Proc. Indian Acad. Sci. (Earth Planet. Sci.), Vol. 100, No. 2, June 1991, pp. 105-126.

Application of principal component analysis to understand variability of rainfall

R N IYENGAR
Centre for Atmospheric Sciences, Indian Institute of Science, Bangalore 560012, India

MS received 25 May 1990; revised **16** February 1991

Abstract. The usefulness of principal component analysis for understanding the temporal variability of monsoon rainfall is s.tudied. Monthly rainfall data of Karnataka, spread on 50 stations for a period of 82 years have been analysed for interseasonal and interannual variabilities. A subset of the above data comprising 10 stations from the coherent west zone of Karnataka has also been investigated to bring out statistically significant interannual signals in the southwest monsoon rainfall. Conditional probabilities are proposed for a few above normal/below normal transitions. A sample prediction exercise for June-July using such a transition probability has been found to be successful.

Keywords. Monsoon; rainfall variability; principal components; empirical orthogonal functions; eigenvalues; spatial structure; predictability; transition probability.

1. Introduction

Rainfall is perhaps the most important variable in the phenomenon of monsoon. The amount of rainfall in a given week, month or season varies from year to year over a wide range. This raises the question: is there an identifiable pattern in these variations, or is the variability purely random. Variability may be defined as a tendency of rainfall to fluctuate around a long-term average (normal) value. It follows that one can consider this variability on several time scales, such as, days, weeks and months, and also on diverse spatial domains, that is, stations, districts or states. As the monsoon is known to be organized spatially on a large scale and is persistent in time for several months, it could be useful to study the data on a few optimal scales. However, the optimal time and space scales for rainfall are unknown and thus one has to accept the data as they are and estimate empirically the existence of patterns. In the present investigation, this is undertaken for the monthly rainfall data of Karnataka. A variety of statistical analyses of rainfall on the monthly scale, have been made earlier by several investigators. Thus, information on the mean, standard deviation, coefficient of variation is available. The autocorrelation and power spectral density of the time series of a few stations have also been obtained (lyengar 1982, 1987; Fleer 1977). It is found that these are white noise (purely random) processes after the annual cycle is removed. No temporal pattern emerges in monthly rainfall at station level. As interstation data are spatially correlated one would ask whether by combining the data from several stations trends could be identified. Identification of coherent zones (Gadgil *et al* 1988) and clustering of stations into groups (Gadgil and lyengar 1978) are examples of such a study. The present work is concerned with both spatial and temporal variation by composing the large scale data into principal components (PC)

in time and empirical orthogonal functions (EOF) in space. Previously Lorenz (1956), Kutzbach (1967), Priesendorfer *et al* (1981), Overland and Priesendorfer (1982), Hastenrath and Rosen (1983), Singh and Kripalani (1986), Bedi and Bindra (1980), Rakhecha and Mandal (1977) among others have utilized this technique. The main emphasis in these studies has been on explaining the spatial structure of the field. In the present study it is shown that principal components can be used to compare and, if necessary, group the 'years'. The PC of monthly and seasonal data reveal interesting information about seasonal, interseasonal and interannual variability. Further, some patterns in predictability hitherto unsuspected are identified.

2. Data

The data analysed in this investigation are the monthly rainfall of Karnataka spread over 50 stations and extending over 82 years, from 1901 to 1982. The State of Karnataka along with the stations considered is presented in figure 1. While it would be useful to consider the all-India data, there are restrictions due to data gaps and unequal length of station time series. It is also not clear whether consideration of a larger area improves or dilutes the temporal signals that may be present. Hence, before studying the all-India rainfall variability, a part of the country is considered in this study. The rainfall in Karnataka by itself is of considerable interest, as three

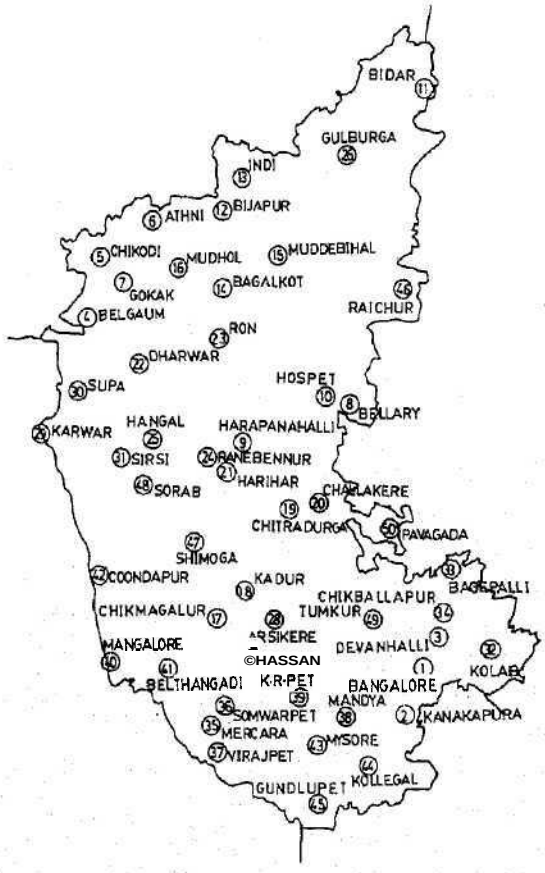

Figure 1. Station data network: Karnataka.

of the rainfall subdivisions of the Meteorological Department of India, namely, Coastal Karnataka (Sub. Div. 31), North Karnataka (Sub. Div. 32) and South Karnataka (Sub. Div. 33) are in Karnataka. Coastal Karnataka receives the highest monsoon rainfall among all the subdivisions. The seventy year mean summer monsoon rainfall for coastal Karnataka is as large as 2907 mm as reported by Shukla (1986). The southwest monsoon (SWM) or the summer season (June-September) accounts for the bulk of the rainfall of the year except in areas south of Bangalore. In areas near the west and south of Bangalore, SWM rainfall is less than 50% of the annual and to the south of Mysore this value is less than 40%. With this in view, the premonsoon and the post-monsoon seasonal rainfall data have also been analysed.

3. Analysis

The State-wide data matrix used here for any month or season is of size 50 x 82. The average, standard deviation, skewness and kurtosis have been computed for each station before further analysis. For principal component analysis (PCA) the centered data are used. Thus, if R_{it} is the actual rainfall at station i $(i = 1, 2 \ldots M)$ in the year $t(t = 1, 2 \ldots N)$ the mean value is

$$m_i = (1/N) \sum_{t=1}^{N} R_{it}, \tag{1}$$

the centered data are

$$r_{it} = (R_{it} - m_i) \tag{2}$$

The covariance matrix is

$$C_{ij} = (1/N) \sum_{t=1}^{N} r_{it} r_{jt}. \tag{3}$$

The eigenvalues λ_j and eigenvectors $\{\phi_{ij}\}$ of this symmetric matrix are extracted. The principal components are

$$p_{jt} = \sum_{i=1}^{M} r_{it}\, \phi_{ij}; (j = 1, 2 \ldots M) \tag{4}$$

This transforms the original time series r_{jt} into the new time series p_{jt}. The first few p_{jt}'s are generally sufficient to account for a large percentage of the spatial variation of the original data. Many of the previous rainfall studies along this line have concentrated on the EOF's or the eigenvectors ϕ_{ij}, which represent spatial patterns. It is found here that p_{it} also contains useful information which can be used to understand temporal variability. At this stage it would be necessary to identify how many p_{jt} have to be retained in the orthogonal representation

$$r_{it} = \sum_{j=1}^{M} p_{jt}\, \phi_{ij} \tag{5}$$

as significant. Preisendorfer *et al* (1981) have discussed different rules which can test the significance of the eigenvalues and the principal components. As pointed out by

them, the tests should be designed depending on the end use of orthogonal decomposition. First the eigenvalues should be tested to verify how significantly the data deviate from purely random noise. If the basic data were spatially uncorrelated with zero mean and unit variance, the eigenvalues would all be equal to unity, each explaining $100/M$ per cent of the total variance. In practice, due to sampling fluctuations the sample eigenvalues will differ from this value. The percentile level of the eigenvalues for several combinations of M and N has been obtained by Preisendorfer (1981) by Monte Carlo simulation of large samples. To test the significance of the eigenvalues they are normalized by

$$\bar{\lambda}_j = M\lambda_j / \sum_{j=1}^{M} \lambda_j \tag{6}$$

and compared with the simulated significance bands. This test is shown in figure 2 for the monthly and seasonal data. For all cases, it is found that the first three terms are significant. The fourth term is marginally significant but its contribution to the total variance is only about 4%. The first three together explain 60-70% of the total variance.

4. Monthly rainfall

Monthly rainfall presents an interesting picture as shown by figure 2. While the first eigenvector dominates the spatial structure, λ_1 increases in May and June to reach

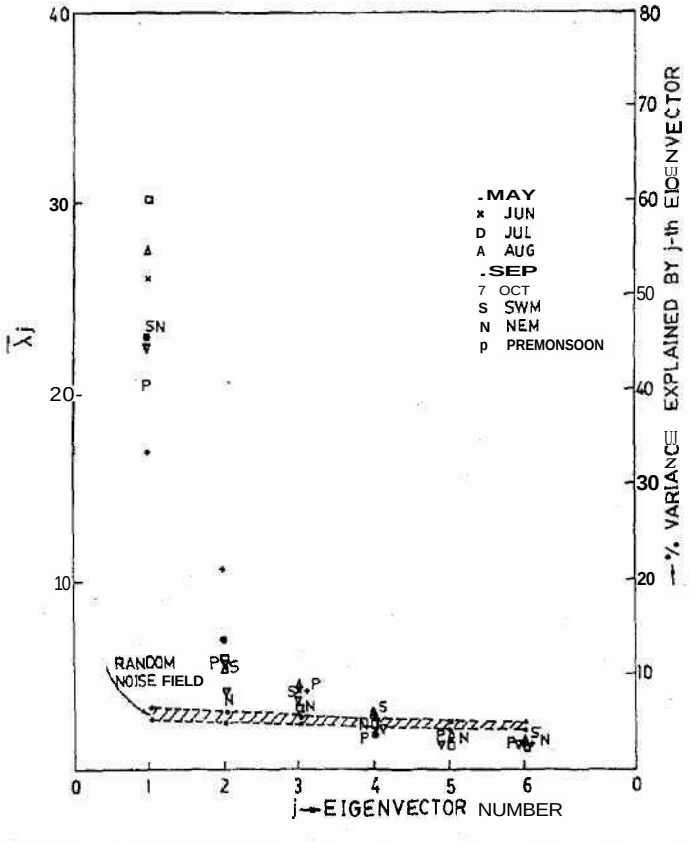

Figure 2. Normalized eigenvalues; Comparison with random noise field.

a peak in July. This is followed by a decrease in August and September. A better view of how the rainfall field is getting reorganized is provided by the eigenvectors (e.v.) shown in figures 3 to 8. Here the first two eigenvectors, multiplied by 100 are shown. Since the first e.v. is always predominant, the month-to-month transition of this would be important. It is seen that in May the whole State remains spatially correlated. This means that above/below normal fluctuations in rainfall along the west coast stations, which have the largest weight, would indicate similar trends in other parts of the State. This picture changes in June when the first e.v. produces a spatial contrast, dividing the State into three regions. As it is difficult to verify the significance of the sign and values of the station weights given by the eigenvectors, the first e.v. in June may be taken to indicate a west-east contrast. Above/below normal rainfall in the west would indicate below/above normal rainfall in the east. This pattern intensifies in July and the contrast decreases, but the east-west divide is still evident in the first e.v. of July and August. In September, positive associations of all stations are restored and this remains stable even in October. An interpretation of the second e.v. would proceed on similar lines. As this accounts for only about 10% of the variance, it is perhaps a local feature not related with atmospheric scales. The second e.v. in June-September indicates a contrast between the west coast and interior stations. The third and fourth eigenvector patterns which are not presented here depict further local scales over which the rainfall is fluctuating about its long-term mean value. The temporal variability of the rainfall is carried over to the PC's in decreasing order of importance. Each p_{jt}, $(j=1,2,3\ldots)$ is a time series sampled annually, and would lead to information on interannual variability. All the first four principal components of the six months have been studied to test the existence of

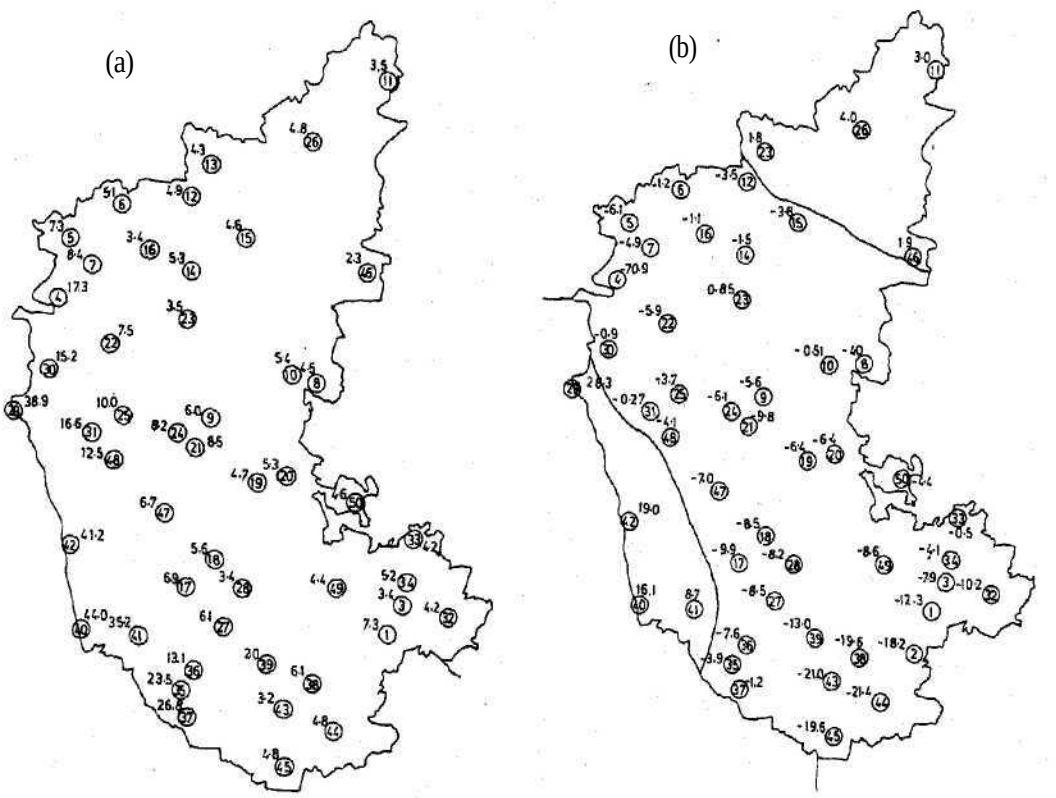

Figure 3. **a.** First eigenvector - May. Variance explained = 45·85%; b. second eigenvector - May. Variance explained = 13·97%.

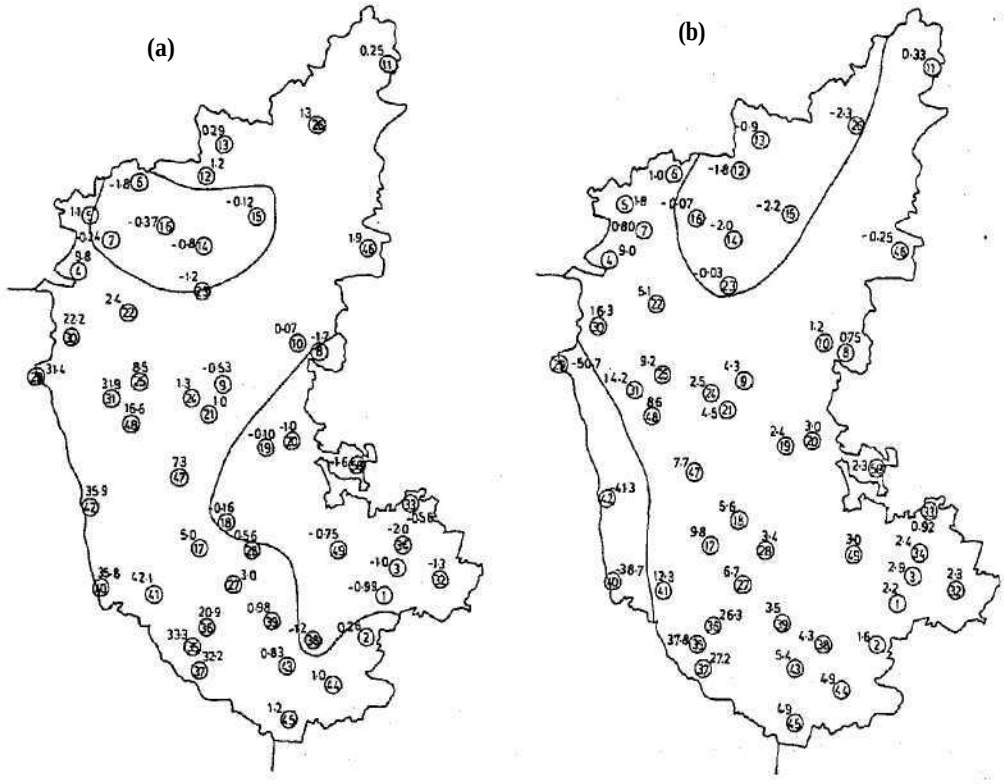

Figure 4. **a.** First eigenvector - June. Variance explained = 52·06%; b. second eigenvector - June. Variance explained = 11·11%.

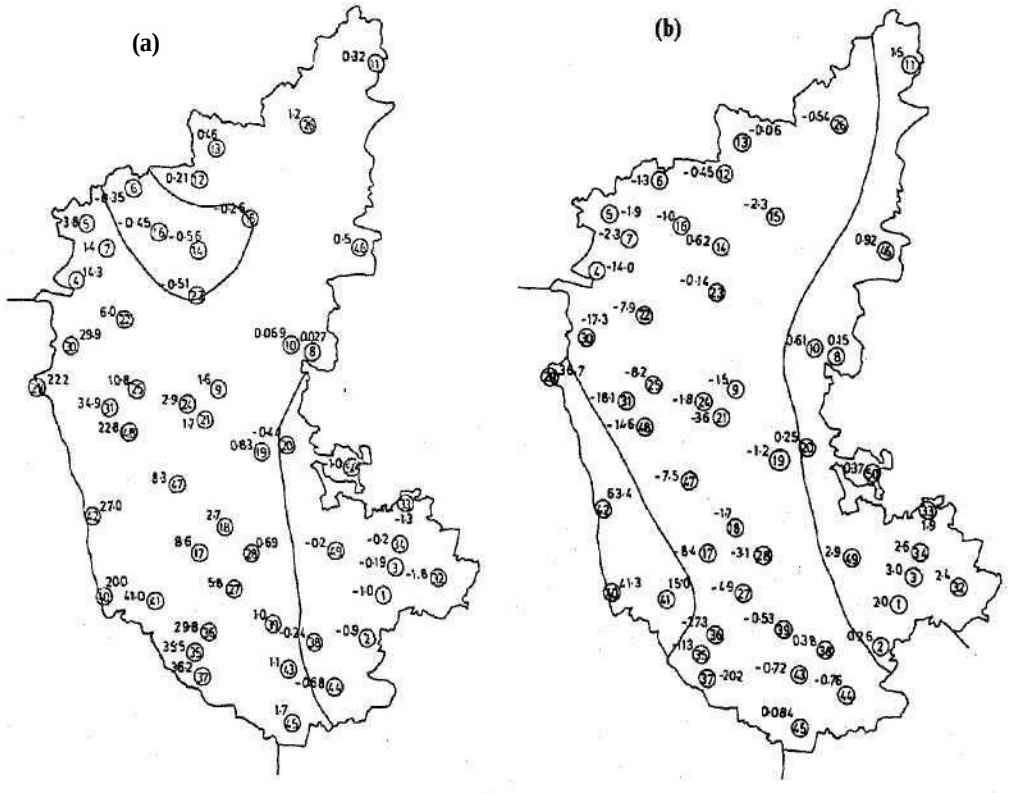

Figure 5. **a.** First eigenvector - July. Variance explained = 60·41%; b. second eigenvector - July. Variance explained = 11·47%.

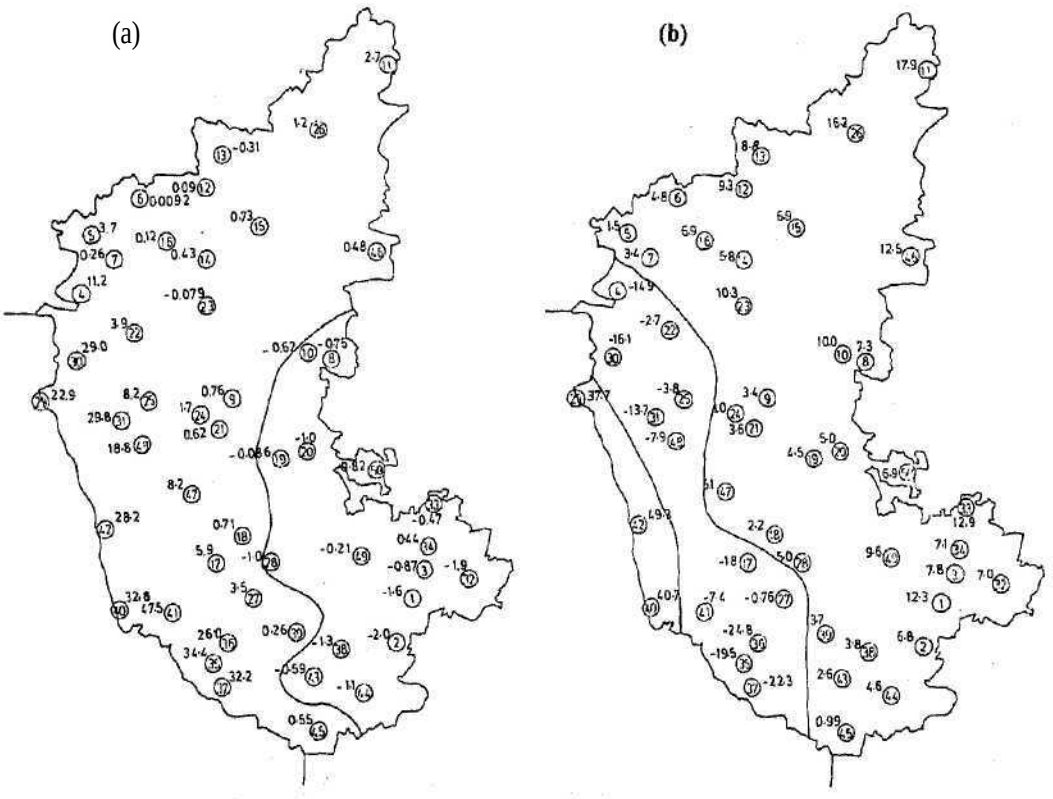

Figure 6. **a.** First eigenvector - August. Variance explained = 54·93%; b. second eigenvector - August. Variance explained = 10·57%.

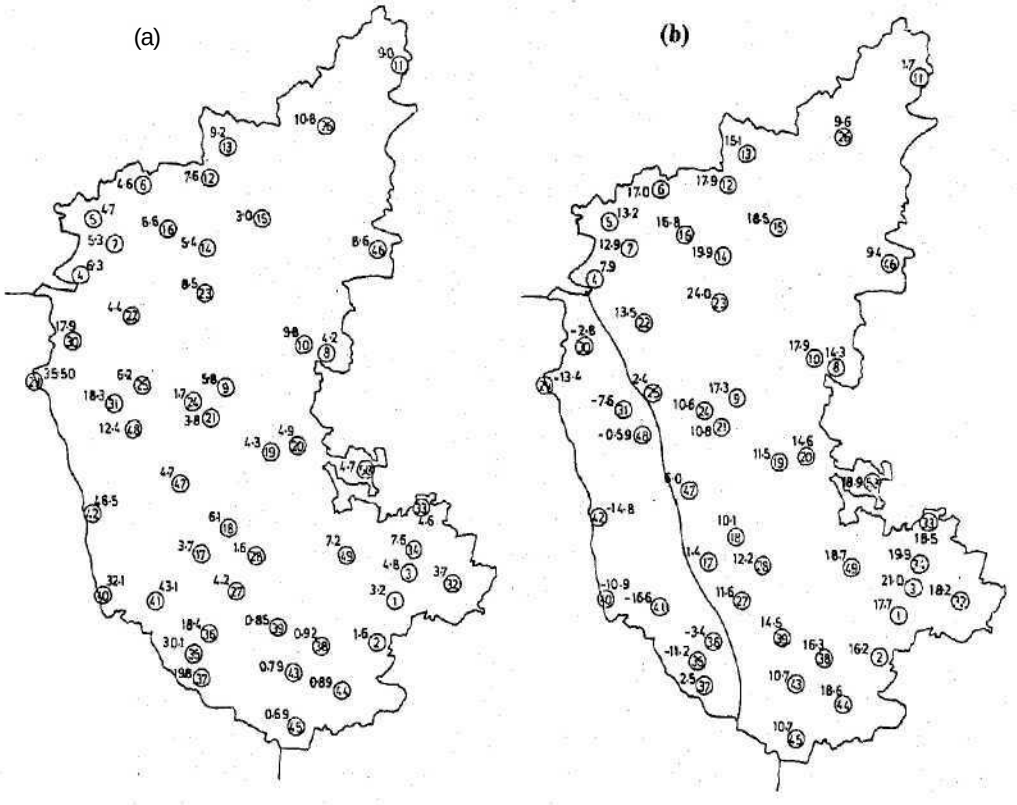

Figure 7. **a.** First eigenvector - September. Variance explained = 33·54%; b. second eigenvector – September. Variance explained = 21·28%.

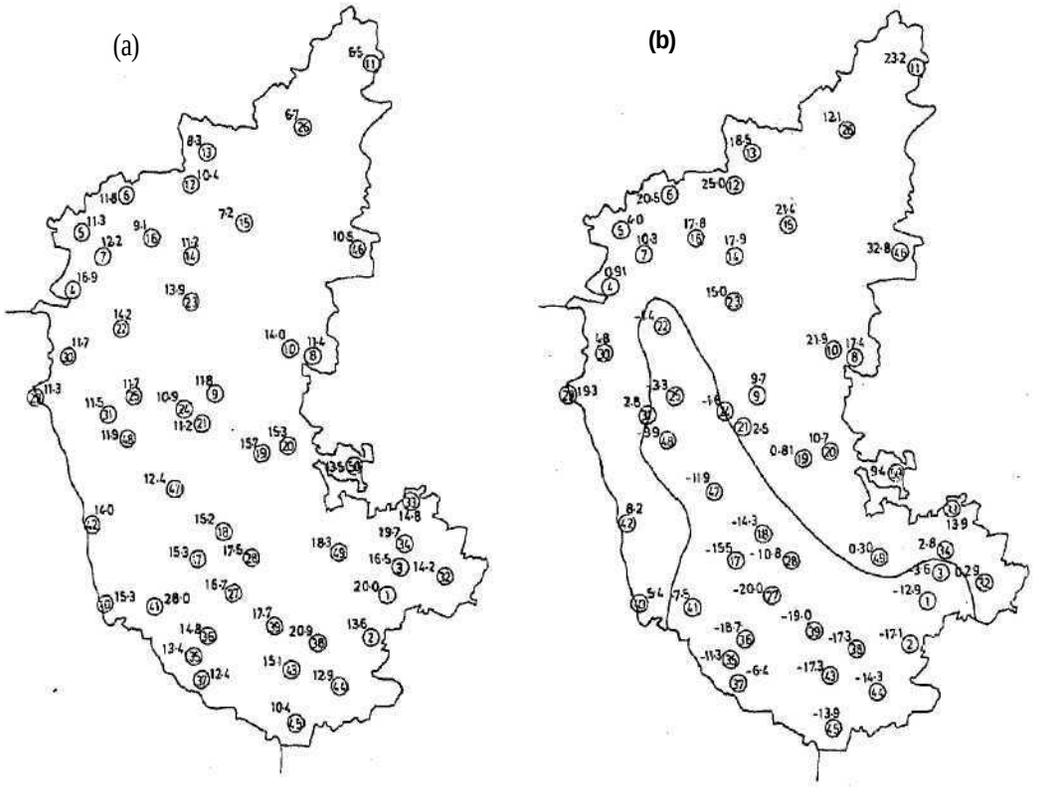

Figure 8. **a.** First eigenvector - October. Variance explained = 45·15%; b. second eigenvector - October. Variance explained = 8·54%.

autocorrelation for a maximum lag of 8 years. No significant autocorrrelation was found in any of the components. As a further test of annual association, the number of changes in the sign of the first two components, namely (++, +−, −+, −−) has been collected in a two-way contingency table. These are tested against the expected number of occurrences if the changes were just due to chance. No significant association in the signs on the annual scale was found for any of the first two monthly principal components.

4.1 *Monthly transitions*

Earlier it has been mentioned that station rainfall does not show month-to-month correlation. This does not exclude the possibility of a correlaton existing in area rainfall time series. The principal components are area-rainfall time series, where the weights are selected in an optimal fashion. However, the question whether the PC's representing the size of a state like Karnataka are able to bring out this feature or not, is still open. But, if monthly associations are present in the basic data one could expect to see them reflected in the way the PC's evolve from month to month. Here one particular indicator of this relation, namely the transition in the signs is studied. If the rainfall in a given month is normal at all sampling stations, all the corresponding PC's will be precisely zero. Since the first PC dominates the spatial variation, whenever it is zero one may expect the rainfall also to be near its own normal value. Thus, dependence if any, in the signs would indicate patterns in the intraseasonal variability of the rainfall. In table 1 the observed number of sequences of ++, +−, −+, −− are listed for the first PC.

For each row in table 1 the persistence or change in the sign can be shown on a 2×2 contingency table. The significance of the association is tested against the number expected, if the sign changes were purely by chance. For example, in May the first PC is + ve, 17 + 12 = 29 times. The corresponding number for June is 17 + 19 = 36. Now, if the PC's of May and June are independent, the expected number of occurrences of the + + sequence in 82 observations would be (36 x 29)/82 = 12·73. These frequencies are also listed in table 1. The full hypothesis H_0 is "there is no dependence in the month-to-month sign changes". The x^2 test is applied to test this hypothesis (Rohatgi 1984). The observed χ^2 values listed in table 1 are compared with the tabulated χ^2 value of 3·84, at one degree of freedom and at 95% significance. Whenever the observed value exceeds the tabulated value, the null hypothesis will be rejected. It is seen that the transitions from May to June and June to July could be accepted as exhibiting a pattern, whereas for the next two months the transitions are purely random. For September to October the null hypothesis is accepted at 95% level, but rejected at 90%. Thus, it is possible that this transition is also not purely due to chance. A similar analysis for the sign changes of the second PC shows that all transitions, except those from September to October are purely random. Cross-correlations between the first and second PC's have also been studied. Again, only the September-October transition is clearly identified as not due to chance. In table 2 all the frequencies observed and those expected due to chance are presented for September-October.

It is interesting to observe that September, which is the last month of the SW monsoon, provides an indication of how the rainfall could be in the first month of

Table 1. Frequency of sign sequences in the I PC of monthly rainfall (N = 82 yrs).

	Sign								
	++		+-		-+		--		χ^2
Month	Obs.	Expt.	Obs.	Expt.	Obs.	Expt.	Obs.	Expt.	Obs.
May-June	17	12·73	12	16·27	19	23·27	34	29·73	3·93
June-July	12	16·68	26	21·32	24	19·32	20	24·68	4·36
July-August	15	14·00	22	23·00	26	17·00	29	28·00	0·21
August-Sept.	11	12·85	20	18·15	23	21·15	28	29·85	0·73
Sept.-Oct.	20	16·22	15	18·78	18	21·78	29	25·22	2·86

Table 2. Frequency of sign sequences for September-October (N = 82 yrs).

	Sign								
	++		+-		-+		--		χ^2
Comp.	Obs.	Expt.	Obs.	Expt.	Obs.	Expt.	Obs.	Expt.	Obs.
PC1–PC1	20	16·22	15	18·78	18	21·78	29	25·22	2·86
PC1–PC2	19	14·50	16	20·50	15	19·50	32	27·50	4·16
PC2-PC1	13	19·00	28	22·00	25	19·00	16	22·00	7·06
PC2-PC2	23	17·00	18	24·00	11	17·00	30	24·00	7·23

Table 3. (a) Observed transition frequencies (b) Expected transition frequencies purely due to chance; Sept.-Oct.

(a)				<b)			
4	9	2	4	3·25	4·63	5·56	5·56
2	8	9	3	3·75	5·36	6·43	6·43
3	2	10	10	4·27	6-1	7·31	7·31
5	1	3	7	2·73	3·90	4·68	4·68

the northeast monsoon season. From figure 2 it is seen that in September both the first and second PC's are important, as they contribute 34% and 21% respectively to the total variance. Thus, it would be more appropriate to depict the state of the rainfall in terms of the first two components. One would ask whether the two components taken as a pair still show a significant relation between September and October. The sign of the first two components taken as a pair can be in any one of the four states, I = + +; II = - +; III =——; IV = + —. To study whether these four states in September and October are dependent, the 4 x 4 contingency table of the corresponding observed frequencies and the expected frequencies due to chance are shown in table 3. The calculated χ^2 value is 22·8, while the tabulated value of χ^2 at 9 degrees of freedom is only 16·9. Thus even with this stronger test it turns out that rainfall in October is related to rainfall in September.

5. Seasonal rainfall

The year can be divided into three seasons, namely, premonsoon (January to May), SWM (June to September) and the Northeast Monsoon (NEM) (October to December). An analysis similar to the monthlies has been carried out on the three seasonal rainfall data spread over the fifty stations. The first six normalized eigenvalues are plotted in figure 2 to test the significance by the dominant variance rule. It is clear that like the monthlies the seasonals also indicate the first three components to be significant. For SWM, the fourth component is also significant with a contribution of 6% to the total variance. The first two eigenvectors for the three seasons are shown in figures 9, 10 and 11. The first e.v. shows a highly correlated field in all three cases. The premonsoon second vector shows a west coast — interior contrast. The SWM second vector seems to accentuate this with further contrast emerging in the SW-NE direction. The second e.v. of the NEM indicates a contrast between stations which predominantly receive the NEM rainfall and those which do not.

5.1 *Interannual variability*

The interannual variability of the three seasonal rainfalls has been investigated as explained in the previous section. The premonsoon PC's do not show any annual relation, as verified by the autocorrelation or through the dependence in the sign sequences. The SWM principal components are, however, interesting because they

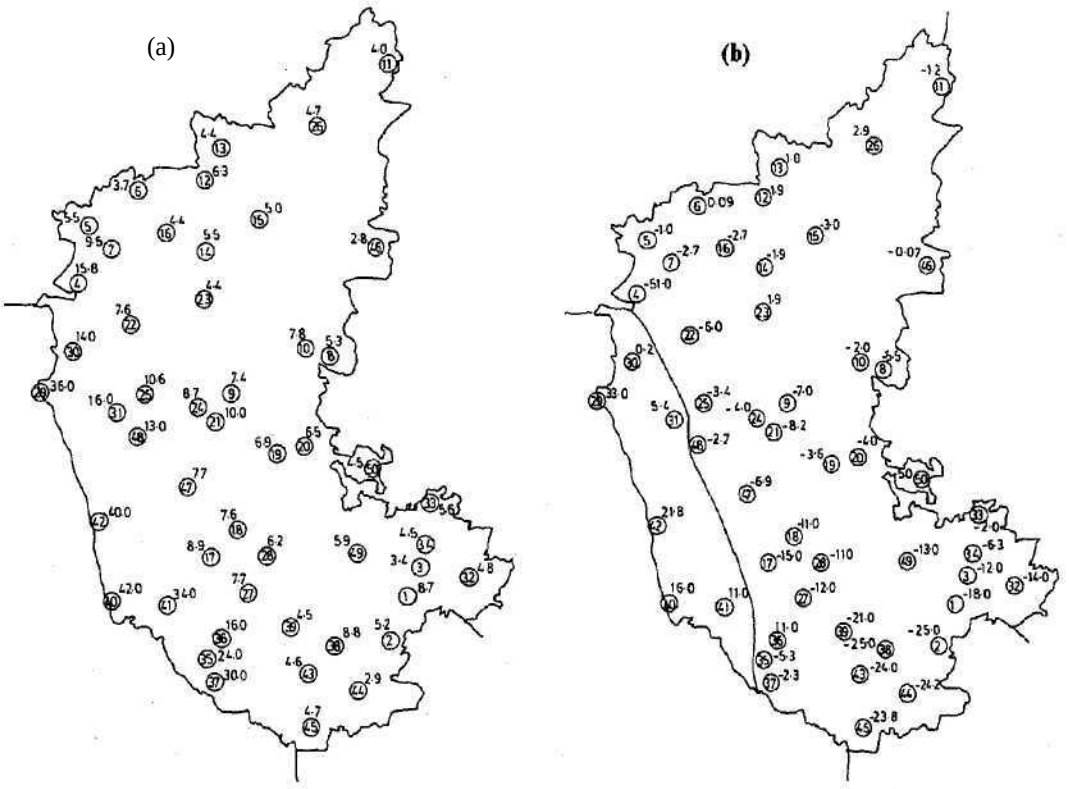

Figure 9. a. First eigenvector - premonsoon. Variance explained = 40·50%; b. second eigenvector - premonsoon. Variance explained = 11·30%.

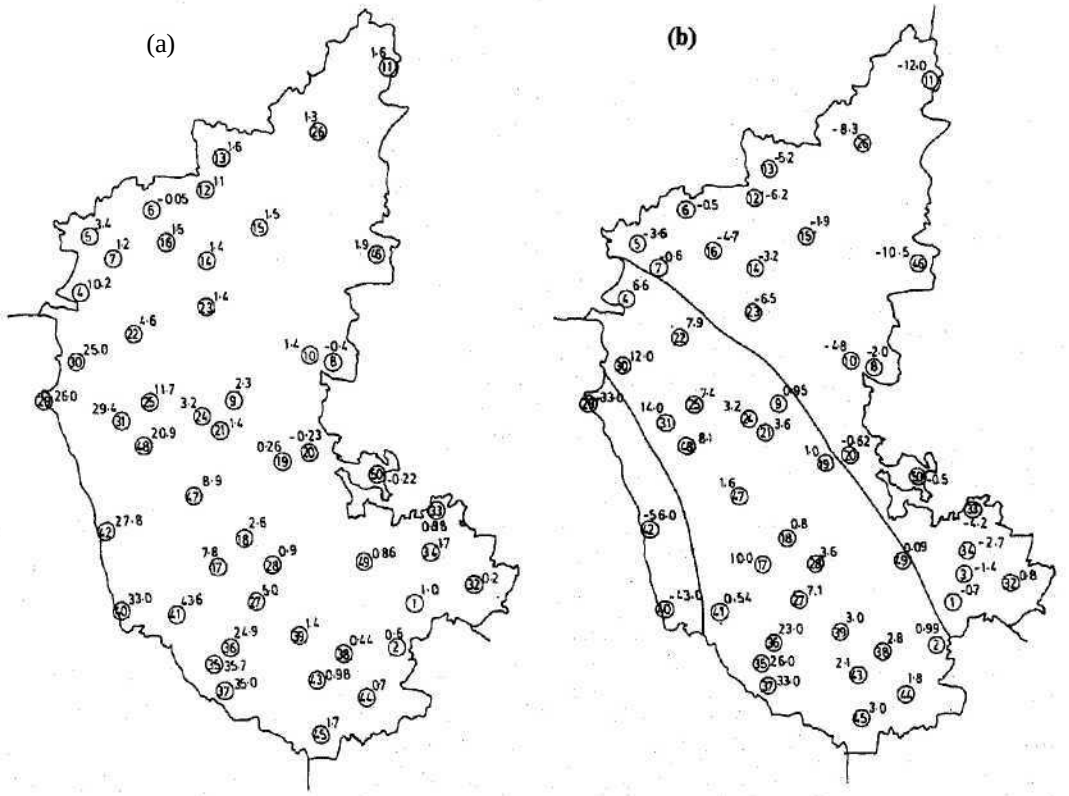

Figure 10. a. First eigenvector - SW monsoon. Variance explained = 47·07%; b. second eigenvector - SW monsoon. Variance explained = 10·85%.

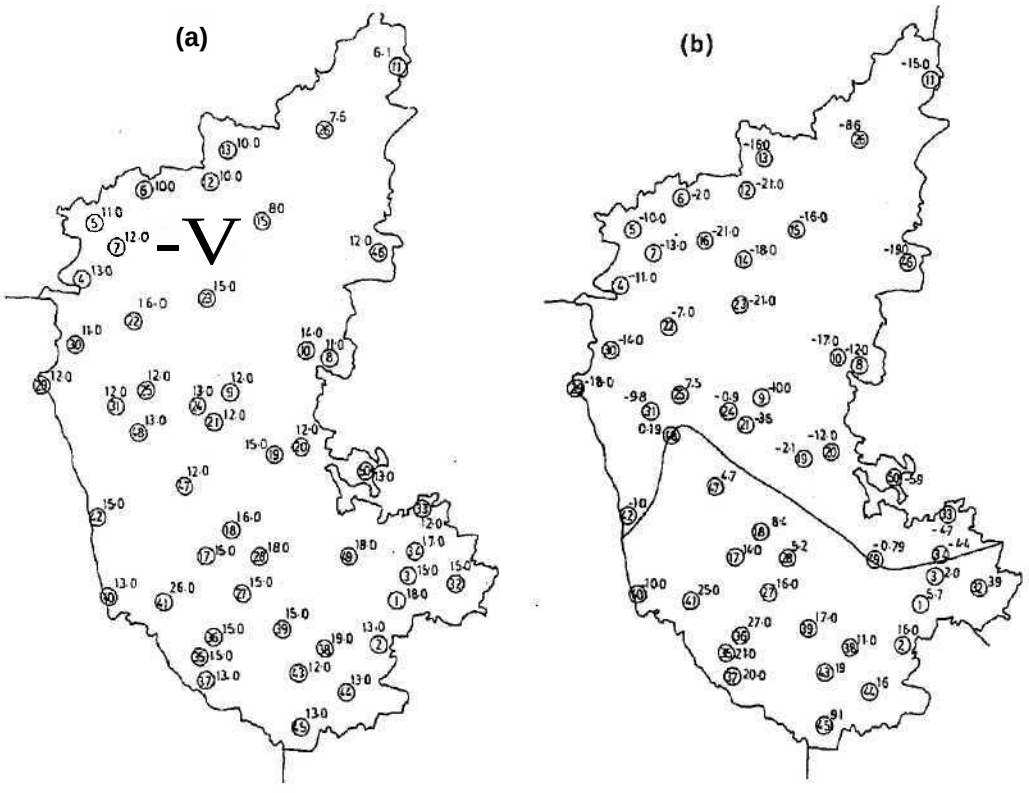

Figure 11. a. First eigenvector - NE monsoon. Variance explained = 46·44%; b. second eigenvector - NE monsoon. Variance explained = 7·86%.

Table 4. Frequency of annual sign sequences (SWM $N = 81$ yrs).

	Sign								
	++		--		−+		+−		
Comp.	Obs.	Expt.	Obs.	Expt.	Obs.	Expt.	Obs.	Expt.	χ^2 Obs.
PC1	13	15·55	22	19·45	23	20·45	23	25·55	1·23
PC2	29	22·30	14	2070	13	19·70	25	18·30	8·90
PC3	20	19·75	20	20·25	20	20·25	21	20·75	0·01
PC4	11	15·12	24	19·88	24	19·88	22	26·12	3·48

indicate the presence of annual signals. In table 4 the frequencies of the sign sequences for the four PC's of the SWM rainfall are shown and their significance is tested. This table shows that the first PC has no pattern on the annual scale. But the evolution of the second component cannot be dismissed as due to chance. Similar tests on NEM components show that again the second PC cannot be a purely random time series. In figure 12 the second PC of the SWM data is presented. The results of the above test can be interpreted to mean that the Karnataka State monsoon rainfall through its first dominant component evolves on scales of the order of a month and less. The second component of the SWM represents a pattern with characteristic time as a year or a multiple of it. In fact, from figure 12 it would seem that this has a predominant

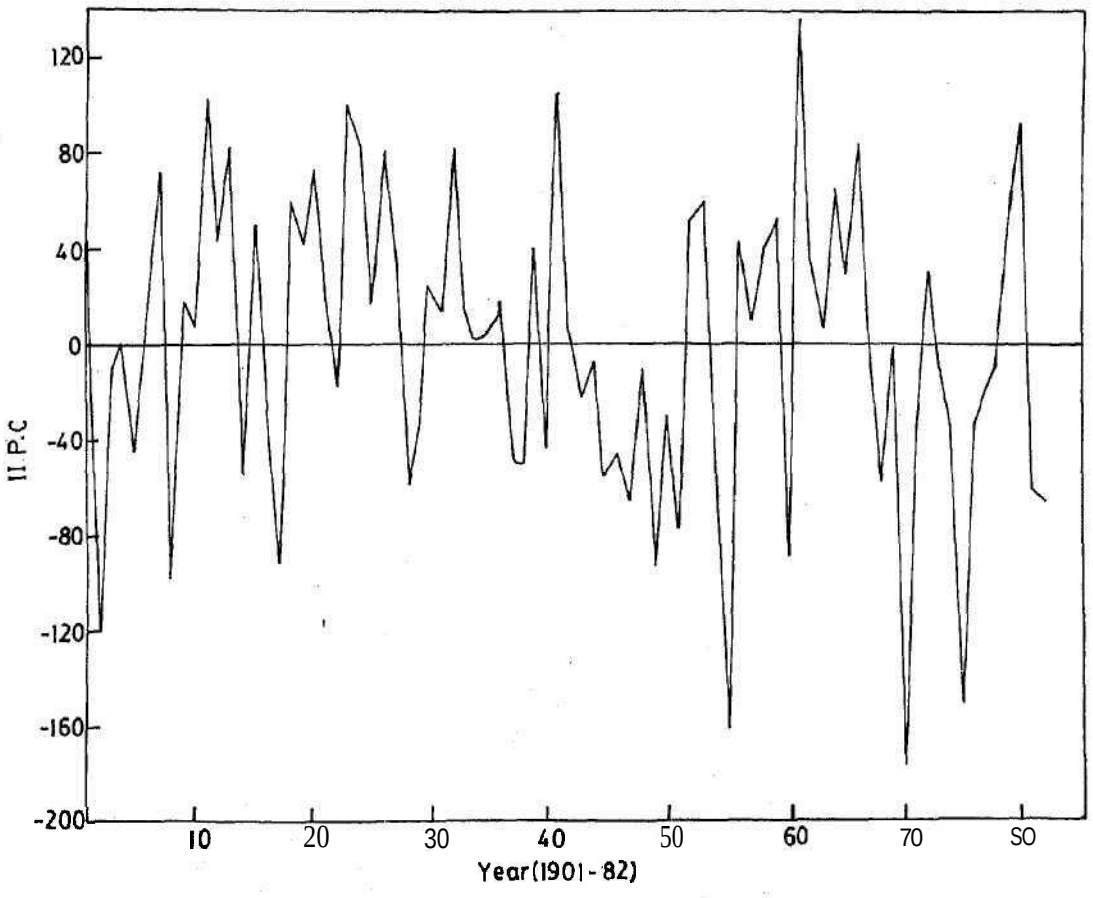

Figure 12. Second principal component of SWM rainfall in Karnataka.

period of nearly three years. This component is seen to persist with the same sign for 2 to 3 years before a change takes place.

5.2 *Interseasonal variability*

The eigenvectors of the premonsoon, SWM and NEM rainfall have been presented in figures 9, 10 and 11. The associations between the seasons can again be studied conveniently through the principal components. It is found that only the second principal components of the SWM and the NEM rainfall show a mutual connection. The sign sequence transition for this case leads to an observed χ^2 value of 4·92 which is significantly higher than the tabulated value of 3·84. This trend is in conformity with the significant dependence in the transitions from September to October as shown in table 3.

6. Grouping the years

When rainfall over a large area is considered, the current practice is to arrive at an area rainfall value as a weighted average of the rainfall at the individual stations. It may be observed that the first PC is already a dominant weighted average of the station rainfall, and is a good measure of the areal rainfall. Further, since the second

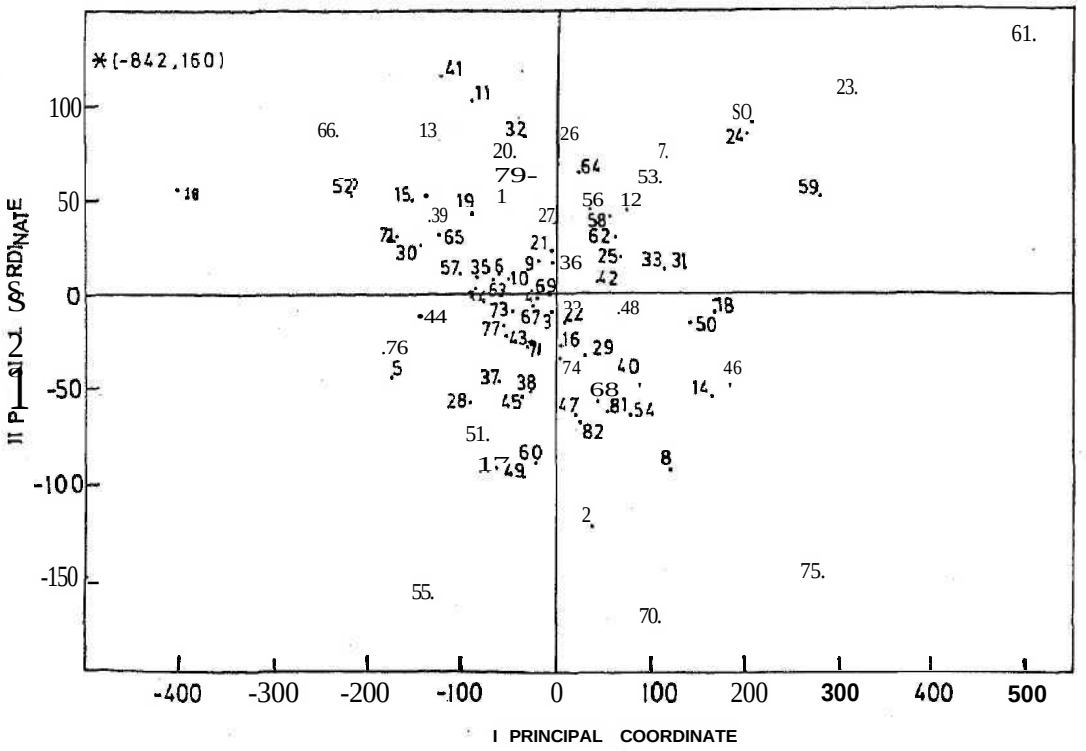

Figure 13. Variability of the principal components of the SWM rainfall in Karnataka 1901-1982. * — zero rainfall.

component is also always statistically significant, PC_1 and PC_2 on any time-scale are the two most important characteristics of rainfall in a given year for the whole network of stations. Thus, with PC_1 and PC_2 as coordinates the past year's data can be represented on a diagram. Such a representation produces a convenient way of comparing the years as in figure 13 for the SWM rainfall. The ideal normal year, i.e. when each station receives exactly its own normal rainfall, has all principal components as zero. Such a year coincides with the origin in figure 13. All the data have been marked in this figure and it is easy to see that the so-called normal years fall around the origin. Years with excessive rainfall like 1961 have large positive PC_1 and PC_2 values. The hypothetical zero rainfall year when there is no rainfall at any of the stations has coordinates $(-842, 160)$. Further one can mark years with prescribed percentage variations about the normal rainfall on this figure. Nearness of two or more years on this diagram indicates that for these years the atmospheric conditions could have been similar. Such information could help in foreshadowing droughts and floods.

7. Analysis in a coherent zone

The study presented so far referred to an area which in terms of either the climate or the topography is not homogeneous. Thus, it would be relevant to ask whether the variability patterns found on different time scales for the State of Karnataka as a whole would be also valid for smaller regions. A more interesting question would be whether the interannual signals which may be too weak to be detected statistically in a large inhomogeneous region become stronger if the principal components are

found for a coherent rainfall zone. With this in view, a set of ten stations from the western region of Karnataka, referred henceforth as the west zone (WZ) is considered. The stations are: Mangalore, Kundapur, Karwar, Supa, Sirsi, Soraba, Belthangadi, Mercara, Somwarpet and Virajpet. The principal components for this set of stations have been found as explained earlier for the period 1901-1980. Here only some limited results regarding the variability trends on monthly and annual scales are studied. in tables 5 and 6 the frequency of sign sequences for the first and second principal components of monthly rainfall is presented. From table 5, it is seen that the transition of PC_1 from June to July shows a significant χ^2-value, a trend also present for the whole State (table 1). While the Karnataka data show significant transitions from September to October in PC_1 and PC_2, this trend is weakened in the west zone data. This behaviour seems reasonable because with the onset of the NEM rainfall in October, the eigenvector patterns (figures 12 and 13) change and the dominance of the western region is reduced. This line of argument would indicate that for the SWM seasonal rainfall, the interannual variability trends, if present, should be better detectable in the PC's of the WZ, than in the PC's of the entire Karnataka data. This hypothesis is verified in table 7 through the sign sequences of the SWM rainfall principal components. It is interesting to observe that while the observed χ^2 value for the all-Karnataka data is only 1·23, for the coherent WZ this value is 4·69, which is conspicuously significant. Thus, it may be seen that the annual signal of the first PC

Table 5. Frequency of sign sequences in the I PC of WZ monthly rainfall.

	Sign								
	++		+−		−+		−−		χ^2
Comp.	Obs.	Expt.	Obs.	Expt.	Obs.	Expt.	Obs.	Expt.	Obs.
April-May	9	11·10	28	25·90	15	12·90	28	30·10	1·05
May-June	12	10·20	12	13·80	22	23·80	34	32·20	0·79
June-July	11	15·30	23	18·70	25	20·70	21	25·30	3·82
July-August	16	13·95	20	22·05	15	17·05	29	26·95	0·89
Aug–Sept	1J	12·01	20	18·99	20	18·99	29	30·01	0·23
Sept-Oct	16	13·18	15	17·83	18	20·83	31	28·17	1·72

Table 6. Frequency of sign sequences in the II **PC** of WZ monthly rainfall.

	Sign								
	++		+−		−+		−−		χ^2
Comp.	Obs.	Expt.	Obs.	Expt.	Obs.	Expt,	Obs.	Expt.	Obs.
April-May	22	22·55	19	18·45	22	21·45	17	17·55	0·06
May-June	21	21·45	23	22·55	18	17·55	18	1845	0·04
June-July	20	19·50	19	19·50	20	20·50	21	20·50	0·05
July-August	19	19·50	21	20·50	20	19·50	20	20·50	0·05
Aug-Sept	14	17·55	25	21·45	22	18·45	19	22·55	2·55
Sept–Oct	22	18·00	14	18·00	18	22·00	26	22·00	3·23

Table 7. Frequency of annual sign sequences WZ SWM rainfall (N = 79 yrs).

	Sign								
	++		+-		-+		--		χ^2
Comp.	Obs.	Expt.	Obs.	Expt.	Obs.	Expt.	Obs.	Expt.	Obs.
PC1	13	17·80	24	19·20	25	20·20	17	21·80	4·69
PC2	23	17·33	14	19·67	14	19·67	28	22·33	6·56
PC3	20	17·80	17	19·20	18	20·20	24	21·80	0·99
PC4	25	22·86	17	19·14	18	20·14	19	16·86	0·94
PC5	23	21·28	18	19·72	18	19·72	20	18·28	0·60

of the SWM rainfall is enhanced in the WZ rainfall data. Moreover, the PC_2 of both the entire State and the WZ rainfall show significant annual transitions.

8. Predictability

A question closely connected with rainfall variability is one of predictability. If the variability, which is a deviation of rainfall about its long-term average value, is not purely random, one expects a temporal relationship to be detectable. The most desirable relationship is a linear one. But, in the present context it has been pointed out that monthly rainfall anomalies show no significant autocorrelations. Thus, linear relationships for time-wise evolution usually get rejected by appropriate statistical tests. On the other hand, it is not obvious what kind of statistical methodology one should adopt to detect and test nonlinear relations. Principal component analysis does not provide a direct answer to this question. But, as principal components are found to posses statistically significant trends it may be more appropriate to first predict the principal components and then foreshadow the rainfall in terms of past data with the help of a diagram like figure 13. The region ideally suited to attempt this kind of predictability is the coherent west zone of Karnataka. In this zone PC_1 and PC_2 of the SWM rainfall show significant annual transitions and one can ask the probability of the next year PC being above/below average (+ or −), if in the present year it is above/below average (+ or −). From table 7 the two-state transition probability matrix for PC_1 and PC_2 are found to be:

$$[P]_{1w} = \begin{bmatrix} 0{\cdot}35 & 0{\cdot}65 \\ 0.60 & 0{\cdot}40 \end{bmatrix} [P]_{2w} = \begin{bmatrix} 0{\cdot}62 & 0{\cdot}38 \\ 0{\cdot}33 & 0{\cdot}67 \end{bmatrix}.$$

Such quantification helps one to understand the physical significance of PCA, which can also be interpreted as a modal decomposition of a multivariate rainfall time series. Now, it is easy to see that PC_1 stands for an annual oscillatory mode, whereas PC_2 stands for a persistence mode. This interpretation is true only for the west zone data. For Karnataka as a whole with the present data, the oscillations in PC_1 are attributable to chance and hence prediction through a transition probability is not justified. PC_2 of the State data has significant transition probability given by

$$[P]_{2s} = \begin{bmatrix} 0{\cdot}67 & 0{\cdot}33 \\ 0{\cdot}34 & 0{\cdot}66 \end{bmatrix}$$

This probability matrix is almost the same as the $[P]_{2w}$ of the west zone. Thus, although the second eigenvector of the State and the WZ are spatially of secondary importance, the corresponding PC_2 time series stands for a stable persistence mode valid for a large spatial region. The question whether the prediction of a secondary component is of importance in forecasting the actual rainfall needs further investigation. But it may be pointed out that even if PC_2 is an atmospheric signal just present in the rainfall time series, it gives one coordinate in locating an year on the PC diagram of figure 13. However, without proper prediction of PC_1 which is of primary importance, knowledge of PC_2 may not be of much practical use. That PC_1 is directly related to the area rainfall is easily demonstrated as follows. Let the area rainfall R_t be defined as the arithmetic average of the rainfall at each station ($j — 1, 2 \ldots M$). Thus,

$$\bar{R}_t = (1/M)\sum_{i}^{M}\sum_{j} P_{jt}\phi_{ij} + (1/M)\sum_{j=1}^{M} m_j$$
$$= (1/M)\Sigma P_{jt}\bar{\phi}_j + \bar{m}. \tag{7}$$

The correlation between $\bar{R}_t$ and the fcth principal component is

$$\langle \bar{R}_t P_{kt} \rangle = \lambda_k \bar{\phi}_k. \tag{8}$$

Hence the linear correlation coefficient between $\bar{R}_t$ and P_{kt} is

$$\rho_k = [\lambda_k^{1/2}\bar{\phi}_k] \Big/ \left[\sum_{j=1}^{M} \lambda_j \bar{\phi}_j^2\right]^{1/2} \quad \phi_i = \sum_{i=1}^{M} \phi_{ij}. \tag{9}$$

Whenever the eigenvector elements are of the same sign, ρ_k will be very nearly equal to + 1. In the present analysis it has been found that the first eigenvector field rarely exhibits spatial contrast. Thus $P_{1t}(PC_1)$ will be highly correlated with the area rainfall time series. In fact PC_1 itself can be taken as a measure of the area rainfall. For the second and higher eigenvectors the elements change sign often leading to small values of $\bar{\phi}_j$. This would lead to lower or insignificant correlation between $\bar{R}_t$ and the higher principal components. However, if instead of the complete data network only part of the stations which have the same sign in their eigenvectors are considered, the area rainfall for these special regions will still have significant correlations with the corresponding principal components. In table 8 the ρ_1 value for the WZ rainfall is presented for monthly and SWM data. The strong correlation between PC_1 and the rainfall leads to the inference that the transition probability $[P]_1$, when significant,

Table 8. Correlation coefficient between rainfall and PC_1 for WZ.

Data	ρ_1
SWM	0·9976
April	0·9945
May	0·9938
June	0·9977
July	0·9810
August	0·9973
September	0·9937
October	0·9926

Table 9. 80 year station average rainfall in cm. for the WZ.

No.	Station	June	July
1	Karwar	95·98	102·57
2	Supa	40·60	95·19
3	Sirsi	52·17	99·59
4	Mercara	60·14	112·76
5	Somwarpet	32·63	77·71
6	Virajpet	55·67	91·55
7	Mangalore	96·26	104·81
8	Belthangadi	94·41	158·52
9	Kundapura	103·44	122·30
10	Soraba	28·72	61·17

can be taken as an overall feature of the rainfall. Thus, for the WZ monsoon rainfall an above-average-rainfall-year will be followed by a below-average-rainfall-year with 65% probability. However, when a given year is below average the following year would be above average with only 60% probability. This skewness in the oscillations of rainfall is an interesting feature which has come out systematically through the present analysis. In the intraseasonal study of table 5 for the WZ only the June-July transition for PC_1 comes out as significant with

$$[P]^{JJ}_{1w} = \begin{bmatrix} 0{\cdot}32 & 0{\cdot}681 \\ 0{\cdot}54 & 0{\cdot}46 \end{bmatrix}.$$

This is an interesting transition in that it states that given the June rainfall to be above normal, July rainfall has a high probability of being below normal. On the other hand, if in June the rainfall is below normal, no predictive tendencies exist, as there is an almost equal chance for July to continue to be below normal, or become above normal. To check the above transition probability, a prediction exercise is undertaken for the 10 stations of the WZ for July. For this purpose the June and July data of 1981 to 1985 not included in the previous analysis are used. In table 9, the information on eighty year normal rainfall for the WZ stations is presented.

In table 10 the prediction of the July rainfall, whenever the June rainfall is above normal is presented and compared with the observed July rainfall. It is to be noted that when the June rainfall is below normal, no prediction is possible according to the June-July transition probability. Such cases are indicated as +/— in table 10. From table 10 it is observed that there have been 28 cases of June rainfall being above normal in the five years considered here. For all these cases based on the transition probability matrix $[P]^{JJ}_{1w}$, July rainfall is predicted to be below average. This prediction is seen to be correct in 27 out of the 28 cases.

9. Discussion

The popular approach in time series studies is that of autocorrelation and power spectrum analysis. One faces several difficulties in understanding monthly rainfall

Table 10. Prediction of July rainfall given rainfall in June.

+: above average
−: below average

Year	1981			1982			1983			1984			1985		
Stn.	June Given	July Pred.	July Obsd.	June Given	July Pred.	July Obsd.	June Given	July Pred.	July Obsd.	June Given	July Pred.	July Obsd.	June Given	July Pred.	July Obsd.
1	+	—	—	—	+/—	+	—	+/−	+	+	—	—	+	—	
2	+	−	−	−	+ / −		− + −			− + −		−	+	−	=
3	− −	+ / −	−	−	+ / −		+ + −			−	+	−	− + −		−
4	—	+/—	—	-	+/−	-	-	+/-	-	+	-	—	+	—	
5	_	+/_	_	_	+/−	_	_	+/−	_	+	_	_	+	—	_
6	+	−	−	−	+/−	−	−	+/−	−	+	−	−	+	−	−
7	+	−	+	+	−	−	−	+/−	+	+	−	−	+	−	−
8	−	+/−	−	+	−	−	−	+/−	−	+	−	−	+	−	−
9	+	-	-	-	+/—	—	-	+/—	—	+	-	—	+	-	−
10	−	+/−	−	−	+/−		+				+/−	−	−	+/−	−

time series data through classical spectrum analysis. First, a large network of station data will have to be simultaneously analysed for their cross-spectral densities as was done by Hartmann and Michelsen (1989). As the sample time series are highly correlated among themselves due to spatial coherence, results of a straightforward spectral technique would be cumbersome, if not difficult to interpret. On the other hand, if each station data are analysed individually, the spatial structure is lost, which may be important in enhancing the temporal signals. In most cases, the monthly station data will be identified as white noise, meaning that the temporal variation is purely due to chance. Since autocorrelation/power spectrum analyses study linear tendencies, they are not strong enough particularly with non-gaussian data to show nonlinear temporal tendencies. This, in turn, demands more complicated higher order spectral analysis like bispectrum computations. Sometimes the argument is put forth that instead of looking at individual station data, as the atmospheric system is organized over large spatial scales, one should analyse area rainfall. While this is reasonable, it is not clear whether the official area rainfall values put forth by government agencies, which are either arithmetic averages or area weighted averages, are the right data for studying the natural variability patterns. Principal component analysis steers clear of these shortcomings, retaining at the same time the simplicity of a linear system analysis. Thus, the first principal component can represent the area rainfall objectively, as the weight for the various stations are assigned by the data itself in an optimal way. Again, in PCA a large number of station data can be simultaneously handled to account for spatial variability, but invariably the final number of components to be studied will be much less than the total number of stations. The classical power spectrum analysis is a Fourier decomposition, wherein the energy contained at many frequencies are found. PCA can be thought of as a generalized Fourier decomposition of a random field. Even though identification of a periodicity is not directly possible, the energy contained in different components is extracted as the eigenvalues of the covariance matrix. The present case study of Karnataka data demonstrates the application of PCA in understanding monthly and seasonal rainfall variability. Figure 2 shows how the significance of eigenvalues can be systematically checked to arrive at the number of principal components to be retained for further work. It is interesting to observe that not more than four components are required to represent the rainfall over the size of a state as large as Karnataka. It may be pointed out here that there is a popular misconcept that unless the cumulative percentage of variance explained by the first few components in very high, say of the order of 90%, PCA is not useful, in rainfall studies. Such a view is, however, unjustified as shown by the present study. In this context, it is important to discriminate between spatial connections and temporal variability. PCA formally represents M-number of given time series data just as a linear combination of another M-number of time series. But, the advantage lies in the fact that since the data are neither perfectly spatially correlated, nor exactly uncorrelated, after the first few terms the decomposition loses its power to discriminate the remainder field from a purely random (white noise) field. Hence the terms within this cut-off limit should contain the temporal variability characteristics valid for the complete station network, although in a transformed fashion. The advantage of this is apparent when one observes that for Karnataka SWM rainfall, the first eigenvector explains less than 50% of spatial variance; but the PC_1 and the area rainfall are correlated with $\rho_1 = 0{\cdot}9805$. Again, for the west zone, this correlation coefficient is consistently very high as shown in table 8. Similarly, the

second and other significant PC's are connected to the area rainfall in regions wherein the corresponding eigenvector elements have the same sign. Thus, temporal signals that may be present over a large spatial regime would be carried over into the first few principal component time series after automatically eliminating what may be termed spatial noise. This interpretation also points out a limitation of PCA, namely, that it is necessary to establish a clear-cut quantitative relationship between rainfall and the PC's before one can effectively use this approach. In this study due to space limitations only a simple representation of the years (figure 13), which gives an intuitive comparison between concepts like drought years, normal years and flood years in terms of the PC's is presented. However, the simple probabilities proposed here for above/below average transitions are found to be significant and consistent.

10. Summary and conclusion

Principal component analysis produces a decomposition of the data field into spatial eigenvectors and a temporal time series. While EOF studies are quite common in meteorological data analysis, the usefulness of the principal component time series in understanding temporal variability of rainfall has not received attention in the past. The present investigation is motivated by the possibility that the first few PC's may contain valuable information regarding the interseasonal, intraseasonal and annual rainfall variability. The monthly rainfall data of Karnataka spread over 50 stations for a period of 82 years show that PCA is a valuable aid in gaining insight into temporal patterns through transition probabilities of the first and second principal components. For the State as a whole, the rainfall variations in May, June, July, September and October are sequentially related. Transitions of fluctuations from July to August and again to September are purely due to chance. The connections-between the variability in the premonsoon, SWM and NEM rainfall are generally attributable to chance, except for the connection between the second principal components of the SWM and NEM data. Again, the Karnataka SWM second PC exhibits significant interannual transitions, whereas the first PC shows no significant trend. However the coherent west zone seems to carry the interannual variation signal of the SWM in a stronger manner since even the first PC of the WZ data shows a statistically significant annual transition, different from chance. The preliminary exercise for predicting the June-July transition in the five years 1981-85 through an estimated transition probability has been surprisingly successful. However, further detailed analysis is required to quantify predictability of the PC's as forecastable signals of impending rainfall variations.

Acknowledgements

The author thanks Prof. R Narasimha, Prof. Sulochana Gadgil and other colleagues for many useful discussions. The author has received help from Ms Yadumani, Ms Asha Guruprasad and Mr P Basak in the computations.

References

Bedi H S and Bindra M M S 1980 Principal components of monsoon rainfall; *Tellus* 32 296-298

Fleer H E 1977 Teleconnections of rainfall anomalies in the tropics and subtropics; in *Monsoon dynamics* 1981 (eds) J Lighthill and R Pierce (Cambridge: Univ Press) pp. 1-18

Gadgil S, Gowri R and Yadumani 1988 Coherent rainfall zones: case study for Karnataka; *Proc. Indian Acad, Sci, Earth Planet Sci,* 97 63-79

Gadgil S and lyengar R N 1978 Cluster analysis of rainfall stations of the Indian peninsula; *Q. J. R. Meteorol. Soc.* 106 873-886

Hartmann D L and Michelsen M L 1989 Intraseasonal periodicities in Indian rainfall; *J. Atmos. Sci,* **46** 2838-2861.

Hastenrath S and Rosen A 1983 Patterns of India monsoon rainfall anomalies; *Tellus* A35 324-331

lyengar R N 1982 Stochastic modelling of monthly rainfall; *J. Hydrol.* 57 375-387

lyengar R N 1987 Statistical analysis of weekly rainfall; *Monsoon* 38 453-458

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

Lorenz E N 1956 Empirical orthogonal functions and statistical weather prediction, Sci, Rept. No. 1, Stat. Fisec. Proj. MIT, Camb. Mass USA

Overland J E and Preisendorfer R W 1982 A significance test for principal components applied to cyclone climatology; *Mon. Weath. Rev.* 110 1-4

Preisendorfer R W 1981 Cumulative probability tables for eigenvalues of random covariance matrices; SIO Rf series 81-2, Scripps Institution of Oceanography

Preisendorfer R W, Zwiers F W and Barnett T P 1981 Foundations of principal component selection rules; SIO Rf Series 81-4, Scripps Inst. of Oceanography

Rakecha P R and Mandal B N 1977 The use of empirical orthogonal functions for rainfall estimates in *Monsoon dynamics* 1981 (eds) J Lighthill and R Pierce (Cambridge: Univ Press) pp. 627-638

Rohatgi V K 1984 *Statistical inference,* (New York: John Wiley)

Shukla J 1986 Interannual variability of monsoon; in *Monsoons* (eds) J S Fein and P L Stephens (New York: Wiley-Interscience)

Singh S V and Kripalani R H 1986 Application of extended empirical orthogonal function analysis to interrelationships and sequential evolution of monsoon fields; *Mon. Weath, Rev.* **114** 1603-1610.