Quart. J. R. Met. Soc. (1980), **106**, pp. 873–886

551.577.32(540) : 519.23

Cluster analysis of rainfall stations of the Indian peninsula

By SULOCHANA GADGIL and R. NARAYANA IYENGAR

Centre for Theoretical Studies *Department of Civil Engineering*
Indian Institute of Science,
Bangalore 560012, India

Received 25 June 1979; revised 19 February 1980. Communicated by Professor R. P. Pearce)

SUMMARY

Principal component analysis is applied to derive patterns of temporal variation of the rainfall at fifty-three stations in peninsular India. The location of the stations in the coordinate space determined by the amplitudes of the two leading eigenvectors is used to delineate them into eight clusters. The clusters obtained seem to be stable with respect to variations in the grid of stations used. Stations within any cluster occur in geographically contiguous areas.

1. INTRODUCTION

Cluster analysis of rainfall stations to yield natural groups which can be considered as homogeneous with respect to the variation of rainfall is a useful first step in the development of stochastic models for prediction as well as in the study of variability over long time-scales using time series (Dyer 1976, Dyer 1975). Such classifications are also required for the choice of appropriate strategies for agriculture and the planning for utilization of water resources of various regions. In this paper we present a cluster analysis of the fifty-three rainfall stations of peninsular India shown in Fig. 4. The mean annual rainfall over the region is shown in Fig. 1.

It may be noted that, climatologically speaking, peninsular India lies between the mean winter location of the intertropical convergence zone (ITCZ) over the Indian ocean and its mean summer location north of 20°N over the Indian longitudes. A large portion of the peninsula receives significant amounts of rainfall during the northward movement of the ITCZ in early summer and its southward migration in autumn and early winter. Thus, from

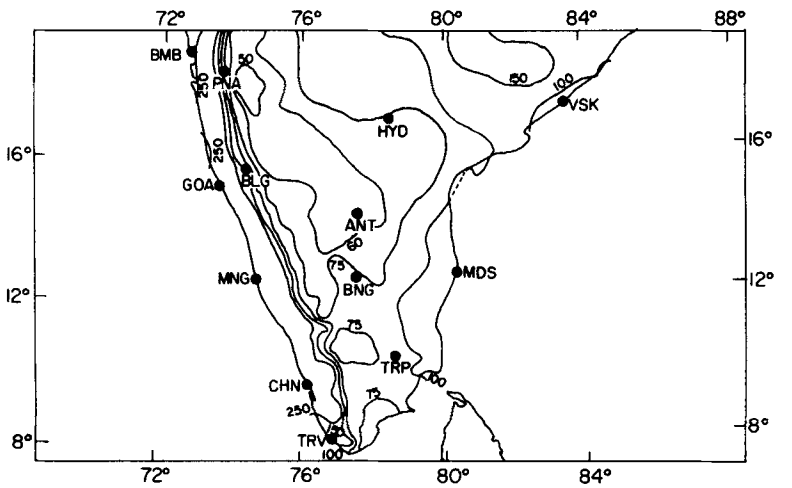

Figure 1. Annual rainfall (cm) of the peninsular region. (From *Rainfall Atlas of India*, India Met. Dept. 1971.)

the viewpoint of large-scale dynamics, the region chosen may be considered as homogeneous as a first approximation. Over this region the spatial variation of the rainfall during different seasons arises mainly from the effects of orographic barriers and the difference in conditions at the oceanic interface over the Arabian sea and the Bay of Bengal. The temporal profile of the rainfall is determined by the interplay of several factors and differs from station to station. A detailed analysis of the observed spatial variation of the rainfall profile can yield an insight into the relative importance of the different factors in different regions, and can also provide an economical presentation of the variations for comparison with the results of physical models incorporating these factors realistically. We undertake one such analysis in this paper.

Principal component analysis (method of empirical orthogonal functions) has been used to derive the spatial and temporal patterns of different meteorological variables (Craddock and Flood 1969, Barnett 1977). Here we apply it to the mean pentad rainfall data at the stations of interest (section 2) to obtain temporal patterns of rainfall over the Indian peninsula (section 3). When the mean rainfall profile at any station is expressed in terms of these patterns, it can be represented as a point in the coordinate space of the associated amplitudes. Clusters are then determined from the distribution of the points representing all the stations in this coordinate space (section 4). The effect of a change in the network density on the clusters obtained is studied in section 5. Results obtained by applying factor analysis in place of principal component analysis are discussed in section 6. Finally, a possible meteorological interpretation of one of the eigenvectors is mentioned in section 7.

2. Data

The mean pentad (5-day) rainfall compiled by Ananthakrishnan and Pathan (1971) from the normals of daily accumulated rainfall for the period 1901–1950, published by the India Meteorology Department, is the basic data set of this study. The temporal profile, at each of the 53 stations chosen, is thus specified in terms of the 73 pentad values implying a data matrix 53×73.

3. Principal component analysis

(a) *Notation and formulation*

We do not expect the 73 pentad values of the mean rainfall profile at any station to be uncorrelated. Hence, it is possible to reduce the dimensionality required to specify the rainfall profile and get a more economical description which is maximally powerful in bringing out the differences between various profiles by using principal component analysis.

The rainfall $R(i, j)$ at station i (i from 1 to N) in pentad j (j from 1 to 73) can be expressed as the sum of the products of the coefficients or amplitudes $A_n(i)$, which vary in space, and associated temporal profiles $B_n(j)$, thus

$$R(i,j) = \sum_{n=1}^{73} A_n(i)\, B_n(j);$$

$$i = 1, \ldots, N; j = 1, \ldots, 73$$

$B_n(j)$ are the eigenvectors of the covariance matrix

$$c(k,m) = \sum_i [\{R(i,k) - \langle R(k)\rangle\}\{R(i,m) - \langle R(m)\rangle\}]$$

where

$$\langle R(k)\rangle = \frac{1}{N}\sum_{i=1}^{N} R(i,k)$$

These eigenvectors are orthonormal and the indices are arranged so that the first $B_1(j)$ corresponds to the largest eigenvalue and in general the k^{th} eigenvector $B_k(j)$ to the k^{th} largest eigenvalue λ_k. The k^{th} principal component is defined as the linear combination of the values at different pentads, with the coefficient of the value of the j^{th} pentad being $B_k(j)$. The total variance v_k explained by the k^{th} principal component is

$$v_k = \lambda_k / \sum_m \lambda_m$$

The first principal component has the largest variance of any linear combination of the pentad values and explains a proportion λ_1 of the variance. Given $B_n(j)$, the amplitudes $A_n(i)$ can be readily found from the data. Then the rainfall profile at any station can be expressed as a point in the n dimensional space of the amplitudes A_n.

(*b*) *Characteristics of rainfall patterns*

On applying principal component analysis to the mean pentad data we find that the first eigenvalue accounts for 85·4% of the variance and the second and the third for 6·4% and 4·5% respectively. Thus the first two principal components account for most of the variance. This result differs from Dyer's (1975) application of principal component analysis to the spatial profiles of the annual rainfall over South Africa in which the first eigenvalue accounted for only 44% of the variance, and twenty-eight components were required to explain 89% of the variance.

The first two eigenvectors are shown in Fig. 2(a, b) for data specified at five-day and

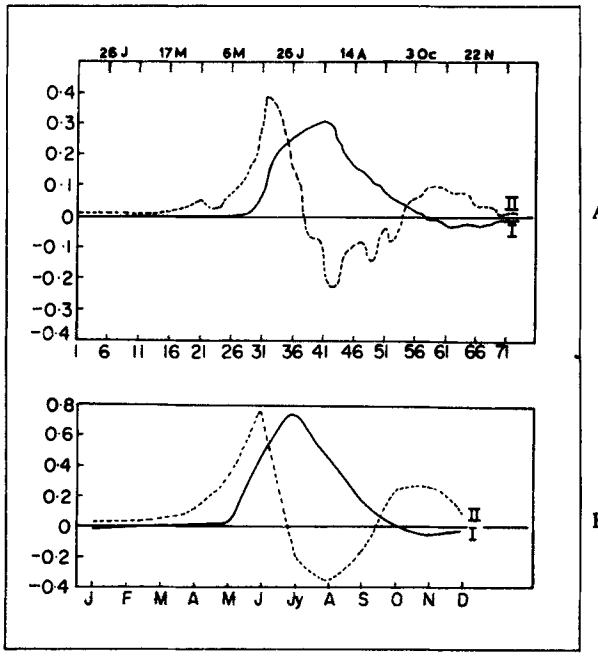

Figure 2. First two eigenvectors for pentad rainfall data (a) and monthly rainfall data (b). I. South-west monsoon component (SWM). II. Pre-monsoon and north-east monsoon (PNEM).

monthly intervals respectively. The first eigenvector is negative and small in magnitude ($\sim 10^{-2}$) for the first 21 pentads, but becomes significantly positive from pentad 29 to pentad 57 (21 May–8 October) with a maximum in pentad 41 (20–24 July). It is negative between pentads 60 and 66 (23 October–26 November) with the maximum negative value occurring in pentad 62 (2–6 November). This indicates that the epoch 23 October–26 November is negatively correlated with that representing the south-west monsoon. This first eigenvector may be interpreted as the south-west monsoon component (SWM). The second eigenvector has positive values over pentads 1 to 37 and 55 to 73 (i.e. 28 September–30 June) with peaks at pentad 61 (28 October–1 November) and 32 (5–9 June). Negative values occur between pentads 39 and 54 with a prominent trough in the pentad 42. We denote this eigenvector representing the pre-monsoon and north-east or post-monsoon seasons by PNEM. Comparison of Figs. 2(a) and 2(b) indicates that the components obtained from monthly data are similar to those obtained from pentad data, apart from the fine scale variations in the latter which are naturally absent in the former. A linear superposition of these two eigenvectors, with a sufficiently large weighting for the second, results in a bi-modal distribution of rainfall with a minimum around August separating the two maxima as pointed out by Ananthakrishnan and Pathan (1971).

4. Clustering by a two-step application of near centroid method

(*a*) *First order clustering*

Since the first two principal components explain 91·8% of the variance, the rainfall profile at any station i can be adequately described in terms of the associated amplitudes $A_1(i)$, $A_2(i)$. Thus every station can be represented as a point in the coordinate space of the amplitudes A_1, A_2, and the Euclidean distance between points representing two stations in this coordinate space is a direct measure of the difference between their profiles. Figure 3 depicts the locations of the stations in this coordinate space.

We use the method of nearest centroid sorting (Anderberg 1973) for determining the clusters. First we identify a set of seed points, which are used as cluster nuclei, subjectively by

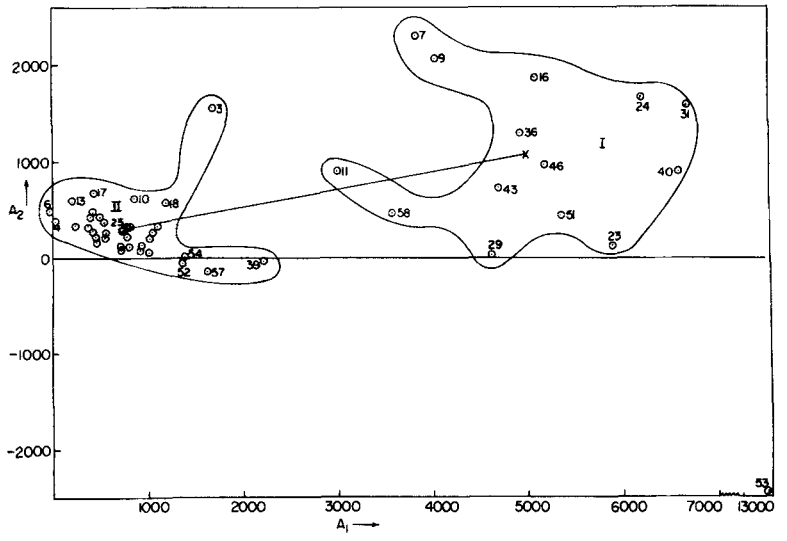

Figure 3. All the stations represented as points in A_1–A_2 space of amplitudes of the first two eigenvectors.

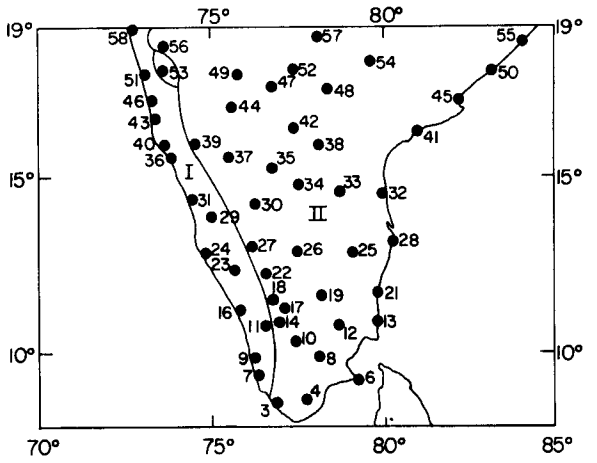

Figure 4. Locations of stations belonging to clusters I and II.

an examination of the distribution of points in the A_1–A_2 space. For example, for the case in Fig. 3 we take points representing stations 46 and 25 as seed points of two possible clusters. In addition, the isolated point 53 is also considered as a third seed point. Next we take any one of the remaining 50 stations, measure the distance between the point representing it and the three seed stations in the A_1–A_2 space, and assign it to the cluster with the nearest seed point. The centroid of the cluster which has gained in membership is then calculated. This process is repeated for the next station and it is also assigned to the cluster with the nearest centroid. After all stations have been assigned, one final pass is made taking the existing centroids of clusters as seed points and assigning each station to the nearest seed point. In the case shown in Fig. 3, this method yields the well-separated clusters I and II and a single member cluster consisting of the isolated point 53.

Clusters I and II consist of 15 stations along the west coast and 37 stations spread over the peninsular region to the east of the Western Ghats respectively (Fig. 4). The west coast stations are characterized by larger than average amplitude of the south-west monsoon component, whereas those belonging to cluster II are by a low amplitude of this component. The station Mahabaleshwar (No. 53) is located on one of the peaks of the Western Ghats and has large rainfall. In the network of stations used it is in a group by itself. However, if more stations similarly situated with respect to orography (such as Agumbe in Karnataka) were included, this group would get a larger membership. The point representing Mahabaleshwar in Fig. 3 is closer to the centroid of the west coast cluster I than that of II and, hence, if only two seed points were chosen it could have been considered as sub-cluster of I rather than an independent cluster. The decision to consider it as an independent cluster is, therefore, subjective.

(*b*) *Second order clustering*

The new variables determined by the principal component analysis have been effective in bringing out the differences between the west coast stations and those on the rest of the peninsula. With respect to these new variables, the profiles of the stations in the latter group are rather similar and they form a densely packed cluster in the A_1–A_2 space. In order to bring out the differences, if any, between the rainfall profiles of these stations, and hence determine the sub-clusters, we repeated the principal component analysis separately for

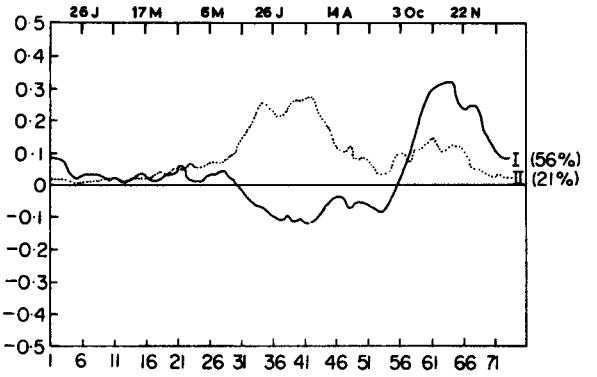

Figure 5. Eigenvectors for stations belonging to the peninsular cluster II in Fig. 4.

stations belonging to clusters I and II. It was found that whereas the first group did not yield any sub-clusters, seven sub-clusters emerged clearly from cluster II.

The first two eigenvectors for this group are shown in Fig. 5. For the majority of the peninsular stations the leading eigenvector is positive in the period 8 October–21 May with a maximum during 28 October–26 November. It is negative during the south-west monsoon season with maximum negative value during 20–24 July. It thus represents the pre- and post-monsoon season with break during the south-west monsoon season and is similar to

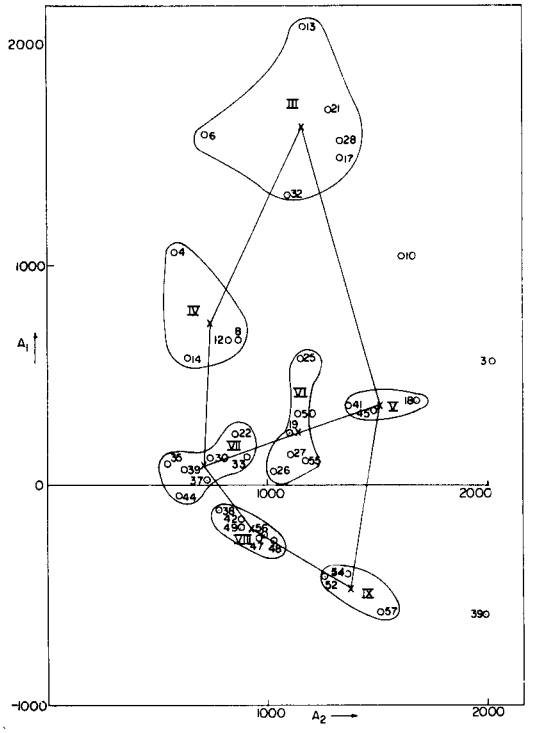

Figure 6. The points in A_1–A_2 space of amplitudes of the eigenvectors in Fig. 5.

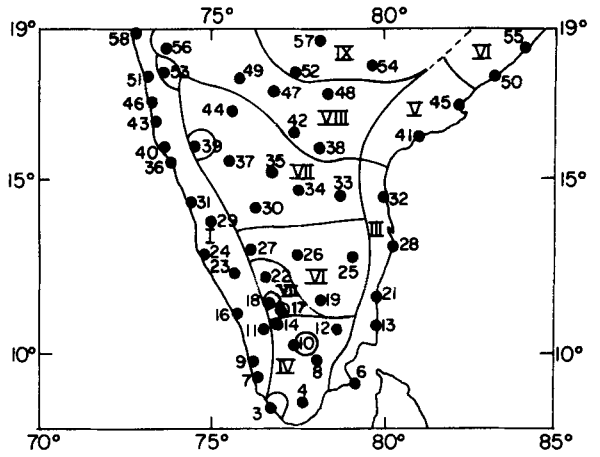

Figure 7. Locations of stations belonging to the different sub-clusters shown in Fig. 6 and Fig. 4.

the second component (PNEM) in Fig. 2, except for the absence of a significant peak in the pre-monsoon season. The second eigenvector is a combination of the south-west and the north-east monsoons having a larger peak in the former season in July and a smaller one in the latter season in November. The first component accounts for 56% while the second component accounts for 21% of the variance. Distribution of the points representing the 37 peninsular stations in the new coordinate space of amplitudes A'_1–A'_2 is shown in Fig. 6. Taking as seed-points the stations 21, 12, 45, 19, 30, 49 and 54 the seven cluster III–IX were determined by the method described earlier. The vast majority of the stations (34 out of 37) can be unambiguously assigned to the different clusters by the method used. However, station 10 is almost equidistant from the centroids of clusters III and V, although slightly closer to that of the latter, and is represented as an isolated point. The stations 3 and 39 which occurred in the periphery of cluster II in the first order analysis (Fig. 3) also appear as isolated points.

The geographical locations of stations belonging to all the clusters identified at the end of the two-step analysis is shown in Fig. 7. It is seen that the clusters III–IX fall into contiguous areas with the exception of stations 10, 17, 18. These stations are located at rather high levels (2·3, 1·7, 2·2 km above m.s.l. respectively) and hence the singular behaviour is understandable. The geographical boundaries between the clusters IX and V, between IX, VIII, and between VIII and V will be better determined if more stations are included north of 16°N. Clusters of stations similar to the isolated stations 53, 3, and perhaps some new clusters, may emerge when a denser network of stations is used along the west coast.

The mean pentad rainfall and the standard deviations at selected pentads for the different clusters are shown in Fig. 8. The west coast clusters I, the northern interior cluster IX, and the southern east coast cluster III are characterized by a sudden onset of the monsoon and a short rainy season in comparison with the other clusters on the peninsula. The clusters VIII and IX experience significant breaks in the 45th pentad (9–13 August). The standard deviation for any pentad is much larger in the west coast cluster I than in the other clusters. In general, the standard deviation is large when the rainfall is large. The ratio of the maximum value of the standard deviation to the maximum value of the pentad rainfall varies between 0·24 and 0·42 for the different clusters.

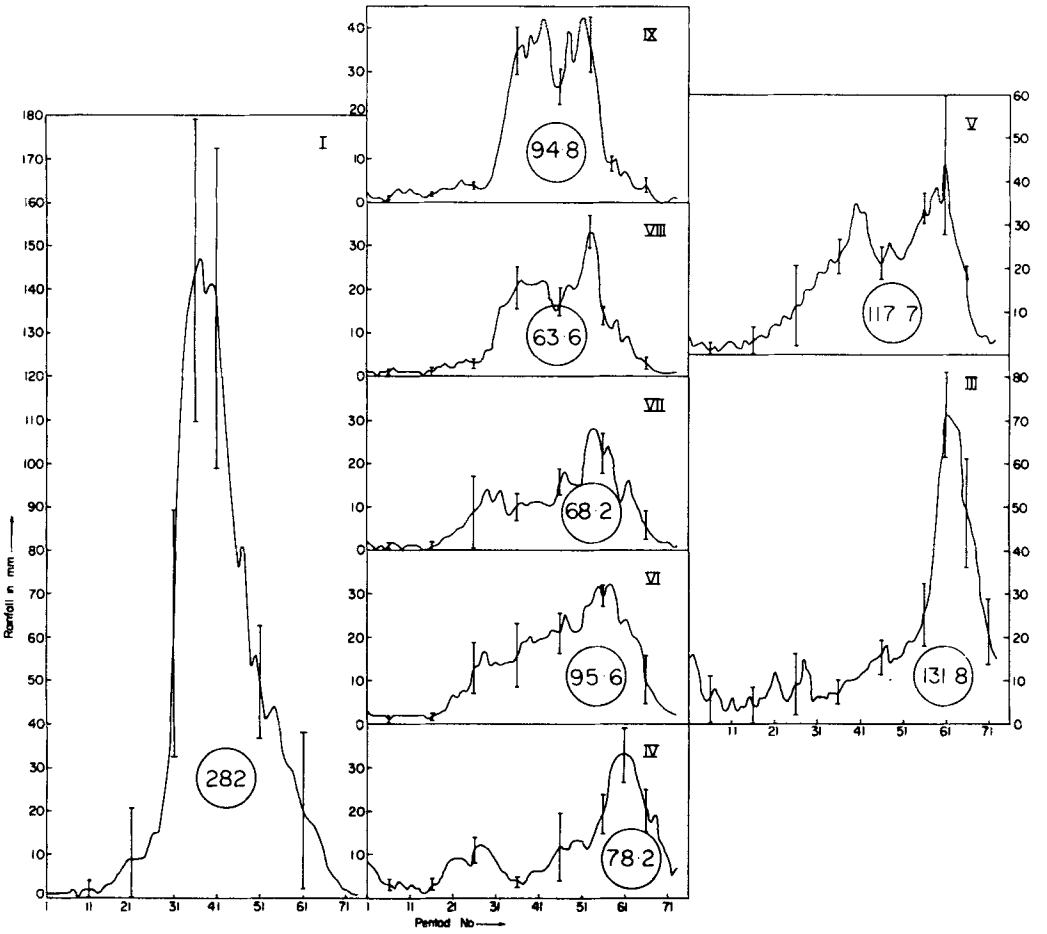

Figure 8. Mean rainfall profiles (in mm per pentad) of the different clusters. The standard deviation is also shown at some points. The annual mean rainfall in cm is indicated below each profile.

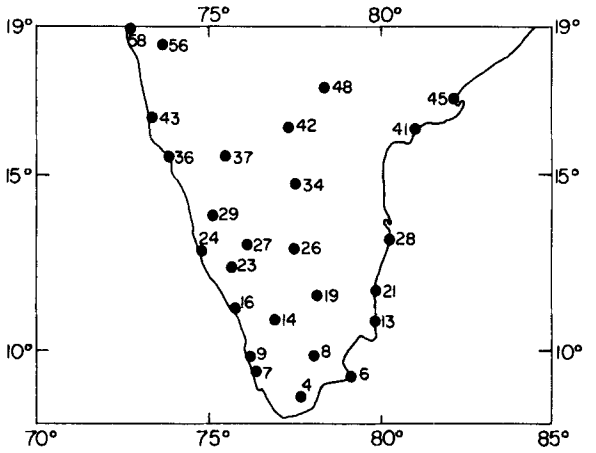

Figure 9. Locations of stations used to study the effect of grid-size.

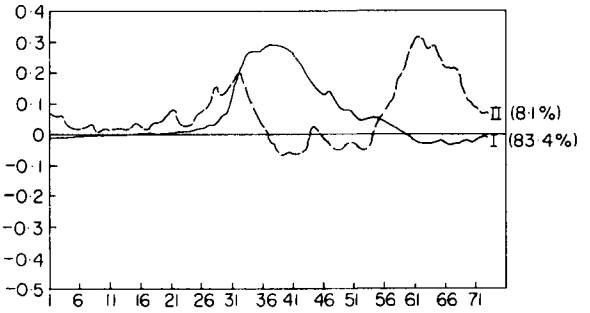

Figure 10. First two eigenvectors for stations in Fig. 9.

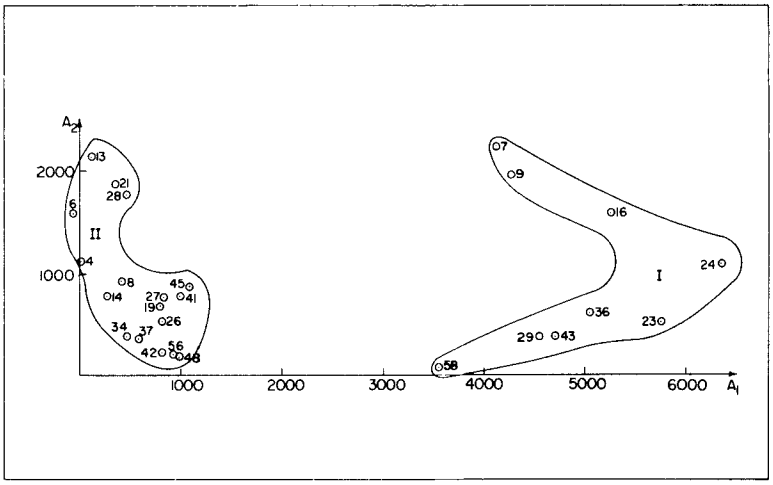

Figure 11. Locations of stations from Fig. 9 in the amplitude space of the eigenvectors in Fig. 10.

5. Effect of density of network on clustering

The effect of the number of stations used on the clustering obtained was tested by repeating the analysis with roughly half the number of stations (Fig. 9) spread over the region south of 19°N. The eigenvectors are shown in Fig. 10 and the locations of the retained stations in the coordinate space of the new amplitudes in Fig. 11. Again, the first application yields two clusters identical to I and II obtained from Fig. 3. The eigenvectors for stations belonging to the cluster II only are shown in Fig. 12 and its sub-clusters in Fig. 13. It is seen that the clusters obtained from the analysis involving fewer stations are better defined with larger distances between any pair of clusters in the amplitude space. We, therefore, expect that for a denser network of stations the distribution of points in the amplitude space would still exhibit high concentrations near the centroids of clusters found here as well as some new clusters; and relatively low concentrations near the peripheries, with less sharp demarcation of the boundaries between the clusters. However, this effect cannot be demonstrated here because of the limitation on the number of stations at which the data is available.

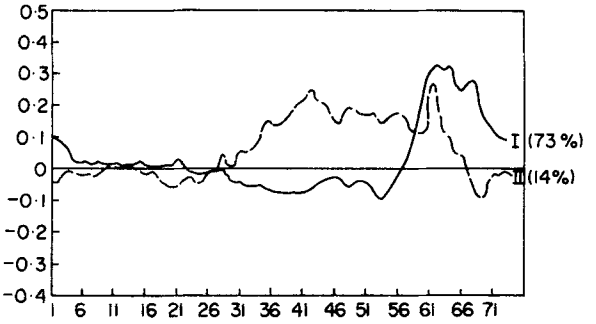

Figure 12. Eigenvectors for the stations in cluster II of Fig. 11.

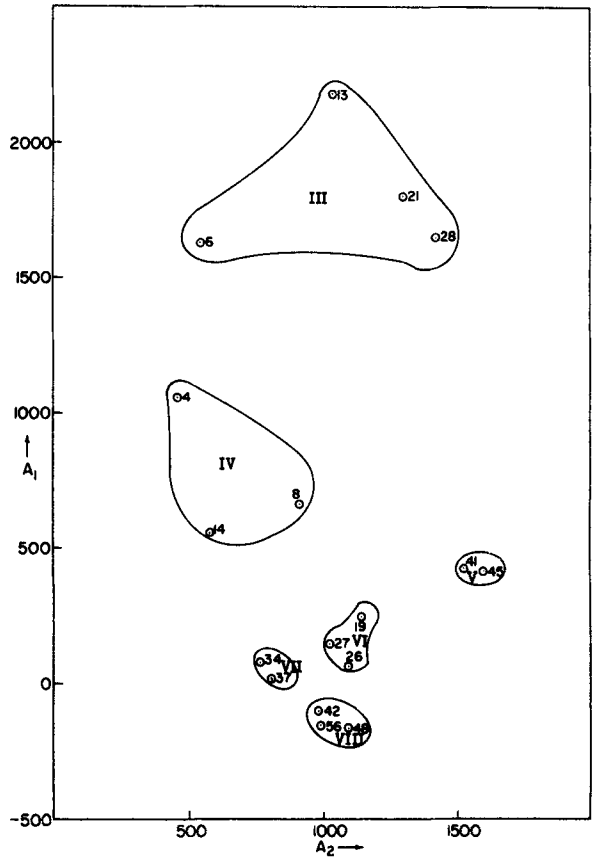

Figure 13. Locations of stations of cluster II of Fig. 11 in amplitude space of eigenvectors of Fig. 12.

6. Factor analysis

When a correlation matrix is used instead of a covariance matrix in the analysis of the fifty-three stations, the eigenvectors have less prominent peaks and appear rather flat (Fig. 14). The first eigenvector is similar to the PNEM component (I in Fig. 5) and the SWM component is now relegated to second place. This is because the effect of the high rainfall stations with large SWM component and large standard deviations is reduced due to the use of correlations rather than covariance. The distribution of stations in the plane of the amplitudes of the first two eigenvectors (Fig. 15) is seen to be similar to that obtained using the covariance matrix (Fig. 3) and clusters which emerge are identical. This is interesting in view of the fact that only 72% of the variance is explained by the first two components in this case.

7. A possible meteorological interpretation

It is of interest to examine the latitudinal variation of the amplitudes of the eigenvectors (SWM) and (PNEM) for the west coast region (Fig. 2) and the rest of the peninsula (Fig. 5)

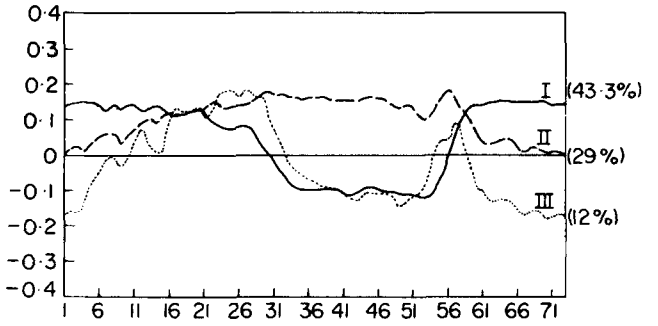

Figure 14. First three eigenvectors obtained by using a correlation matrix for all the 53 stations.

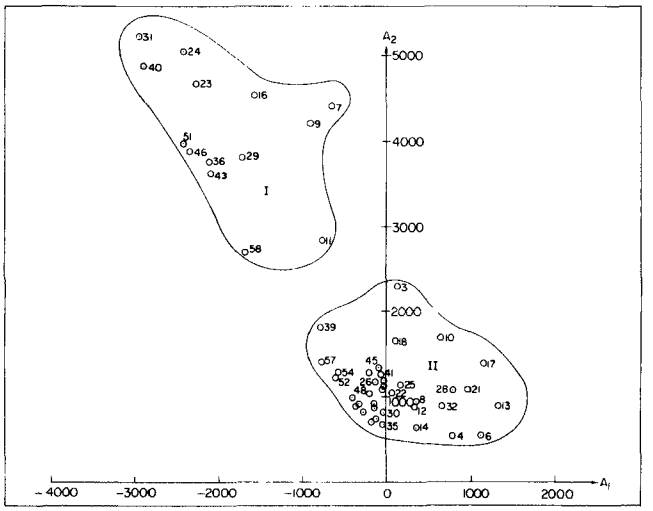

Figure 15. Location of stations in the amplitude space of the first two eigenvectors of Fig. 14.

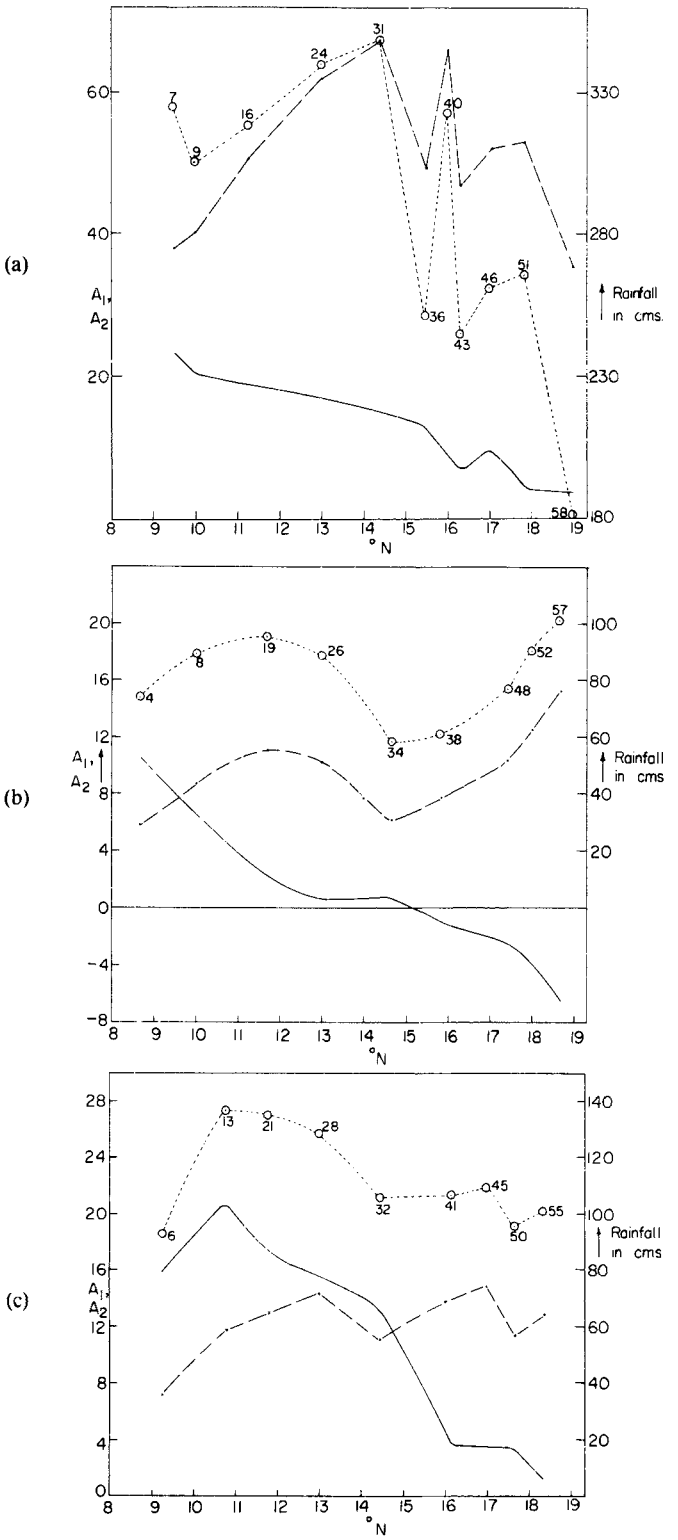

Figure 16. Latitudinal variation of the annual rainfall in cms (.....o.....o.....o.....), of the amplitudes of the SWM (- - - -) and the PNEM (————) components: along west coast (a); 78°E (b); and east coast (c).

in relation to that of the annual rainfall. This is depicted in Fig. 16(a, b, c) along the west coast, along the central meridian of 78°E, and along the east coast respectively. It is seen that the latitudinal variation in the amplitude of the SWM component is remarkably similar to that of the rainfall in all cases. This close correspondence between the values of the amplitude of the SWM and rainfall is somewhat surprising in the last two cases since SWM represents not the leading but the second eigenvector for these regions.

The amplitude of the PNEM component is seen to decrease almost monotonically with increasing latitude along the west coast, along 78°E and north of 11°N along the east coast. This suggests that PNEM represents rainfall associated with a system located somewhere in the region south of about 10°N. It is interesting that the ITCZ is located in this region in the pre-monsoon and post-monsoon seasons. Further, even in the summer monsoon, a secondary cloud band is located in this region (Hubert *et al.* 1969). Thus the specific weighting to the different pentads implied by the eigenvector PNEM could be looked upon as an empirically derived rainfall pattern associated with the oceanic ITCZ.

8. Conclusion

This application of principal component analysis to the mean temporal profiles of rainfall at the different stations has shown that 91% of the variance is explained by the first two components. Thus the profile at any station can be adequately described in terms of the two associated amplitudes. This brings about a great reduction in the dimensionality required to specify the rainfall profile (e.g. from 73 pentad values to the 2 values of these amplitudes). Natural groups or clusters emerge from an analysis of the distribution of points representing the stations in the coordinate space of the amplitudes. A two-step application of this procedure has yielded eight clusters. One important feature of the clustering obtained is that generally stations within any cluster also occur in geographically contiguous areas, although no information about geographical location was used in the analysis. This, together with the stability of the clusters with respect to change in network suggests that a reasonable clustering has been achieved.

Acknowledgments

We thank R. Narasimha and the referees for their valuable suggestions.

References

Ananthakrishnan, R. and Pathan, J. M. 1971 Rainfall patterns over India and adjacent seas, *India Met. Dept. Sci. Rep. No.* 144.

Anderberg, M. R. 1973 *Cluster analysis for applications*, Academic Press, New York, London.

Barnett, T. P. 1977 The principal time and space scales of the Pacific Trade Wind Fields, *J. Atmos. Sci.*, **34**, 221–236.

Craddock, J. M. and Flood, C. R. 1969 Eigenvectors representing the 500 mb geopotential surface over the northern hemisphere, *Quart. J. R. Met. Soc.*, **95**, 576–593.

Dyer, T. G. J. 1975 The assignment of rainfall stations into homogeneous groups: an application of principal component analysis, *Ibid.*, **101**, 1005–1013.

1976 On the components of time series; the removal of spatial dependence, *Ibid.*, **102**, 157–165.

Hubert, L. F., Krueger, A. F. and Winston, J. S. 1969 Double ITCZ – fact or fiction? *J. Atmos. Sci.*, **26**, 771–773.

LIST OF STATIONS USED

Code No.	Name of station	Latitude N	Longitude E
3	Trivandrum	8–29	76–57
4	Palayamkottai	8–44	77–45
6	Pamban	9–16	79–18
7	Alleppey	9–33	76–20
8	Madurai	9–55	78–07
9	Cochin	9–58	76–14
10	Kodai Kanal	10–14	77–28
11	Palghat	10–46	76–39
12	Tiruchirapalli	10–46	78–43
13	Nagapattinam	10–46	79–51
14	Coimbatore	11–00	76–58
16	Kozhikode	11–15	75–47
17	Coonoor	11–21	76–48
18	Ootacamund	11–24	76–44
19	Salem	11–39	78–10
21	Cuddalore	11–46	79–46
22	Mysore	12–18	76–42
23	Mercara	12–25	75–44
24	Mangalore	12–52	74–51
25	Vellore	12–55	79–09
26	Bangalore	12–58	77–35
27	Hassan	13–00	76–09
28	Madras	13–04	80–15
29	Balehonnur	13–22	75–27
30	Chitaldurga	14–14	76–28
31	Honnavar	14–17	74–27
32	Nellore	14–27	79–59
33	Cuddapah	14–29	78–50
34	Anantpur	14–41	77–37
35	Bellary	15–09	76–51
36	Marmagoa	15–25	73–47
37	Gadag	15–25	75–38
38	Kurnool	15–50	78–04
39	Belgaum	15–51	74–32
40	Vengurla	15–52	73–38
41	Masulipatnam	16–11	81–08
42	Raichur	16–12	77–21
43	Devgarh	16–23	73–21
44	Bijapur	16–49	75–43
45	Kakinada	16–57	82–14
46	Ratnagiri	16–59	73–20
47	Gulbarga	17–21	76–51
48	Hyderabad	17–27	78–28
49	Sholapur	17–40	75–54
50	Visakhapatnam	17–41	83–18
51	Harnai	17–49	73–06
52	Bidar	17–55	77–32
53	Mahabaleshwar	17–56	73–40
54	Hanamkonda	18–01	79–34
55	Calingapatnam	18–20	84–08
56	Poona	18–32	73–51
57	Nizamabad	18–40	78–06
58	Bombay (Colaba)	18–54	72–49