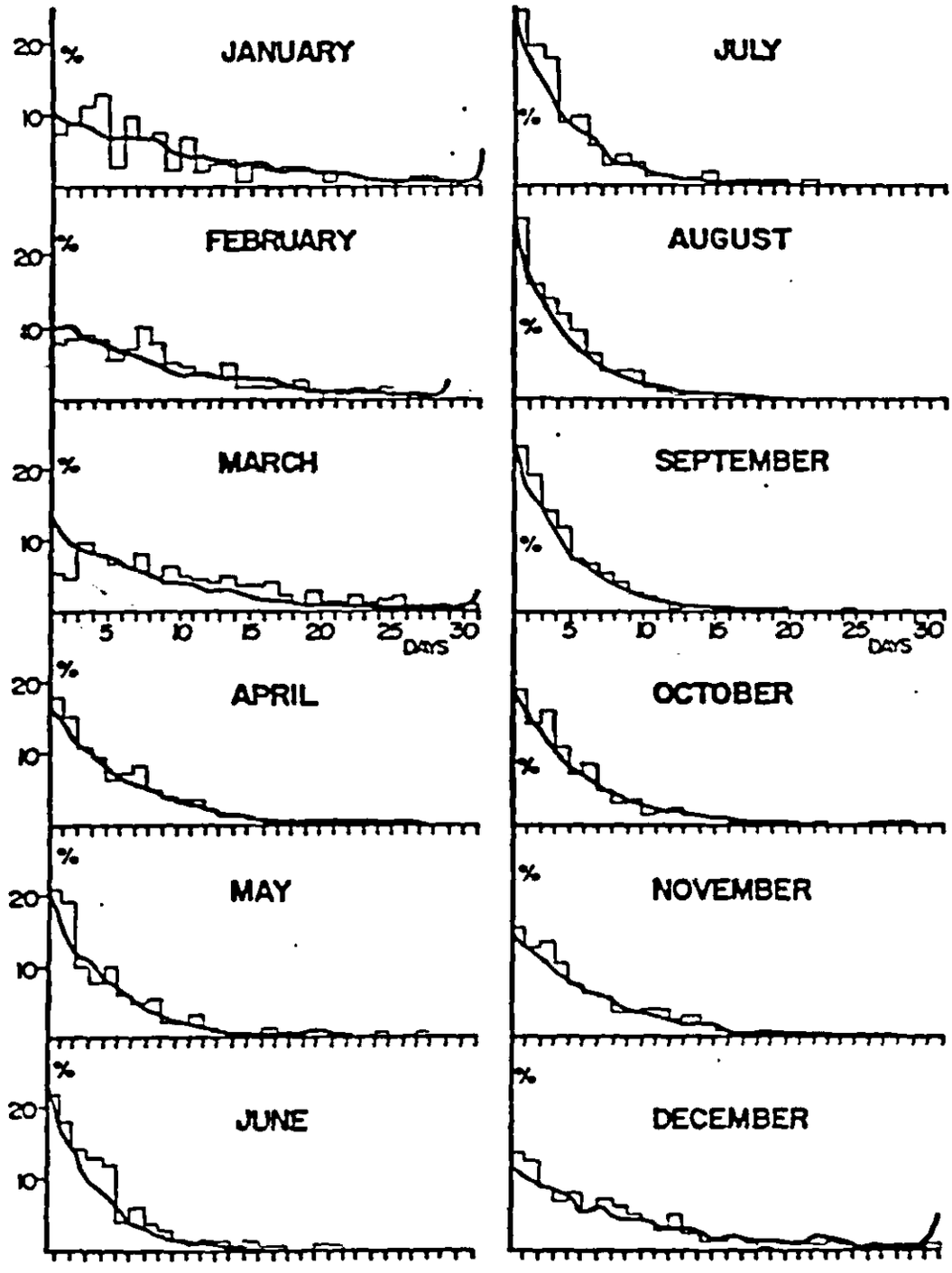


FIGURE 5.13.4 Histograms and simulated frequency distributions of run lengths of dry days by month : Stellenbosch

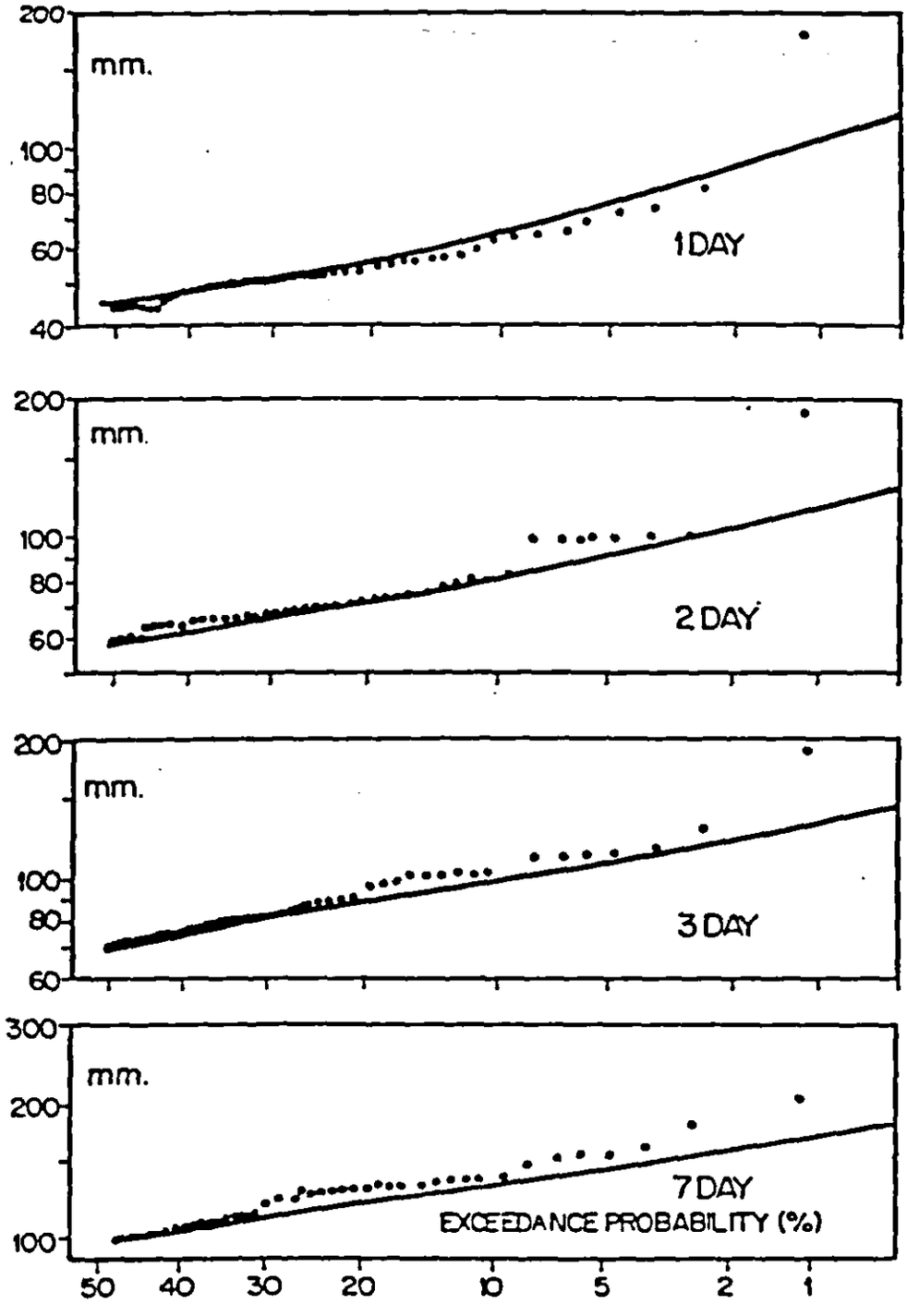


would be relatively rare.

A relatively severe test of model performance is to assess its fidelity in preserving the distribution of annual maximal n-day rainfalls. This assessment is severe in the sense that one would, in order to preserve such extremes, apply considerable emphasis to the seasonal periodicity of the variance and the selection of a univariate model of daily rainfall depths. The model as it stands is not tied to any particular univariate model in so far as one could from the moments estimate the parameters of any number of likely models, e.g. the gamma or extreme type I. The Weibull, as it happens, appears to perform well so far.

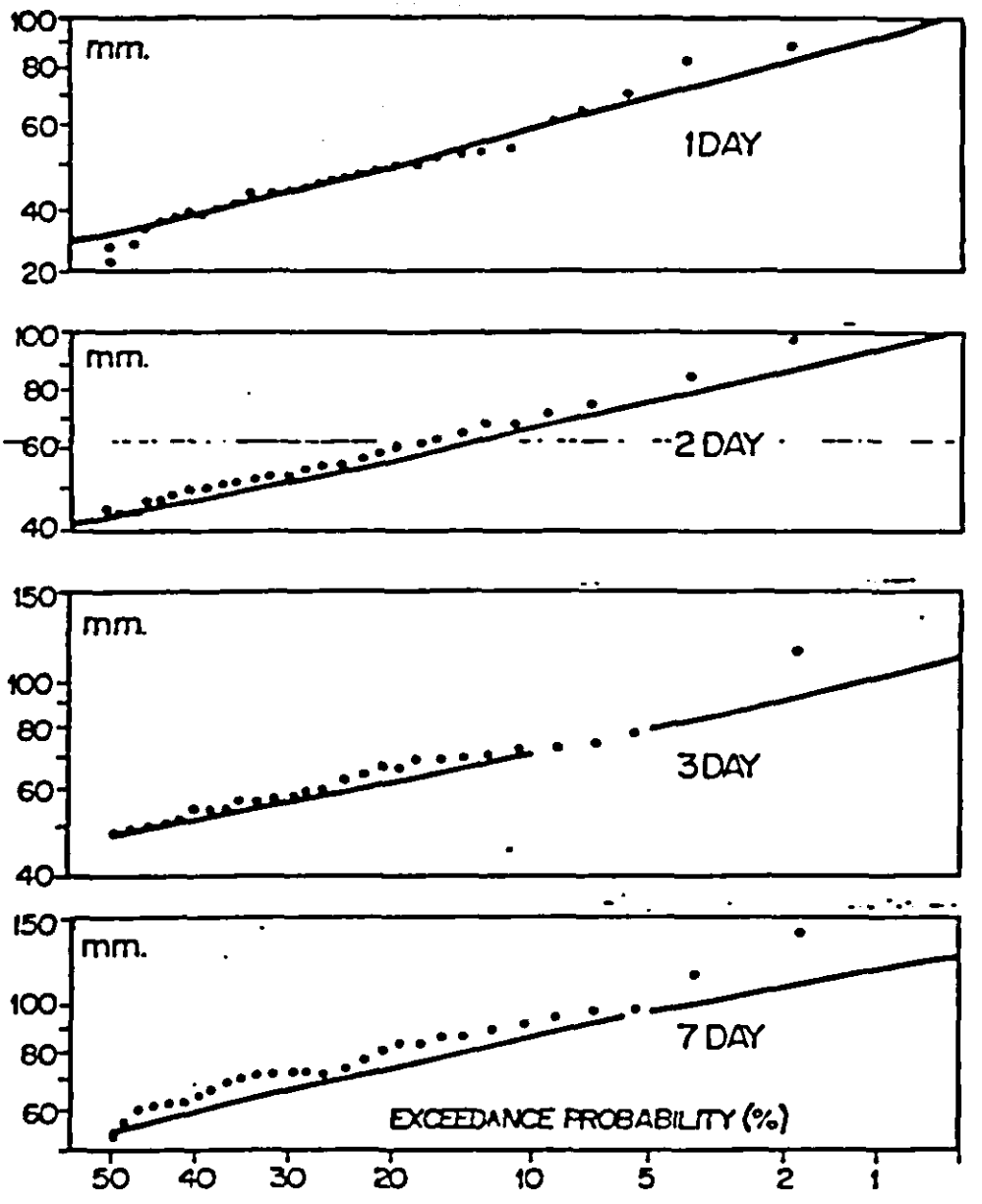
From the historical data we sample the n-day annual maxima and plot them in the usual way according to the Weibull plotting position (rank/(N+1)). We then simulate their distribution and consider the 1, 2, 3 and 7 day events. Figures 5.14.1 to 5.14.6 show the results for the six stations. Generally the results are good, given the approach used to model the periodicity of the variance of daily rainfalls and the assumption of a constant coefficient of variation over the year. As can be seen for Durban and Pretoria in particular, the model cannot preserve the distribution of extremes which are obviously drawn from a mixture of synoptic generating processes. That this is so is manifested in the decided break in the frequency curve which at Durban for example would be associated with the incidence of intense storms related to the rare influx of cyclones over the coastal regions. In order to accommodate this characteristic of extremes in some areas of the country a more complex approach to the modelling of the seasonal variance of daily rainfalls would be required . Also the univariate model fitted to the depths would need be of the mixed distribution type as proposed by Woolhiser and Pegram (1979).

FIGURE 5.14.1 Sample points and simulated distribution function of annual maximum n-day rainfalls : Stellenbosch



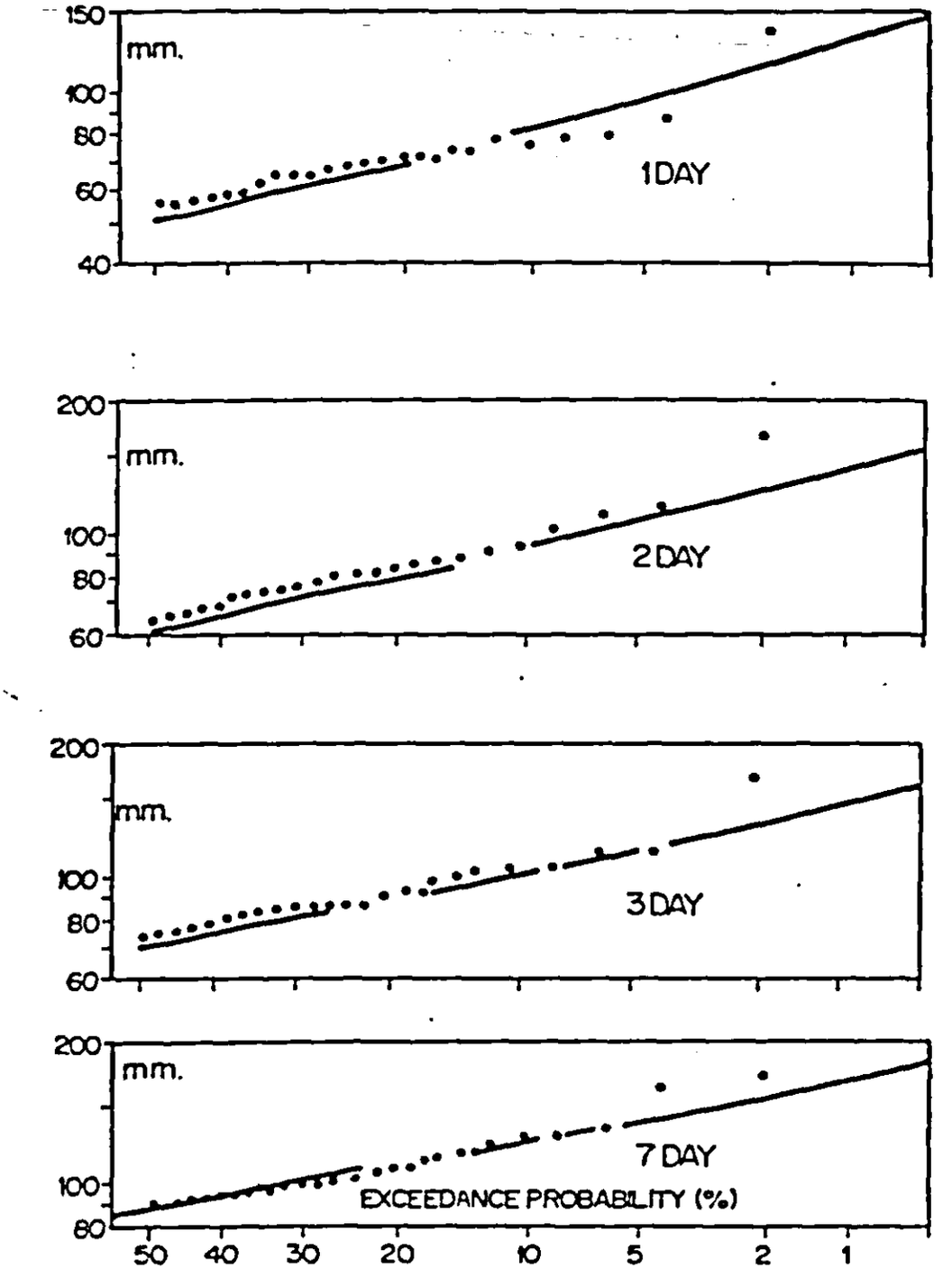
STELLENBOSCH

FIGURE 5.14.2 Sample points and simulated distribution function of annual maximum n-day rainfalls : Middelburg



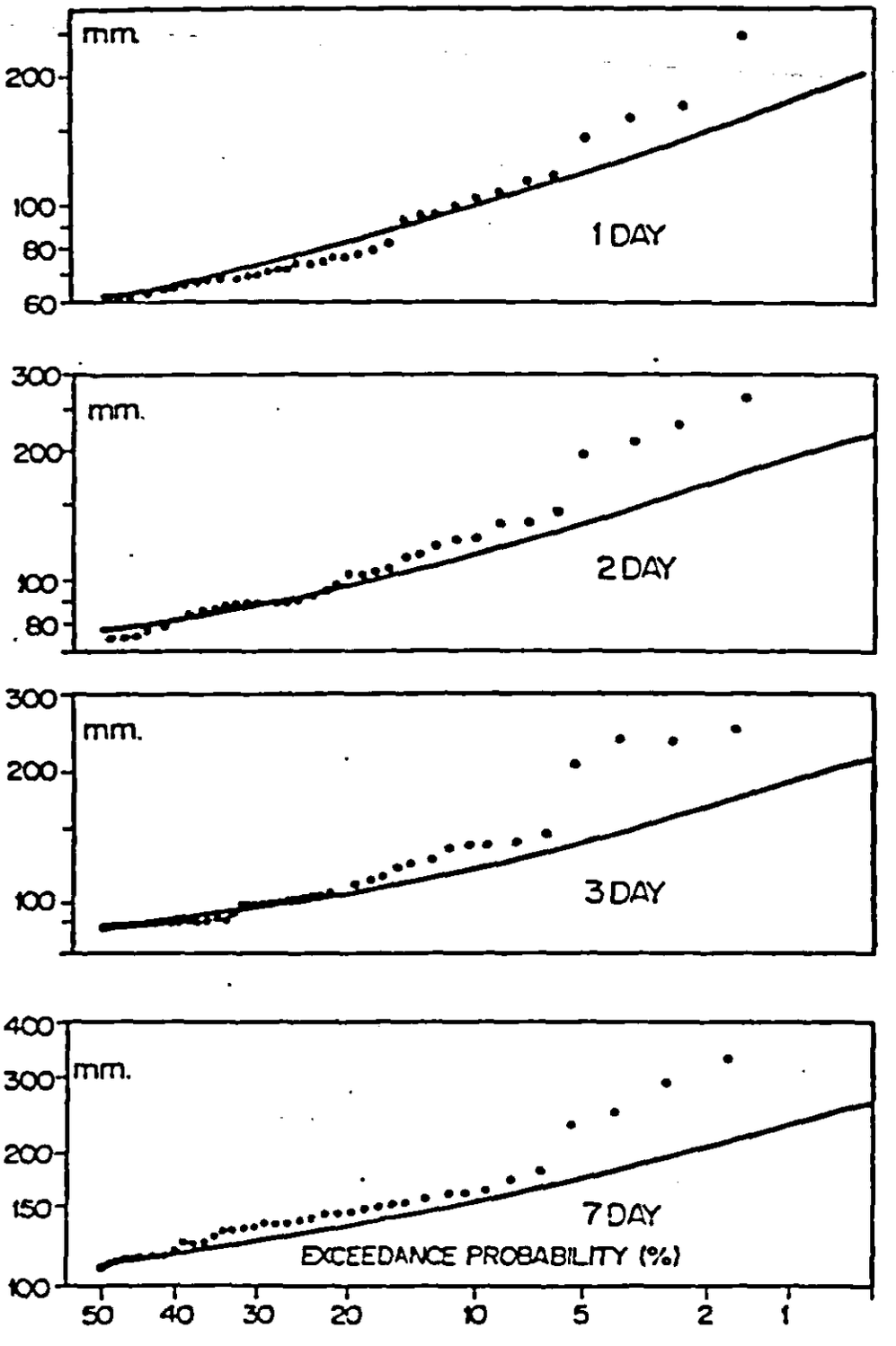
MIDDELBURG

FIGURE 5.14.3 Sample points and simulated distribution function of annual maximum n-day rainfalls : Pietersburg



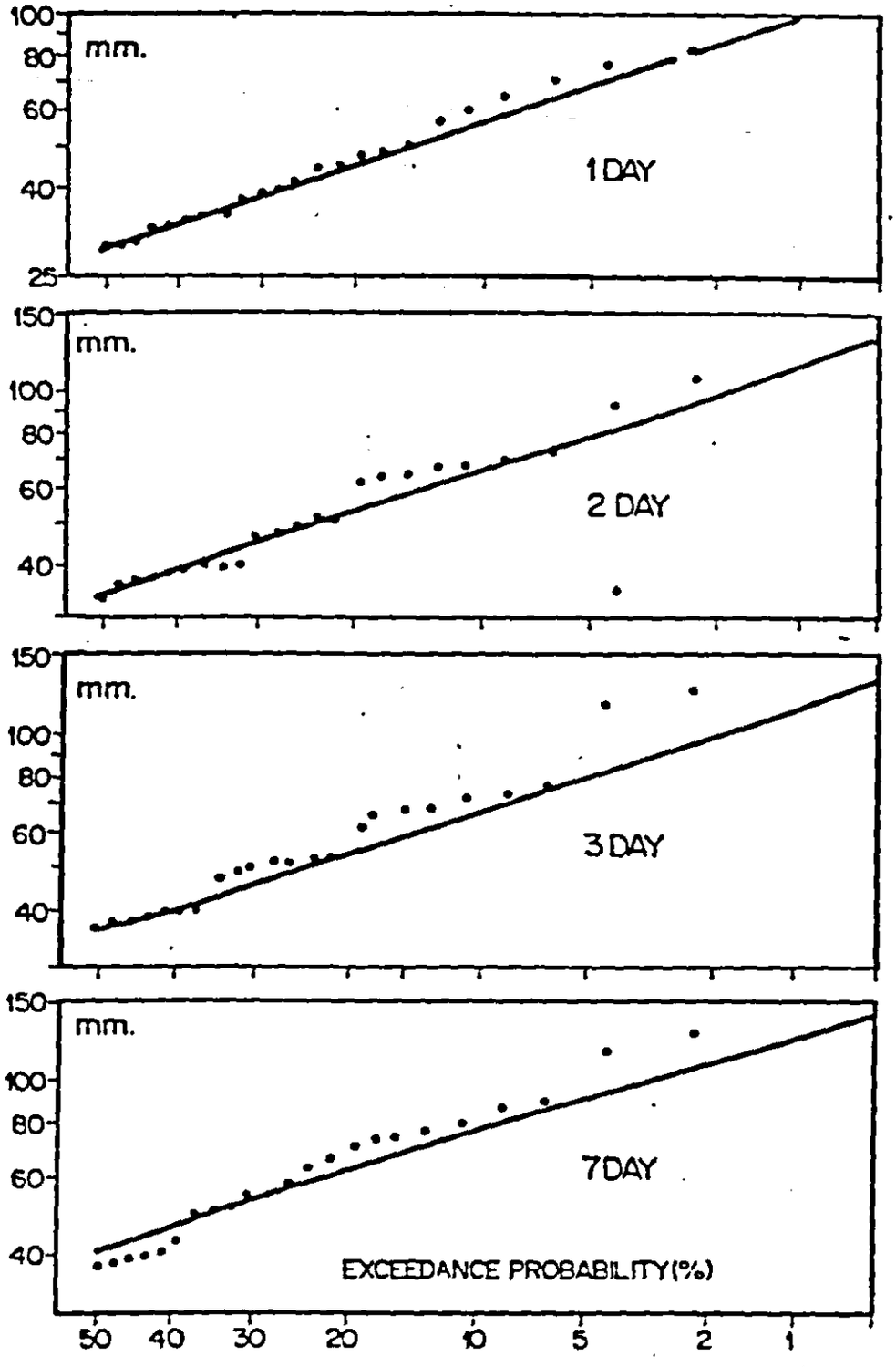
PIETERSBURG

FIGURE 5.14.4 Sample points and simulated distribution function of annual maximum n-day rainfalls : Pretoria



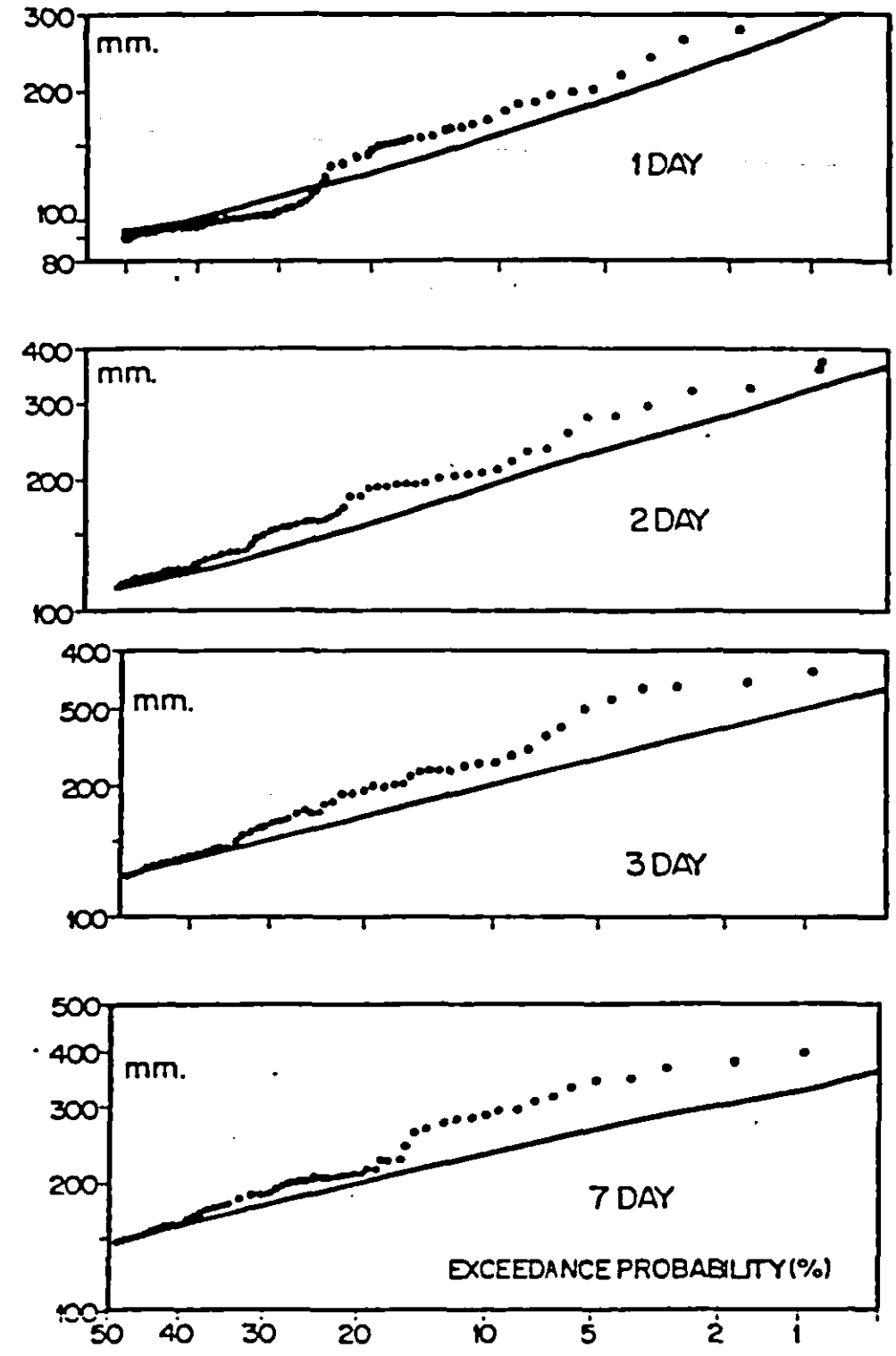
PRETORIA

FIGURE 5.14.5 Sample points and simulated distribution function of annual maximum n-day rainfalls : Kakamas



KAKAMAS

FIGURE 5.14.6 Sample points and simulated distribution function of annual maximum n-day rainfalls : Durban



DURBAN

Conclusions

The results of the model validation reveal that the assumptions initially made about the structure of daily rainfall data, the rationale of model structure and the parameter estimation techniques have been particularly successful in providing a model that can faithfully reproduce the properties of daily rainfall sequences. A higher order Markov model combined with a more complex univariate model for rainfall depths may have led to even better results but its contribution to the practical application of the model would probably not justify such additional complexity. The criterion for adopting a model depends as much on the use to which it is to be put as on statistical tests of its validity (Gabriel 1984) and significance. The proposed model may be seen to provide rainfall sequences that are suitably accurate for a vast number of applications.

6. APPLICATIONS OF THE DAILY RAINFALL MODEL

Statistical models provide a concise summary of data sets which in themselves may be unavailable to the majority of potential users. This is particularly true of daily rainfall which requires very large data bases for efficient storage. In order to be made manageable the rainfall sequences are usually summed to form monthly or annual series which are then used to portray point or regional rainfall characteristics, such as mean monthly or annual rainfall. Alternatively they may provide the input to models of streamflow or soil moisture. But a good statistical model does more than just summarise the data - it provides insight into the underlying process of which the observed data set is only one manifestation.

The advantage of the present model, as of any parsimonious model of a stochastic process, is that all the properties of the process, in this case daily rainfall, are encapsulated in a relatively small number of parameters. The probability distribution of events of importance thereby become accessible to users who have access only to modest computing facilities and at very little cost or inconvenience. The parameters themselves can be mapped to provide insight into the properties of daily rainfall characteristics of a region; stochastic sequences of daily, weekly, monthly or annual rainfall can be simulated to compute the distribution of some characteristic of the point rainfall process. We can simulate annual totals, for example, and estimate their probability distribution or we can simulate the probability of runs of dry days within the period of critical growth for a commercial crop.

A daily rainfall model as presented here and as fitted on the scale of 2550 locations throughout South Africa is of

great value and has a very broad range of potential applications: from the assessment of the geography of the rainfall climate of the country from a daily to an n-annual time period, to the generation of stochastic realizations of the process as input to further models. From this stochastic input we can compute the distribution of some event in the output such as that of a run of deficient flows over some discrete interval using a rainfall-runoff model.

Before proceeding to the consideration of a model of point drought and its applications in conjunction with the daily rainfall model, we use the latter per se: firstly to paint a picture of the spatial characteristics of the South African rainfall climate; and secondly to illustrate several types of results of interest in agricultural planning where the seasonal distribution of daily rainfall amounts is of considerable importance. We concentrate on deficiencies in rainfall in these examples.

There are three methods of obtaining results from the model. The first is analytic, in which a formula is derived to give the required result in terms of the parameters of the model. Todorovic and Woolhiser (1975) give such a formula for the total rainfall in a period, using a very simple model, with a single constant probability of rain and exponentially distributed rainfall amounts. However, even for such a simple model, the equations are complex and not feasible for more realistic models. We can, however, as a second method, derive results directly from the proposed model by using its Markovian structure. A very simple example would be that the estimate of the mean annual number of wet days is simply the sum of the 365 probabilities of a wet day. Recurrence relations (Stern 1980, 1981, 1982, Stern and Coe 1984) can be derived which use numerical procedures to solve a set of equations, one set for each day, to build up

results over a period of interest. The method can be used to provide results on the total rainfall in a given period, the probability of long dry spells and the distribution of the start of the rainy season. This approach rapidly leads to mathematical complications if some of the more subtle properties of the rainfall process are investigated.

The third and most convenient method of obtaining results from the model is simulation. Sufficiently long records are generated so that a smooth approximation to the distribution function of the event of interest is obtained. The generation of sequences of wet and dry days from a Markov chain is straightforward whilst the generation of random variables distributed as Weibull is perfectly economic in terms of computer time. It is quite easy to prepare a computer program which accepts the parameters of the rainfall process as input and then generates an artificial rainfall sequence of any desired length - an algorithm to do this is given in Chapter 4. Once one has this program it can be used as the basis for simulating *any* desired property of the rainfall process.

The rainfall regime of South Africa

South Africa displays a large variety of regional climatic characteristics and transitions. The element most critical in its effects on land use and economic development is rainfall and by association, streamflow and the availability of water. There have in consequence been numerous studies toward the identification of the major climatic and agricultural regions of South Africa largely based on the areal distribution and seasonality of precipitation. These classifications have generally been based on a relatively simple view of seasonality and mean annual precipitation with monthly means constituting the basic unit of analysis (Dove 1888 , Schumann and Thompson 1934 ,

Schumann and Hofmeyr 1938 , Schulze 1947 , 1958 , Jackson 1951 , Wellington 1955). These regional delimitations were systematically reviewed and adopted by the South African Weather Bureau (1954-1963) to produce a system of maps for regional climatic classifications.

These traditional means of portraying precipitation climatology can hardly achieve more than a qualitative description of core regions of different precipitation regimes. McGee and Hastenrath (1966) considered the spatial continuity of precipitation characteristics and their gradual transition over South Africa using the results of a harmonic analysis of mean monthly rainfalls at 513 locations. Amplitudes, phases and the variances attributable to the various harmonics were mapped and the seasonal march of the rainfall climate across the country well illustrated. Welding and Havenga (1974) considered the spatial correlation of monthly rainfall sequences and, using a hierarchial classification technique, delimited regions wherein the correlation between stations was above a critical value. This allowed for the detailed identification of precipitation regions which share a significant level of temporal association.

We can then view the rainfall climatology of South Africa either as a spatial continuum where gradual zones of transition are clearly identifiable, or as a system of discrete regions within which selected characteristics of the time series of monthly or annual rainfalls vary within some acceptable bounds.

A model of daily rainfalls at 2550 locations provides the potential for a particularly detailed study of the rainfall climate of South Africa. The strictures imposed by the need to analyse hundreds of data sets are removed and replaced by parameter sets and a simple generating model

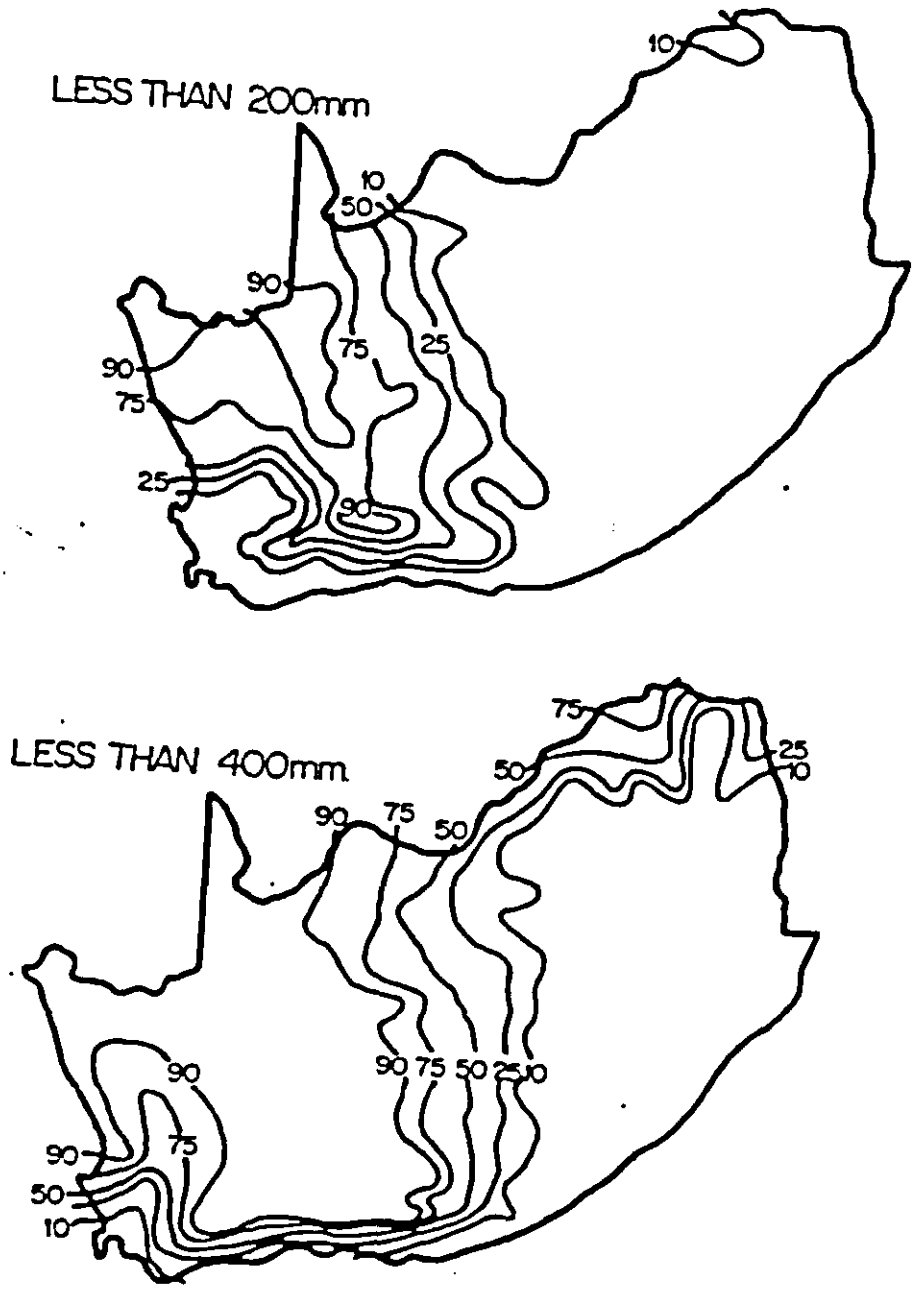
from which any statistic of interest and its distribution can be estimated by simulation. The possibilities for mapping the march of the precipitation climate across South Africa are legion. In order to estimate the probability of the summer rainfall season beginning by a certain date we may simulate the distribution of say 25 mm or more rain in 5 days or less over pentads starting on 1 October. We may be particularly interested in the "storminess" of the climate in the form of the mean percentage of seasonal rainfall attributable to days on which the rainfall exceeds 25 mm. We may be interested in the frequency of dry-day runs exceeding some critical length during the wet season. Such detail is possible in addition to the more familiar monthly and annual statistics that we may wish to investigate.

Figures 6.1.1 to 6.1.4 show a sequence of mappings of annual rainfall percentiles computed at 540 sites over the country. The distribution functions were estimated from a simulation of 500 years of daily data summed to provide annual totals at each point. The maps contain far more information than the usual presentation of mean annual rainfall which, particularly in arid and semi-arid regions given the high variability of the annual regime, conveys a minimum of information. The continuity of the decline in annual rainfalls from the coastal regions towards the subcontinental interior is quite clear and the maps provide a useful spatial presentation of the distribution of annual rainfall totals.

Simulation may be used for the agroclimatic classification of a region according to some or other aspect of rainfall. Such classifications can be quite complex since in dry land agriculture edaphic (soil) factors play a role in the modification of regional rainfall characteristics. Figure 6.2

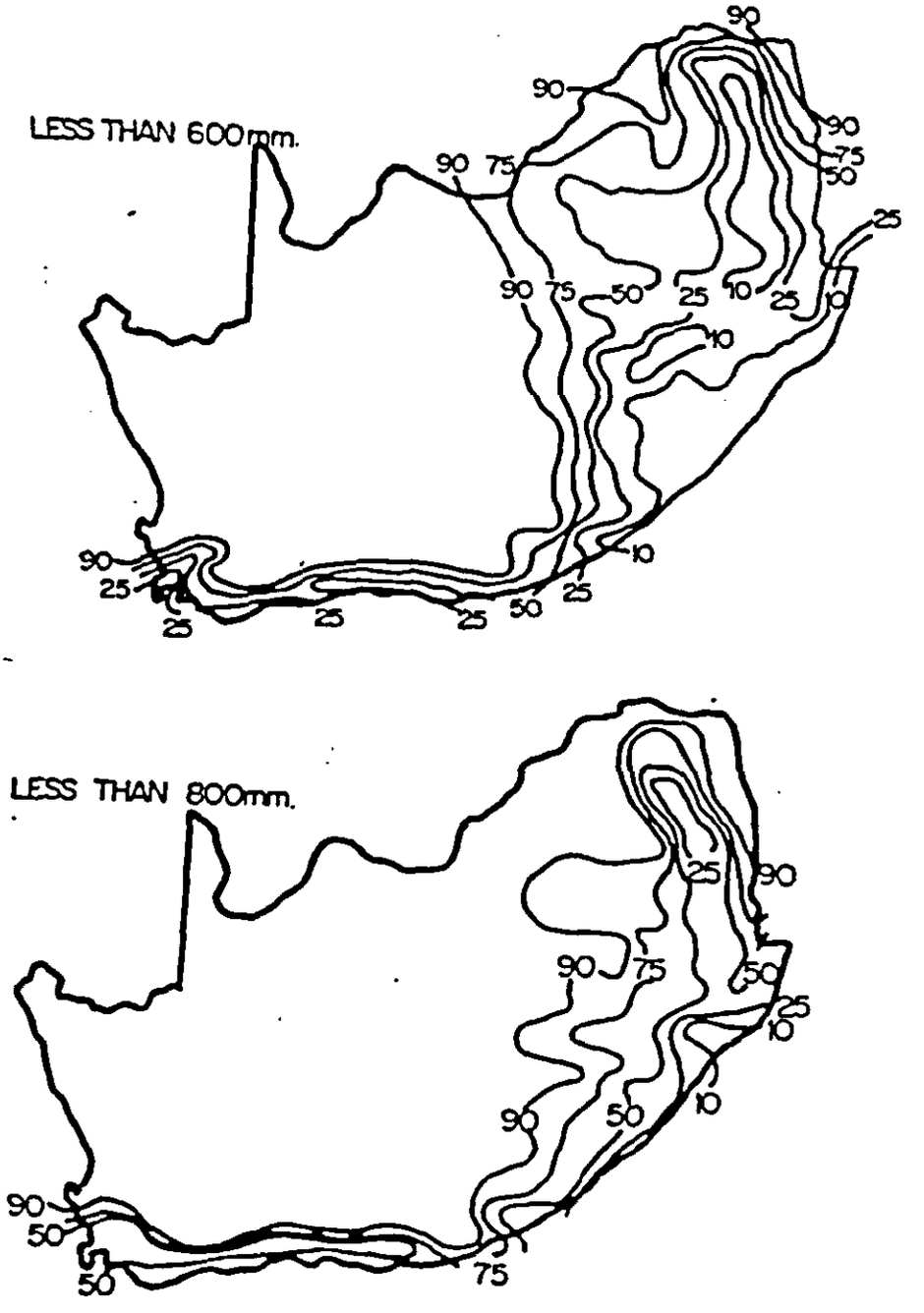
FIGURES 6.1.1 and 6.1.2

Contours of equal probability (expressed as percentages) for the event that the annual rainfall total is less than 200 mm (top map) and less than 400 mm (bottom map).



FIGURES 6.1.3 and 6.1.4

Contours of equal probability (expressed as percentages) for the event that the annual rainfall total is less than 600 mm (top map) and less than 800 mm (bottom map).



shows a very simplistic classification of South Africa into arid, semi-arid, sub-humid and humid regions. It is based on the median (50% percentile) of annual rainfall and the actual figures chosen to discriminate between the regions are very largely arbitrary. The classification does, however, delimit the major maize producing areas of the Transvaal, Orange Free State and Northern Natal which account for 95% of national production (Gillooly and Dyer 1982) and as encompassed by the 500 mm isohyet. Outside this area the frequency of moisture stress conditions during the growing season reduces yields quite considerably. Regions of South Africa that may, according to this scheme, be classified as humid, are confined to the Central Natal coastal belt and the South Western Cape Peninsula.

The inherent variability of annual rainfalls is of obvious importance in many applications from agriculture to water resources planning and is usually mapped as the coefficient of variation of the annual rainfalls. Five hundred years of data were simulated at five hundred sites over South Africa and the coefficient of variation computed and mapped. The result is shown in Figure 6.3 with, as expected, the higher variability associated with the more arid regions.

The seasonality of precipitation is the tendency for a place to have more rainfall in certain months or seasons than in others and a rather efficient way of mapping the tendency is given by Markham (1970). The assumption is made that the mean monthly rainfall values are vector quantities with both direction and magnitude, magnitude being the amount and direction being the month of the year expressed in units of arc. Vector direction for mean monthly rainfall is thus 015° for January, 044° for February, 074° for March etc. The next step is to add the twelve monthly vectors. The vector resultant is a measure of the seasonality of precipitation, its magnitude re-

FIGURE 6.2

Classification of South Africa in terms of the median annual rainfall total.

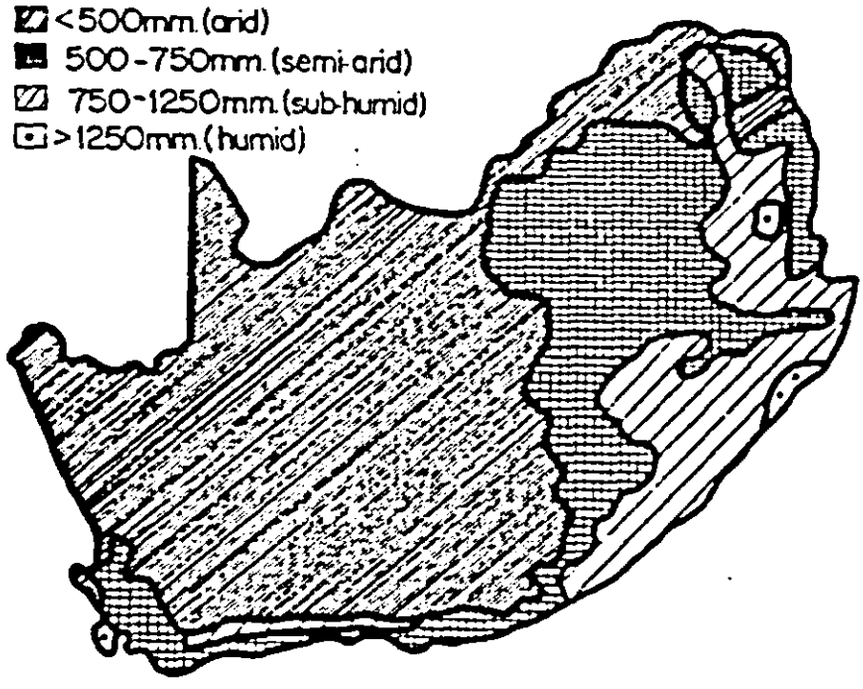
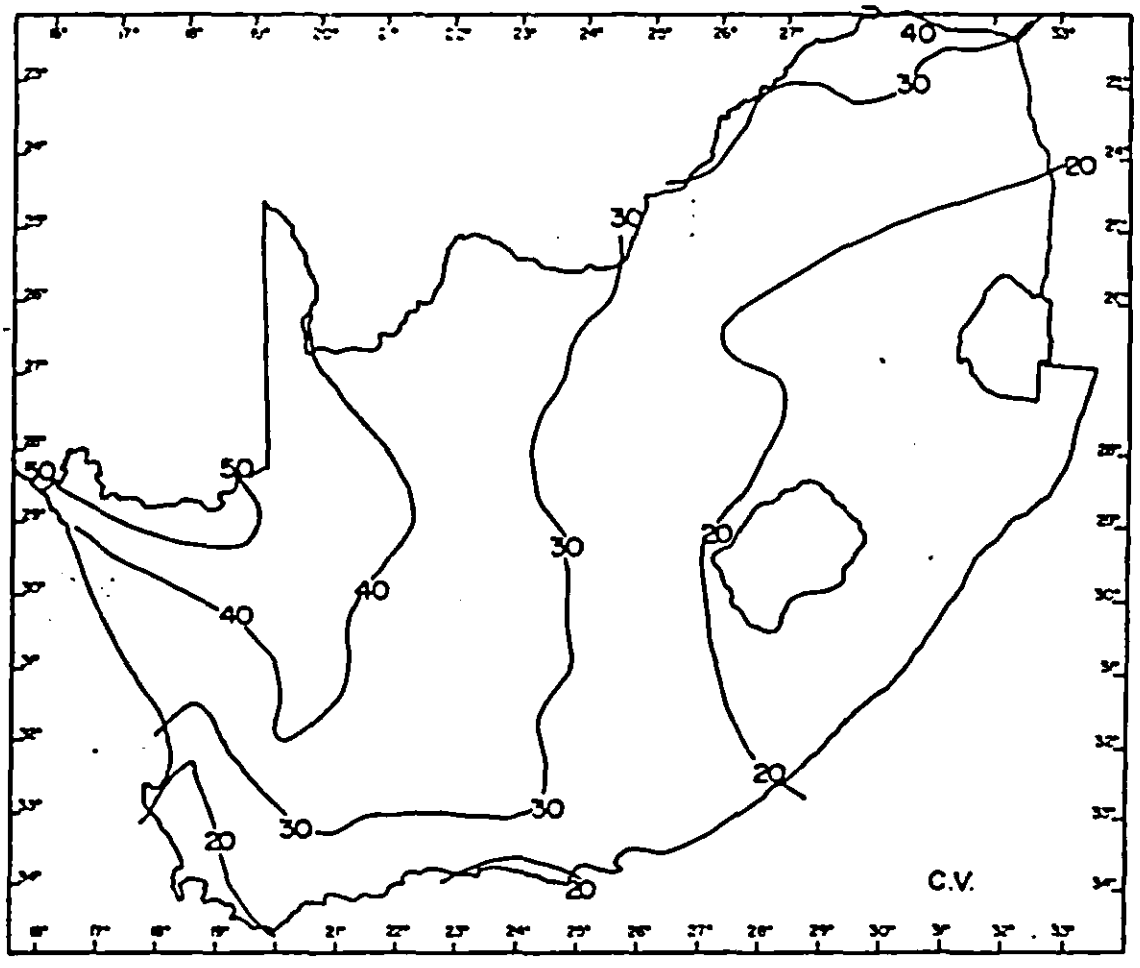


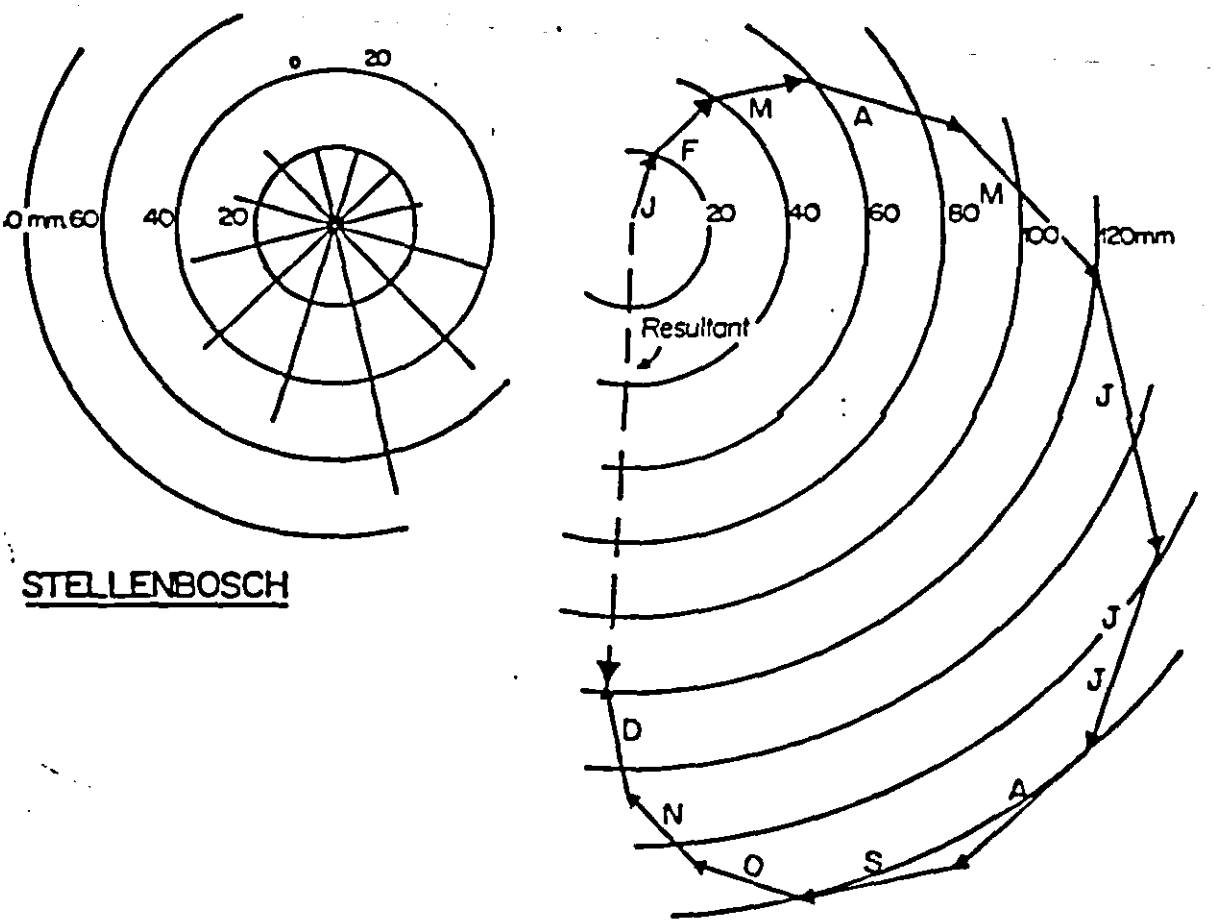
FIGURE 6.3

Coefficient of variation of the annual rainfall total.



FIGURES 6.4.1 and 6.4.2

Construction of the seasonality indices for Stellenbosch and Kakamas, following Markham (1970).



STELLENBOSCH

KAKAMAS

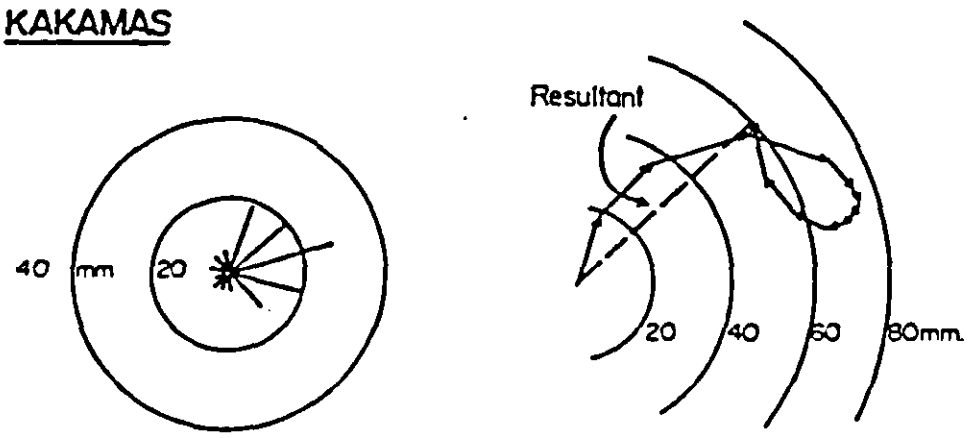
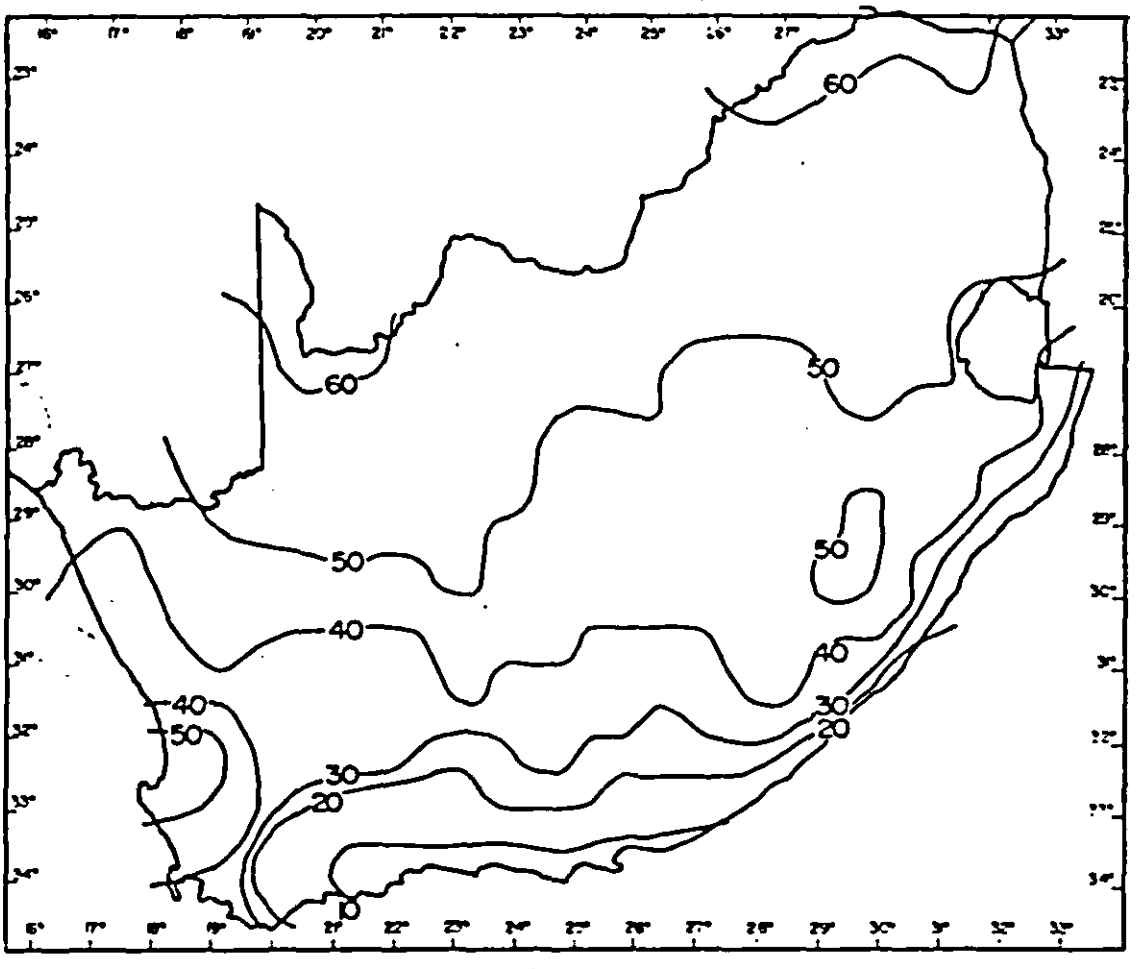


FIGURE 6.5

Contours of equal seasonality index.



presenting the degree of seasonality, and its direction the period of seasonal concentration. The ratio between the magnitude of the resultant and the total mean annual precipitation, expressed as a percentage, is called the Seasonality Index. The maximum possible value is 100% and would occur if all the precipitation came in a single month. The minimum value is 0%, occuring if precipitation is evenly distributed throughout the year. Two graphical examples of the construction of the index are shown for Kakamas and Stellenbosch. (Figures 6.4.1 and 6.4.2.) A map of South Africa is presented using the 500 sets of simulated data used above (Figure 6.5). Clearly high seasonality indices are associated with the more arid regions, reaching a maximum of 60%. The lowest values (<10%) are confined to the Southern Cape coastal region where rainfall can be expected all year round.

The direction of the vector could have been mapped to show the period within the year over which most rainfall can be expected. However, since we are attempting to portray the value of a simulation model of daily rainfalls, we can show this aspect of the rainfall climate in more detail. The truncated Fourier series as fitted to the 365 probabilities of a wet day and the 365 mean rainfalls allows us to estimate very simply (at each of the 2550 locations at which the model was fitted) that period when the probability of a wet day reaches a maximum and that period when the mean daily rainfall reaches a maximum. The two would not of necessity coincide and the shift may be of interest. Figures 6.6 and 6.7 show such mappings based on all 2550 points. We see, for example, that in Pretoria more wet days are expected in December but more rainfall in January, for Cape Town June would be expected to produce a greater amount of precipitation but August a greater number of wet days. The approach allows such

FIGURE 6.6

Period of the year when the probability of a rain day is maximum.

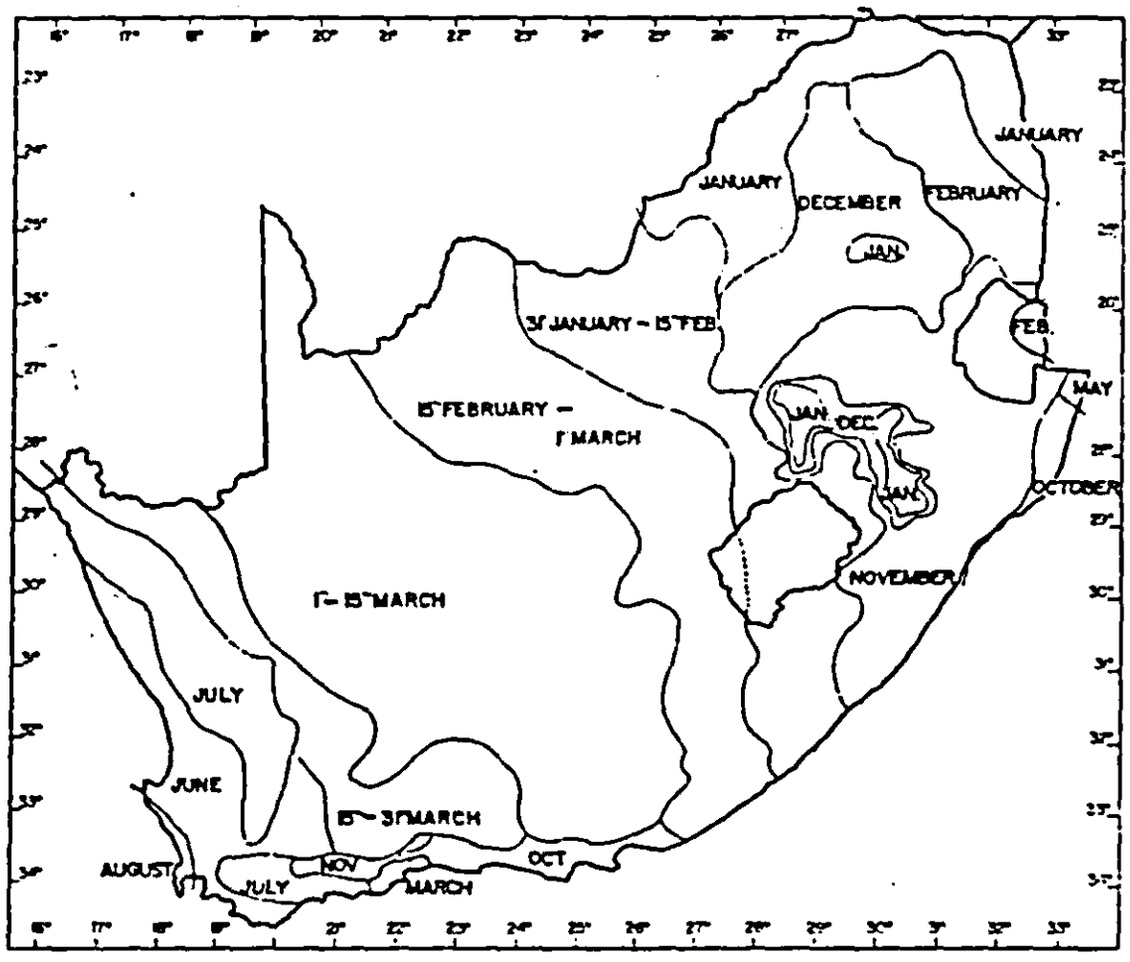
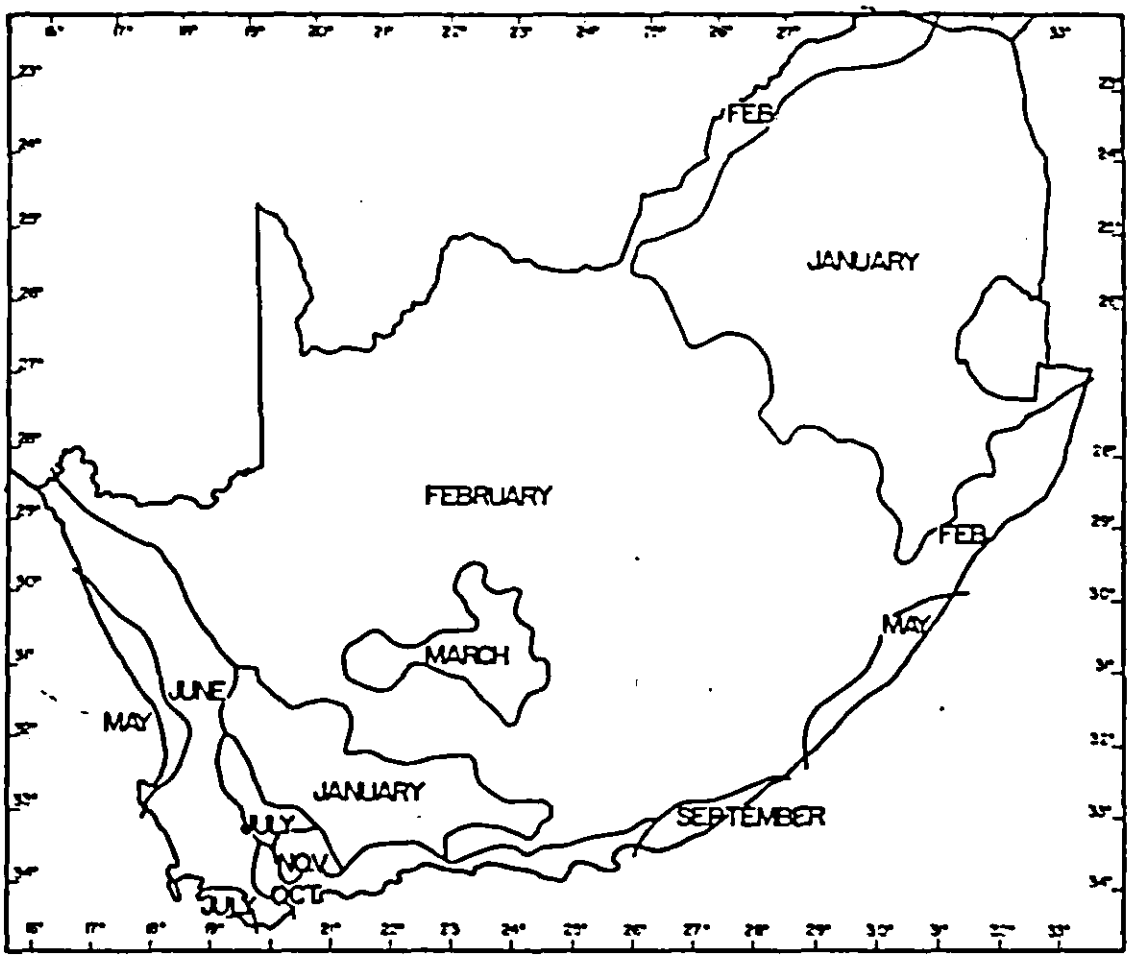


FIGURE 6.7

Period of the year when the mean daily rainfall is maximum.



detail to be derived from daily rainfall histories that within the Karoo regionalizations based on the period of maximum probability of a wet day could be made as small as 2 weeks. The distinction between the summer and winter rainfall regions is quite clear from such maps.

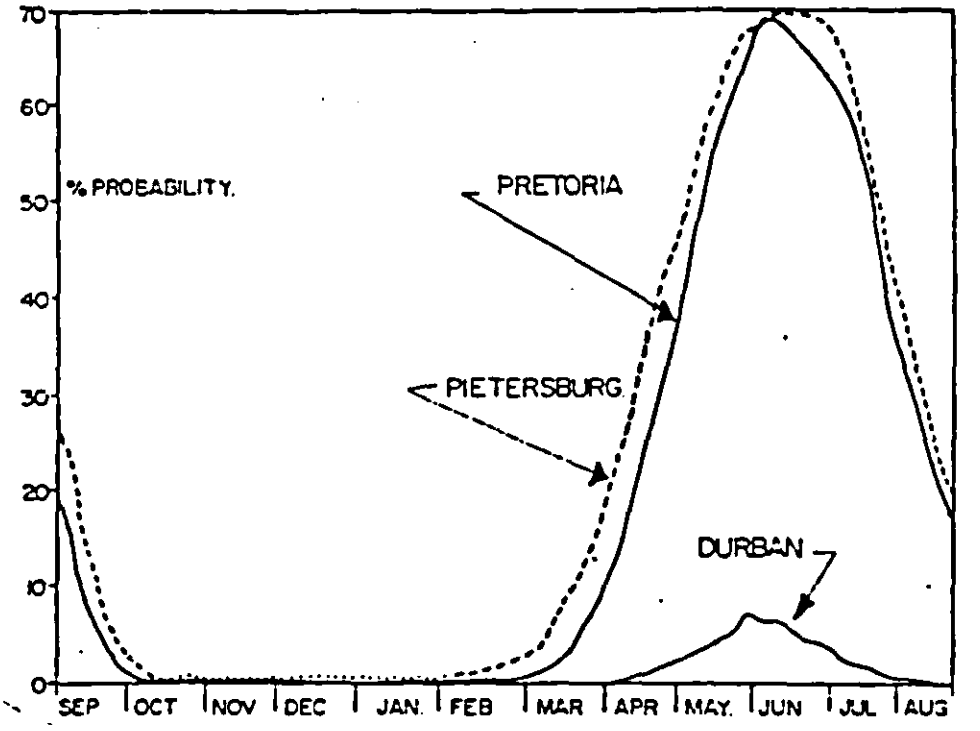
As already pointed out, the potential of the model in the assessment of regional rainfall regimes is very wide indeed. The criteria that may be proposed for regional discrimination can be tailored to suit the needs of a particular investigation with emphasis on agriculture, streamflow or some distinct aspect of the daily rainfall from the frequency of drought runs to the seasonality of storm rainfalls.

Some illustrative applications of the model to point rainfall characteristics

In many parts of the world the occurrence of long dry spells during the growing season of a crop is a major agricultural hazard (Stern and Coe 1982). Using the proposed model the probability of such periods of any arbitrary length can be simulated. Figures 6.8.1 and 6.8.2 show the seasonal probability of a dry run of 30 days starting on each day of the year from 1 September. 1000 years of simulated daily data were used to compute the result at each station. It can be seen that at Durban the probability of such a dry spell is comparatively low and confined to the period between April and July. The other results portray the distinct seasonality of the rainfall regime. Complementary to such a result would be one showing the probability of receiving more than x mm of rainfall over the next 30 days. We may be in just such a dry run and it would be useful to be able to estimate the probability that it will break over some future period of days. Figure 6.9 shows a simulation result for Pretoria and

FIGURES 6.8.1 and 6.8.2

Probability of a dry run of 30 days from a given starting date (abscissa).



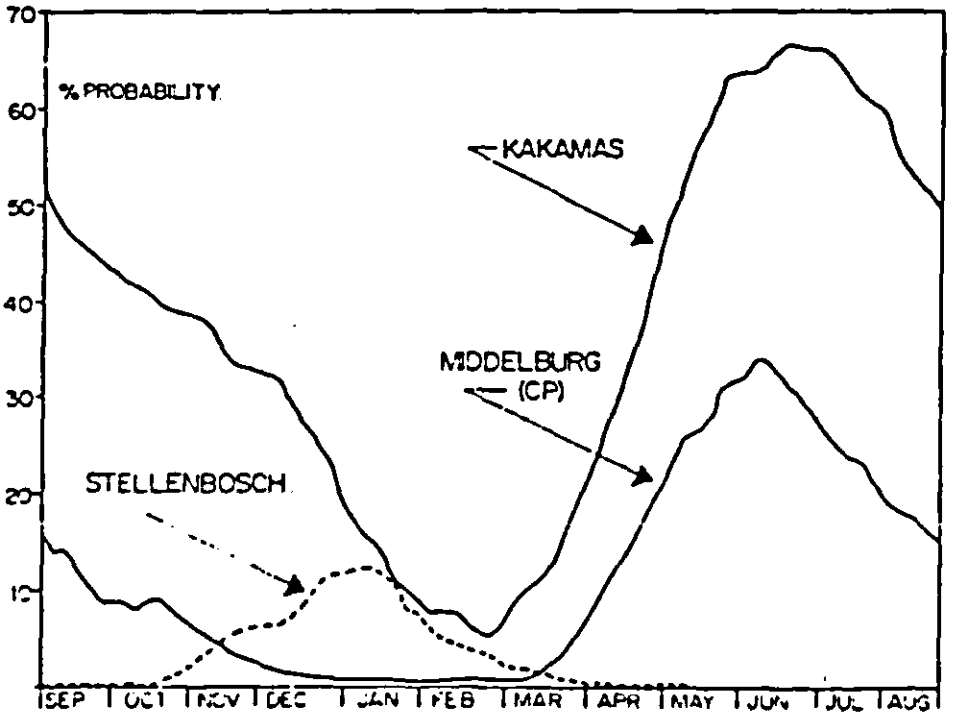
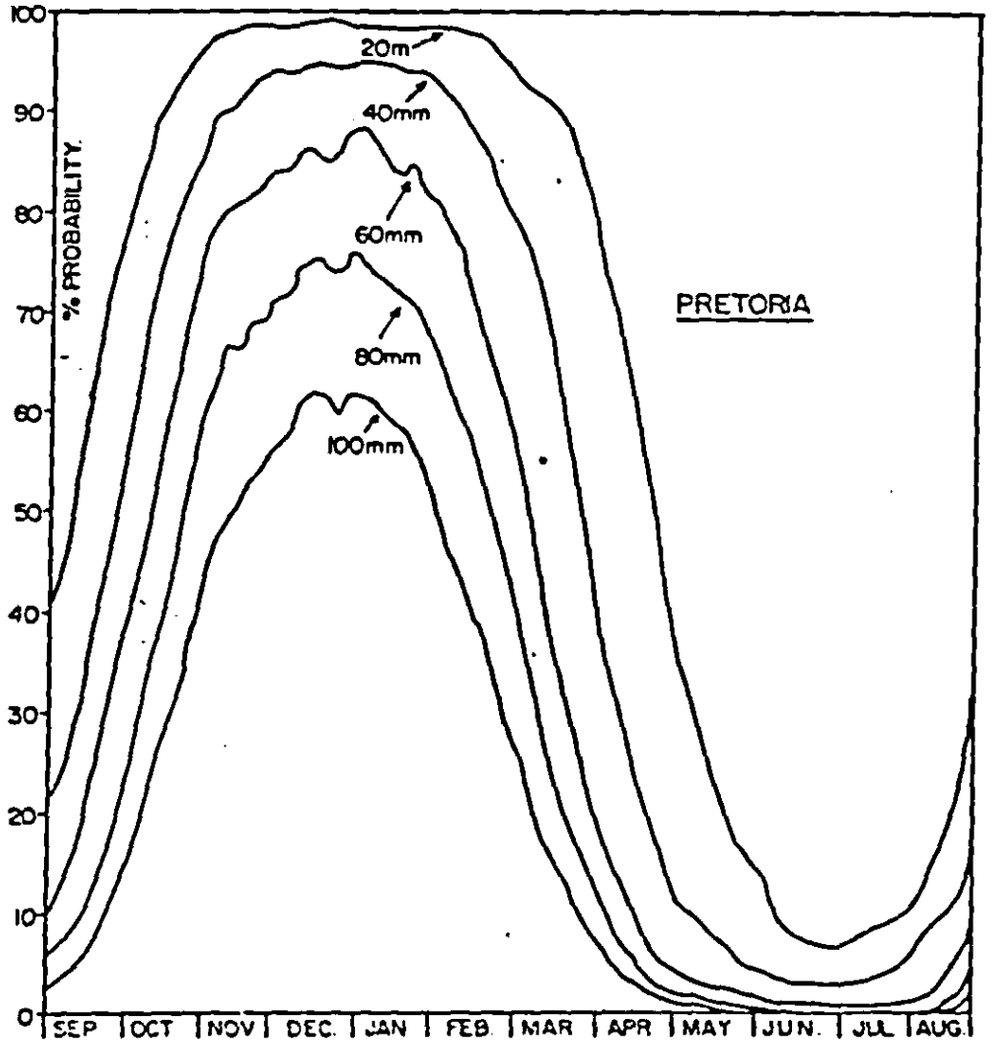


FIGURE 6.9

Probability of receiving more than 100, 80, 60, 40 and 20 mm rainfall over a 30 day period from a given starting date (abscissa).



the probability of receiving more than 100, 80, 60, 40 and 20 mm over the next 30 days starting on any given day of the year.

The yield of a crop is obviously related to the amount of precipitation that falls during some critical period of growth. Late-maturing cereal cultivars usually have the highest yield potential, but this is often not realized because of moisture stress during the grainfilling period, resulting in poor grain size or a high proportion of shrivelled grains (Dennet et al 1983). Grainfilling in maize takes approximately 30 days and the probabilities of receiving less than 50 mm during this period are shown in Figure 6.10 for crops maturing at any date at four locations within the major maize-producing region of South Africa. The time axis shows planting date and the probabilities refer to the 30 day period between the 70th and 100th day after planting. We see that for Lichtenburg in the Western Transvaal, "optimal" planting dates are reached as early as October whilst for the Northern Cape as represented by Edenburg, planting dates should be confined to December when the probability of grainfilling rains even so is significantly lower than those to be found in and near the Northern Orange Free State (Potchefstroom) and Western Transvaal (Lichtenburg).

In the Eastern Orange Free State winter wheat is an important winter crop, the success of which depends on two factors: firstly that the soil moisture content is sufficient to last the crop through the dry winter months, and secondly that early spring rains will boost yields subsequent to the exhaustion of soil moisture usually by the end of July. "Optimal" planting dates will in consequence depend upon the timing of the end of the summer rains and the likelihood of early spring rains. Figure 6.11 shows the

FIGURE 6.10

Probability of receiving more than 50 mm rainfall between 70 and 100 days (inclusive) following the given starting dates (abscissa).

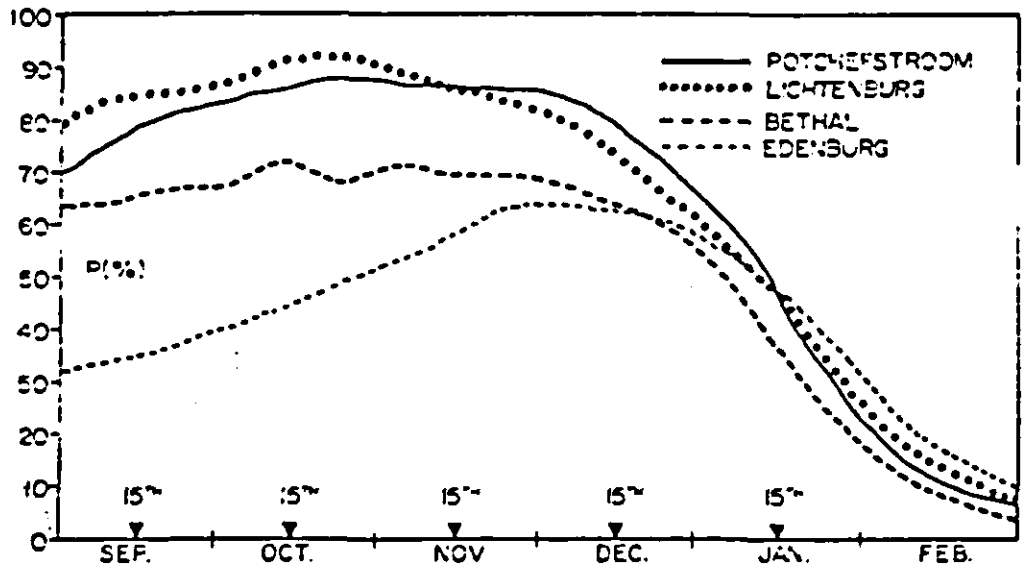


FIGURE 6.11

Probability of a dry run of 30 days : Vrede and Ficksburg.

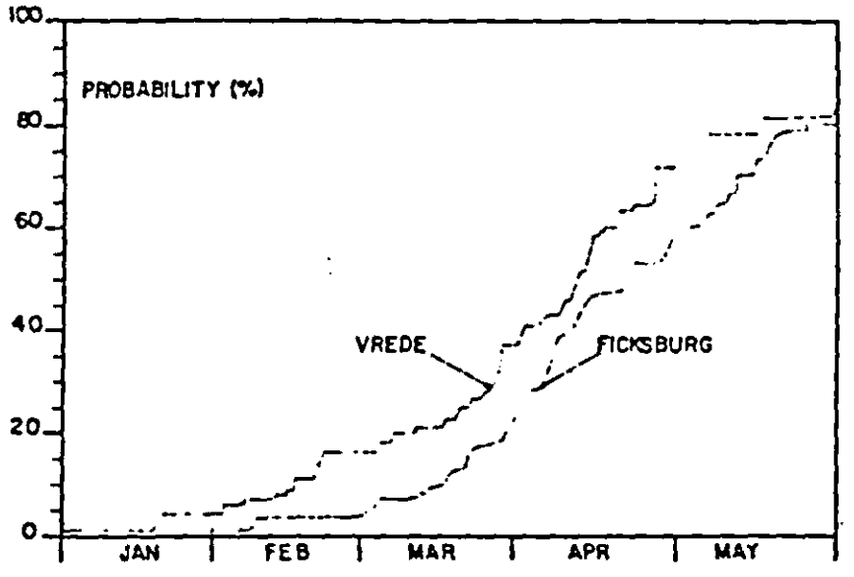
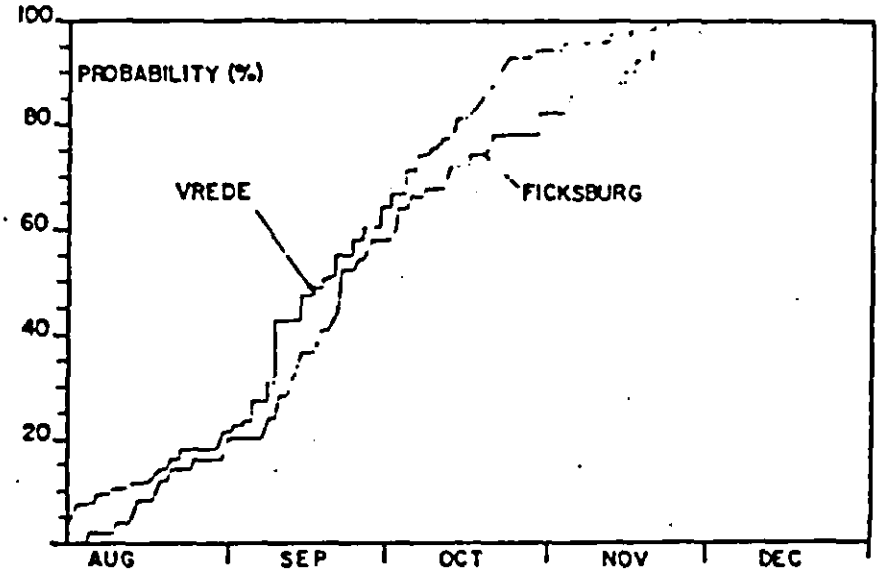


FIGURE 6.12

Probability of receiving more than 25 mm in 5 days or less: Vrede and Ficksburg.

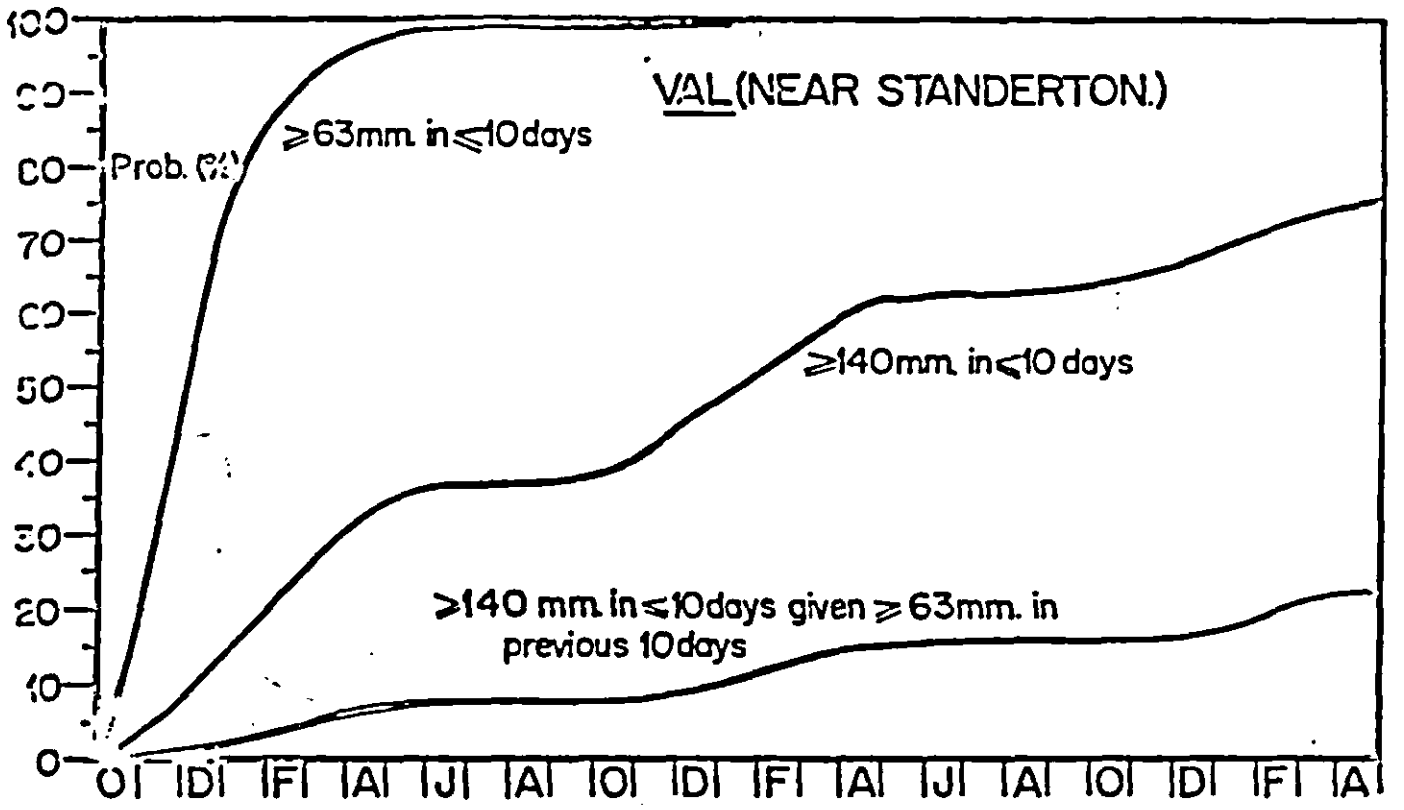


probability of the summer rains ending early at two locations in the Eastern Orange Free State as defined by a dry run of 30 days starting on 1 January. Such a dry run late in the summer season heralding an early start to the winter season would indicate a deficiency of soil moisture for the winter wheat crop. Figure 6.12 shows the probability of receiving more than 25 mm in 5 days or less starting on 1 August and for the same two stations. Such a rainfall is generally considered to announce the beginning of the summer rainfall season. From the two graphs it can be seen that the "dry season" starts 2 to 3 weeks earlier at Vrede than at Ficksburg whilst spring rains can be expected less than a week later at the latter with the same probability.

A most important aspect of drought is the magnitude of storm event that effectively "breaks" the rainfall deficiency. For reservoirs suffering from low storage levels only a flood-producing sequence of rainfalls over a period of several days is likely to contribute towards an effective recovery of storage levels. The recipe for such an event is quite clear, being several days of soaking rains over the catchment to provide antecedent conditions for a subsequent sequence of storm events that will generate the best possible level of surface runoff. The antecedent rainfall will saturate the soil moisture profile such that the later storms will provide considerable volumes of streamflow rather than be absorbed into soil moisture storages. Such events generally "broke" the drought of the early thirties over the Vaal catchment in early October 1933. At Val, near Standerton, 63 mm was recorded in a 10-day period followed by 140 mm during the subsequent ten days. Given Vaal Dam to be in a deficient state of storage it is of interest to know the probability of receiving such a storm sequence over a forecast period of interest, say the next three seasons. Figure 6.13 shows the results of such

FIGURE 6.13

Probabilities of three specific storm sequences starting from 1 October.



a simulation and reveals that the thirties drought was indeed broken by a particularly rare sequence of storm events. Although the antecedent condition would almost certainly occur before the end of the first season, its combination with a subsequent storm period yielding 140 mm in ten days is unlikely, being only 20% after three seasons.

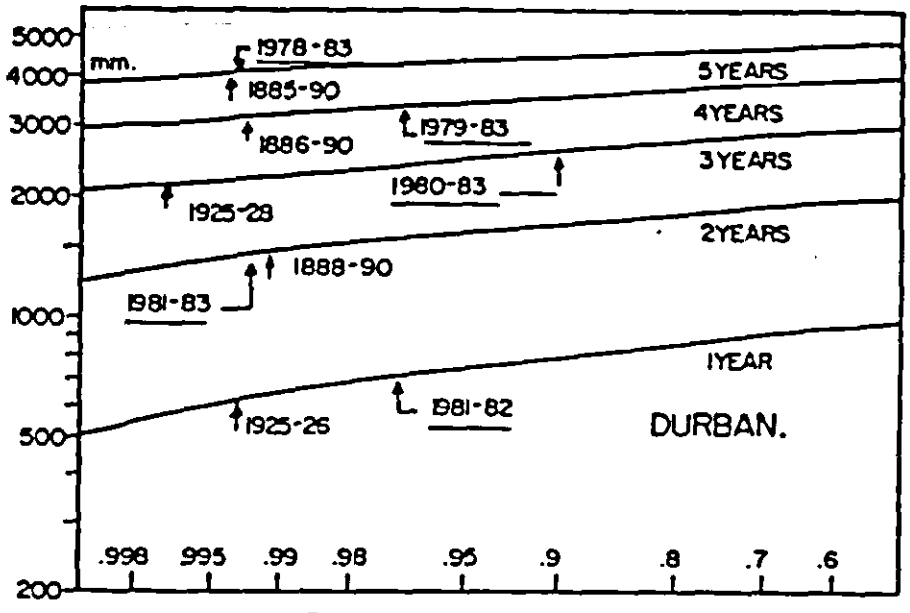
The drought history of South Africa

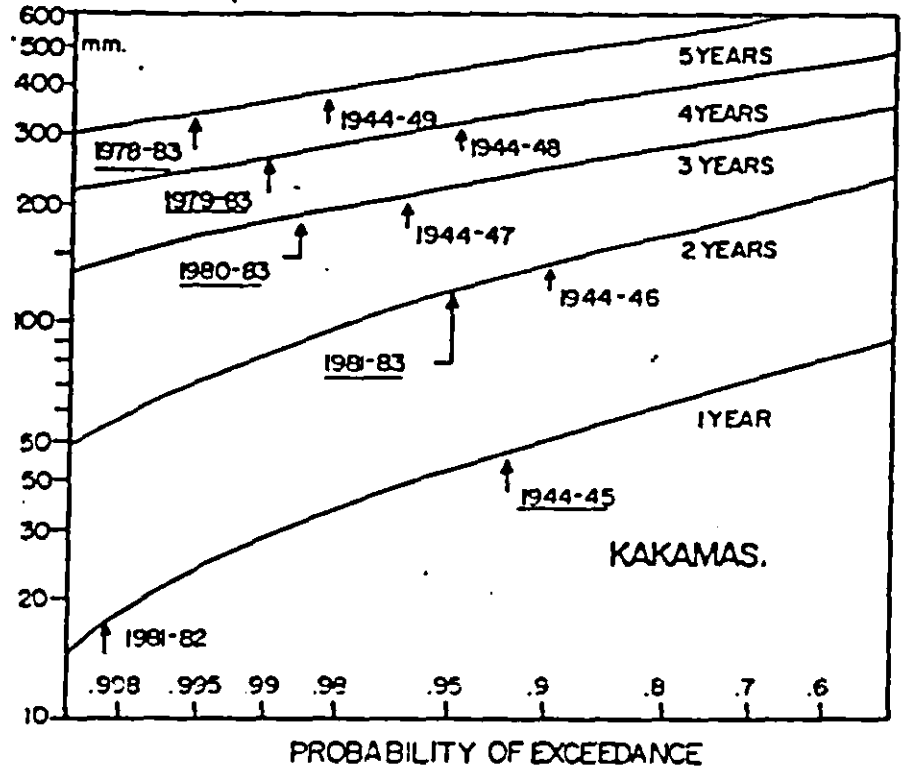
In order to review the recent chronology of drought in South Africa and to place historical and current events in perspective, two approaches were made. The first illustrates a simple application of simulation results from the model, and the second looks at the spatial history of drought using the historical data.

For each of the six stations upon which we have been concentrating so far for illustrative purposes, namely Pretoria, Durban, Stellenbosch, Kakamas, Middelburg and Pietersburg, 1000 years of data were simulated and the distribution functions of n-year totals from 1 to 5 years plotted. On each plot the two worst n-year runs are shown (Figures 6.14.1 to 6.14.6). Three periods dominate, namely the present drought (1978-1983), the mid-forties and the early nineteen thirties. With the exception of Stellenbosch which being in the winter rainfall region is unaffected by the present drought, the last five-year period (1978-1983) is seen to contain the majority of the driest runs recorded. A characteristic of the current drought is its duration and although there have been more severe one- and two-year periods, it does represent one of the worst periods as viewed in probabilistic terms. An interesting comparison is that between 1885-1890 and 1978-1983 at Durban. In the North Western Cape as represented by Kakamas, the last five-year run is by far more deficient from a rainfall point of view than anything recorded

FIGURES 6.14.1 and 6.14.2

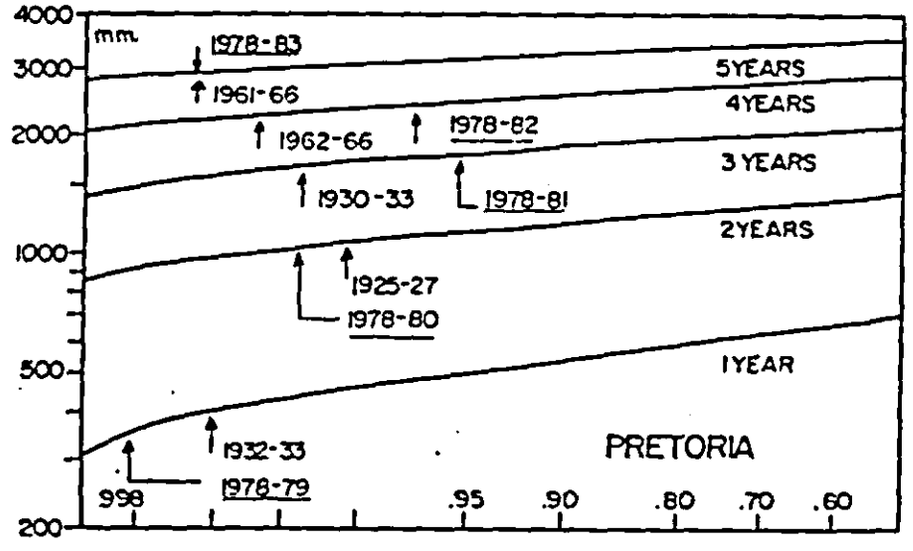
The most severe historical n-year droughts and their estimated probability of exceedance.

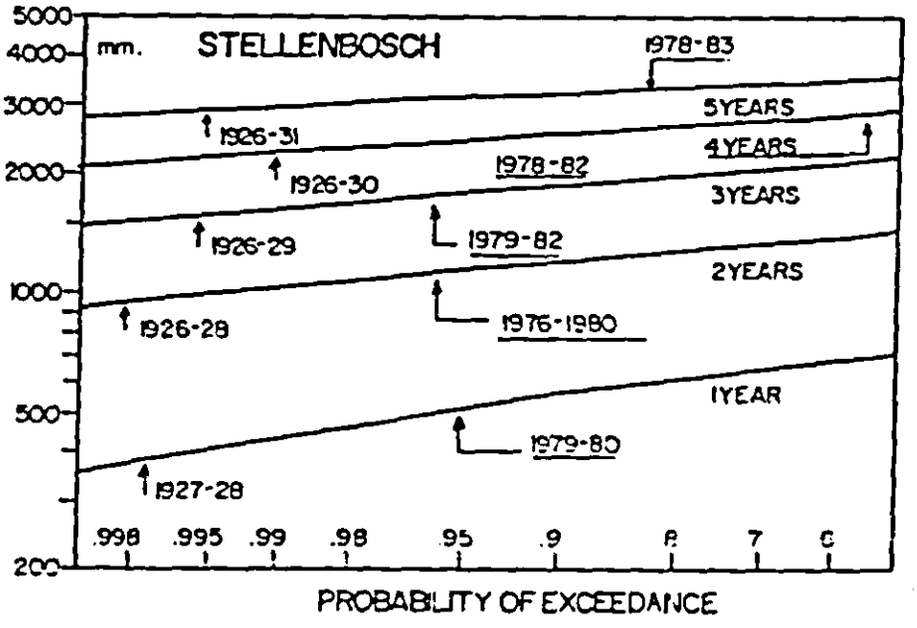




FIGURES 6.14.3 and 6.14.4

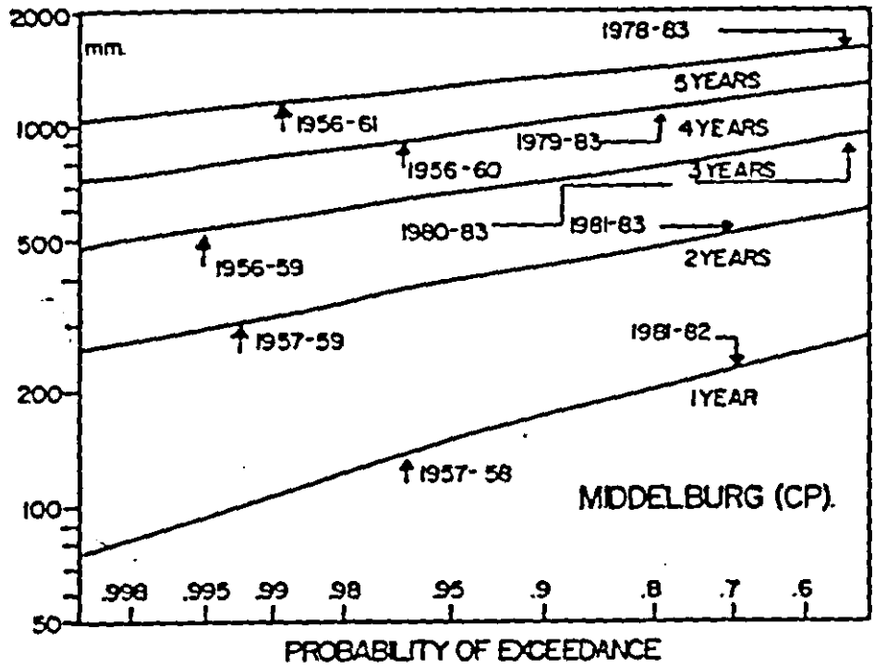
The most severe historical n-year droughts and their estimated probability of exceedance.

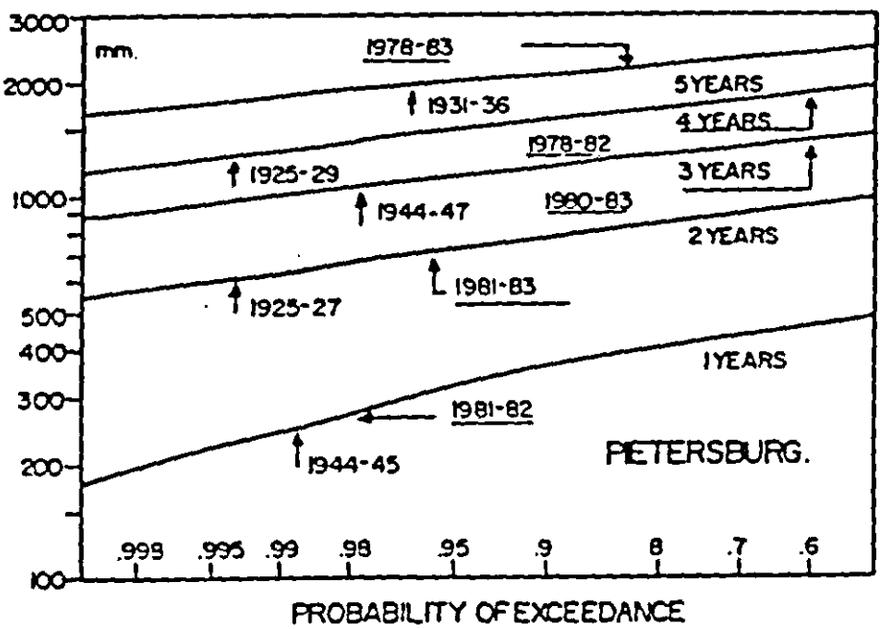




FIGURES 6.14.5 and 6.14.6

The most severe historical n-year droughts and their estimated probability of exceedance





historically. At Middelburg (Cape Province) the late nineteen fifties represents the driest period on record, whilst at Pretoria although 1978-1979 represents the driest single year on record, the period from 1978-1983 is only equally as bad as that from 1961-1966. In the Northern Transvaal (Pietersburg) the period from 1931-1936 was in fact drier than the severe drought being experienced in the area at present.

Such a view of drought at a point, useful as it is, fails to give any indication of the spatial nature of the deficiency and an attempt to do just this is now made. Five hundred and fifty locations were chosen at regular intervals over South Africa and the history of spatial drought was mapped on an annual basis (October-September) from 1920-1921 to 1979-1980. In order to overcome distributional problems a different approach was used in which percentile values were computed and the annual rainfalls classified as being above or below these selected percentiles. The univariate model selected for the study is the Gamma which has found wide application in the study of precipitation totals (Barger and Thom 1949 , Gupta and Panchapakesan 1980 , Mooley and Crutcher 1968 , Neymann and Scott 1967 , Shenton and Bowmann 1973) and was separately validated within the present study. The mean length of record used was 54 years with a minimum length of 37 years. The reliability of precipitation probabilities estimated from the Gamma distribution for such sample sizes is quite high (Bridges and Haan 1972) and the risk of considerable error greatly reduced. For each sample of annual rainfall the Gamma parameters were estimated by maximum likelihood and the historical record then screened year by year. Yearly totals that fell below the 50%, 20% and 5% percentiles were noted and the areas suffering such a level of drought mapped, one map for each year. Simplistically we can view these percentiles as

representing the 1 in 2, 1 in 5 and 1 in 20 year annual rainfall deficiency. The sequence of maps so obtained is shown in Appendix 5. It became apparent during the mapping exercise that the method of approach was reasonably successful with juxtaposed stations showing very similar percentile levels. When this did not happen the historical record was examined, generally to reveal a short intense cluster of rainfall days which boosted the annual total. That this should occur and be quite localized is to be expected in a region where summer rainfall is largely a consequence of local instability and convection and the associated thunderstorms. Thus regional drought is not necessarily broken by one or a few recording stations reporting relatively high annual figures since these almost certainly would be the result of a very limited number of storms, the generating mechanisms of which were quite localised.

In viewing the sequence of sixty maps a number of physiographic controls are apparent in their influence on the areal extent of rainfall deficiency, the most obvious being the Eastern Transvaal and Natal escarpments. Another is the Outeniqua mountains of the Southern Cape coastal belt. The maps bear close scrutiny in conjunction with the work of Harrison (1983) who, using principal components analysis, proposed a generalised classification of South African rain-bearing synoptic systems. These systems control the regional rainfall and it is qualitatively apparent that areas with one dominant rain-bearing mechanism generally coincide with areas where historically rainfall deficiencies have tended to be centred. A major feature of the maps is the dominance of deficiency along a north-south axis over the centre of the country. It is attractive to associate this feature with the major rain-bearing disturbance which normally occurs over the centre of South Africa and is the major contributor to summer rainfall over

the interior. This is a system of cloud bands which connect the tropical circulations with mid-latitude cyclones and which satellite imagery has generally shown to be weak or absent during drier periods(Harrison 1983). The deficiency of particular rainbearing systems such as cyclonic disturbances over Natal no doubt accounts for serious drought confined to this particular region.

The major value of the maps is that they provide a view of the extent, severity and frequency of sub-continental drought from 1920. The first period of such deficiency begins in 1925-1926 with severe shortfalls in annual precipitation over most of the central interior with the exception of the Eastern Transvaal highveld and lowveld. We note that Natal and the South Western Cape show no deficiency during 1925-1926, confirming the fact that over these regions the rainfall generating mechanism is distinct from that over the major part of the country. The drought tended to linger over the southern Cape interior during 1926-1927. During 1927-1928 the severe drought that affected the South Western Cape (cf. Figure 6.14.1 for Stellenbosch) is apparent with the central interior of Natal and Transvaal also suffering shortfalls.

The drought of the early nineteen thirties reached maximum development during the 1932-1933 season, with vast areas of the country severely affected whilst large areas were similarly affected during the 1948-1949 season. An interesting period is that from 1963-1964 to 1965-1966 where the influence of the Eastern Transvaal escarpment is clearly apparent. Over the Northern Transvaal in particular these years represent some of the driest experienced.

The worst drought in the South Western Cape from a water resources point of view occurred during the 1972-1973 season. The event was very largely confined to the coastal regions

and the southern interior with serious shortfalls in dam levels. Over most of the interior the early and mid seventies were particularly wet with serious flooding during the 1974-1975 season. The present drought is seen to have developed over the Northern Transvaal during 1978-1979 and over Natal during 1979-1980.

The maps illustrate the spatial nature of interannual rainfall variability over South Africa and show that drought is an inherent part of the climate. The search for a source for such variability has attracted considerable attention given the seriousness of the current drought. Much research in South Africa and elsewhere has given support to the results of Arkin (1982), Pan and Oort (1982) and Winston (1982) which found that the aperiodic warmings of the equatorial tropical sea surface temperatures can have a dramatic impact on the planetary scale circulations. Rasmussen and Carpenter (1982) have shown that these warmings may take 12-18 months to develop from outset to maturity to their final disappearance, with the area of the ocean which is warmer than normal of the order of 5-10% of the earth's surface. That these aperiodic sea surface temperature anomalies have a predictable quasi-periodicity, and therefore may be useful in the prediction of the outset of perennial periods of drought, is an attractive field of research - particularly as streamflow and annual rainfalls from Indonesia to India to South Africa have an apparent quasi-periodicity (Quinn *et al* 1978 , Angell 1981 , Tyson and Dyer 1978). From a purely statistical viewpoint such quasi-periodicity is rarely apparent from the correlogram of the annual time series of interest; filters, usually in the form of moving averages, are resorted to.

The process of filtering introduces serial correlation in the resulting time series (even if the original series

is serially uncorrelated) and in consequence this manipulation alone often leads to the creation of pseudo-cycles. This fact constitutes one of the main obstacles in establishing the existence of true cycles.

In order to investigate the possible existence of drought cycles from a somewhat novel point of view the total areas covered by each of the three severity categories were determined by means of a planimeter for each of the 60 maps given in Appendix 5. Table 1 gives the areas, represented as a percentage of the total land area of South Africa, corresponding to the 50%, 20% and 5% percentiles respectively. The serial correlation coefficients were estimated for lags 1 to 20 and are given in Table 2. The critical values (95% significance level) for the null hypothesis that a particular (population) serial correlation coefficient is zero are approximately $-2/\sqrt{60}$ and $2/\sqrt{60}$ (see Box and Jenkins 1970), i.e. -0,26 and 0,26. The only estimate which falls in the critical region is that of lag 19, for the 20% percentile series; a few others are quite close to the critical values. Unfortunately this does not really establish that the serial correlation for this lag is significantly different from zero. It would have done so if we had specified that lag 19 was the (single) lag under consideration *a priori*. What we actually did, however, was to compute all the serial correlations from lags 1 to 20 and then we looked around for the largest one. Now the distribution of *largest* estimated serial correlation coefficient under the null hypothesis is quite different from that of any of the individual ones and leads to quite different critical points. Only very coarse approximations based on Bonferroni bounds are available and it can be established for example that the upper critical point is lower than (approximately) 0,45*.

---

*We wish to thank Professor C.G. Troskie, Department of Mathematical Statistics, University of Cape Town for carrying out this computation.

But whatever the exact critical value may be, it is certain that it will be a good deal larger than the observed value of 0,27. We have therefore insufficient evidence to establish the existence of a 19-year or any other cycle. Interestingly enough Thompson (1981) postulates a 19,2 cycle based on data from 18 meteorological stations between Margate in the south and Hluhluwe in the north and as far inland as Greytown.

A somewhat puzzling feature of the data in Table 1 and which we are not able to explain is that the means for the three series are 41,17%, 11,68% and 2,23%. We would have expected these to be closer to 50%, 20% and 5%. A test of the hypothesis that the observed averages differ significantly from these values is difficult to construct in this case because there is a great number of complicating factors on which the distribution of a suitable test statistic would depend, e.g. the crosscorrelation coefficient between each of the stations considered, the relative "coverage" of each station, and so on. It is therefore difficult to gauge precisely how unlikely these averages are.

TABLE 1

Year	50%	20%	5%	Year	50%	20%	5%
1920/21	26,46	4,30	0,19	1950/51	61,64	8,84	1,08
1921/22	53,30	17,13	2,69	1951/52	69,66	28,07	3,11
1922/23	42,59	4,49	0,00	1952/53	39,98	7,87	0,77
1923/24	32,83	9,60	1,19	1953/54	26,11	5,80	0,92
1924/25	10,64	2,46	0,50	1954/55	29,99	5,68	0,51
1925/26	81,14	36,33	15,48	1955/56	26,34	2,92	0,23
1926/27	79,80	20,31	3,92	1956/57	23,66	1,38	0,00
1927/28	33,91	11,98	2,73	1957/58	38,59	12,25	1,27
1928/29	24,62	5,38	0,61	1958/59	32,41	4,03	0,65
1929/30	64,48	12,56	0,65	1959/60	47,89	17,13	2,00
1930/31	70,93	15,86	2,73	1960/61	15,32	1,23	0,50
1931/32	72,89	15,40	1,57	1961/62	37,14	11,83	1,61
1932/33	91,67	60,56	18,86	1962/63	25,19	3,57	0,00
1933/34	33,41	4,38	0,08	1963/64	53,30	15,51	2,11
1934/35	41,74	12,48	0,31	1964/65	34,98	8,60	0,77
1935/36	65,82	17,97	0,69	1965/66	6,34	32,10	12,25
1936/37	44,66	5,57	0,54	1966/67	20,24	1,80	0,00
1937/38	46,58	8,91	1,23	1967/68	46,16	16,28	4,76
1938/39	13,48	2,65	0,12	1968/69	45,55	15,02	1,84
1939/40	27,65	4,69	0,81	1969/70	59,41	15,17	2,96
1940/41	39,29	8,03	0,88	1970/71	28,80	5,38	0,88
1941/42	63,10	18,09	2,88	1971/72	28,57	5,65	0,00
1942/43	16,40	1,46	0,04	1972/73	41,05	14,02	3,38
1943/44	21,66	3,46	0,00	1973/74	7,83	0,77	0,00
1944/45	90,32	27,23	4,95	1974/75	9,87	1,69	0,00
1945/46	60,22	15,67	3,76	1975/76	5,07	0,42	0,00
1946/47	76,46	19,47	4,11	1976/77	6,76	1,19	0,00
1947/48	23,50	3,88	0,35	1977/78	31,68	6,30	1,34
1948/49	91,74	36,18	13,44	1978/79	64,32	20,31	1,69
1949/50	19,97	2,69	0,15	1979/80	45,12	21,12	3,49

Percentage (by area) of South Africa with total annual rainfall below the 50%, 20% and 5% percentiles of the local annual total distribution.

TABLE 2

lag	50%	20%	5%
1	19	- 8	-10
2	8	10	7
3	3	2	- 9
4	21	2	- 2
5	10	-11	- 7
6	2	-18	- 7
7	- 4	9	25
8	-18	-24	-17
9	10	- 5	- 9
10	5	-16	-19
11	8	- 5	-10
12	-10	- 8	- 5
13	3	6	- 2
14	6	11	4
15	17	2	- 6
16	10	20	23
17	- 7	4	8
18	11	- 3	- 5
19	24	27	16
20	2	3	2

Estimated serial correlation coefficients (times 100) for the three time series given in Table 1.

7. A FAMILY OF DROUGHT INDICES

Fundamentally drought is a deficiency of rainfall relative to water requirements. This shortfall manifests itself in many different ways from crop failure to deficiencies in reservoir levels. Its impact is directly related to its duration and severity as well as its temporal nature. For example a three-week dry run during the active growing season of a commerical crop may constitute a particular agricultural drought whereas for a system of reservoirs such a shortfall may be insignificant. This implies that any single universal definition of drought would be inappropriate for all but a few water consumers.

Ideally the definition of drought is user specific but even at this level the problem is complex. For example, a particular farmer may grow two different crops each of which demandsdifferent water usage at different times of the year; an industry may consume water for different purposes, for example for cooling and for the dilution of effluent. It is not feasible in a single study to analyse the specific requirements associated with each of the different water-related activities, but on the other hand it is clear that a single notion of drought is inadequate. We need a simple working tool which is sufficiently flexible to accommodate a range of situations. We propose a family of drought indices as a minimal requirement. The particular family we consider is one of many that are possible but it enjoys the kind of properties that are associated with the commonly accepted image of the drought process. Furthermore the indices have analogies in nature, for example streamflow, soil moisture, etc ... .

Essentially a drought index is some measure of wetness/dryness at each point in time which simultaneously takes

account of the antecedent conditions. Mathematically such an index is constructed by passing the primary rainfall process through a suitable filter. In the next section we will discuss a particular type of filter for constructing a general measure of wetness/dryness. Examples of application are discussed in Chapter 8.

7.1 LINEAR FILTERS; THE EXPONENTIAL FAMILY OF DROUGHT INDICES

An ideal way to measure drought severity would be by proxy where although rainfall is the driving mechanism the impact is expressed in terms of crop yield, reservoir storage, soil moisture storage, water table level, streamflow, etc .. . In practice, however, historical records are either unavailable or too short for assessing the risk associated with a particular drought event and until such records become available one will be obliged to work with rainfall data. By suitable selection of a filter applied to the primary driving process, i.e. rainfall, we can imitate in part at least the secondary or tertiary process of interest. The precise way in which a process such as water table level is related to rainfall is complex and is a function of climatic, physiographic and geological factors as well as soil conservation practice and other human activities. The expense of accurately determining the precise nature of such relationships, many of which we do not yet clearly understand and all of which differ from place to place, is considerable. It is more appropriate (at least as far as the assessment of drought risk is concerned) to employ some simple approximating relationships which preserve the general character of the process of interest. In this respect the filter described below is well suited to a wide range of practical situations.

In what follows the basic unit of time is taken to be one day. This is done for convenience; any other unit of time

would also do. By the "response" to a particular rainfall event we mean the contribution, over time, due to that event to the level of the process. The term process here is used to represent the level over time of some variable of interest, such as streamflow, soil moisture and so on.

In constructing a measure for one of the processes mentioned above it has to be kept in mind that the response to a rainfall event continues after the event. For example, streams continue to flow, soil continues to contain water after it has stopped raining. In other words, the level of the process on a given day depends not only on the amount of rain which falls on that day but also on the amounts which fell on preceding days. It is also quite obvious that the level of the process on a given day does not depend on rainfall events which occur on subsequent days, that is, the response function is zero for such events. This observation may appear trivial but this condition is not satisfied for ordinary moving averaging, a common index for the state of wetness/dryness particularly in studies relating to drought cycles. A further property of the response function is that it decays with time. The time taken for it to reach zero (or some negligible quantity) can be as short as a few hours for a process such as overland flow or it can be as long as several months for soil moisture or base flow.

The model which we construct is based on three further conditions. Strictly speaking these are not met by the processes under consideration which are in fact much more complex than is implied by the conditions. The model must therefore be regarded as an approximation. Its accuracy will, for each process, depend on the degree to which the following conditions are met:

(a) LINEARITY: This requires that the level of the process on a given day be the sum of the individual contributions due to rainfall events up to and including that day and that these contributions do not themselves depend on the level of the process.

(b) PROPORTIONALITY: The contribution to the level of the process due to a particular rainfall event is proportional to the rainfall depth.

(c) STATIONARITY: The contribution to the level of the process on day t due to rainfall on day j depends only on the interval (t-j) and not on the particular times t and j. (We note that this does not mean we are assuming that the process must be stationary, but only that the response function does not depend on the time of the year.)

Processes which meet condition (a) are called linearly filtered processes. Conditions (b) and (c) define particularly simple types of such processes. We denote the rainfall depth by R(t) and the level of the filtered process by F(t) where t denotes the day starting from some arbitrary origin. It can be shown that filtered processes which satisfy the above conditions can be represented in the form

$$F(t) = \sum_{i=0}^{\infty} r(i)\, R(t-i) \qquad (1)$$

where r(x) is the response function which we have assumed to have the properties

$$r(x) = 0 \quad \text{for} \quad x < 0, \qquad (2a)$$

$$\lim_{x \to \infty} r(x) = 0 \; . \qquad (2b)$$

Strictly speaking we must also require that $\sum_{x=0}^{\infty}|r(x)|$ is finite. For the types of processes which we consider this requirement is always met.

The response function, r(x), describes the contribution to the level of the process which arises from a unit of rainfall x days after the rainfall occurs. For example the response function associated with streamflow is the well-known unit hydrograph, for soil moisture there are different response functions depending on the depth and type of soil. Near the surface the response decays exponentially; deeper down the decay is generally not exponential. If the response function were known then it would be possible to calculate F(t) from rainfall records using equation (1). In many situations r(x) is not known but, given some direct measurements on the process F(t), it can be estimated. For example if soil moisture measurements over even a short period of time are available and a long rainfall record is also available then one can use the concurrent records to estimate r(x) and thereby make use of (1) to estimate F(t) over the whole period for which rainfall data are available. A suitable estimation procedure is given in Appendix 4.

It is not necessary to associate a particular physical process to models of the type given in (1); they can also be used as general measures of the state of wetness/dryness. Naturally the properties of each measure depends on the particular response function (or filter) which is used. For example a simple filter is $r(0) = r(1) = \ldots = r(L-1) = 1/L$ and $r(x) = 0$ for $x > L$. This is called a rectangular filter and is illustrated in Figure 7.1. Equation (1) then reduces to

$$F(t) = \frac{1}{L}\sum_{i=0}^{L-1} R(t-i) \quad , \quad t = L, L+1, \ldots$$

FIGURE 7.1 Rectangular filter

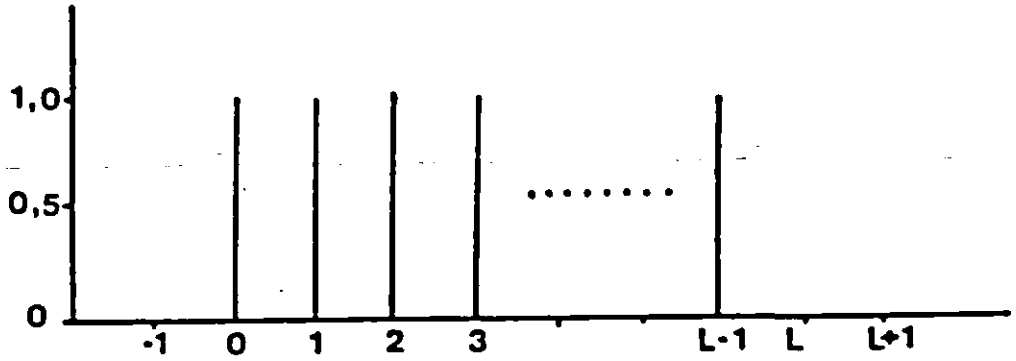


FIGURE 7.2.1 Exponential filter with ρ = 0,3

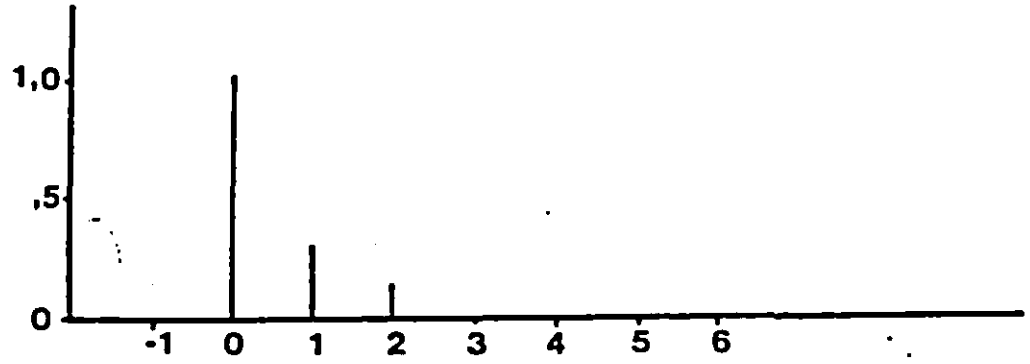


FIGURE 7.2.2 Exponential filter with ρ = 0,8

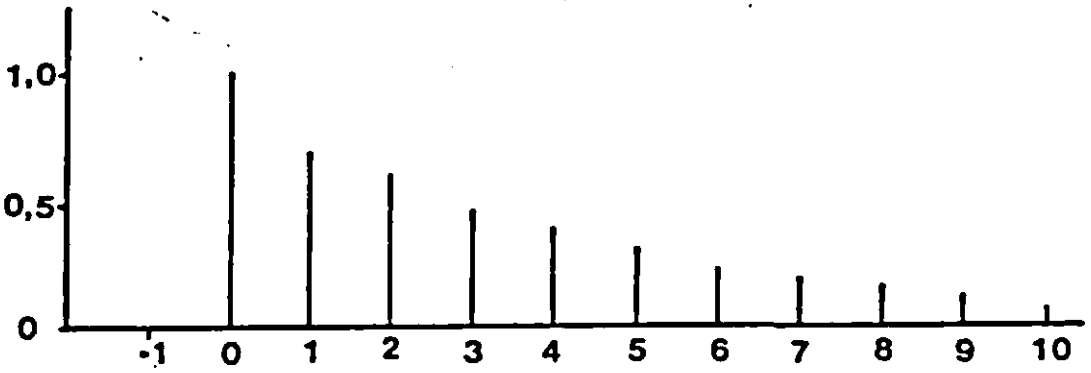
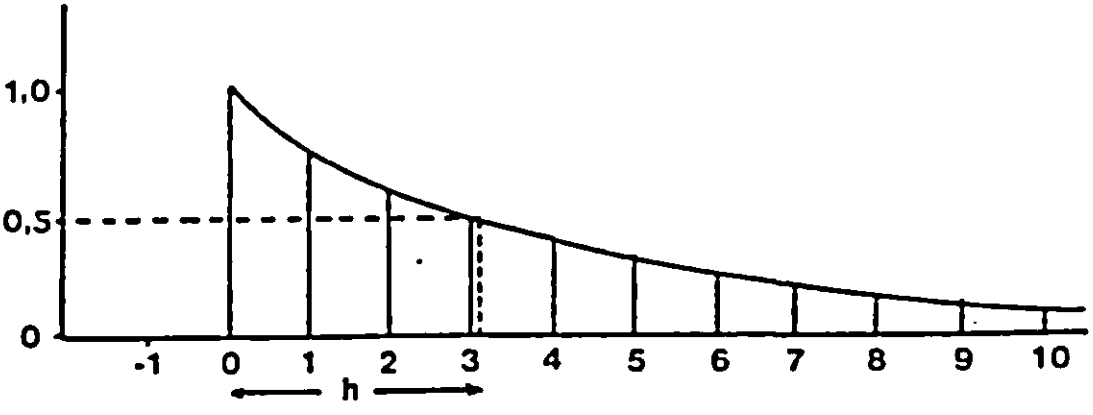


FIGURE 7.3 Half-life of exponential filter



that is, the quantity F(t) describes the average rainfall over days t to t-L-1 (inclusive). The so-called "bandwidth", L, can be selected to suit one's specific purpose.

Any general index of the state of wetness/dryness is essentially arbitrary and one is free to select whatever may be convenient. However, a linearly filtered process with *exponential* response function enjoys a number of useful properties both theoretical and practical. It is defined by

$$f(x) = \rho^x \quad , \qquad x = 0,1,2,\ldots, \qquad 0 < \rho < 1 \ , \tag{3}$$

where the parameter ρ determines the rate of decay of the response. If ρ is close to zero then the decay is rapid whereas if it is close to one the decay is slow. This is illustrated in Figures 7.2.1 and 7.2.2.

The main advantage of the exponential filter is that many of the physical processes which we have mentioned happen to have response functions which are at least approximately of this form. In other words an exponentially filtered process can be used as an approximate model for processes such as streamflow, run-off, soil moisture, etc ... . It also conforms quite well to our generally accepted notions of how an index of wetness/dryness should behave.

Ordinary moving averages, i.e. rectangular filters, are quite often used as measures of wetness/dryness but they are not, in our opinion, suitable for the purpose. By using this index one is implicitly making the assumption that at on a given day, say t, x units of rainfall on day t are equivalent to having had x units of rainfall on day t-1, or on day t-2, and so on back to day t-L+1. In practice

20 mm of rain which fell say one week ago is not equivalent to 20 mm of rain today. Morevoer it is assumed that L days after the event the "effect" of the rainfall suddenly vanishes. Natural processes simply do not behave in this way; their response functions do not remain constant for several days, and they decay gradually over time, not abruptly.

A second useful property of the exponential filter is that it is determined (except for a scaling factor which may be necessary in some applications) by a single parameter, ρ. This not only keeps matters simple but as every parameter in a model has to be estimated or guessed it is obviously important to have as few of them as possible. This parameter has a simple interpretation, it describes the rate of decay of the response. In cases where no observations whatsoever are available on the process of interest one can guess the value of ρ by guessing another quantity which determines ρ, namely the *half-life*, h, of the filter. This is related to ρ by the following equations:

$$\rho = \exp(\ell n(\tfrac{1}{2})/h) \qquad (4a)$$

$$h = \ell n(\tfrac{1}{2})/\ell n(\rho). \qquad (4b)$$

The half-life is the time required for the response to a rainfall event to decay to exactly half of its original value, see Figure 7.3. Here response can be taken to stand for "benefit", "effect", "level" or whatever may be appropriate for the problem at hand. Experienced hydrologists usually have a fairly good idea of what the half-life of the response function of a process should be. This can then be used to determine ρ by means of (4a).

A further convenient property of the exponential filter leads to a wetness/dryness index which has a simple structure. Equation (1) reduces to

$$F(t) = \rho F(t-1) + R(t) \quad , \quad t = 2,3,... \qquad (5a)$$

where only the case t = 1 has to be computed separately:

$$F(1) = \sum_{i=0}^{L-1} \rho^i R(1-i) \qquad (5b)$$

and L is selected so that ρ^x is negligibly small for $x > L$.

This recurrence formula is convenient to compute F(t) which (except for F(1)) is only a function of R(t), ρ and F(t-1) and is consequently convenient to update as fresh rainfall data become available. As an index, F(t) is also easy to interpret: the state of wetness/dryness on a given day is a fraction ρ of the state on the preceding day plus the current rainfall depth.

Once the constant ρ (or equivalently the half-life, h) has been specified the distributional properties of F(t) are completely determined by those of the rainfall process. In other words once a model for the rainfall process is available one can answer any question relating to the behaviour of the filtered process. So for example one can define a drought as being a time when F(t) falls below some required level which need not be constant over the year. Suppose that at time t a level D(t) is required for the effective operation of some water related activity, t = 1,2,... . Then the process of negative deviations from D(t) provide an appropriate index of drought:

$$D\_(t) = \min(0, D(t)-F(t)) \quad , \quad t = 1,2,... .$$

The duration of a drought is then the time which D_(t) remains below zero. Alternatively it may be convenient to associate a cost function with deviations of F(t) from D(t) :

$$C(t) = f(D(t)-F(t)) \quad , \quad t = 1,2,...$$

and so on. The rainfall model described earlier in this chapter can be used to derive the quantities of interest relating to $D_{\_}(t)$, $C(t)$ or any other such process. For example one could compute the probability that a drought begins at a particular time, that it will be broken within a certain time, the distribution of the average drought severity per year, the average cost attributable to shortage of rainfall and so on.

The filtered process described above can be used as a general-purpose measure of wetness/dryness. It is flexible because the user is free to select the half-life to meet his specific needs. A short half-life would be suitable in situations where regularity of rainfall is important, for example in many agricultural applications; a long half-life where the amount rather than the regularity of rainfall is important, for example in applications relating to reservoir storage levels.

In the initial stages of this project it was our intention to find stochastic models to describe the exponential filtered process directly (rather than as by-products of the rainfall model). In particular we considered the following types of "standardised" drought indices:

$$I(t) = \{F(t)-EF(t)\}/S.D.(F(t)) \qquad (6a)$$

$$J(t) = F(t)/EF(t) \quad , \quad t = 1,2,... \qquad (6b)$$

where E denotes expectation and S.D. the standard deviation.

It was established that the severity of droughts as measured in terms of such indices was very strongly correlated to its duration. This confirmed the findings of Hulley (1980)

and it was therefore reasonable to describe droughts in terms of only one of these two variables. We then attempted to find probability distributions to describe the durations. The Sichel distribution (Sichel 1971) was found to be suitable for a number of stations and for half-lives in the range one to six weeks, approximately (Zucchini 1974). For many records, particularly those associated with highly seasonal rainfall no suitable distribution could be found. The problem is that in some areas there is an appreciable probability that it will not rain at all during the dry season and in the event of drought one has to wait until the following rainy season before the drought can be broken. Consequently the probability distribution of drought duration is quasi-periodic in spite of the standardisations (6a) or (6b). It would perhaps have been possible to postulate new probability distributions but these would be inevitably quite complex and furthermore their parameters would have to vary seasonally.

A second and equally troublesome drawback of standardised indices in regions with a marked dry season is that rainfall events in the dry season are disproportionately inflated by such indices. So, for example, a relatively insignificant amount of rainfall in the dry season can "break" a drought according to the index. One could of course restrict attention to the rainy season but this would involve one in arbitrary definitions of what constitutes this season. Moreover storms in the dry season can significantly contribute to reservoir storage levels, i.e. break certain types of drought, and so it would be unsatisfactory to ignore dry seasons altogether.

The whole question of what might be the best way of constructing a general drought index becomes rather unimportant once a model for the rainfall process is available, because the properties of any rainfall-based index can be

derived with relative ease. The family of indices proposed in this chapter is one of many possibilities. As was pointed out, it has a number of desirable features and if a "general purpose" family of indices is required we would recommend that it be used, but in its untransformed form, i.e. as given in (5a) and (5b). Applications of this index are discussed in the next chapter.

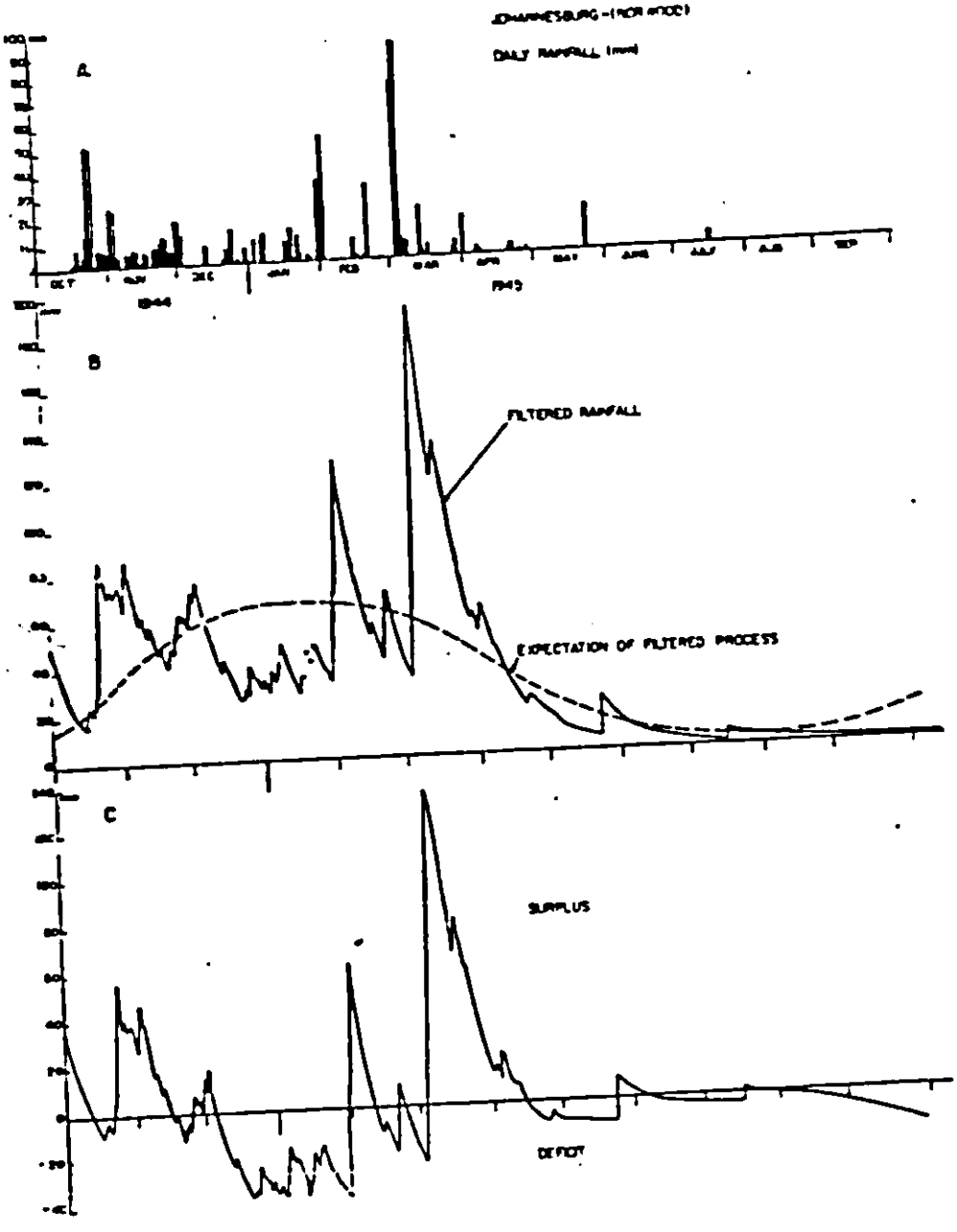
8. APPLICATIONS OF THE EXPONENTIAL FAMILY OF DROUGHT INDICES

For the illustration of and in order to assess the utility of the proposed drought model a considerable number of applications are given. The primary intention is to show that such a modelling scheme provides us with a very wide spectrum of potential applications from the retrospective assessment of historical droughts to the ability to forecast the likelihood of recovery over a particular horizon from any given or current state of rainfall deficiency. We can view the history or future of a drought either as a process of shortfalls from day to day or as a sum of shortfalls over some discrete time interval, for example a month or a year. Except where otherwise stated the half-life considered is ten days, which is largely arbitrary although having some physical justification in that the average life of a hydrograph in the Vaal River at Standerton is ten days whilst at intermediate depths many of the more common soil types found in South Africa "dry out" at a rate of this order of magnitude. (see e.g. Beukes and Weber 1981).

We are obviously in a position to change the half-life and therefore the rate of decay of the filtered process at will, given some knowledge of the particular physical system that we are attempting to imitate. Figure 8.1 portrays the operation of the model given a one-year sequence of daily rainfalls at a station in Johannesburg. Even at this primary level of application the treatment of rainfall in such a way is of considerable value.

A visual inspection of the daily rainfalls as measured reveals the obvious features of surpluses in October, late January/early February and early March. But to what degree is the three-month run from November to January deficient,

FIGURE 8.1 An illustration of the level of the filtered process and its expectation over a single year at Johannesburg

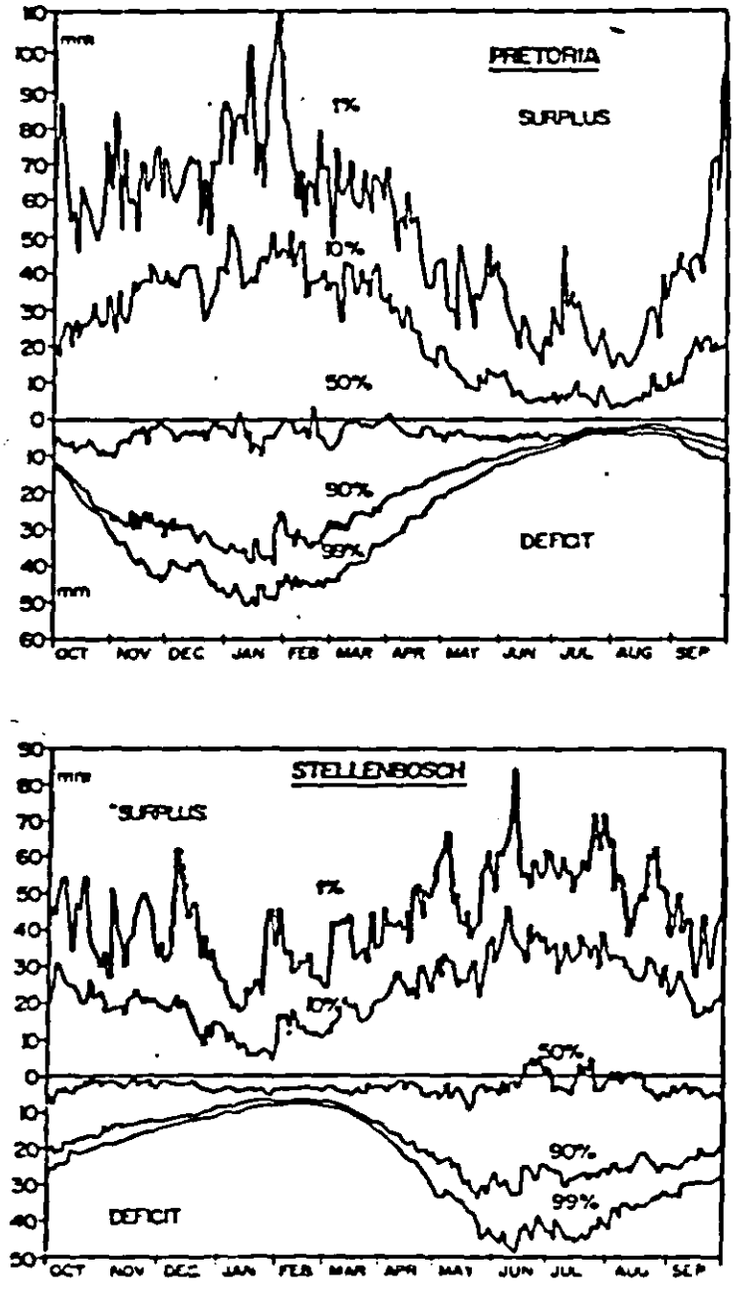


since no daily rainfall during this period exceeded 20 mm? That it is a deficit run is clear from B and C with the difference between expectation and historical rainfall reaching a shortfall of 20 to 40 mm. We would accumulate the sum of deficits over the period and thus accord the event a run sum, thereby having information on the duration in days, the total deficit and maximum deficit of this drought.

The obvious next step would be to associate a probability of occurrence with the event either as an entity in itself or by way of evaluating the risk of being x mm in deficit on each of these particular days of the year. Figure 8.2 shows the daily percentiles of surplus and deficit for Pretoria and Stellenbosch estimated from 1000 years of simulated data. For Pretoria on 1 February there is an equal chance of being more than 48 mm in deficit or 80 mm in surplus with the median position of this day being a slight deficit. Even though computed from a relatively large number of simulations the percentile estimates are far from smooth and illustrate a wide sampling variation from day to day, particularly at the extremes, as expected. However, if one is satisfied that the estimates are reasonable it is straightforward to fit a smooth periodic function through each percentile from day 1 to 365. Other points that are noteworthy are the skewed nature of the distribution of surplus/deficit on any particular day and the seasonality of the percentiles over the year. During the wet season our expectations of rainfall are higher with the consequence that any unseasonal run of dry days would be responsible for larger deficits whilst the possibility of larger surpluses is naturally greater than in the dry season.

Given these percentiles of surplus/deficit it is of interest to examine historical years during which drought was ex-

FIGURE 8.2 Percentiles of surplus and deficit computed on a daily basis from 1000 years of simulated data



perienced and establish the level of the filtered rainfall from day to day. Figure 8.3 shows just such an analysis, with the percentiles now smooth after fitting a harmonic function through those shown in Figure 8.2. In both examples, the status of the index of surplus/deficit is plotted for the last day of each month and shows for Pretoria that by the end of January 1933 it had almost reached the 1% level whilst on the same day in 1966 some considerable recovery was evident due to good rains over the previous 60 days. Most analyses so far (cf. Figure 6.14.4) have shown 1927/8 to be the driest year on record at Stellenbosch and the severity of the situation can clearly be seen. During 1972/3, the second driest year, some unseasonal rainfalls gave a slight surplus but the earlier winter rains are seen to be particularly deficient. On both occasions maximum deficit was reached at the end of June.

The model can be further employed to forecast the surplus/deficit situation over a horizon of any given length from days to years and from any given initial condition. Consider a hypothetical surplus of 10 mm at Stellenbosch on 1 January. We see from Figure 8.2 that the probability of being in such a state on this day is 10% or less. From Figure 8.4, where the state is forecast over the next 365 days and given at the end of each week, we see that the probability of improving on this surplus is very small for the next six weeks and that there is a 90% chance of being in deficit within two weeks. By mid-year (week 26) there is an equal chance of being more than 45 mm in deficit and 70 mm in surplus. By the end of the year the chance of still being in surplus is slightly less than 50% and that of ending the year as it began, with a 10 mm surplus, is about 10%.

Having explored some initial applications of the drought

FIGURE 8.3 Historical levels of surplus/deficit (plotted for the last day of each month) with their associated daily percentiles

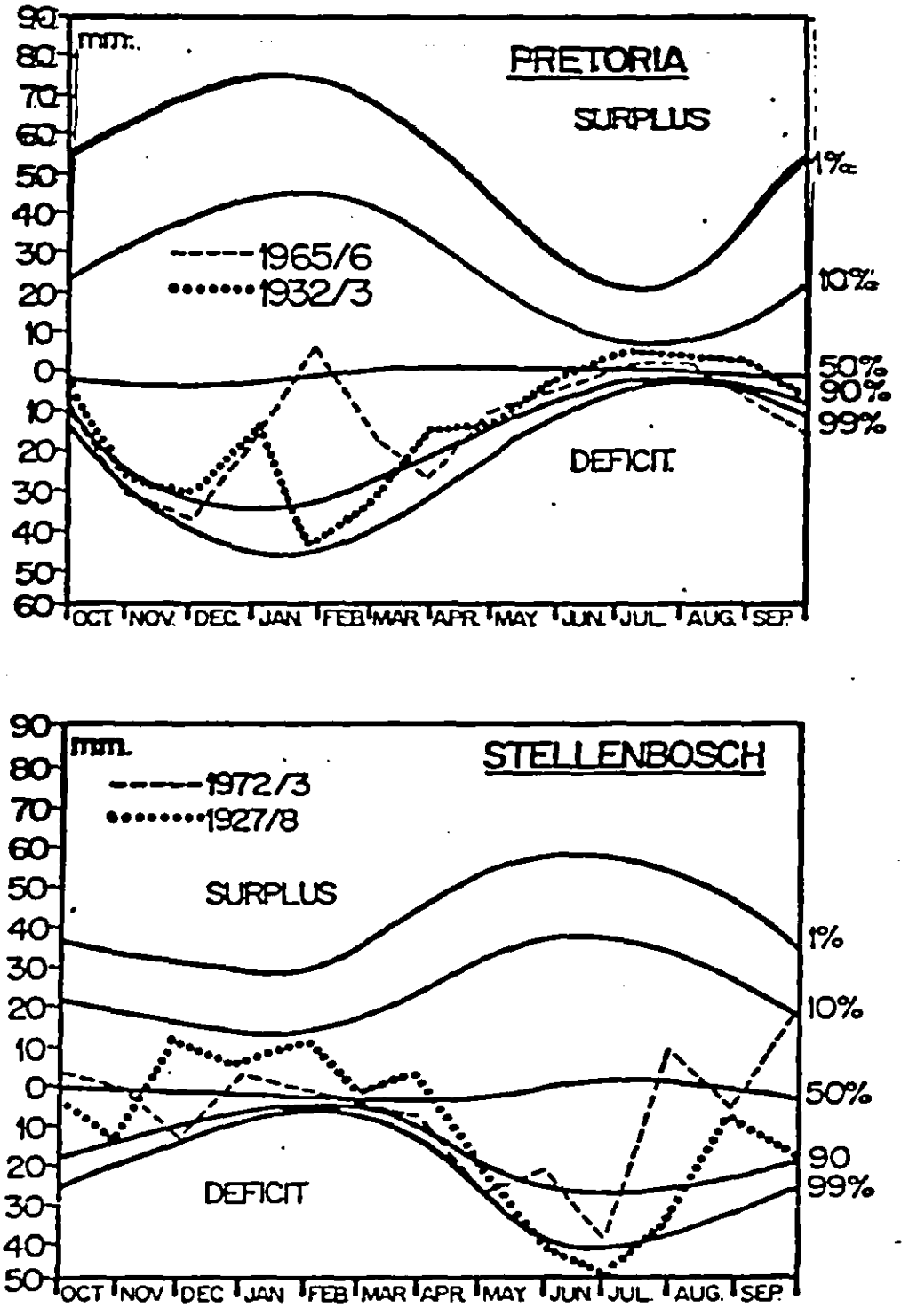
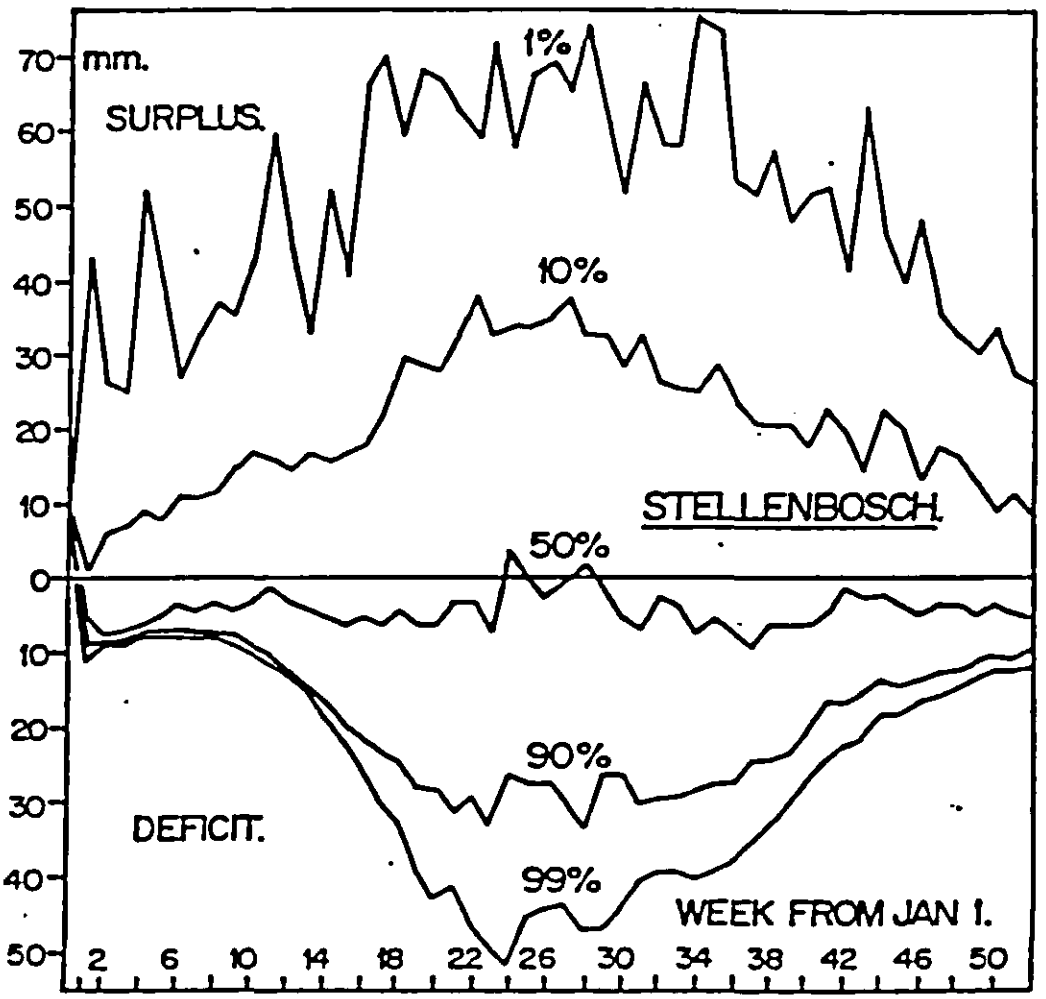


FIGURE 8.4 Forecast of surplus/deficit over a 365-day horizon given a surplus of 10 mm on 1 January at Stellenbosch. Selected percentiles are shown



model by example, it is pertinent at this point to examine how the various aspects of deficit (run length, run sum and maximum run deficit) are affected by the selection of a half-life for the filtered rainfall. It is conceivable that we may have some idea of these properties of the physical process that we are attempting to imitate. For example, the data from lysimeter studies may reveal some information on the distribution of drought runs or run sums for a particular plant/soil system under conditions of natural rainfall. It would be useful to be able to choose a half-life in advance in the knowledge that the distribution of these aspects of the lysimeter study are preserved. Choosing Pretoria as an example Figure 8.5 shows that an increase in half-life will increase the run length at a particular probability level and similarly affect the run sum and maximum deficit within a run. A longer half-life will, in consequence, produce more severe droughts of longer duration with the implication that only considerable rainfall over a period will lead to recovery. A physical analogy would be the performance of a water table over an extended (perhaps n-year) period of rainfall deficit. Ground water levels, having fallen to some uncommonly low level may require an extended period of surplus prior to any recovery to levels that would be considered normal.

So far we have considered the level of the process of surplus/deficit which would be pertinent to the imitation of a soil moisture régime, for example. In considering river flow and, in particular, reservoir storages it is perhaps more relevant to emphasise run sums. We could accumulate the daily surplus/deficit over each year of a historical record and so identify periods of drought. Two such examples are given in Figures 8.6 and 8.7 and an estimate of the percentiles of surplus/deficit obtained by simulation (1000 replicates) adds further to the utility of the

FIGURE 8.5 Effect of choice of half-life on the distribution of various aspects of drought runs

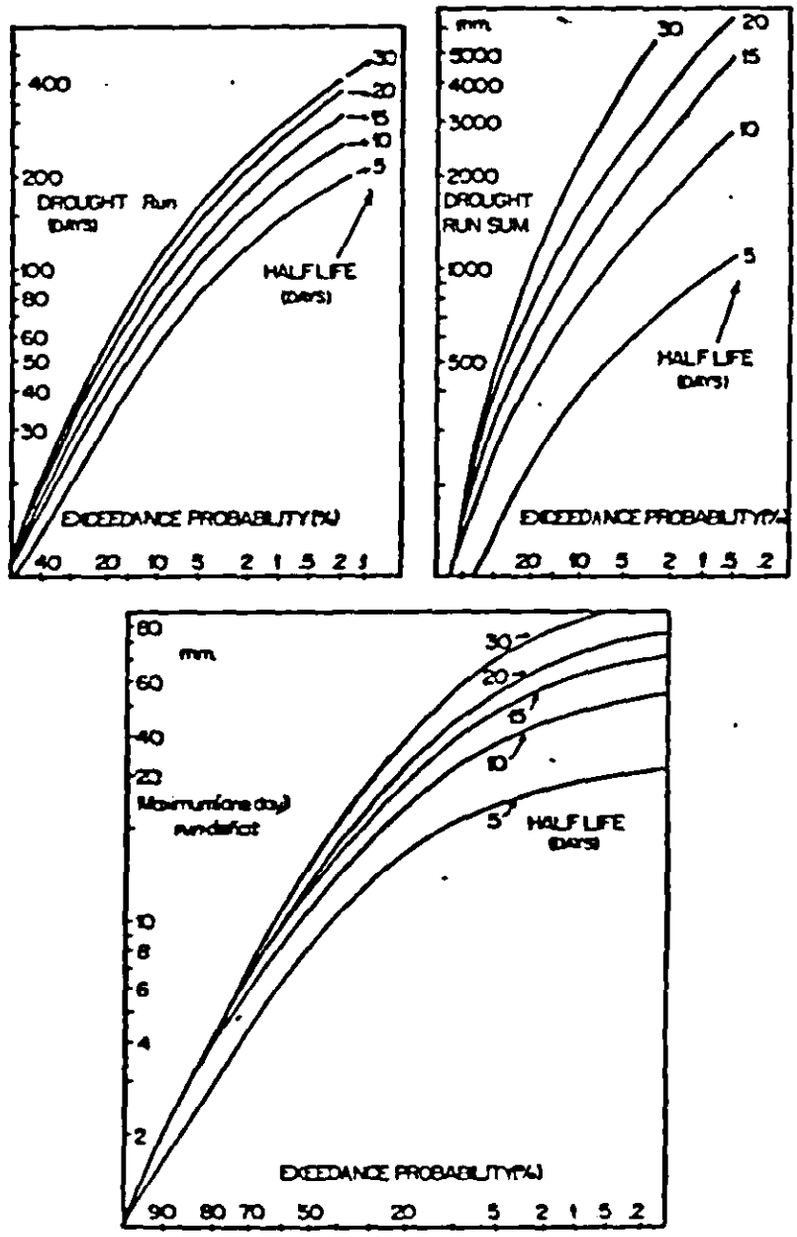


FIGURE 8.6 Annual sums of surplus/deficit accumulated on a daily basis with associated percentiles.
A: ½ life = 10 days; B: ½ life = 30 days

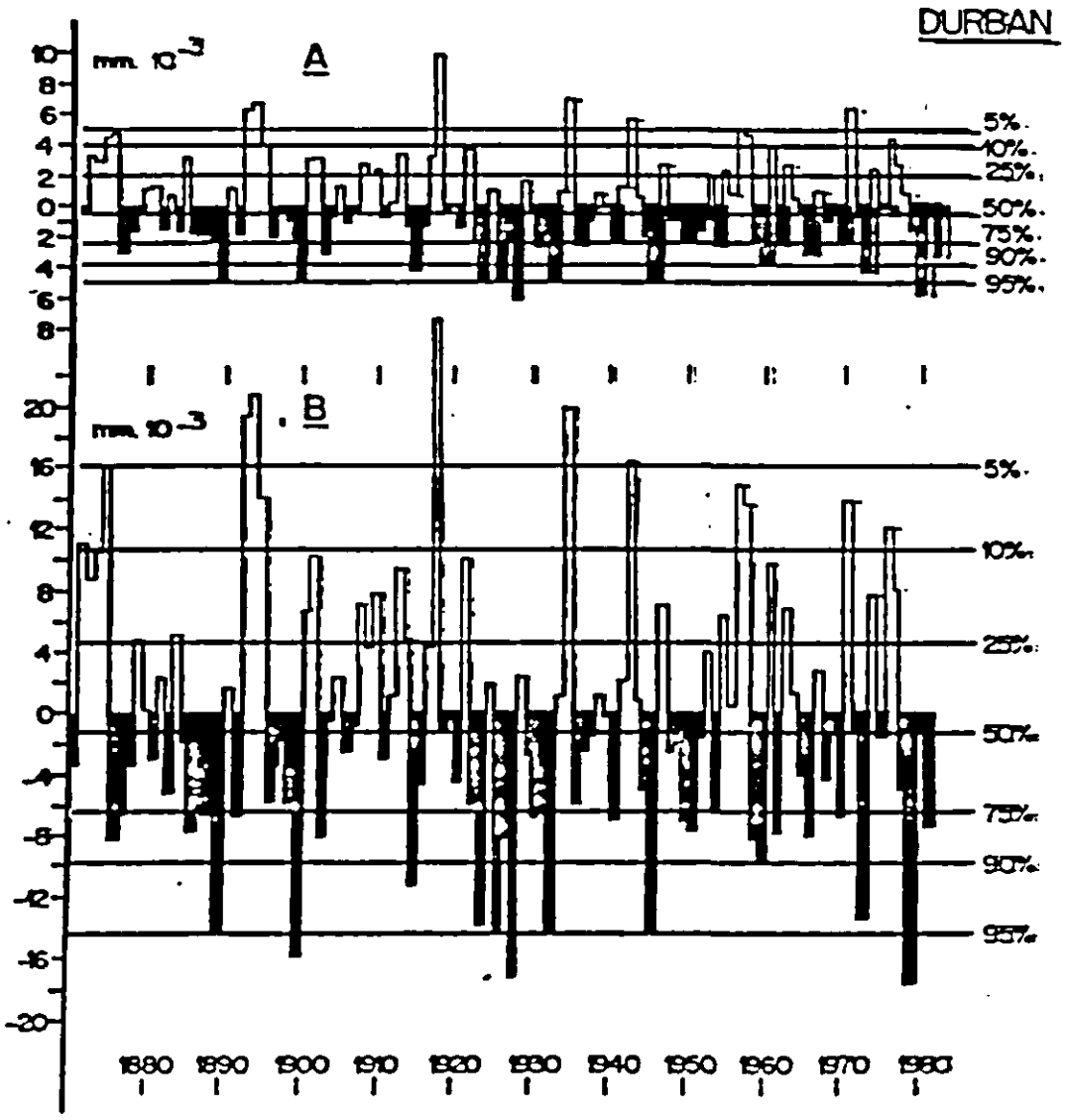
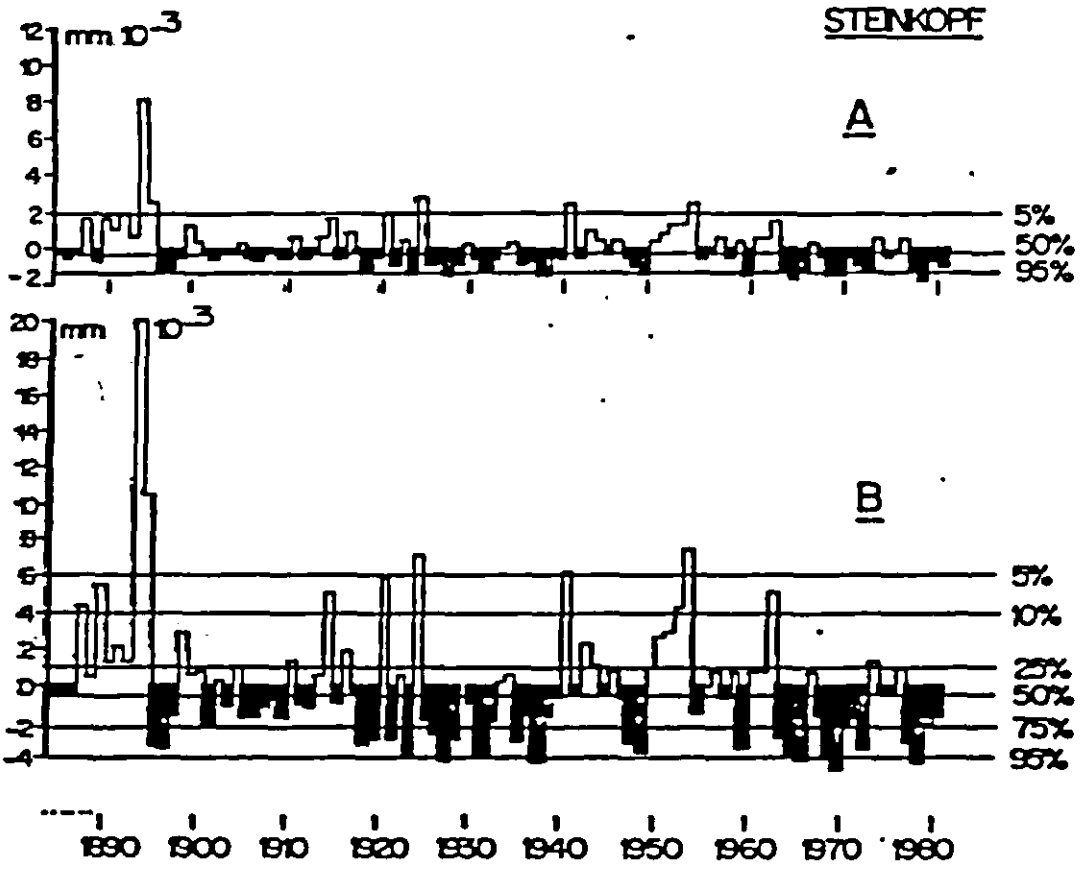


FIGURE 8.7 Annual sums of surplus/deficit accumulated on a daily basis with associated percentiles.
A: ½ life = 10 days; B: ½ life = 30 days



presentation. Durban (mean annual precipitation : 1000 mm) and Steinkopf (mean annual precipitation : 140 mm) portray extremes of rainfall régimes within South Africa but both illustrate drought as an integral part of the precipitation climate. The debit or credit is accumulated on a daily basis and the total annual position computed in this way would not necessarily correspond with some classification of years based simply on total annual rainfall. If there was only one wet day in a particular year on which 1½ times the annual mean rainfall occurred then the year would normally be considered wet. Such an occurrence could happen in an extremely arid environment. However, within the scheme presented here such a year would show a deficit since with a short half-life, say ten days, the surplus gained from such an event would soon decay. Similarly, unseasonal rainfall, given the failure of the seasonal rains, may amount to the annual mean and thus the classification of the year as normal. However, the accumulated deficit due to the failure of the seasonal rains would, in terms of our model, be so large as to negate the surplus gained unseasonally and thus the year would show a deficit. In other words a deficit occurs not only as a consequence of an overall lack of rainfall but additionally because the rainfall occurs at an unexpected time or is attributed to unusually few rain days. Such rainfall is generally less effective for the generation of streamflow and therefore of reservoir storage. A considerable volume of such events would be taken up in filling depleted soil moisture storage prior to the generation of surface runoff.

It is true to say that these aspects of drought have largely been ignored in favour of a relatively simplistic view of credit and debit assessed relative to a mean or some familiar view of the total expected rainfall account over the year. Paradoxically, we subjectively view a year differently in expecting certain times to be wet and any

deviations or unseasonality in daily rainfalls is seen as a shortfall whatever the end of the year account may reveal. The model behaves in precisely the same way.

That the level of the annual sum of surplus/deficit is not simply related to the corresponding number of wet days that occurred is shown in Figures 8.8 and 8.9. The correlation improves significantly as the definition of a wet day is changed from one upon which more than 0 mm fell to one which might be considered to be a storm day (more than 20 or 40 mm at Durban; more than 10 mm at Steinkopf). This implies that deficits are more significantly related to a lack of storms rather than to a lack of wet days *per se*.

For periods shorter than one year we can either consider the process or the sum of the process, depending upon the field of interest. Figure 8.10 shows the level of the index at Durban on the last day of each month from October 1871 to September 1927. Although periods of surplus are easily seen, periods of deficit are less apparent and if it is our intention to view the monthly history of surplus/deficit then its monthly sum would be more revealing. The driest five year period at Durban prior to 1980 was that from October 1885 to September 1890 (cf. Figure 6.14.1) and the sum of the daily index for each of these 60 months is shown in Figure 8.11. The period involved the overall failure of the summer rainfall in all years with the possible exception of 1888/9. A characteristic of the drought was that the last year in the run (1889/90) was by far the worst which, following upon four unusually dry years, would make the event unusually severe. A second example, the nine-year sequence from October 1925 to September 1934, is characterised by a lack of surplus in any month from June 1927 to March 1929.

The monthly sums of the index are now used to portray the

FIGURE 8.8 Relationship between annual frequency of various types of wet day and total annual deficit/surplus

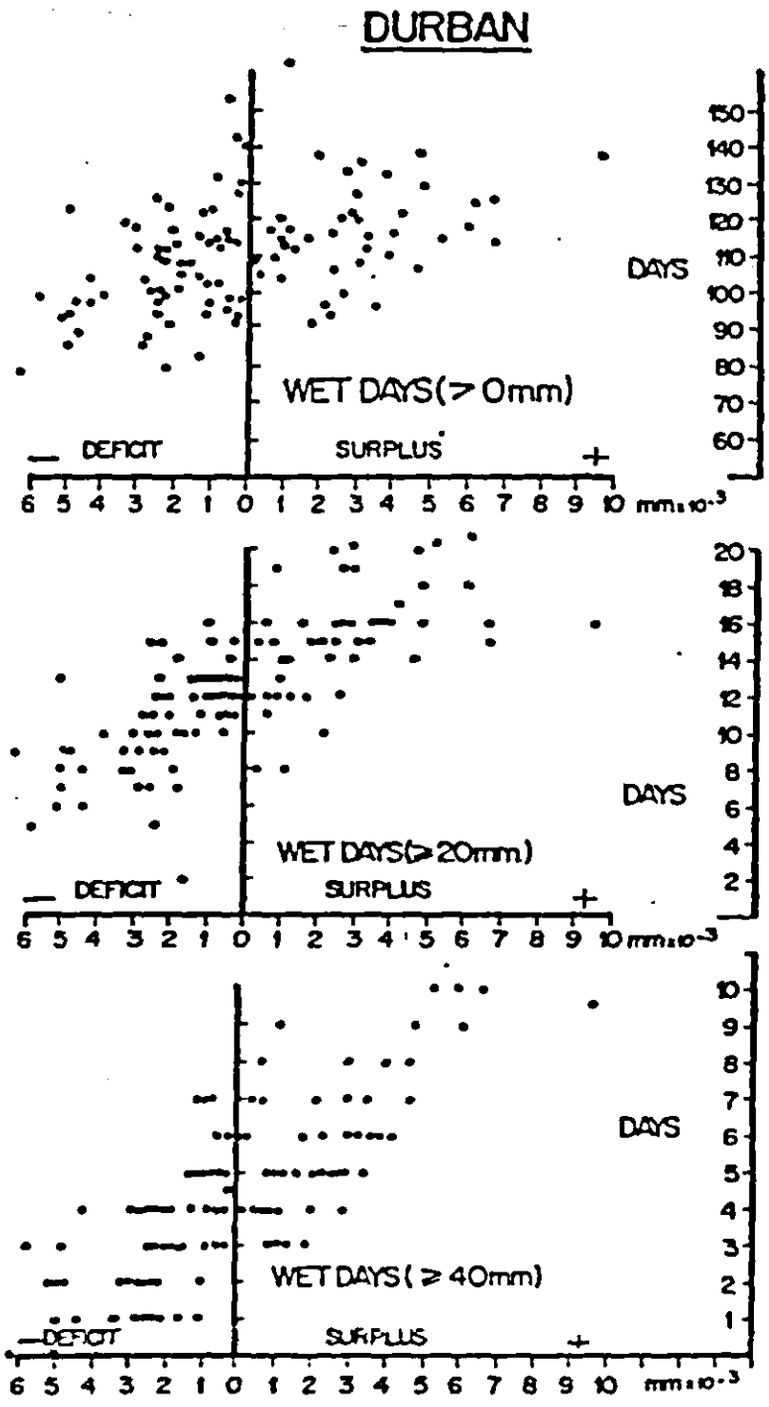


FIGURE 8.9 Relationship between annual frequency of various types of wet day and total annual deficit/surplus

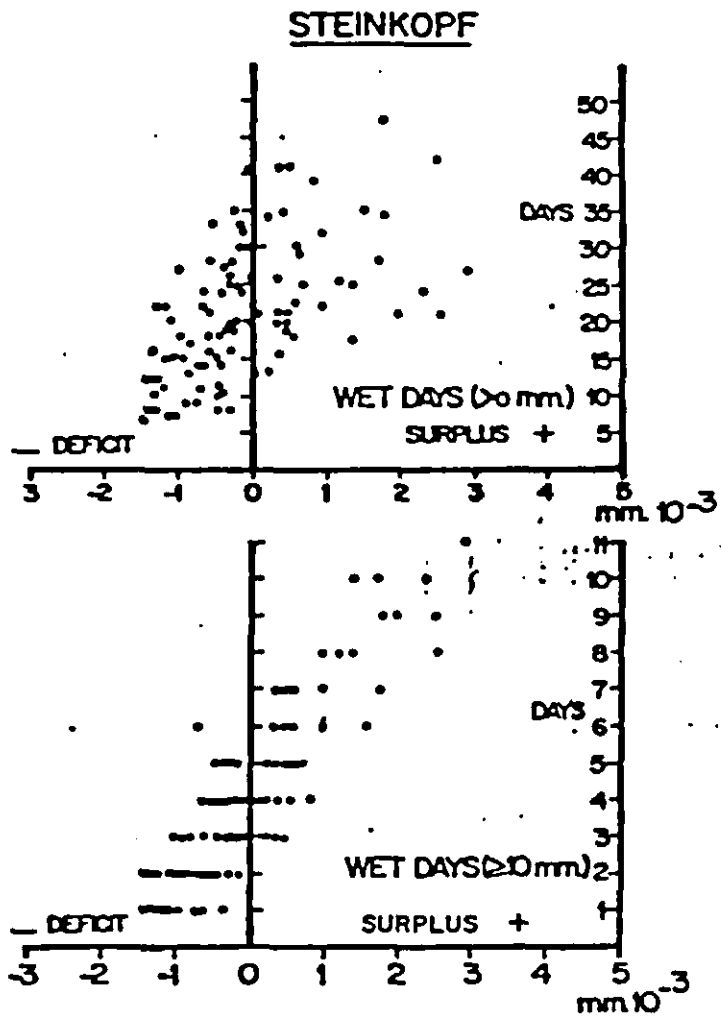


FIGURE 8.10 Index of surplus/deficit on the last day of each month at Durban from October 1871 to September 1927, with associated 5% and 95% percentiles

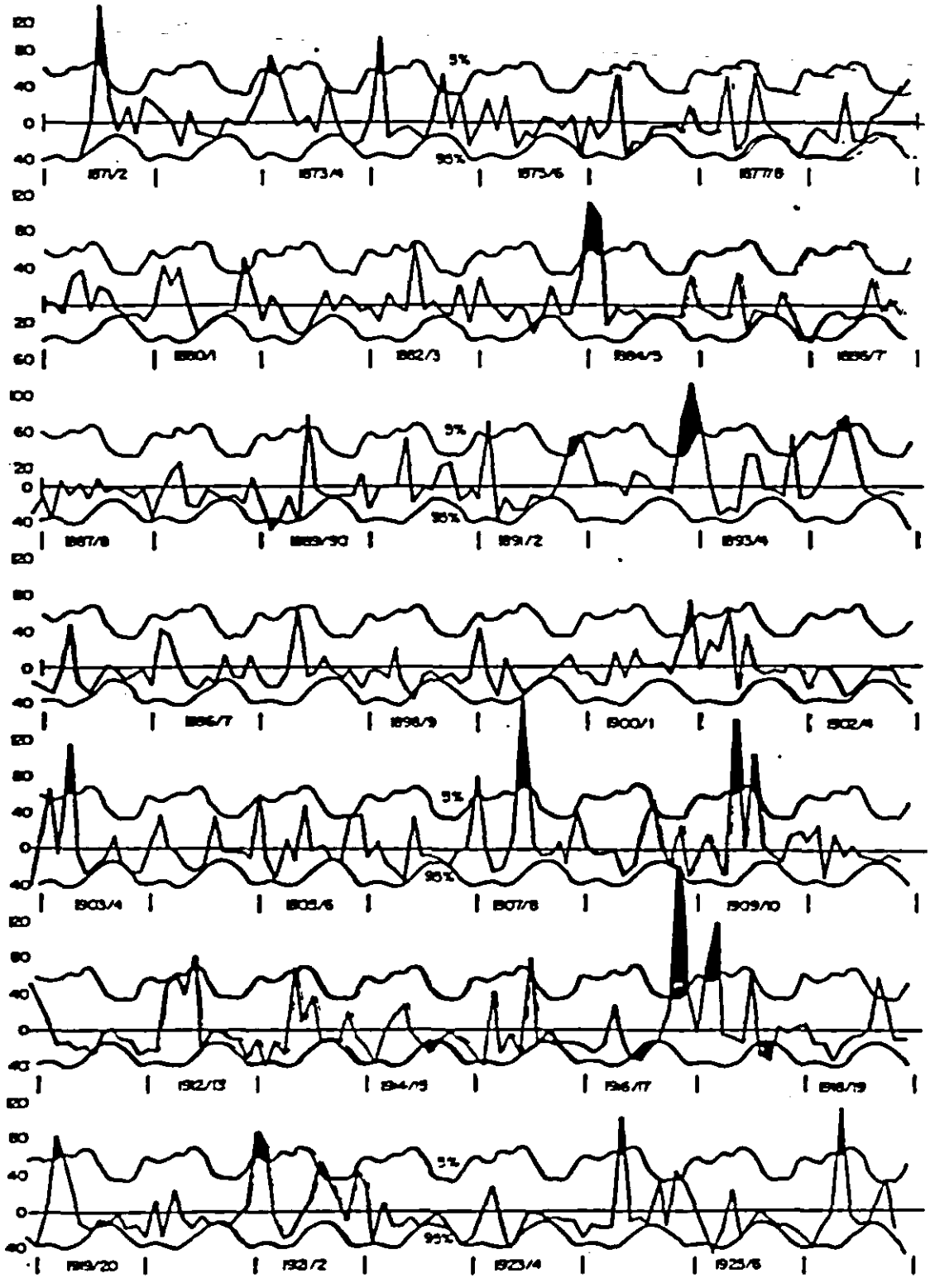
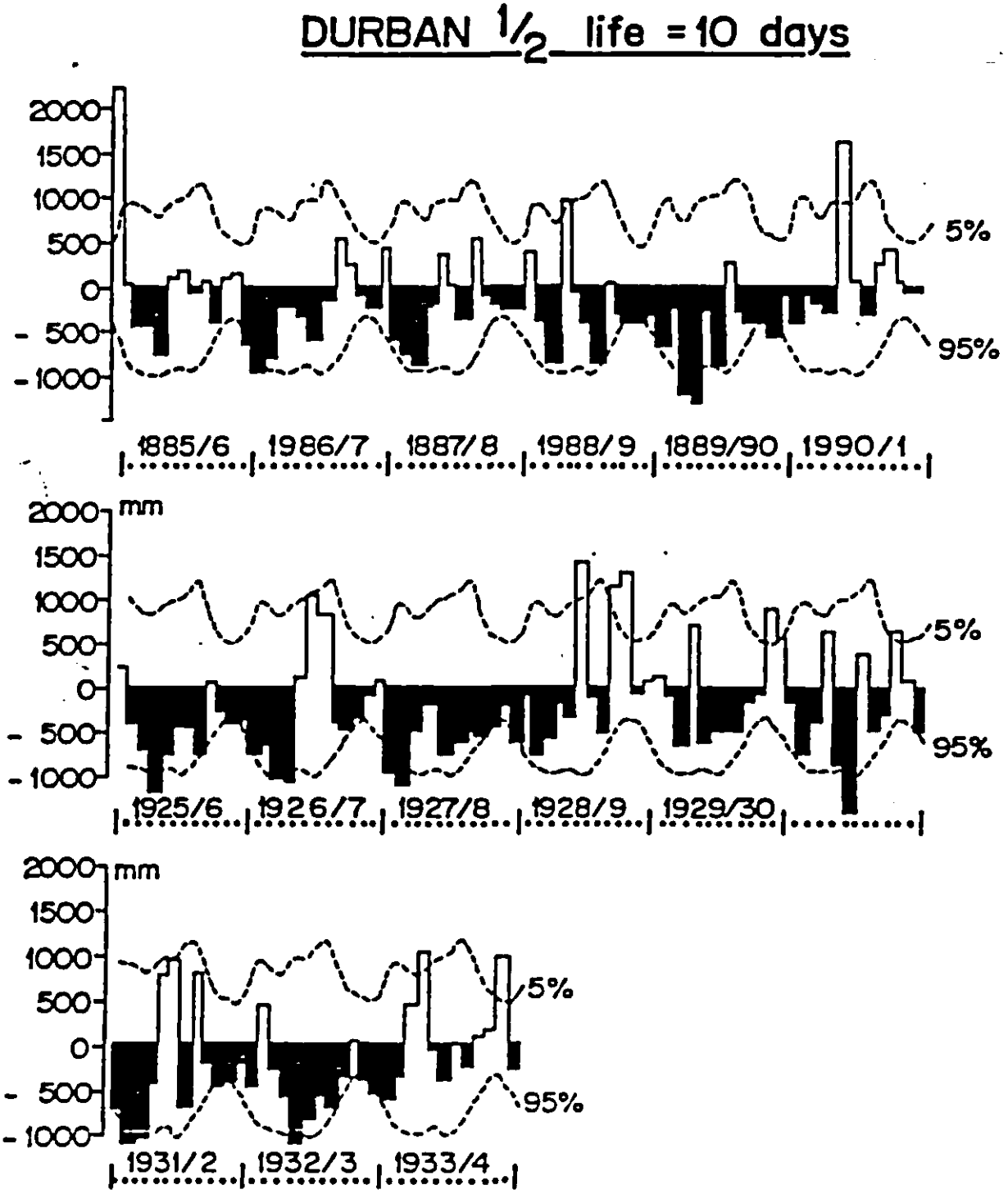


FIGURE 8.11 Monthly sum of surplus/deficit at Durban for two drought periods with associated 5% and 95% percentiles



history of what is generally seen as the most severe drought over the summer rainfall region to have occurred prior to 1978. The impact of the event of the early nineteen thirties was made all the more memorable combined as it was with severe economic depression. It is particularly significant in that until recently (1980) it provided the critical inflow period for reservoir design and operation upon which much official planning was based. The "assured yield" of water schemes was almost totally founded on their ability to survive a drought of the same magnitude. Figures 8.12, 8.13 and 8.14 illustrate the monthly history of the drought from October 1929 to September 1935 at eight stations selected from various parts of the summer rainfall region. A number of points emerge. The drought generally began in early 1930 with a serious deficiency in the seasonal rainfall, that is with the exception of Logaging, near Mafeking in the North Western Transvaal, where no surplus at all had been recorded from September 1929. The worst period of deficit was from October 1931 to September 1933 when during the two seasons the monthly deficiencies reached extreme levels. With the exception of the North Western Transvaal, the drought broke simultaneously over the region with exceptional surpluses during the summer of 1933/4.

In the South Western Cape (winter rainfall region) the worst drought, at least from a water resources point of view, occurred during the 1972/3 season. Figures 8.15 and 8.16 show the monthly history of the event at four stations and illustrate that the deficiency effectively lasted for three years starting in the winter of 1970. The drought broke with considerable late winter rainfalls providing exceptional surpluses during 1974.

We can look at such historical droughts in a way that emphasises their risk of occurrence rather than their

FIGURE 8.12 Monthly sum of surplus/deficit for the period October 1929 to September 1935 for selected stations and with associated 5% and 95% percentiles.

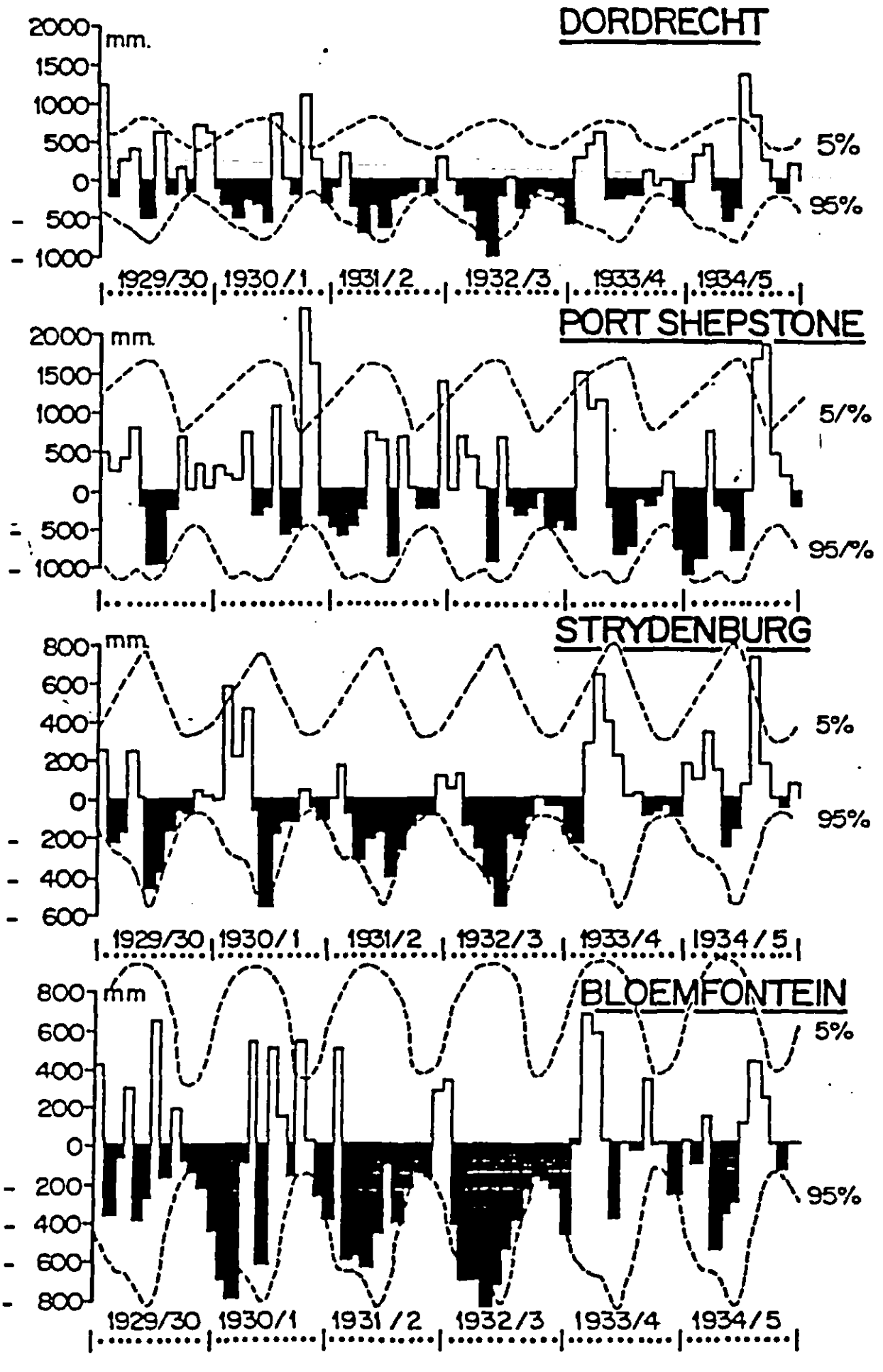


FIGURE 8.13 Monthly sum of surplus/deficit for the period October 1929 to September 1935 for selected stations and with associated 5% and 95% percentiles

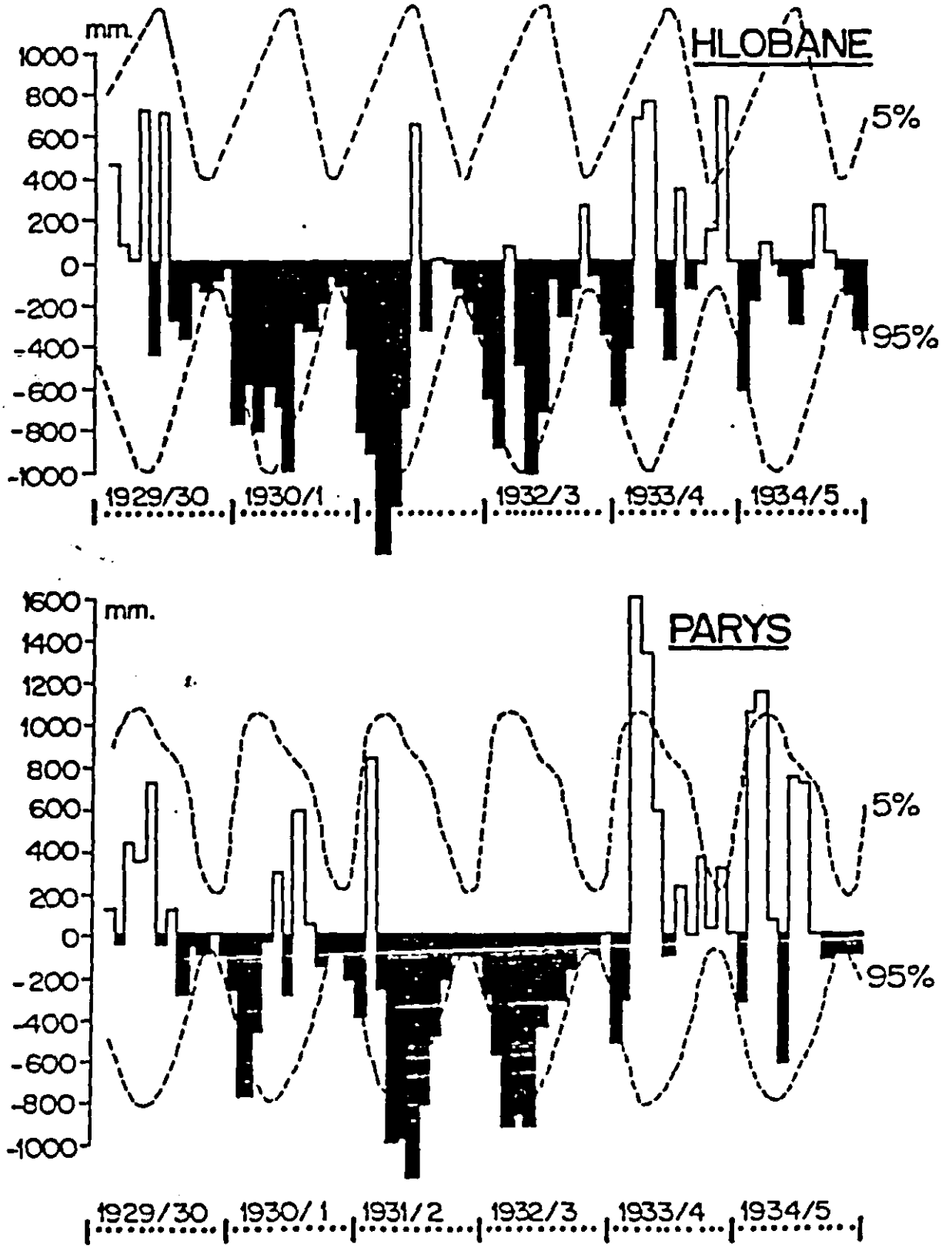


FIGURE 8.14 Monthly sum of surplus/deficit for the period October 1929 to September 1935 for selected stations and with associated 5% and 95% percentiles

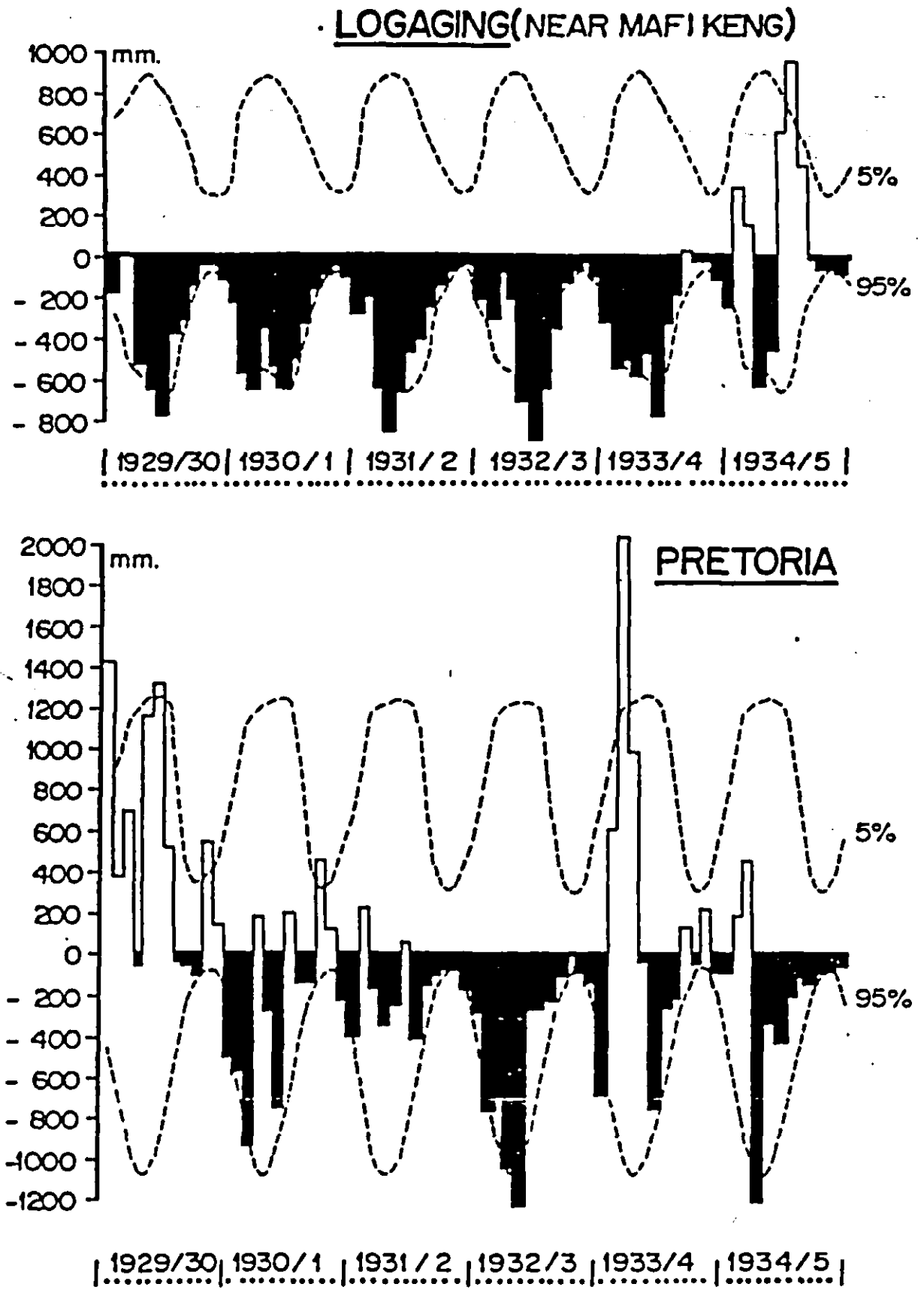


FIGURE 8.15 Monthly sum of surplus/deficit for the period October 1970 to September 1974 at Cape Town and Stellenbosch with associated 5%and 95% percentiles

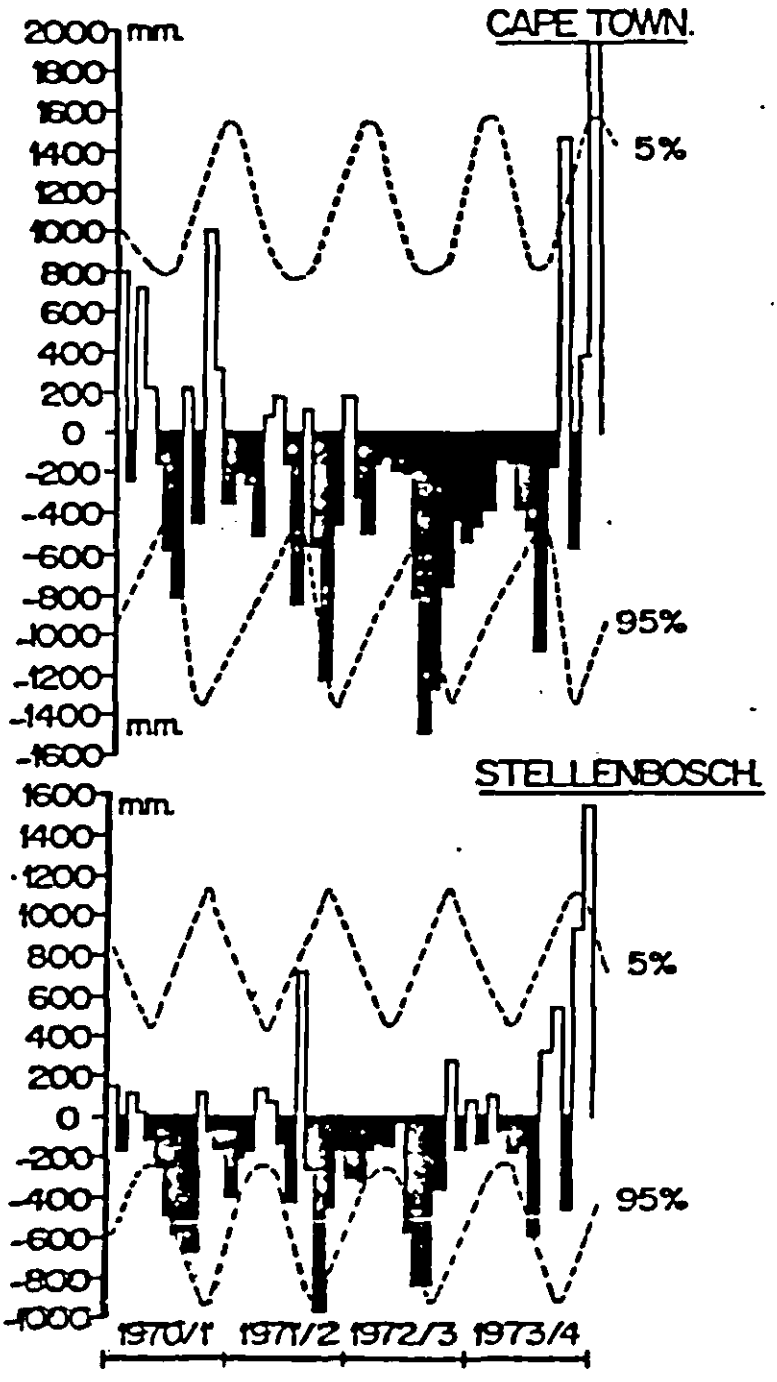
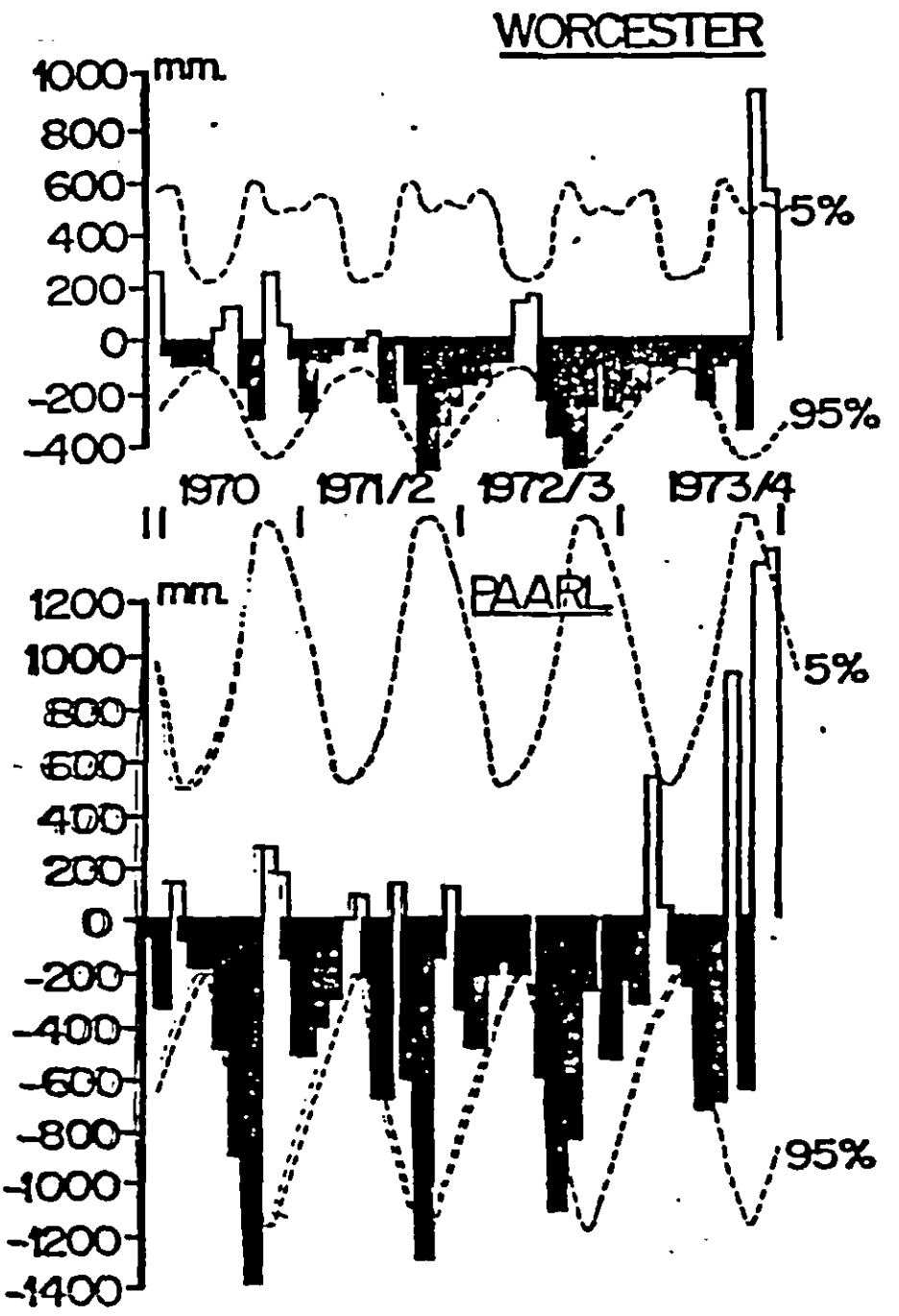


FIGURE 8.16 Monthly sum of surplus/deficit for the period October 1970 to September 1974 at Worcester and Paarl with associated 5%and 95% percentiles



chronology. By proposing a starting and ending date for the 1930/1 to 1933/4 event at Pretoria according to some criterion, we can simulate the distribution of surplus/deficit over the particular period of months. For example, we might choose 1 October 1930 as the starting date when serious deficits began to develop and 31 January 1934 as the ending date when the sum of daily surpluses over the previous 31 days exceeded the 5% level. We then simulated this period (1000 times) starting at the historical position of the process on 30 September 1930 and computed the distribution of the sum of surplus/deficit over this 40-month period. Such a result is shown in Figure 8.17 which reveals that the historical deficit was indeed an extreme event. A similar result for Worcester for the event from 1/4/72 to 31/7/74 is shown in Figure 8.18.

Yet another way of viewing historical droughts is to consider the month-by-month development of the accumulated deficit over the critical period. Given the position at the end of September 1930, we simulate the distribution of the accumulated process of surplus/deficit at the end of each month over a 36 month period and then plot the history of cumulative deficit. The results of the exercise for the period October 1930 to September 1933 are shown in Figures 8.19 and 8.20 respectively.

As a final application of the drought model we may wish to forecast the development of cumulative surplus/deficit on a weekly basis over some time horizon of interest. Figures 8.21 and 8.22 show two such results for Pretoria starting in surplus and deficit. The starting points relate the sum of the index for September and would indicate the occurrence of early spring rains (surplus) or their non-occurrence (deficit).

FIGURE 8.17 Simulated distribution of surplus/deficit over a 40-month period at Pretoria starting on 1 October. The exceedance probability of the deficit that occured from 1/10/30 to to 31/1/34 is shown.

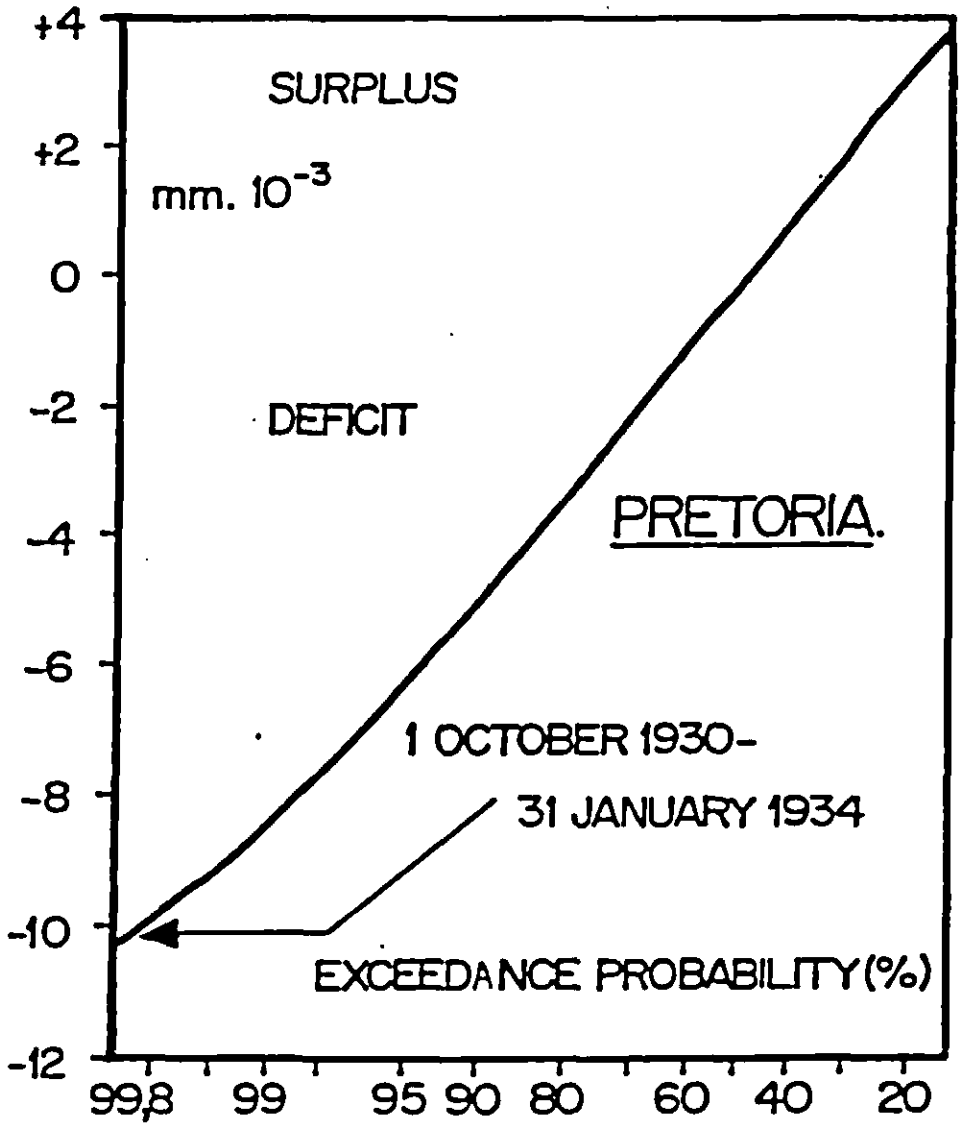


FIGURE 8.18 Simulated distribution of surplus/deficit over a 28-month period at Worcester starting on 1 April. The exceedance probability of the deficit from 1/4/72 to 31/7/74 is shown.

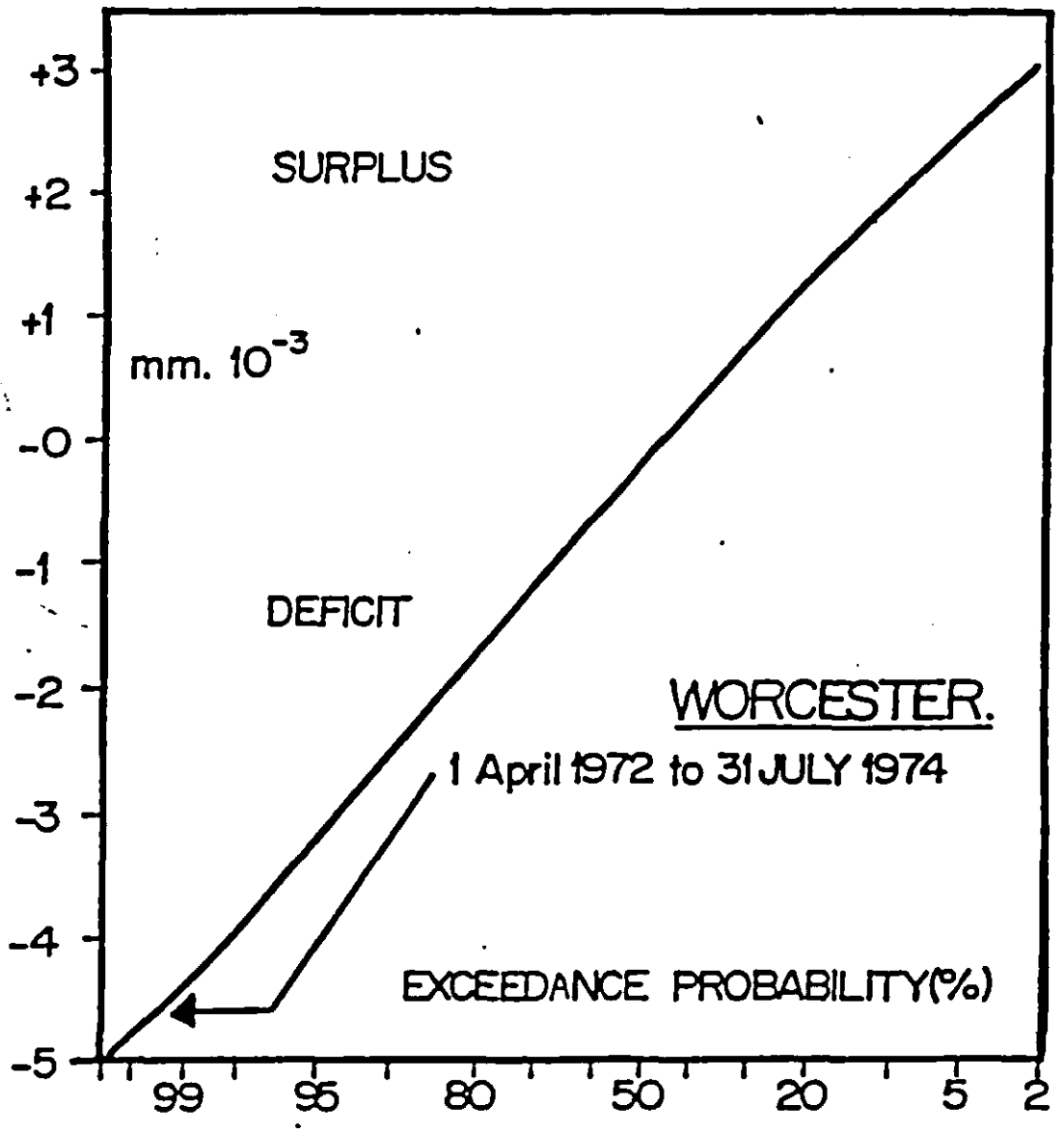


FIGURE 8.19 Simulated percentiles of the distribution of cumulative surplus/deficit over a 36-month period at Pretoria with the event of October 1930 to September 1934 (----)

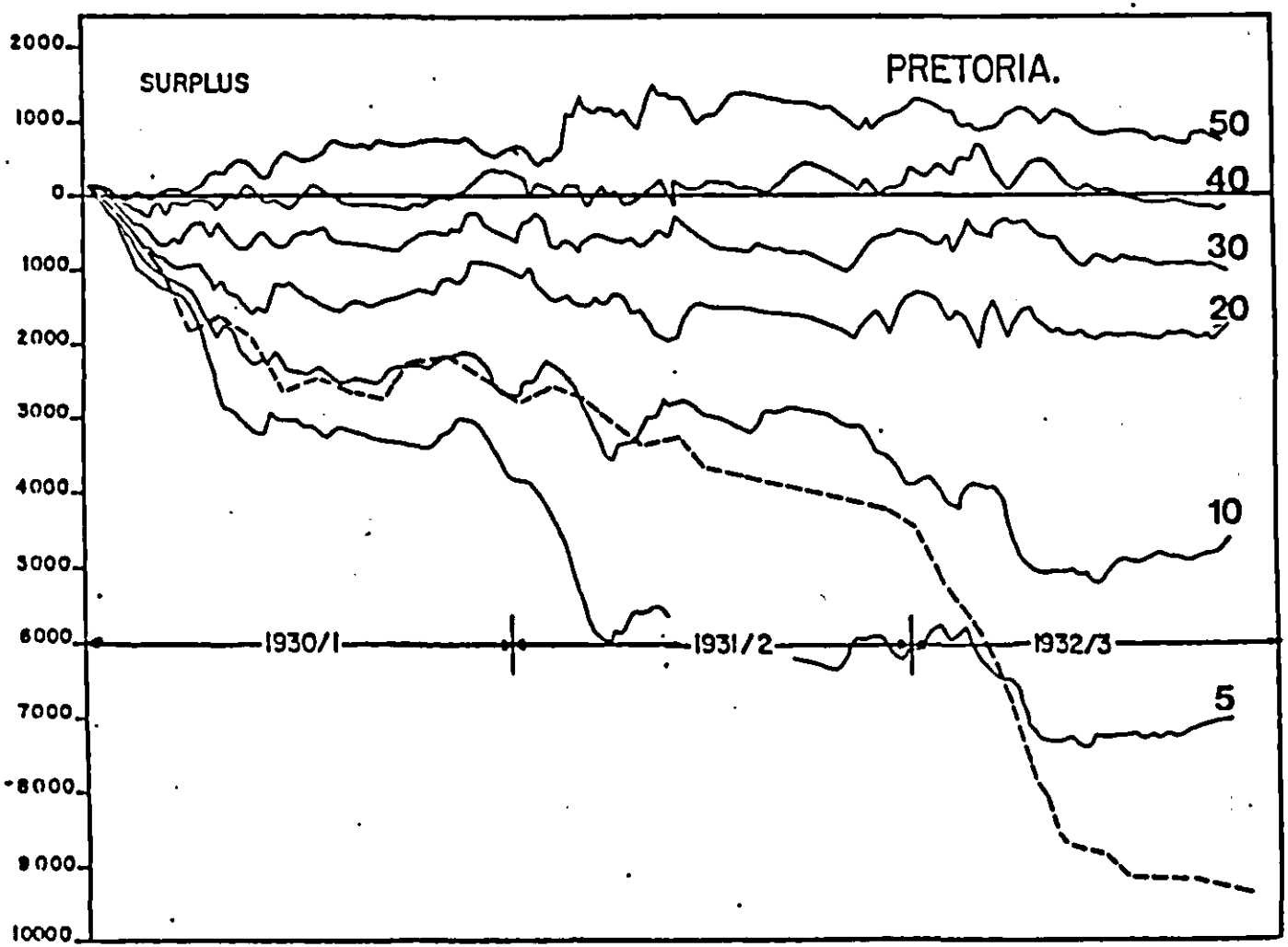


FIGURE 8.20 Simulated percentiles of the distribution of cumulative surplus/deficit over a 36-month period at Parys with the event of October 1930 to September 1934 (----)

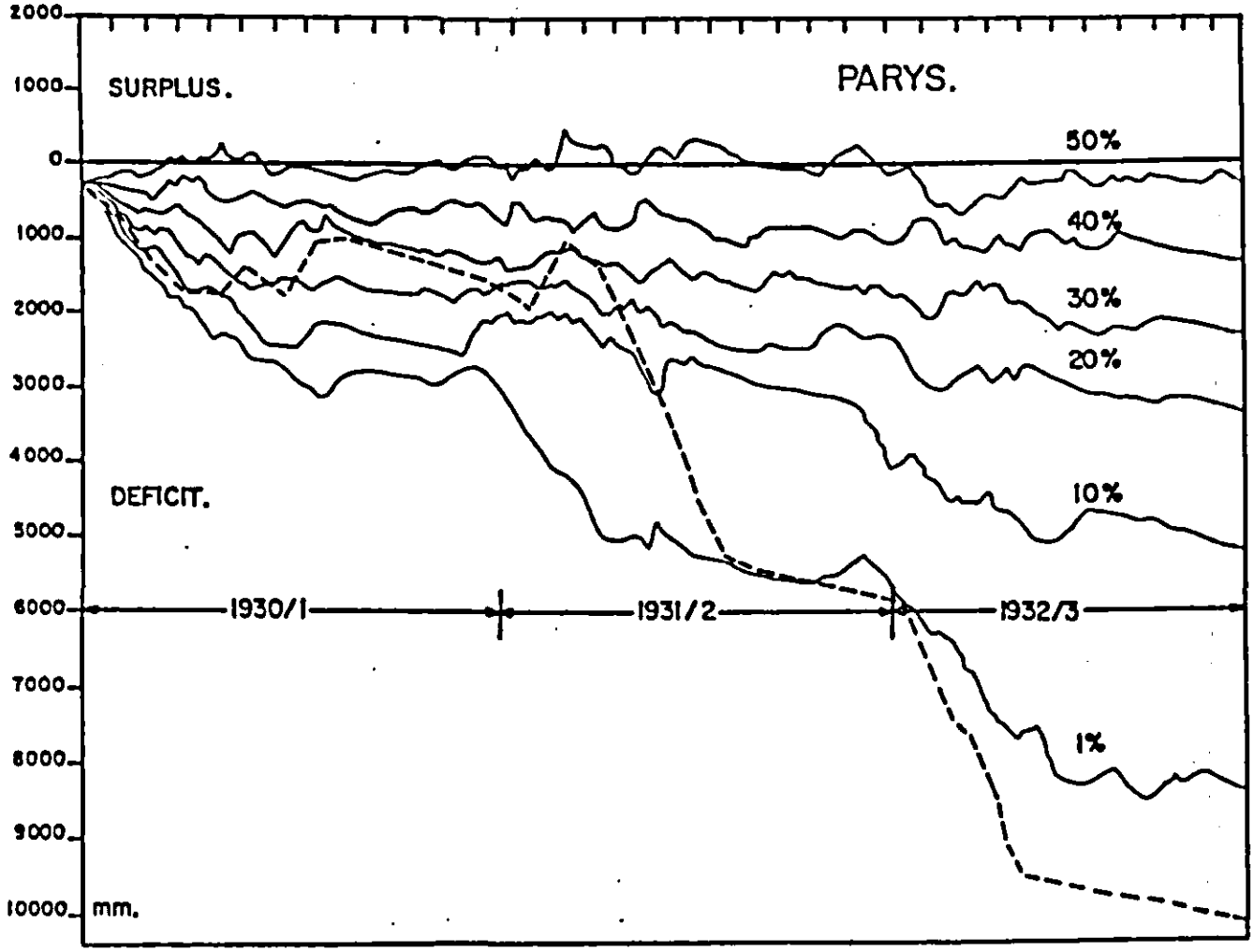


FIGURE 8.21 Estimates of cumulative weekly surplus/deficit over 52 weeks starting 1 October with a deficit of 200 mm at Pretoria.

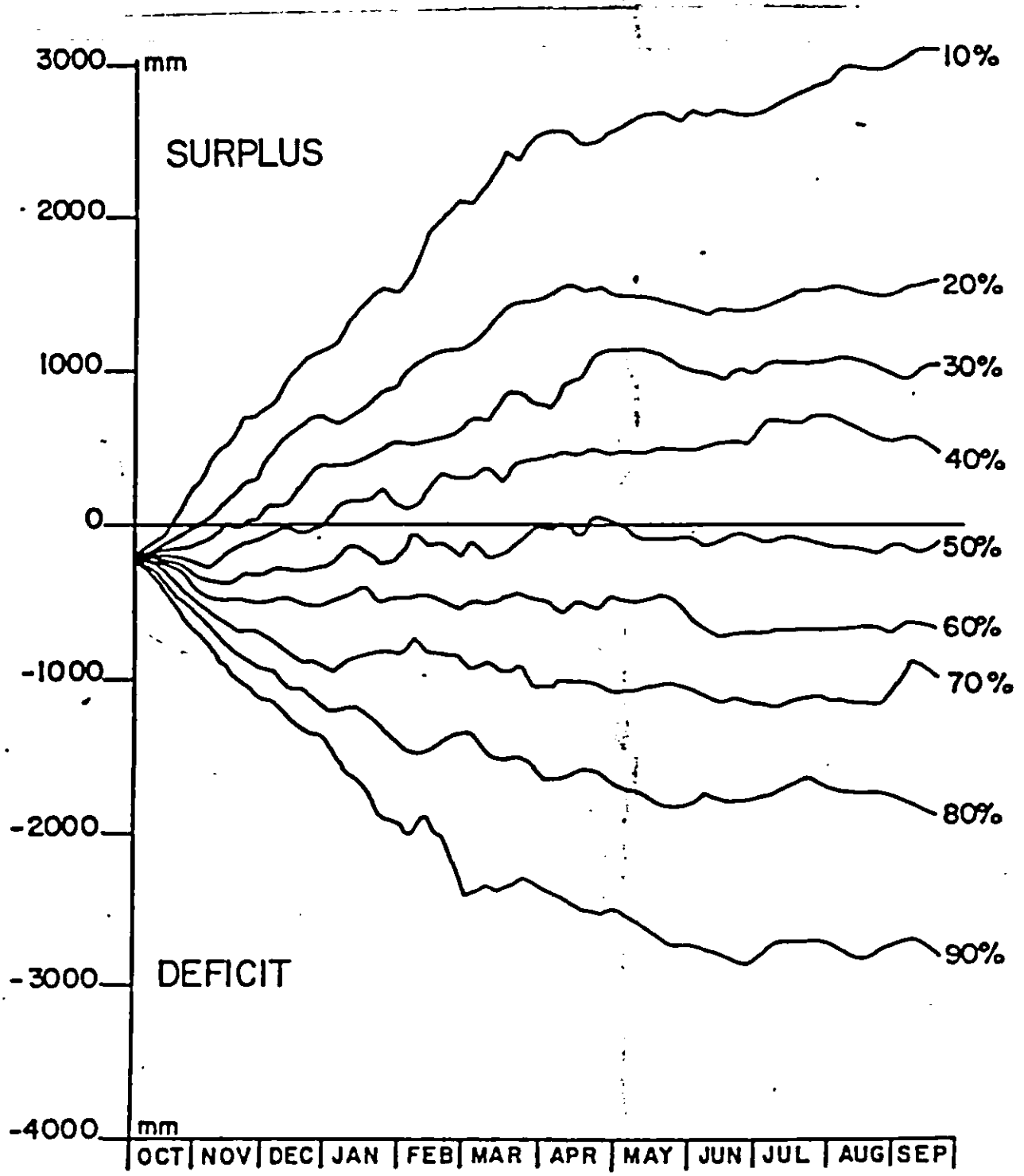
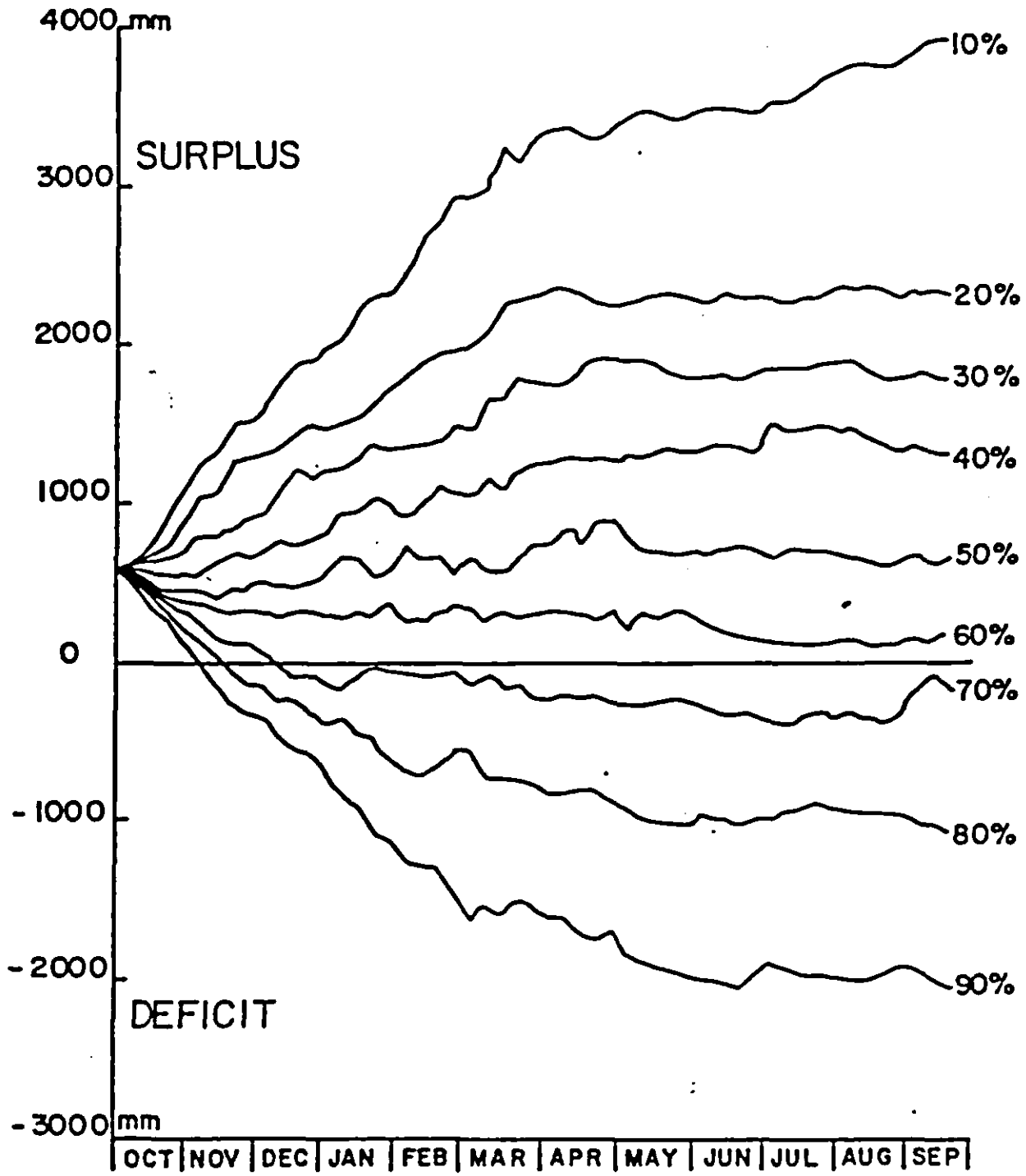


FIGURE 8.22 Estimates of cumulative weekly surplus/deficit over 52 weeks starting 1 October with a surplus of 600 mm at Pretoria.



The above applications of the drought model are mainly intended to be illustrative. There are many more which may be tailored to meet the requirements of almost any investigation of surplus/deficit with regard to the rainfall process. We have shown that the model can be successfully applied on a daily to annual basis, that it can be used to investigate the *risk* of events over days, weeks, months or years or the *chronology* of historical events on any time scale. It can be used to make assessments from any arbitrary starting point and can imitate a variety of physical processes of interest from soil moisture to streamflow. Above all the model provides a contribution towards the long-standing problems of drought research. When does a drought begin and end? What is its definition in terms of run length, run-sum and maximum deficit? What is its risk of occurrence? How long before recovery from a state of deficit occurs and with what risk will the drought last another season? The proposed model is capable of a considerable contribution to the answers to all of these questions and more.

APPENDIX 1

A SEASONAL LOGNORMAL MODEL

In this appendix we derive the maximum likelihood estimates for fitting a lognormal model to seasonal data. The same basic approach can be used to fit any other model to such data. The question of how many parameters should be included in the model is also discussed.

Suppose that the year is divided into NT intervals, e.g. 52 weeks, 365 days, etc ..., denoted by $T = 1,2,\ldots,NT$. Let $N(T)$ represent the number of times that it rained at time T and $R(I,T)$, $I = 1,2,\ldots,N(T)$, the rainfall depth on the Ith year that it rained at time T. Let $\mu(T)$ and $\sigma(T)$ represent the parameters of the lognormal distribution at time $T = 1,2,\ldots,NT$.

For precisely the same reasons which were given in Chapter 2 it is undesirable to estimate $\mu(T)$ and $\sigma(T)$ separately for each T. Instead we will again make use of the approximation obtained by truncating their respective Fourier Series which are of the form

$$\mu(T) = \sum_{i=1}^{NT} \mu_i \, \phi_i(T)$$

$$\sigma(T) = \sum_{i=1}^{NT} \sigma_i \, \phi_i(T) \qquad , \qquad T = 1,2,\ldots,NT$$

where $\phi_i(T)$ is defined in Section 2.4 and $\mu_1,\mu_2,\ldots,\mu_{NT}$; $\sigma_1, \sigma_2,\ldots,\sigma_{NT}$ are the Fourier coefficients of $\mu(T)$ and $\sigma(T)$ respectively.

Truncating the series to L(μ) and L(σ) terms respectively we define

$$\mu(T,L(\mu)) = \sum_{i=1}^{L(\mu)} \mu_i \, \phi_i(T) \quad ,$$

$$\sigma(T,L(\sigma)) = \sum_{i=1}^{L(\sigma)} \sigma_i \, \phi_i(T) \quad , \qquad T = 1,2,\ldots,NT \quad .$$
$$L(\mu),\ L(\sigma) < NT.$$

The approximation which we make is

$$\mu(T,L(\mu)) \approx \mu(T)$$

$$\sigma(T,L(\sigma)) \approx \sigma(T) \quad , \qquad T = 1,2,\ldots,NT \ .$$

The effects of varying L(μ) and L(σ) are analogous to those of varying L in section 2.4 and so will not be repeated in detail here. Briefly L(μ) must be large enough for the above approximation to be accurate but as small as possible in order to keep the uncertainties associated with sampling variation to a minimum. The same applies to L(σ). We suppose for the moment that both L(μ) and L(σ) are fixed and derive the equations whose solutions give the maximum likelihood estimates of the parameters, viz $\mu_1, \mu_2, \ldots, \mu_{L(\mu)}$ and $\sigma_1, \sigma_2, \ldots, \sigma_{L(\sigma)}$.

For convenience we use the notation

$$X(I,T) = \ell n \ R(I,T) \quad , \qquad I = 1,2,\ldots,N(T),$$
$$T = 1,2,\ldots,NT \ .$$

Since we are only dealing with wet days the rainfall depths are necessarily positive and so the above logarithms are always defined. As L(μ) and L(σ) are for the moment taken to be fixed we will use the briefer notation μ*(T) and

$\sigma^*(T)$ to represent $\mu(T,L(\mu))$ and $\sigma(T,L(\sigma))$ respectively in the derivation to follow.

Under the assumption that the rainfall depths are independently distributed the log likelihood function is given by

$$\ell n\ L\ (\mu,\sigma)$$

$$= \ell n\ L\ (\mu_i,\ i = 1,2,\ldots,L(\mu);\ \ \sigma_i,\ i = 1,2,\ldots,L(\sigma);$$

$$X(I,T),\ I = 1,2,\ldots,N(T),\ T = 1,2,\ldots,NT)]$$

$$= -\tfrac{1}{2} \sum_{T=1}^{NT} N(T)\ \ell n\ 2\pi - \sum_{T=1}^{NT} N(T)\ \ell n\ \sigma^*(T)$$

$$- \tfrac{1}{2} \sum_{T=1}^{NT} \left\{ \sum_{I=1}^{N(T)} (X(I,T)-\mu^*(T))^2 \right\} / \sigma^*(T)^2$$

It can be shown that

$$k(T) = \sum_{I=1}^{N(T)} (X(I,T)-\mu^*(T))^2 = s(T) + N(T)(m(T) - \mu^*(T))^2$$

where

$$s(T) = \sum_{I=1}^{N(T)} (X(I,T) - m(T))^2 \quad ,$$

$$m(T) = \frac{1}{N(T)} \sum_{I=1}^{N(T)} X(I,T) \qquad , \qquad \text{for } N(T) > 0 \ .$$

The maximum likelihood estimates are those values of the μ_i and σ_i which maximise $\ell n\ L\ (\mu,\sigma)$, i.e. they are the solutions to the $L(\mu) + L(\sigma)$ "normal equations", obtained by setting its partial derivatives equal to zero:

$$\frac{\partial L(\mu,\sigma)}{\partial \mu_a} = \sum_{T=1}^{NT} \{N(T)(m(T)-\mu^*(T))/\sigma^*(T)^2\}\ \phi_a(T),$$

$$a = 1,2,\ldots,L(\mu) \ .$$

$$\frac{\partial L(\mu,\sigma)}{\partial \sigma_a} = \sum_{T=1}^{NT}\{k(T)/\sigma^*(T)^3 - N(T)/\sigma^*(T)\}\ \phi_a(T),$$

$$a = 1,2,\ldots,L(\sigma)\ .$$

This system of equations cannot be solved analytically and so numerical methods have to be used. For the Newton-Raphson iteration method (and also for the purpose of estimating standard errors for the estimates) the second partial derivatives are required. These are given by:

$$\frac{\partial^2 L(\mu,\sigma)}{\partial \mu_a \partial \mu_b} = -\sum_{T=1}^{NT}\{N(T)/\sigma^*(T)^2\}\ \phi_a(T)\ \phi_b(T)\ ,$$

$$a,b, = 1,2,\ldots,L(\mu)$$

$$\frac{\partial^2 L(\mu,\sigma)}{\partial \mu_a \partial \sigma_b} = -2\sum_{T=1}^{NT}\{N(T)\ (m(T)-\mu^*(T))/\sigma^*(T)^3\}\ \phi_a(T)\ \phi_b(T)\ ,$$

$$a = 1,2,\ldots,L(\mu)\ ,$$
$$b = 1,2,\ldots,L(\sigma)\ ,$$

$$\frac{\partial^2 L(\mu,\sigma)}{\partial \sigma_a \partial \sigma_b} = \sum_{T=1}^{NT}\{N(T)/\sigma^*(T)^2 - 3k(T)/\sigma^*(T)^4\}\ \phi_a(T)\ \phi_b(T)\ ,$$

$$a,b = 1,2,\ldots,L(\sigma)\ .$$

The following initial estimates which are based on the method of least squares can be used to start the iteration:

$$\hat{\mu}_a^{(o)} = \sum_{\substack{T=1\\N(T)>0}}^{NT} m(T)\ \phi_a(T) \Big/ \sum_{\substack{T=1\\N(T)>0}}^{NT} \phi_a(T)^2\ ,\quad a = 1,2,\ldots,L(\mu)$$

$$\hat{\sigma}_a^{(o)} = \sum_{\substack{T=1\\N(T)>1}}^{NT} \{s(T)/N(T)\}^{\frac{1}{2}}\phi_a(T) \Big/ \sum_{\substack{T=1\\N(T)>1}}^{NT} \phi_a(T)^2\ ,$$

$$a = 1,2,\ldots,L(\sigma)\ .$$

Problems can arise in the iteration routine if $\hat{c}^*(T)$ becomes negative at any stage. This seldom occurs, and if it does, it can usually be remedied by experimenting a little with the starting values. Should the problem persist then one must increase either $L(\mu)$, $L(\sigma)$ or both.

We now give an outline of the algorithm to carry out the estimation. We will denote the arrays of first and second partial derivatives of $\ell n\, L(\mu,\sigma)$ at the kth iteration by $f^{(k)}$ and $F^{(k)}$ respectively, i.e. $f^{(k)}$ is a (column) vector with entries

$$f_i^{(k)} = \begin{cases} \partial \ell n\, L\,(\mu^{(k)},\sigma^{(k)})/\partial \mu_i & , \\ & i = 1,2,\ldots,L(\mu) \\ \partial \ell n\, L\,(\mu^{(k)},\sigma^{(k)})/\partial \sigma_{i-L(\mu)} & , \\ & i = L(\mu)+1,\ldots,L(\mu)+L(\sigma). \end{cases}$$

The (i,j)th entry of the matrix $F^{(k)}$ is given by

$$F_{ij}^{(k)} = \begin{cases} \partial^2 \ell n\, L\,(\mu^{(k)},\sigma^{(k)})/\partial \mu_i \partial \mu_j & , \\ & i = 1,2,\ldots,L(\mu) \\ & j = 1,2,\ldots,L(\mu) \\ \partial^2 \ell n\, L\,(\mu^{(k)},\sigma^{(k)})/\partial \mu_i \partial \sigma_{j-L(\mu)} & , \\ & i = 1,2,\ldots,L(\mu) \\ & j = L(\mu)+1,\ldots,L(\mu)+L(\sigma) \\ \partial^2 \ell n\, L\,(\mu^{(k)},\sigma^{(k)})/\partial \mu_{i-L(\sigma)} \partial \sigma_j & , \\ & i = L(\sigma)+1,\ldots,L(\sigma)+L(\mu) \\ & j = 1,2,\ldots,L(\sigma) \\ \partial^2 \ell n\, L\,(\mu^{(k)},\sigma^{(k)})/\partial \mu_{i-L(\mu)} \partial \sigma_{j-L(\mu)} & , \\ & i = L(\mu)+1,\ldots,L(\mu)+L(\sigma) \\ & j = L(\mu)+1,\ldots,L(\mu)+L(\sigma) \end{cases}$$

where $\mu^{(k)}$ and $\sigma^{(k)}$ are vectors representing the estimates of $\mu_1, \mu_2, \ldots, \mu_{L(\mu)}$ and $\sigma_1, \sigma_2, \ldots, \sigma_{L(\sigma)}$ respectively at the kth iteration. One substitutes these values into the formulae for the derivatives given earlier in this appendix and hence obtains $f^{(k)}$ and $F^{(k)}$.

ALGORITHM

STEP 1 Obtain initial estimates, $\mu^{(0)}$ and $\sigma^{(0)}$, and set $k = 0$.

STEP 2 Compute $f^{(k)}$ and $F^{(k)}$

STEP 3 Compute the vector $\delta^{(k)}$, the solution to the system of $L(\mu) + L(\sigma)$ linear equations given by

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

STEP 4 Set $\begin{pmatrix} \mu^{(k+1)} \\ \sigma^{(k+1)} \end{pmatrix} = \begin{pmatrix} \mu^{(k)} \\ \sigma^{(k)} \end{pmatrix} - \delta^{(k)}$

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

To speed up the algorithm one should make use of the fact that the matrix $F^{(k)}$ is symmetric, i.e. it is only necessary to actually compute the entries of the upper triangle of the matrix. Subroutines to solve linear equations directly are generally more efficient than those to compute the inverse of a matrix and it is therefore recommended that the equations in Step 2 be solved directly rather than by premultiplying the $f^{(k)}$ with the inverse of $F^{(k)}$.

Selection criteria

To select $L(\mu)$ and $L(\sigma)$, the number of terms in the approximations to $\mu(T)$ and $\sigma(T)$, the methods described in Linhart and Zucchini (1986) could be used, but in this case we recommend that the Akaike Information Criterion (AIC) be used instead. As we are estimating the parameters of the model by the method of maximum likelihood, the natural discrepancy on which to base model selection is the Kullback-Leibler discrepancy (cf. Appendix 2 in the report "Assessing the Risk of Deficiencies in Streamflow"). For this discrepancy and for the values of $L(\mu)$ and $L(\sigma)$ usually required, the two methods are practically equivalent except that the AIC is easier to compute. It is given by

$$AIC = - \ell n\, L\,(\hat{\mu},\hat{\sigma}) + L(\mu) + L(\sigma).$$

To apply the method one begins by setting $L(\mu) = L(\sigma) = 1$, fits the model and then computes the value of AIC associated with this set of estimates. Keeping in mind that both $L(\mu)$ and $L(\sigma)$ should be odd numbers, these are then systematically increased and at each stage the AIC is computed for the corresponding models. The values of $L(\mu)$ and $L(\sigma)$ which minimise AIC are selected as being the most appropriate.

This procedure was applied to 100 test stations in South Africa. The optimal values of $L(\mu)$ and $L(\sigma)$ which were obtained are illustrated in Figures A1.1 and A1.2.

It can be seen that for most stations the best $L(\mu)$ was equal to either 3 or 5. (There are twelve cases where 1 was best and a single case where 7 was best). There seems to be no clear systematic pattern in the distribution of these numbers. It should be noted that length of record

plays a major role in determining which $L(\mu)$ is estimated to be best. Were one to use a single $L(\mu)$ for all the stations in the country (because to select $L(\mu)$ for each of a very large number of stations is costly), $L(\mu) = 5$ would be the obvious choice. Although this would increase the discrepancy due to estimation for several of the stations, we are fitting so few parameters that this increase is quite small. The alternative of using $L(\mu) = 3$ is not very safe because it is not possible to assume that the increase in the discrepancy of approximation will be small if this number of terms is used for those stations where $L(\mu) = 5$ is estimated to be best.

The results for $L(\sigma)$ are much more consistent. In 86 out of the 100 cases, $L(\sigma) = 1$ was found to be best. By examining the records for which $L(\sigma) = 3$ was selected one can detect that many of these occur at stations where exceptionally long records are available. It is mainly this influence which gives rise to the additional parameters being selected rather than any marked seasonal variation in $\sigma(T)$, and so there is little danger of any large increase in the discrepancy of approximation if $L(\sigma) = 1$ is used for all the stations.

An important conclusion can be drawn from the fact that $L(\sigma) = 1$ is estimated to be the best choice: this is that the coefficient of variation for the untransformed rainfall totals must be very nearly constant. If one keeps in mind that this type of model selection criterion can be (roughly) compared with a statistical test using a 50% rejection level, then it follows that the hypothesis of constant coefficient of variation could certainly *not* be rejected using a more conventional significance level, of say 10%, for any of the cases with $L(\sigma) = 1$, and probably not for any of those with $L(\sigma) = 3$ either.

To demonstrate that the coefficient of variation of the untransformed data is a constant if $\sigma(T)$ is constant is very easy. This coefficient is not a function of $\mu(T)$, but only of $\sigma(T)$; it is given by

$$C(T) = (e^{\sigma(T)^2}-1)^{\frac{1}{2}}.$$

Clearly, if $\sigma(T)$ is constant for all T then so is $C(T)$.

FIGURE A1.1

Optimum L(μ) for each of 100 test stations.

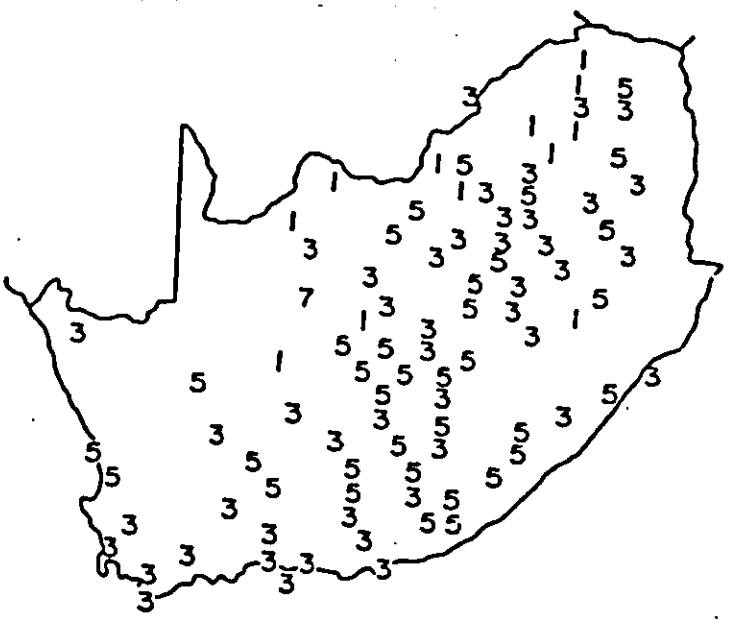
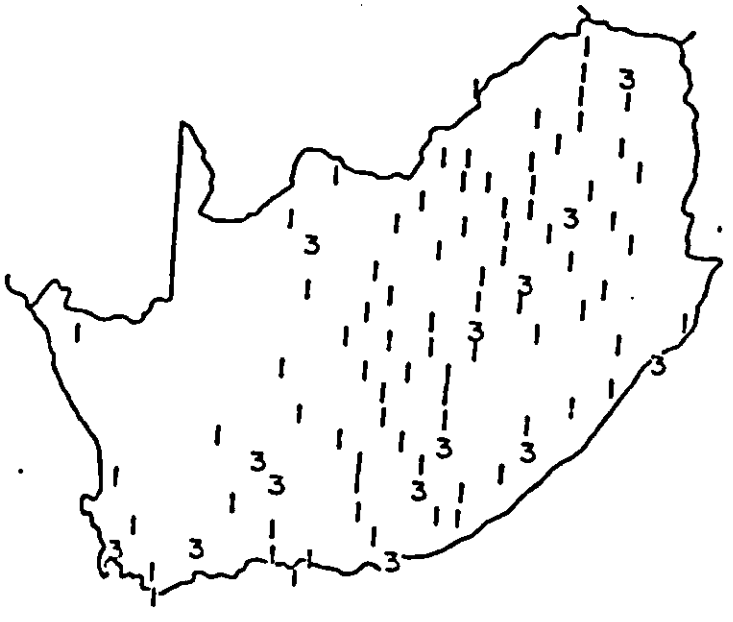


FIGURE A1.2

Optimum L(σ) for each of 100 test stations.



APPENDIX 2

A RATIONAL FUNCTION APPROXIMATION TO COMPUTE THE SHAPE PARAMETER OF THE WEIBULL DISTRIBUTION FROM THE COEFFICIENT OF VARIATION

The distribution function of the 2-parameter Weibull distribution is given by

$$F(x) = 1 - \exp\{-(x/\alpha)^{\beta}\} \quad , \quad x > 0, \tag{1}$$

where α is the scale parameter and β the shape parameter. The expectation (E), variance (V) and coefficient of variation (C) are given by

$$E = \alpha\, \Gamma(1 + 1/\beta) \tag{2}$$

$$V = \alpha^2\{\Gamma(1 + 2/\beta) - \Gamma(1 + 1/\beta)^2\} \tag{3}$$

$$C = \frac{\sqrt{V}}{E} = \sqrt{\frac{\Gamma(1+2/\beta)}{\Gamma(1+1/\beta)^2} - 1} \tag{4}$$

Note that the coefficient of variation is a function of the shape parameter but does not depend on the scale parameter.

Estimates of E, V and C can be computed from the observed sample $x_1, x_2, \ldots, x_n$ using the usual formulae:

$$\hat{E} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{5}$$

$$\hat{V} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{E})^2 \tag{6}$$

$$\hat{C} = \hat{V}^{\frac{1}{2}}/\hat{E} \ . \tag{7}$$

The method of moments to estimate the parameters α and β consists of replacing E and C in equations (2) and (4) by $\hat{E}$ and $\hat{C}$, and then solving these two equations to obtain estimates $\hat{\alpha}$ and $\hat{\beta}$. Equation (4) is solved first to obtain $\hat{\beta}$ which can then be used in equation (2) to obtain

$$\hat{\alpha} = \hat{E}/\Gamma(1 + 1/\hat{\beta}) \qquad (8)$$

The only difficulty which arises in this estimation algorithm is that of solving equation (4). It is not possible to give a closed expression for β in terms of C.

There are several ways to solve the equation numerically. For example Table 1 below gives pairs (β, C) for $\beta = 0{,}20(0{,}01)2{,}54$. This table covers the range which is likely to occur in practice. Linear or quadratic interpolation can be used to estimate β for a value of C which falls between the values given. Note that C is a monotonically decreasing function of β and so no problems of multiple values arise. Fig.A2.1 has been drawn on the basis of Table 1. This can be used to estimate β directly.

Whereas the above method is convenient for hand computation, a more accurate, more convenient and less space-consuming method is available for use on digital computers. The method consists of approximating the inverse function of (4) by means of a rational function approximation over the interval of interest, viz. $\beta \in [0{,}2 \; ; \; 2{,}5]$. We approximate β using

$$\beta \approx \frac{a_0 + a_1 C + a_2 C^2 + \ldots + a_k C^k}{1 + b_1 C + b_2 C^2 + \ldots + b_k C^k} \; . \qquad (9)$$

To compute the coefficients $a_0, a_1, \ldots, a_k, b_1, b_2, \ldots, b_k$ for a given value of k one proceeds as follows:

Select $(2k+1)$ distinct points $\beta_1, \beta_2, \ldots, \beta_{2k+1}$ in the range $[0{,}2 ; 2{,}5]$. These points should be approximately equally spaced and should cover the whole range. For each of these points compute the corresponding values of C. One then has $(2k+1)$ pairs of points (β_1, C_1), (β_2, C_2), $\ldots, (\beta_{2k+1}, C_{2k+1})$. The required coefficients are then the solutions to the *linear* system of equations:

$$\beta_i(1+b_1C_i+b_2C_i^2 + \ldots + b_kC_i^k) = a_o + a_1C_i + \ldots + a_kC_i^k ,$$

$$i = 1,2,\ldots,2k+1 ,$$

that is

$$b_1(\beta_iC_i) + b_2(\beta_iC_i^2) + \ldots + b_k(\beta_iC_i^k) - a_o - a_1C_i \ldots - a_kC_i^k = -\beta_i ,$$

$$i = 1,2,\ldots,2k+1 \quad (10)$$

There are $2k+1$ linear equations in $2k+1$ unknowns, viz $b_1, b_2, \ldots, b_k, a_o, a_1, \ldots, a_k$. These are computed and substituted into equation (9).

In theory this approximation increases in accuracy as the number of coefficients $(2k+1)$ is increased. In practice if k is too large the accuracy actually decreases due to rounding errors on the computer. For the approximation considered here $k = 3$ (7 coefficients) is appropriate. The corresponding coefficients are given in Table 2 and the error involved using this approximation can be gauged from the values given in Table 3. From this table it can be deduced that the absolute error is less than 0,0027. This is accurate enough for most purposes, but if greater accuracy is required then iterative methods are available, in particular the Newton-Raphsonmethod. The above rational function approximation can be used as a starting value for the iteration.

TABLE 1

β	C	β	C	β	C	β	C	β	C
0.20	15.8430	0.21	13.5794	0.22	11.8066	0.23	10.3930	0.24	9.2477
0.25	8.3066	0.26	7.5236	0.27	6.8646	0.28	6.3043	0.29	5.8236
0.30	5.4077	0.31	5.0451	0.32	4.7267	0.33	4.4455	0.34	4.1955
0.35	3.9721	0.36	3.7714	0.37	3.5904	0.38	3.4264	0.39	3.2771
0.40	3.1409	0.41	3.0159	0.42	2.9011	0.43	2.7952	0.44	2.6972
0.45	2.6064	0.46	2.5219	0.47	2.4431	0.48	2.3695	0.49	2.3007
0.50	2.2361	0.51	2.1753	0.52	2.1182	0.53	2.0642	0.54	2.0133
0.55	1.9650	0.56	1.9193	0.57	1.8759	0.58	1.8347	0.59	1.7955
0.60	1.7581	0.61	1.7224	0.62	1.6883	0.63	1.6558	0.64	1.6246
0.65	1.5948	0.66	1.5661	0.67	1.5386	0.68	1.5122	0.69	1.4869
0.70	1.4626	0.71	1.4389	0.72	1.4162	0.73	1.3944	0.74	1.3733
0.75	1.3529	0.76	1.3332	0.77	1.3141	0.78	1.2957	0.79	1.2778
0.80	1.2605	0.81	1.2437	0.82	1.2275	0.83	1.2117	0.84	1.1964
0.85	1.1815	0.86	1.1670	0.87	1.1530	0.88	1.1393	0.89	1.1260
0.90	1.1130	0.91	1.1004	0.92	1.0881	0.93	1.0761	0.94	1.0644
0.95	1.0530	0.96	1.0419	0.97	1.0311	0.98	1.0205	0.99	1.0101
1.00	1.0000	1.01	0.9901	1.02	0.9804	1.03	0.9710	1.04	0.9618
1.05	0.9527	1.06	0.9438	1.07	0.9352	1.08	0.9267	1.09	0.9184
1.10	0.9102	1.11	0.9022	1.12	0.8944	1.13	0.8867	1.14	0.8792
1.15	0.8718	1.16	0.8646	1.17	0.8575	1.18	0.8505	1.19	0.8436
1.20	0.8369	1.21	0.8303	1.22	0.8238	1.23	0.8174	1.24	0.8112
1.25	0.8050	1.26	0.7989	1.27	0.7930	1.28	0.7871	1.29	0.7814
1.30	0.7757	1.31	0.7701	1.32	0.7647	1.33	0.7593	1.34	0.7540
1.35	0.7487	1.36	0.7436	1.37	0.7385	1.38	0.7335	1.39	0.7286
1.40	0.7238	1.41	0.7190	1.42	0.7143	1.43	0.7096	1.44	0.7051
1.45	0.7006	1.46	0.6961	1.47	0.6917	1.48	0.6874	1.49	0.6832
1.50	0.6790	1.51	0.6748	1.52	0.6707	1.53	0.6667	1.54	0.6627
1.55	0.6588	1.56	0.6549	1.57	0.6511	1.58	0.6473	1.59	0.6436
1.60	0.6399	1.61	0.6363	1.62	0.6327	1.63	0.6291	1.64	0.6256
1.65	0.6222	1.66	0.6188	1.67	0.6154	1.68	0.6120	1.69	0.6087
1.70	0.6055	1.71	0.6023	1.72	0.5991	1.73	0.5959	1.74	0.5928
1.75	0.5897	1.76	0.5867	1.77	0.5837	1.78	0.5807	1.79	0.5778
1.80	0.5749	1.81	0.5720	1.82	0.5691	1.83	0.5663	1.84	0.5636
1.85	0.5608	1.86	0.5581	1.87	0.5554	1.88	0.5527	1.89	0.5501
1.90	0.5475	1.91	0.5449	1.92	0.5423	1.93	0.5398	1.94	0.5373
1.95	0.5348	1.96	0.5323	1.97	0.5299	1.98	0.5275	1.99	0.5251
2.00	0.5227	2.01	0.5204	2.02	0.5181	2.03	0.5158	2.04	0.5135
2.05	0.5112	2.06	0.5090	2.07	0.5068	2.08	0.5046	2.09	0.5024
2.10	0.5003	2.11	0.4982	2.12	0.4960	2.13	0.4940	2.14	0.4919
2.15	0.4898	2.16	0.4878	2.17	0.4858	2.18	0.4838	2.19	0.4818
2.20	0.4798	2.21	0.4779	2.22	0.4760	2.23	0.4740	2.24	0.4721
2.25	0.4703	2.26	0.4684	2.27	0.4665	2.28	0.4647	2.29	0.4629
2.30	0.4611	2.31	0.4593	2.32	0.4575	2.33	0.4558	2.34	0.4540
2.35	0.4523	2.36	0.4506	2.37	0.4489	2.38	0.4472	2.39	0.4455
2.40	0.4438	2.41	0.4422	2.42	0.4406	2.43	0.4389	2.44	0.4373
2.45	0.4357	2.46	0.4341	2.47	0.4326	2.48	0.4310	2.49	0.4294
2.50	0.4279	2.51	0.4264	2.52	0.4249	2.53	0.4234	2.54	0.4219

The coefficient of variation, C, for the Weibull distribution corresponding to selected values of the shape parameter, β .

FIGURE A2.1

The shape parameter, β, of the Weibull distribution as a function of the coefficient of variation, C.

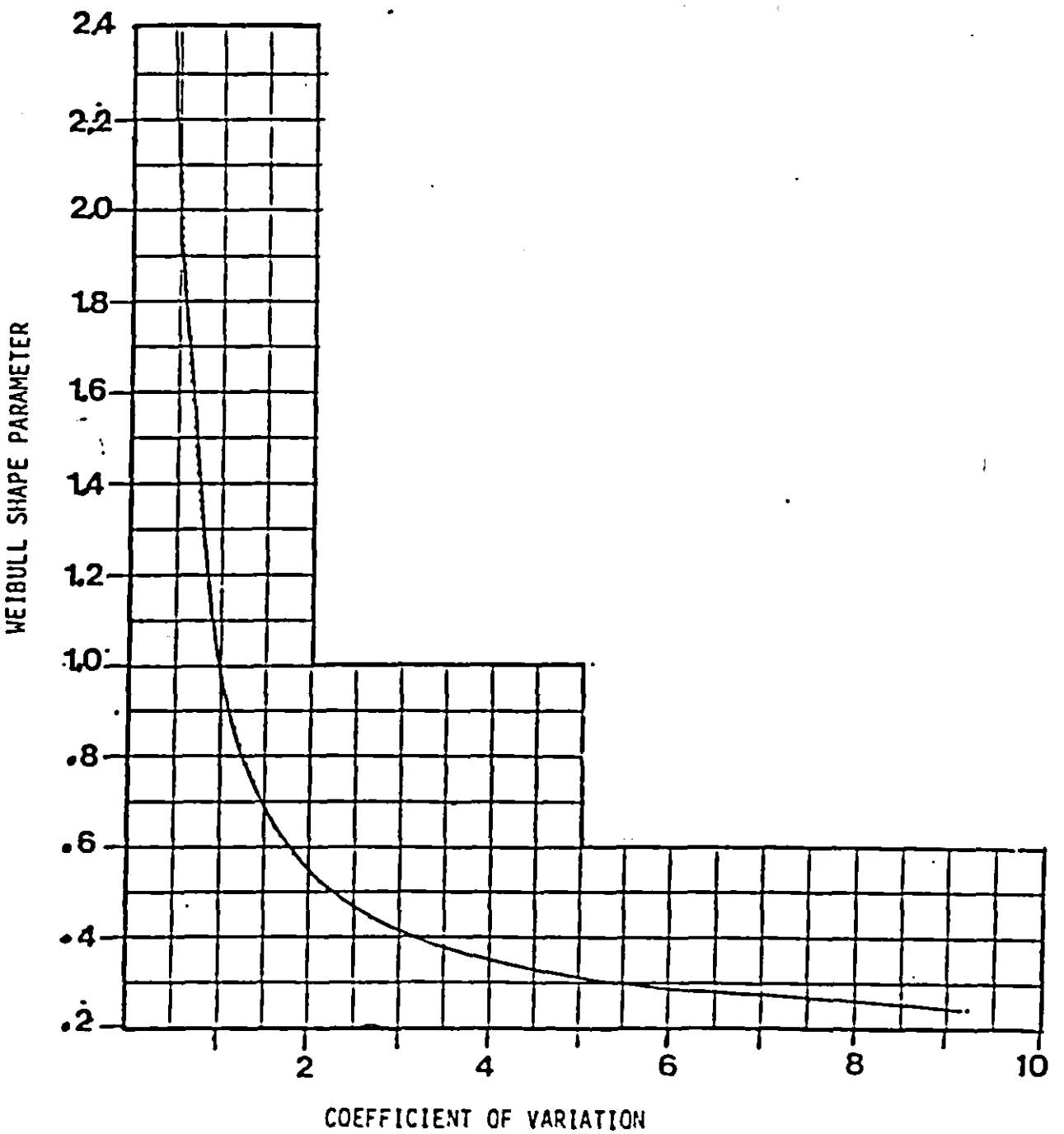


TABLE 2

i =	0	1	2	3
α_i	339,5410	148,4445	192,7492	22,4401
β_i	(1,0000)	257,1162	287,8362	157,2230

Coefficients in the rational approximation for β as a function of C for the Weibull distribution.

TABLE 3

C	β(exact)	β(approx)	difference
0.40	2.6956	2.6956	0.0000
0.80	1.2582	1.2582	-0.0000
1.20	0.8376	0.8376	-0.0000
1.60	0.6482	0.6482	0.0000
2.00	0.5427	0.5426	-0.0000
2.40	0.4756	0.4755	-0.0002
2.80	0.4291	0.4288	-0.0004
3.21	0.3949	0.3942	-0.0007
3.62	0.3685	0.3675	-0.0010
4.03	0.3474	0.3461	-0.0013
4.44	0.3302	0.3286	-0.0016
4.86	0.3158	0.3139	-0.0018
5.27	0.3035	0.3015	-0.0021
5.69	0.2930	0.2907	-0.0023
6.11	0.2838	0.2814	-0.0024
6.53	0.2757	0.2732	-0.0025
6.96	0.2685	0.2659	-0.0026
7.38	0.2621	0.2595	-0.0026
7.80	0.2563	0.2537	-0.0026
8.22	0.2511	0.2484	-0.0026
8.63	0.2463	0.2437	-0.0026
9.05	0.2420	0.2394	-0.0025
9.46	0.2380	0.2355	-0.0025
9.87	0.2343	0.2319	-0.0024
10.28	0.2309	0.2286	-0.0023
10.69	0.2277	0.2256	-0.0021
11.09	0.2248	0.2228	-0.0020
11.49	0.2221	0.2202	-0.0019
11.89	0.2195	0.2178	-0.0017
12.28	0.2171	0.2155	-0.0016
12.67	0.2149	0.2134	-0.0014
13.05	0.2127	0.2115	-0.0013
13.43	0.2108	0.2097	-0.0011
13.81	0.2089	0.2079	-0.0009
14.18	0.2071	0.2063	-0.0008
14.55	0.2054	0.2048	-0.0006
14.91	0.2038	0.2034	-0.0004
15.27	0.2023	0.2020	-0.0003
15.63	0.2008	0.2007	-0.0001
15.98	0.1995	0.1995	0.0001

The shape parameter, β, for the Weibull distribution computed for selected values of the coefficient of variation, C, using the rational approximation compared with the exact value of β for each C.

APPENDIX 3

AN EFFICIENT METHOD TO COMPUTE THE SINE AND COSINE TERMS IN FOURIER EXPANSIONS

In order to fit the rainfall model described in Chapters 2 and 3, to carry out the algorithms for model selection and to apply the model for generating artificial rainfall sequences, one has to make repeated use of Fourier series representations. This involves the computation of a large number of sine and cosine terms, especially when one is dealing with daily rainfall series. The evaluation of sine and cosine functions is slow on a computer and it is therefore of some importance that such computations be carried out efficiently.

For daily data we need to compute the terms

$$\phi_i(T) = \begin{cases} \cos\left(\omega(T-1)i/2\right) & , \quad i = 2,4,\ldots,L-1, \\ \sin\left(\omega(T-1)(i-1)/2\right) & , \quad i = 3,5,\ldots,L \quad , \end{cases} \qquad T = 1,2,\ldots,365,$$

where $\omega = 2\pi/365$ and L represents the number of terms in the expansion which in our application is always taken to be odd. $\phi_1(T)$ is simply equal to 1.

Where storage requirements allow it, it is strongly recommended that these computations be carried out only once for the maximum L which may be required (typically L is less than 25 for applications involving model selection) and that the results be stored in an array which we will call PHI(I,T) in the algorithm given below. To compute $\phi_i(T)$ efficiently one makes use of the following recurrence relation:

$$\phi_i(T) = a_i\phi_i(T-1) - \phi_i(T-2) \quad , \quad T = 3,4,5,\ldots$$

where

$$\phi_i(1) = 1 \quad ,$$

$$\phi_i(2) = \begin{cases} a_i/2 & i = 2,4,\ldots,L-1 \\ \sin(\omega(i-1)/2) & i = 3,5,\ldots,L \end{cases}$$

$$a_i = \begin{cases} 2\cos(\omega i/2) & i = 2,4,\ldots,L-1 \\ a_{i-1} & i = 3,5,\ldots,L \; . \end{cases}$$

This relationship follows from well-known properties of the sine and cosine functions (see e.g. Abramowitz and Stegun 1972). The following algorithm can be used to compute PHI(I,T).

```
ALGORITHM

STEP 1  INPUT  L (an odd integer)

STEP 2  SET  W = 0,01721421
             K = (L-1)/2
             PHI(1,T) = 1,  T = 1,2,...,365

STEP 3  LOOP  FOR  J = 1,2,...,K

           SET  J1 = 2*J
                J2 = J1+1
                THETA = W*J
                A = 2*COS(THETA)
                PHI(J1,1) = 1
                PHI(J1,2) = A/2
                PHI(J2,1) = 0
                PHI(J2,2) = SIN(THETA)

        LOOP  FOR  T = 3,4,...,365

           SET  PHI(J1,T) = A*PHI(J1,T-1) - PHI(J1,T-2)
                PHI(J2,T) = A*PHI(J2,T-1) - PHI(J2,T-2)

        END  OF  T  LOOP

        END  OF  J  LOOP

STEP 4  OUTPUT  ARRAY PHI
```

APPENDIX 4

ESTIMATING THE RESPONSE FUNCTION OF A LINEARLY FILTERED PROCESS

This appendix describes a procedure to estimate the response function of a process which satisfies the conditions given in section 7.1. We suppose that a rainfall record $R(t)$, $t = 1,2,\ldots,n$ is available and that a shorter record of observations on the filtered process $F(t)$, $t = n_1, n_1+1,\ldots,n$, where $1 < n_1 < n$ is also available. We assume here that the concurrent portion occurs at the end of the rainfall record because this is what would generally happen in practice, but the methods outlined below can be applied in situations where the concurrent part occurs in the middle of the rainfall record. Our object here is to estimate the response function $r(x)$ and then to use the estimate to reconstruct the values of $F(t)$ over the non-concurrent part of the rainfall record. Having done this it is then possible to define a drought index directly in terms of $F(t)$, the process of interest, and one has enough data to construct models for this process in order to assess drought risk.

As the smallest unit of time considered is one day it is sufficient to estimate $r(x)$ for $x = 0,1,2,\ldots$. We have also assumed that $r(x)$ decays with time; suppose that $r(x)$ is effectively zero for $x > L$ and that $L < \min(n_1, n-n_1)$.

From (1) of section 7.1 it is assumed that $F(t)$ and $R(t)$ are related by

$$F(t) = \sum_{i=0}^{\infty} r(i)\, R(t-i) \quad , \quad t = n_1, n_1+1, \ldots, n \ ,$$

which we have now assumed can be approximated by

$$F(t) = \sum_{i=0}^{L-1} r(i)\, R(t-i) \quad , \quad t = n_1, n_1+1, \ldots, n \ .$$

In practice this relationship will not be exact because of measurement errors and random disturbances. What we have in in fact is of the form

$$F(t) = \sum_{i=0}^{L-1} r(i)\, R(t-i) + e(t)$$

where e(t) represents the combined effect of these deviations from the model on day t.

We will use the following notation:

$$F = \begin{pmatrix} F(n) \\ F(n-1) \\ \vdots \\ F(n_1) \end{pmatrix} \ , \quad R = \begin{pmatrix} R(n) & R(n-1) & \ldots & R(n-L) \\ R(n-1) & R(n-2) & \ldots & R(n-1-L) \\ \vdots & & & \\ R(n_1) & R(n_1-1) & \ldots & R(n_1-L) \end{pmatrix} \ ,$$

$$e = \begin{pmatrix} e(n) \\ e(n-1) \\ \vdots \\ e(n_1) \end{pmatrix} \quad \text{and} \quad r = \begin{pmatrix} r(0) \\ r(1) \\ \vdots \\ r(L-1) \end{pmatrix}$$

In terms of the model the observations can be now represented in the form

$$F = Rr + e \ ,$$

which is a linear model of the type encountered in re-

gression analysis or the analysis of variance. The standard method of estimating the unknown parameters, i.e. $r(0),r(1),..,r(L-1)$ is that of ordinary least squares. One obtains

$$\hat{r} = (R^T R)^{-1} R^T F \quad .$$

This not only provides us with an estimator of the response function but also a means of assessing the accuracy of the model. By examining the residuals

$$\hat{e}(t) = F(t) - \sum_{i=0}^{L-1} \hat{r}(i)\, R(t-i) \quad , \quad t = n_1+1, n_1+2, \ldots, n \ ,$$

one can clarify a number of issues. For example their standard deviation provides a measure of the accuracy of the model; a small value indicates a high accuracy whereas if the standard deviation is large then the model may have to be discarded. Individual values of $\hat{e}(t)$ may be large, thereby indicating anomalies on the corresponding days. Explanations for such anomalies can be sought and may provide important insights which further our understanding of the process. The presence of serial correlation in the residual series may also indicate lack of fit or perhaps some other feature of the process. The techniques to carry out these analyses are well documented in the statistical literature (see e.g. Draper and Smith 1966 , Box and Jenkins 1970), and so will not be repeated here. Having estimated the response function and assuming that the model has been proved to provide an acceptable fit, one can then estimate the missing F(t) by using

$$\hat{F}(t) = \sum_{i=0}^{L-1} \hat{r}(i)\, R(t-i) \quad , \quad t = L, L+1, \ldots, n_1-1 \ .$$

These, together with the directly observed values of the process, can then provide the basis of an analysis of drought risk on the process of interest.

APPENDIX 5

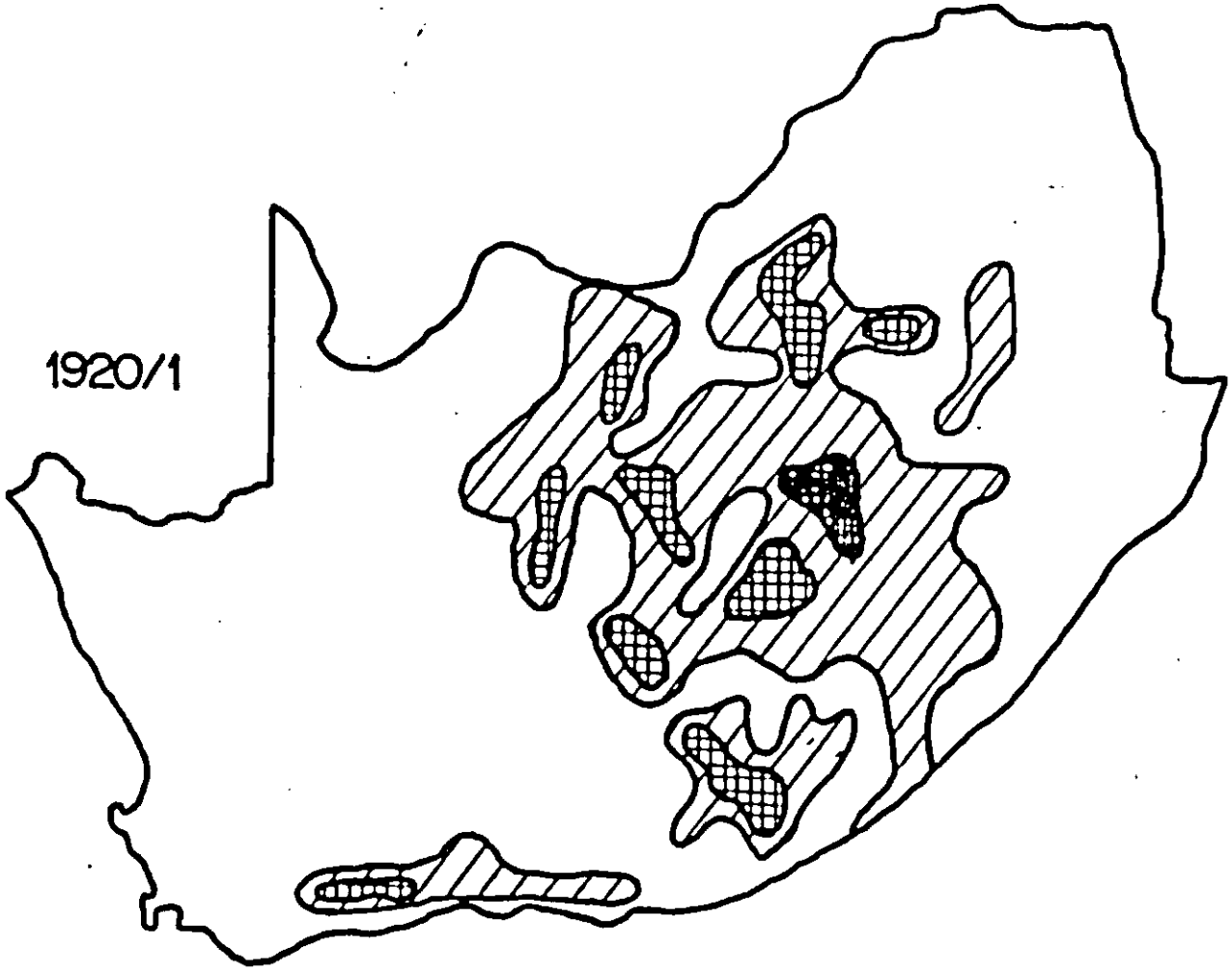
A SPATIAL HISTORY OF DROUGHT OVER SOUTH AFRICA

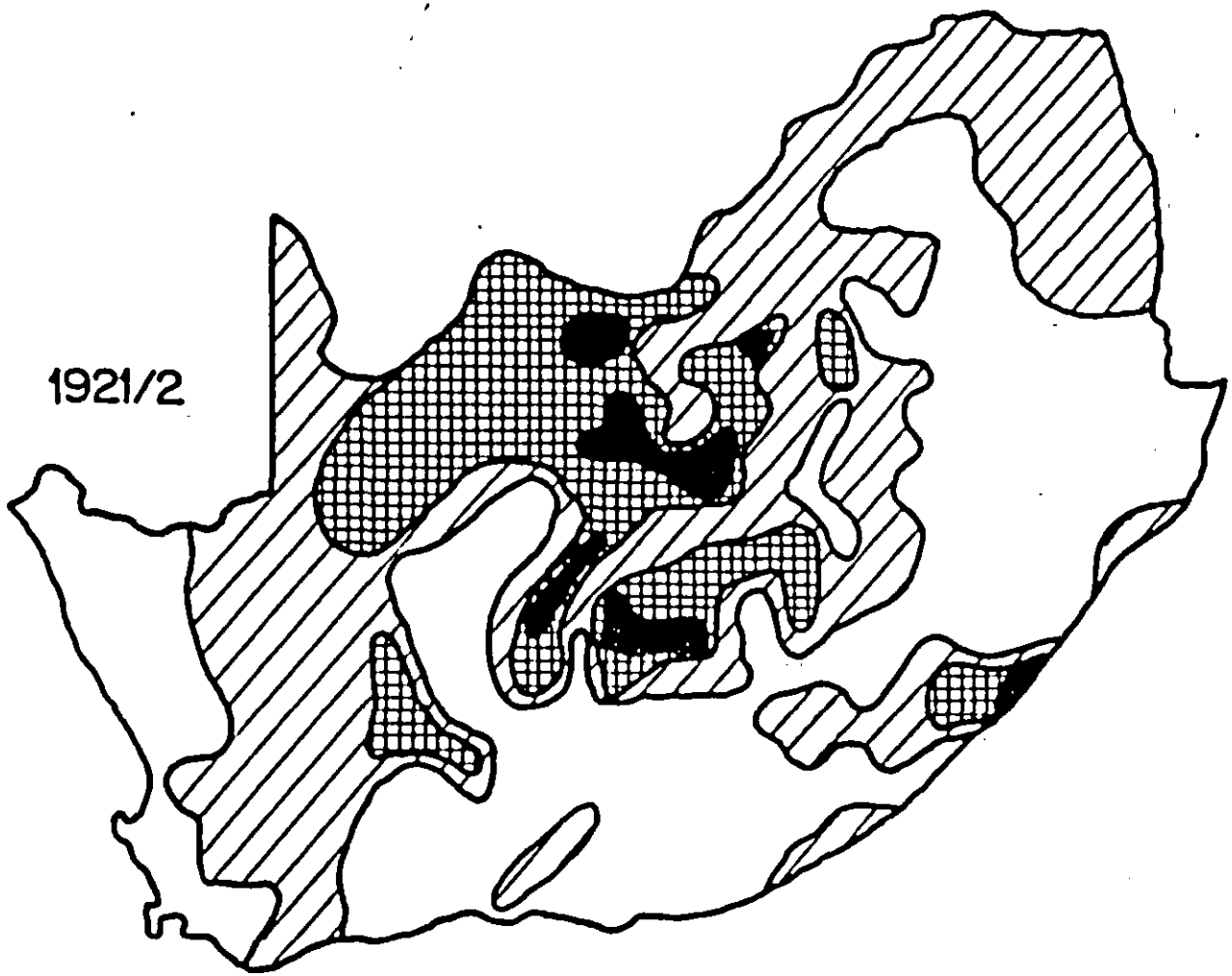
This appendix contains the sequence of drought maps described in Chapter 6. The chronological sequence is shown in water years starting on 1 October and ending on 30 September. The key to the shading of the maps is as follows:

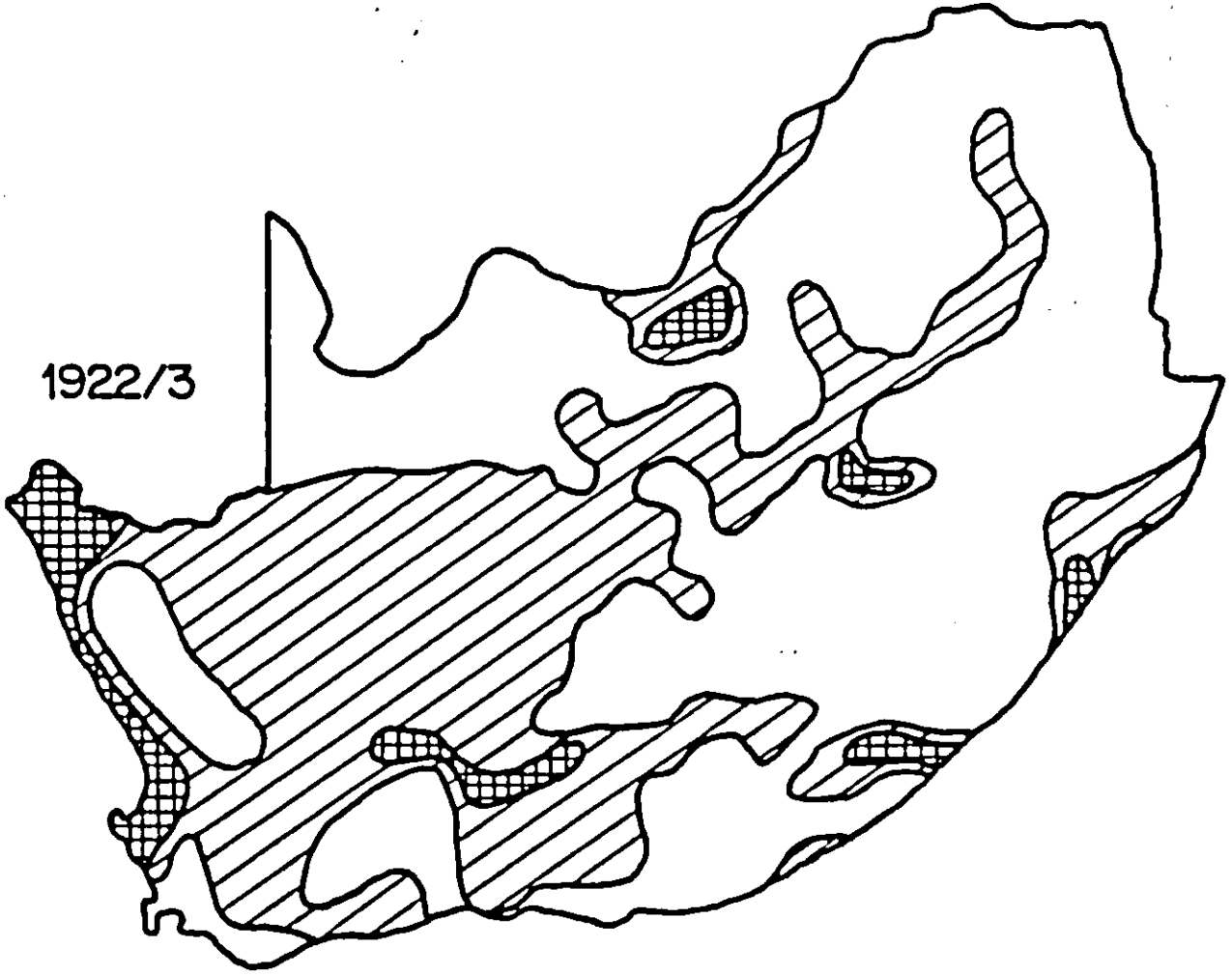
Annual rainfall less than 50% percentile

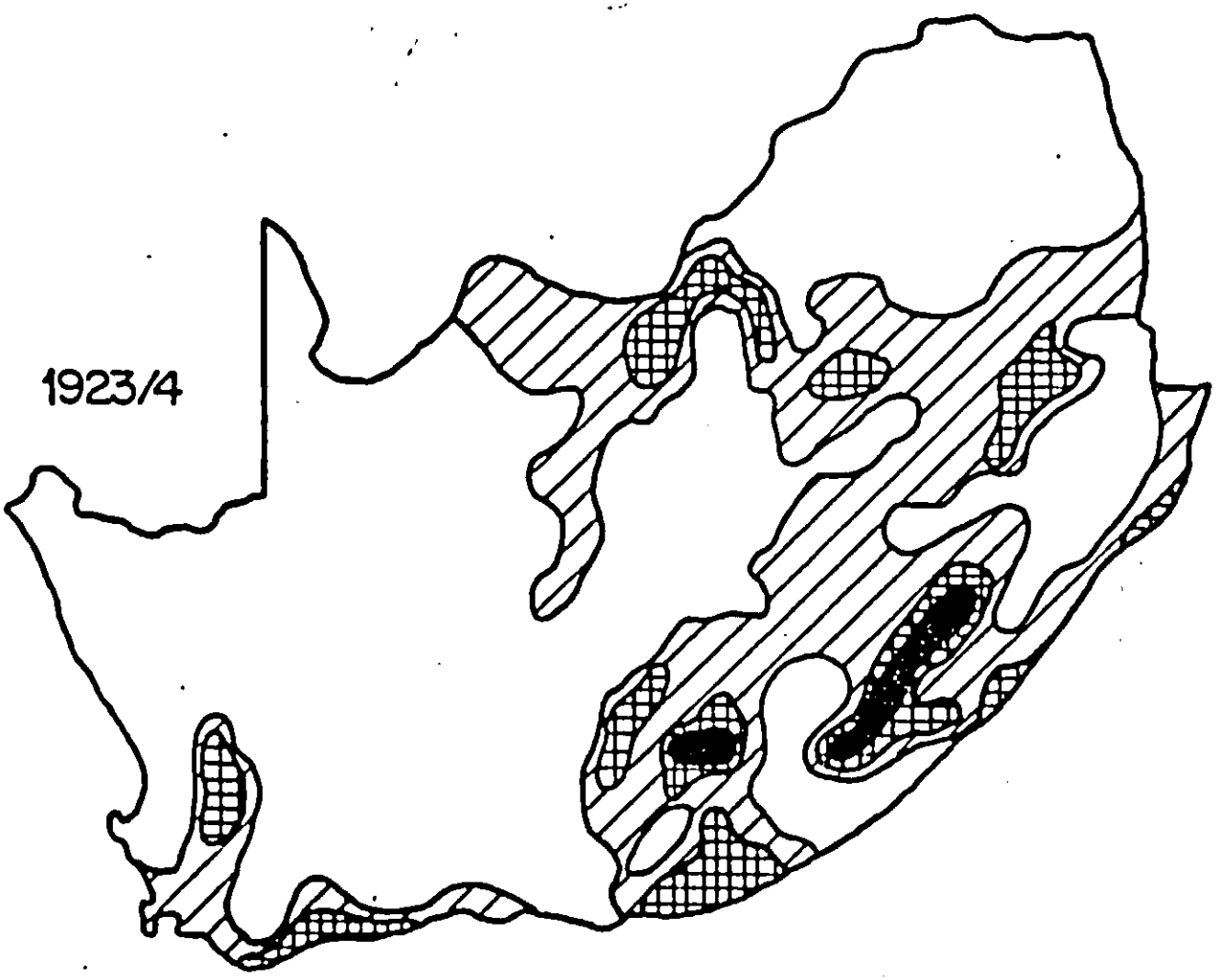
Annual rainfall less than 20% percentile

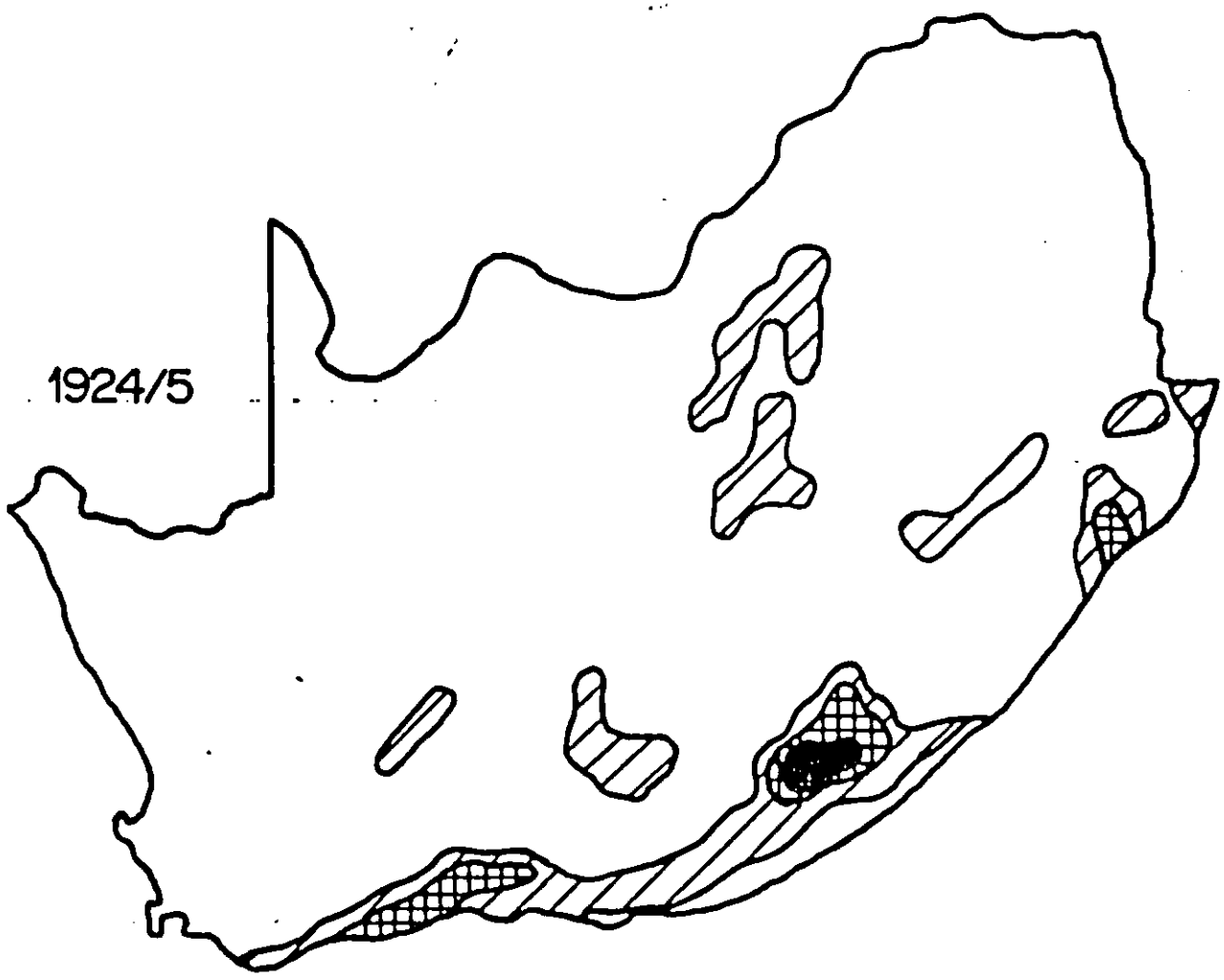
Annual rainfall less than 5% percentile

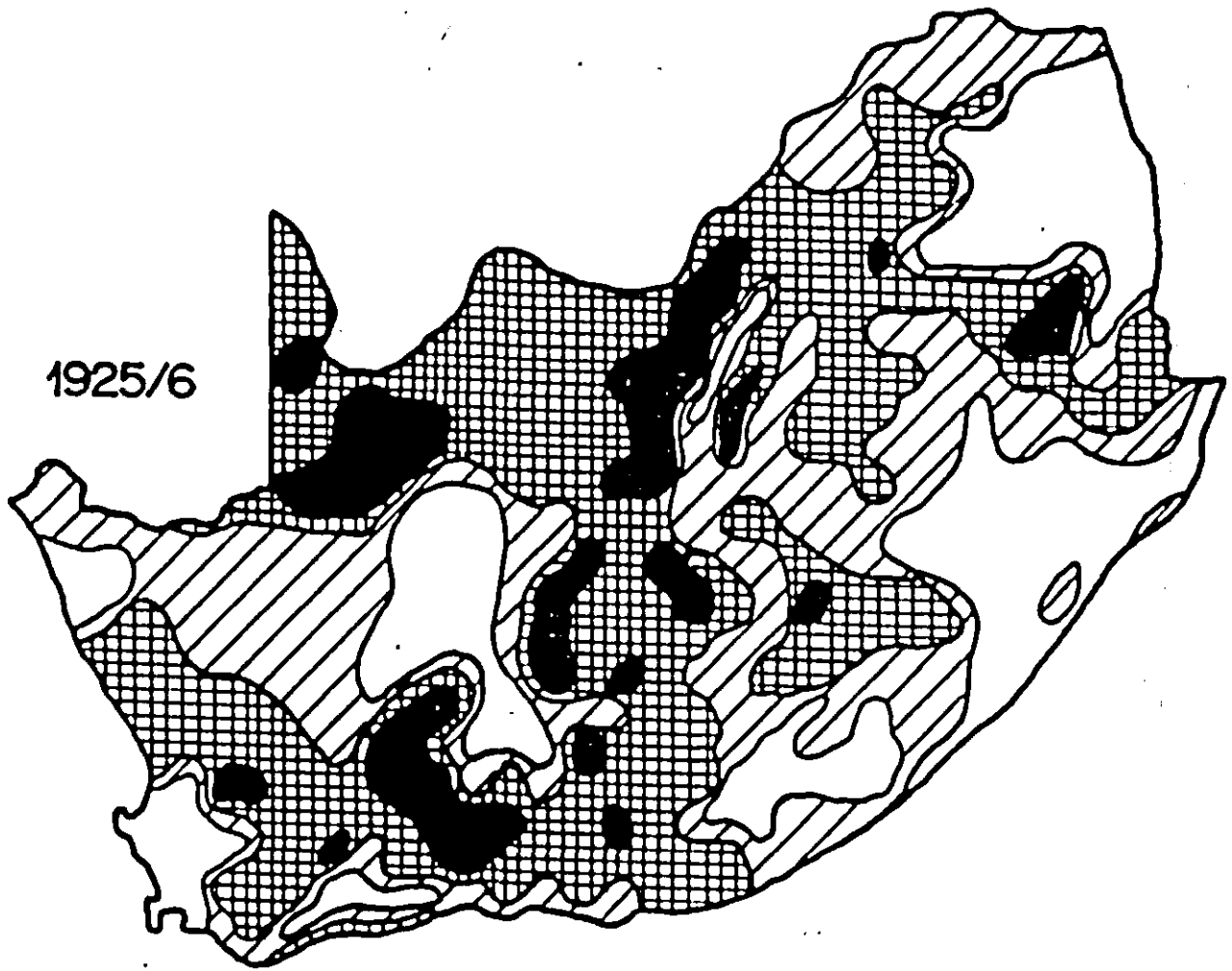

1920/1

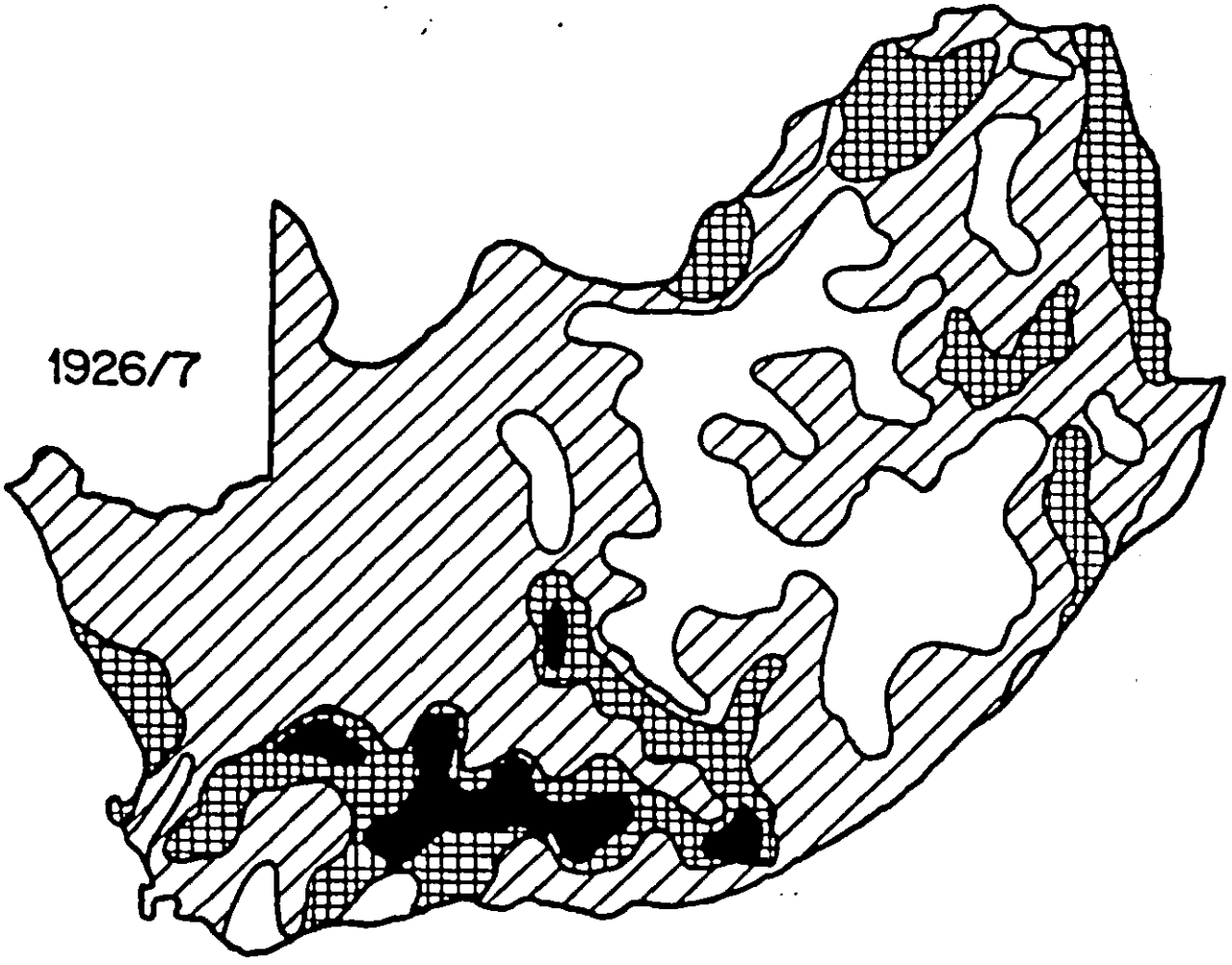

1921/2

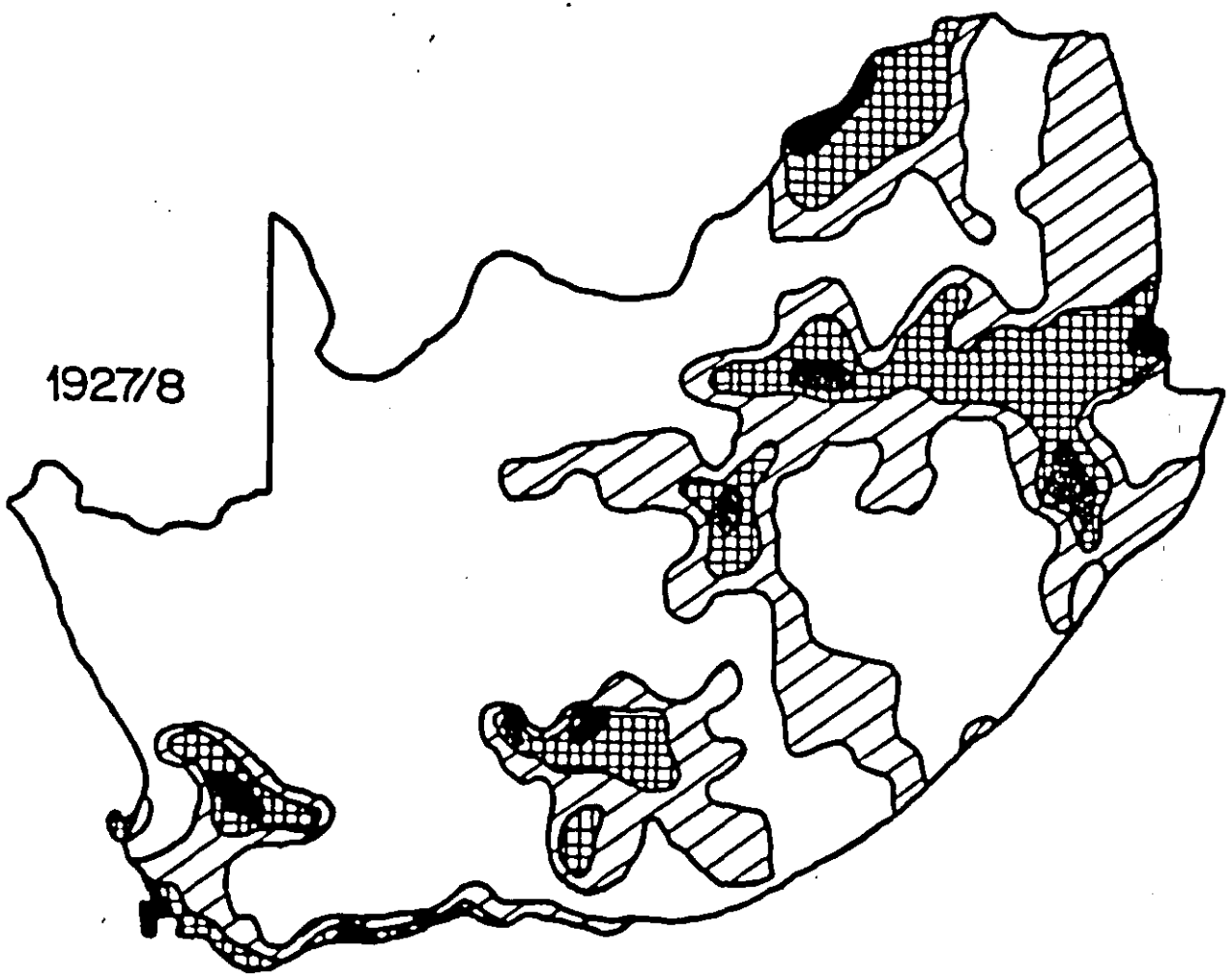

1922/3

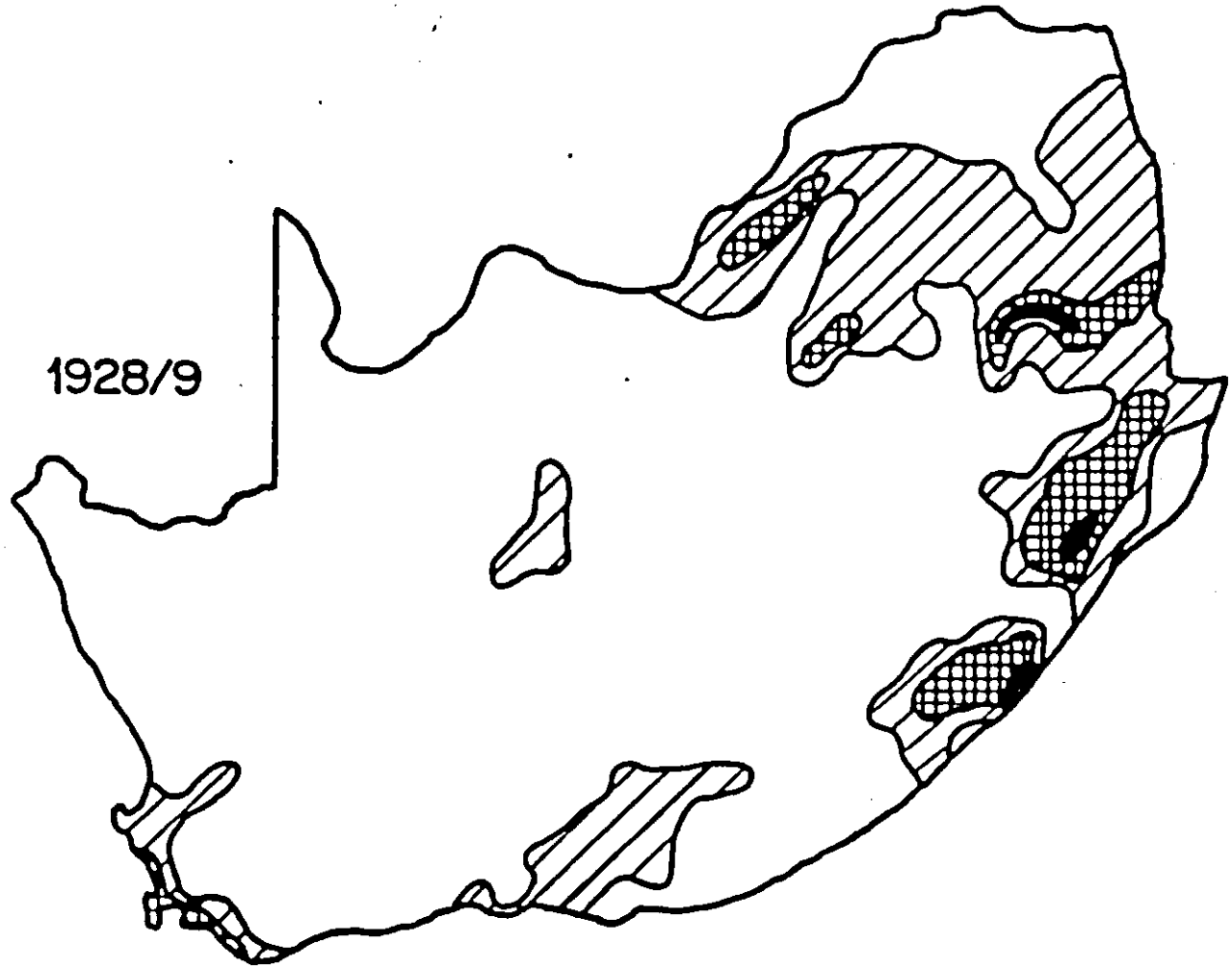

1923/4

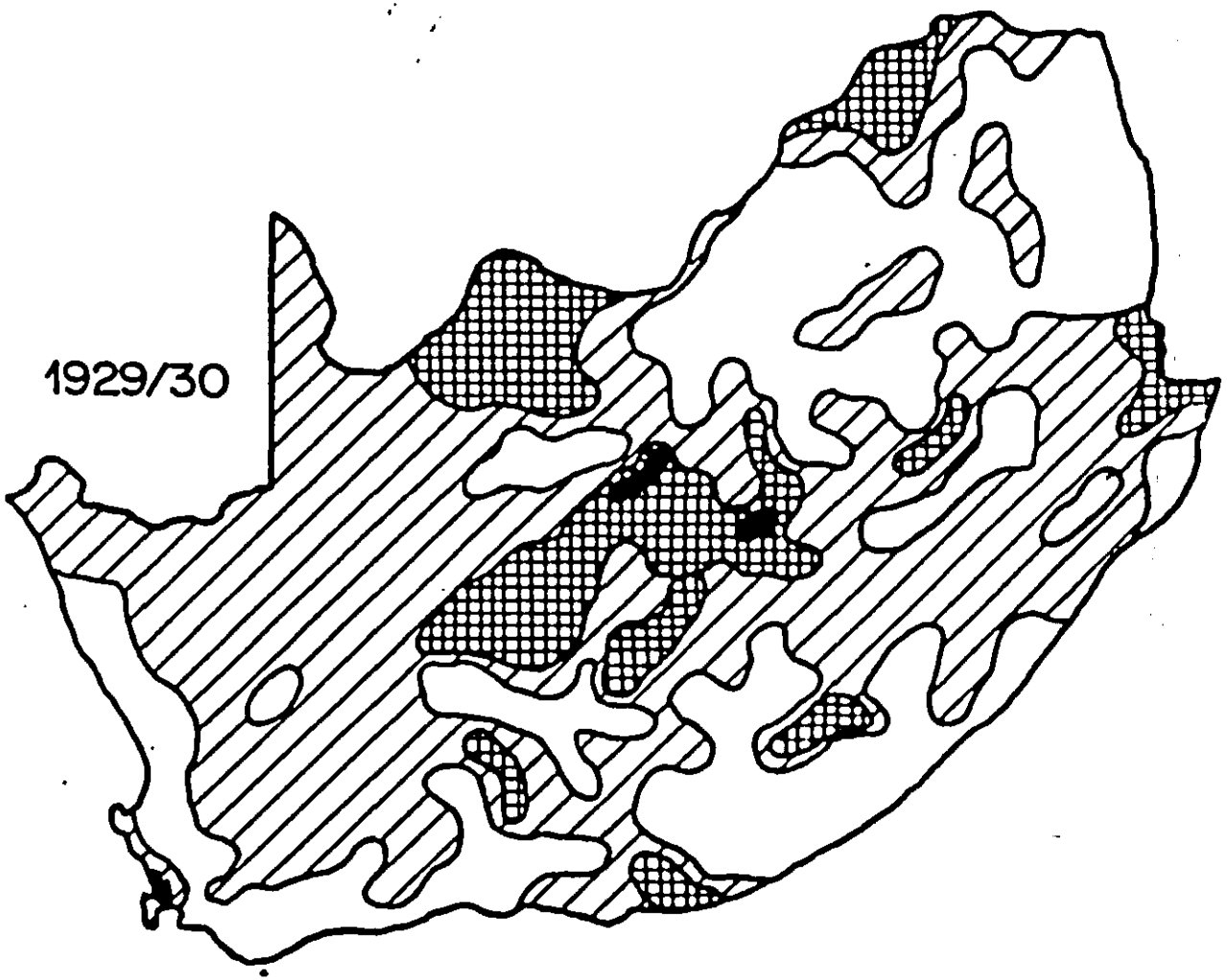

1924/5

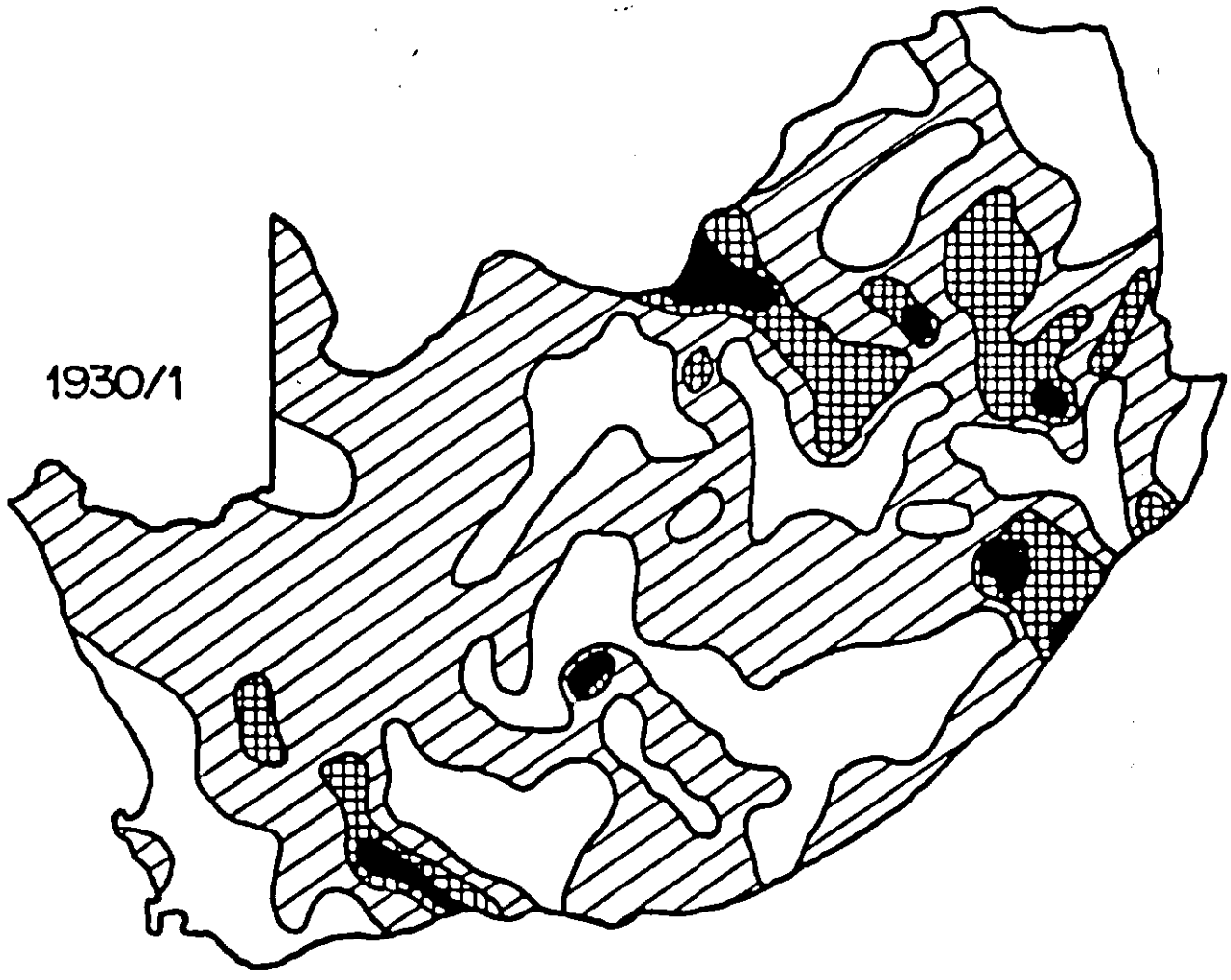

1925/6

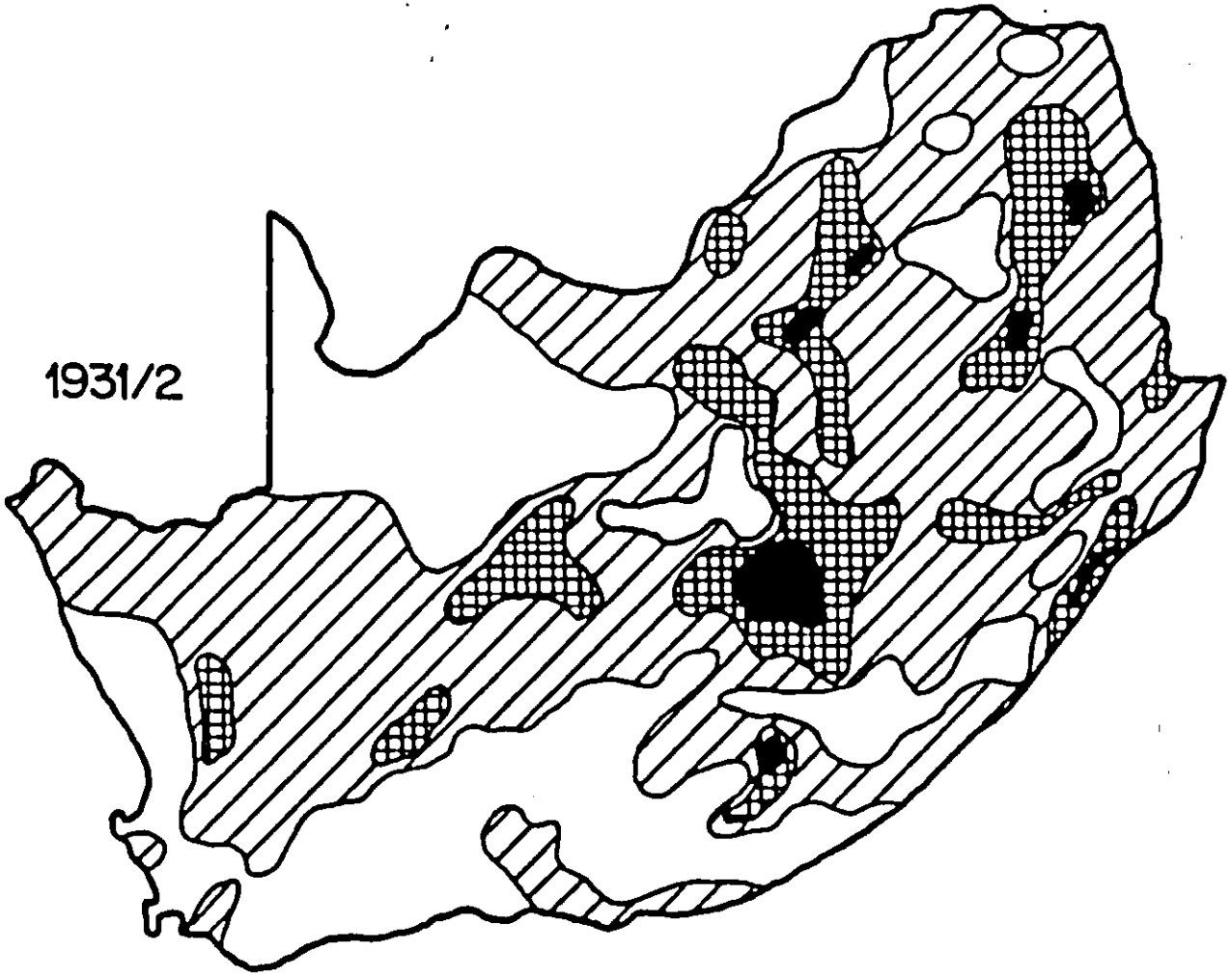

1926/7

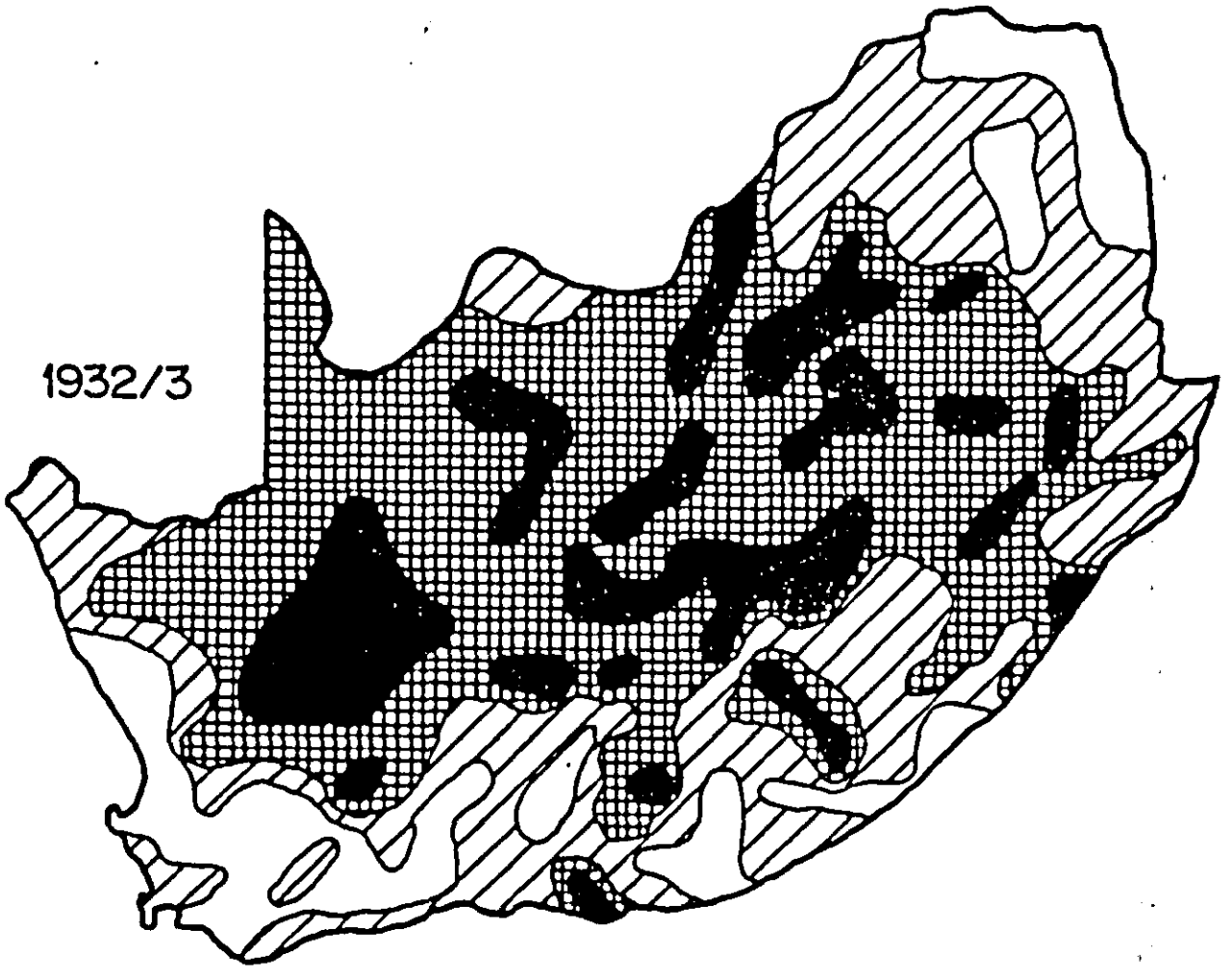

1927/8

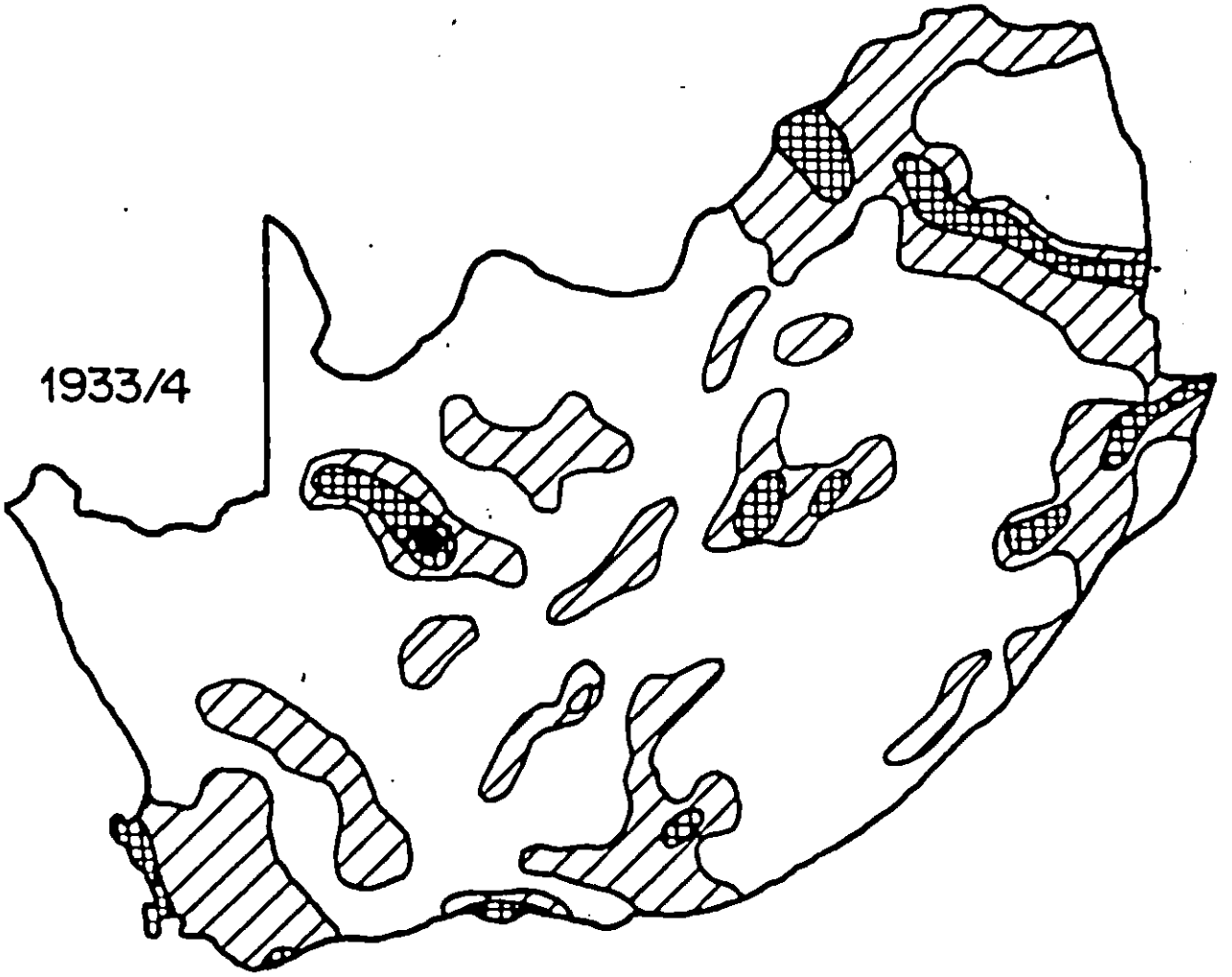

1928/9

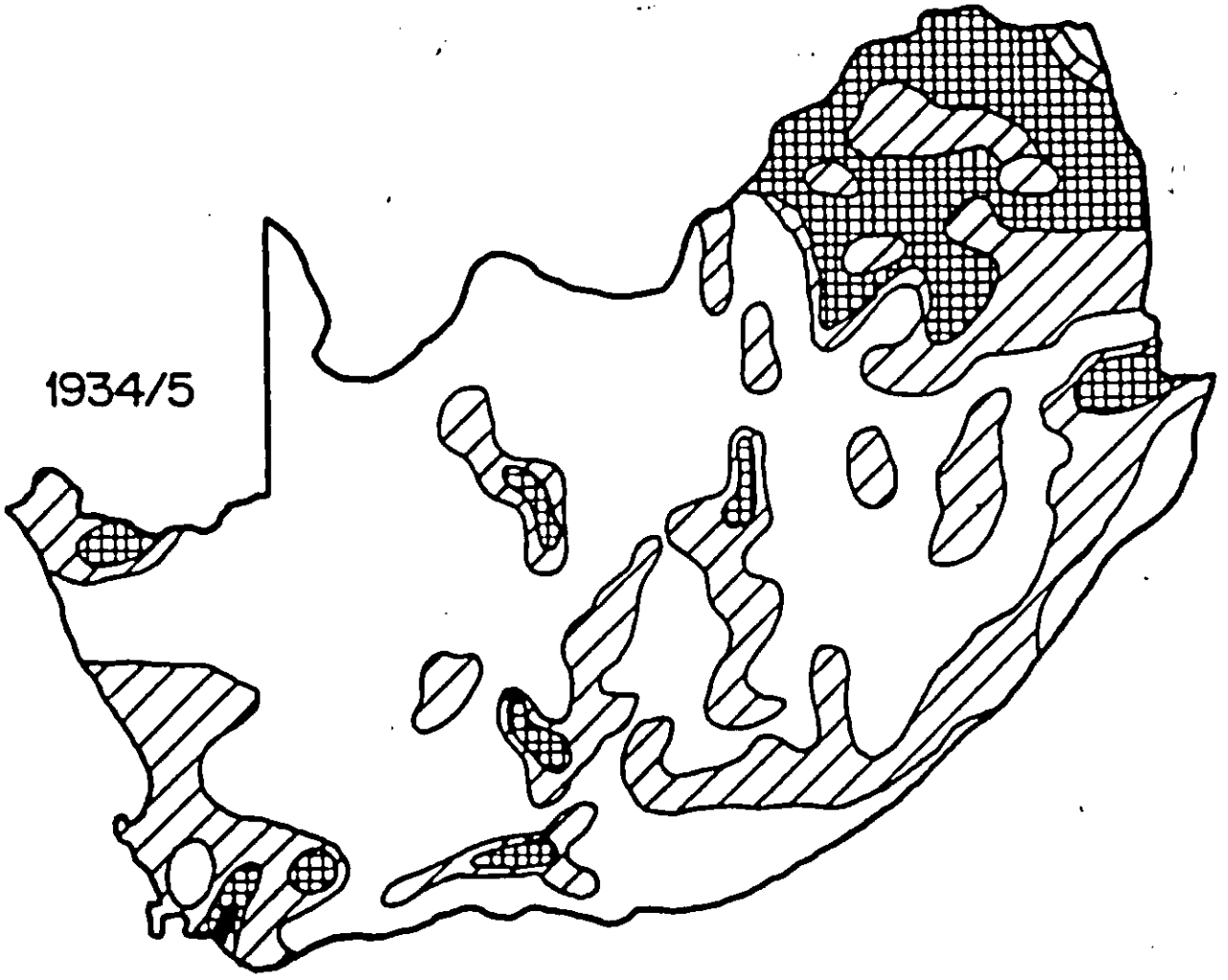

1929/30

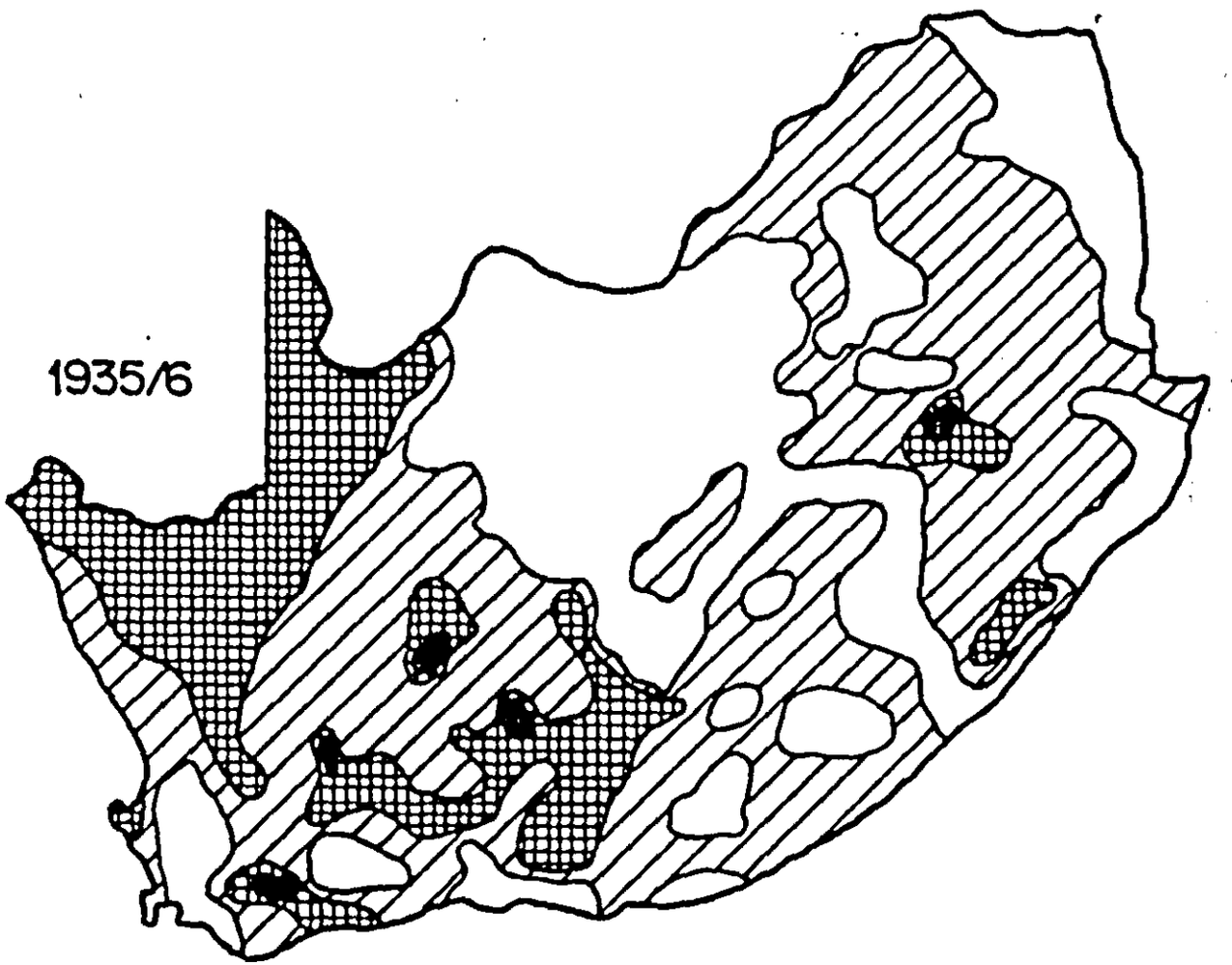

1930/1

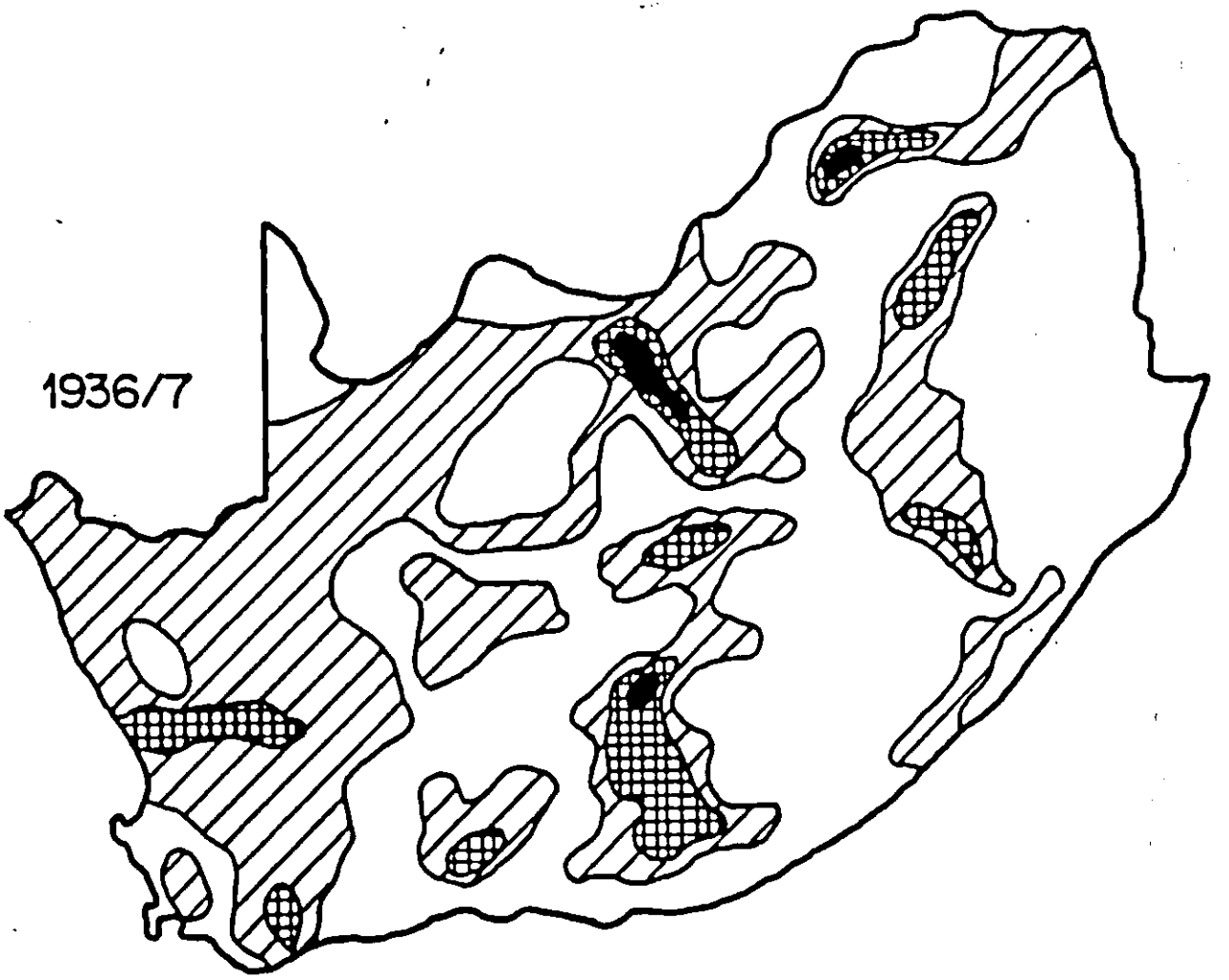

1931/2

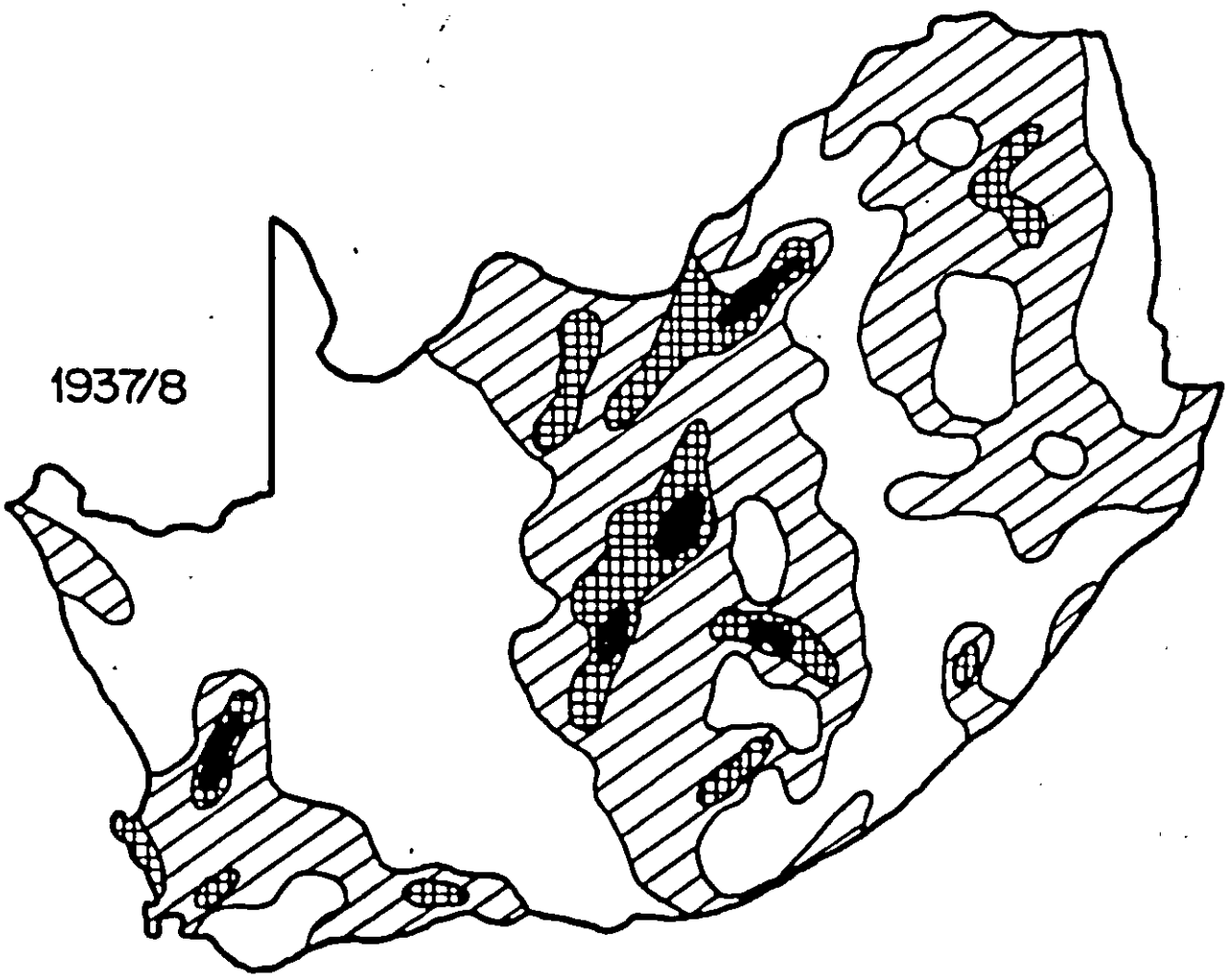

1932/3

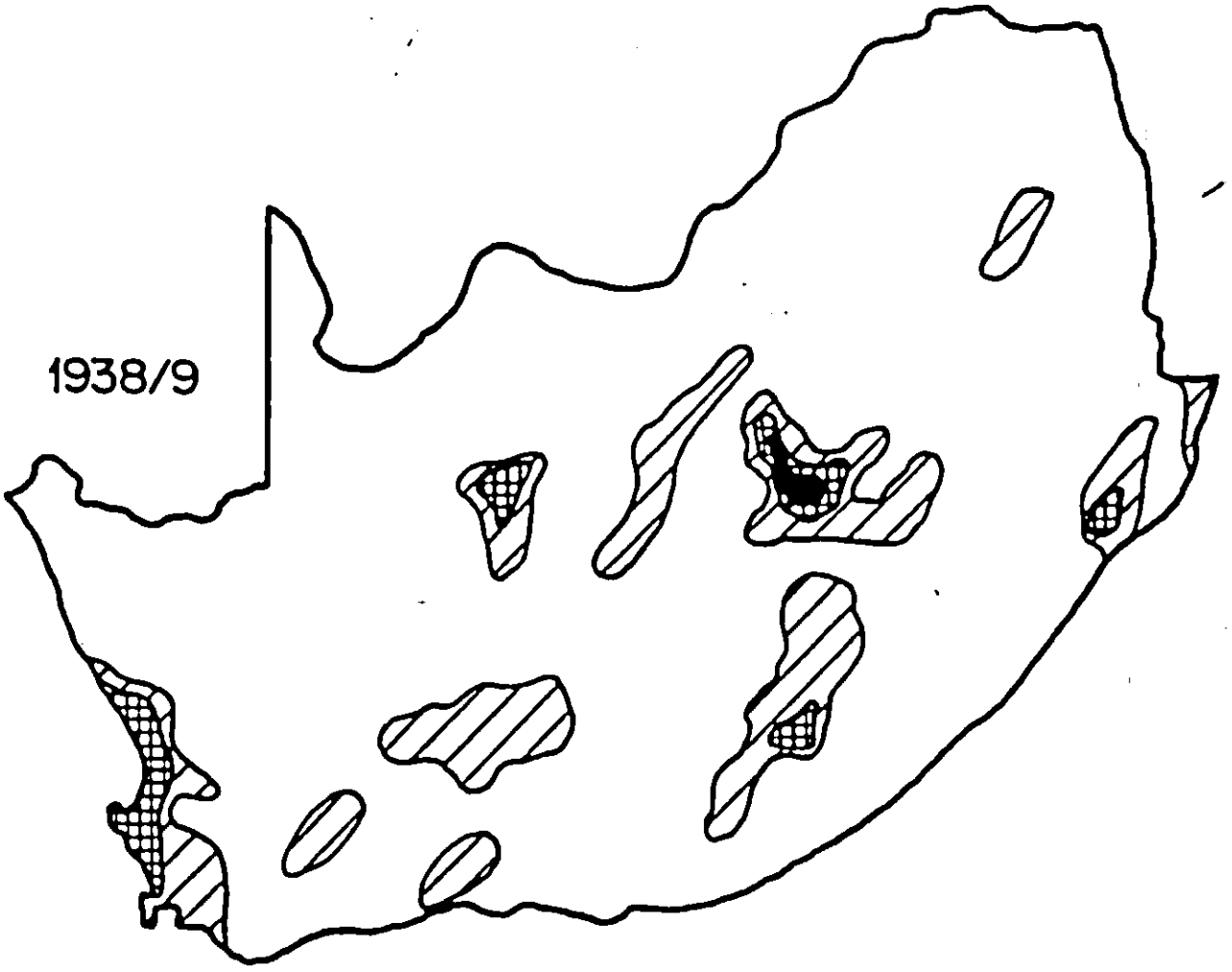

1933/4

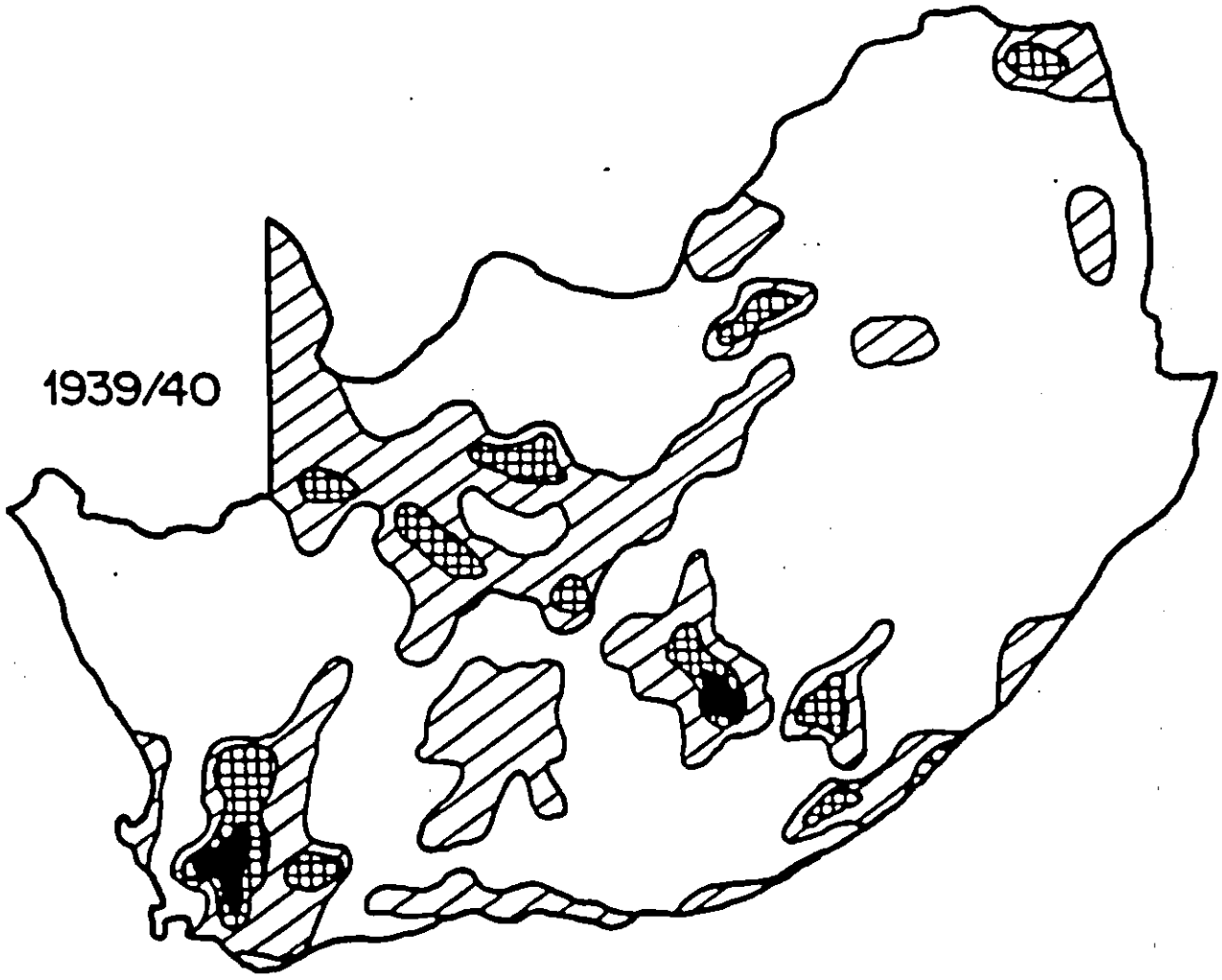

1934/5

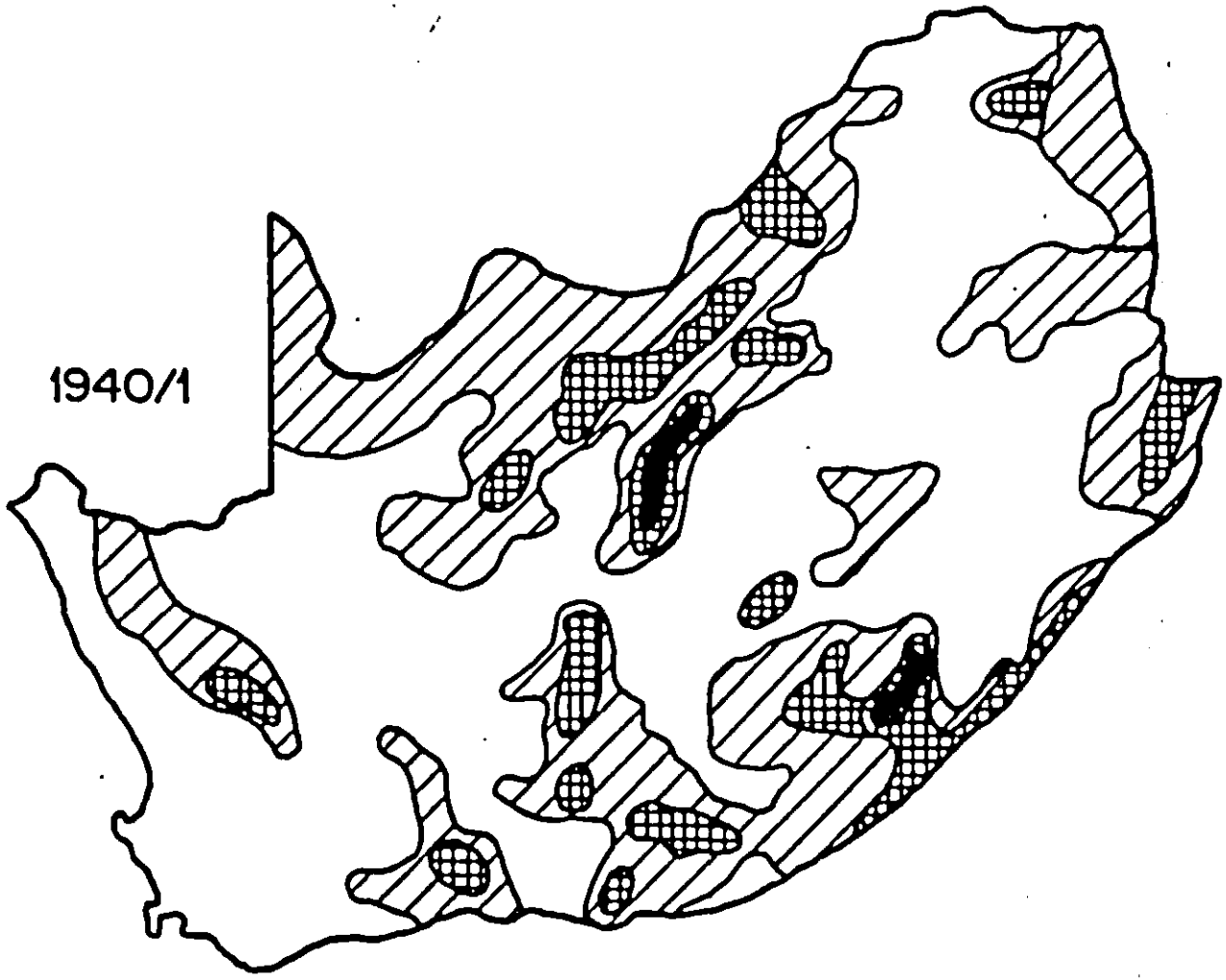

1935/6

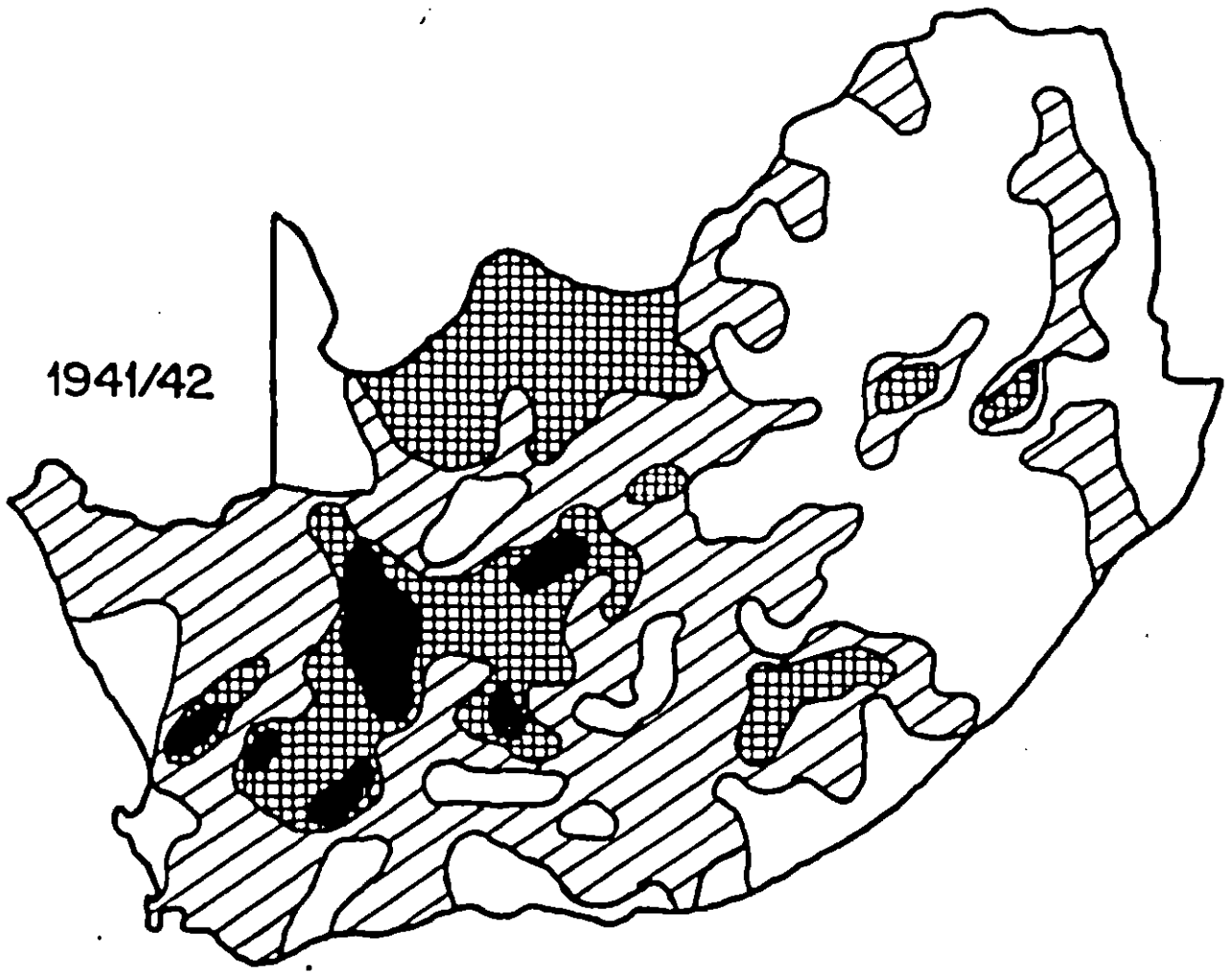

1936/7

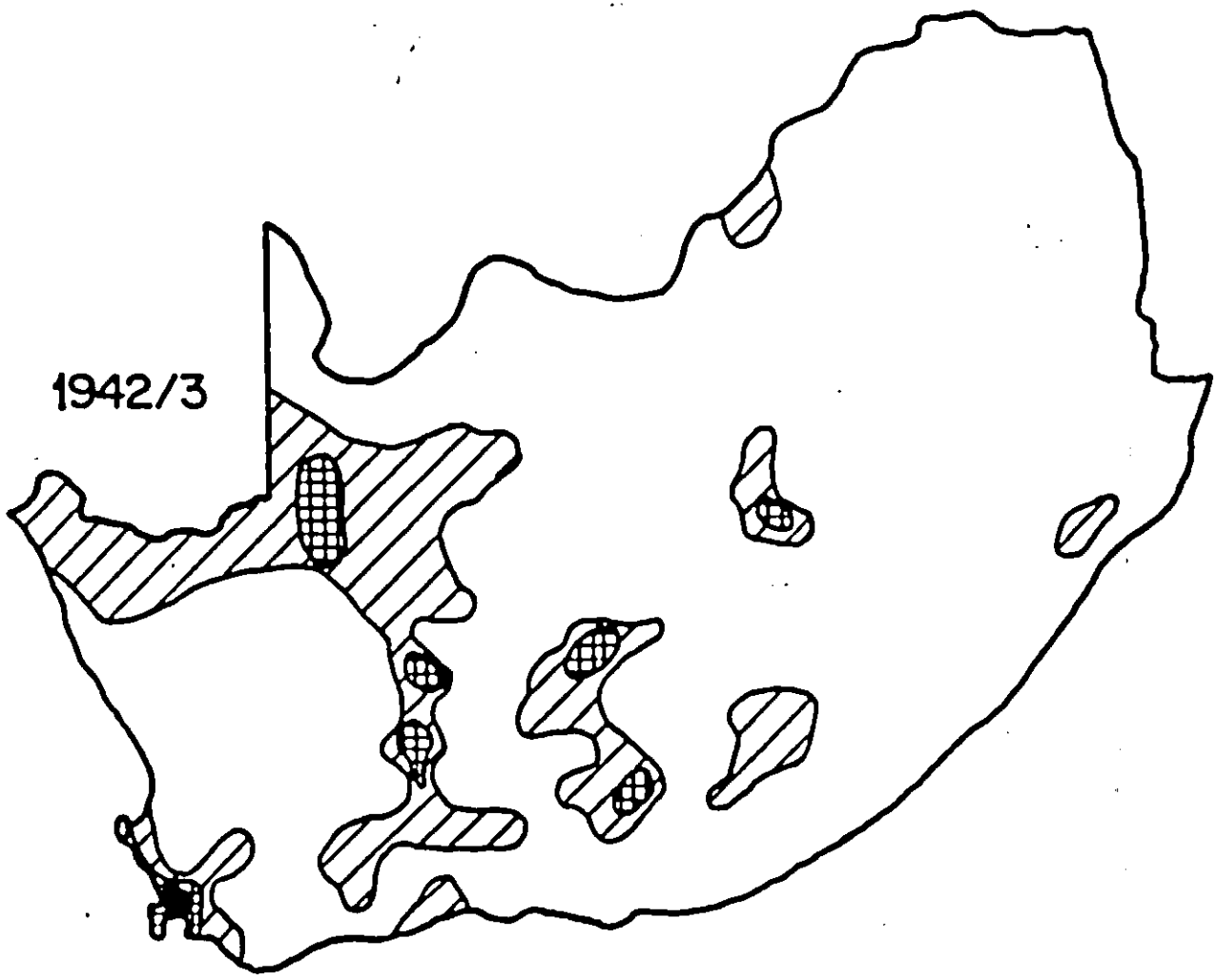

1937/8

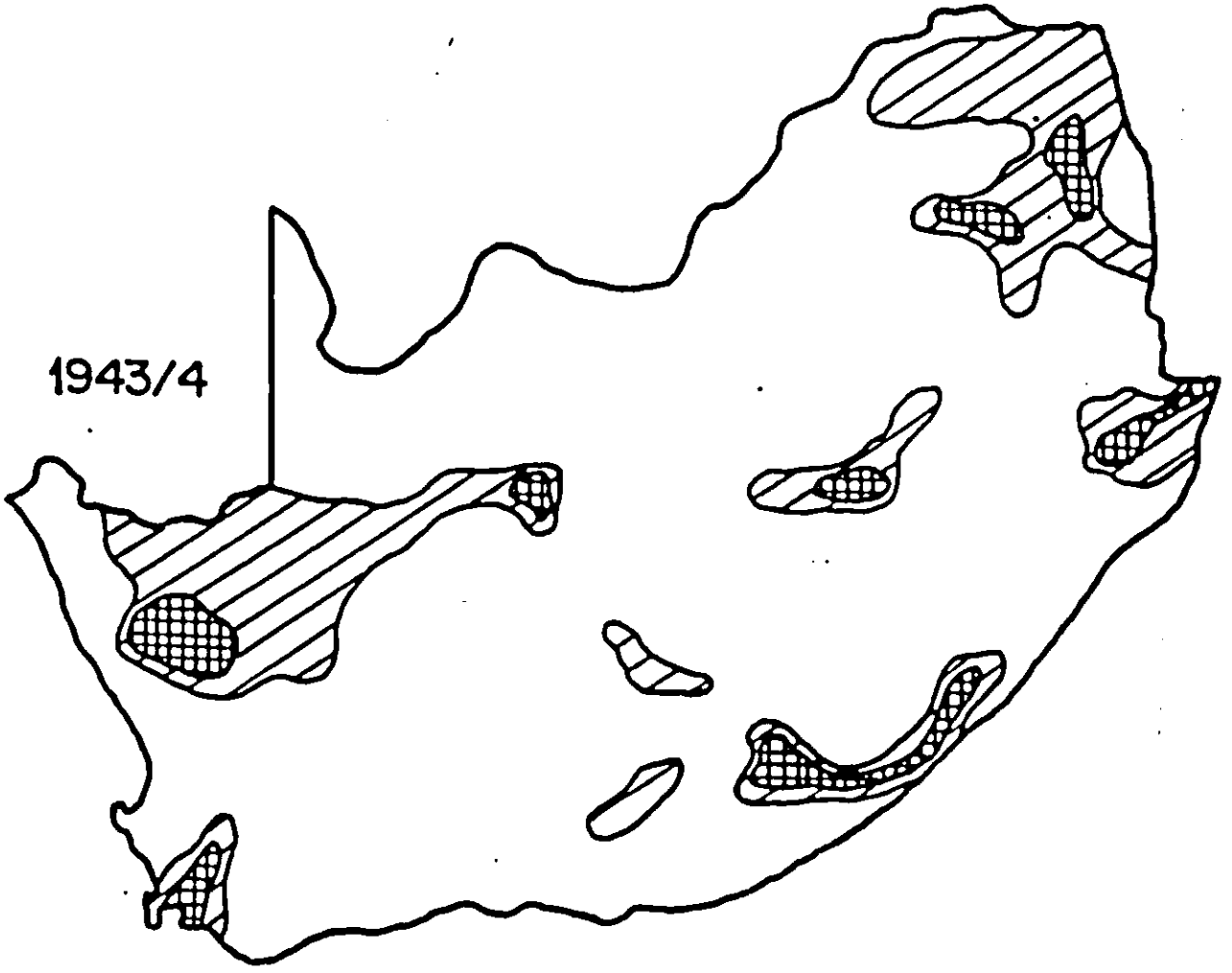

1938/9

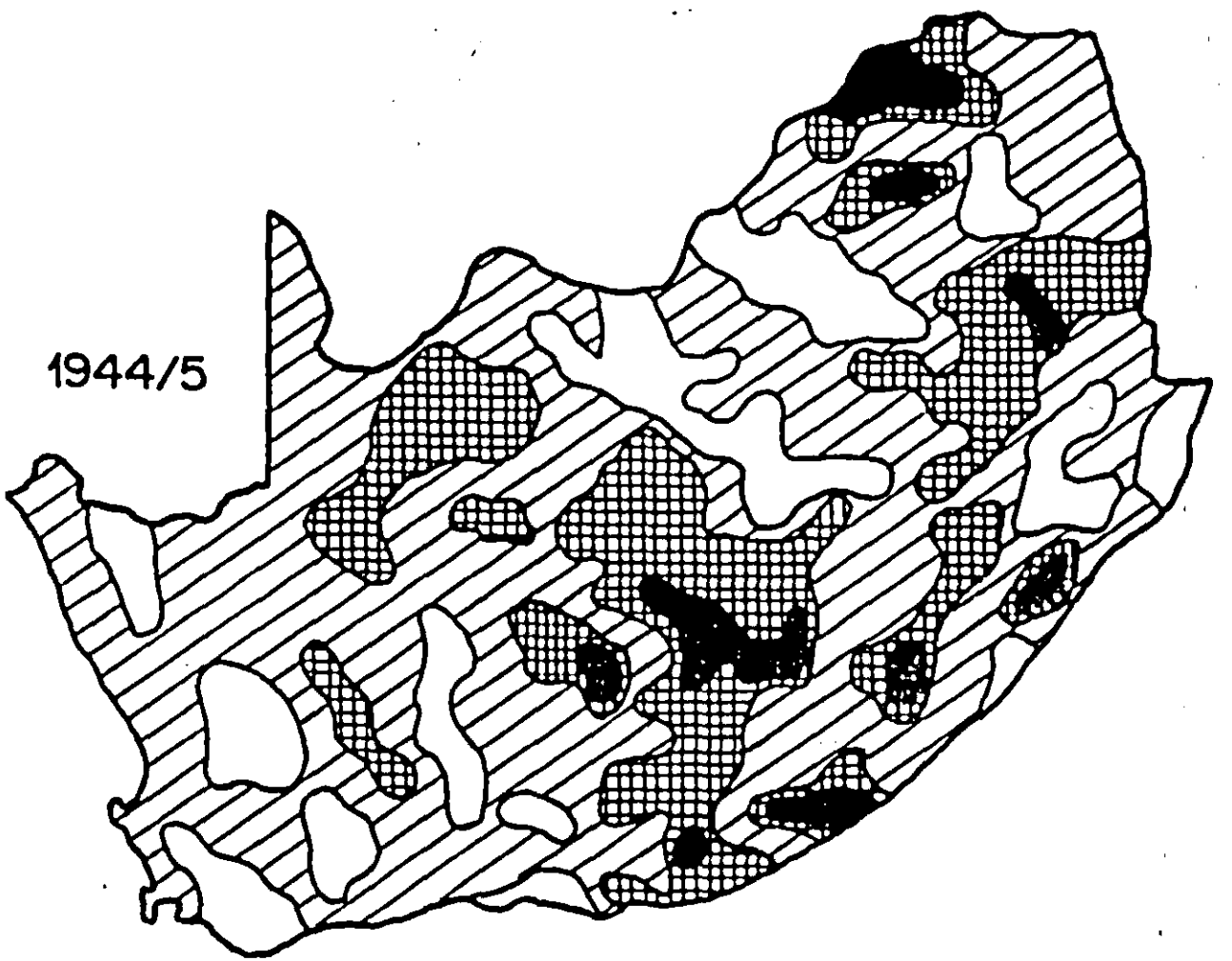

1939/40

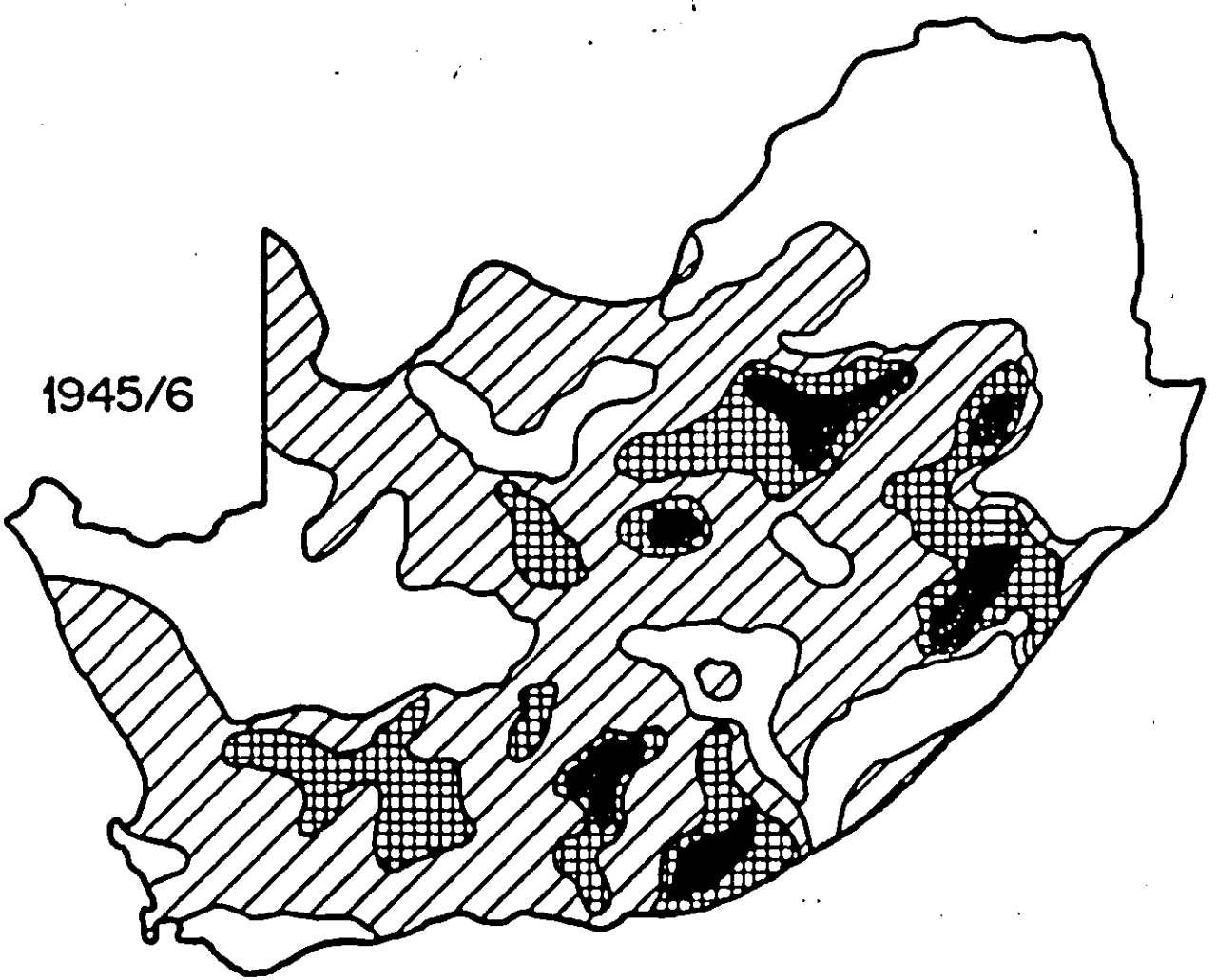

1940/1

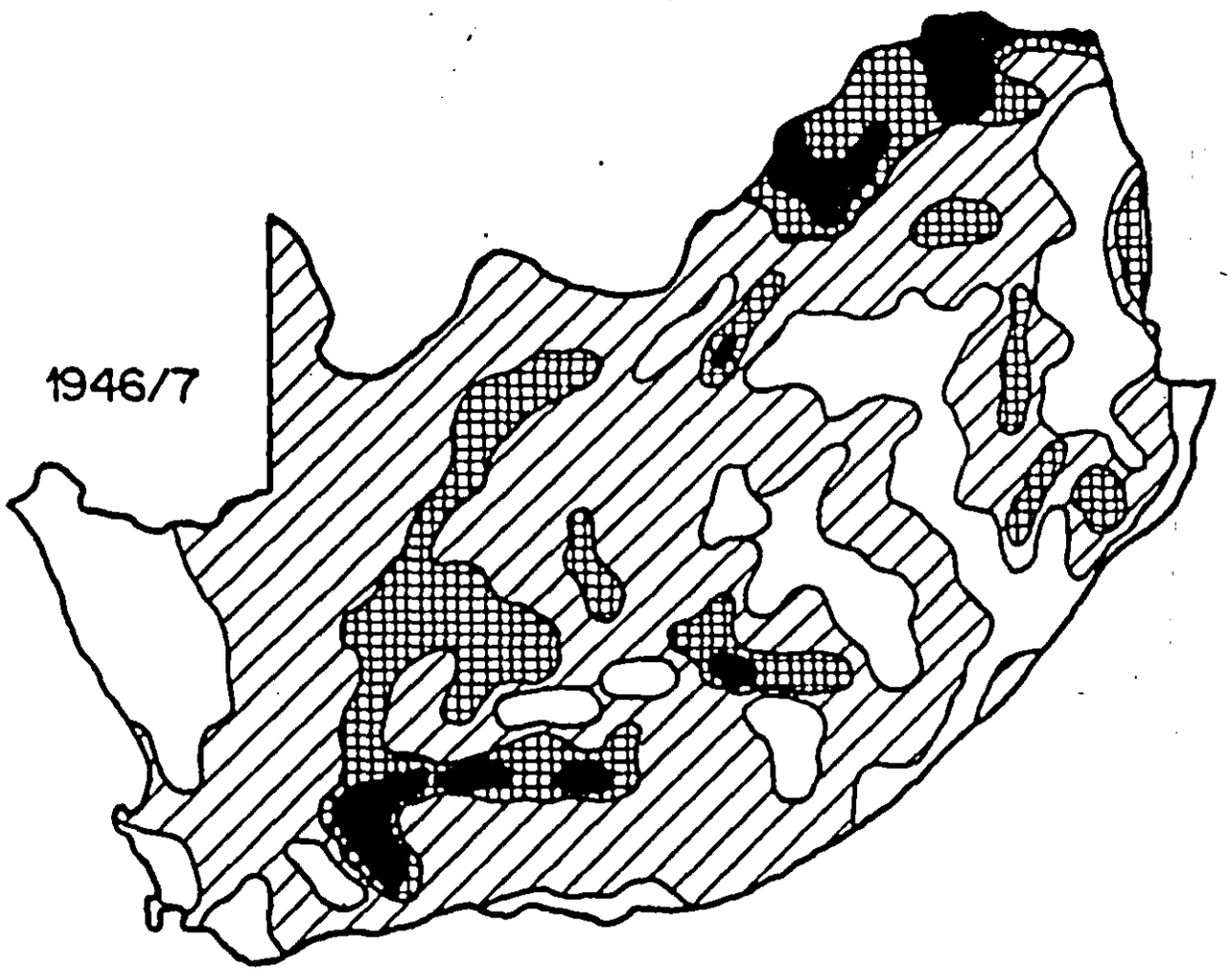

1941/42

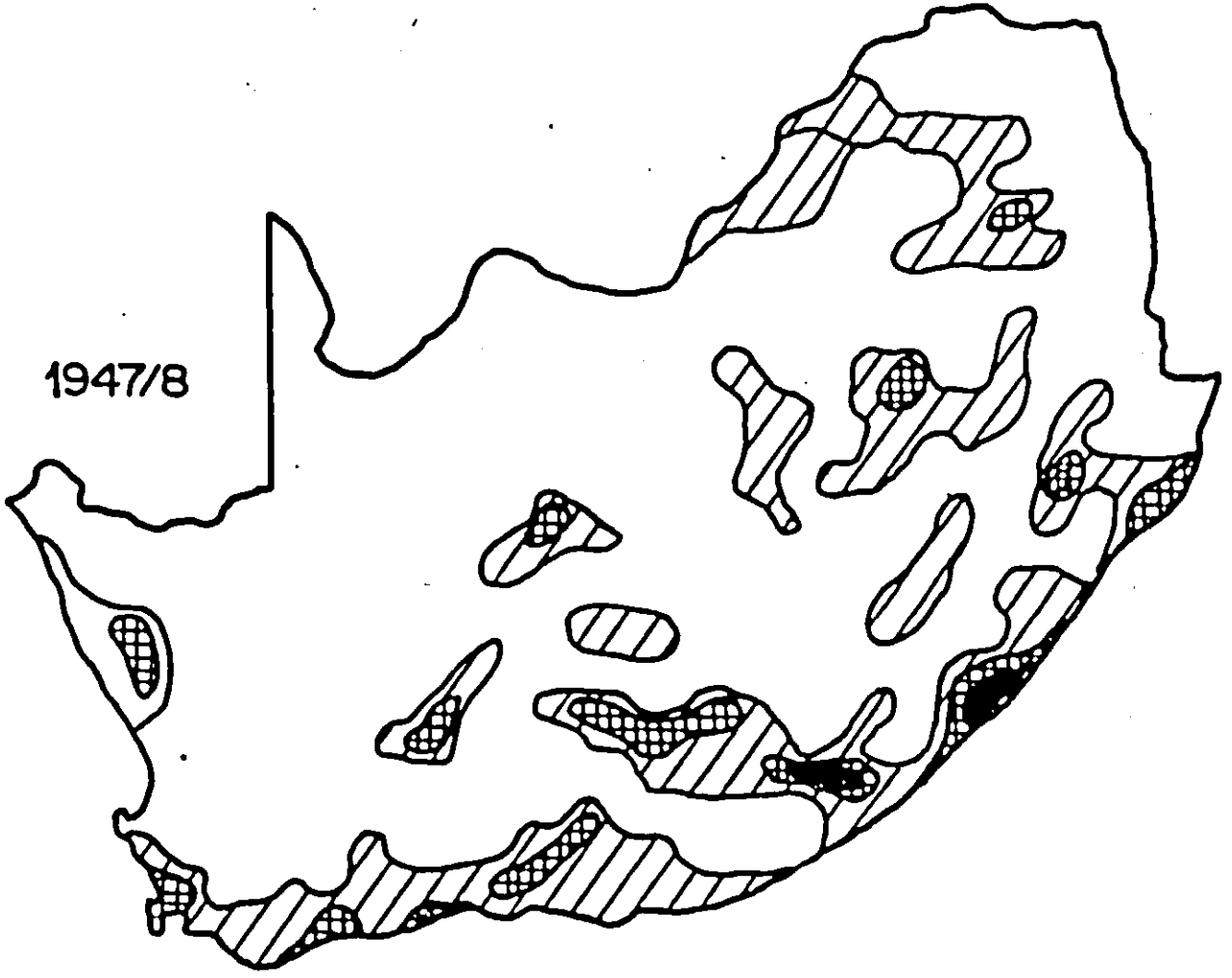

1942/3

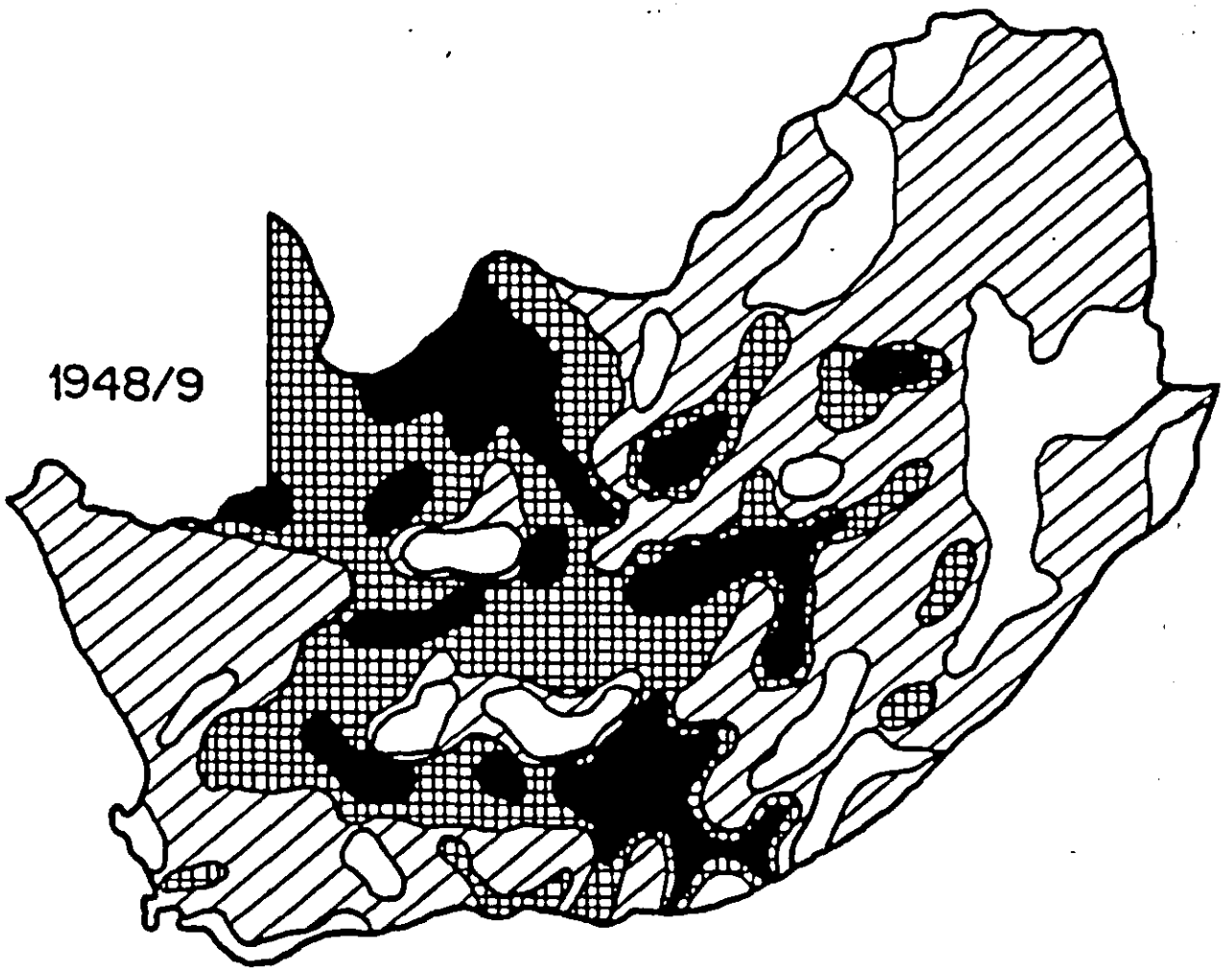

1943/4

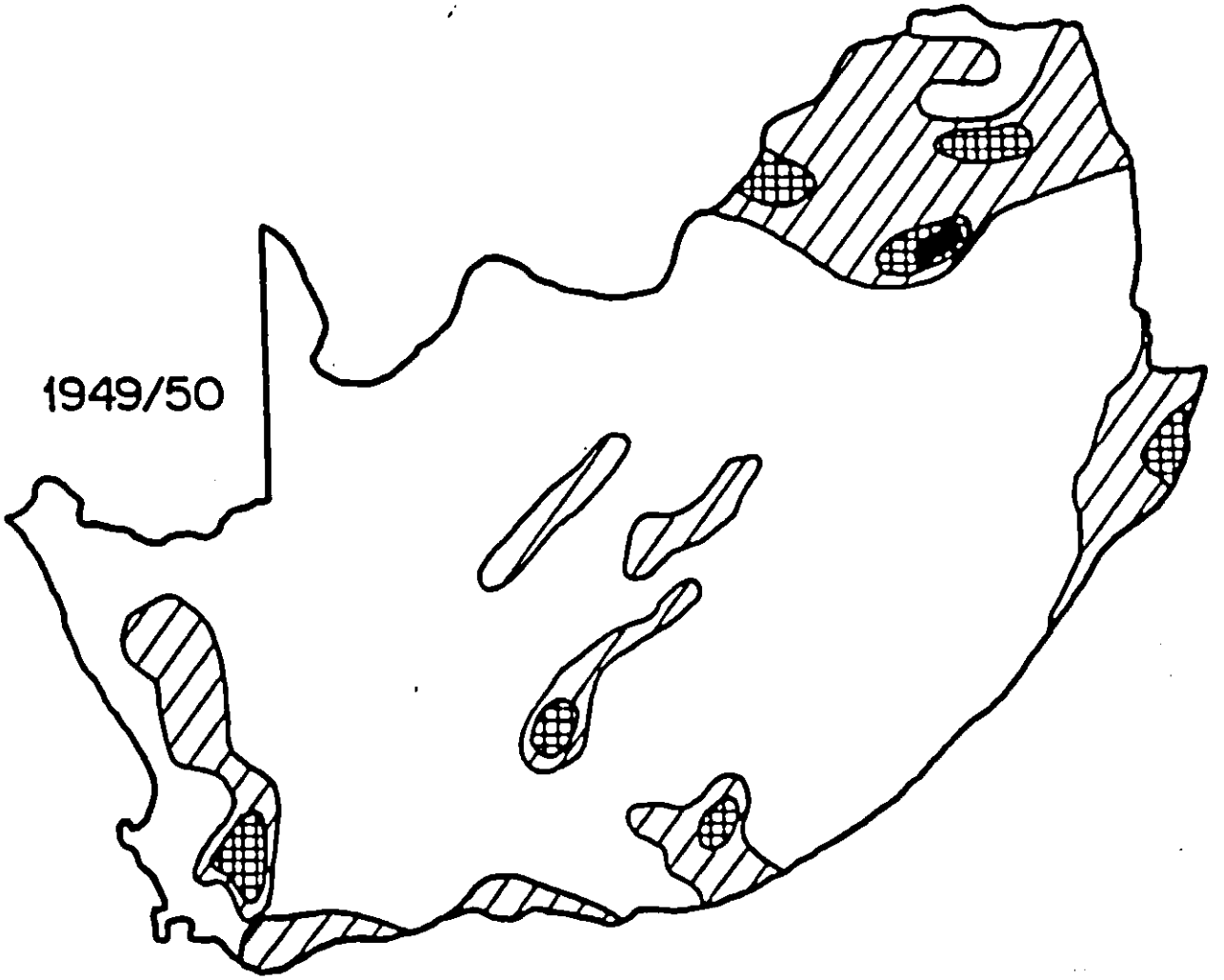

1944/5

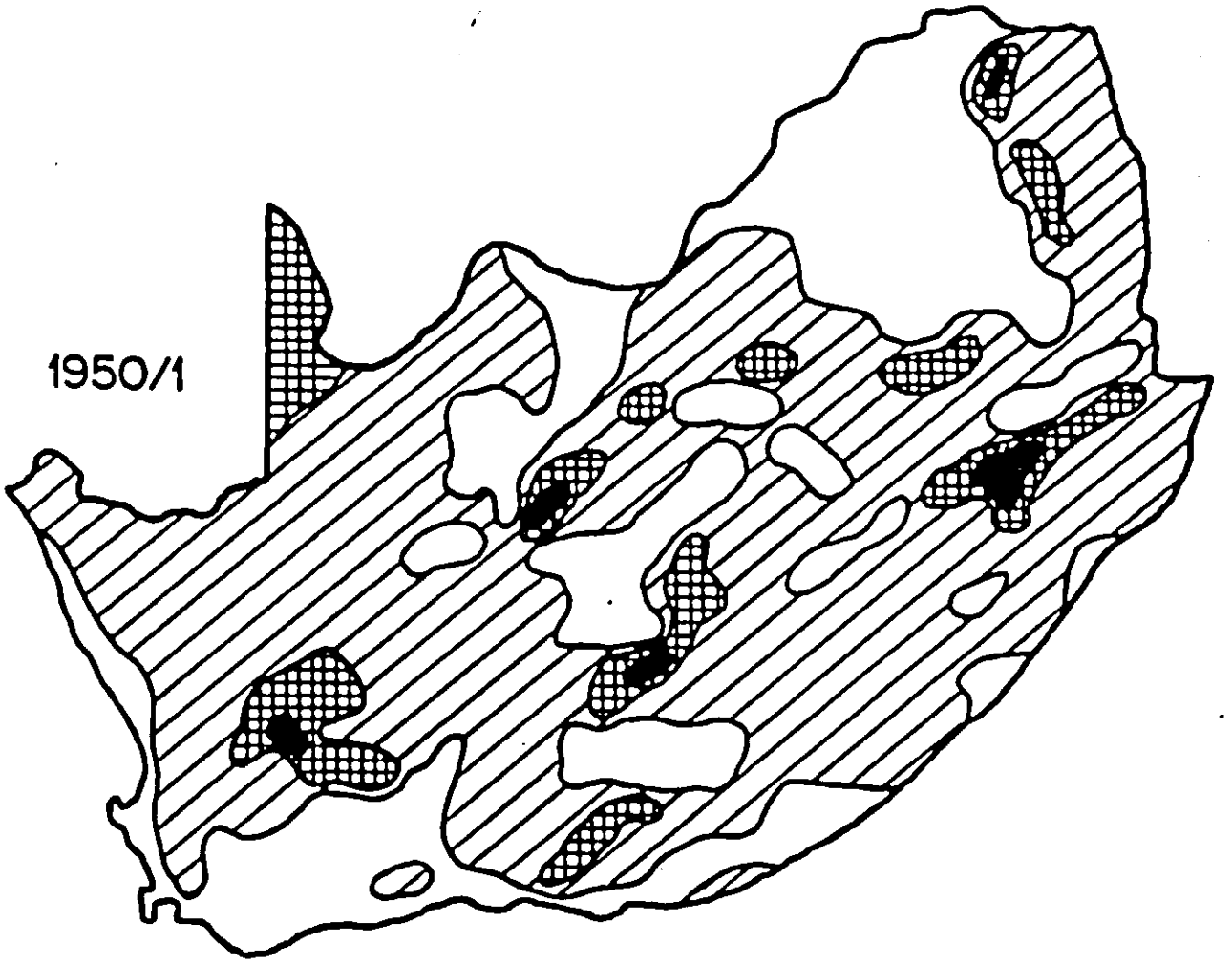

1945/6

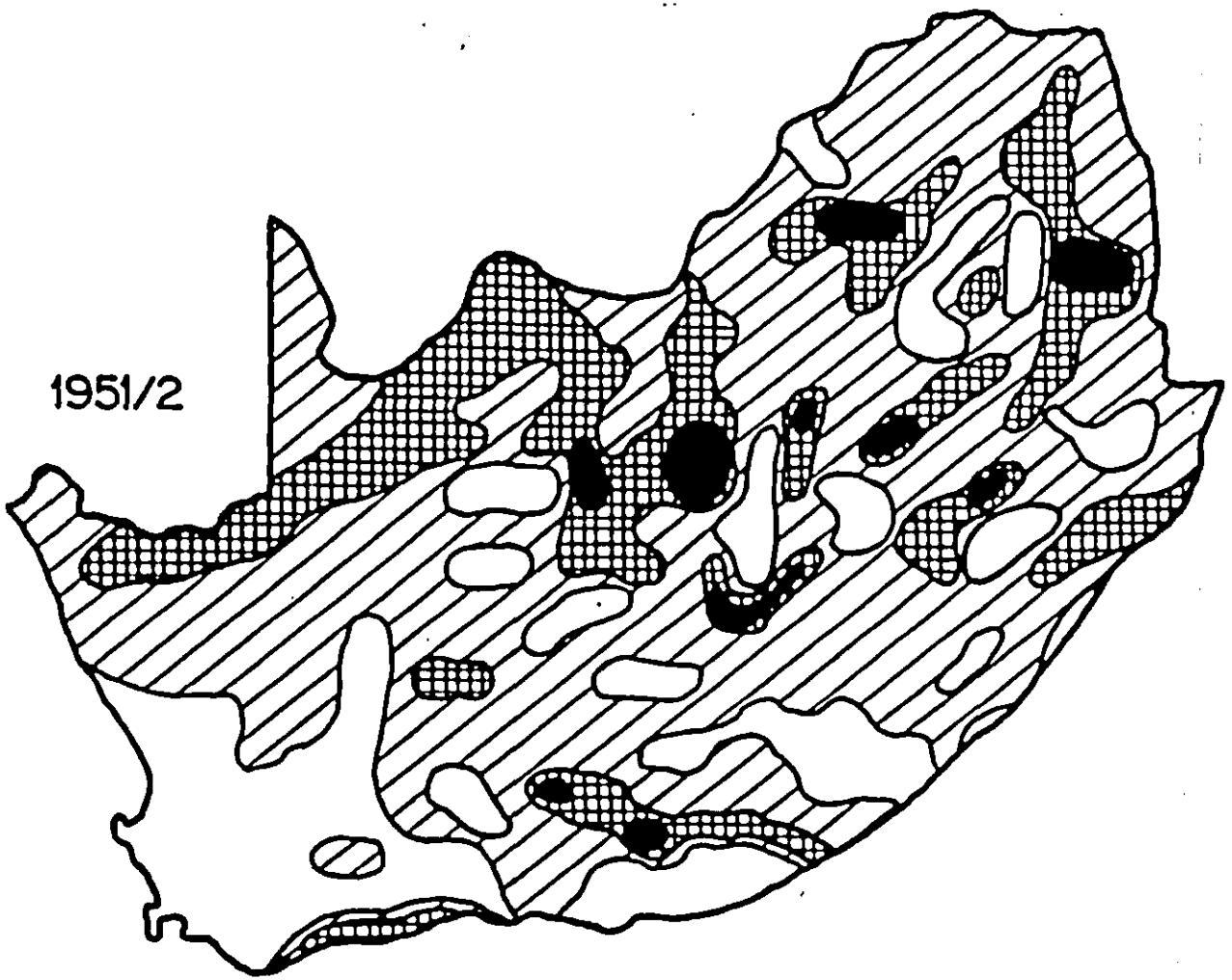

1946/7

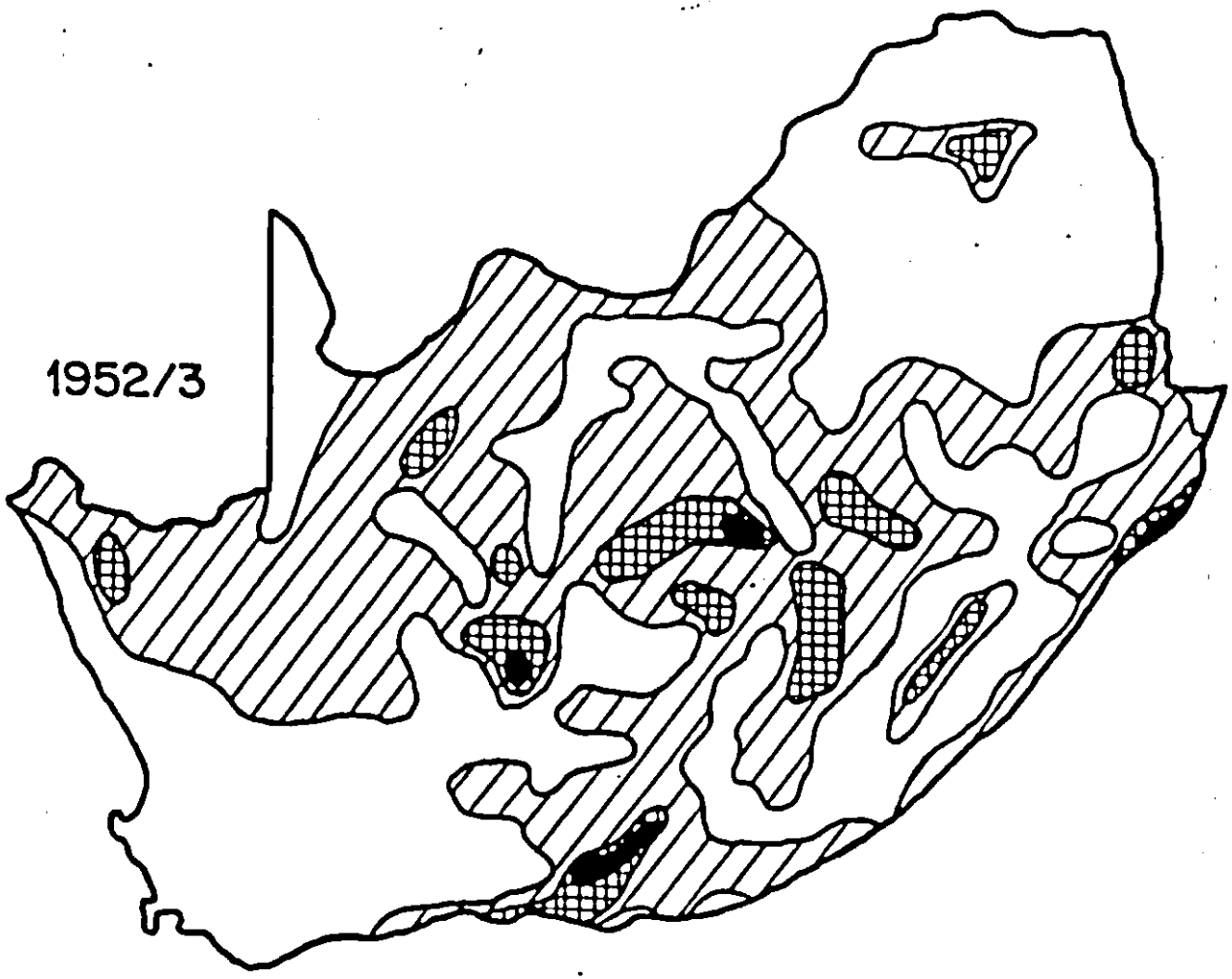

1947/8

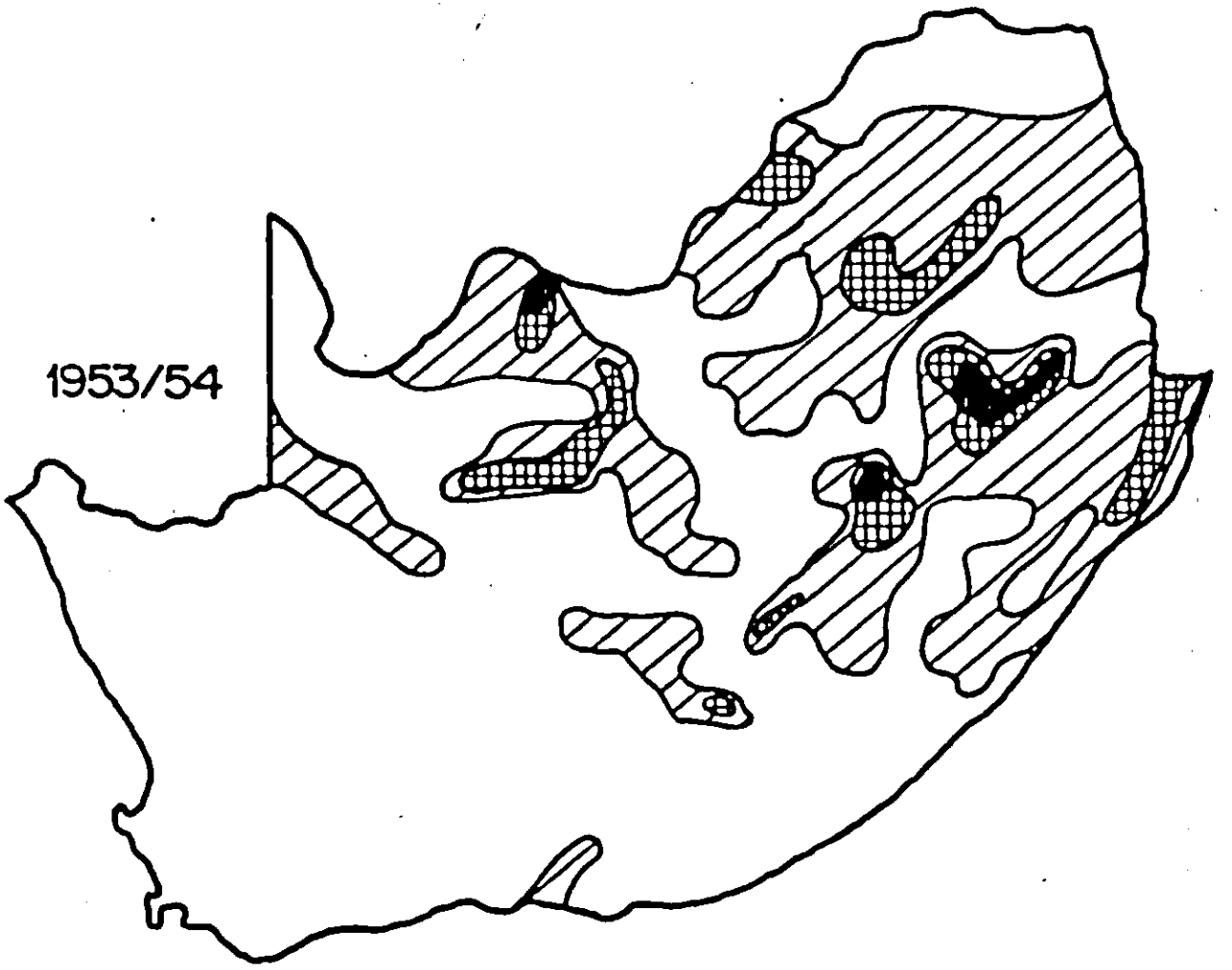

1948/9

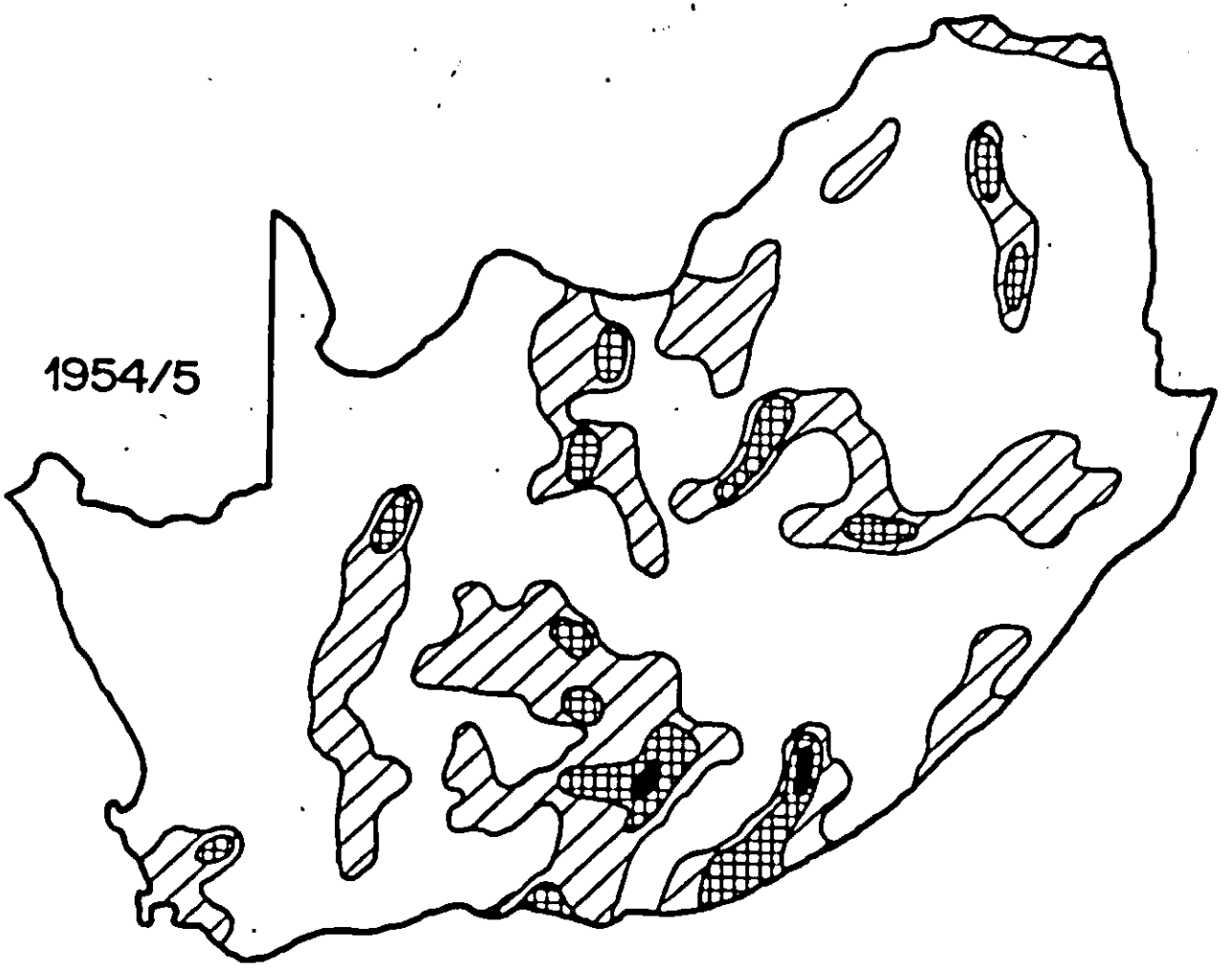

1949/50

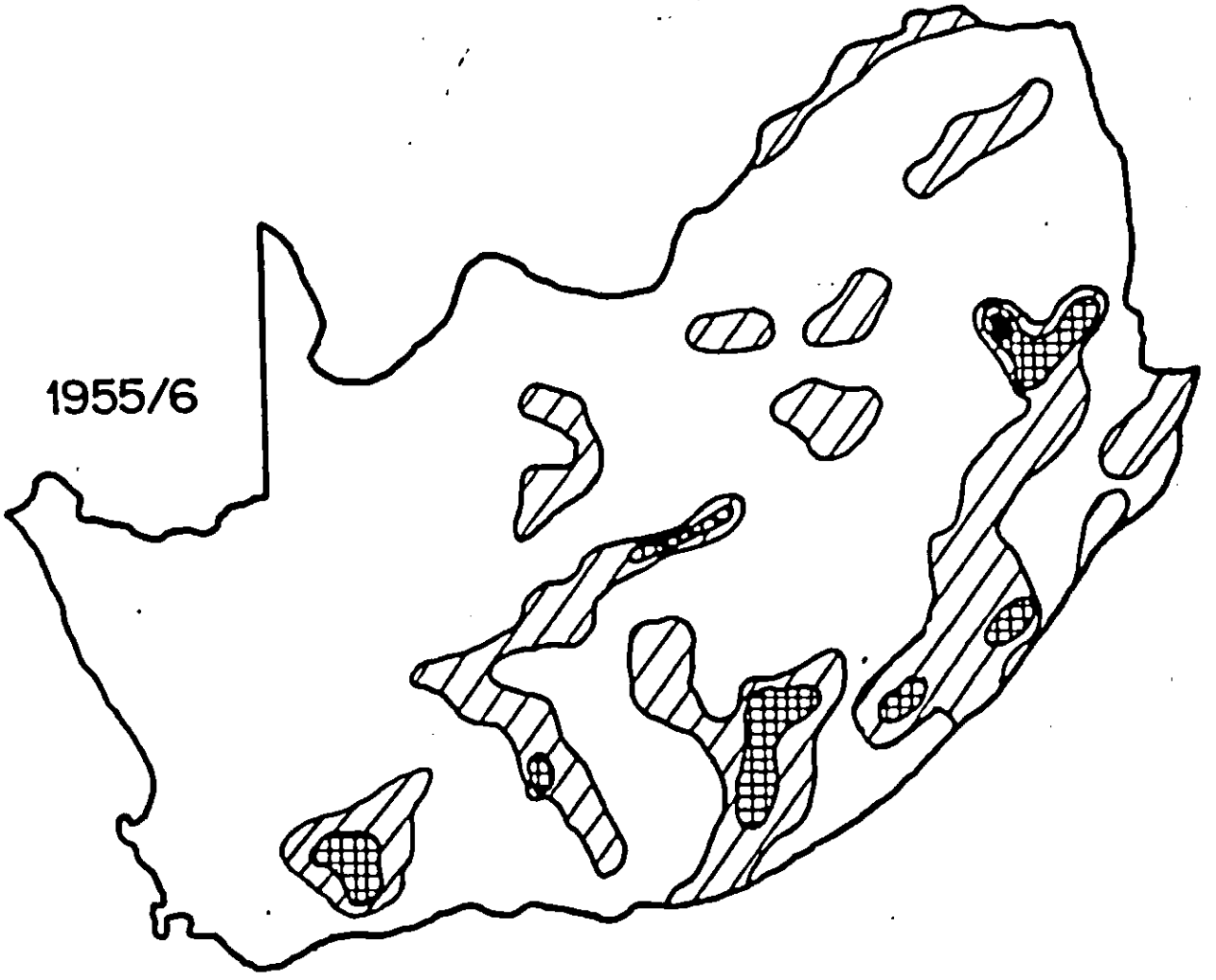

1950/1

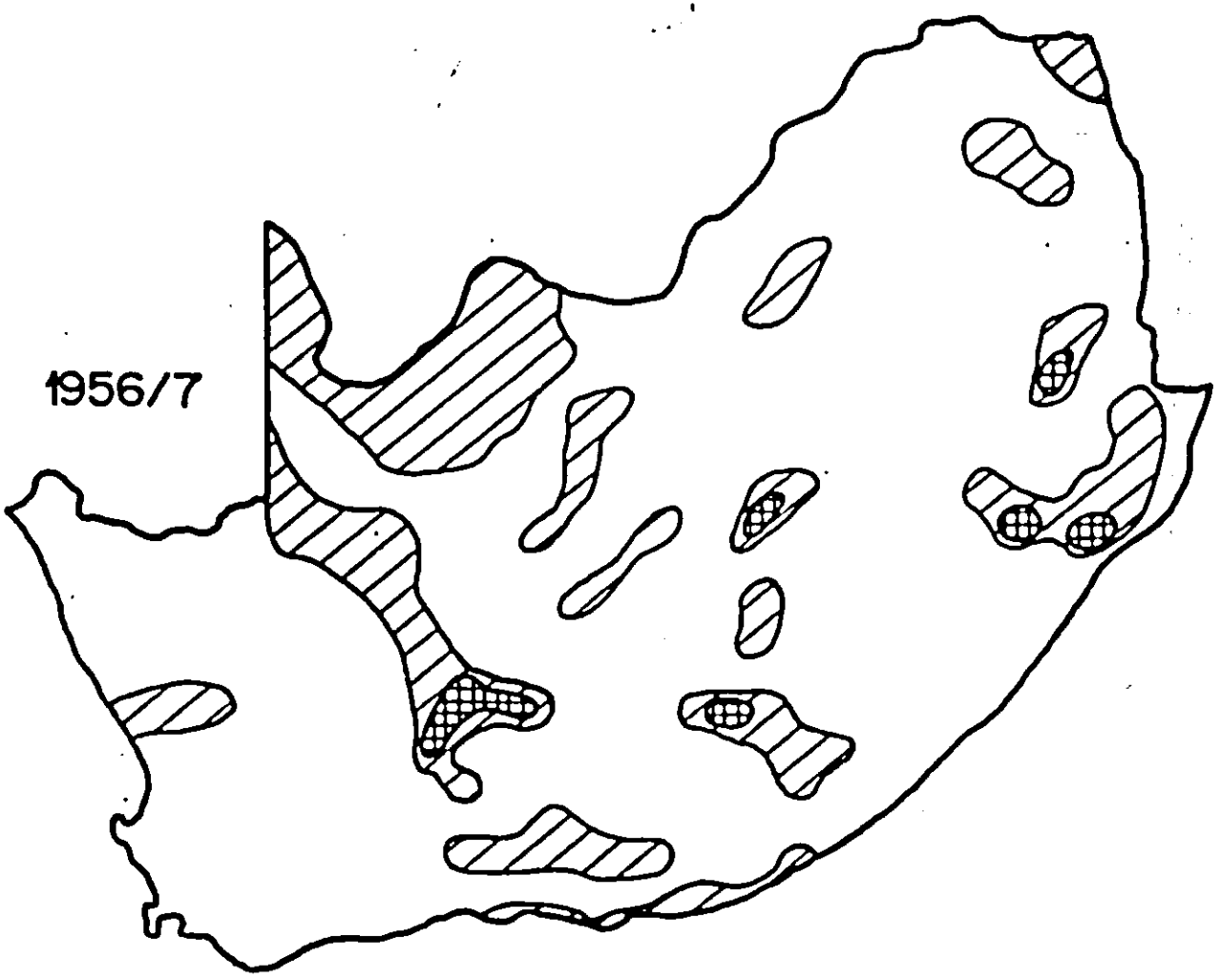

1951/2

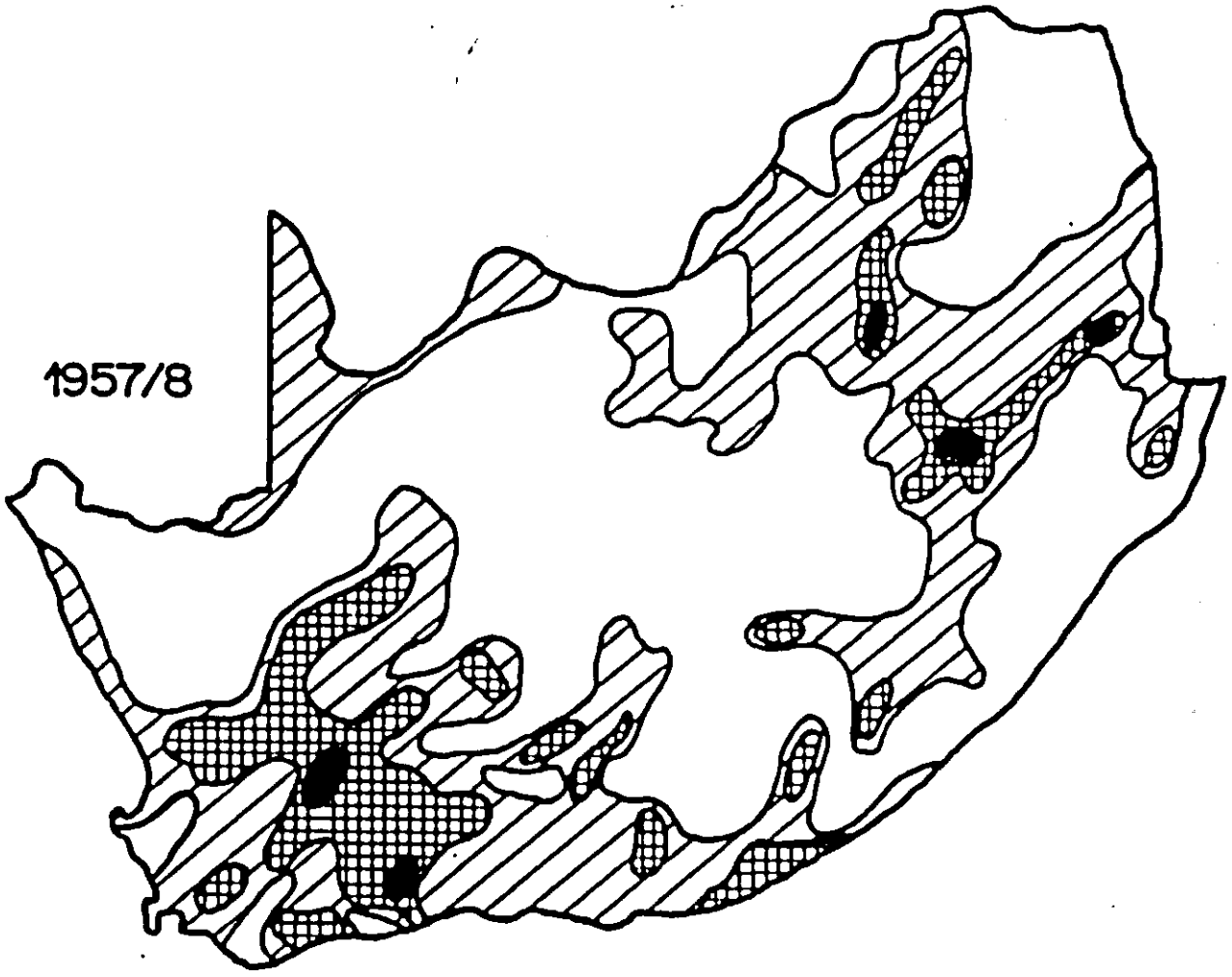

1952/3

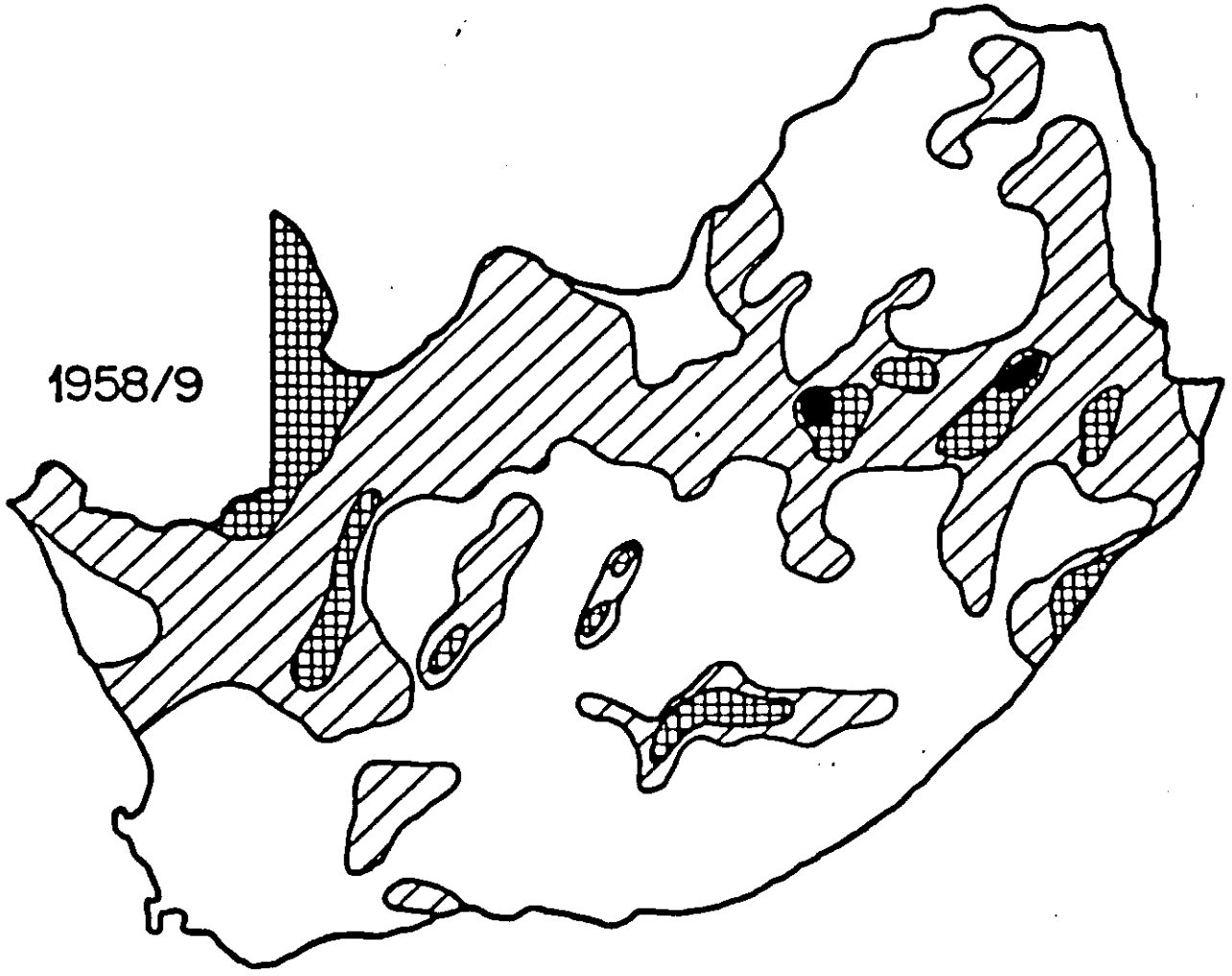

1953/54

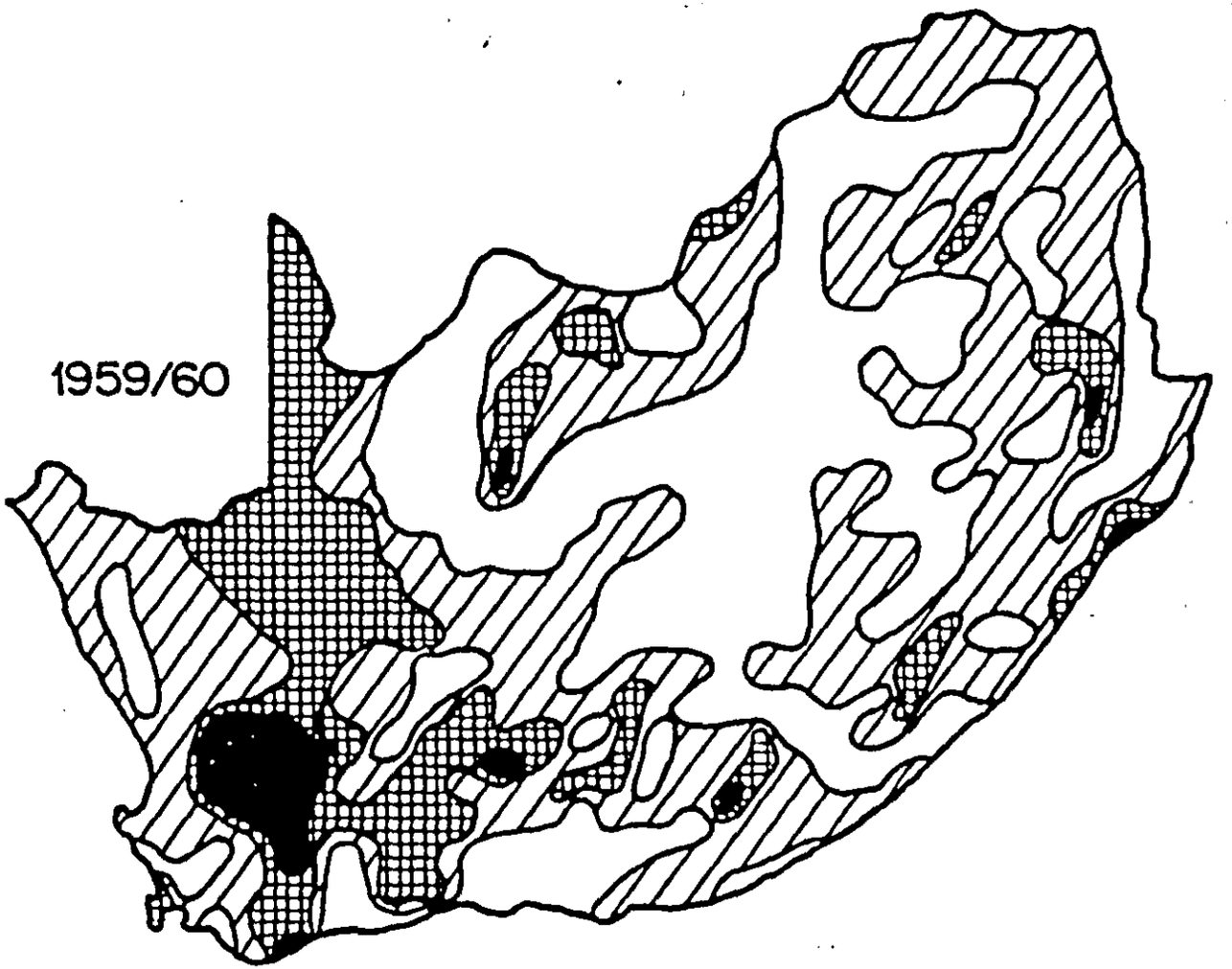

1954/5

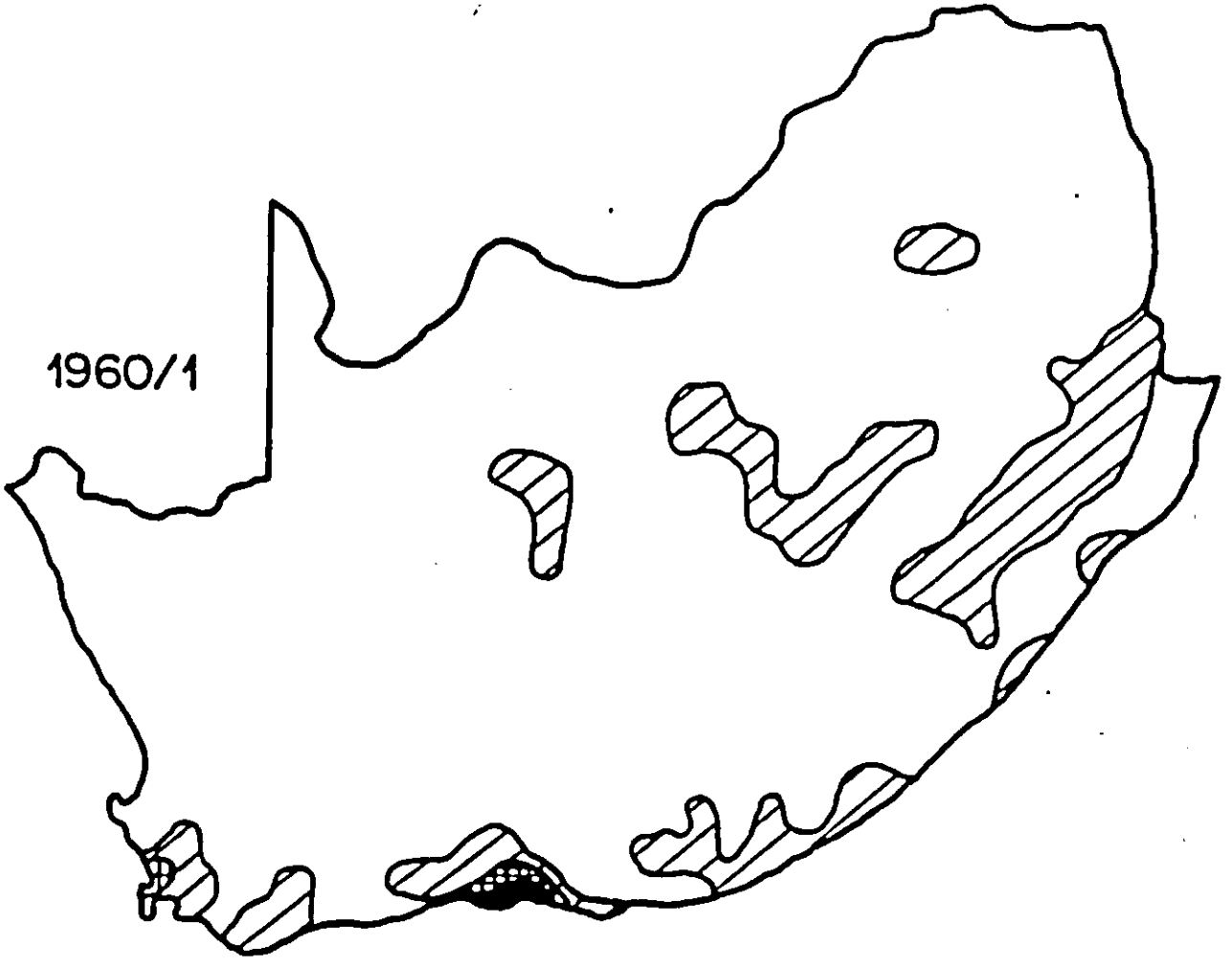

1955/6

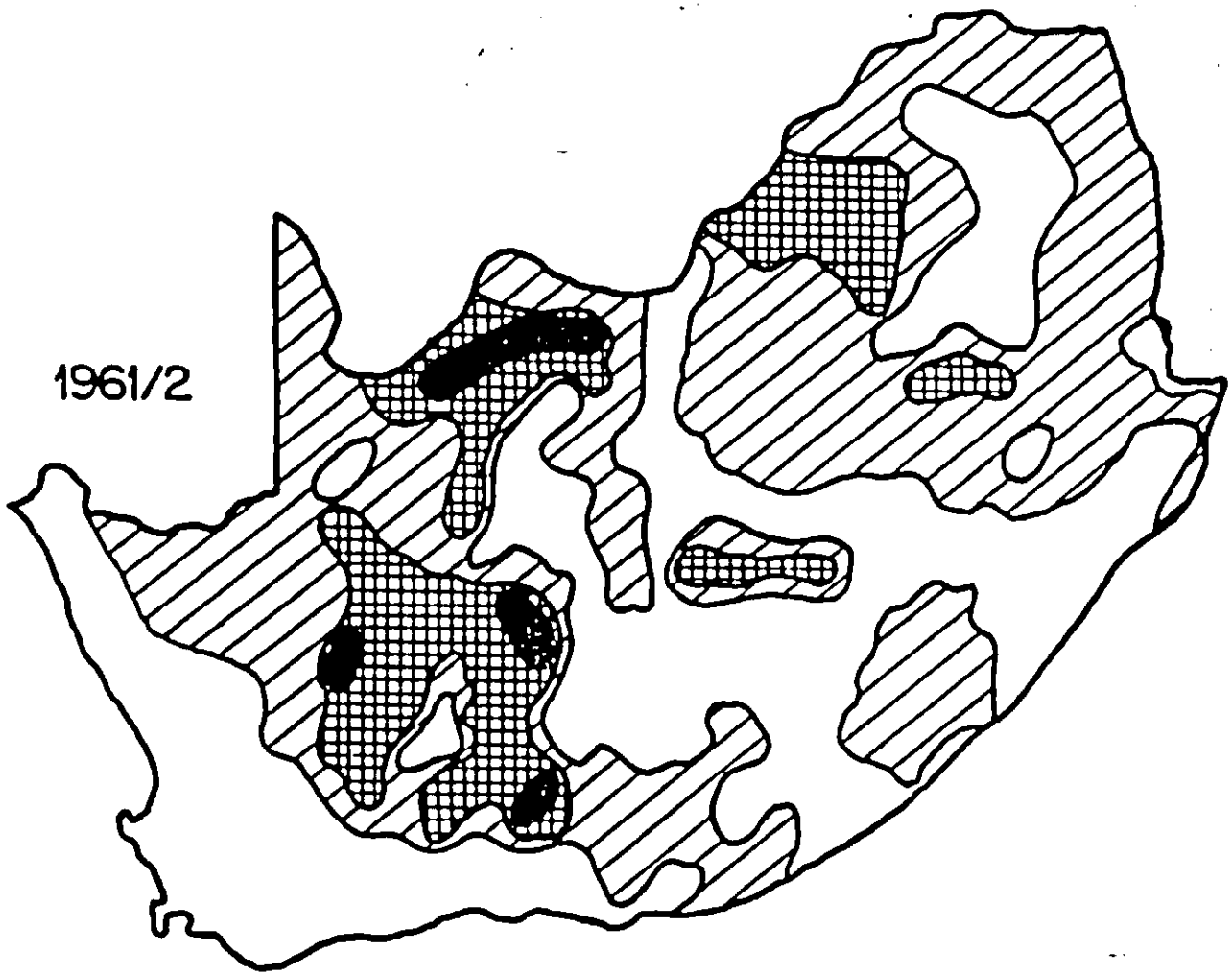

1956/7

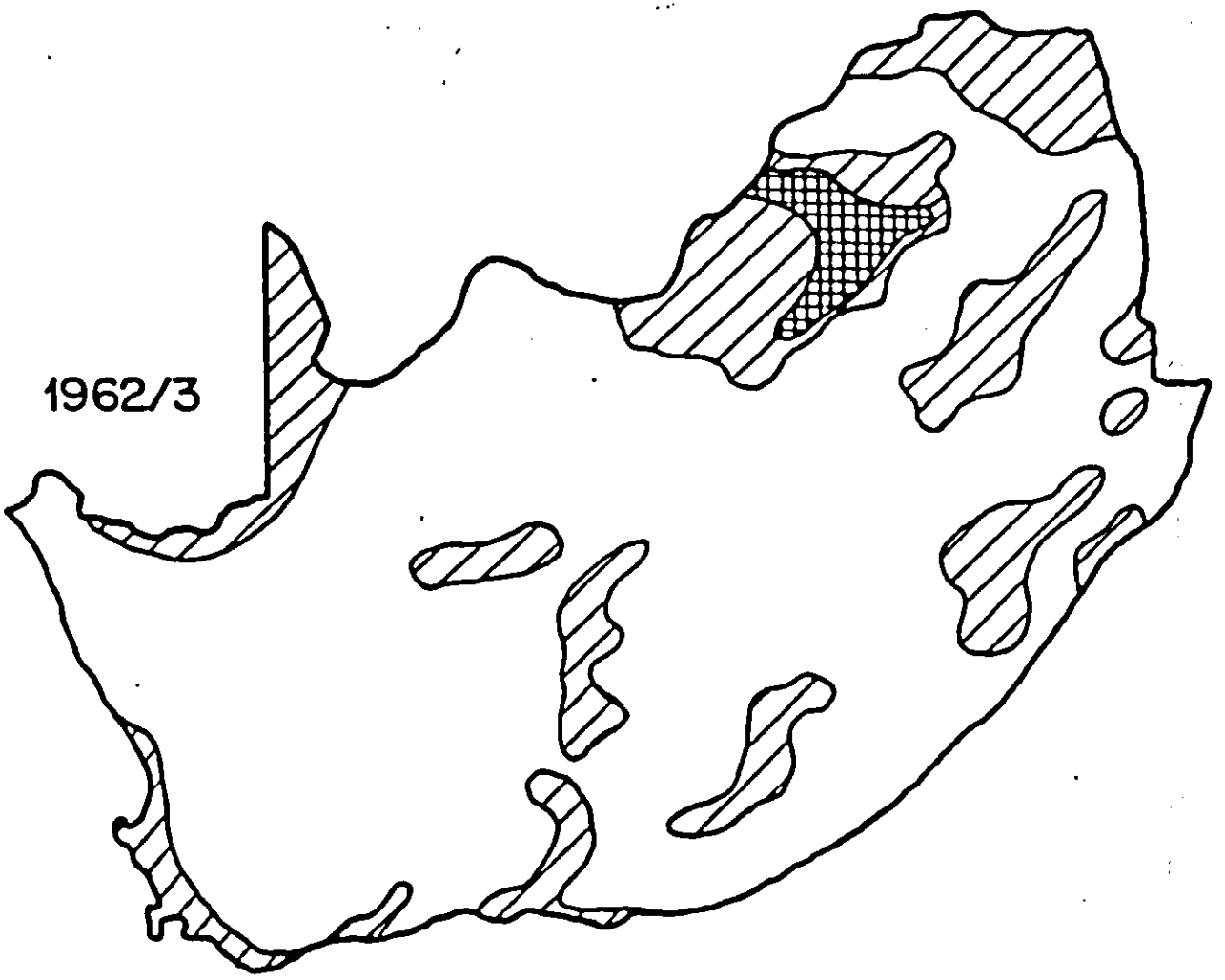

1957/8

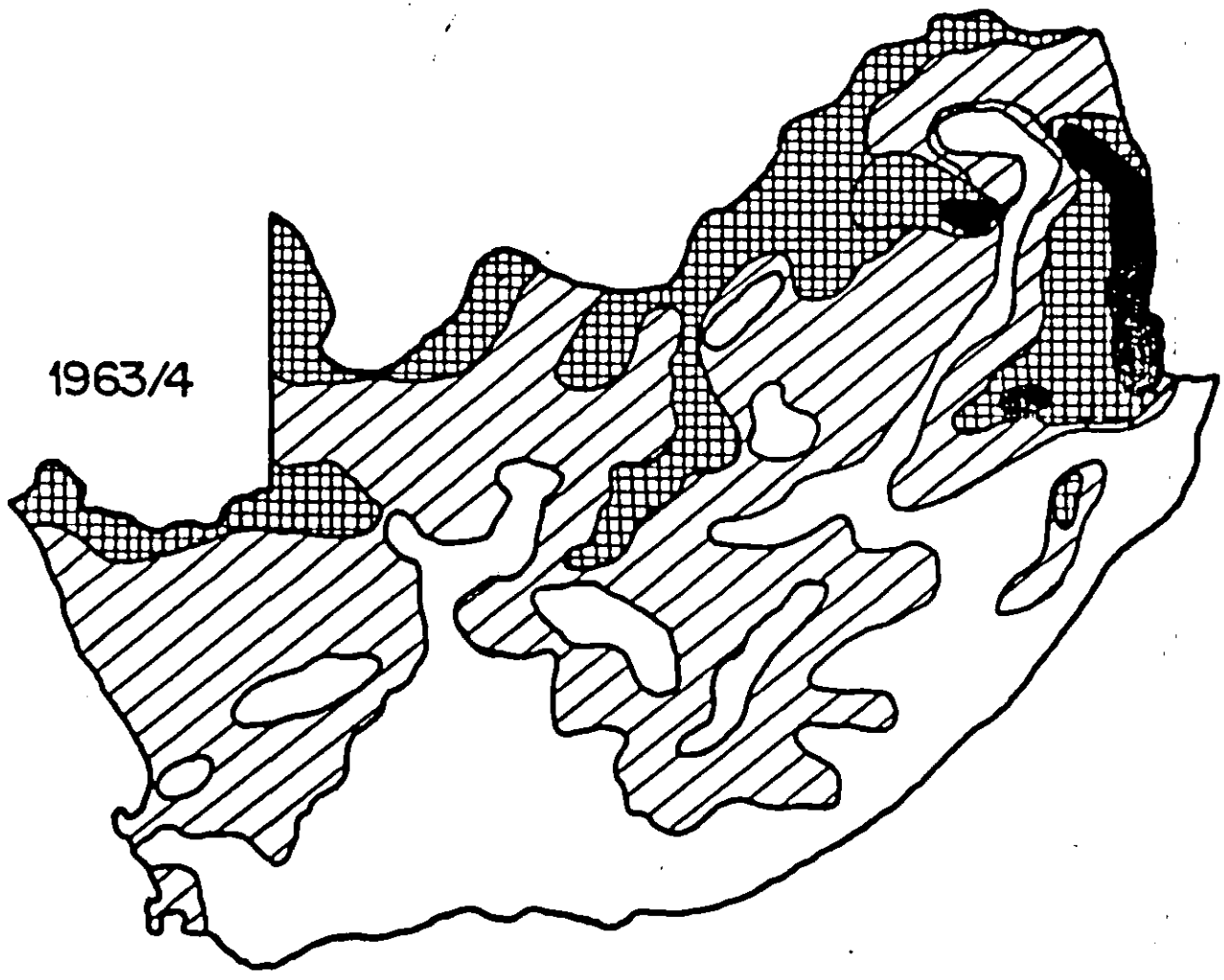

1958/9

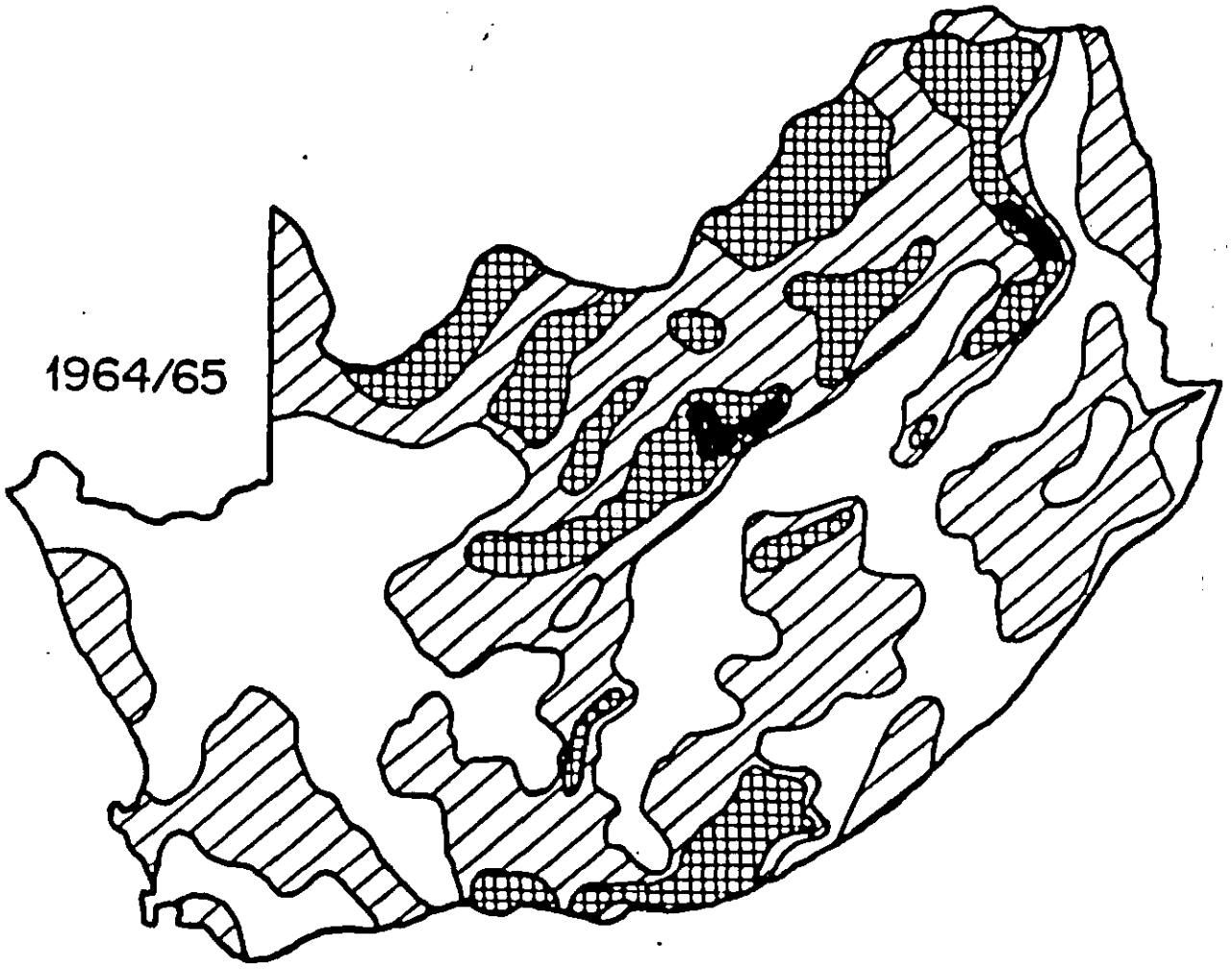

1959/60

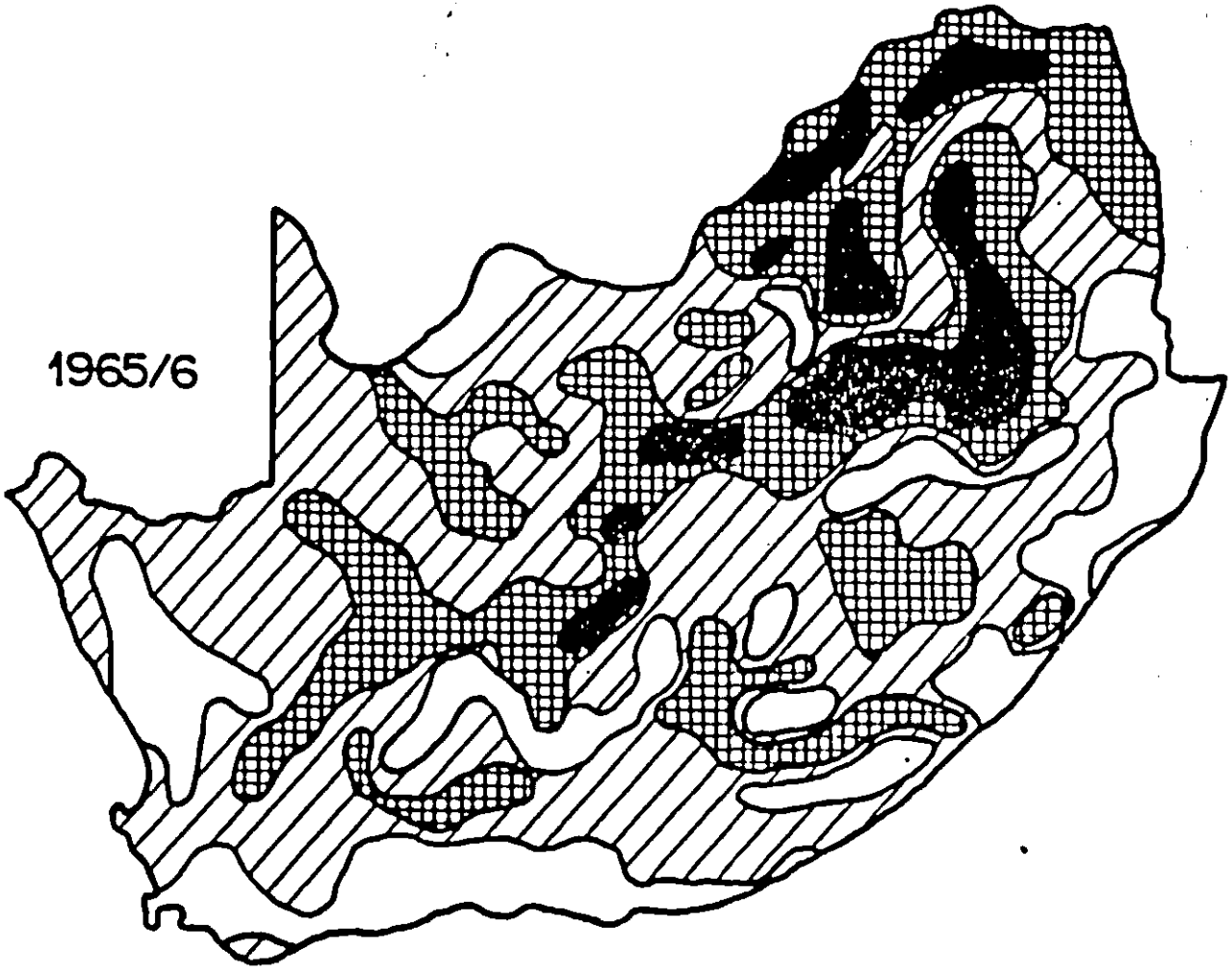

1960/1

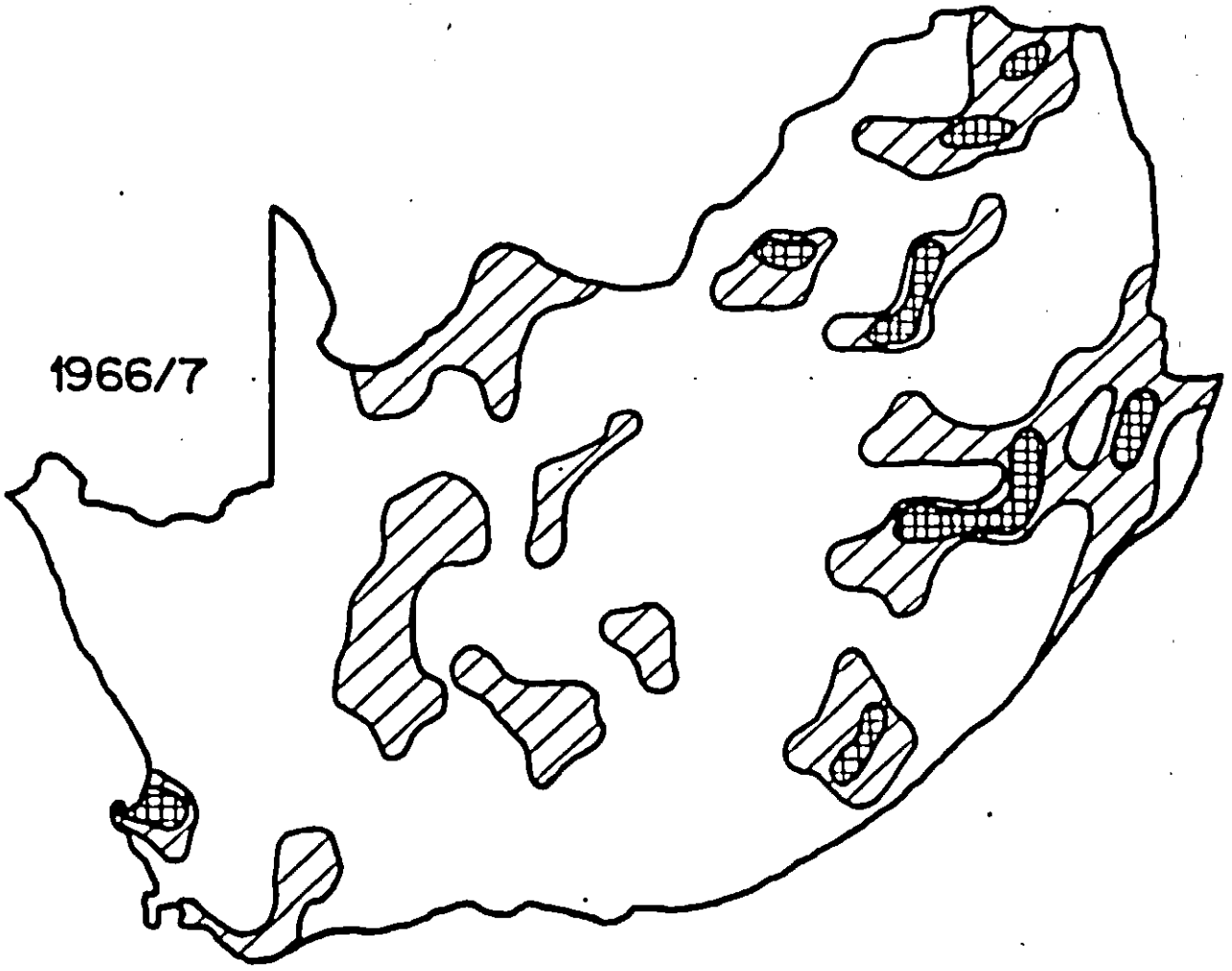

1961/2

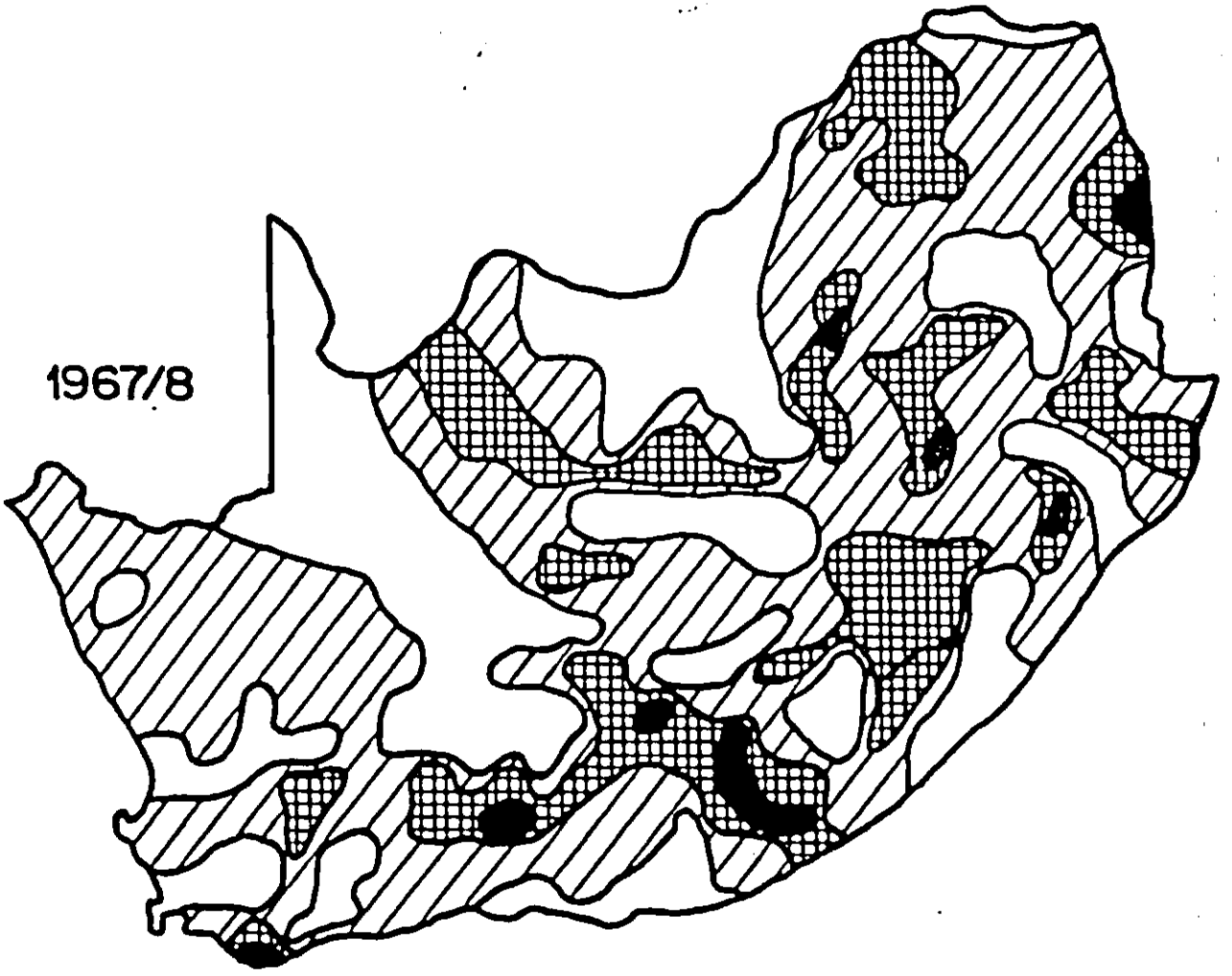

1962/3

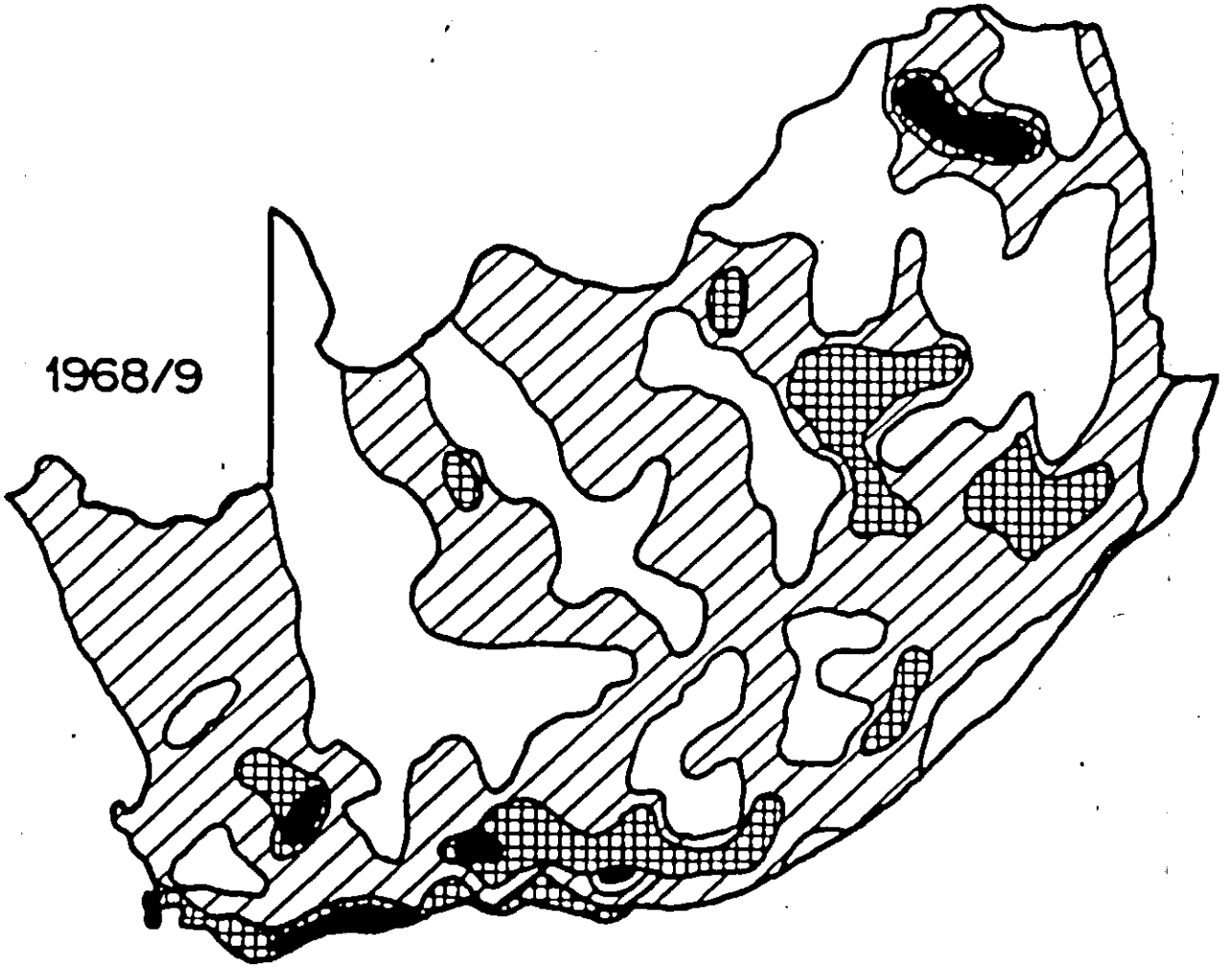

1963/4

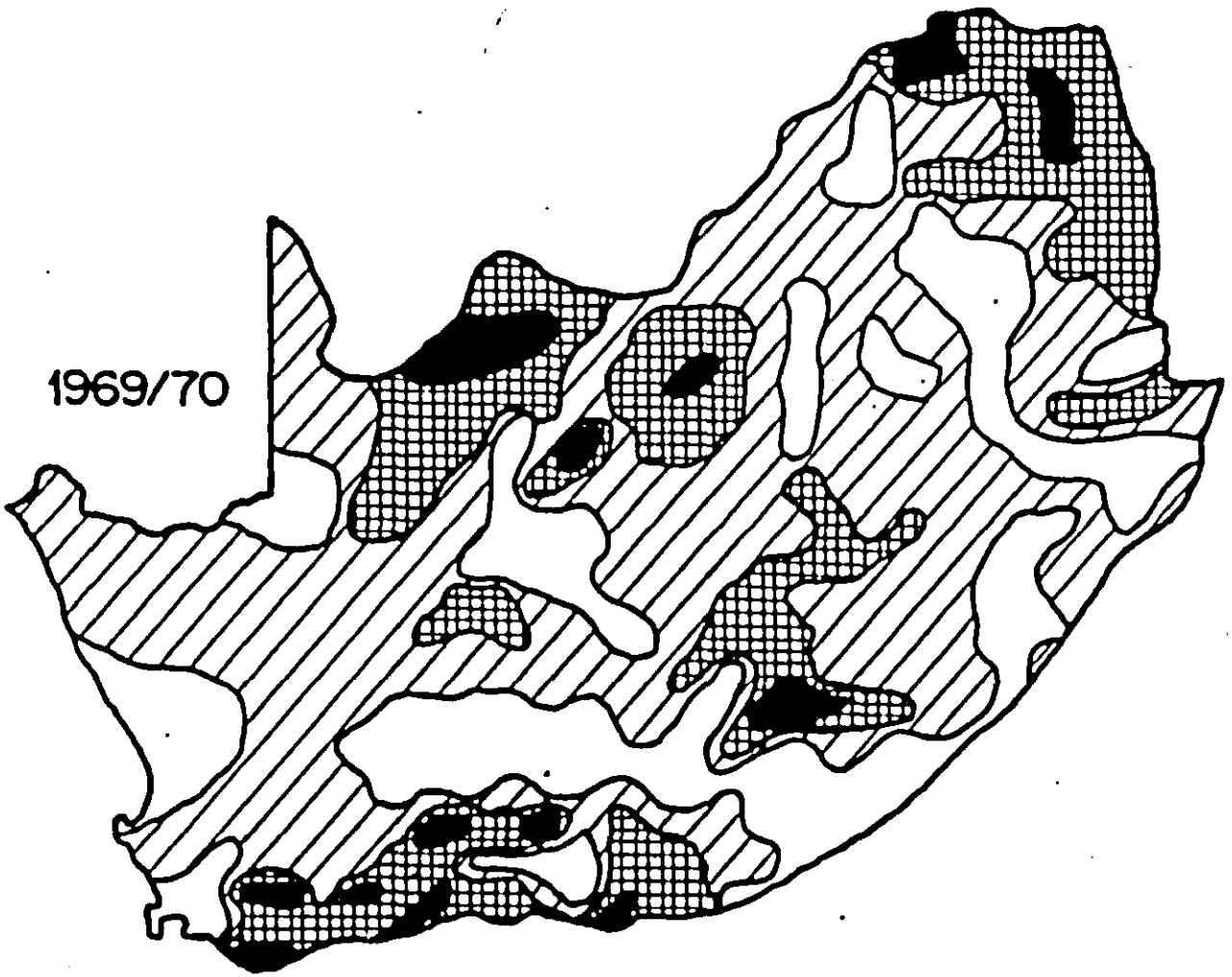

1964/65

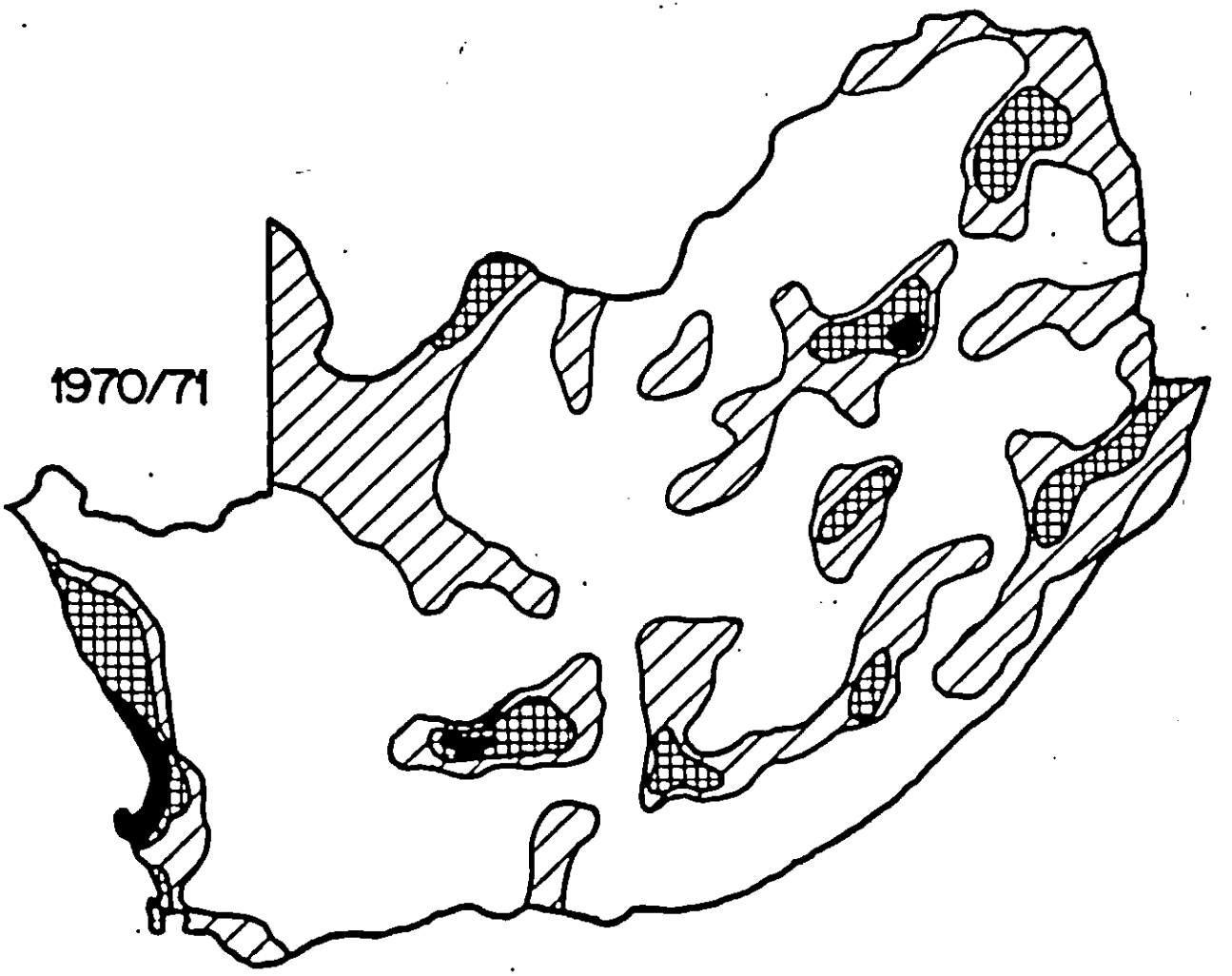

1965/6

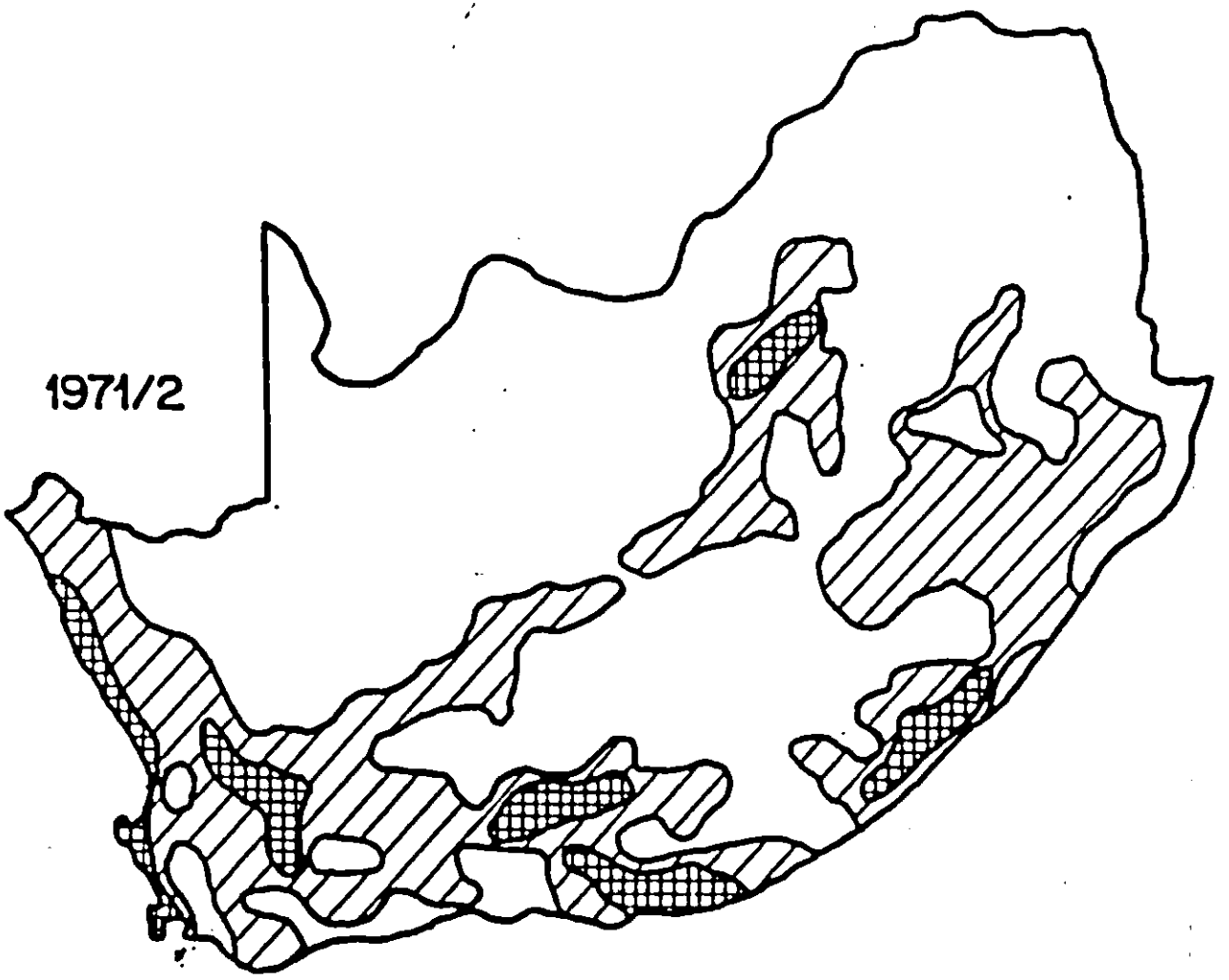

1966/7

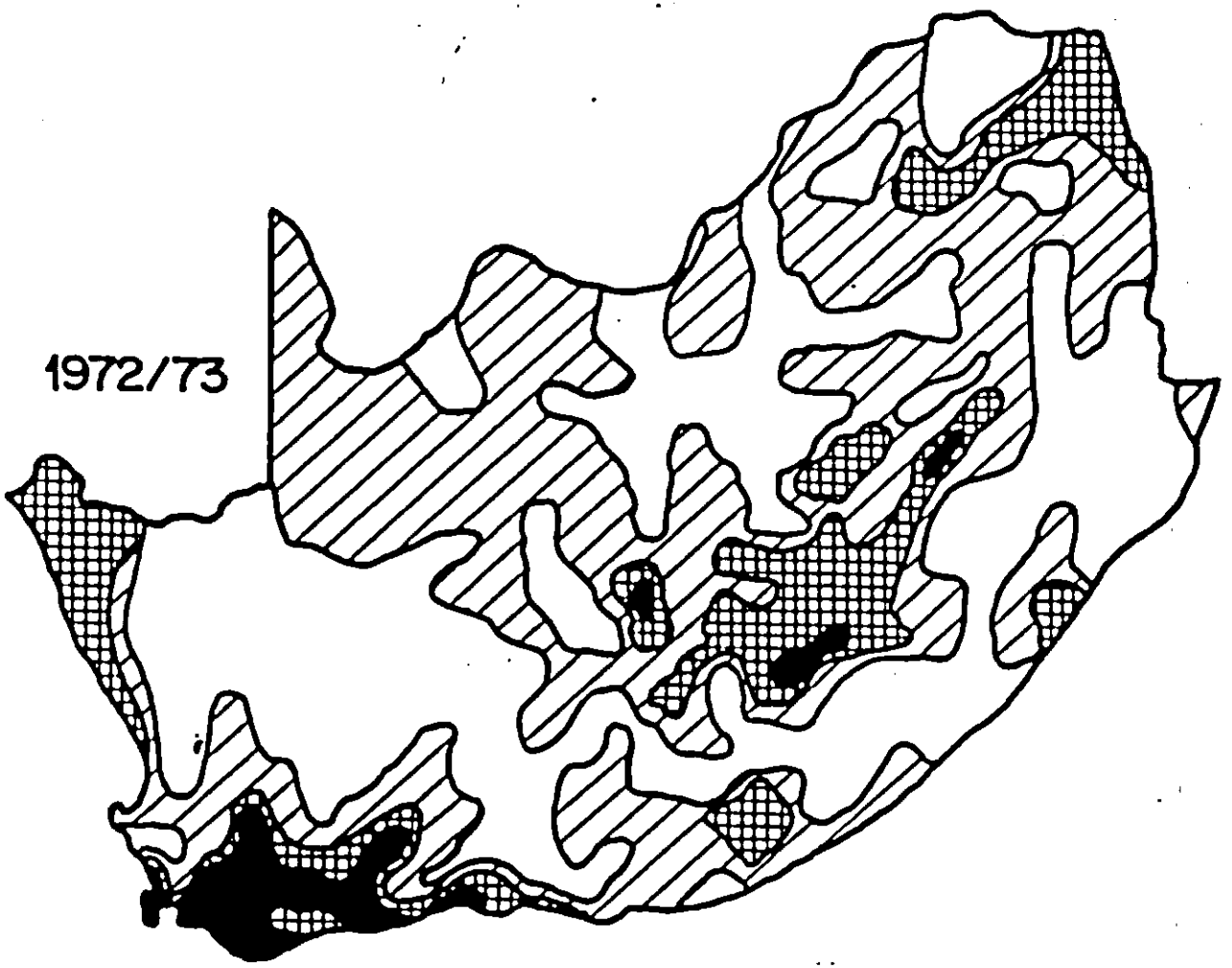

1967/8


1968/9

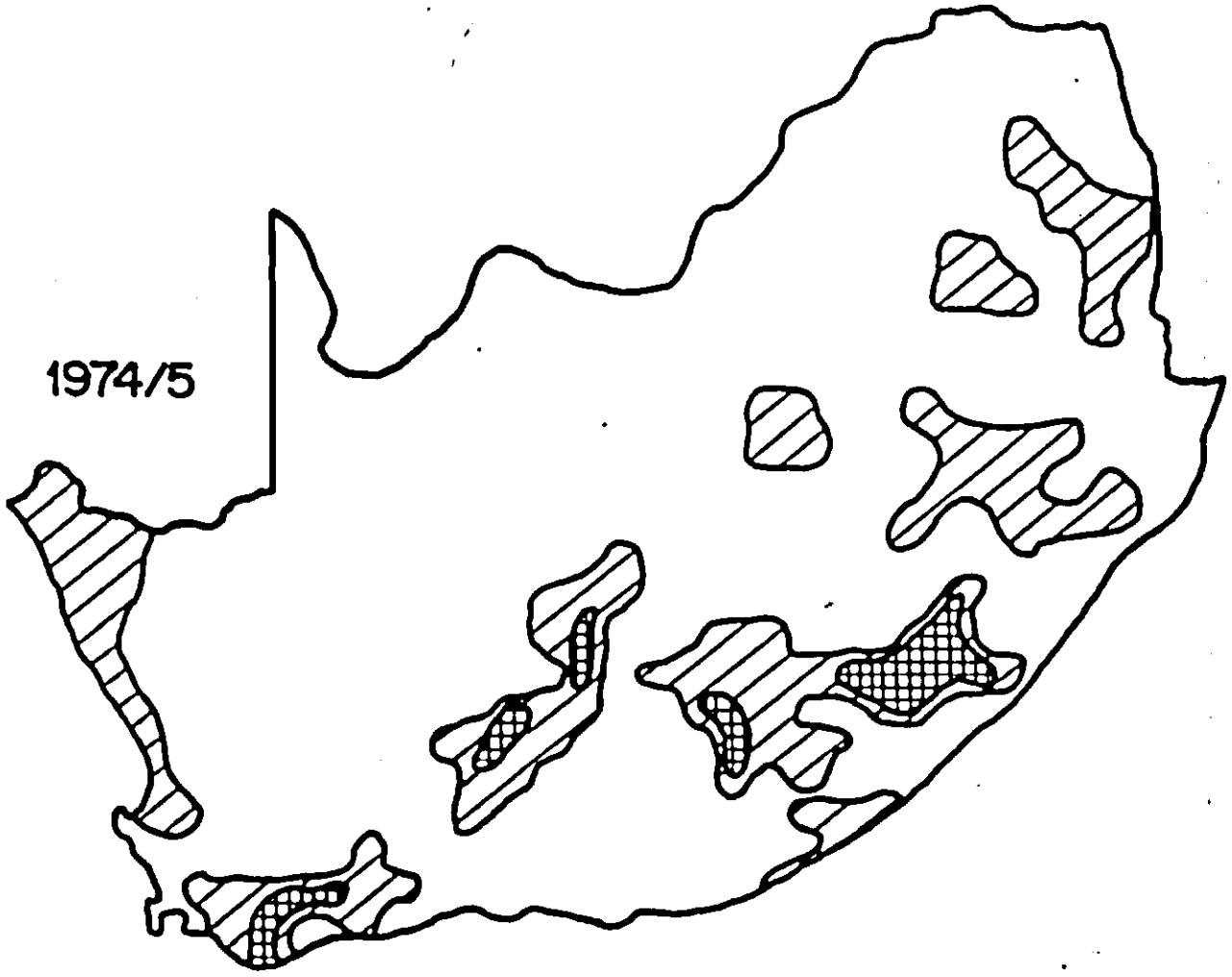

1969/70


1970/71

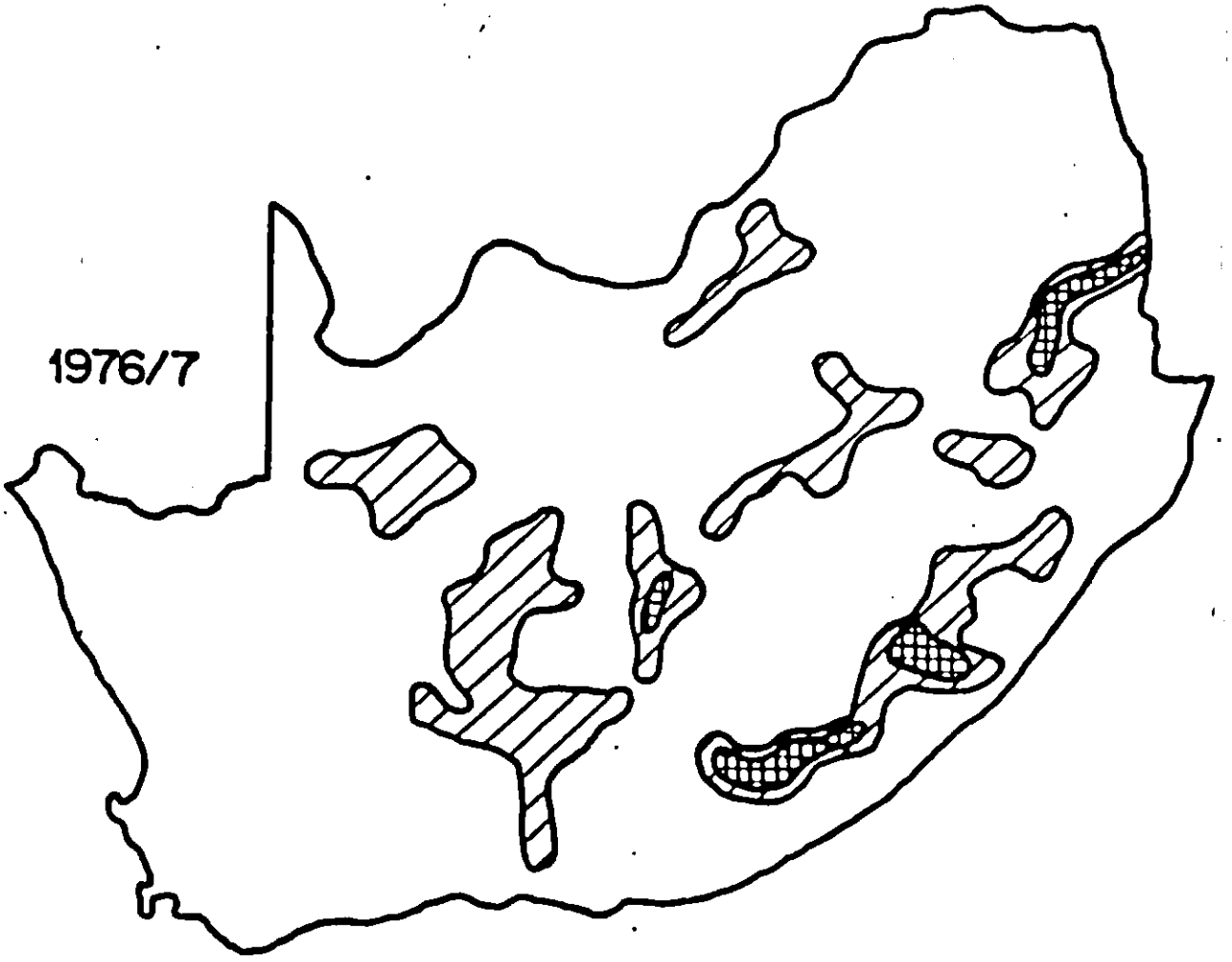

1971/2


1972/73


1973/4

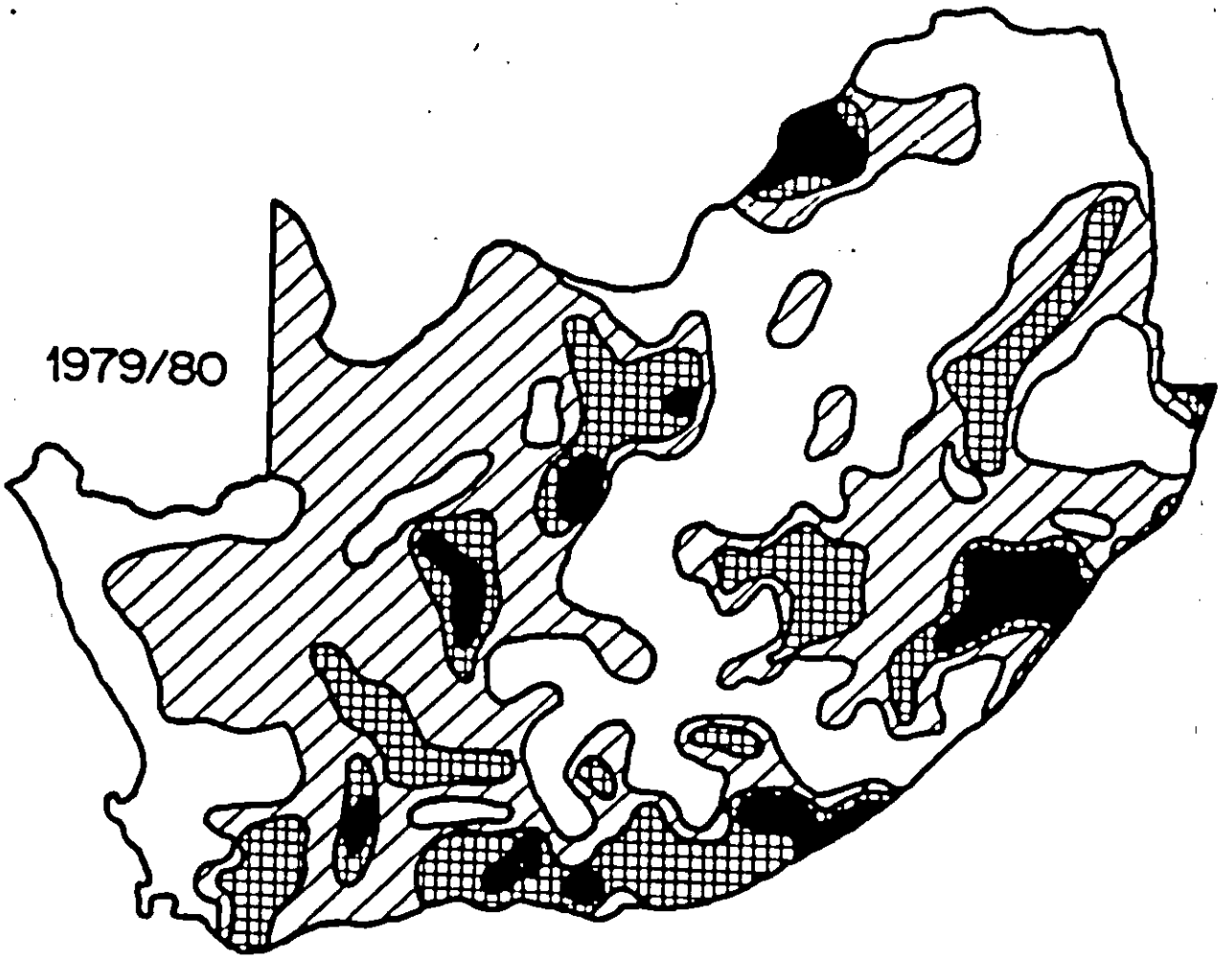

1974/5

1975/6


1976/7

1977/8

1978/9


1979/80

REFERENCES

ABRAMOVITZ, M. and STEGUN, I.A. (1972). *Handbook of Mathematical Functions*. Dover Publications, New York.

ANGELL, J.K. (1981). Comparison of variations in atmospheric quantities with sea surface temperature variations in the equatorial Pacific. *Monthly Weather Review*, 109, 230-243.

ARKIN, P.A. (1982). The relationship between interannual variability in the 200 MG tropical wind field and the southern oscillation. *Monthly Weather Review*, 110, 1393-1404.

BARGER, G.L. and THOM, H.C.S. (1949). Evaluation of drought hazard. *Agronomy Journal*, 41, 519-526.

BEUKES, D.J. and WEBER, H.W. (1981). Soil water studies on small lucerne plots. *Water SA*, 7(3), 166-174.

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

BRIDGES, T.C. and HAAN, C.T. (1972). Reliability of precipitation probabilities estimated from the gamma distribution. *Monthly Weather Review*, 100, 607-611.

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

DENNET, M.D., RODGERS, J.D. and KEATINGE, J.D.H. (1983). Simulation of a rainfall record for the site of a new agricultural development : An example from Northern Syria. *Journal of Agricultural Meteorology*, 29, 247-258.

DOVE, K. (1888). *Das Klima des Aussertropischen SudAfrika.* Vandenhoeck und Ruprecht, Göttingen.

DRAPER, H.R. and SMITH, H. (1966). *Applied Regression Analysis*, 2nd edition. Wiley Publications in Probability and Mathematical Statistics. John Wiley, New York.

FLACHS, J. and MARTIN, D. (1982). On smoothing non-negative data by means of non-negative functions, National Research Institute for Mathematical Sciences, Internal Report 1417, *Scientia*, Pretoria, July 1982.

GABRIEL, K.R. (1984). Discussion of the paper by Stern and Coe (1984). *Journal of the Royal Statistical Society, Series A*, 147, 28.

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.

GILLOOLY, J.F. and DYER, T.G.J. (1982). Variations in moisture deficits over the maize growing regions of South Africa : 1. Spatial aspects. *Water SA*, 8, 1-8.

GUPTA, S.S. and PANCHAPAKESAN, S. (1980). Some statistical techniques for climatological data. *Statistical Climatology : Developments in the Atmospheric Sciences*, 13, Elsevier.

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

HARRISON, M.S.J. (1983). A generalised classification of South African rain-bearing synoptic systems. *Climatology Research Group.* University of the Witwatersrand, Johannesburg. Submitted.

HERBST, P.H., BREDENKAMP, D.B. and BARKER, H.M.G. (1966). A technique for the evaluation of drought from rainfall data. *Journal of Hydrology*, 4, 264-272.

HOPKINS, J.W. and ROBILLARD, P. (1964). Some statistics of daily rainfall occurrence from the Canadian Prairies provinces. *Journal of Applied Meteorology*, 3, 600-602.

HUBER, P.J. (1977). *Robust Statistical Procedures*. CBMS Regional Conference Series in Applied Mathematics, No. 27, Society for Industrial and Applied Mathematics, Philadelphia.

HULLEY, A.G. (1980). Drought modelling in South Africa. Unpublished Masters Thesis, Department of Civil Engineering, University of Natal, Durban.

JACKSON, S.P. (1951). Climates of Southern Africa. *South African Geographical Journal*, 33, 17-37.

LINHART, H. and ZUCCHINI, W. (1982a). On model selection in analysis of variance. In B. Fleischmann *et al*, editors, *Oper. Res. Proceedings*, 1981, Springer Verlag, Berlin, 483-493.

LINHART, H. and ZUCCHINI, W. (1982b). A method for selecting the covariates in analysis of covariance. *South African Statistical Journal*, 16, 97-112.

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

MARKHAM, C.G. (1970). Seasonality of precipitation in the United States. *Annals of the Association of American Geographers*, 60, 593-597.

McGEE, O.S. and HASTENRATH, S.L. (1966). Harmonic analysis of rainfall over South Africa. *Notos*, 15, 79-90.

MOOLEY, D.A. and CRUTCHER, H.L. (1968). An application of the gamma distribution function to Indian rainfall. *Essa Technical Report*, EDS5. US Department of Commerce, Environmental Data Service, Silver Springs, Maryland.

NEYMANN, J. and SCOTT, E.L. (1967). Some outstanding problems relating to rain modification. Proceedings. *Fifth Berkley Symposium on Probability and Statistics*, University of California Press, 293-325.

PALMER, W.C. and DENNY, L.M. (1971). Drought bibliography. *NOAA Technical Memorandum*, EDS20. US Department of Commerce, Silver Springs, Maryland.

PAN, H.Y. and OORT, A.H. (1982). Global climate variations connected with sea surface temperature and anomalies in the equatorial Pacific Ocean for the period 1958 to 1973. *Monthly Weather Review*, 111, 1244-1258.

QUINN, W.H., ZOPF, D.O., SHORT, K.S. and KUO-YANK, R.T.W. (1978). Historical trends and statistics of the southern oscillation El Nino and Indonesian droughts. *Fisheries Bulletin*, 76, 663-678.

RASMUSSEN, E.M. and CARPENTER, T.H. (1982). Variations in tropical sea surface temperature and surface wind fields associated with the southern oscillation/ El Nino. *Monthly Weather Review*, 110, 354-384.

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

ROLDAN, J. and WOOLHISER, D. (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 classifications of Koppen and Thornthwaite. *South African Geographical Journal*, 29, 32-42.

SCHULZE, B.R. (1958). The climate of South Africa according to Thornthwaite's rational classification. *South African Geographical Journal*, 40, 31-53.

SCHUMANN, T.E.W. and THOMPSON, W.R. (1934). A study of South African rainfall secular variations and agricultural aspects. *Pretoria University Occasional Series*, No. 1.

SCHUMANN, T.E.W. and HOFMEYR, W.R. (1938). The partition of a region into rainfall districts with special reference to South Africa. *Journal of the Royal Meteorological Society*, 64, 482-488.

SHENTON, L.R. and BOWMANN, K.O. (1973). Remarks on Thom's estimators for the gamma distribution. *Monthly Weather Review*, 98, 154-160.

SICHEL, H.S. (1971). On a family of discrete distributions particularly suited to represent long-tailed frequency data. *Proc. 3rd Symp. on Math. Statistics* (N.F. Laubscher, ed.), S.A. C.S.I.R., Pretoria, 51-97.

SOUTH AFRICAN WEATHER BUREAU (1954-1963). Climate of South Africa, Parts 1-7. *Department of Transport*, Pretoria.

STERN, R.D. (1980). The calculation of probability distributions for models of daily precipitation. *Arch. Meteorol. Geophys. Bioklimatol*, Series B, 28, 137-147.

STERN, R.D. (1981). The start of the rains in West Africa. *Journal of Climatology*, 1, 59-68.

STERN, R.D. (1982). Computing a probability distribution for the start of the rains from a Markov chain model for precipitation. *Journal of Applied Meteorology*, 21, 420-423.

STERN, R.D. and COE, R. (1982). The use of rainfall models in agricultural planning. *Agricultural Meteorology*, 26, 35-50.

STERN, R.D. and COE, R. (1984). A model fitting analysis of daily rainfall data. *Journal of the Royal Statistical Society*, Series B, 147, 1-34.

THOMPSON, G.D. (1981). Water use of sugar cane. South African Sugar Industry Agronomists Association. *Review Paper* No. 8. Experimental Station, Mount Edgecombe, Natal.

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

TONG, H. (1975). Determination of the order of a Markov chain by Akaike's information criterion. *Journal of Applied Probability*, 12, 488-497.

TYSON, P.D. and DYER, T.G.J. (1978). The predicted above normal rainfall in the seventies and the likelihood of droughts in the eighties in South Africa. *South African Journal of Science*, 74, 372-377.

WEISS, L.L. (1964). Sequences of wet and dry days described by a Markov chain model. *Monthly Weather Review*, 92, 169-175.

WELDING, M.C. and HAVENGA, C.M. (1974). The statistical classification of rainfall stations in the Republic of South Africa. *Agrochemophysica*, 6, 5-24.

WELLINGTON, J.H. (1955). *Southern Africa : A Geographical Study, Vol. 1, Physical Geography*. Cambridge University Press.

WINSTON, J.S. (1982). Climate of spring, 1982, a season of abnormally strong westerlies. *Monthly Weather Review*, 110, 1729-1744.

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.

YEVJEVICH, V. and DYER, T.G.J. (1983). Basic structure of daily precipitation series. *Journal of Hydrology*, 64, 49-67.

ZUCCHINI, W. (1974). Statistical models for droughts and floods. Paper read at annual conference of the South African Statistical Association, Bloemfontein.