Defining Homogeneous Precipitation Regions by Means of Principal Components Analysis

DIRK MALLANTS AND JAN FEYEN

Laboratory of Land Management, Faculty of Agricultural Sciences, Catholic University of Leuven, Leuven, Belgium

(Manuscript received 15 October 1989, in final form 2 March 1990)

ABSTRACT

The spatial patterns of precipitation over the IJzer watershed (in western Belgium and northern France) are investigated for three years: 1973 (dry), 1977 (wet), and 1978 (average). Analyses were performed using daily precipitation data for the three years together and for the three years separately. The first principal component explains most of the variance (about 75%) and is uniformly distributed over the whole region. Higher-order components delineate subregions of consistent rainfall. No essential difference in pattern in the three years occurs between the first four components. The components patterns seem to express maritime and topographic effects, and form a basis for dividing the IJzer catchment into four coherent subregions. For representative stations inside each subregion, cross correlations with the remaining stations over the entire watershed indicate that at a level of $r = 0.85$ three stations define three homogeneous precipitation regions. The result is three representative time series of daily rainfall, one for each region.

1. Introduction

This paper describes the temporal and spatial variability of rainfall series over the IJzer watershed (1112 km^2). It is the aim to define homogeneous precipitation areas in order to reduce the number of rainfall series that have to be taken into account for the field water balance model SWATRER (Dierickx et al. 1986). In the SWATRER model, daily precipitation, irrigation, and interception are considered as input. Estimation of the remaining terms of the water balance, such as the capillary rise and the percolation at the bottom of the soil profile, are made from hydraulic properties of the soil in combination with meteorological data. In order to calculate the water movement toward the groundwater table in the IJzer watershed, the point model SWATRER has to be integrated over the whole watershed. For each climatologic parameter and soil parameter different homogeneous regions have to be defined. Therefore, the model has to be implemented as many times as there are such regions. Reduction of total computer time is achieved by reducing the total number of regions for which the SWATRER model has to be implemented.

2. Data collection

Daily precipitation data were collected for 11 stations and served as a basic data network. The station histories, whether rainfall data over different periods are comparable or not, were inspected for changes of location or procedure (the so-called homogeneity test) using the residual mass curve (Buishand 1982). Annual precipitation principal components patterns were calculated for the three years together (1973, 1977, and 1978) and for the separate years. The stations used are shown in Fig. 1 and listed in Table 1. Only four stations are inside the IJzer watershed; the remaining were included to counteract possible boundary effects. Because of minor topographical variations in the watershed that covers more than 1100 km^2, these four stations provide a fairly dense network for further investigations.

3. Principal components analysis (PCA)

Principal components analysis (PCA) is used to simplify the original data by representing the same objects (observations) in fewer than the original number of variables. It is also referred to as eigenvector analysis (Kutzbach 1967; Gray 1981), empirical orthogonal functions (Lorenz 1956; Gilman 1957), and singular decomposition (Rasmusson et al. 1981). These so-called reduced space techniques attempt to find a smaller number of dimensions (variables) that contain most of the information in the original space (Green and Caroll 1978; Jolliffe 1986). Principal components try to find linear combinations of the original variables. The most important properties of these combinations are 1) the factor scores have maximal variance, and 2) the combinations are uncorrelated with previously computed combinations. These linear combinations can be found by rotation of the initial configuration of

Corresponding author address: Dr. Jan Feyen, Laboratorium voor Landbeheer, Katholieke Universiteit Leuven, Kardinaal Mercierlaan 92, B-3030 Leuven, Belgium.

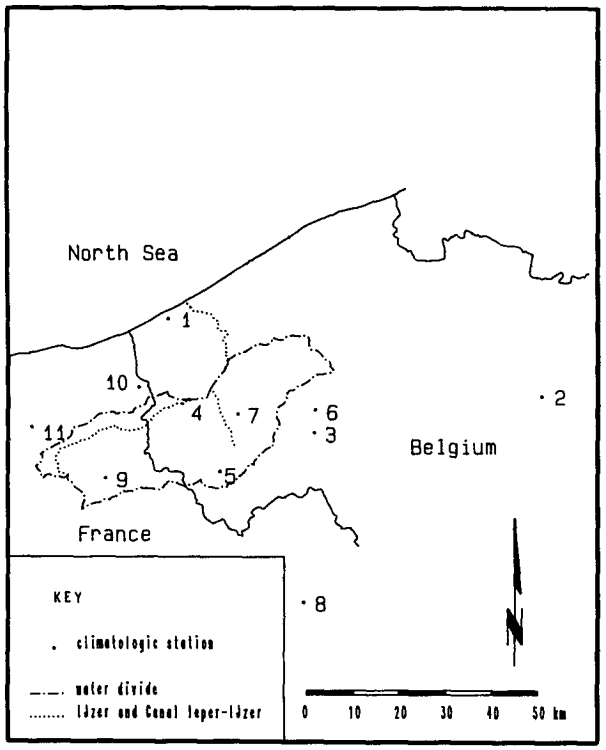

FIG. 1. Situation map of the stations used in the principal components analysis.

points (observations) to a new orientation of the same dimension. This new orientation displays mutually orthogonal dimensions with sequentially maximal variance. The first dimension (first principal component) exhibits the largest variance of points projections. The second dimension (second principal component) exhibits the next largest variance and is orthogonal to the first, and so on.

PCA can be performed using a covariance or a correlation matrix. The former puts greater weight on the most variable sites; the principal component scores have almost the same pattern as the standard deviations of annual precipitation. The latter gives equal weight to each station: PCA are calculated using standardized variables. In this study the correlation matrix was used.

For each station the 1- and 3-year mean was calculated and subtracted from each daily value giving a dataset of daily departures. These departures were in turn standardized; i.e., normalized, by dividing each value by their standard deviation.

Principal components calculations display two sets of numbers: 1) component scores, and 2) component loadings. The matrix of unstandardized component scores Z can be obtained as

$$Z = X_s U \tag{1}$$

where X_s is the standardized data matrix, U is an orthogonal ($U'U = UU' = I$) matrix of eigenvectors of R, the correlation matrix. The matrix Z_s of standardized (to unit variance) component scores is computed by

$$Z_s = X_s U D^{-1/2} \tag{2}$$

where D is the diagonal matrix of (ordered) eigenvalues of R. The unstandardized components loadings matrix F is obtained from

$$F = UD^{1/2}. \tag{3}$$

By "loading" we mean the correlation of some original variable X_{si} with a principal component.

4. Results of principal components analysis

Analyses were carried out using daily precipitation data for the three years together and are shown in Table 2 and for each of the years alone (see Table 3). The first four component patterns are shown in Figs. 2a–d for the combined data and in 3a,b, 4a,b and 5a,b for the individual years.

For the three years together the first four components explained about 90% of the variance. The remaining

TABLE 1. List of stations used for principal components analysis.

No.	Name	Lat	Long	Elevation (m)
1	Koksijde	51°7′0″N	2°40′0″E	5
2	Melle	50°58′35″N	3°49′30″E	17
3	Beitem	50°53′56″N	3°7′26″E	28
4	Lo	50°57′28″N	2°44′10″E	6
5	Vlamertinge	50°49′8″N	2°50′12″E	17
6	Roeselare	50°56′38″N	3°7′37″E	18
7	Boezinge	50°53′30″N	2°52′5″E	17
8	Lille	50°34′0″N	3°6′0″E	49
9	Cassel	50°47′30″N	2°30′54″E	139
10	Hondschoote	50°58′49″N	2°34′45″E	6
11	Merckeghem	50°51′40″N	2°17′45″E	4

TABLE 2. Standardized (to unit variance) factor loadings for the first four components, eigenvalue (λi) and variance explained (1973, 1977, 1978).

Station	PC1	PC2	PC3	PC4
1	0.303	−0.344	−0.099	0.153
2	0.256	0.493	−0.320	0.673
3	0.312	0.250	−0.295	−0.258
4	0.320	−0.114	0.040	−0.228
5	0.313	0.117	0.251	−0.226
6	0.308	0.243	−0.395	−0.260
7	0.326	0.068	0.057	−0.259
8	0.270	0.335	0.742	0.223
9	0.310	−0.170	0.145	−0.065
10	0.315	−0.312	−0.052	0.062
11	0.274	−0.496	−0.050	0.403
λi	8.233	0.822	0.443	0.406
Cumulative variance accounted for	0.748	0.823	0.863	0.900

seven components explained only 10% of the variance. The first principal component (PC1) shows variations (Fig. 2a) affecting the region as a whole, with maximum loadings at the center of the watershed, and hence represents rainfall events common to all stations; i.e., all places wet or all places dry. All variables exhibit positive correlations with the first principal component. The correlation coefficient between the precipitation at any point and the principal component is obtained from the product of the factor loading at that point and the square root of the eigenvalue. In the dry year (1973) the first four components again account for 90% of variance (Table 3). The same results are found in factor loadings of the wet (1977) and average (1978) year. In both years the first four components explain about 90% of variance in the original data.

The second principal component (PC2) shows variations (Fig. 2b) west and east of a zero line extending from the northern part of the watershed to the southern part, with largest anomalies in the eastern part. This zero line can be interpreted as dividing the stations into two groups: the first group (stations 1, 4, 9, 10 and 11) has a strong marine influence while the second group (stations 2, 3, 5, 6, 7 and 8) has a reduced marine (or an increased continental) influence. About 7.5% of variance is explained by PC2. Similar components loadings can be found for 1973 and 1978 (see Figs. 3a and 5a). In the wet year (1977, Fig. 4a), however, the negative sign of the loadings has shifted from the left of the zero line to the right, indicating that in wet years the precipitation at stations in the eastern part of the watershed is negatively correlated with the second component. This suggests that in a wet year the marine influence causes the area between the coast and the zero line to receive higher rainfall amounts than the area to the right of that line.

For PC3 the zero line (Fig. 2c) separates the north of the watershed from the south. This division is due to some topographic effects, the southern part being the highest part of the watershed. The more extreme years 1973 (dry) and 1977 (wet) display a similar north–south division (Figs. 3b and 4b); the negative loadings have shifted again from above the zero line for 1978 (Fig. 5b) to below this line.

The PC4 pattern (Fig. 2d) separates the southwestern part from the remaining part. It describes smaller-scale variations in rainfall. However, the reason for this pattern is not so clear. Almost the same component pattern appears for the average year (1978). The dry and wet year patterns also show small-scale variability explaining less than 4% of the variance in the original data (see Table 2).

Average years (1978) have more variance (80%) associated with the first principal component than dry (75%) and wet (72%) years, and more variance is associated with the second, third, and fourth components in wet and dry years. This suggests that spatial coherence of rainfall is greatest in an average year and least in wet and dry years.

Another way to investigate the relationships among variables is shown in Fig. 6a,b. Each variable (station) can be represented by factor loadings for two principal components (the so-called dimensions). In Fig. 6a all variables are plotted with respect to the first (PC1) and second (PC2) component. All variables are positively correlated with the first component. Stations S2, S3, S5, S6, S7 and S8 are positively correlated with the second component, whereas stations S1, S4, S9, S10 and S11 are negatively correlated with PC2. To obtain the correlation coefficient between the principal component and the variables, each variable has to be multiplied by the square root of the eigenvalue of that principal component.

Variables that point in the same direction are highly correlated whereas variables pointing in the opposite direction have negative correlations.

With respect to PC1 and PC2, we can state that stations S2 and S11, which lie 110 km apart, have low correlations: S2 is located at the east and S11 at the west side of the region (Fig. 1). The same holds for S8 and S1, which are located 70 km apart, S1 on the coast and S8 far inland, and S10. Stations S1 and S10 are at the northern side of the watershed, whereas S8 is situated in the south of the region. The remaining stations in Fig. 6a point in the same direction, and thus are highly positively correlated. These stations are situated in the central part of the watershed. Inside the watershed two different groups can be distinguished: S5 and S7, S4 and S9. Within each group the correlation between variables is very high. These relations serve as an additional indication for defining homogeneous precipitation areas.

As mentioned before, the higher-order principal components delineate small scale variations in rainfall. Figure 6b shows a better partitioning of the stations into groups with respect to PC2 and PC3. Four major groups, one in each quadrant, are distinguished. The first group (S5-S7-S8) is positively correlated with both PC2 and PC3. Stations S5 and S8 are situated in the southern part of the region and S7 in the central part.

TABLE 3. Cumulative variance accounted for and eigenvalue (between parentheses) for the first four components.

Year	PC1	PC2	PC3	PC4
1973	0.747 (8.215)	0.823 (0.838)	0.865 (0.462)	0.901 (0.393)
1977	0.717 (7.887)	0.814 (1.069)	0.862 (0.533)	0.902 (0.430)
1978	0.800 (8.805)	0.859 (0.639)	0.895 (0.400)	0.919 (0.264)

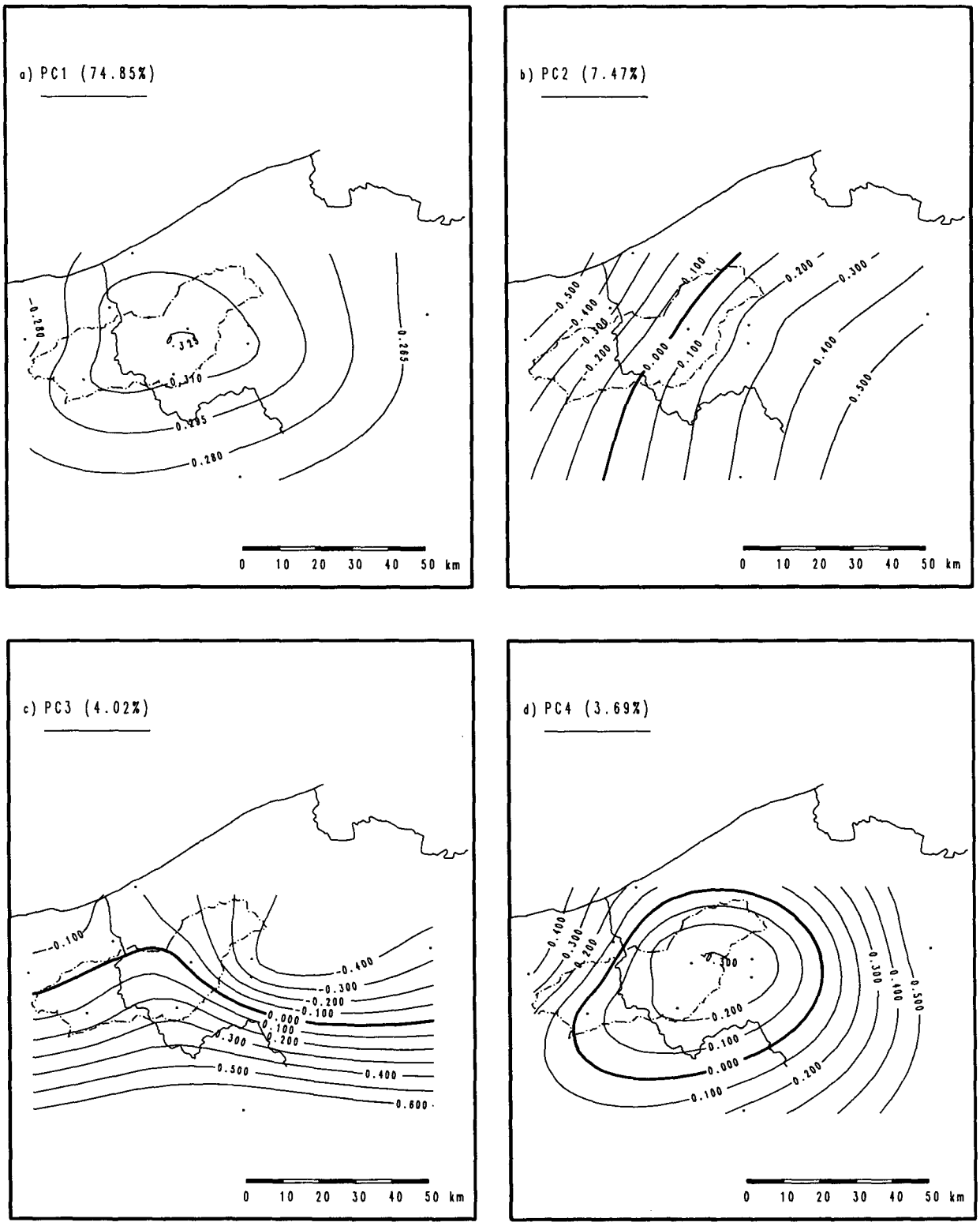

FIG. 2. Principal component patterns for the combined three years of daily data.

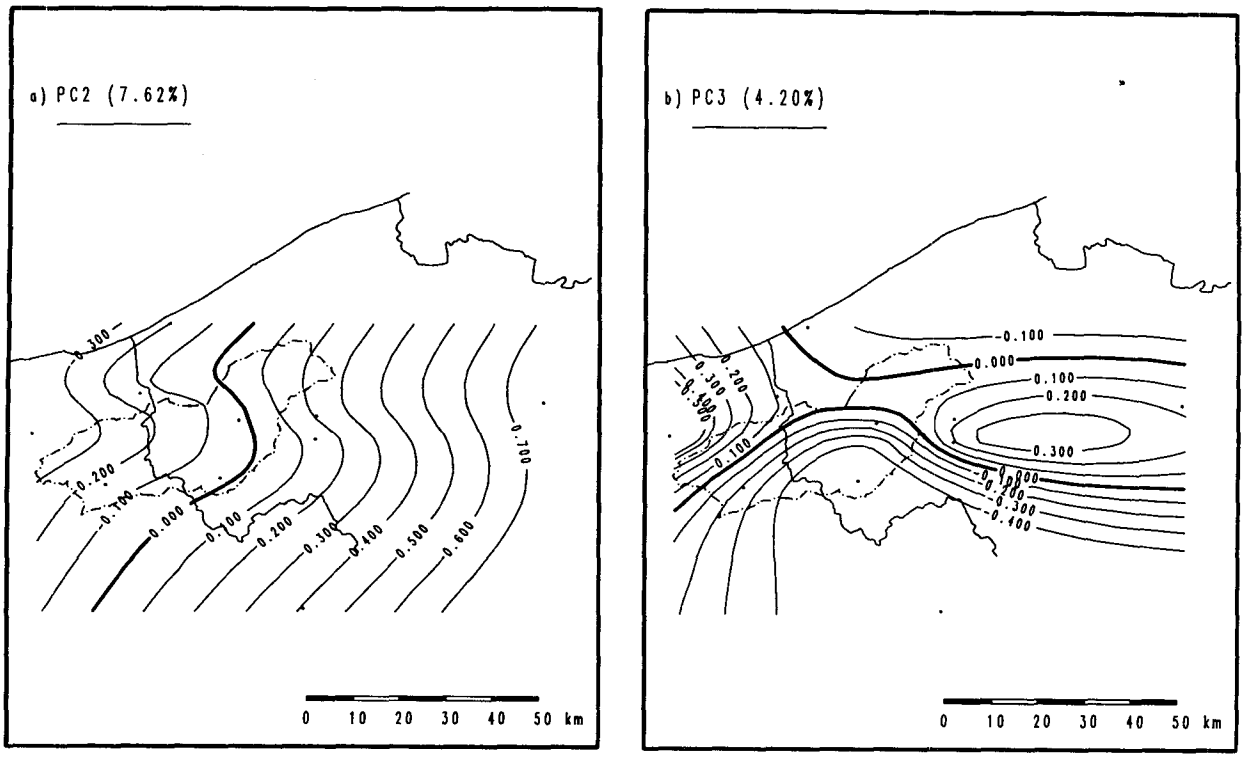

FIG. 3. Principal component patterns in 1973.

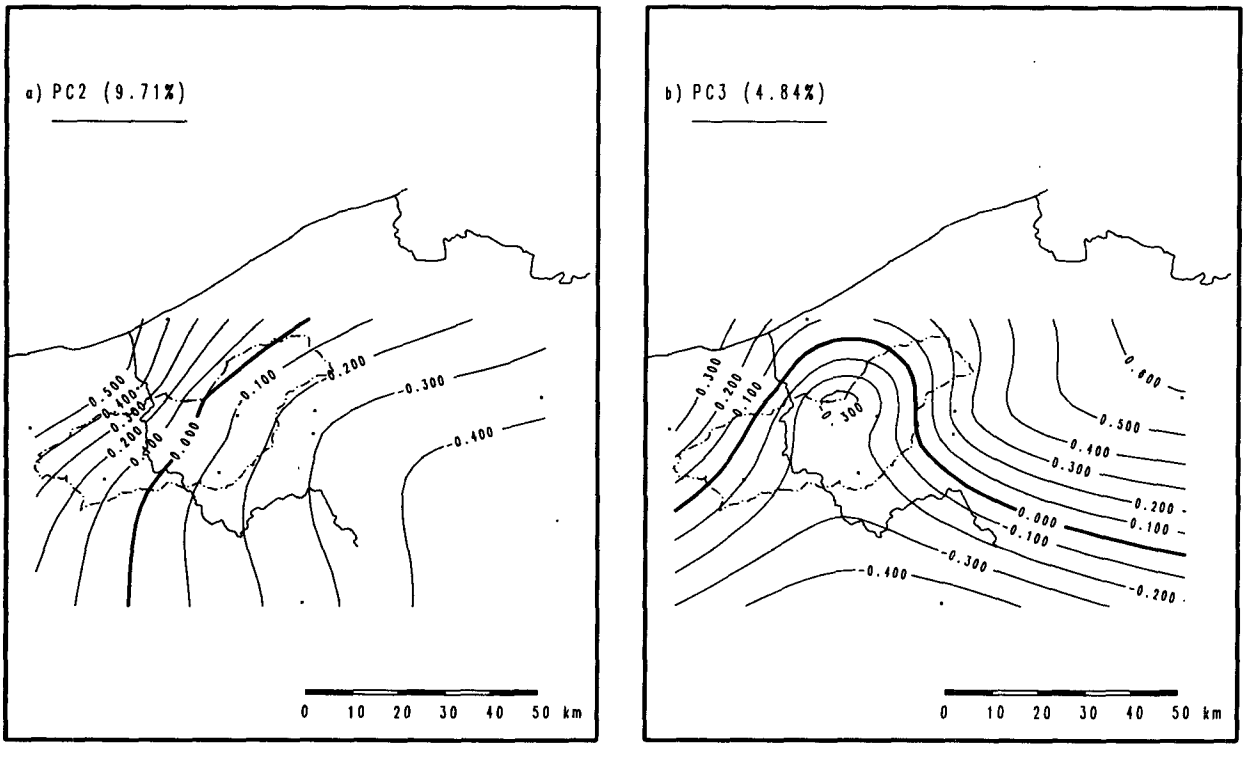

FIG. 4. Principal component patterns in 1977.

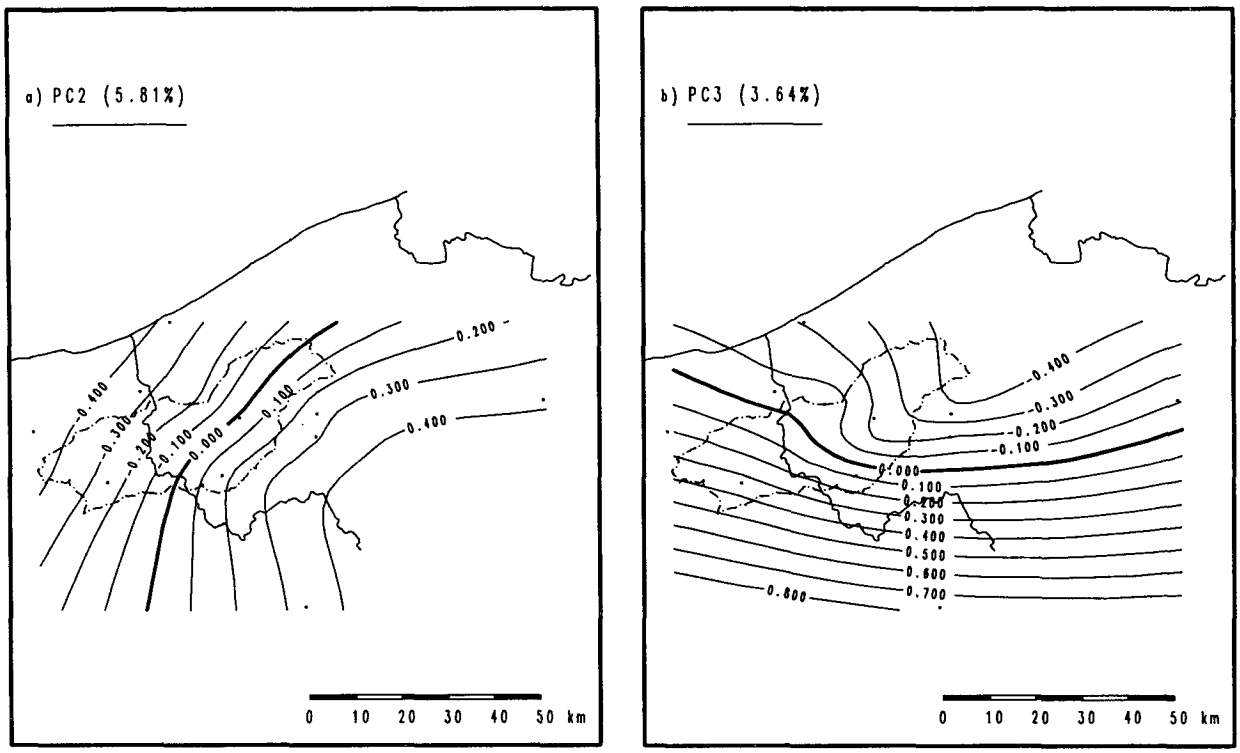

FIG. 5. Principal component patterns in 1978.

The second group (S2-S3-S6), located in the eastern part, is also positively correlated in PC2 but negatively in PC3. In the third quadrant, S1-S10-S11 form the third group. Again, they belong to one region, an area in the northwest of the region. The last group consists of S4 and S9.

This spatial grouping of the stations into four zones can be seen in Fig. 7, where the zero line from the

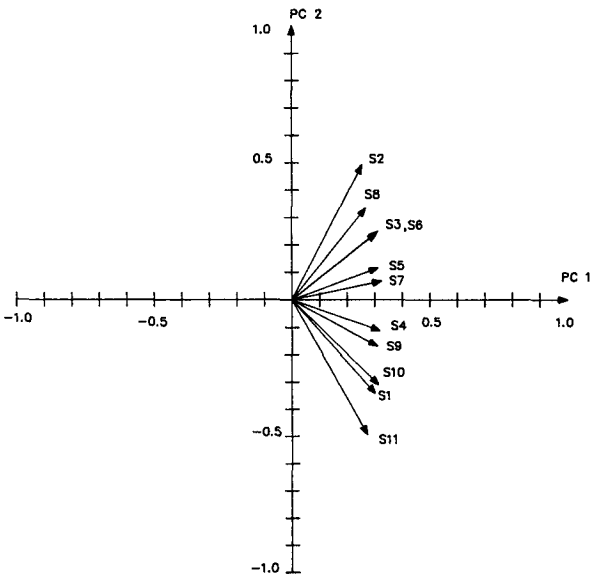

FIG. 6a. Factor loadings for 11 stations (S1–S11) with respect to the first (PC1) and second (PC2) principal component (three year data).

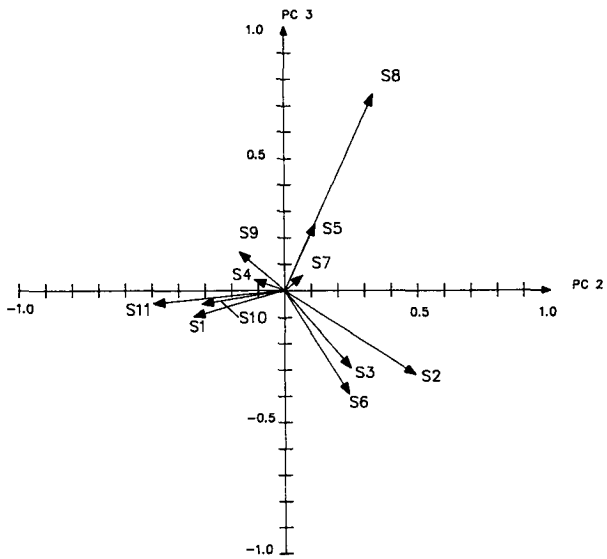

FIG. 6b. Factor loadings for 11 stations (S1–S11) with respect to the second (PC2) and third (PC3) principal component (three year data).

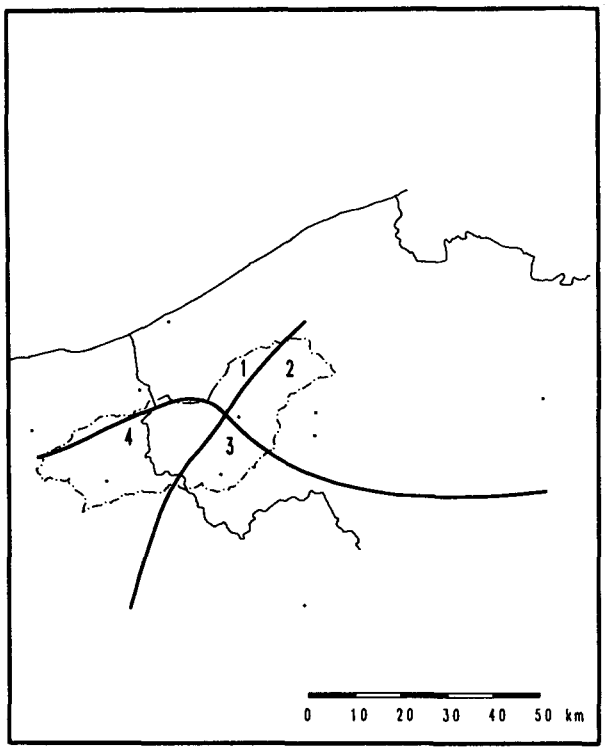

FIG. 7. The coherent precipitation regions as identified by principal components analysis.

second and third principal component divides the watershed into four subregions. The four groups of variables distinguished in Fig. 6b are exactly the same as those in Fig. 7. Except for the first group from Fig. 6b (S5-S7-S8), variable S7 moved toward the second group (S2-S3-S6).

5. Identification of coherent precipitation regions

The principal component patterns suggest four main regions into which rainfall variations over the IJzer watershed can be divided: northeast, southeast, southwest, and northwest (Fig. 7). Out of each subregion a station was chosen to calculate the cross-correlation with the remaining stations. This was done for the three years together (1973-1977-1978: the main spatial patterns of the principal components are consistent from year to year, so the correlations were examined only for the three years together). The stations used in each region were Lo (northwest, number 4), Boezinge (northeast, number 7), Vlamertinge (southeast, number 5) and Cassel (southwest, number 9). The cross-correlation maps (Figs. 8a–d) show that precipitation at each of the four chosen stations is strongly correlated with precipitation at most of the other stations. The correlation matrix between all stations is listed in Table 4. A critical level of $r = 0.85$ was chosen to define three coherent precipitation regions. The fourth region, the land represented by Boezinge, was skipped because it coincides almost entirely with the first region (Lo); the latter was preferred because it covers a greater area (at the $r = 0.85$ level) northeast of the watershed. Within each region, the areal precipitation is put equal to the point precipitation of the reference station. These three reference stations, Lo, Vlamertinge, and Cassel each define a homogeneous precipitation region with known accuracy ($r = 0.85$). A graphical representation of the three regions is shown in Fig. 9.

6. Discussion

A similar study (Wigley et al. 1984) considered a much wider area with 55 stations in England and Wales over the period 1861–1970. The first component explains appreciably less of the total variance (around 50%) than we found in Flanders. Unlike the study in this paper, Wigley used annual precipitation. However, our aim was to investigate only the rainfall patterns in three years (wet, dry, average). In a study produced by Tabony (1981) for Western European rainfall, approximately 28% of the variance was accounted for by the first pattern. Salinger (1980) found that for New Zealand rainfall only 12.5% of the variance is explained by the first component. In both studies the first principal component comprises a fairly uniform field. These findings show that the relative importance of the patterns depends among other things on the size of the area covered. As the area grows the uniform pattern (accounting for most of the variance) will decrease in importance and will be replaced as first principal component by one of the other patterns. Although in our study the first principal component accounts for almost 75% of the variance, the higher order PCs (PC2, PC3 and PC4) are still important for delineating small-scale heterogeneities. In this respect the second principal component pattern is related to geographical effects and the third principal component pattern to topographical effects. Another determining factor, though not investigated, is the surface pressure. Although this factor might be of little relevance in a small study area, Salinger (1980) found that higher-order components are significantly correlated with a pressure eigenvector. The fourth principal component might be related to this factor.

Although the analysis indicated that spatial coherence of rainfall is greatest in an average year (1978) and least in wet (1977) and dry (1973) year, the differences were found not to be important enough to define different coherent regions for each separate year. Therefore the division of the watershed into four regions was based on the three years component pattern. No further analyses were carried out to identify seasonal fluctuations patterns. For Western European rainfall, other researchers (Tabony 1981) found that the principal components were essentially the same in all seasons. The same findings were reported for Western European temperature patterns (Gray 1981). It was con-

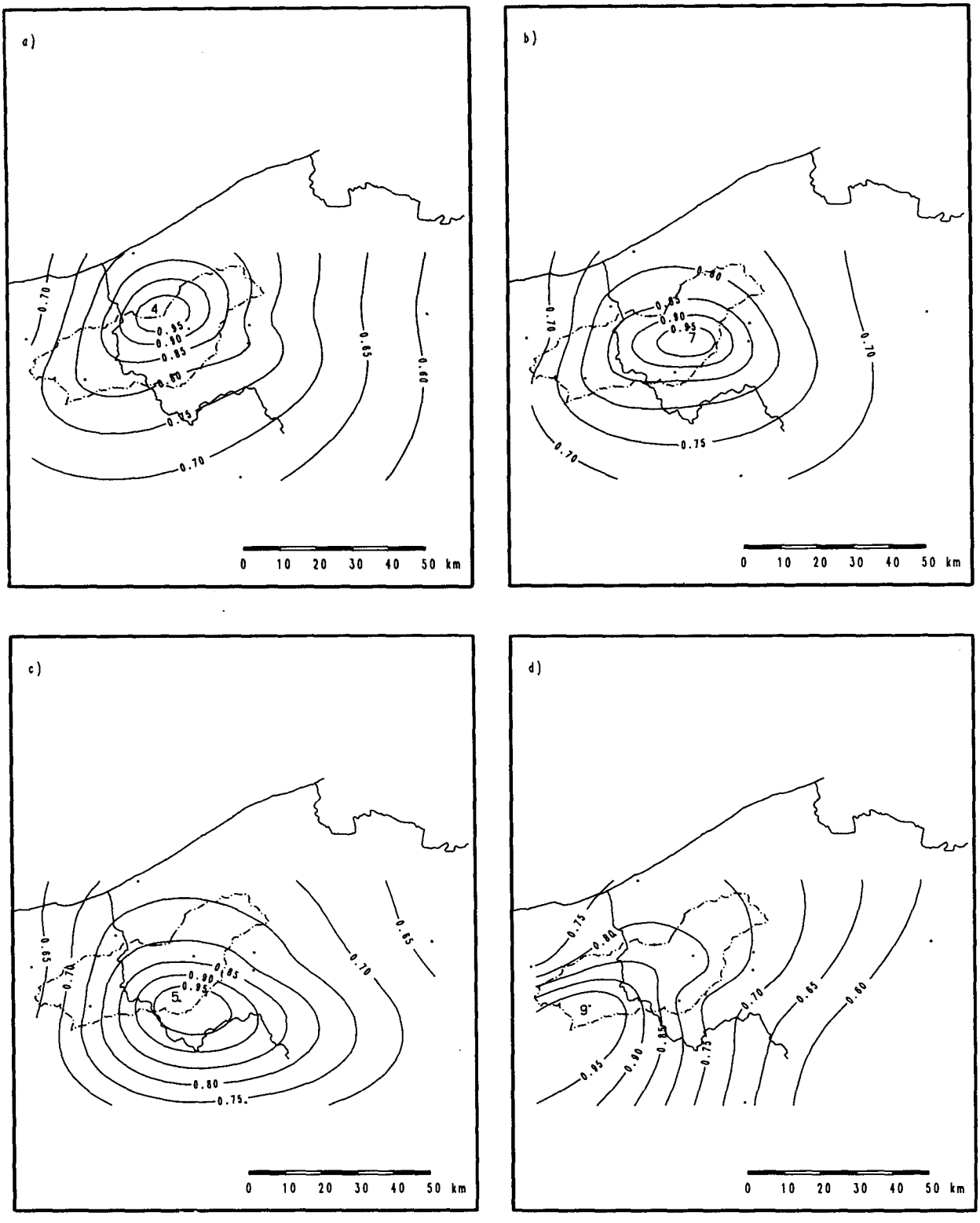

FIG. 8. Spatial patterns of the correlations between the annual precipitation at stations listed in Table 1 and (a) Lo, (b) Boezinge, (c) Vlamertinge, and (d) Cassel (three year data).

TABLE 4. Correlation matrix between all 11 stations for three years of daily rainfall data.

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11
S1	1.00	0.55	0.69	0.81	0.72	0.68	0.77	0.57	0.76	0.86	0.78
S2		1.00	0.69	0.58	0.63	0.70	0.65	0.62	0.56	0.57	0.46
S3			1.00	0.79	0.79	0.92	0.84	0.67	0.73	0.73	0.61
S4				1.00	0.80	0.77	0.86	0.66	0.81	0.86	0.69
S5					1.00	0.77	0.87	0.74	0.77	0.77	0.63
S6						1.00	0.82	0.63	0.74	0.72	0.60
S7							1.00	0.71	0.81	0.82	0.65
S8								1.00	0.65	0.60	0.51
S9									1.00	0.80	0.74
S10										1.00	0.80
S11											1.00

cluded that patterns that are time-invariant have physical significance.

The correlation patterns for the four representative stations out of these four regions divide the watershed into almost the same regions. If the division is based on the level $r = 0.85$, instead of four regions, only three remain. The first and second region, defined with the principal component patterns, have merged together in the correlation pattern method to form one region. The correlation method is independent of the principal component analysis in the sense that the resulting spatial patterns are not influenced by boundary effects as the component method is to some extent. Of course, there is no "correct" division, but the one identified by correlation analysis is suitable for our purposes and has an objective basis.

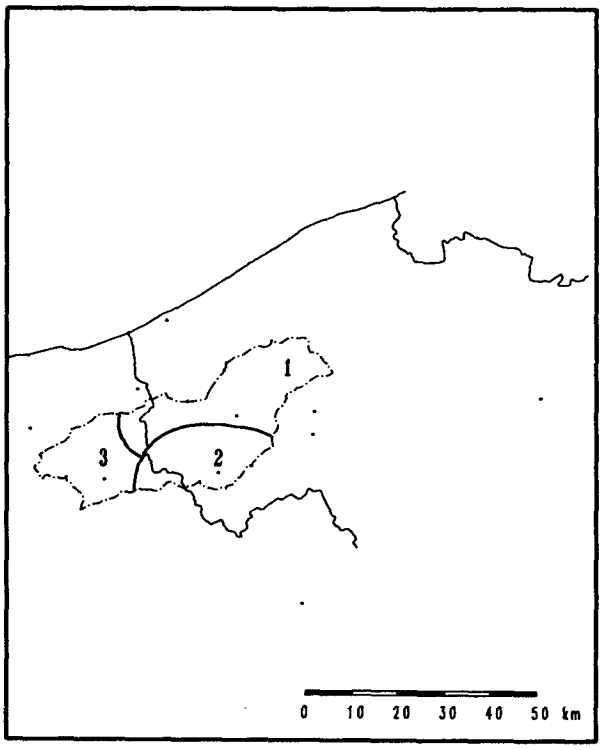

FIG. 9. The coherent precipitation regions identified by correlation analysis.

The field water balance model SWATRER not only requires daily precipitation as input data but also other climatological variables, such as temperature, relative humidity, etc. There is some evidence, however, that due to its high spatial variability, rainfall point measurements are less representative of an area than is the case for other meteorological variables. Further analyses have to prove whether these other variables exhibit similar patterns and whether these patterns are displayed on the same regional scale. These analyses also could be carried out using the geostatistical approach presented in this paper, which has proved to be an objective and exploratory method for deriving regions with high coherence in rainfall pattern.

7. Conclusions

Principal components analysis has proved to be a very useful tool in investigating the relationships between the different stations. Moreover, the geographical distribution of the principal components can be related to factors that influence the rainfall distribution. In this study the two most important factors were the coastal and continental influence on rainfall and the orographic effect. They were related to the second and the third principal component, respectively. Higher principal components (PC4, PC5, . . .) are difficult to interpret: they refer to combined effects of different factors and/or small-scale factors.

PCA can be used, as we have demonstrated here, to delineate regions that exhibit comparable rainfall patterns. Once a representative station out of each region has been chosen, it can be cross-correlated with the remaining stations. Regional patterns of the correlation coefficients serve as a final tool for dividing the watershed into homogeneous areas with known accuracy.

PCA gives a deeper understanding of the relationships between the different variables and the physical characteristics of the data are clearly reflected.

Acknowledgments. Our thanks go particularly to Els Abts for collecting and processing of the rainfall data, and to Harry Vereecken for the discussions on data analysis.

REFERENCES

Buishand, T. A., 1982: Some methods for testing the homogeneity of rainfall records. *J. Hydrol.*, **58,** 11–27.

Dierickx, J., C. Belmans and P. Pauwels, 1986: SWATRER: a computer package for modelling the field water balance. Laboratory of Soil & Water Engng., K. U. Leuven, Reference manual 2, 114 pp.

Gilman, D. L, 1957: Empirical orthogonal functions applied to thirty-day forecasting. Sci. Rep. No. 1, Contract AF19-(604) 1283, Dept. Meteor., MIT, 129 pp.

Gray, B. M., 1981: On the stability of temperature eigenvector patterns. *J. Climatol.*, **1,** 273–281.

Green, P. E., and J. Douglas Caroll, 1978: *Analyzing Multivariate Data.* John Wiley and Sons, 519 pp.

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

Kutzbach, J. E., 1967: Empirical eigenvectors of sea-level pressure, surface temperature and precipitation complexes over North America. *J. Appl. Meteor.*, **6,** 791–802.

Lorenz, E., 1956: Empirical orthogonal functions and statistical weather prediction. Sci. Rep. No. 1, Contract AF19-(604) 1566, Dept. Meteor., MIT.

Rasmusson, E. M., P. A. Arkin, W-Y Chen and J. B. Jalickee, 1981: Biennal variations in surface temperature over the United States as revealed by singular decomposition. *Mon. Wea. Rev.*, **109,** 587–598.

Salinger, M. J., 1980: New Zealand climate: I. Precipitation patterns. *Mon. Wea. Rev.*, **108,** 1892–1904.

Tabony, R. C., 1981: A principal component and spectral analysis of European rainfall. *J. Climatol.*, **1,** 283–294.

Wigley, T. M. L., J. M. Lough and P. D. Jones, 1984: Spatial patterns of precipitation in England and Wales and a revised, homogeneous England and Wales precipitation series. *J. Climatol.*, **4,** 1–25.