NOTES AND CORRESPONDENCE

Choice of Distance Matrices in Cluster Analysis: Defining Regions

GILLIAN M. MIMMACK

Department of Statistics and Actuarial Science, University of the Witwatersrand, Johannesburg, South Africa

SIMON J. MASON

Scripps Institution of Oceanography, University of California, San Diego, La Jolla, California

JACQUELINE S. GALPIN

Department of Statistics and Actuarial Science, University of the Witwatersrand, Johannesburg, South Africa

26 June 2000 and 6 December 2000

ABSTRACT

Cluster analysis is a technique frequently used in climatology for grouping cases to define classes (synoptic types or climate regimes, for example), or for grouping stations or grid points to define regions. Cluster analysis is based on some form of distance matrix, and the most commonly used metric in the climatological field has been Euclidean distances. Arguments for the use of Euclidean distances are in some ways similar to arguments for using a covariance matrix in principal components analysis: the use of the metric is valid if all data are measured on the same scale. When using Euclidean distances for cluster analysis, however, the additional assumption is made that all the variables are uncorrelated, and this assumption is frequently ignored. Two possible methods of dealing with the correlation between the variables are considered: performing a principal components analysis before calculating Euclidean distances, and calculating Mahalanobis distances using the raw data. Under certain conditions calculating Mahalanobis distances is equivalent to calculating Euclidean distances from the principal components. It is suggested that when cluster analysis is used for defining regions, Mahalanobis distances are inappropriate, and that Euclidean distances should be calculated using the unstandardized principal component scores based on only the major principal components.

1. Introduction

Cluster analysis is used to assign a set of observations into groups or clusters that have similar characteristics as measured by a set of classifying variables (Everitt 1980). When defining climatic regions, for example, the observations typically are a set of stations or grid points, while the variables may be measurements of rainfall and/or temperature over a number of years. The grid points are then grouped into clusters, which hopefully are spatially coherent so that regions can be identified. Examples of using cluster analysis to define climate regions are numerous (e.g., Fovell and Fovell 1993; Gadgil et al. 1993; Jackson and Weinand 1995; Bunkers et al. 1996; DeGaetano 1996; Gerstengarbe et al. 1999).

Alternatively, cluster analysis can be used to identify synoptic types or weather regimes. Here the observations to be grouped into similar types are the various days or months, while the variables could be the grid-point or station measurements of sea level pressure (e.g., Mo and Ghil 1988; Stone 1989; Dorling et al. 1992; Werner and Gerstengarbe 1997). Additional applications of the synoptic-typing approach include the definition of seasons (Alsop 1989), and the identification of sets of distinct forecasts in an ensemble (Brankovic et al. 1990; Ferranti et al. 1994; Molteni et al. 1996; Toth et al. 1997; Atger 1999; Stephenson and Doblas-Reyes, 2000).

There are many possible methods of performing cluster analysis, and the relative merits of the alternative approaches have been reviewed extensively (Fovell and Fovell 1993; Gong and Richman 1995; Jackson and Weinand 1995). Recommendations for clustering methods have concentrated on the choice between hierarchical and nonhierarchical methods, and between the various linkage options, but insufficient attention has been given to the importance of the method of calcu-

Corresponding author address: Dr. Simon J. Mason, Climate Research Division, Scripps Institution of Oceanography, University of California, San Diego, La Jolla, CA 92093-0230.
E-mail: smason@ucsd.edu

lating the distance matrix. Given that all clustering procedures are based on some form of distance measure, the method of calculating this matrix can have an important effect on the clustering results. In this paper, the applicability of the Mahalanobis distance matrix and of the commonly used Euclidean distance matrix are examined in the context of defining climate regions from a single meteorological parameter using hierarchical cluster analysis. Applications using the transposed data matrix are discussed in a companion paper (Mimmack et al. 2000, manuscript submitted to *J. Climate*).

2. Distance measures

Cluster analysis assigns a set of n cases to groups or clusters on the basis of measurements of dissimilarity (or distance) between the various cases, as measured on a set of p variables. The distance measure forms the basis for defining how similar or dissimilar the different cases are. There is no agreement as to the most appropriate distance measure to use (Sokal 1977; Seber 1984), but all have the three properties below. Let d_{ij} denote the distance between points x_i and x_j in the p-dimensional space.

1) Symmetry. The distance from x_i to x_j is the same as the distance from x_j to x_i, that is, $d_{ij} = d_{ji}$.
2) Nonnegativity. Distance is measured as a nonnegative quantity, that is, $d_{ij} \geq 0$.
3) Identification. The distance between x_i and x_i is zero, that is, $d_{ii} = 0$.

It is generally considered desirable for the distance measure to be a metric (Mielke 1985), in which case the measure has the additional properties below (Mardia et al. 1979; Krzanowski 1988).

4) Definiteness. If the distance between x_i and x_j is zero, then x_i and x_j are the same—that is, $d_{ij} = 0$ only if $x_i = x_j$.
5) Triangle inequality. The length of one side of the triangle formed by any three points cannot be greater than the total length of the other two sides—that is, $d_{ij} \leq d_{ik} + d_{jk}$.

It is evident that these properties express characteristics that are fundamental to a measure of distance. Distances that are not metrics have the problem that one can have a zero distance without the points being coincident, and also that a projection of the n points into lower-dimensional space may be problematic (Krzanowski 1988).

a. Euclidean distance

Geometrically the Euclidean distance between two points is the shortest possible distance between the two points. In addition to the five properties above, the Euclidean distance measure is invariant under orthogonal transformations of the variables (rotating the points does not change the distances). Since principal component analysis is just a centering of the data, followed by a rotation of the axes, it follows that the Euclidean distances between principal component scores are the same as those in the original space.

It is because of the many useful properties of the Euclidean metric that by far the majority of cluster analyses in the climatological literature have been based upon Euclidean distances, and recent developments in clustering algorithms predominantly have involved the use of Euclidean distances (e.g., Fovell 1997; Yao 1997; Gerstengarbe et al. 1999). Euclidean squared distances, which are less widely used, do not have the property of triangular inequality and so are considered inappropriate for use in most climatological applications (Mielke 1985, 1987).

One problem with the Euclidean distance measure is that it does not take the correlation between variables into account. Where there are highly correlated variables, these variables measure essentially the same characteristic (Fovell and Fovell 1993). In this situation, Euclidean distance assigns equal weight to each variable, thereby according additional weight to the single characteristic that is measured by the correlated variables. In effect, Euclidean distance gives excess weight to correlated variables (Jolliffe 1986).

b. Mahalanobis distance

A measure that incorporates correlations between variables as well as differences in variances is the Mahalanobis distance. The Mahalanobis distance gives less weight to variables with high variance and to highly correlated variables, so that all characteristics are treated as equally important. The Mahalanobis distance between two points $\mathbf{x}_i$ and $\mathbf{x}_j$ is defined as

$$d_{ij}^2 = (\mathbf{x}_i - \mathbf{x}_j)' \, \mathbf{S}^{-1}(\mathbf{x}_i - \mathbf{x}_j),$$

where $\mathbf{S}$ is the $p \times p$ covariance matrix of $\mathbf{X}$, and $\mathbf{X}$ is assumed to be of full rank so that $\mathbf{S}^{-1}$ exists. Apart from accounting for correlations between variables and for differences in variances, Mahalanobis distances have other attractive properties, such as being related to the log-likelihood of multivariate normal distributions, and to multidimensional scaling (Greenacre and Underhill 1982; Stephenson and Doblas-Reyes 2000). Mahalanobis distances have not explicitly been widely used in climatology, although recently applications have become more frequent (e.g., Stephenson 1997; Remund and Long 1999; Stephenson and Doblas-Reyes 2000).

Clustering of principal components using Euclidean distances is performed frequently, and under certain conditions this is equivalent to using Mahalanobis distances. If all principal components are retained and the principal component scores are standardized, Euclidean distances calculated from the principal component scores are the same as Mahalanobis distances calculated from the original data (Jolliffe 1986; Fovell and Fovell

1993; Stephenson 1997). In this situation, the new variables (the principal components) are uncorrelated and have equal variance, and so each has equal weight in the calculation of Euclidean distances.

If the original data matrix is not of full rank, either because there are more variables than there are observations or because one variable is a linear combination of others, Mahalanobis distance cannot be calculated directly because the covariance matrix is not of full rank (and therefore $\mathbf{S}^{-1}$ does not exist). In these problematic situations, a pseudoinverse can be used, resulting in a truncated Mahalanobis distance (Stephenson 1997) or, equivalently, principal components analysis can be used to reduce the dimensionality of the data and (Euclidean) distances can be calculated using a reduced set of principal axes.

By standardizing the principal component scores the smaller eigenvalues, which are generally related to noise, have an exaggerated effect on the distances. The effects of noise on the distances can be eliminated by dropping some of the principal components (Stephenson 1997). Alternatively, unstandardized principal component scores could be used, although in applications of synoptic typing, Euclidean distances calculated from the standardized principal component scores are less sensitive to problems of uneven station distributions (Karl et al. 1982) and sensitivity to interpolation of data to different grid resolutions (Stephenson 1997). However, the applicability of standardizing the principal components when using Euclidean distances in cluster analysis to define regions is less clear, and is the focus of discussion in this paper.

3. Data and methods

Monthly rainfall totals for the 30-yr period 1961–90 were obtained for 517 stations across South Africa and Lesotho. The distribution of the stations is shown in Fig. 1, which indicates a reasonably even distribution except for the dry northwestern part of South Africa.

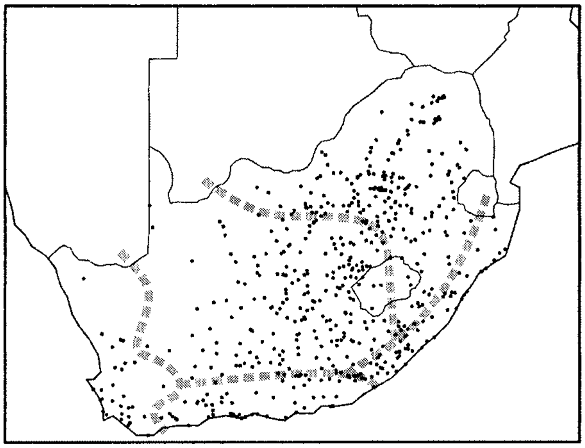

FIG. 1. Distribution of 517 rainfall stations across Lesotho and South Africa. Dotted lines show the approximate distribution of the five seasonal rainfall regimes.

Before clustering the stations, the need for standardization of the monthly rainfall values was examined. Since Euclidean distances between stations with low mean and variance are likely to be small simply because differences in rainfall are small in absolute terms, these stations are likely to be clustered together if they are not standardized, even if the stations are poorly correlated. Rather than standardizing the data by the usual method of subtracting the mean and dividing by the standard deviation, which can be inappropriate for zero-bound data (Wilks 1995), the data were standardized by dividing by the annual mean for that station, as given by

$$x_{ij}^{*} = \frac{x_{ij}}{\frac{12}{p}\sum_{j=1}^{p} x_{ij}}, \qquad (1)$$

where the denominator represents the annual mean rainfall at station i, calculated from the $p = 12 \times 30$ monthly observations. Monthly rainfall is then expressed as a proportion of the mean annual total. In this manner, all stations have an equal mean value (of 1), dry months maintain a value of zero, and the annual cycle is retained. Some differences in variance will remain because of differences in the degree of seasonality in rainfall, and because of differences in the coefficient of variation before standardizing the data. Areas with strongest seasonality and with highest coefficient of variation occur in the driest regions of northwestern South Africa (Tyson 1986). Stations in these areas are likely to be clustered last.

Principal components of the scaled rainfall data were computed, based on the correlation, the covariance, and the cross-products matrices. When using station (or gridpoint) rainfall data to define regions, the variables are the monthly rainfall totals, and the cases are the individual stations. Therefore the standardization as defined in Eq. (1) is across the rows or variables, and the principal components analysis is in the T mode, rather than the more common S mode (Richman 1986). Principal components analyses from the cross-products matrices have not been widely used in climatology (Molteni et al. 1983). Whereas the difference between the use of the correlation matrix and that of the covariance matrix is that the latter accounts for differences in variance, the cross-products matrix accounts for differences in variance *and* differences in means (Jackson 1991). This option can be attractive given noncentered data, and is particularly appropriate for use with precipitation data, which is zero bound. In our context, the cross-products matrix gives greater weight to months with higher spatial mean rainfall and higher spatial variance. The effect is to give added weight to the summer months (for summer rainfall areas) in the definition of the regions.

Clustering was performed using an agglomerative hierarchical approach and Ward's minimum variance method (Gong and Richman 1995). Other linkage methods were tried, but did not give meaningful results. The clustering was performed using Euclidean distances calculated from standardized and unstandardized principal component scores, retaining varying numbers of components. For the sake of ease of comparison of the results, the same number of clusters (five) was retained each time.

4. Examples of regionalization

Examples are shown of attempts to identify rainfall regions over South Africa and Lesotho based on the 360 monthly rainfall values over the period 1961–90. Successful clustering should identify the main seasonal rainfall regimes of South Africa and Lesotho, and should identify important features of interannual variability. The seasonal rainfall regimes of South Africa and Lesotho include the winter rainfall region in the southwest of the country, the all-season rainfall region of the south coast, and the summer rainfall region of the eastern and inland areas (Tyson 1986). The summer rainfall region can be divided into eastern and western halves, with the western half having a later rainfall maximum, and into the eastern coastal strip where there is a definite summer maximum, but also significant rainfall in spring. The approximate delimitation of these five areas is shown in Fig. 1. Because interannual rainfall variability is included in the clustering variables (the 360 monthly rainfall values), the regions identified are likely to reflect a northeast-to-southwest alignment (Mason and Jury 1997).

a. Standardized principal component scores

When all principal components with nonzero eigenvalues are retained and the principal component scores are standardized (i.e., Mahalanobis distances are used) no contiguous regions are defined. An example showing five regions is given in Fig. 2a from the cross-products matrix. The regionalization is based on the occurrence of spatial rainfall patterns that occur throughout the 360 months. By standardizing the principal component scores, all the temporal modes are given equal weight, and so rainfall distribution patterns that occur frequently are treated as equal to unusual patterns and to noise components. Since it makes sense to weight more heavily those distributions that occur most frequently, standardizing the principal component scores is not recommended. In addition, by leaving the principal component scores unstandardized, the lowest-order modes, which define the noise element of the data, are given minimal weighting. Analysis of unstandardized scores is discussed below.

The problems introduced by the exaggeration of the noise component of the data could be eliminated or minimized by retaining the standardized scores of only the first few principal components (i.e., using truncated Mahalanobis distances). Reasonably contiguous regions are defined when a subset of the principal components is retained. Figure 2b, for example, illustrates the regions defined when ten components of the cross-products matrix are retained, which explain 86.3% of the variance. The regions correspond much more closely with those shown in Fig. 1, although the western interior extends to the south coast, and also includes the far northeastern part of the country, and so the regionalization is not entirely satisfactory. In addition, when a subset of the principal components is selected, the defined regions become sensitive to the number of components selected. There is some adjustment of the regional boundaries when one additional component is retained (Fig. 2c). This eleventh component explains only an additional 0.4% of the variance. A sensitivity of the clustering results to the number of principal components retained has been noted elsewhere (e.g., Bunkers et al. 1996; DeGaetano 1996).

Because of the sensitivity of the clustering results to the number of retained principal components, the ‘‘correct’’ number of components to retain needs to be identified. It is important that the variation between the clusters is represented in the direction of at least one of the principal components (Jolliffe 1986), and so it is best to err toward retaining too many principal components rather than too few (Chang 1983). If there are too few components, observations that are not well represented will be clustered together because they have low scores on all the components: important differences between such observations are possibly evident on components that have been dropped. The smaller components are therefore important for assigning observations to clusters that do not otherwise fall obviously into any one cluster. Small, but quite distinct regions may be represented by only lower-order principal component(s) simply because these regions are represented by only a few stations or grid points. The importance of retaining sufficient components is illustrated in Fig. 2d, in which only three principal components have been retained (explaining 79.0% of the variance). The clustering fails to distinguish between the eastern and southern coastal regions, and breaks the northeastern region into discontinuous zones. On the other hand, inclusion of too many principal components inflates the importance of noise, and results in poorly defined regions, as illustrated in Fig. 2a.

The problems of random clustering when all principal components are retained, of sensitivity to the number of principal components retained, and of the need to retain sufficient components, all occur whether the principal components analysis is based on the correlation, covariance, or cross-products matrix. Although there are clear differences in results depending on which matrix is used, the sensitivity to the number of principal components retained appears to be at least as important as which matrix is used. Regional definitions obtained by

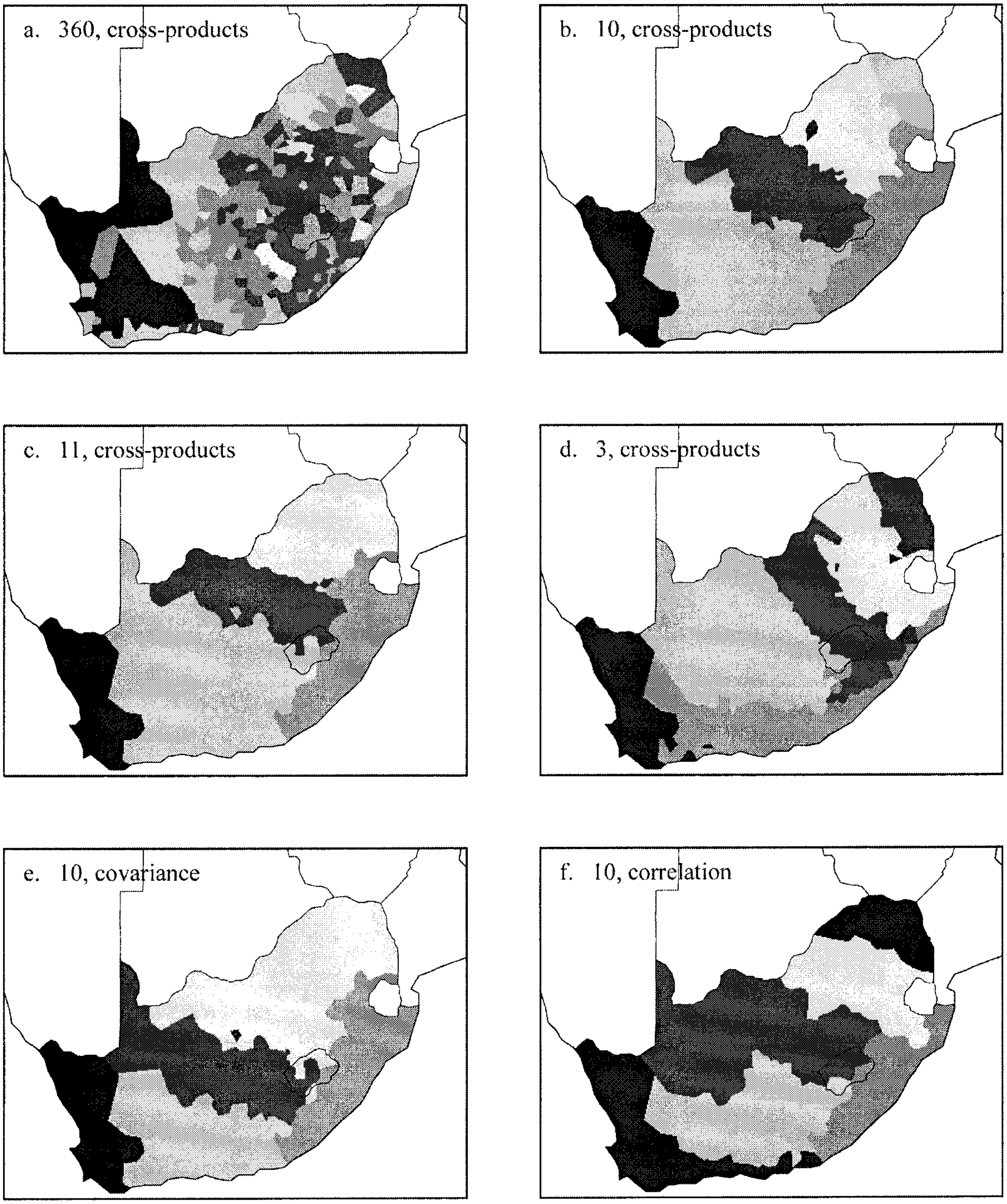

FIG. 2. Rainfall regions over Lesotho and South Africa, defined by hierarchical clustering of standardized principal component scores, using the first (a) 360, (b) 10, (c) 11, and (d) 3 principal components of the cross-products matrix of monthly rainfall, and using the first 10 principal components of the (e) covariance and (f) correlation matrices.

retaining the first 10 principal components of the covariance and correlation matrices are shown in Figs. 2e and 2f, respectively. In both cases, these 10 components explain approximately 60% of the variance. For the covariance matrix (Fig. 2e) the west coast region is identical to that defined using the cross-products matrix (Fig. 2a), and the east coast region is very similar. For the correlation matrix (Fig. 2f), problems resulting from an insufficient number of retained principal components appear to occur: there is an inability to distinguish between the rainfall regimes of the south and west coasts, and the far northeastern part of the country is grouped with a different part of the country (cf. Figs. 2b and 2d).

b. Unstandardized principal component scores

If the principal components are left unstandardized, and all principal components with nonzero eigenvalues are retained (i.e., Euclidean distances are calculated using the raw station data), the clustering of the rainfall stations results in contiguous regions (Fig. 3a) that reflect differences in the climatology of the region (Fig. 1; Tyson 1986). Results obtained with the correlation and covariance matrices are almost identical. Because the combined effect of the principal components is indicated by the variance explained, it is evident that the lower-order components have minimal effect on the distance matrix, and so the regional definitions are not dominated by noise components in the original data. For similar reasons, clustering results are not as sensitive to the number of principal components retained when the component scores are left unstandardized compared to when they are standardized. In Figs. 3b and 3c, for example, regions are shown when 10 and 11 principal components are retained, respectively. The adjustment in the regional boundaries are considerably less than for the standardized principal components (Figs. 2b and 2c), and the regions are similar to those defined by retaining all 360 components (Fig. 3a).

If only a few components are retained (Fig. 3d), the regional definitions remain similar to those defined using alternative numbers of retained principal components. However, the definition of the east coastal region becomes robust only after at least seven principal components are retained. As with the standardized principal components, therefore, it is important to retain a sufficient number of components.

Differences between the clustering results based on the correlation, covariance, and cross-products matrices are relatively minor compared with the differences between standardized and unstandardized principal component scores. Results for the covariance and correlation matrices when 10 principal components are retained (approximately 60% of the variance) are shown in Figs. 3e and 3f, respectively. In both cases, the west coast region is identical to that defined when using the cross-products matrix (Fig. 3b), while adjustments to the south and east coasts are relatively small. The similarity of the results shown in Figs. 3b,e and 3f is largely an effect of the method of standardization of the station data prior to calculating the principal components. Differences become more pronounced if the data are not standardized by station or grid point.

5. Conclusions and recommendations

Euclidean distance–based cluster analysis is frequently used to define climate regions. The clustering of grid-point or station data to define regions can be highly sensitive to the distance measure used, and so careful attention needs to be given to the appropriate method of calculating the distance matrix, in addition to the choice of the clustering algorithm used. When using cluster analysis to define regions based on a single meteorological parameter measured over a number of time steps, the following recommendations are made. It should be emphasized that these recommendations may not be appropriate if regions are being defined using more than one meteorological parameter (e.g., temperature as well as rainfall).

1) The principal component scores should not be standardized. Standardizing the principal component scores has the effect of weakening the influence of climate/synoptic regimes that occur frequently, and of inflating the influence of unusual patterns and noise elements represented by the lower-order components. If the principal component scores are standardized, the clustering results become sensitive to the number of components retained, and the results become dominated by noise components if all the components are retained (Fovell and Fovell 1993). Given that Euclidean distances calculated from standardized principal component scores are equivalent to Mahalanobis distances (including truncated Mahalanobis distances when a subset of the principal components is retained), it is concluded that Mahalanobis distances are inappropriate for defining climate regions from a single meteorological parameter.
2) When prefiltering the data using principal components analysis, the choice of the distance measure appears to be much more important than deciding whether the principal components analysis should be based on the correlation, covariance, or cross-products matrix. However, if the data is not standardized by station or grid point, the choice of matrix in the principal component analysis may become more important, and so this question should not be ignored.
3) Although, it is computationally inefficient to retain all principal components, it is best to err on the side of retaining too many components, otherwise stations that are significantly differentiated only on a lower-order component may be clustered together. Given that the component scores are unstandardized, the inclusion of the lower-order components should have little effect on the distance matrix.

In making the above recommendations, it is not being advocated that the use of clustering of principal component scores is the correct, or even necessarily the preferred, method of regionalization. The disadvantages and merits of alternative approaches to defining regions have not been considered in this paper. Instead, the objective is only to provide some guidelines to those researchers who elect to define regions by using cluster analysis. In addition, the recommendations not to use Mahalanobis or truncated Mahalanobis distances apply only in the context of defining regions; this distance

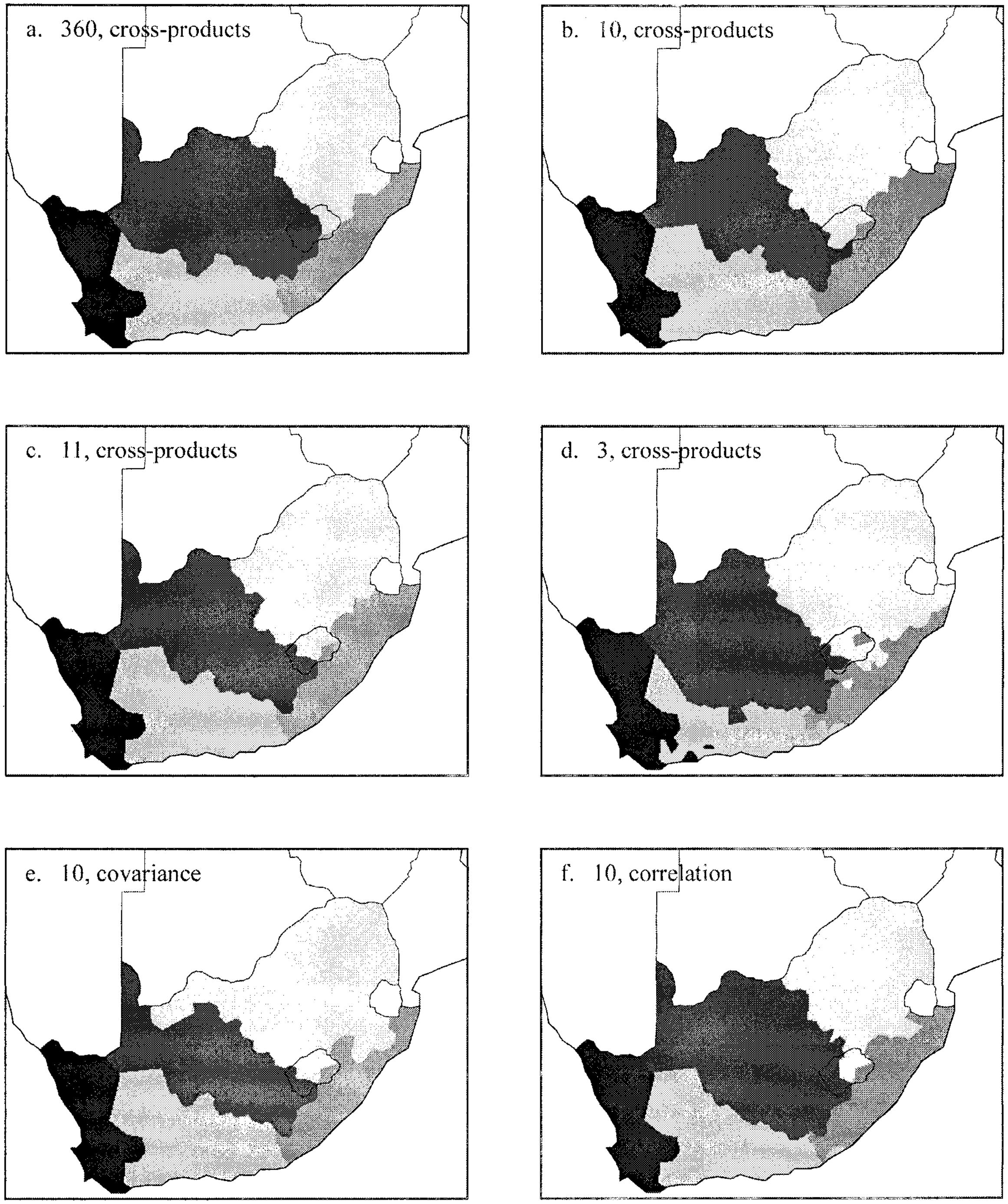

FIG. 3. Rainfall regions over Lesotho and South Africa, defined by hierarchical clustering of unstandardized principal component scores, using the first (a) 360, (b) 10, (c) 11, and (d) 3 principal components of the cross-products matrix of monthly rainfall, and using the first 10 principal components of the (e) covariance and (f) correlation matrices.

metric has been shown to be highly suitable in the identification of synoptic types and climate regimes (Stephenson 1997; Stephenson and Doblas-Reyes 2000).

Acknowledgments. This paper was funded in part by the Scripps Institution of Oceanography. Initial funding for the research was provided in part by Technology Research and Investigations, a division of the Technology Group of Eskom. Helpful discussions with D. Bradu, I. Abramson, and L. Goldstein are gratefully acknowledged. The comments of two anonymous referees contributed to a notable improvement to this paper. The rainfall data were obtained from the South African Weather Bureau and Lesotho Meteorological Services.

REFERENCES

Alsop, T. J., 1989: The natural seasons of western Oregon and Washington. *J. Climate,* **2,** 888–896.

Atger, F., 1999: Tubing: An alternative to clustering for the classification of ensemble forecasts. *Wea. Forecasting,* **14,** 741–757.

Brankovic, C., T. N. Palmer, F. Molteni, S. Tibaldi, and U. Cubasch, 1990: Extended-range predictions with ECMWF models: Time-lagged ensemble forecasting. *Quart. J. Roy. Meteor. Soc.,* **116,** 867–912.

Bunkers, M. J., J. R. Miller, and A. T. DeGaetano, 1996: Definition of climate regions in the Northern Plains using an objective cluster modification technique. *J. Climate,* **9,** 130–146.

Chang, W.-C., 1983: On using principal components before separating a mixture of two multivariate normal populations. *J. Appl. Stat.,* **32,** 267–275.

DeGaetano, A. T., 1996: Delineation of mesoscale climate zones in the northeastern United States using a novel approach to cluster analysis. *J. Climate,* **9,** 1765–1782.

Dorling, S. R., T. D. Davies, and C. E. Pierce, 1992: Cluster analysis: A technique for estimating the synoptic meteorological controls on air and precipitation chemistry—Method and applications. *Atmos. Environ.,* **26A,** 2575–2581.

Everitt, B. S., 1980: *Cluster Analysis.* Halsted, 136 pp.

Ferranti, L., F. Molteni, C. Brankovic, and T. N. Palmer, 1994: Diagnosis of extratropical variability in seasonal integrations of the ECMWF model. *J. Climate,* **7,** 849–868.

Fovell, R. G., 1997: Consensus clustering of U.S. temperature and precipitation data. *J. Climate,* **10,** 1405–1427.

——, and M. C. Fovell, 1993: Climate zones of the conterminous United States defined using cluster analysis. *J. Climate,* **6,** 2103–2135.

Gadgil, S., Yadumani, and N. V. Joshi, 1993: Coherent rainfall zones of the Indian region. *Int. J. Climatol.,* **13,** 547–566.

Gerstengarbe, F. W., P. C. Werner, and K. Fraedrich, 1999: Applying non-hierarchical cluster analysis to climate classification: Some problems and their solutions. *Theor. Appl. Climatol.,* **64,** 143–150.

Gong, X., and M. B. Richman, 1995: On the application of cluster analysis to growing season precipitation data in North America east of the Rockies. *J. Climate,* **8,** 897–931.

Greenacre, M. J., and L. G. Underhill, 1982: Scaling a data matrix in a low-dimensional Euclidean space. *Topics in Applied Multivariate Analysis,* D. M. Hawkins, Ed., Cambridge University Press, 183–266.

Jackson, I. J., and H. Weinand, 1995: Classification of tropical rainfall stations: A comparison of clustering techniques. *Int. J. Climatol.,* **15,** 985–994.

Jackson, J. E., 1991: *A User's Guide to Principal Components.* Wiley, 569 pp.

Jolliffe, I. T., 1986: *Principal Component Analysis.* Springer-Verlag, 271 pp.

Karl, T. R., A. J. Koscielny, and H. F. Diaz, 1982: Potential errors in the application of principal component (eigenvector) analysis to geophysical data. *J. Appl. Meteor.,* **21,** 1183–1186.

Krzanowski, W. J., 1988: *Principles of Multivariate Analysis: A User's Perspective.* Oxford Science Publications, 563 pp.

Mardia, K. V., J. T. Kent, and J. M. Bibby, 1979: *Multivariate Analysis.* Academic Press, 521 pp.

Mason, S. J., and M. R. Jury, 1997: Climate variability and change over southern Africa: A reflection on underlying processes. *Prog. Phys. Geogr.,* **21,** 23–50.

Mielke, P. W., 1985: Geometric concerns pertaining to applications of statistical tests in the atmospheric sciences. *J. Atmos. Sci.,* **42,** 1209–1212.

——, 1987: L_1, L_2 and L_∞ regression models: Is there a difference? *J. Stat. Planning and Inference,* **13,** 430.

Mo, K., and M. Ghil, 1988: Cluster analysis of multiple planetary flow regimes. *J. Geophys. Res.,* **93,** 10 927–10 952.

Molteni, F., P. Bonelli, and P. Bacci, 1983: Precipitation over Northern Italy: A description by means of principal components analysis. *J. Climate Appl. Meteor.,* **22,** 1738–1752.

——, R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The ECMWF Ensemble Prediction System: Methodology and validation. *Quart. J. Roy. Meteor. Soc.,* **122,** 73–119.

Remund, Q. P., and D. G. Long, 1999: Sea ice extent mapping using Ku band scatterometer data. *J. Geophys. Res.,* **104,** 11 515–11 527.

Richman, M. B., 1996: Rotation of principal components. *J. Climatol.,* **6,** 293–335.

Seber, G. A. F., 1984: *Multivariate Observations.* Wiley, 686 pp.

Sokal, R. R., 1977: Clustering and classification: Background and current directions. *Classification and Clustering,* J. van Rysin, Ed., Academic Press, 155–173.

Stephenson, D. B., 1997: Correlation of spatial climate/weather maps and the advantages of using the Mahalnobis metric in predictions. *Tellus,* **49A,** 513–527.

——, and F. J. Doblas-Reyes, 2000: Statistical methods for interpreting Monte Carlo ensemble forecasts. *Tellus,* **52A,** 300–322.

Stone, R., 1989: Weather types at Brisbane, Queensland: An example of the use of principal components and cluster analysis. *Int. J. Climatol.,* **9,** 3–32.

Toth, Z., E. Kalnay, S. M. Tracton, R. Wobus, and J. Irwin, 1997: A synoptic evaluation of the NCEP ensemble. *Wea. Forecasting,* **12,** 140–153.

Tyson, P. D., 1986: *Climatic Change and Variability over Southern Africa.* Oxford University Press, 220 pp.

Werner, P. C., and F. W. Gerstengarbe, 1997: Proposal for the development of climate scenarios. *Climate Res.,* **8,** 171–182.

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

Yao, C. S., 1997: A new method of cluster analysis for numerical classification of climate. *Theor. Appl. Climatol.,* **57,** 111–118.