Application of Principal Component Analysis in Understanding Variability of Monsoonal Rainfall in West Bengal

Pijush Basak
Narula Institute of Technology, Kolkata
E-mail: pijushbasak@rediffmail.com

ABSTRACT

The Principal Component Analysis (PCA) is utilized in realizing the temporal variability of South Monsoon (SWM) rainfall. The monthly rainfall data of West Bengal spread over 21 stations for period of 60 years have been analyzed for intra-seasonal and inter-annual variability. The study brings out statistically significant inter-annual signals in the monthly and SWM rainfall. Conditional probabilities are prepared for a few normal / below normal transitions. A sample prediction exercise for July-August using such transition probability has been found to be successful.

Keywords: *South West Monsoon rainfall, Rainfall variability, Principal Components, Empirical orthogonal function, Eigenvalue, Spatial structure, Predictability, Transition probability.*

1. Introduction

Rainfall is one of the most important phenomenona of monsoon season. The amount of rainfall fluctuates in days, weeks, months and seasons over a wide range. The question however remains: whether the variations are purely random or there remains an identifiable pattern in variations. A variation is perhaps, the fluctuation about a long term average value. The variability may be on several time scales, such as days, weeks and months or diverse spatial scales, such as station, district or state. The South West Monsoon (SWM), organized spatially over large scale, persistent in time for several months. It would thus be useful to study the data on few optimum scales. In the present investigation, the monthly rainfall data of West Bengal has been considered. The preliminary statistical information is available. The autocorrelations and power spectral densities of the stations are obtained (Chandra and Dhar, 1975; Basu, 2001; Basu et. al., 2004 and Basak, 2014). It is revealed that they are mostly white noise processes except a few cases. In fact, no temporal pattern emerges in monthly rainfall at station level. However, as the inter-station data are known to be spatially correlated, there may be some kind of trend that could be identified. The present work is connected with both spatial and temporal variation by decomposing the large scale data into principal components (PCs) in time and empirical orthogonal function (EOF) in space. Earlier, few works in this respect in All India level are Bedi and Bindra (1980), Hastenrath and Rosen (1983), Iyenger and Basak (1994); for north-east India Mahapatra et al. (2001); for Karnataka, Iyenger (1991); for West Bengal, Basak (2014). The main emphasis in this paper is to locate spatial structure in the field and the temporal pattern detectable in the data. In the present study it is shown that PCs can be used to compare and if necessary, group the 'years'. The PC of monthly and seasonal data reveals interesting information about intra-seasonal and inter-annual variability.

2. Data

The analyzed data in current investigation are the monthly rainfall data of 21 selected stations spread over West Bengal and extending over 60 years from 1901 to 1960. The stations considered in West Bengal are presented in Fig. 1 and corresponding details are presented in Table 1. While it would be reasonable to consider more number of stations, there are restrictions due to data-gaps and unequal length of time series. Moreover, it is not clear whether inclusion of more number of stations would enhance or dilute the signal that may be present. Thus, a skeleton number of stations are considered in the study. The state of West Bengal is of considerable interest, as two meteorological subdivisions of Indian Meteorological Department, namely, Gangetic West Bengal (GWB) and Sub-Himalayan West Bengal (SHWB) are in West Bengal. The GWB receives about 60% of SWM rainfall namely 9000 mm. Regarding monthly analysis, SHWB receives maximum rainfall in July, accounting for 40% of total

TABLE 1
Stations detail with tests of Gaussianness and trend

Sl. No.	Station Name	Latitude/Longitude	Sub-division	K-S statistics	Mann-Kendall ®
1.	Jalpaiguri	26.53N,88.72E	SHWB[#]	-0.5182	0.0249
2.	Alipurduar	26.47N,89.55E	SHWB	-1.7619	0.1729*
3.	Darjeeling	27.10N,88.30E	SHWB	-1.1400	-0.1910*
4.	Kalchini	26.41N,89.25E	SHWB	-2.0728*	-0.0260
5.	Malda	25.03N,88.13E	SHWB	-0.5182	0.0791
6.	Kishanganj	26.12N,87.93E	SHWB	-0.5182	0.0667
7.	Mongpo	26.90N,88.50E	SHWB	0.1036	0.0249
8.	Mathabhanga	26.35N,89.22E	SHWB	-0.5182	0.1582
9.	Amta	22.58N,88.02E	GWB[$]	-1.4510	0.0655
10.	Arambag	22.88N,87.78E	GWB	0.4146	-0.1612
11.	Budge Budge	22.48N,88.18E	GWB	-0.5182	-0.0124
12.	Bongaon	23.07N,88.82E	GWB	1.1400	0.0576
13.	Burdwan	23.25N,87.85E	GWB	0.1036	-0.0927
14.	Ghatshila	22.60N,86.50E	GWB	0.0364	0.0226
15.	Sagar Island	21.65N,88.05E	GWB	1.1400	0.1175
16.	Kukrahati	22.18N,88.12E	GWB	0.1036	-0.0689
17.	Ranaghat	23.18N,88.55E	GWB	0.1036	-0.1559
18.	Uluberia	22.47N,88.12E	GWB	-1.4509	0.1175
19.	Vishnupur	23.08N,87.32E	GWB	-2.0728*	0.0339
20.	Kharagpur	25.12N,86.55E	GWB	0.7255	0.0508
21.	Silda	63N,86.80E	GWB	-1.1400	0.0847

*Significant at 5% level. [#]Sub-Himalayan West Bengal [$] Gangetic West Bengal

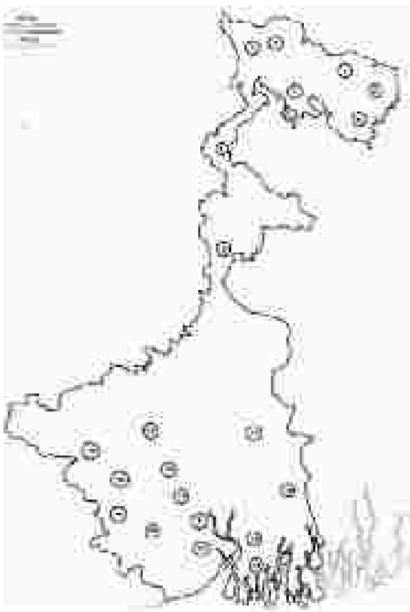

Fig.1 Station Data Network of West Bengal

SWM period. It is followed by June, August and September. With this in view, monthly PCA analysis is carried out for SWM period (June-September) of the stations of West Bengal.

3. Method of Analysis

The state wise data matrix of size 21x60 has been used for analysis. Firstly, the statistical properties such as mean, standard deviation, skewness and kurtosis are evaluated for each station. For Principal Component Analysis (PCA) the mean centered data time series is analyzed to find the principal components (PC) as presented (Gnanadesikan, 1977).

Let, R_{it} be the actual rainfall at station i (i=1, 2..., M) in the year t (t=1,2,...,N), then, the centered data series are

$$r_{it} = (R_{it} - m_i), \quad m_i = \left(\frac{1}{N}\right)\sum_{t=1}^{N} R_{it}$$

The covariance matrix is constructed as

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

The orthonormal eigenvalues (λ_j) of the symmetrical matrix are extracted such that the j^{th} vector $\{\phi_{ij}\}$ corresponds to the j^{th} largest eigenvalue λ_j of the covariance matrix.

The rainfall anomaly at station i in year t can be represented as orthogonal decomposition in terms of principal components in time and empirical orthogonal function (EOF) in space, namely,

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

The principal components are defined as

$p_{jt} = \quad r_{it}\phi_{ij}$

This transforms the original time series r_{it} into the new time series p_{jt}, which also reflects the spatial variation of the original series. The first few principal component series p_{jt} usually account for a large proportion of the spatial variation contained in the data set. It is found that p_{jt} can be used to extract the temporal variability in the data while the eigenvectors $\{\phi_{ij}\}$ represent spatial patterns underlying the data.

The percentage of variance explained by the eigenvalues for each of the months is presented in Table 2. It is found that for all the months first 4 eigenvalues (λ_j, j=1,..,4) accounts for 27-34%, 15-19%, 7-12% and 6-8% of total variance (Table 2). The third and fourth eigenvalues contribute to only 7-12% and 6-8% respectively. The first four eigenvalues taken for all the months June-September contribute about 60-70% of total variance.

4. Monthly Rainfall

Monthly rainfall patterns present an interesting feature as indicated in Table 2 for the first and second eigenvalues respectively. While the first eigenvector (e.v.) dominates the spatial structure, it is observed thatis maximum in June. The feature is followed by a gradual decrease from July to September. Also, for the second eigenvalue, the next dominant spatial structure increases from May to reach a peak in June. This is followed by a decrease in August.

A better view of how the rainfall field is getting organized is provided by the eigenvectors (e.v.) shown in Fig. 2(a)-(b) to Fig. 5(a)-(b). Here, first two e.v.s are shown. As the first e.v. is always predominant, the month–to-month transition would be of importance. It is seen that the whole state is spatially correlated (except Jalpaiguri, Alipurduar and Kalchini in northern part) in June. This means that above/below normal fluctuation along the southern part which has the largest weight, would indicate similar trends in other part of the state. The picture changes in July when the first e.v. develops a spatial contrast dividing the state into 3 regions. In the northern part of the state, there are two regions (with positive and negative e.v.) and in southern part, a region of positive e.v. It may be

TABLE 2

Result of Monthly PCA of Stations: First Ten Eigen-values& Cumulative percentage of variance explained

Sl.No.	JUNE		JULY		AUGUST		SEPTEMBER	
	Eigen-val.	% var. expl.	Eigen-val.	% var. expl.	Eigen-val.	% var expl.	Eigen-val.	% var expl.
1.	7.1296	33.9500	6.3312	30.1485	6.1916	29.4840	5.7487	27.3748
2.	3.2237	49.0909	3.8987	48.7135	3.6590	46.9079	3.2662	42.9281
3.	2.4307	60.8758	1.7398	56.9984	1.6364	54.7005	2.0409	52.6452
4.	1.5353	68.1868	1.3682	61.5136	1.4067	61.3993	1.4042	59.3320
5.	1.0270	73.0774	1.2616	69.5211	1.1603	66.9248	1.3158	65.5977
6.	0.7453	76.6266	1.0529	74.5349	1.1172	72.2451	1.1298	70.9778
7.	0.6775	79.8528	0.9425	79.0230	0.9899	76.9590	0.9402	75.4549
8.	0.6088	82.7519	0.7303	82.5008	0.7925	80.7238	0.7513	79.0323
9.	0.6088	85.5254	0.6267	85.4849	0.6213	83.6915	0.6308	82.0363
10.	0.5825	87.7682	0.5726	88.2117	0.5681	86.3969	0.5773	84.7489

interpreted wherein above/below rainfall in region of positive e.v. would indicate below/average rainfall in the region of negative e.v. The pattern intensifies in August and contrast matures to grow to two regions of contrast. From the southern part to the fringe of the northern hill and from the northern hills along with doors area, there are two regions of contrast. In September, the pattern of August gets restored. It clearly indicates growth, maturity and development of dominant pattern of rainfall.

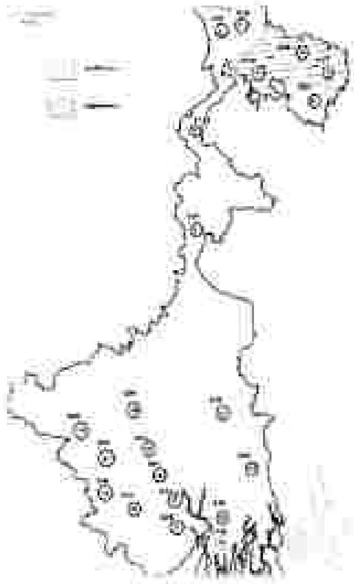

Fig.2a. First eigenvector - June. Variance explained = 33.9500%.

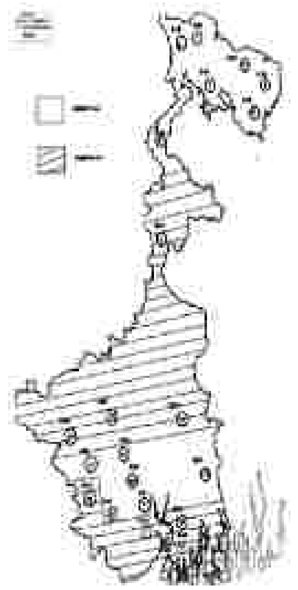

Fig.2 b Second eigenvector -June. Variance explained = 15.3509%.

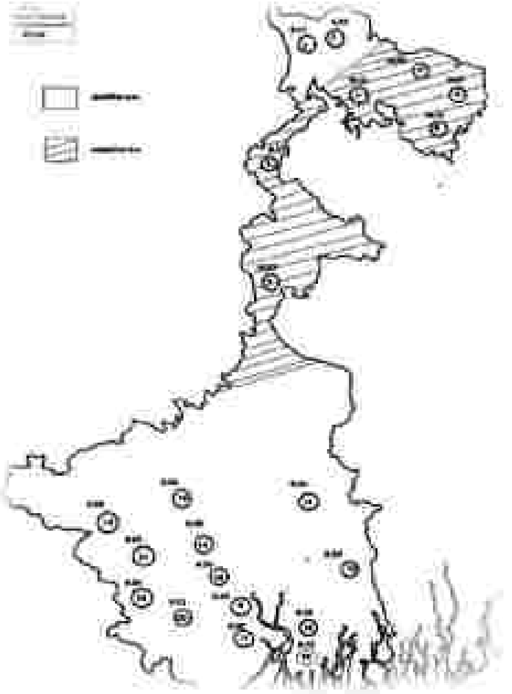

Fig.3a First eigenvector - July. Variance explained = 30.1485%

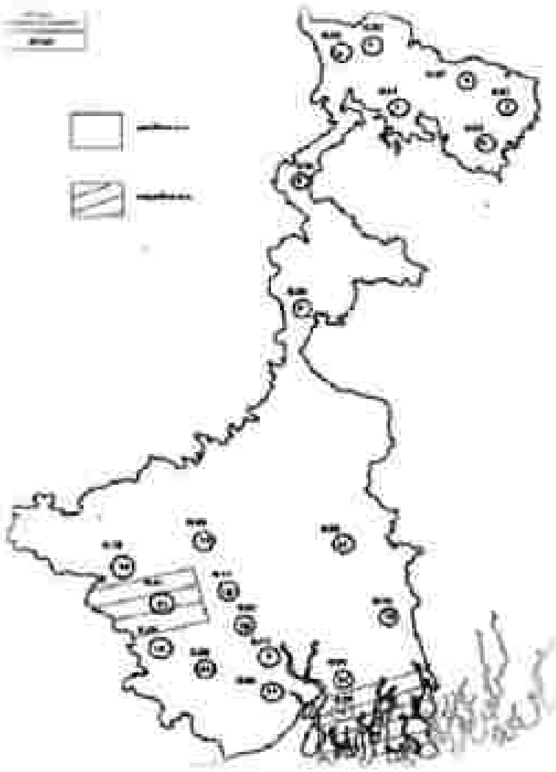

Fig.3b Second eigenvector - July. Variance explained = 18.5651%

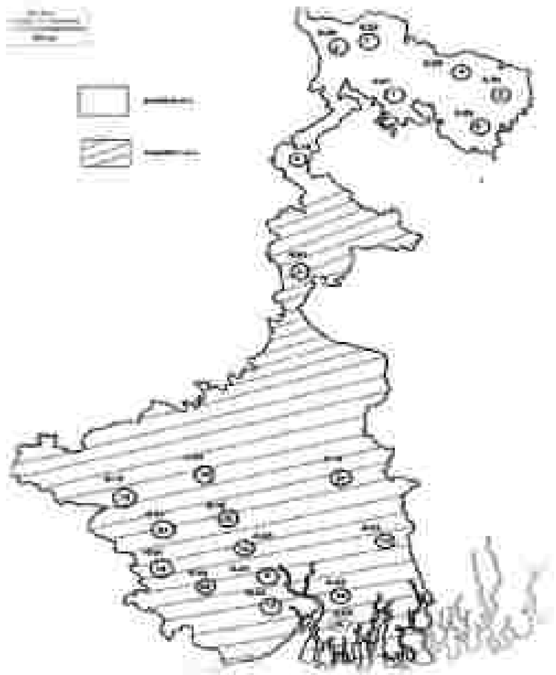

Fig.4a First eigenvector - August. Variance explained = 29.48402%.

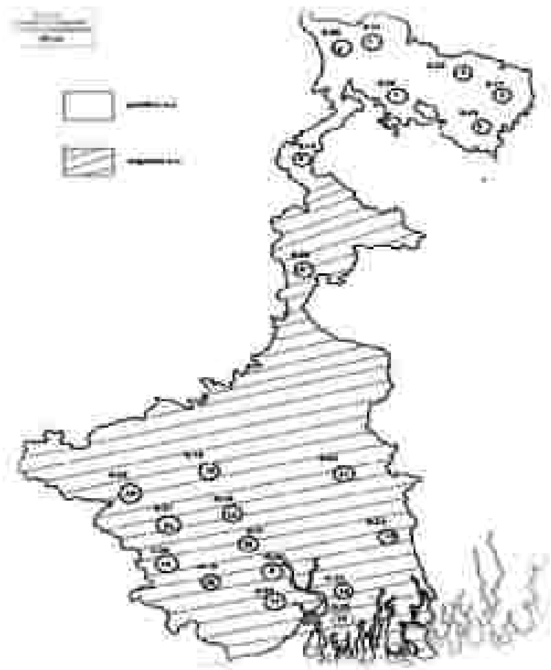

Fig.5a First eigenvector - September. Variance explained =27.3748%.

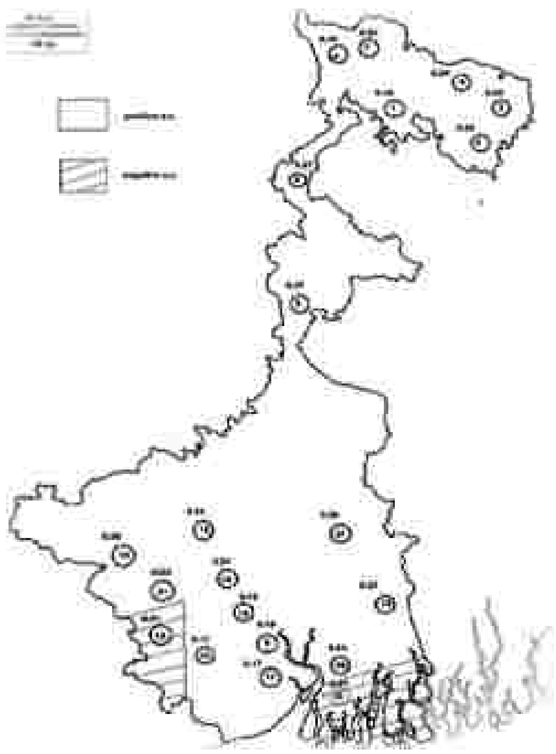

Fig.4b Second eigenvector -August. Variance explained = 17.4239%.c

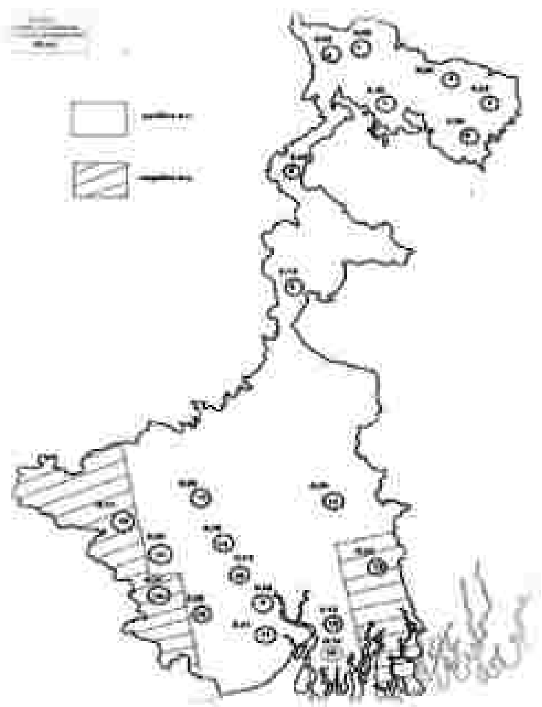

Fig.5b Second eigenvector -September. Variance explained =15.5534%.

An interpretation of the 2nd e.v. would proceed in the similar lines. As this accounts for about 15-18% of the variance, it is perhaps the local feature not related to atmospheric scales. The 2nd e.v. in June indicates a contrast of four regions namely, northern hills, mid-central part, south-central and extreme south respectively. In July, it indicates a whole state in the same state with contrast develops on the West part. This is intensified in August. In September, the contrast develops in west and east parts of the state. The 3rd and 4th e.v. pattern 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 PCs in descending order of importance. Each p_j (j=1, 2 ...) is a time series sampled annually and would lead to information on inter-annual variability. All the first 4 PCs of the 4 months have

TABLE 3

Frequency of sign sequences in PC1-PC4 of monthly PCA (N=60 years)

	Sign								
	++		+-		-+		--		
	Obs.	Expt.	Obs.	Expt.	Obs.	Expt.	Obs.	Expt.	Chi-sq. obs.
Month: JUNE									
PC1	10	9.7627	14	14.2373	14	14.2373	21	20.7627	0.0164
PC2	17	16.2881	14	14.7119	14	14.7119	14	13.2881	0.1381
PC3	15	13.2882	13	6.7241	13	14.7119	18	16.2881	0.7988
PC4	20	15.2543	10	14.7458	10	14.7458	19	14.2542	6.1112**
Month: JULY									
PC1	14	10.1694	11	14.8305	10	13.8305	24	20.1695	4.2206**
PC2	16	15.7627	15	15.2372	14	14.2373	14	13.7627	0.0153
PC3	15	13.2881	13	14.7119	13	14.7119	18	16.2881	0.7988
PC4	13	15.2542	17	14.7458	17	14.7458	12	14.2542	1.3788
Month: AUGUST									
PC1	13	11.0169	12	13.9831	13	14.9831	21	19.0169	1.1074
PC2	8	10.1695	16	13.8305	17	14.8305	18	20.1695	1.3539
PC3	12	14.7458	17	14.2542	18	15.2542	12	14.7458	2.0457
PC4	17	14.2542	12	14.7458	12	14.7459	18	15.2542	2.0457
Month: SEPTEMBER									
PC1	13	15.2542	17	14.7458	17	14.7458	12	14.2542	1.3788
PC2	12	10.1695	13	14.8305	12	13.8305	22	20.1694	0.9638
PC3	17	14.7458	13	15.2542	12	14.2542	17	14.7457	1.3788
PC4	19	15.2542	11	14.7457	11	14.7457	18	14.2542	3.8071**

****Values significant at 5% level.**

been studied to test the existence of auto-correlation for a maximum lag of 6 years. Only a marginally significant auto-correlation such as JunePC4 (lag2, lag4), JulPC2 (lag5), JulPC3 (lag2 and lag5), AugPC3 (lag6), SepPC3 & PC4 (lag1) at 5% level are observed. The auto-correlations though sometimes significant would be of little importance in forecasting PCs.

As a further test of annual association, the number of changes in the sign of the first 4 components, namely (++, +-, -+, --) has been collected in a two-way contingency table. These are tested against the expected number of occurrence if the changes were due to chance (Table 3). The PC series such as PC4 of June, PC1 of July and PC4 of September are found significant at 5% level. Hence, year-to-year association in changes in sign in the above monthly PC series can be accepted as exhibiting a pattern and cannot be dismissed as simply due to chance at 5% level.

5. Monthly Transition

It has been verified and mentioned that station rainfall does not show month-to-month correlation. This does not exclude the possibility of a correlation existing among principal component (PC) series. PC series are, in fact area rainfall series where weights of stations are assigned in an optimal way. However, the possibility of whether the PCs representing the size of West Bengal can bring out a feature is still open. If the monthly associations are present in the rainfall data, it is expected to reflect into the concerned PCs. Here, one particular indicator of this relation, namely, the transition in sign is examined. If rainfall in a given month is normal at all sampling stations, all the corresponding PCs would be essentially zero. As the first PC dominates the spatial variation, when it is zero, it is expected the rainfall also to be near its own normal value. Thus, the dependence, if any, in the signs would indicate patterns in the inter-month variation of rainfall. In Table 4(a), the observed number of sequences of ++, +-, -+, — are listed for the first PC. For each row in Table 4(a), the persistence or change in the sign can be shown on a 2 x 2 contingency table. The significance of the association is tested against the number expected, if the sign changes are purely by chance. For example, for June, the first PC is + ve, 16 + 15 = 31 times. The corresponding number for July is 16 + 16 = 32 Now, if the PC's of June and July are independent, the expected number of occurrences of the ++ sequence in 60 observations would be (31x32)/60=16.53. These frequencies are also listed in Table 4(a). The null hypothesis H_0 is "there is no dependence in the month-to-month sign changes". The chi-square ($x2$) test is applied to test this hypothesis (Rohatgi 1984). The observed $x2$ values listed in Table 4(a) are compared with the tabulated $x2$ value of 3.84, at one degree of freedom and at 95% significance. Whenever the observed value exceeds the tabulated value, the null hypothesis is rejected. However, it is observed that monthly transitions for the first PC do not exhibit a pattern and are purely due to chance.

However, it is observed that for first PC, it is surprisingly observed that the transition from July to August could be accepted as exhibiting a pattern at 5% level and cannot be simply dismissed as being random, whereas the other monthly transitions are purely random.

A similar analysis for the sign changes of the 2nd PC is also performed and is presented in Table 4(b). As indicated in the table, all the transitions are purely random at 5% level.

The 3rd and 4th PC series, though are of secondary importance explaining about 11% and 7% of variance, the persistence or change in sign are also tested and are presented in Tables 4(c) and 4(d) respectively. It is observed that for PC4, from June to July transitions exhibit a pattern of sign sequence at 10% level of confidence. All the other transitions may be accepted as purely random.

In Table 5, all frequencies observed and the corresponding expected due to chance are presented for the inter-month PC transitions, namely, June-July, July-August and August-September are presented. It is interesting to note that in case of June-July, PC4-PC3 and in case of August-September, PC1-PC2 is clearly identified as not due to chance at 10% level, whereas the other transitions can be accepted as purely random.

It is already noted that first four PCs may be considered for monthly analysis; in the month June-July, when SWM is in developing form and in August-September, when SWM is in fully matured form, the significant transition provides an indication of how the rainfall could be in the process of matured form and would be an interesting phenomenon.

6. South West Monsoon (SWM) variation

An analysis similar to monthlies has been carried out on the SWM rainfall over 21 stations for

TABLE 4 (A)

Frequency of sign sequences in PC1 of monthly rainfall (N=60 years)

Sign →	++		+-		-+		--		
Month ↓	**Obs.**	**Expt.**	**Obs.**	**Expt.**	**Obs.**	**Expt.**	**Obs.**	**Expt.**	**Chi-sq.obs**
Jun-Jul	16	16.5333	15	14.4667	16	15.4667	13	13.5333	0.0763
Jul-Aug	9	13.3333	23	18.6667	16	11.6667	12	16.3333	5.1735**
Aug-Sep	8	10.4167	17	14.5833	17	14.5833	18	20.4167	1.6477

*Values significant at 5% level.

TABLE 4 (B)

Frequency of sign sequence in PC2 of monthly rainfall (N=60 years)

Sign →	++		+-		-+		--		
Month ↓	**Obs.**	**Expt.**	**Obs.**	**Expt.**	**Obs.**	**Expt.**	**Obs.**	**Expt.**	**Chi-sq.obs**
Jun-Jul	9	10.3999	15	13.6000	17	15.6000	19	20.3999	0.5543
Jul-Aug	10	11.2667	16	14.7333	16	14.7333	18	19.2667	0.4434
Aug-Sep	13	13.0000	13	13.0000	17	17.0000	17	17.0000	0.0000

TABLE 4 (C)

Frequency of sign sequences in PC3 of monthly rainfall (N=60 years)

Sign →	++		+-		-+		--		
Month ↓	**Obs.**	**Expt.**	**Obs.**	**Expt.**	**Obs.**	**Expt.**	**Obs.**	**Expt.**	**Chi-sq.obs**
Jun-Jul	15	13.5333	14	15.4667	13	14.4667	18	16.5333	0.5768
Jul-Aug	15	14.0000	13	14.0000	15	16.0000	17	16.0000	0.2679
Aug-Sep	15	15.0000	15	15.0000	15	15.0000	15	15.0000	0.0000

TABLE 4 (D)

Frequency of sign sequences in PC4 of monthly rainfall (N=60 years)

Sign →	++		+-		-+		--		
Month ↓	**Obs.**	**Expt.**	**Obs.**	**Expt.**	**Obs.**	**Expt.**	**Obs.**	**Expt.**	**Chi-sq.obs**
Jun-Jul	19	15.5000	12	15.5000	11	14.5000	18	14.5000	3.2703*
Jul-Aug	12	15.0000	18	15.0000	18	15.0000	12	15.0000	2.4000
Aug-Sep	15	15.5000	15	14.5000	16	15.5000	14	14.5000	0.0667

*Values significant at 10% level.

the period 1900-1960. The first five components explain about 70% of total variance as inspected in the analysis (Basak, 2014). The spatial organization of first two eigenvectors (e.v.) for SWM is presented in Figures 6 and 7 respectively. Inspection of the Fig. 6 for the first e.v. indicates a North-South contrast having negative loadings beneath the Malda station, namely Gangetic West Bengal (GWB) and essentially positive loading north of it, namely Sub-Himalayan West Bengal (SHWB). As an interpretation, it may be thought of above/below normal rainfall in the Southern stations throughout

TABLE 5

Frequency of sign sequences for inter-month transition of Principal Components
(N=60 years)

	Sign								
	++		+-		-+		--		
	Obs.	Expt.	Obs.	Expt.	Obs.	Expt.	Obs.	Expt.	Chi-sq. obs.
JUNE-JULY									
PC1-PC2	12	12.8000	12	11.1999	20	19.2000	16	16.7999	0.1786
PC2-PC1	15	13.4333	16	17.5667	11	12.5667	18	16.4333	0.6671
PC2-PC3	16	14.4667	15	16.5333	12	13.5333	17	15.4667	0.6305
PC3-PC2	13	15.4667	16	13.5333	19	16.5333	12	14.4667	1.6315
PC3-PC4	13	14.5000	16	14.5000	17	15.5000	14	15.5000	0.6007
PC4-PC3	11	14.4667	20	16.5333	17	13.5333	12	15.4667	3.2226*
JULY-AUG									
PC1-PC2	12	10.8333	14	15.1667	13	14.1667	21	19.8333	0.3801
PC2-PC1	14	13.8667	18	18.1333	12	12.1333	16	15.8667	0.0048
PC2-PC3	14	16.0000	18	16.0000	16	14.0000	12	14.0000	1.0714
PC3-PC2	10	11.6667	18	16.3333	15	13.3333	17	18.6666	0.7653
PC3-PC4	16	14.0000	12	14.0000	14	16.0000	18	16.0000	1.0714
PC4-PC3	16	15.0000	14	15.0000	14	15.0000	16	15.0000	0.2867
AUG-SEP									
PC1-PC2	8	10.8333	18	15.6667	17	14.6667	17	19.8333	2.2418*
PC2-PC1	12	12.5000	13	12.5000	18	17.5000	17	17.5000	0.0686
PC2-PC3	13	12.5000	12	12.5000	17	17.5000	18	17.5000	0.0686
PC3-PC2	15	12.5000	15	17.5000	10	12.5000	20	17.5000	1.7143
PC3-PC4	17	15.5000	13	14.5000	14	15.5000	16	14.5000	0.6007
PC4-PC3	15	15.0000	15	15.0000	15	15.0000	15	15.0000	0.0000

*Values significant at 10% level.

PC2-PC3	13	12.5000	12	12.5000	17	17.5000	18	17.5000	0.0686
PC3-PC2	15	12.5000	15	17.5000	10	12.5000	20	17.5000	1.7143
PC3-PC4	17	15.5000	13	14.5000	14	15.5000	16	14.5000	0.6007
PC4-PC3	15	15.0000	15	15.0000	15	15.0000	15	15.0000	0.0000

***Values significant at 10% level.**

the SWM season which has largest weight would indicate a similar trend in Northern stations and a below/above normal in the North stations.

For the 2^{nd} e.v., a dominant positive and negative loading is observed in the southern region (below Malda station) and also in Northern region of the state (Fig.7). Clearly, it indicates straight-forward two regions among Northern stations (with positive and negative in loadings) indicating variation of SWM among the Northern stations. In together, the e.v.s shows a highly correlated field in case of SWM.

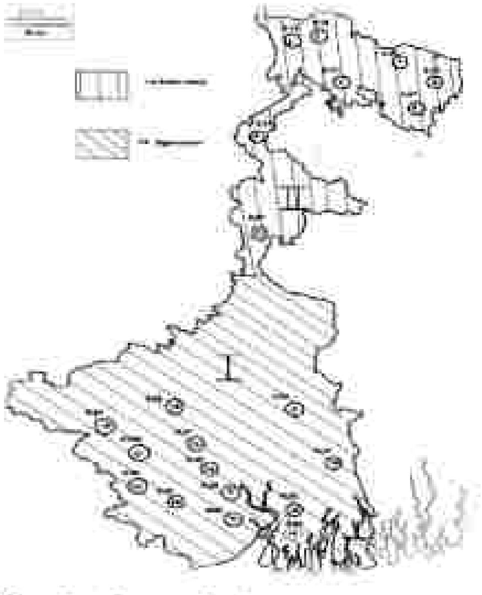

Fig.6 Station Network wit

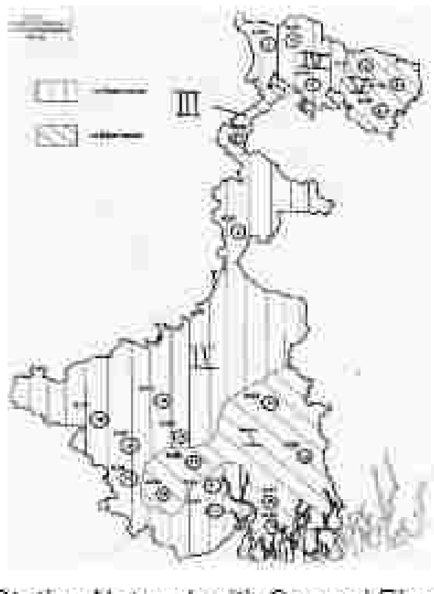

Fig.7 Station Network with Second Eigen-vector (SWM.)

7. Inter-annual variability

The SWM PCs are, however, important because they indicate the presence of annual signal. A similar analysis for the sign sequence changes as in Section 5 is performed for the SWM principal components (PC) series p_{ji} (j=1, 2, .., 5). The 3^{rd} PC series shows predominantly significant transition in changes in sign at 1% level of significance; also first PC series possess the significant transition at 5% level. All the other PC series are clearly identified as purely random. Thus, the first and 3^{rd} component of PCA of the SWM rainfall contributing 23.90% and 10.53% respectively of total variance represents a pattern with characteristic term as a year or a multiple of it. As an example, the time series of the first PC shows a predominant period of nearly 2 to 7 years meaning that the same sign persist for 2 to 7 years before a change in sign takes place (Fig. 8).

8. Grouping the year

When rainfall over a large area is considered, it is desirable to arrive at an area rainfall value as a weighted average of the rainfall at the individual stations. It may be mentioned that first PC is a dominant weighted average of the station rainfall and is a good measure of area rainfall. Further, since the second component is predominantly second in order, PC1 and PC2 on any time-scale are the two most important characteristics of rainfall in a particular year for the whole network of stations. Thus, with PC1 and PC2 as coordinates the yearly data may be represented on a diagram. Such a representation as in Fig. 9 produces a meaningful way of comparing the years for the SWM rainfall. When each station receives exactly its own normal rainfall, all principal components is zero. Such a year coincides with the origin in Fig. 9. The nearly normal years fall around the origin. Years with excessive rainfall that is flood years such as 1917, 1922 etc. have large positive PC1 and PC2 values are placed far away from origin in the first quadrant. Also, years with deficit rainfall years (draughts), namely 1918, 1920, 1941 etc. possess negative PC1, PC2 values and are placed on the fourth quadrant. Nearness of two or more years on this diagram indicates almost similar atmospheric conditions. Such information may help in manipulating the behavior of rainfall.

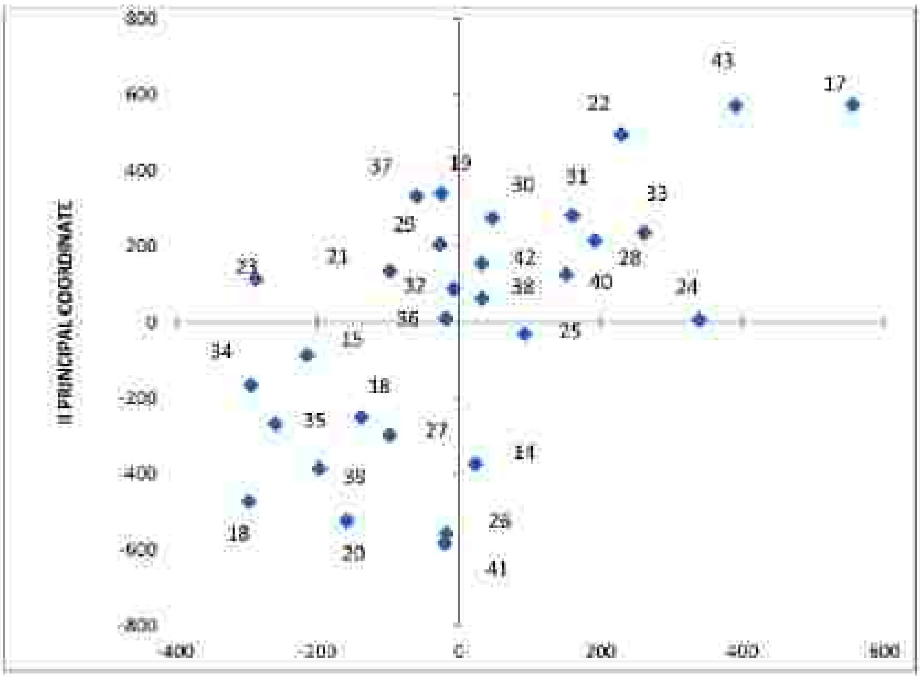

Fig.8 First PC time series of WB rainfall in West Bengal.

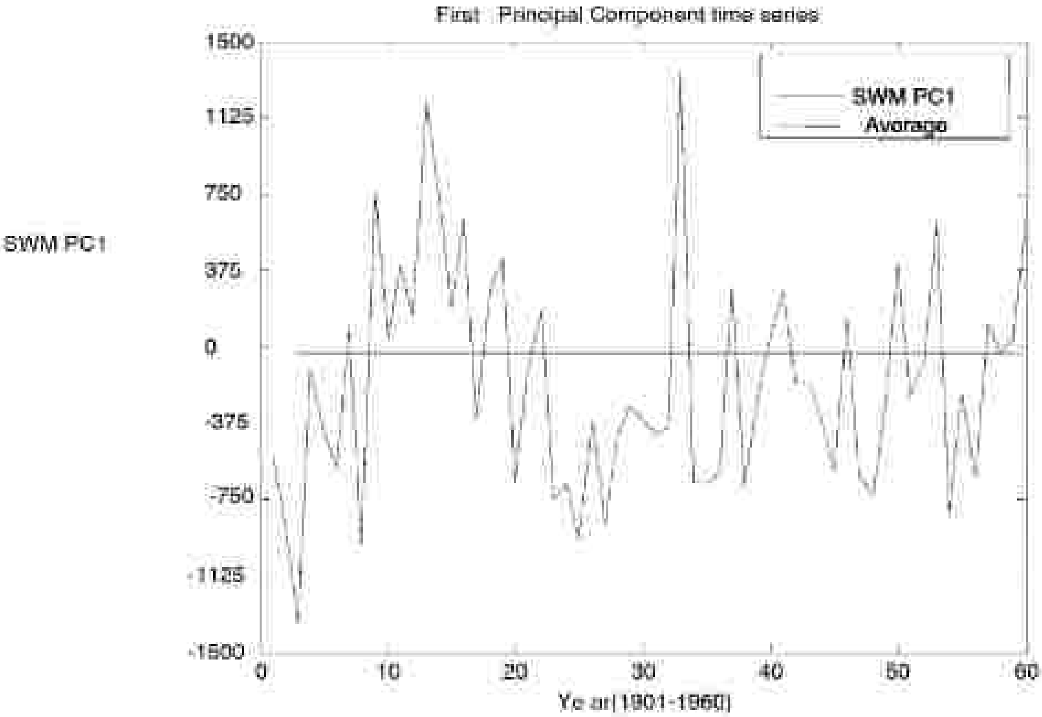

Fig.9.Variability of the principal components of the SWM rainfall in West Bengal (1901-1960)

9. Predictability of SWM rainfall

The question next to variability is perhaps predictability. If the variability, which is a deviation of rainfall about its long-term average value, is not purely random, it is expected to possess a temporal relationship to be detectable; the most probable relationship is a linear one. But, in the present context it has been pointed out that monthly rainfall anomalies exhibit no or marginally significant autocorrelations. Thus, linear relationships for time-wise evolution usually do not hold with suitable statistical test. Alternatively, as non-linear relation is complicated, the kind of statistical test to detect the relationship is cumbersome and difficult to interpret. Moreover, the kind of statistical methodology to detect nonlinear relations is not obvious. Principal component analysis (PCA) may be utilized in this connection. As PCs are found to possess statistically significant trends in some cases, it may be appropriate to first predict the PCs and then estimate the rainfall in terms of past data. First and 3rd PCs of the SWM rainfall show significant annual transitions and it is plausible to ask the probability of the next year PC being above/ below average (+ or —), if in the present year it is above/below average (+ or —). The two-state transition probability matrix for first and third PC is found to be:

$$[P]^1 = \begin{bmatrix} 0.6253 & 0.38460 \\ 0.25 & 0.75 \end{bmatrix}$$

$$[P]^3 = \begin{bmatrix} 0.25 & 0.75 \\ 0.6470 & 0.3529 \end{bmatrix}$$

Now, it is easy to see that the first PC stands for annual persistence mode and third PC stands for annual oscillatory mod. For West Bengal as a whole with the present data, the oscillation in the 2nd PC is attributable to chance and thus prediction through a transition may not be justified.

TABLE 6

Correlation of station with the first PC for South West Monsoon (SWM)

Sl.No.	Correlation	Stn. Average (July) (mm)
1	0.5872	2668.5017
2	0.6843	2371.2549
3	0.7555	2311.1583
4	0.6324	3030.6015
5	0.5581	1100.1399
6	0.5014	1748.8213
7	0.5191	2546.4435
8	0.6571	2369.8715
9	0.1062	1234.6466
10	0.6290	1025.2199
11	0.0002	1273.3282
12	0.4647	1145.7432
13	0.2338	1089.3658
14	-0.4510	1129.5833
15	-0.5346	1335.6848
16	-0.2317	1251.1499
17	-0.5098	914.4081
18	0.1657	1197.8098
19	-0.4117	1073.5714
20	-0.3884	1102.0283
21	-0.2677	1054.1980

The first PC in the SWM data is the major PC component contributing to about 25% of variance through PCA. It may be accepted as area rainfall as the weights area optimally assigned. The transition probability of first PC, that is, as obtained from the annual association of signs. Also, it is noticed that the most of the stations are fairly highly correlated (except a few) with the first PC of SWM rainfall (Table 6). The strong correlation between first PC and SWM rainfall leads to the inference that when significant can be taken as the considerable part of SWM rainfall. Then, for example, for SWM rainfall, the above average rainfall will be followed by an above average rainfall with 63% probability. Thus, a kind of prediction exercise is fruitful for those station SWM rainfalls which are fairly highly correlation with SWM.

A prediction exercise is undertaken for the SWM rainfall of the 21 stations wherein the SWM rainfall of the years 1961-1965 which are not included in the PCA has been considered. The number of counts of above/below SWM rainfall (i.e. ++, +-, -+, --) are counted and percentage of success is evaluated with respect to transition matrix. It is observed that the percentage of ++, +-, -+, -- transitions closely match with the transition matrix, for example, for ++, 0.55 from 1961-65 data against 0.63 from analysis and for --, 0.63 from 1961-65 data against 0.75 from analysis (transition matrix).

10. Predictability of monthly rainfall

In earlier section, it has been pointed out that regarding SWM, the significant first PC has brought forward some existence of predictability for SWM rainfall.

However, it has been observed that in Table 4(a) that transition of first PC from July to August is not random and exhibit a pattern. The concerned transition probability for the first PC is

$$[P]^1JA = \begin{bmatrix} 0.28 & 0.72 \\ 0.57 & 0.43 \end{bmatrix}$$

TABLE 7

SWM rainfall of stations above/below average with percentage of transition

+ Above Average; - Below Average No. of signs

Station Name	Aver. SWM (mm)	1981	1982	1983	1984	1985	++	+-	-+	--
Jalpaiguri	2592.75	-	-	+	+	-	1	1	1	1
Darjeeling	2311.16	+	-	+	+	-	1	2	1	0
Kalchini	3030.60	+	+	-	+	+	2	1	1	0
Malda	1100.14	-	-	-	+	+	1	0	1	2
Mongpo	2546.44	+	-	-	+	+	1	1	1	1
Mathabhanga	2369.87	+	+	-	-	+	1	1	1	1
Amta	1234.65	-	-	+	+	-	1	1	1	1
Arambag	1025.22	+	-	-	-	-	0	1	0	3
Budge Budge	1273.33	-	-	-	-	-	0	0	0	4
Burdwan	1089.37	-	-	-	-	-	0	0	0	4
Ghatshila	1129.56	-	-	-	+	-	0	1	1	2
Sagar Island	1335.68	+	-	+	-	-	0	2	1	1
Kukrahati	1251.15	-	-	+	-	-	0	1	1	2
Ranaghat	914.41	-	+	-	+	+	1	1	2	0
Kharagpur	1102.03	+	+	-	+	-	1	2	1	0
Silda	1054.20	+	-	-	-	-	0	1	0	3
					Percent of transition		55	45	37	63

This represents an oscillatory mode. The above average rainfall in July is expected to be following by a below average with 72% probability. This skewness of the transition is very much interesting feature that comes out systematically in present analysis. It has been verified that in case of monthly analysis, majority of the stations are highly correlated with the first PC of July and August (Tables 8a and 8b respectively).

This implies that a kind of predictability for first PC for July and August would be valid for station rainfall. In Table 9, for all the stations, the transitions of July to August are presented. It has been found that transition from July to August when the SWM is in matured form, matches fairly well with the transition probability. It is observed that the percentage of ++, +-, -+, -- transitions closely match with the transition matrix, for example, for --', 0.53 from 1961-65 data against 0.47 from analysis (transition matrix).

11. Discussion

The approach in common in understanding the time series studies of both monthly and SWM rainfall is that of autocorrelation and power spectrum analysis. However, the main difficulty arises in the analysis is the fact that the series are mostly purely random. Moreover, the rainfall stations being widely spread and being correlated among themselves, straight forward time series analysis is complicated and cumbersome. The analysis of individual the time series of station would also neglect the spatial structure that are inadvertently present in a large area like state. To overcome the ensuing difficulty, one needs non-linear techniques such as bi-spectrum analysis (Hartmann and Michelsen, 1989). This definitely asks for a demarcation of area of station rainfall as performed by Iyenger and Basak (1994) for All India and Iyenger (1991) for Karnataka. PCA provides some sort of solution for the difficulty. A large number of stations spreading

TABLE 8 A

Correlation of stations with the first PC for July

Sl.No.	Correlation	Stn. Average (July) (mm)
1	0.5872	2668.5017
2	0.6843	2371.2549
1	-0.4937	787.2166
2	-0.5898	886.2615
3	0.1542	761.1501
4	-0.3760	914.4133
5	-0.1473	295.5099
6	-0.5573	540.7315
7	-0.0886	819.5567
8	-0.5230	689.4133
9	0.7586	342.4866
10	0.7401	316.4349
11	0.7487	362.0000
12	0.4313	318.8933
13	0.5951	330.3250
14	0.5777	340.1783
15	0.3786	382.4116
16	0.6083	364.2917
17	0.5458	240.5766
18	0.8158	352.9349
19	0.6448	312.7799
20	0.2718	314.1499
21	0.5414	294.2767

over long area is handled simultaneously as well as the number of components for studying variability becomes comparatively less than the total number of stations.

PCA can be considered to be a generalized Fourier decomposition of a random field. In this technique, a large number of station data may be handled simultaneously to account for spatial variability but undoubtedly the number of components to be studied will be less than the total number of stations. In the present study, PCA technique is used to understand SWM rainfall variability. The station data which are neither uncorrelated nor perfectly correlated gets transformed into PCA and extract the temporal variable characteristic for complete network of stations.

The advantage of this is apparent when we observe that for West Bengal rainfall for SWM period, the first e.v. explains less than 50% of spatial variance but the first PC and area rainfall are highly correlated (r= 0.8 as observed) and the first e.v. demarcates two separate zones, namely north and south zone (Fig.6). Similarly, the 2nd and other significant PCs are connected to the area rainfall in regions where in the corresponding e.v. has the same sign. The temporal signals that may be present over large spatial regions would be carried over into the first few PC time series after automatically eliminating noises retaining in other PCs, called spatial noise.

Regarding SWM rainfall predictability, whenever the PCs show any kind of relationship in auto-correlation or Power Spectral that are significant can be predicted as it is commonly done in time series analysis. However, a prediction exercise with the help of transition probability (above/below normal) of PC1 results in a fairly good prediction of SWM rainfall (Table 7).

Moreover, for the predictability of monthly rainfall, a prediction exercise based on PC1 July-August transition resulted in reasonable prediction of July-August monthly rainfall (Table 9).

TABLE 8 B

Correlation of stations with the first PC for August

Sl.No.	Correlation	Stn. Average (July) (mm)
1	0.7653	655.9783
2	0.7326	653.3233
3	0.5475	598.1383
4	0.8309	739.8449
5	0.0174	281.4116
6	0.4937	444.0948
7	0.6216	634.7050
8	0.7221	508.8983
9	-0.5420	349.6900
10	-0.5951	299.6182
11	-0.5522	352.5399
12	-0.4722	323.3200
13	-0.3715	303.6833
14	-0.4951	337.1666
15	-0.5738	380.4882
16	-0.5201	340.3983
17	-0.2572	254.0050
18	-0.6300	343.9783
19	-0.4024	330.6932
20	-0.1304	164.8908
21	-0.0346	316.2817

TABLE 9

July/August rainfalls of Stations above/below average with percentage of transition

+ Above Average; - Below Average J=July: A=August

Station Name	Aver. SWM J/A (mm)	1981 J/A	1982 J/A	1983 J/A	1984 J/A	1985 J/A	++	+-	-+	--
Jalpaiguri	787.12/ 655.98	++	--	++	+-	++	2	1	1	1
Darjeeling	761.15/ 598.14	++	-+	-+	++	-+	2	0	3	0
Kalchini	914.41/ 739.84	-+	++	++	+-	++	3	1	1	0
Malda	295.51/ 281.41	-+	+-	+-	+-	+-	0	3	1	1
Kishanganj	540.73/ 444.09	--	-+				0	0	1	1
Mongpo	819.56/ 634.71	++	-+	++	++	-+	3	0	2	0
Mathabhanga	689.41/ 508.90	++	-+				1	0	1	0
Amta	342.49/ 349.69		--	+-	+-	-+	0	2	0	2
Arambag	316.43/ 299.62	--	--	--	+-	--	0	1	2	2
Budge Budge	362.00/ 352.54	-+	--	--	+-	-+	0	1	2	2
Burdwan	330.32/ 303.68	--	--	--	++	+-	1	1	0	3
Ghatshila	340.18/ 337.17	--	--	+-			0	1	0	2
Sagar Island	382.41/ 380.49	-+	--	-+	+-	-+	0	1	3	1
Kukrahati	364.29/ 340.40	-+	--	--	+-	-+	0	1	1	3
Ranaghat	240.58/ 254.00	--	-+	--	-+	-+	0	0	3	2
Vishnupur	312.78/ 330.69	--	--				0	0	0	2
Kharagpur	314.15/ 164.89	-+	-+	-+	++	-+	1	0	4	0
Silda	294.28/ 316.28	--	-+	+-	+-	--	0	2	1	2
					Percent of transition		48	52	47	53

12. Summary and Conclusions

PCA are sometimes used in meteorological data analysis, producing a decomposition of the data field into spatial eigenvectors (e.v.s) and a temporal time series. Whilst e.v. pattern is used in meteorological field, the usefulness of the PC time series has received limited attention in the state-wise analysis, especially in West Bengal for rainfall variability. The present investigation is motivated by the possibility that few PCs may contain valuable information regarding the maturity, genesis and variability of rainfall during SWM period. The monthly rainfall data of West Bengal spread over 21 stations for a period of 60 years show that PCA is a valuable tool in grooving insight into temporal patterns through transition probabilities of the first and 3rd PCs. For the state, the rainfall variations in June, July, August and September are related in sequence. Transitions of fluctuations except from July to August are due to chance. For the state as a whole for the SWM period, the first and 3rd PC exhibit significant inter-annual transition whereas the 2nd PC shows no significant trend.

A prediction exercise for predicting the July-August in the 5 years in 1961-65 through an estimated transition probability has been surprisingly successful. However, further detailed analysis is required to quantify predictability of the PCs as forecast able signals of impending rainfall variations.

Acknowledgements

The author sincerely thanks National Data Centre ADGM (R) Pune for providing SWM rainfall data of 21 stations of West Bengal.

References

Basak P., 2014, "Variability of south west monsoon in West Bengal: An application of principle component analysis", Mausam, 65, 4, 559-568.

Basu G. C., 2001, "A feature of monsoon rainfall and its variability with comparative study in meteorological sub-divisions of West Bengal", Mausam, 52, 736-746.

Basu G. C., Bhattacharjee U. and Ghosh R., 2004, "Statistical analysis of rainfall distribution and trend of rainfall anomalies district wise during monsoon period over West Bengal", Mausam, 55, 409-418.

Bedi H. S. and Bindra M.M.S., 1980,"Principle components of monsoon rainfall", Tellus, 32, 296-298.

Chanda B.N. and Dhar O.N. 1975,"A study of incidence of droughts in the Gangetic West Bengal", www.new.dli.ernet.in/rawdataupload/upload/insa/.../20005bae_22.pd.

Gnanadesikan R., 1977, "Methods for statistical data analysis of multivariate observation", Wiley, New York, p. 311.

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 Indian monsoon rainfall anomaly", Tellus, A35, 324-331.

Hays W. H. and Winkler R. L., 1970, "Statistics: Probability, Inference and Decision", Rinehart and Winston, inc., New York, p. 937.

Iyenger R. N., 1991, "Application of principal component analysis to understand variability of rainfall", Proc. Ind. Acad. Sc. (Earth Planet Sc.), 100, 2, 105-126.

Iyenger R. N. and Basak P., 1994, "Regionalization of Indian monsoon rainfall and long term variability signals", Int. J. Climatol., 14, 1095-1114.

Mohapatra N., Biswas H. R. and Sawaisarje G. K., 2011, "Spatial variability of daily rainfall over Northeast India during summer monsoon season", Mausam, 62, 2, 215-228.

Rohatgi V. K., 1984, "Statistical Inference" John Wiley, New York.

World Meteorological Organization, 1966b, "Some methods in climatologically analysis", *WMO Technical Note No.81, WMO No.* 199, p. 53.