SEASONAL FORECASTING OF SOUTH AFRICAN RAINFALL USING A NON-LINEAR DISCRIMINANT ANALYSIS MODEL

S. J. MASON*

*Climatology Research Group, University of the Witwatersrand, P.O. Wits, Johannesburg 2050
email:simon@crg.bpb.wits.ac.za*

ABSTRACT

Statistical models have been used to provide operational seasonal forecasts of rainfall over southern Africa since 1992. The Climatology Research Group has been using a quadratic discriminant analysis model since November 1994. The model relates rainfall over different areas of South Africa to principal components of sea-surface temperatures in the Indian, South Atlantic and Pacific Oceans. Details of the model are described in this paper. High forecast-skill levels can be achieved for much of the country throughout the year. The mostly successful performance of the model over a 15-year independent testing period indicates that the model can be used successfully in an operational environment. © 1998 Royal Meteorological Society.

KEY WORDS: discriminant analysis; linear error in probability space; rainfall; sea-surface temperatures; seasonal forecasting; South Africa; skill-score measures.

1. INTRODUCTION

Partly in response to the devastating effects of the 1991–1992 drought in southern Africa (Vogel and Drummond, 1993), the South African Weather Bureau and a number of research groups at some of the South African universities have begun to release operational seasonal rainfall forecasts in the last few years (Mason *et al*., 1996). In October 1994 the South African Weather Bureau founded the South African Long-lead Forecast Forum (SALFF) with the aims of developing and co-ordinating the seasonal forecasting capabilities of the country. The forecasts produced by the SALFF member organizations are based on statistical associations between sea-surface temperatures, outgoing longwave radiation and atmospheric circulation indices in the tropics, subtropics and mid-latitudes (Jury, 1996; Mason *et al*., 1996; Mason, 1997). Upper atmospheric circulation indices are used in seasonal forecasting over southern Africa in addition to surface features and include the Quasi-Biennial Oscillation (Jury *et al*., 1994; Mason *et al*., 1994) and the zonal component of upper tropospheric winds over the equatorial Atlantic Ocean (Jury, 1996). The main source of predictability over southern Africa is from the tropical atmospheric circulation, which responds directly to sea-surface temperature anomalies (Walker, 1990; Jury and Pathack, 1993; Mason, 1995), including El Niño events (Lindesay, 1988; Jury *et al*., 1994). As a result, highest forecast skills are obtainable during the period December–February (Barnston *et al*., 1996) when rainfall amounts reach a maximum over most of the subcontinent (Tyson, 1986) and when tropical atmospheric influences on the region predominate (Harrison, 1984).

Although general circulation models do not appear to indicate high levels of forecastability of the atmosphere on seasonal time-scales over southern Africa (Dix and Hunt, 1995; Stern and Miyakoda, 1995), theoretical considerations suggest that there should be reasonable prospects for obtaining useful forecast skill and lead-times

---

*Correspondence to: S. J. Mason, Scripps Institute of Oceanography, La Jolla CA 92093-0224, USA.

over most of the subcontinent (Mason *et al*., 1996; Mason, 1997). The predictability of the atmosphere is probably underestimated by the general circulation models partly because of inadequate/incorrect simulated responses to sea-surface temperature anomalies (Latif *et al*., 1994). Modelled tropical convection responses to positive sea-surface temperature anomalies are generally too weak (Mo and Wang, 1995; Smith, 1995), including in the southern African region (Joubert, 1997). Further, the importance of atmospheric chaos is exaggerated by model systematic errors (Dix and Hunt, 1995), which again are responsible for underestimates of potential predictability. The continued use of statistical methods in seasonal forecasting of rainfall over southern Africa should therefore be encouraged at least until the dynamic models can respond more realistically to sea-surface temperature anomalies (Mason, 1997).

Each member of the SALFF uses different sets of predictors and statistical techniques in producing the operational long-range rainfall forecasts. The forecasts produced by the Research Group for Seasonal Climate Studies (RGSCS) of the South African Weather Bureau are based on a canonical correlation model, which uses sea-surface temperatures as predictors. The RGSCS, additionally produce one- and three-month forecasts using a general circulation model. The Climate and Weather Research Laboratory of the University of Cape Town uses a multiple regression model, relating rainfall variability to sea-surface temperatures, the Southern Oscillation Index, various surface and upper atmospheric wind indices and outgoing longwave radiation (Jury, 1996). The Climatology Research Group began to release Seasonal Rainfall Outlooks in September 1992, which were originally based on multiple linear regression models relating sea-surface temperatures in the Indian and Atlantic Oceans and the Southern Oscillation Index to rainfall over South Africa (Mason, 1995; Mason *et al*., 1996). Since November 1994 the Seasonal Rainfall Outlooks have been based on the results of a discriminant analysis model. In addition to the models used by the SALFF members, a neural network model has been developed for South African rainfall and indicates useful forecast skills (Hastenrath *et al*., 1995). The discriminant analysis model used by the Climatology Research Group is adapted from a method developed at the United Kingdom Meteorological Office for producing seasonal forecasts for the Sahel region of West Africa and north-east Brazil (Ward and Folland, 1991). The forecasting model is described in this paper and the seasonal and spatial dependence of skill scores are outlined. A description of the different skill-score measures used is included.

2. DATA AND METHODS

The Climatology Research Group's seasonal rainfall forecasting model is trained by relating principal components of sea-surface temperature to rainfall for different areas of South Africa. The data and statistical methods involved in the training of the model are outlined in the sections below.

2.1. Rainfall data and regionalization

Monthly station rainfall data for the 45-year period January 1951 to December 1995 were obtained from the Computing Centre for Water Research, University of Natal. The stations were pre-selected on the basis of completeness of records. Stations were only selected if rainfall recordings were kept over the entire 45-year period and if there was a maximum of 10 per cent missing values. A total of 430 stations across South Africa were obtained (Figure 1). Missing values were replaced by the respective monthly mean over the 45-year period.

The 430 stations were grouped into homogeneous rainfall regions based on the standardized monthly rainfall totals over the full 45-year period using cluster analysis. The aim of the cluster analysis was to group the rainfall stations into regions with similar interannual rainfall variability, within each of which the mechanisms responsible for the rainfall variability should be similar. The data were first standardized to eliminate problems of non-linearity that arise if the variables (monthly rainfall) do not have equal variance (Everitt, 1980). Euclidean distances are most commonly used when calculating the distance matrix for cluster analysis (Gong and Richman, 1995), but if the variables are correlated Mahalanobis distances should be calculated instead (Everitt, 1980). The Mahalanobis distance, D_{ij} between objects i and j can be calculated from

$$D_{ij} = \sqrt{(x_i - x_j)^T \mathbf{R}^{-1} (x_i - x_j)} \tag{1}$$

where x_i is the ith sample point in p dimensional space and **R** is the $p \times p$ correlation matrix. Because the number of objects to be clustered (430 stations) was less than the number of variables (45×12 months), **R** would be singular and so its inverse does not exist. The total number of months was therefore reduced by calculating principal components from the correlation matrix: eight of the principal components each explain more than 5 per cent of the total variance and were accordingly retained. These eight principal components together explain 65 per cent of the total variance. The components were not rotated because their physical interpretability was not important: the analysis was performed purely to reduce the number of variables. The principal components are by definition orthogonal so that **R** is an identity matrix, in which case the Mahalanobis distance is then equivalent to the Euclidean distance. Once the distances were calculated, hierarchical clustering was performed using Ward's minimum variance criterion (as recommended by Gong and Richman, 1995). The number of clusters was selected by identifying a 'significant' jump in the distance between clusters merged at each step, which implies that two relatively dissimilar clusters have been merged. A total of eight regions were defined (Figure 1), each of which contains at least 20 stations.

For each station, running three- and six-month totals were calculated from the raw data, starting in March–May and March–August 1951 and ending February–April and February–July 1995. Given n stations in region l, let x_{ijkl} represent the three- or six-month rainfall at station k in year i and beginning in month j. The average January–March rainfall over the 45-years, for example, can then be expressed by $\bar{x}_{1kl}$ and the standard deviation σ_{1kl}. The totals were standardized for each seasonal period separately and then averaged across all stations in each region, using equation (2),

$$z_{ijl} = \frac{1}{m} \sum_{l=1}^{m} \frac{x_{ijkl} - \bar{x}_{ikl}}{\sigma_{jkl}} \tag{2}$$

to give a regional index z_{ijl}. The procedure was repeated for all eight regions. For each region and starting month, the three-month indices were then grouped into five equi-probable rainfall categories. The driest 20 per cent of the 45 years of three-month rainfall indices were thus classified as 'very dry', the next driest 20 per cent as 'dry' and so on for the 'average', 'wet' and 'very wet' categories. By definition, the *a priori* probability of any three-month period from the training period falling within one of the categories is 20 per cent, and each category contains nine values. The six-month indices were similarly categorized, but using only three equi-probable categories, with a priori probabilities of 33 per cent, and 15 cases in each category.

2.2. Sea-surface temperature data

Sea-surface temperature data for the Atlantic and Pacific Oceans south of 20°N and the entire Indian Ocean were extracted from the Meteorological Office Historical Sea Surface Temperature version 6 (MOHSST6) data set for the 46-year period January 1950 to December 1995. The data consists of quality controlled monthly sea-surface temperature anomalies for 5° latitude by 5° longitude grids (Parker, 1987; Parker and Folland, 1988) and the anomalies are calculated with respect to the 1961–1990 monthly means. The data were averaged into 10° by 10° grids, partly to reduce the number of missing values and partly to reduce the dimensionality of the data set. Any remaining missing values were replaced with Gaussian noise (Mason, 1995) provided that at least one-third of the observations were present.

Rotated principal components analysis was then performed on the sea-surface temperature data for each of the three oceans separately. For the Atlantic and Pacific Oceans only data south of 20°N were considered, whereas all available data for the Indian Ocean were included. The analyses were based on the variance–covariance matrix for each ocean and spatial loadings were rotated using the varimax criterion to generate robust and physically interpretable components (Richman, 1986, 1987, 1993; Cheng *et al.*, 1995). Only those components that explain at least 5 per cent of the total variance were retained. Five components were retained for the Atlantic Ocean, four for the Indian and two for the Pacific Oceans. Their spatial loadings were divided by the square root of the eigenvalue and the resultant spatial score coefficients are illustrated in Figure 2–4. For the Atlantic Ocean, components 1–4 and component 6 were retained: after rotation component five explained approximately 3 per cent of the total variance only, and the spatial scores pattern indicated high coefficients in just a few grid areas
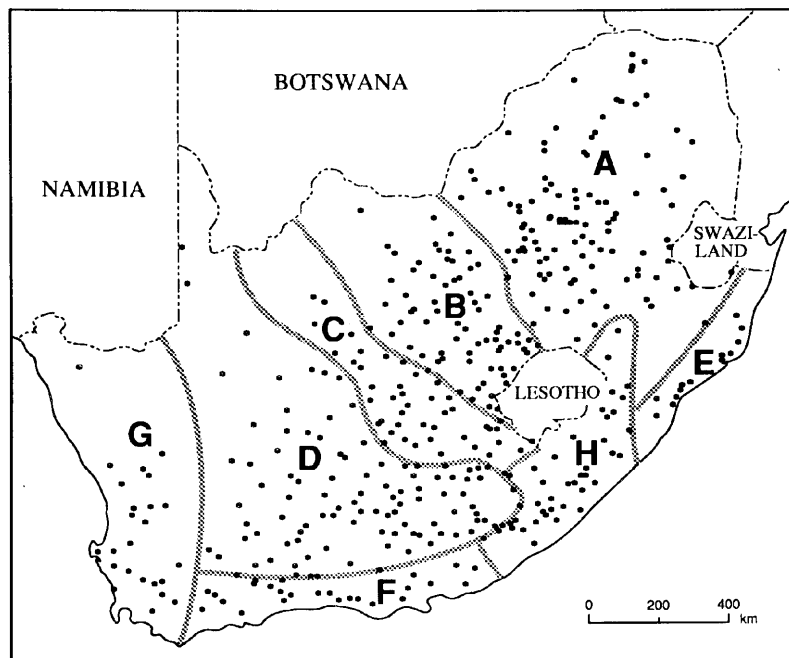
Figure 1. Location map of the 430 rainfall stations and the eight rainfall regions defined by clustering of the spatial loadings of the first eight non-seasonal principal components of monthly rainfall for the period 1951–1995

where data quality is poor. The first of the Pacific Ocean principal components (Figure 4(a)) represents variability in the NIÑO3 region (5°N to 5°S, 90°W to 150°W). Sea-surface temperatures in this region are sometimes used as a less noisy index of the El Niño–Southern Oscillation phenomenon than the Southern Oscillation Index (Cane *et al.*, 1994).

The first components of each of the three oceans represent, to a large extent, equatorial sea-surface temperature variability. An additional three predictors were defined by calculating the differences between the scores on these first components to give indices of contrasts in equatorial sea temperatures between the three oceans. Differences in equatorial sea-surface temperatures between the Atlantic and Pacific Oceans are thought to have an important influence on rainfall over West Africa because of an effect on the Walker circulation (Janicot *et al.*, 1996). Zonal sea-surface temperature gradients may be similarly important for southern Africa (Mason and Jury, 1997) and could account for the strong statistical association between upper tropospheric zonal winds over the central equatorial Atlantic Ocean and rainfall over southern Africa (Jury, 1996).

2.3. The non-linear discriminant analysis model

The general problem addressed by discriminant analysis is to assign an individual to a category based on the values of the vector of 'independent' variables, x. It is assumed that each individual has to be assigned to one of the categories and that no other categories are allowed. A classification rule, or discriminant function, is defined to make the assignment. Given a population P with a proportion p_i in category G_i, the probability density function of observations within G_i can be defined as $f_i(x)$. The assignment is made to the group for which the value $f_i(x)p_i$ is maximized. It is usually assumed that the underlying population is multivariate normal with mean vector μ and covariance matrix Σ. Because the mathematics becomes exceptionally complicated if this assumption is relaxed, multivariate normality was assumed throughout. In linear discriminant analysis the further assumption of equality of variance between the different groups is adopted, but if this assumption is relaxed, the discriminant function
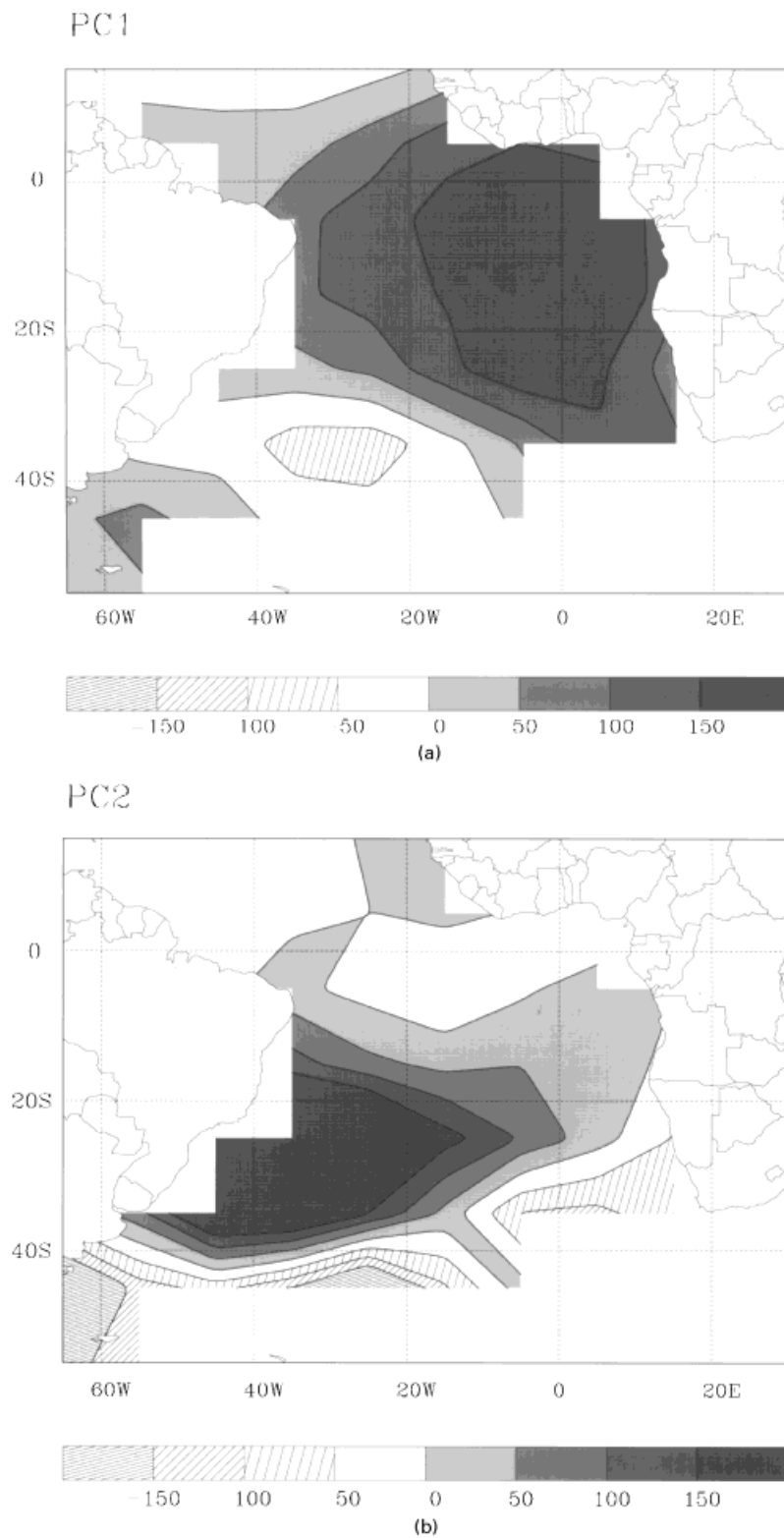
PC1

PC2

Figure 2. Spatial core coefficients of the first four and the sixth (a–e) principal components of monthly sea-surface temperature anomalies in the Atlantic Ocean south of 20°N for the period 1950–1995. The principal components were calculated from the variance–covariance matrix and were rotated using the varimax method
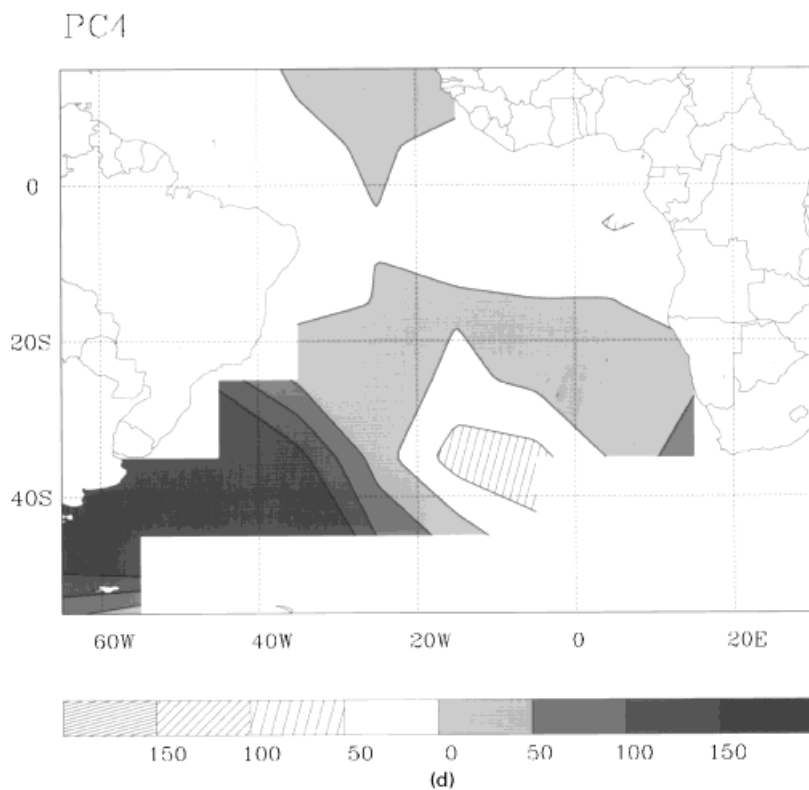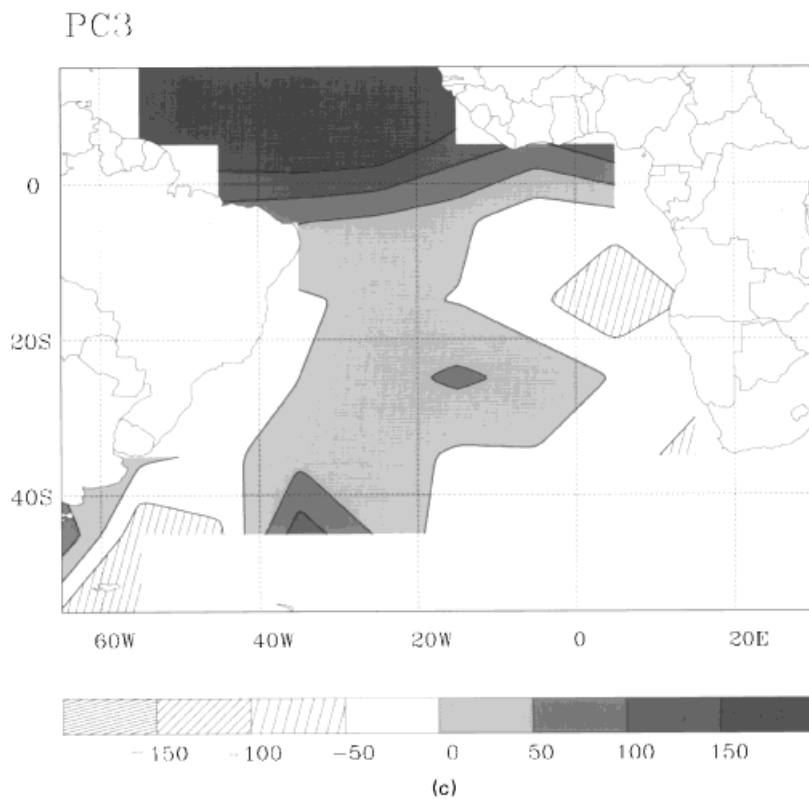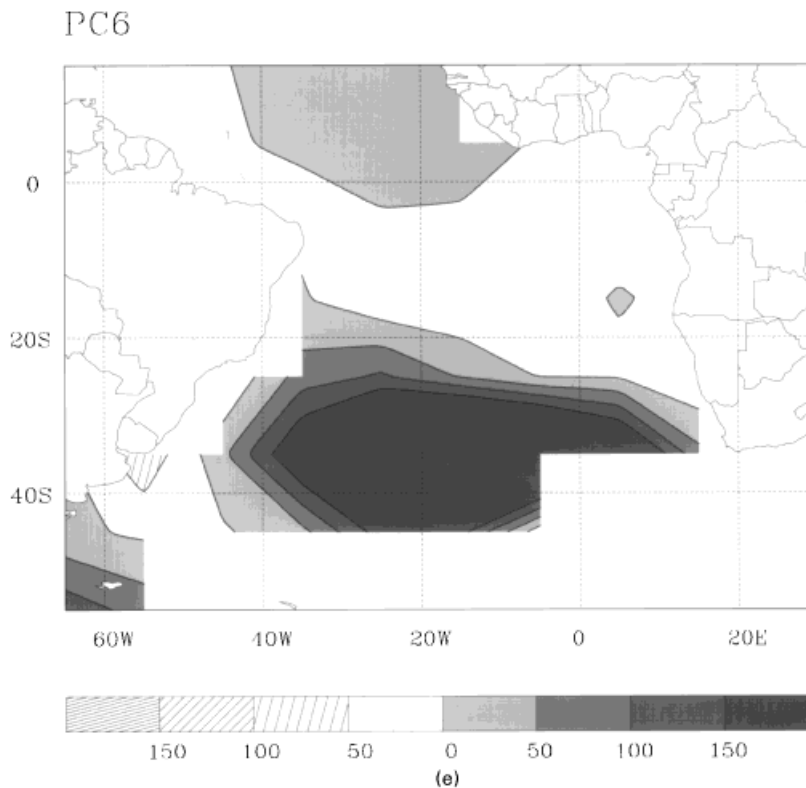
PC3

(c)

PC4

(d)

Figure 2 c–d

PC6

Figure 2e

becomes quadratic (Manly, 1994). The NAG routines G03DAF and G03DCF were used (Numerical Algorithms Group, 1996).

There are two established methods of estimating the probability distribution function $f_i(x)$: an estimative and a predictive approach. There are, therefore, four possible alternatives: an estimative approach that assumes equality of variance (Ee) and one that does not assume equality of variance (Eu); a predictive approach that assumes equality of variance (Pe) and one that does not (Pu). It has been shown that the estimates derived using Pu are 'far superior' when the assumption of equal variances is not valid, but that Pe is preferable when equality is obtained (Aitchison *et al.*, 1977; Moran and Murphy, 1979). Therefore, a predictive approach was used throughout and tests for the equality of variance were conducted to decide between Pe and Pu.

The two-month mean sea-surface temperature principal component scores and seasonal total regionalized rainfall indices were used as input into the non-linear discriminant analysis model. Two-month mean sea-surface temperature values were used to reduce the effects of noise that result from sampling variability. The predictor variables and rainfall indices were lagged so that the model was trained for potential operational use as a forecast tool. All operational forecasts are produced for three- and six-month totals and in operational use the forecasts for January–March and January–June, for example, can be made only with the latest available observed sea-surface temperature data. Hence, the January–March rainfall is forecast using November–December sea-surface temperature principal component scores. The model was trained with these operational restrictions in mind and, consequently, was constructed to relate seasonal rainfall totals to sea-surface temperatures of the two months prior to the first month of the rainfall index.

As with multiple regression, not all the candidate predictors contribute to a significant improvement in the discriminant analysis model. Unfortunately, however, there is no established method for assessing which predictors are useful when the assumption of equal variance is dropped. Typically, stepwise linear multiple
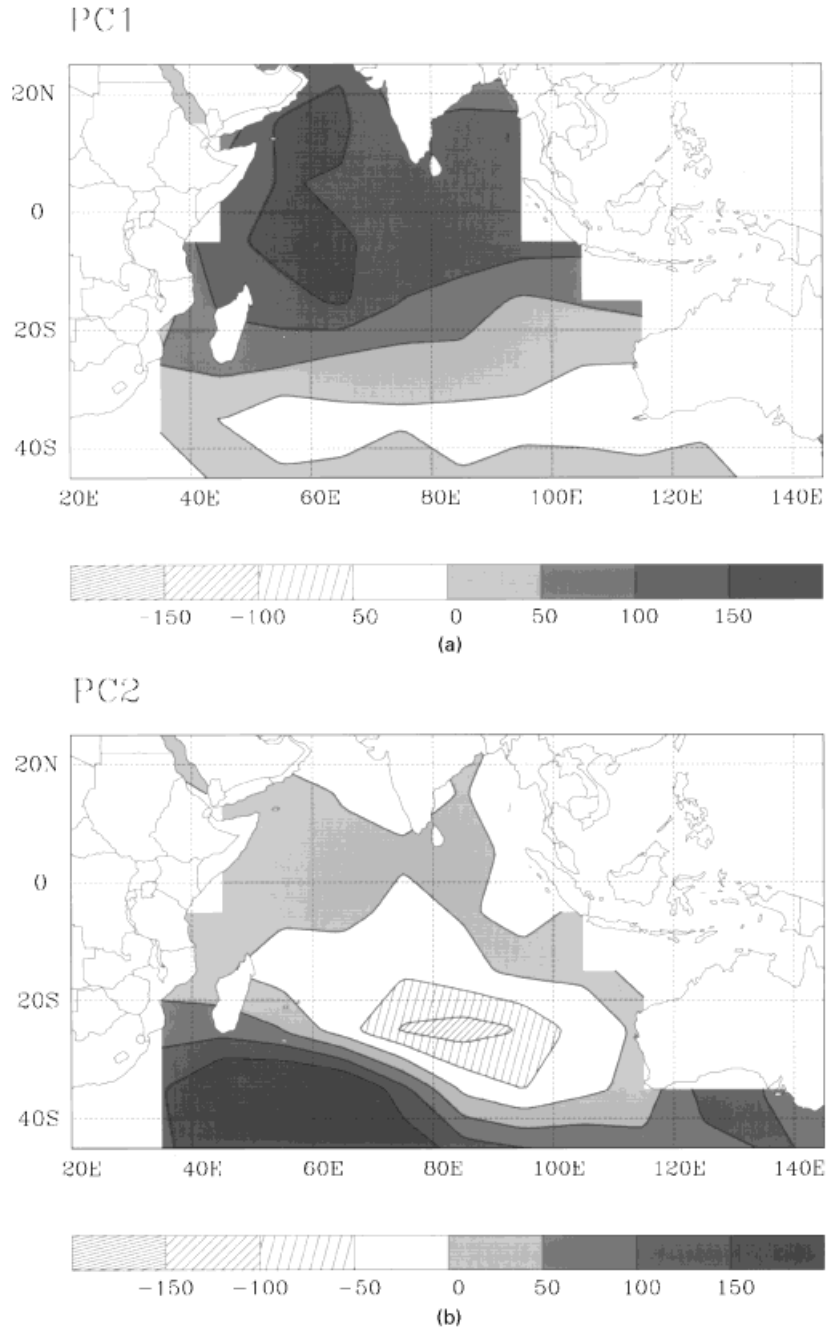
Figure 3. Spatial score coefficients of the first four (a–d) principal components of monthly sea-surface temperature anomalies in the Indian Ocean for the period 1950–1995. The principal components were calculated from the variance–covariance matrix and were rotated using the varimax method

regression is used as a first step to indicate which variables to input to the discriminant analysis model (e.g. Ward and Folland, 1991), but since the model used here is non-linear, this procedure was considered inappropriate. Instead, it was decided that the number of predictors would be kept to a maximum of three. Experience with using four predictors suggested that the fourth variable over-parameterizes the model. All possible permutations of one, two and three predictor variables were considered.
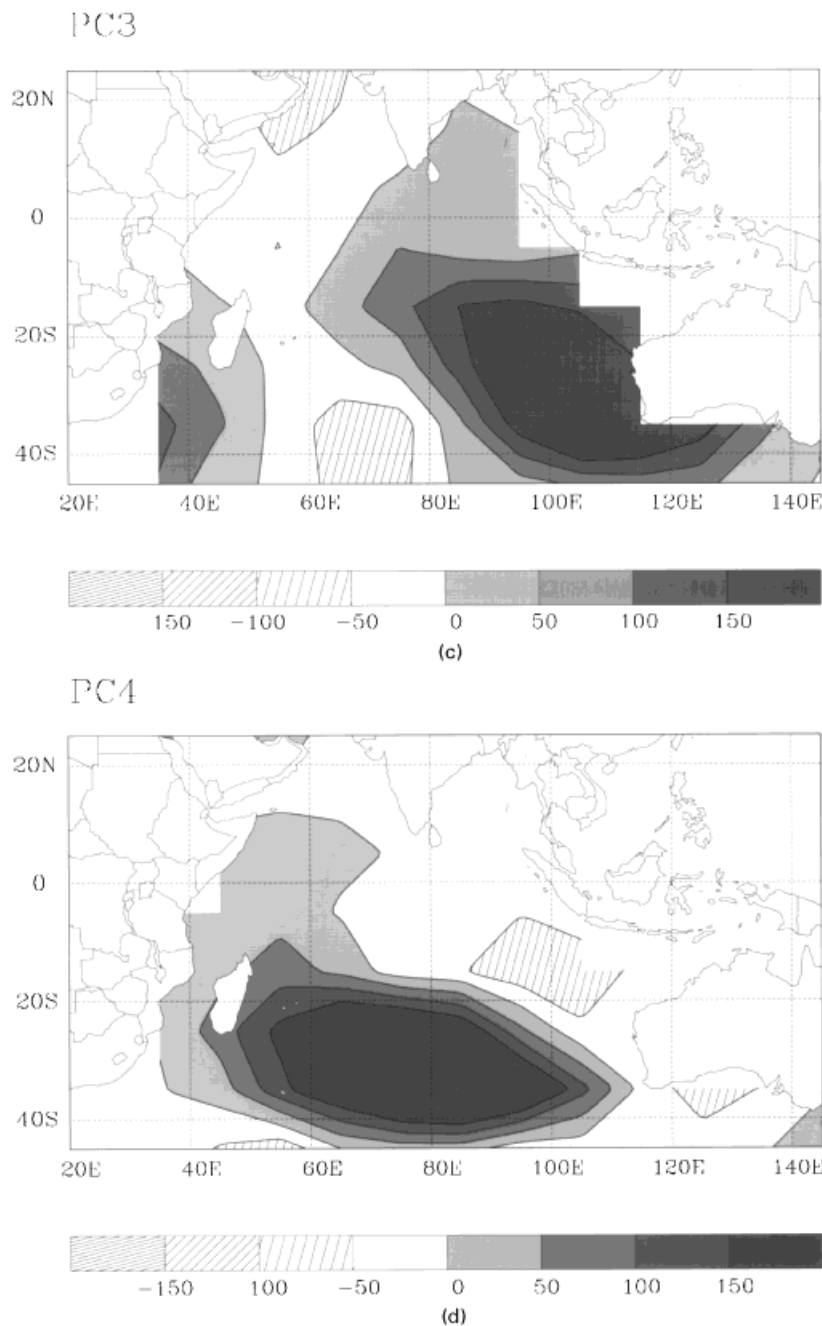
PC3

(c)

PC4

(d)

Figure 3c–d

2.4. Model validation

Validation is an important component of model construction, particularly with discriminant analysis, which lacks an adequate measure of significance akin to the R^2 statistic used in regression procedures (Montgomery and Peck, 1992; Manly, 1994). Two approaches to model validation were adopted. The most recommendable method is to construct the model using a subset of the data available and to then test the model over the remaining periods (Wilks, 1995). Unfortunately, this approach reduced the number of observations available to train the model. As
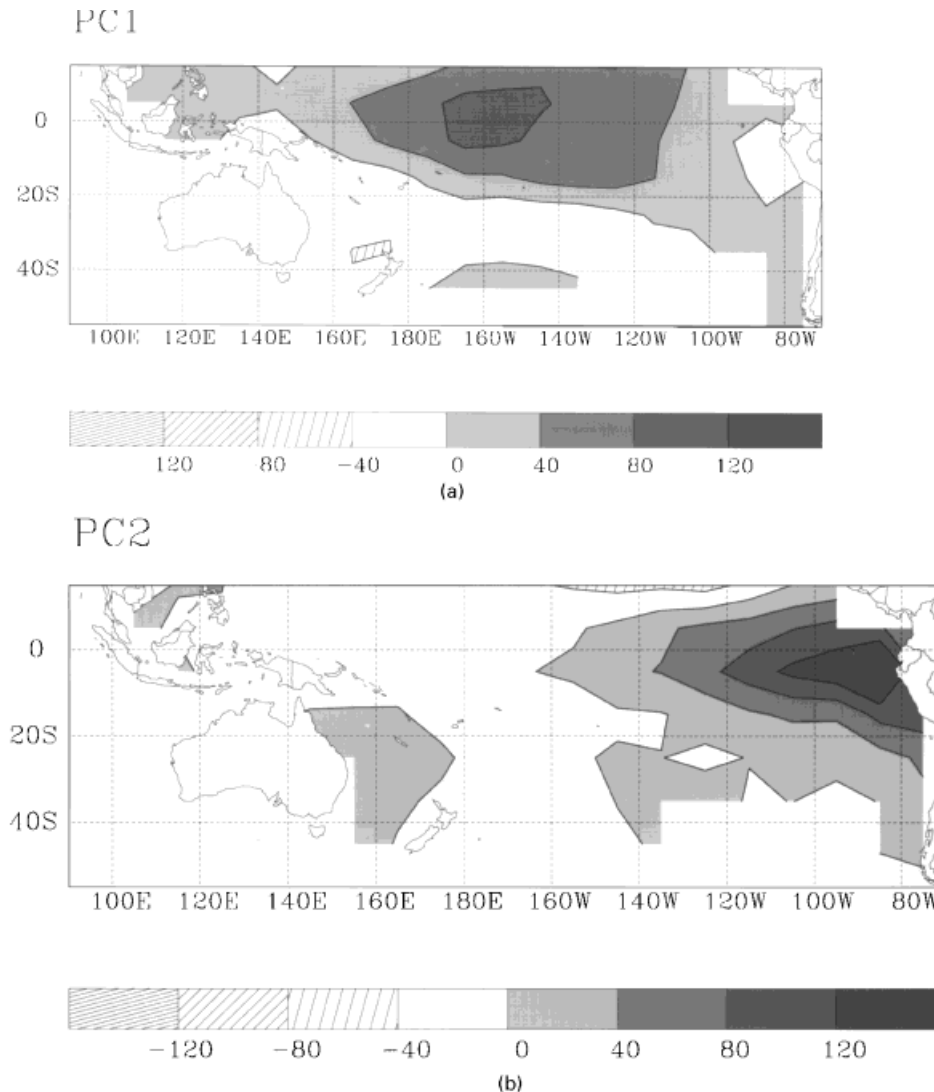
Figure 4. Spatial score coefficients for the first two (a and b) principal components of monthly sea-surface temperature anomalies in the Pacific Ocean south of 20°N for the period 1950–1995. The principal components were calculated from the variance–covariance matrix and were rotated using the varimax method

discriminant analysis (depending on the number of categories used) generally requires a longer training period than regression-based methods, sufficiently long independent testing periods could not be defined given five rainfall categories. Independent cross-validation has therefore been performed for the six-month three-category hindcasts only. For the six-month hindcasts the models were trained over the period 1951–1980 and validated over 1981–1995.

An alternative approach is to perform jack-knife validation, which involves the successive deletion of each observation from the model (Michaelsen, 1987). The deleted observation is then hindcasted and compared with the observed value. The discriminant functions are recalculated at each step so that the model used in the hindcast is constructed without any information about the deleted observation. Recalculation of the discriminant functions is important with small data sets, otherwise the forecast skill, especially of the most extreme values, is exaggerated (van den Dool, 1987; Barnston and van den Dool, 1993). The constraint that only those variables in the full model could be used in the jack-knife models was imposed to avoid year-to-year changes in predictors, which can occur given a large pool of candidate predictors (Hastenrath, 1995). Hindcasts were produced using the

jack-knife method for each of the 45 years. Jack-knife hindcast skill scores were then calculated. The assessment of the hindcast skill of discriminant analysis models is not as straightforward as it is of regression-based techniques. A number of skill-score measures were used and are discussed below.

3. SKILL-SCORE MEASURES

Skill scores were used to express the accuracy of the model over the training period. A number of skill-score measures were calculated and are detailed below.

3.1. Heidke score

The simplest skill measure is the Heidke, or 'hits', score. The score measures the total number of correct forecasts (H) and expresses this as a percentage of the total obtainable with a perfect forecast model (T). The number of hits (H) should be compared with the number expected by chance (E): with five categories a score of 20 per cent should be obtained, with three the *a priori* probability is 33 per cent. The score is often rescaled so that random guessing gives 0 per cent and hence the equation for the score is given as:

$$S = \frac{H - E}{T - E} \times 100 \text{ per cent} \tag{3}$$

The Heidke score is simple to calculate and interpret, but gives no credit for 'near-misses'. For example, given 'very dry' conditions, it would have been better to have forecasted 'dry' than 'very wet', but the Heidke score does not acknowledge this possibility. A more serious problem is that the score is very sensitive to the number of categories used. With more categories, the score usually decreases because with a larger number of categories greater accuracy is required to obtain a perfect hit rather than a near-miss.

3.2. Average distance score

In order to give credit for near-misses a simple extension of the Heidke score can be suggested. The skill of each forecast is determined by the number of categories it differs from the observed forecast. For example, if the forecast (F) is for category 2 ('dry') while the observed category (O) was 3 ('average'), the forecast missed by one. Clearly a perfect forecast has a score of zero and the worst possible score, given five categories, is four. The average score over the validation period can be calculated using

$$S = \frac{\sum_{i=1}^{n} |F_i - O_i|}{n} \tag{4}$$

and can be compared with a perfect score of zero and to the expected score obtained by chance. It can be shown that the expected score obtained by random guessing is 1·60 with five categories and 0·89 with three categories. However, with perpetual forecasts of average, the expected number of hits remains unchanged, but the worst possible score for any individual forecast in a five category system, for example, is 2·00. Consequently, perpetual forecasts of average will give a better average distance score (expected values are 1·20 and 0·67 for five and three categories respectively) without any real improvement in forecast skill. Only models with an average distance score of better than 1·20 should therefore be considered. Significance levels for the average distance score were calculated using 5000 simulations of 45 randomly selected pairs of observed and forecasted categories. The 45 simulated observed categories were constrained to be equi-probable, but the simulated forecasted categories were unconstrained. The 99 per cent significance levels were identified after ranking the 5000 simulated average distance scores and corresponded approximately with scores of 1·22 and 0·64. For the five category system, the 99 per cent significance level for random guessing is higher than the score achievable by perpetual forecasts of average. The stricter score of 1·20 was used as a criterion for minimum acceptable forecast skill.

The absolute distance score is an improvement on the Heidke score and is relatively simple to interpret. It is possible to convert the score to a percentage, but in the form of equation (3) it is probably easier to interpret and

avoids some of the problems encountered by non-scientific readers in understanding skill scores expressed as percentages. The ability to obtain an apparently significant score through perpetual forecasts of average is a limitation.

3.3. Average position

The average position score is calculated by taking the rank of the *a posteriori* probabilities of the categories for which the actual rainfall was observed and averaging over the 45 years. In the event that the rainfall category with the highest *a posteriori* probability was not observed, the score gives credit if the observed category corresponds to the forecasted category with the second highest *a posteriori* probability. The score is of interest in the case of bimodal probabilities, or in the case of a non-extreme forecast when the user wishes to know whether it is more likely for the forthcoming rainfall to be on the wetter or drier side of the actual forecast.

3.4. Average confidence

The average confidence score is calculated by taking the *a posteriori* probabilities of the categories for which the actual rainfall was observed and averaging over the 45 years. The skill score gives credit for correctly forecasting a category with a high degree of confidence and also does not penalize the forecaster heavily in the case of a bimodal *a posteriori* rainfall probability. The *a priori* probability for each category is 0·20 and 0·33 for five and three categories respectively and so a skillful forecast would have to have an average confidence exceeding this.

3.5. LEPS scores

A more sophisticated scoring system has been developed to cater for the shortcomings of the Heidke and average distance scores. As with the average distance score, the LEPS (linear error in probability space) score of the forecast defines error in terms of the 'distance' between the observed value and the forecast (Ward and Folland, 1991; Potts *et al.*, 1996). However, instead of measuring the distance in terms of the number of categories, it is measured in terms of the difference in the cumulative probabilities of the observed and forecast categories. A table of skill coefficients can be constructed and in the case of five equi-probable categories, the scores are presented in Table I. Corresponding scores for three categories are given in Table II. The rescaled LEPS scores (Potts *et al.*, 1996) are used because of the 'bending back' effect that is observed with the standard LEPS scores as defined by Ward and Folland (1991).

If the LEPS score for an individual forecast is defined as S_i, then the score for a set of forecasts can then be calculated using

$$S = \frac{\sum_{i=1}^{n} S_i}{\sum_{i=1}^{n} S_p} \times 100 \text{ per cent} \tag{5}$$

Table I. LEPS score coefficients for five equi-probable categories (from Potts *et al.*, 1996)

Forecast	Observed				
	1	2	3	4	5
1	1·28	0·52	−0·20	−0·68	−0·92
2	0·52	0·56	0·04	−0·44	−0·68
3	−0·20	0·04	0·32	0·04	−0·20
4	−0·68	−0·44	0·04	0·56	0·52
5	−0·92	−0·68	−0·20	0·52	1·28

Table II. LEPS score coefficients for three equi-probable categories (from Potts *et al.*, 1196)

Forecast	Observed		
	1	2	3
1	0·89	−0·11	−0·78
2	−0·11	0·22	−0·11
3	−0·78	−0·11	0·89

where S_p is the score for a perfect forecast. It is evident from Table I that the coefficients in any given row sum to zero. The expected LEPS score for a single forecast is then 0 per cent. More importantly, the expected score for a perpetual forecast of average will also be 0 per cent. Theoretical estimates of the significance of LEPS scores are not currently available (Potts *et al.*, 1996), but given a total of *n* forecasts it is possible to calculate the empirical distribution function of the scores, from which the significance can be obtained. The significance of the LEPS score was calculated in the same way as that of the average distance score, using 5000 simulations. Given 45 independent forecasts, the 99 per cent significance level for the LEPS score is about 30·0 per cent for three categories and 26·7 per cent for five categories.

3.6. Model selection

Ideally, although the atmosphere in the southern African region is known to be associated with sea-surface temperature variability in areas highlighted by the principal components illustrated in Figures 2–4 (Mason and Jury, 1997), the predictors should be selected by theoretical rather than statistical considerations. However, further research into ocean–atmosphere interaction in the region is required before an *a priori* selection can be made. Instead, the vigorous skill-score testing performed should hopefully minimize the probability of selecting a model with high skill scores.

In identifying models with useful forecast skill, only the significance of the LEPS and the average distance score has been considered, although there is a high correlation between all the different skill scores calculated. For a model to be selected both the LEPS and the average distance scores had to exceed the 99 per cent significance level. In practice, because of the similarity of the two scores, it is unusual for a model to achieve significance on only one of the scores. The significance of the LEPS score was considered because of its theoretical superiority, whereas the average distance score was considered because of the fact that the end-user finds it relatively easy to understand. In most cases, more than one skillful model can be defined, often with totally independent predictors. The model with the highest LEPS score has usually been selected, although consideration is also given to the consistency of the predictors of alternative models, to those of the models for neighbouring regions and for adjacent periods. In addition, preference has been given to models with fewer predictors, when the addition of a second or third predictor provides only a minimal improvement in the LEPS score. It is insufficient to focus only on the significance of the LEPS scores because models with high hindcast variance will achieve a higher LEPS score without an improvement in forecast skill (Potts *et al.*, 1996). The other scores and the atypicality index (section 5) have been used only in helping to decide between alternative skillful models with near-equal LEPS scores. When the operational forecasts are produced, all the alternative models are considered and an indication of the level of agreement between the models is expressed. Occasionally, no skillful model can be defined, in which case no forecast is offered.

4. MODEL PERFORMANCE

4.1. Three-month forecast skill

The seasonal variation of the jack-knife-validated LEPS scores for the three-month forecasts in each of the eight regions is shown in Figure 5. For some months, insufficient forecast skill is obtainable to release an
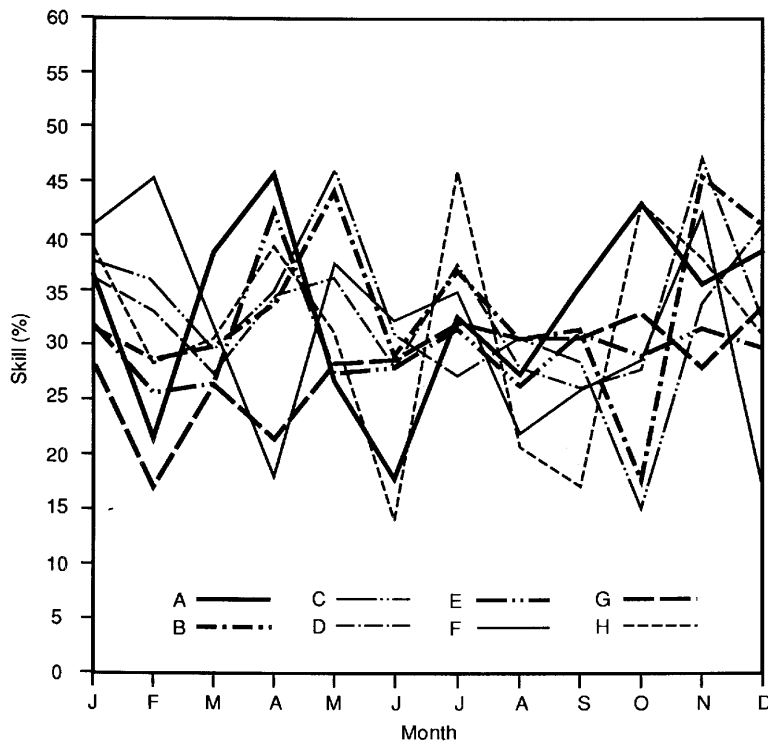
Figure 5. Seasonal variation of the LEPS score for the cross-validated three-month rainfall hindcasts over the training period 1951–1995 for each of the eight rainfall regions

operational forecast, but the highest skill levels obtainable are still shown. There is no obvious seasonal cycle in forecast skill, although peaks are experienced in all regions except for the west coast (region G) during the autumn season in about April or May, when forecasts for the coming boreal winter are released. The LEPS scores never exceed 35 per cent for region G at any time of the year and skillful three-month forecasts can be provided only in austral autumn. The assumption of normally distributed rainfall for this mostly arid area is least valid (Dyer, 1974; Onesta and Verhoef, 1976) and may explain the poor performance of the model here. Over the summer rainfall regions (regions A–E) high forecast skills are obtainable during the late spring or early summer before the main rains and are an indication of the relatively high predictability of the tropical atmosphere (Mason et al., 1996). In almost all cases, forecast skill is achieved by the correct forecasts of the 'very dry' and 'very wet' categories.

The first principal components of all three oceans (Figures 2(a), 3(a) and 4(a)) are the most frequently used predictors. The Pacific Ocean and northern Indian Ocean are important sources of predictability throughout the austral summer season for all regions except the east and west coasts (E and G) and reflect an influence on the tropical atmospheric circulation (Mason and Jury, 1997). The first principal component of the Atlantic Ocean is an important predictor for rainfall over the eastern half of the country. The mechanisms of the association are not well-understood, but probably involve an influence on atmospheric convergence in the Zaire Air Boundary. The general spatial and temporal patterns of forecast skill are consistent with the theoretical levels of higher predictability during the summer and over the summer rainfall region of South Africa (Mason et al., 1996).

4.2. Six-month forecast skill

The seasonal variation of the jack-knife-validated LEPS scores for the six-month forecasts in each of the eight regions is shown in Figure 6. Forecast skill levels are encouragingly high and are in general higher than for the three-month forecasts, presumably because of the smaller number of categories. Over the summer rainfall regions
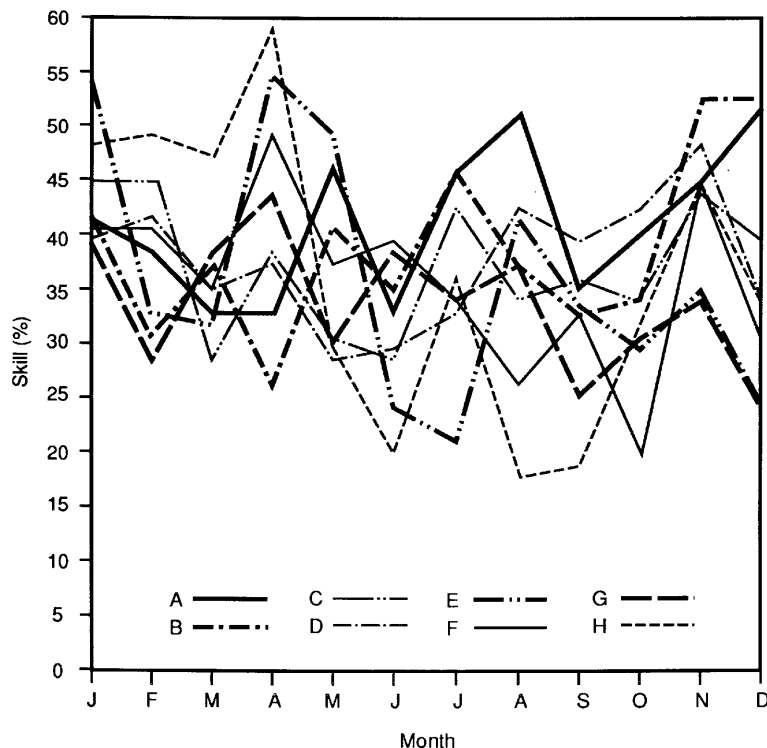
Figure 6. Seasonal variation of the LEPS score for the cross-validated six-month rainfall hindcasts over the training period 1951–1995 for each of the eight rainfall regions

(A–E), highest skill levels are achievable for the six months beginning in early- to mid-summer (November–February), suggesting that the most reliable forecasts of the season's rainfall can be made only shortly after the season has actually started. However, over the former Transvaal (region A) high skill levels are achievable as early as July and August. The Pacific Ocean is the main source of predictability. Along the south and east coasts (regions E, F and H) forecast skill levels peak in April shortly before the winter rains begin. Regions E and H are generally dry during April–September, but region F does receive significant winter rains. Again, in almost all cases, forecast skill is achieved by the correct forecasts of the two extreme categories 'dry' and 'wet'.

The independent validations for the period 1981–1995 are encouragingly high for many of the models (Figure 7), indicating that the identified forecast skill is likely to be real and not a Type I error (a false rejection of the null hypothesis that there is no forecast skill provided by the sea-surface temperature—Wilks, 1995). (Note that the LEPS score coefficients given in Table II for simplicity have not been adjusted to account for the possibility that the observed categories may not be equi-probable.) The distribution of skill scores shows a distinct seasonality in the forecastability of the rainfall. Of greatest interest are high skill scores for the summer rainfall regions A–E exceeding 40 per cent (with the exception of region C) during July and August, providing useful skill before the beginning of the rains. Forecast skill levels for the summer rainfall region peak again in November, after the start of the rains. For the south coast (region F) the winter rains can be forecast successfully from April and May and the summer rains in November. Similarly, the winter rains of the west coast (region G) can be forecast from March to May. Over region H, the forecasts have been consistently good throughout most of the year. The highest Heidke score was for region A, hindcasting for the winter rains of May–October, when perfect scores were achieved for 12 of the 15 years. However, at this time of the year rainfall is minimal and the distributional assumptions are least valid and so the model may be problematic. In general, however, the performance of the scheme over the 15-year independent validation period is encouraging and demonstrates the value of the model as an operation tool.
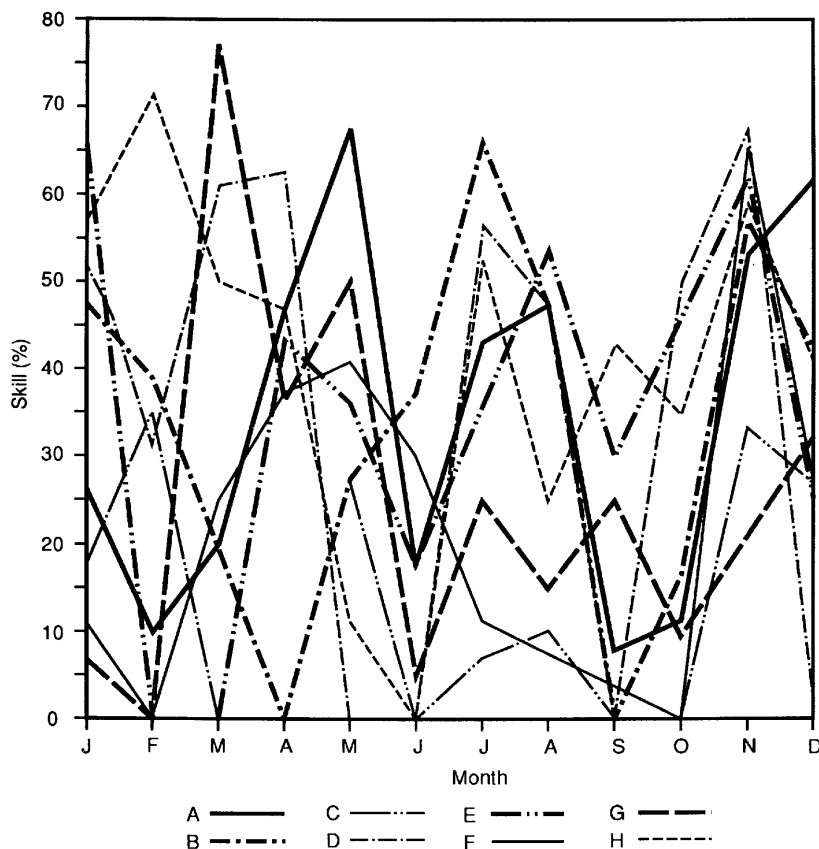
Figure 7. Seasonal variation of the LEPS score for the independently validated six-month rainfall hindcasts for the period 1981–1995 for each of the eight rainfall regions. Negative LEPS scores are indicated as zero, which is the expected score given random forecasts. Note that the vertical scale differs from Figures 6 and 7

5. THE OPERATIONAL FORECASTS

Once the discriminant functions are defined from the data for the training period, a new vector of observations on the independent variables can be submitted and the model will define the *a posteriori* probabilities of each of the categories. In practice, the model defines the probability of three- (six-) month rainfall totals falling within each of the five (three) categories, given the sea-surface temperature fields averaged over the last two months. The model uses the information in the predictor variables to improve on the *a priori* probabilities of 20 per cent. In calculating the *a posteriori* probabilities an atypicality index is also computed that defines the probability of obtaining an observation vector that is more typical of group *I* than the observation vector (Aitchison *et al.*, 1977). The atypicality index provides an indication of the confidence that can be placed in the forecast (Moran and Murphy, 1979) and is a useful statistic in the selection of models when more than one model is available.

For the operational forecasts, near real-time estimates of the time scores on the sea-surface temperature principal components are required. The time scores are calculated from the spatial loadings and the 1° by 1° optimum interpolated (OI) data available from the Climate Prediction (Reynolds and Smith, 1995). The OI data are first averaged into 10° by 10° grids for compatibility with the data from the training period. Although some inconsistencies may result from using a different data set to produce the operational forecasts from that used to train the model, these are likely to be small and the problem is balanced by the fact that the OI data provides global coverage. Once the OI data has been averaged into 10° by 10° grids, principal component scores are calculated for the most recent two months and then averaged. These scores are then used as input into the discriminant analysis model to give *a posteriori* probabilities for the following three- or six-month period. It is these *a posteriori* probabilities that are published in the Seasonal Rainfall Outlooks.

7. CONCLUSIONS

A quadratic discriminant analysis model has been constructed to produce a three- and six-month rainfall forecasting model for South Africa. It provides a useful supplement to the regression-based models used by other seasonal forecasting groups within the South African Long-lead Forecast Forum. The model relates rainfall over eight different areas of South Africa to principal components of sea-surface temperatures in the Indian, South Atlantic and Pacific Oceans. High forecast skill levels can be achieved for much of the country throughout the year and there are indications of useful forecast skill shortly before or soon after the start of the summer rainfall season, which could therefore have significant operational value. Independent validation of the model over the 15-year period 1981–1995 confirms that the model can be used successfully in an operational environment. The independent validation indicates that there is seasonality in the forecastability of rainfall over South Africa, reflecting, in part, the higher forecastability of the tropical atmosphere, which dominates the region during the austral summer. It may be possible to improve on lead-times and skill levels by using forecast sea-surface temperatures in a two-tiered approach. Plans to adapt the model to forecast rainfall in neighbouring countries are being discussed with the respective national meteorological services.

REFERENCES

Aitchison, J., Habbema, J. D. F. and Kay, J. W. 1977. 'A critical comparison of two methods of statistical discrimination', *Appl. Stat.*, **26**, 15–22.
Barnston, A. G. and van den Dool, H. M. 1993. 'A degeneracy in cross-validated skill in regression-based forecasts', *J. Climate*, **6**, 963–977.
Barnston, A. G., Thiao, W. and Kumar, V. 1996. 'Long-lead forecasts of seasonal precipitation in Africa using CCA', *J. Climate*, **11**, 506–520.
Cane, M. A., Eshel, G. and Buckland, R. W. 1994. 'Forecasting Zimbabwean maize yield using eastern equatorial Pacific sea surface temperature', *Nature*, **370**, 204–205.
Cheng, X., Nitsche, G. and Wallace, J. M. 1995. 'Robustness of low-frequency circulation patterns derived from EOF and rotated EOF analyses', *J Climate*, **8**, 1709–1713.
Dix, M. R. and Hunt, B. G. 1995. 'Chaotic influences and the problem of deterministic seasonal predications', *Int. J. Climatol.*, **15**, 729–752.
Dyer, T. G. J. 1974. 'The frequency distribution of annual precipitation of annual precipitation totals for South Africa', *S. Afr. J. Sci.*, **70**, 180–182.
Everitt, B. 1980. *Cluster Analysis*, Halsted, New York, 136 pp.
Gong, X. and Richman, M. B. 1995. 'On the application of cluster analysis to growing season precipitation data in North America east of the Rockies', *J. Climate*, **8**, 897–931.
Harrison, M. S. J. 1984. 'The annual rainfall cycle over the central interior of South Africa', *S. Afr. Geogr. J.*, **66**, 47–64.
Hastenrath, S. 1995. 'Recent advances in tropical climate prediction', *J. Climate*, **8**, 1519–1532.
Janicot, S., Moron, V. and Fontaine, B. 1996. 'Sahel droughts and ENSO dynamics', *Geophys. Res. Lett.*, **23**, 515–518.
Joubert, A. M. 1997. 'Simulations by the Atmospheric Model Intercomparison Project of atmospheric circulation over southern Africa', *Int. J. Climatol.*, **17**, 1129–1154.
Jury, M. R. 1996. 'Regional teleconnection patterns associated with summer rainfall over South Africa, Namibia and Zimbabwe', *Int. J. Climatol.*, **16**, 135–153.
Jury, M. R. and Pathack, B. M. R. 1993. 'Composite climatic patterns associated with extreme modes of summer rainfall over southern Africa: 1975–1984', *Theor. Appl. Climatol.*, **47**, 137–145.
Jury, M. R., McQueen, C. A. and Levey, K. M. 1994. 'SOI and QBO signals in the African region', *Theor. Appl. Climatol.*, **50**, 103–115.
Latif, M., Sterl, A., Assenbaum, M., Junge, M. M. and Maier-Reimer, E. 1994. 'Climate variability in a coupled GCM, Part II: the Indian Ocean and monsoon', *J. Climate*, **7**, 1449–1462.
Lindesay, J. A. 1988. 'South African rainfall, the Southern Hemisphere semi-annual cycle', *J Climatol.*, **8**, 17–30.
Manly, B. F. J. 1994. *Multivariate Statistical Methods: a Primer*, Chapman and Hall, London, 215 pp.
Mason, S. J. 1995. 'Sea-surface temperature – South African rainfall associations, 1910–1989', *Int. J. Climatol.*, **15**, 119–135.
Mason, S. J. 1997. 'Recent developments in seasonal forecasting of rainfall', *WaterSA*, **23**, 57–62.

Mason, S. J. and Jury, M. R. 1997. 'Climatic change and inter-annual variability over southern Africa: a reflection on underlying processes', *Prog. Phys. Geog.*, **21**, 24–50.

Mason, S. J., Lindesay, J. A. and Tyson, P. D. 1994. 'Stimulating drought in southern Africa using sea surface temperature variations', *WaterSA*, **20**, 15–22.

Mason, S. J., Joubert, A. M., Cosijn, C. and Crimp, S. J. 1996. 'Review of the current state of seasonal forecasting techniques for southern Africa. Assessment of the prospects for improving forecast skill and lead-time', *WaterSA*, **22**, 203–209.

Michaelsen, J. 1987. 'Cross-validation in statistical climate forecast models', *J. Clim. Appl. Meteorol.*, **26**, 1589–1600.

Mo, K. C. and Wang, X. L. 1995. 'Sensitivity of the systemic error of extended range forecasts to sea surface temperature anomalies', *J. Climate*, **8**, 1533–1543.

Montgomery, D. C. and Peck, E. A. 1992. *Introduction to Linear Regression Analysis*, Wiley, Chichester, 527 pp.

Moran, M. A. and Murphy, B. J. 1979. 'A closer look at two alternative methods of statistical discrimination', *Appl. Stat.*, **28**, 223–232.

Numerical Algorithms Group. 1996. *NAG Fortran Library Manual, Mark 17*, Numerical Algorithms Group, Oxford.

Onesta, P. A. and Verhoef, P. 1976. 'Annual rainfall frequency distributions for 80 rainfall districts in South Africa', *S. Afr. J. Sci.*, **72**, 120–122.

Parker, D. E. 1987. 'The Meteorological Office historical sea surface temperature data set', *Meteorol. Mag.*, **116**, 250–254.

Parker, D. E. and Folland, C. K. 1988. 'The Meteorological Office historical sea surface temperature data set', in Gregory, S. (ed.), *Recent Climatic Change*', Belhaven Press, London, pp. 41–50.

Potts, J. M., Folland, C. K., Jolliffe, I. T. and Sexton, D. 1996. 'Revised ''LEPS'' scores for assessing climate model simulations and long-range forecasts', *J Climate*, **9**, 34–53.

Reynolds, R. W. and Smith, T. M. 1995. 'A high-resolution global sea surface temperature climatology', *J. Climate*, **8**, 1571–1583.

Richman, M. B. 1986. 'Rotation of principal components', *J. Climatol.*, **6**, 293–335.

Richman, M. B. 1987. 'Rotation of principal components: a reply', *J. Climatol.*, **7**, 511–520.

Richman, M. B. 1993. 'Comments on: ''The effect of domain shape dependence on principal components analysis''', *J. Climatol.*, **13**, 203–218.

Smith, I. N. 1995. 'A GCM simulation of global climate interannual variability: 1950–1988', *J. Climate*, **8**, 709–718.

Stern, W. and Miyakoda, K. 1995. 'Feasibility of seasonal forecasts inferred from multiple GCM simulations', *J. Climate*, **8**, 1071–1085.

Tyson, P. D. 1986. *Climate Change and Variability over Southern Africa*, Oxford University Press, Cape Town, 220 pp.

van den Dool, H. M. 1987. 'A bias in skill in forecasts based on analogues and antilogues', *J Clim. Appl. Meteorol.*, **26**, 1278–1281.

Vogel, C. H. and Drummond, J. 1993. 'Dimensions of drought: South African case studies', *Geojournal*, **30**, 93–98.

Walker, N. D. 1990. 'Links between South African summer rainfall and temperature variability of the Agulhas and Benguela Current Systems', *J. Geophys. Res.*, **95**, 3297–3319.

Ward, M. N. and Folland, C. K. 1991. 'Prediction of seasonal rainfall in the north Nordeste of Brazil using eigenvectors of sea-surface temperatures', *Int. J. Climatol.*, **11**, 711–743.

Wilks, D. S. 1995. *Statistical Methods in the Atmospheric Sciences*, Academic Press, San Diego, 467 pp.