

¹ Indian Institute of Science, Center for Atmospheric and Oceanic Sciences, Bangalore, India

² Indian Institute of Science, Department of Civil Engineering, Bangalore, India

Intrinsic mode functions and a strategy for forecasting Indian monsoon rainfall

R. N. Iyengar^{1,2} and S. T. G. Raghu Kanth²

Summary

Indian monsoon rainfall data is shown to be decomposable into six empirical time series, called intrinsic mode functions. This helps one to identify the first empirical mode as a nonlinear part and the remaining as the linear part of the data. The nonlinear part is handled by artificial neural network (ANN) techniques, whereas the linear part is amenable for modeling through simple regression concepts. It is found that the proposed model explains between 75 to 80% of the interannual variability (IAV) of eight regional rainfall series considered here. The model is efficient in statistical forecasting of rainfall as verified on an independent subset of the data series. It is demonstrated that the model is capable of foreshadowing the drought of 2002, with the help of only antecedent data. The statistical forecast of All India rainfall for the year of 2004 is 80.34 cms with a standard deviation of 3.3 cms. This expected value is 94.25% of the longterm climatic average.

1. Introduction

The summer monsoon also called the southwest monsoon (SWM) delivers the major component of annual rainfall in India. The sum of quantified rainfall during June, July, August and September is the seasonal monsoon rainfall on a yearly basis. This time series on different spatial regimes is among the most widely investigated meteorological data in India. Efforts have been

made in the past to understand the connections between SWM and other global phenomena such as El Niño (EN), southern oscillation (SO) and sunspot cycle. In addition, there have been studies investigating the rainfall data series to characterize temporal variability on monthly, seasonal and annual scales. Even though the monsoon season in terms of wind direction, cloud formation and other physical features has recurred annually since ancient times, the amount of rainfall it brings varies considerably year to year. This raises the question whether the quantum of rainfall for the season can be forecast keeping in view the regularity with which the monsoon season appears. In the past, this issue has been addressed in two different ways. In the first approach, rainfall is thought to be the effect of other antecedent meteorological parameters. Walker (1923) was perhaps the first to study teleconnections between Indian monsoon and meteorological parameters at other parts of the earth. The models of Gowariker et al (1989), Thapliyal (1990), and Sahai et al (2003) belong to this category. Among these, the model of Sahai et al (2003), which links global SST data with Indian monsoon seasonal data appears to be the most successful one. In the second approach, rainfall time series is supposed to carry the imprint of all causes in itself. In a mechanical

system with memory, the response can be notionally represented as a convolution between external forces and the system impulse response or Green's function. Hence, even when the causes are not known, with a sufficiently long data series, the rainfall can be modeled as a generalized function of the known past. A model capable of successfully replicating the past can be used to forecast the rainfall a year ahead with a known error band. The studies of Sahai et al (2000), Iyengar and Raghukanth (2003) are examples of this approach. Apart from these, there are several studies (Hastenrath and Grieschar, 1993) on the variability of monsoon on annual, seasonal and other time scales. Some of these elucidate how particular dominant periods observed in the Fourier spectrum of the data may arise due to quasi-biennial oscillation (Rao and Lakhole, 1978), tidal forcing (Campbell et al, 1983), ENSO (Shukla and Paolino, 1983), sunspot cycles (Bhalme and Jadhav, 1984) and intraseasonal periodicities (Cadet and Daniel, 1988; Hartmann and Michelson, 1989). Narasimha and Kailas (2001) have used wavelets to understand temporal variability of monsoon rainfall. Their analysis reveals that the annually sampled seasonal data is characterized by near periodic oscillations of 3, 5.8, 11.6, 20.8, 37 and 80 year, periods. The existence of oscillatory trends very near the above periods can be surmised from the power spectral density function of the basic data series. However, translation of this knowledge to effectively forecast or extrapolate data by one year has not been possible. The present paper studies forecasting of Indian rainfall data with the above points in the background. A new representation of the data series, in terms of a finite number of empirical time series is presented. These time series are simpler than the original data for modeling and forecasting.

2. Empirical modes

Huang et al (1998) have presented a method for decomposing time series data into a finite number of empirical modes called Intrinsic Mode Functions (IMF), which can be shown to be uncorrelated among them, except for round off errors. An IMF is a data derived function such that, in its interval of definition the number of zeros and extrema are either equal or differ at most by one. Further, at any point, the mean

value of the local positive and negative envelopes of the IMF would be zero. Each IMF so obtained is a narrowband time series with an identifiable central period around which the oscillations take place. The amplitude and period of an IMF will be slowly varying. This simple form also provides a physical basis, for relating monsoon rainfall with meteorological parameters that show similar periods as a particular IMF. The present paper studies eight sets of Indian rainfall time series to decompose the observed data into their basic IMF's. It is demonstrated that in all the cases the data exhibit six modes of temporal variation. The last mode always represents the climatic average, which remains almost constant. The traditional method of investigating rainfall data has depended on models of stationary random processes with Gaussian properties. Application of the corresponding statistical tests to verify the auto-correlation or power spectral density functions of SWM data leads to the result that signals, if any, in these data are very weak. This is the main reason, why SWM data cannot be modeled in terms of a linear time series. On the other hand, the particular form of the nonlinear model to be used is not so obvious. Previously (Iyengar and Raghukanth, 2003) it has been demonstrated that a nonlinear model with variable frequency harmonic terms can be effectively used to explain about 50% of the IAV. This would indicate that the basic data, if it is not a white noise, should carry the decomposed IMF's as signals. For all the eight cases studied here, following the work of Wu and Huang (2003), it is found that the white noise test fails meaning that the IMF's are signals. The first IMF is a high-frequency mode that is also strongly non-Gaussian. It needs a nonlinear framework for modeling and forecasting. The remaining IMF's are progressively slower and tend to become less random. This property can be used to extend them by one year using linear regression techniques. This amounts to a new strategy for long range forecasting of Indian monsoon rainfall.

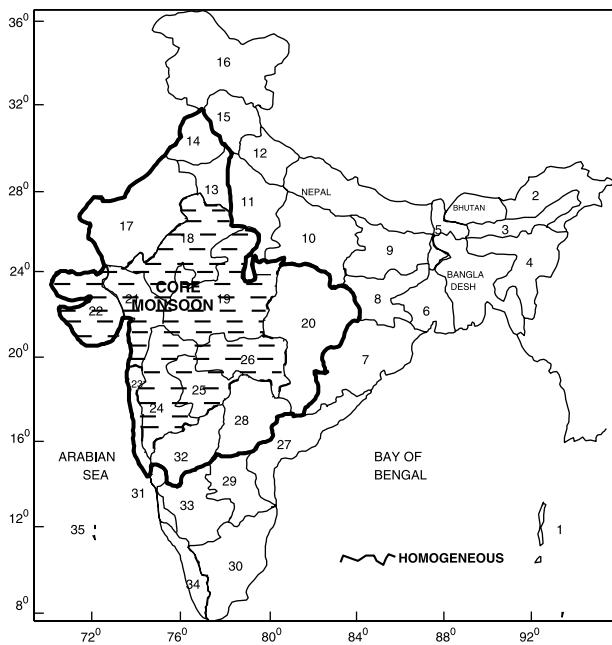
3. Rainfall data

Large numbers of raingauge stations are operative in India to continuously monitor the daily, weekly, monthly and seasonal rainfall all through the year. From the station data, area rainfall

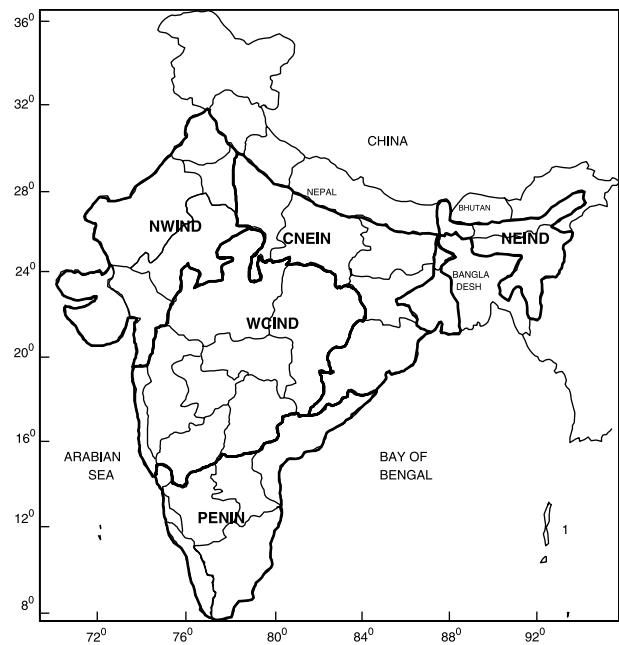
Table 1. SWM rainfall data (1871–1990)

Region	Area (Sq. km)	m_R (cm)	σ_R (cm)	Skewness	Kurtosis
All India	2,880,000	85.2424	8.4686	-0.5670	2.9574
Homogeneous	1,596,970	75.7212	11.8958	-0.4387	3.0699
Core monsoon	776,942	85.8222	14.7662	-0.4980	3.0586
WCIND	962,698	93.3198	12.5899	-0.4383	3.0351
CNEIN	573,006	100.2477	11.1060	-0.2082	4.4050
NEIND	267,444	141.9222	12.1318	0.1422	3.0917
NWIND	634,272	49.0082	13.2373	-0.2861	2.8866
Peninsular	442,632	65.9455	9.8273	0.3105	3.1164

WCIND: Western-Central India; CNEIN: Central-Northeast India; NEIND: Northeast India; NWIND: Northwest India



a



b

Fig. 1a. Meteorological subdivisions of India; **b** Five non-overlapping regions contributing to AIRF

values are computed for various regions and made available at the website www.tropmet.res.in of Indian Institute of Tropical Meteorology (IITM), Pune. Here, eight sets of such regional SWM rainfall data, which is the sum of the monthly values of June, July, August and September, are chosen for detailed study. The different regions covered by the database, some of which are overlapping, are listed in Table 1 and shown in Fig. 1a, b. The different data series may be correlated but taken individually the regions are coherent within themselves. It has been customary to compute an All India Rainfall (AIRF) value as a weighted sum of rainfall in non-

overlapping regions. This is invariably cited as an index of the monsoon performance at the national level. Thus, AIRF series can be viewed as an indicator of the agricultural economy of the country. Some basic statistics of the data such as the climatic normal (m_R) and climatic deviation about the normal (σ_R) are presented in Table 1. It may be noted that monsoon rainfall is negatively correlated among some of the contributing regions, as a consequence of which, AIRF series shows lesser variability in terms of (σ_R/m_R) than the other data series. Here all the eight time series are retained since they represent different spatial regimes of the vast country.

4. Intrinsic mode functions

The approach of IMF decomposition to investigate monsoon rainfall is new. With this in view, the method of extraction of IMF's is briefly described below with reference to the first data series of Table 1. Following Huang et al (1998) the consecutive peaks and consecutive valleys of the data series are joined by cubic splines (Fig. 2). At every time step the average of the positive (E^+) and negative (E^-) envelopes are found. This is shown in Fig. 2 with a bold line. This average $m_0(t)$ which is the bias of the data about the zero level, is subtracted from the raw data to get $R_3(t) = R(t) - m(t)$. This new time series is further processed as in the previous step to get $R_2(t) = R_1(t) - m_1(t)$. This process is repeated m times till the sieved data $R_m(t)$ is centered symmetrically such that with every zero only one peak or valley occurs. Such an $R_m(t)$ is the first intrinsic mode denoted as IMF_1 . In Fig. 2,

IMF_1 of the AIRF series is extracted after six iterations. To extract the second IMF, the first IMF is subtracted from the original data and the process is repeated. On similar lines IMF_3, IMF_4, \dots are hierarchically extracted until the sieved data shows no oscillations or is an IMF by itself. Thus, long-term climate trends, center-line drifts, and long period nonstationary features come out as the last IMF. For the time series of Fig. 2, six IMF's can be extracted. Interestingly, all the eight regional rainfall data of Table 1 are decomposed into six IMF's. In Fig. 3a-h, the IMF's of the eight data are arranged in the order they are extracted. It is observed that the last IMF is invariably positive and is a mode slowly varying around the long-term average. This may be thought of as the normal or climatic component about which the IAV of the monsoon rainfall appears. This IAV itself can be decomposed into five dynamic modes each evolving around a

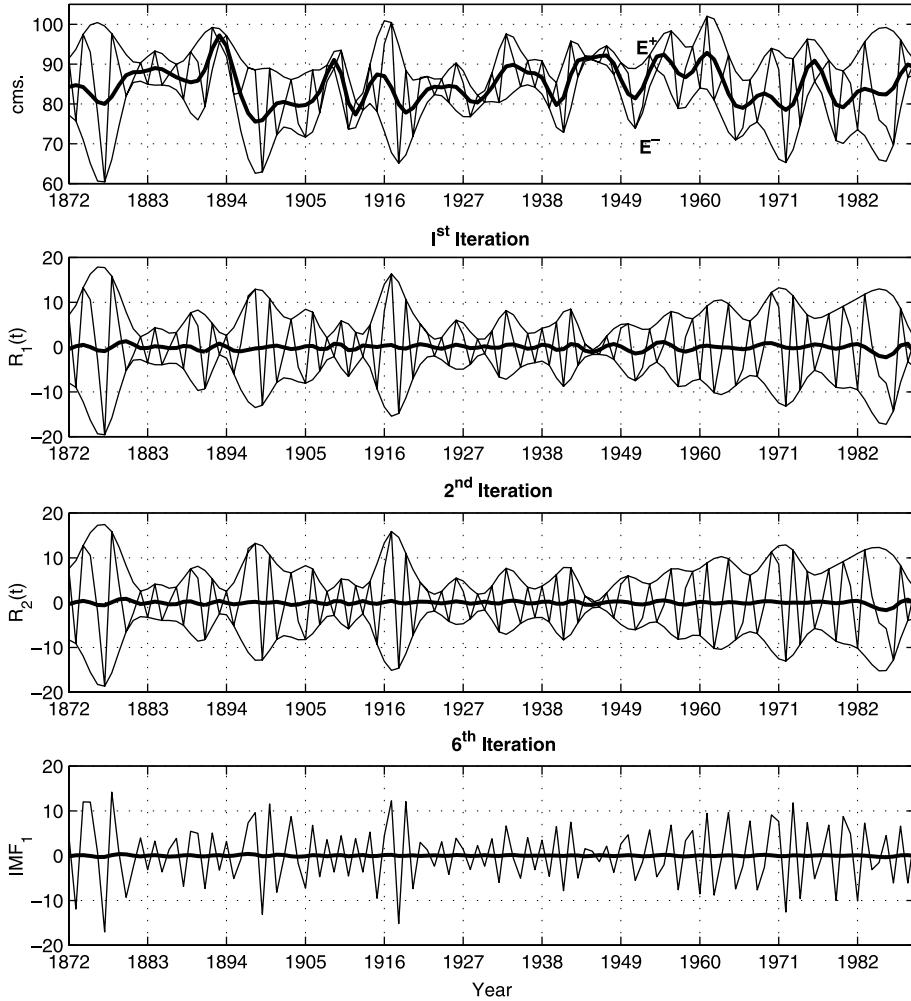


Fig. 2. Extraction of IMF_1 from AIRF data

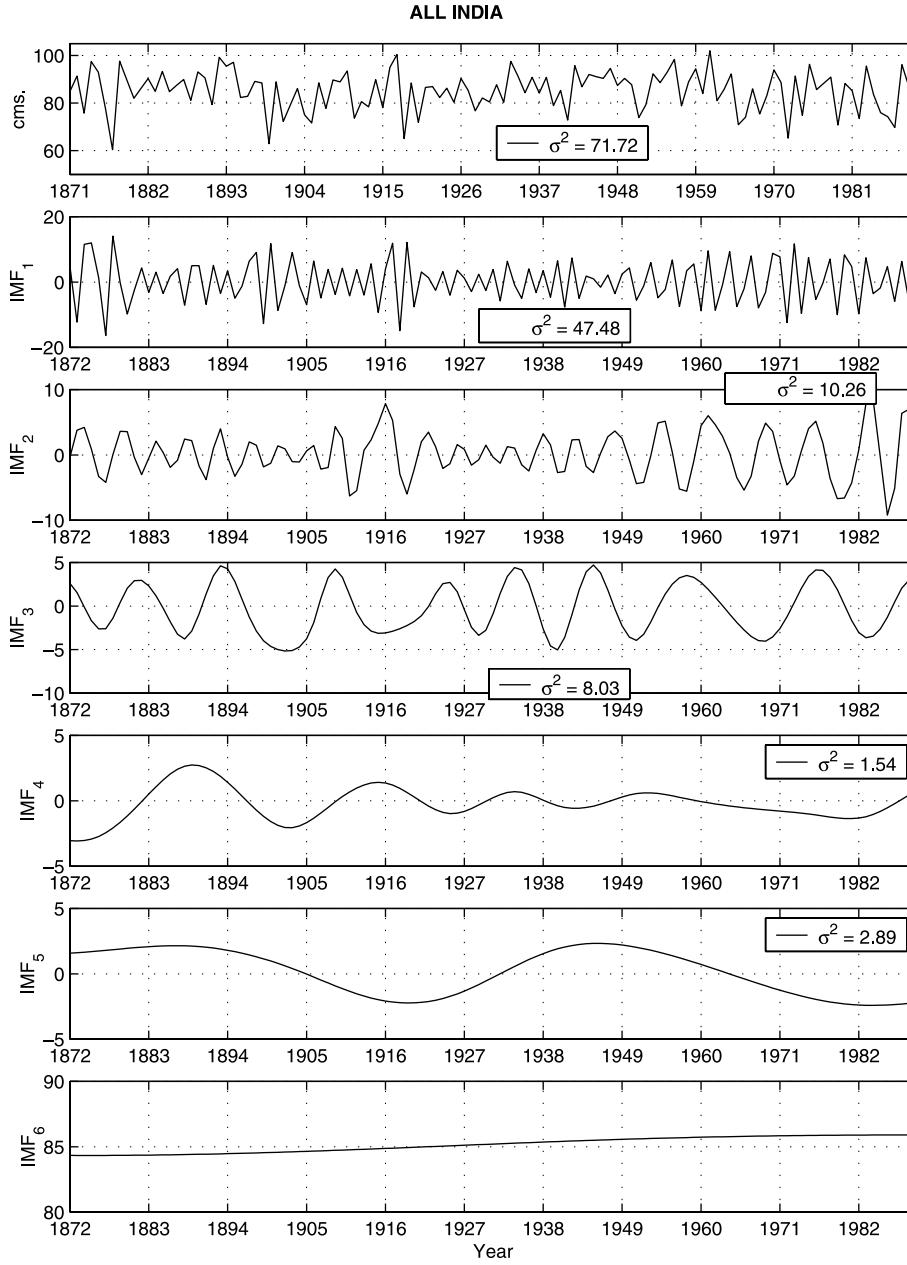


Fig. 3a. IMF's of All India rainfall

specific frequency or period. In all the figures, the variance of the basic data series and that of the IMF's found by time averaging is shown to indicate the relative contribution of an IMF to the total variability of the rainfall. It is easily observed that all IMF's exhibit slowly varying amplitudes and frequencies. Hence, these are narrow band processes with well-defined Hilbert transforms. However, even without such a representation the dominant period of oscillation can be found by counting the zeros and the extrema in an IMF. In Table 2, the central period found by this simple method is listed for all the regional

IMF's. The percentage of variance explained by each IMF, or the contribution of each IMF to IAV of the regional rainfall is also presented in Table 2. It is observed that all the regions exhibit IMF₁ to be the predominant mode with an average period of 2.7 years contributing to 50–70% of the inter-annual variability. IMF₂ is a mode with a dominant period of 5–6 years. This is the second most important mode in all the cases. These two modes are most probably connected with the quasi-biennial oscillation (QBO) and ENSO phenomenon, which show similar quasi-periodic structure around a period of 2–5 years. Similarly,

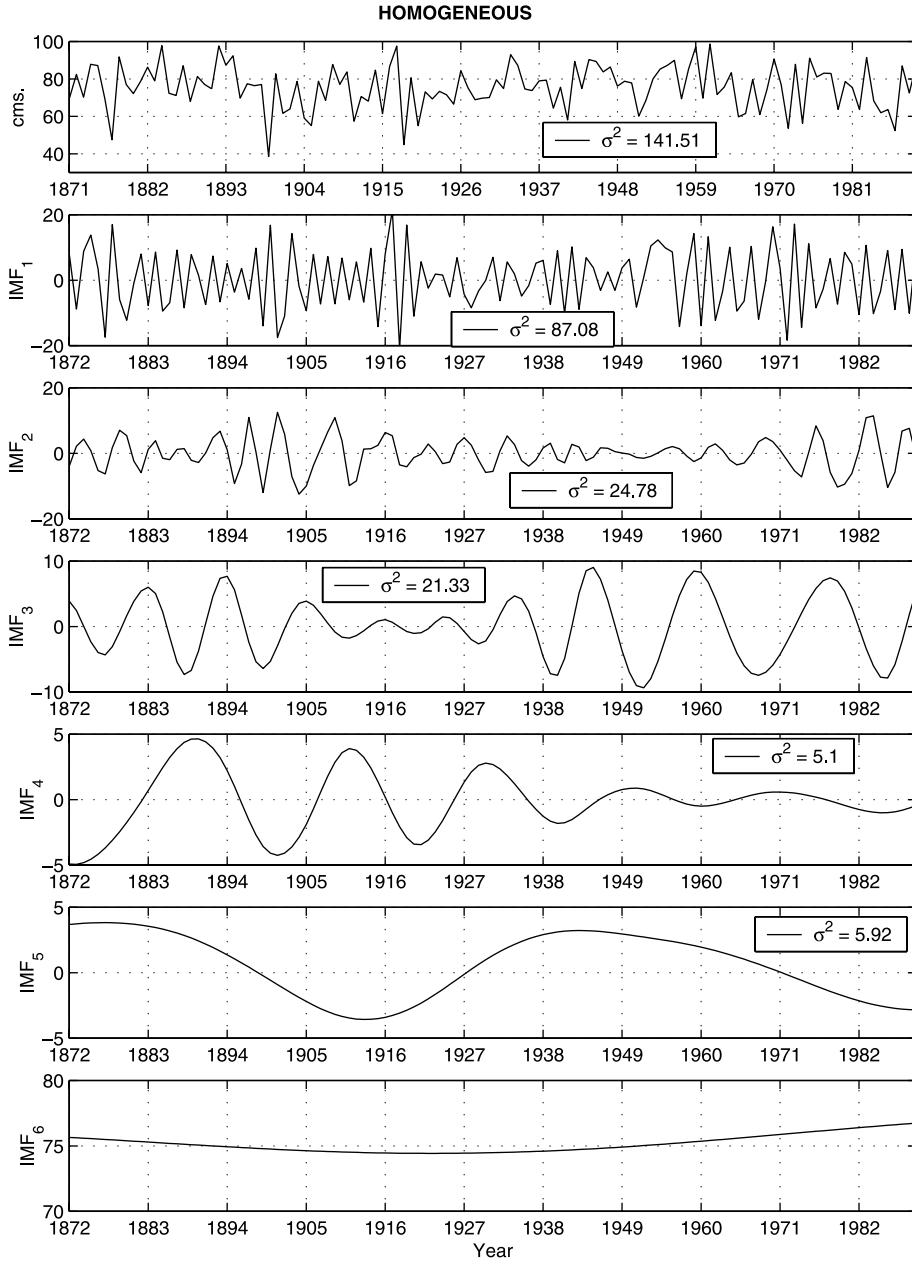


Fig. 3b. IMF's of homogeneous region rainfall

IMF₃ can be associated with the sunspot cycle of about 11 years period. Demonstration of this relation, in a physical sense, between IMF₃ and the sunspot data is presented in Fig. 4. In this figure, IMF₃ of AIRF, is compared with the annual European sunspot index available at www.sidc.oma.be. A case for linking SWM rainfall with sunspot cycle had been previously made by Bhalme and Jadhav (1984). However, physically quantifiable oscillations around 11 years in the rainfall data have been extracted and presented here. It may be noted here, that the same year or zero lag correlation between IMF₃ and sun spot

data is not significant due to drifting of phases in the time series. Nevertheless, the cross correlation peaks at 13 years lag to 0.4, which is statistically significant. Thus, association between the two data series is a distinct possibility. The central period of IMF₄ is about 20–24 years, which can be related to tidal forcing. Previously, Campbell et al (1983) have presented evidence for the presence of this quasi-cycle in Indian monsoon rainfall. The present analysis is able to trace this influence also from the data in a quantitative fashion. The fifth IMF shows an elongated period of the order of 60 years. The

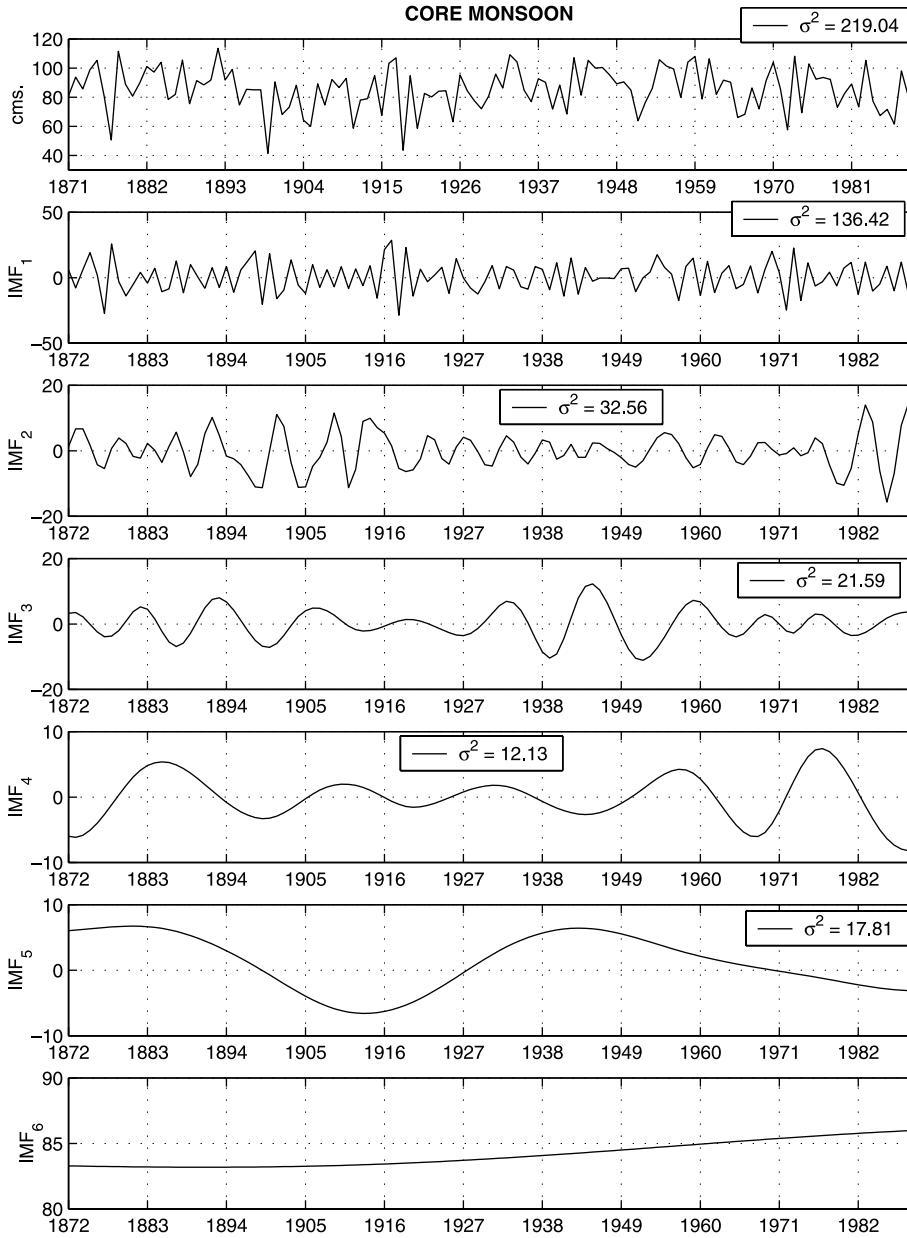


Fig. 3c. IMF's of core monsoon region rainfall

last component, which is the residue, as per Huang et al (1998), is here taken as the slowly varying climate mode. This way, IMF_6 is here identified as the deterministic long-term behavior. It may be mentioned here that wavelet analysis of monsoon rainfall data by Narasimha and Kailas (2001) indicated the presence of six quasi-cycles (modes) at nearly the same average periods obtained here. The present study has been able to identify the time histories of the embedded modes also in the form of various IMF's. The representation obtained for any of the data series is of the type $R(t) = \sum IMF_i(t)$. The sum of the IMF's leads to the original data,

as can be easily verified, in all cases. For example, the error between the sum of the six IMF's and the AIRF data series has an average value of 10^{-16} with a standard deviation of 10^{-14} .

5. IMF statistics

For understanding the statistical relation between the IMF's and the data, one has to construct the correlation matrix of the time series. In Table 3, the (6×6) correlation matrix of the AIRF data and the five variable IMF's is shown. It is immediately clear that correlation values between the data and the IMF are statistically significant

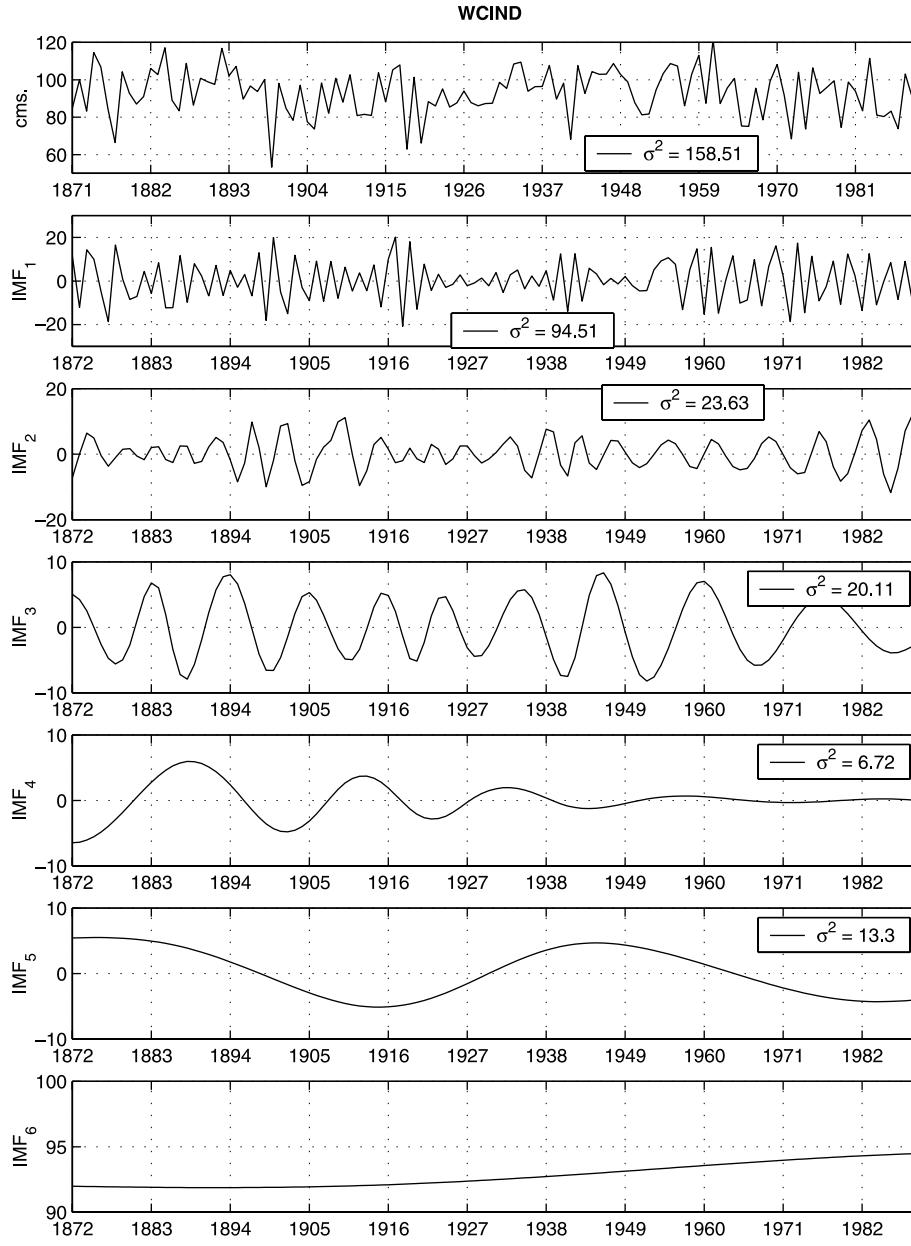


Fig. 3d. IMF's of WCIND region rainfall

and hence are physically meaningful. Further, among themselves the IMF's are statistically uncorrelated or orthogonal. Thus, we can expect the sum of the variances of the IMF's to be nearly equal to the total variance of the data. However, due to sample size effects and round off errors there can be small differences between the two-variance figures. For example, the sum of the variances of the IMF's of AIRF adds up to 70.5, whereas the data variance is 71.7. Wu and Huang (2003) have recently proposed an indirect way of testing the importance of the IMF's by comparing them with the IMF's of a known

white noise data series. In Fig. 5, this test is shown graphically for the above data. For a strict white noise, the variance of the IMF's and their corresponding central periods vary linearly on a log-log plot. Hence, for the data to be declared as pure noise, all the sample variances have to lie within a band of confidence. It is seen, from the results presented in Fig. 5, the five IMF's do not lie within the 99% confidence interval. Thus, the null hypothesis that the original data has no signals is rejected. The above characteristic has been verified to be true for the other seven sets of data also. These considerations point towards

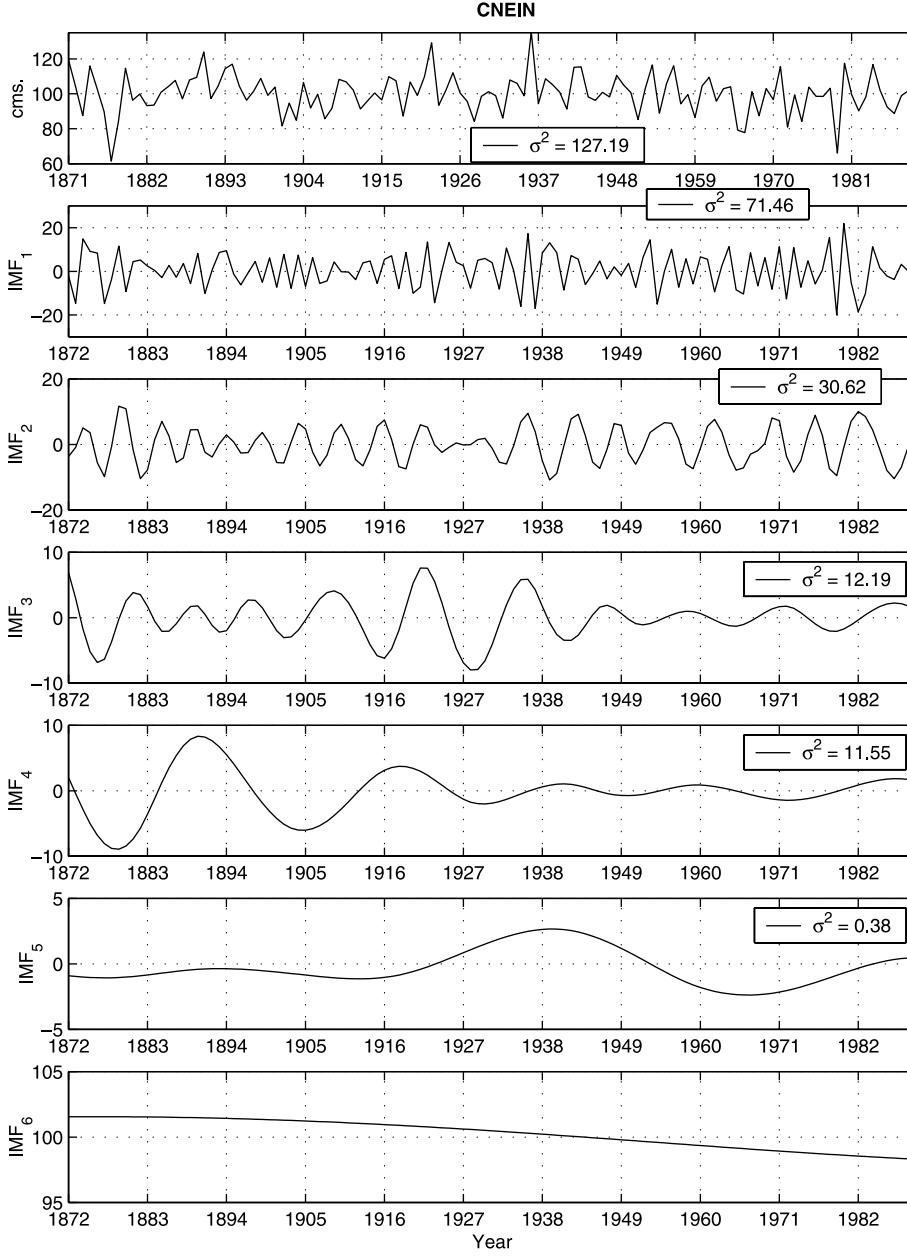


Fig. 3e. IMF's of CNEIN region rainfall

the possibility of statistical forecasting of SWM rainfall incorporating the IAV signatures through the IMF's. The first IMF carries the higher frequency end of the information and hence is expected to be more random than others. One way of describing uncertainty in rainfall is through the probability density function of the data. It is known that rainfall, as a random variable is non-Gaussian. This is true of the data studied here (Table 1) even though, being the sum of several individual variables, the seasonal data has a tendency towards being Gaussian. However, when decomposed into IMF's, a new

feature emerges clearly. It is the bimodality of the IMF_1 , which is also the most important component. To demonstrate this, the first IMF of five non-overlapping regions (WCIND, CENIN, NEIND, NWIND, PENIN) of Fig. 1b, are scaled by their standard deviation and pooled together. The resulting relative frequency structure is shown in Fig. 6. It may be noted here that, even without combining the samples, the bimodal behavior of IMF_1 is apparent from the time histories of Fig. 3a-h. This behavior indicates strong nonlinearity in the dynamics of the process and rules out the possibility of a

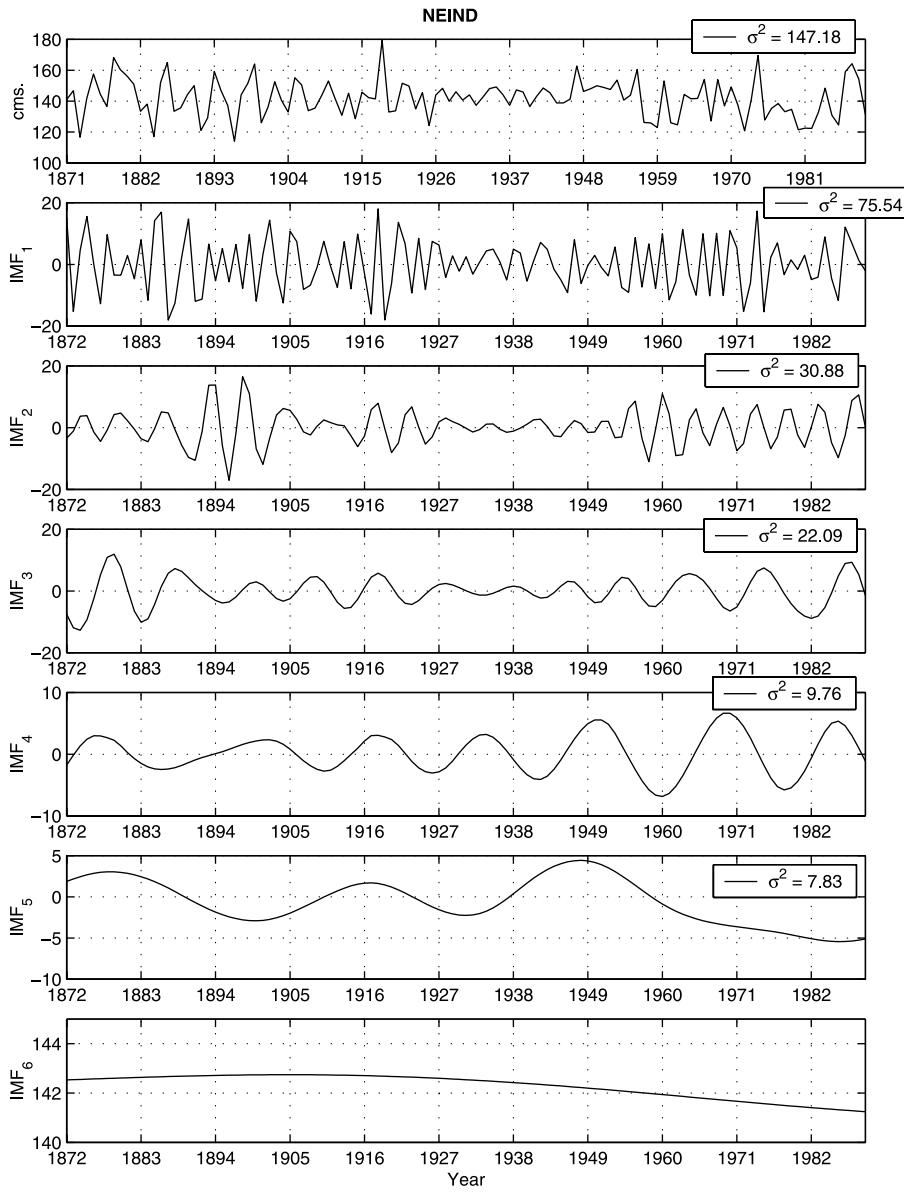


Fig. 3f. IMF's of NEIND region rainfall

linear time series model for IMF_1 . The remaining IMF's are unimodal and not strongly non-Gaussian. This raises the hope that they may be amenable for linear autoregressive representation.

6. Forecasting strategy

There is great interest among the agriculture, industrial and policy-making sectors in India to know in advance how the monsoon in a particular year behaves as far as rainfall is concerned. Thus, considerable literature exists on the various strategies adopted by the India Meteorology Depart-

ment (IMD) in producing a long range forecast for the All India seasonal rainfall (Rajeevan et al, 2000; Rajeevan, 2001). Forecasting may be seen as extending the data series by one step. This exercise, for simple functions with an analytic form can be easily carried out by Talyor's series expansion. However, rainfall data is highly erratic and no simple function can be fitted to the whole data series. Hence, the approaches taken have been statistical whether explicitly stated to be so or not. The decomposition of data into IMF's presents another approach for forecasting Indian monsoon rainfall. It is clear that one can attempt modeling and forecasting the IMF's,

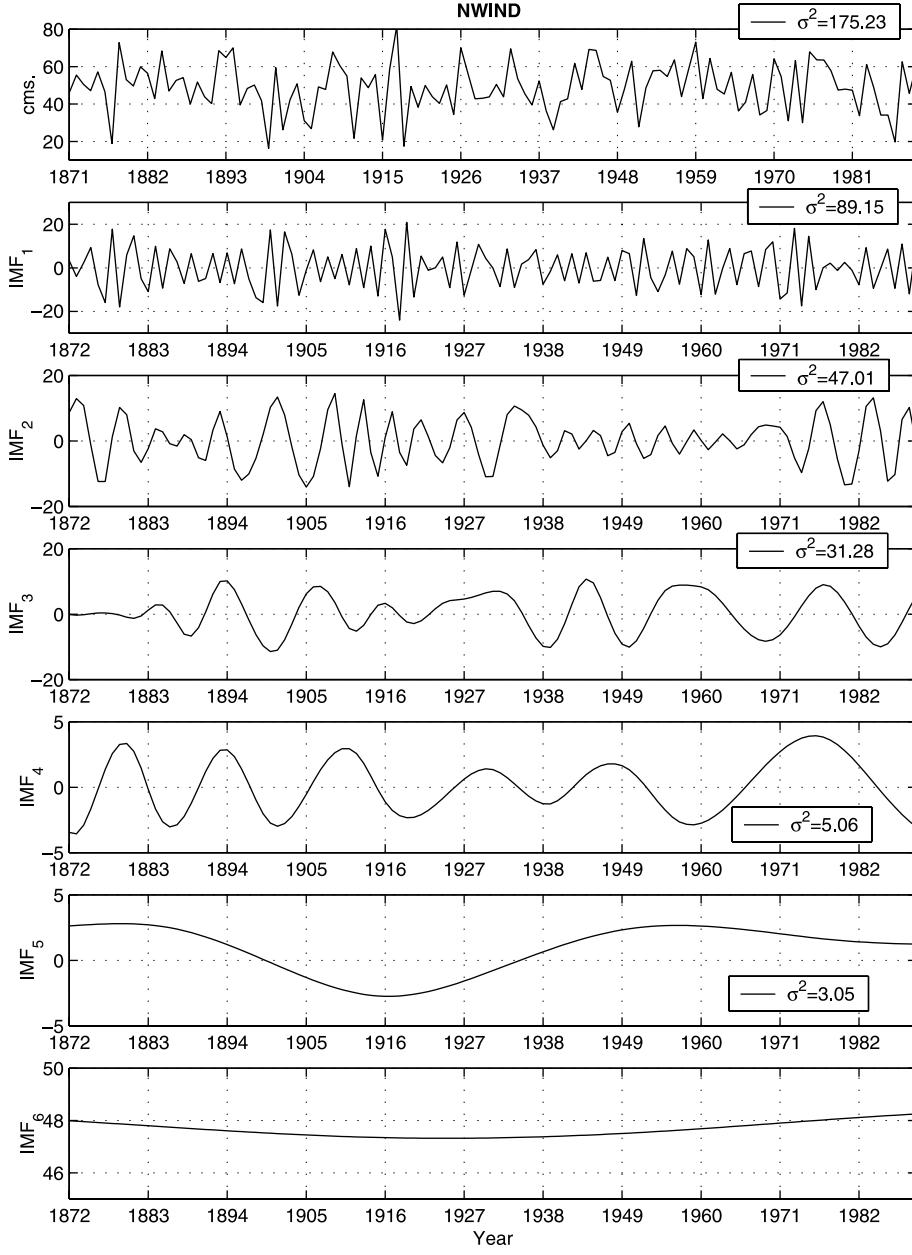


Fig. 3g. IMF's of NWIND region rainfall

which are simpler, instead of the original data. The sum of the predicted values of IMF's, leads to a forecast for the rainfall. However, difficulties arise in finding IMF's at the end of a record. This is because the envelop on both the sides at a point are not defined without the subsequent data point. Use of a mirror image of the past data for the next value, as suggested by Huang et al (1998) would not be acceptable in a forecasting exercise. Thus, if the data R_j is given for $j = 1, 2, \dots, n$, IMF's can be found only for $j = 2, \dots, (n - 1)$. Hence, for forecasting the value

of R_{n+1} one has to work without accurate values of IMF_{in} . This difficulty can be overcome by recognizing that except for the first IMF, others can be modeled through linear regression on their own past values. In fact for purposes of forecasting it is found easier to handle the data R_j as consisting of a nonlinear part and a linear part. The first IMF_{1j} represents the nonlinear part, whereas $y_j = (R_j - IMF_{1j})$, $j = (2, 3, \dots, n - 1)$; represents the linear part of the data. The stationarity of this part has also been verified by the standard run test on decadal variance values. For

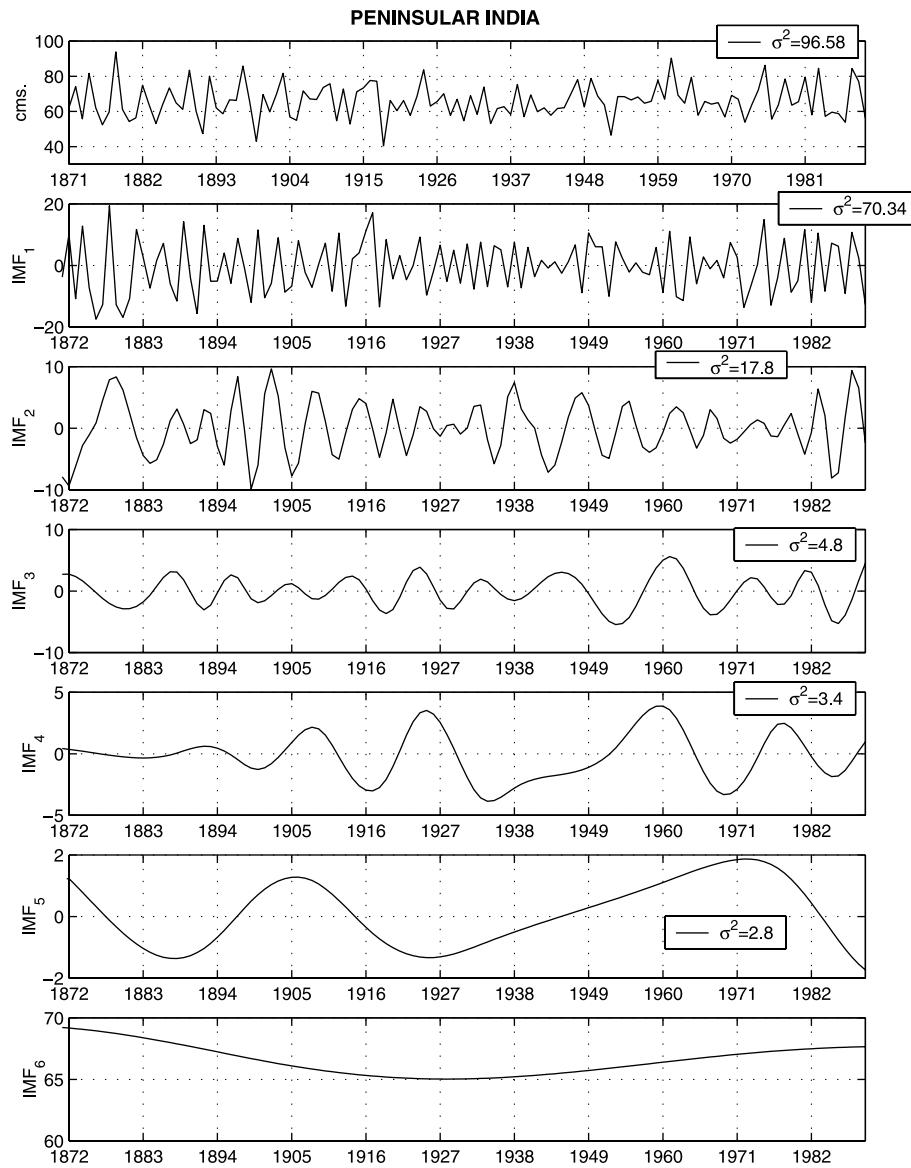


Fig. 3h. IMF's of peninsular region rainfall

Table 2. Central period of the IMF's in years and % variance contributed to IAV

Region	IMF ₁		IMF ₂		IMF ₃		IMF ₄		IMF ₅	
	T	IAV%								
All India	2.67	66.2	5.45	14.3	12.00	11.2	30	2.2	60	4.0
Homogeneous	2.72	61.5	5.71	17.5	12.00	15.1	24	3.6	60	4.1
Core monsoon	2.79	62.7	6.00	14.9	10.90	9.9	24	5.5	60	8.1
WCIND	2.61	59.6	5.71	15.5	12.00	12.7	24	4.2	60	8.4
CNEIN	3.08	56.2	6.32	24.1	10.90	9.6	24	9.1	60	0.3
NEIND	3.08	51.3	6.00	20.9	10.90	15.1	30	6.6	40	5.3
NWIND	2.79	59.2	5.45	25.4	13.30	19.2	20	1.5	60	0.4
Peninsular	2.72	72.8	6.00	18.4	10.00	4.97	20	3.5	60	2.9

y_j associated with AIRF, with $N = 13$, there are seven runs about the median value of the decadal variance. For the remaining data, in the order

listed in Table 1, the runs are [6, 7, 8, 8, 5, 6, 8]. The null hypothesis that the variance remains constant in time, implying stationarity, is

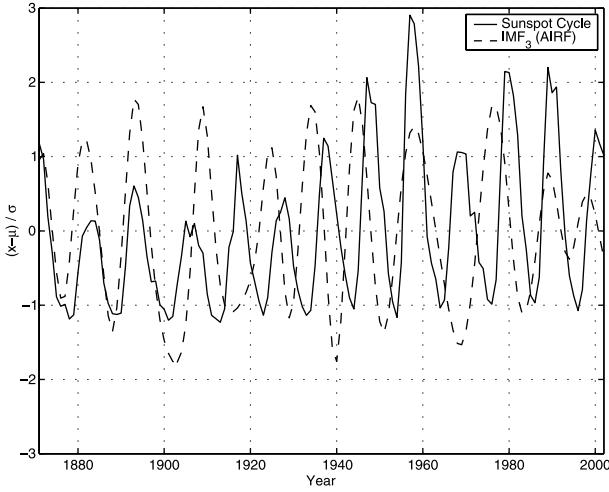


Fig. 4. Similarity between standardized IMF_3 and sunspot time series

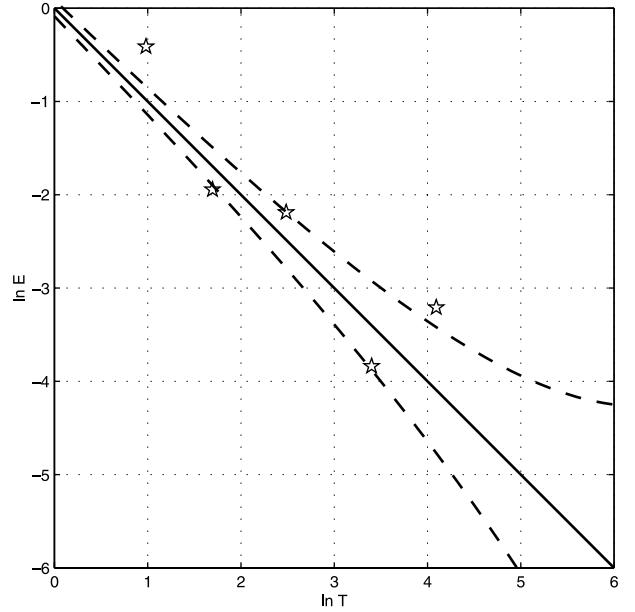


Fig. 5. White noise test for AIRF. \star IMF, — expected for white noise. - - - 99% confidence bands

Table 3. Correlation matrix of AIRF

Data	IMF ₁	IMF ₂	IMF ₃	IMF ₄	IMF ₅	
Data	1.0000	0.8019	0.4177	0.3502	0.1675	
IMF ₁		1.0000	0.0158	-0.0447	-0.0169	
IMF ₂			1.0000	0.0558	0.0368	
IMF ₃				1.0000	0.0967	
IMF ₄					1.0000	
IMF ₅						1.0000

accepted since the tabulated runs at 5% significance level are between 4 and 11. With this in view, the representation for the linear part avoiding y_n , is chosen as

$$y_{n+1} = C_1 R_n + C_2 y_{n-1} + C_3 y_{n-2} + C_4 y_{n-3} + C_5 y_{n-4} + C_6. \quad (1)$$

It is found that in all the eight cases, this equation provides an excellent fit for the linear part of

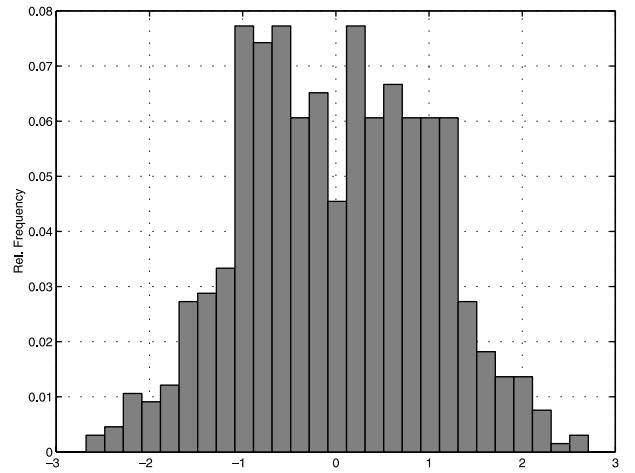


Fig. 6. Bimodal probability density function of IMF_1

Table 4. Regression coefficients of Equation 1

Region	C ₁	C ₂	C ₃	C ₄	C ₅	C ₆	$\sigma_y(\varepsilon)$	CC
All India	0.1053	1.1161	-1.8983	1.4354	-0.5281	64.5648	2.2458	0.9058
Homogeneous	0.1407	0.8528	-1.4464	1.2093	-0.5420	56.9314	3.5412	0.7802
Core Monsoon	0.1137	0.9231	-1.3782	1.1410	-0.4681	56.4207	3.9918	0.9003
WCIND	0.1199	1.2444	-2.0069	1.6592	-0.6767	61.2830	2.9735	0.9262
CNEIN	0.0801	1.3239	-2.1073	1.5395	-0.4399	60.8022	2.9904	0.9148
NEIND	0.0007	1.0650	-2.0051	1.6800	-0.6907	134.839	3.9708	0.8413
NWIND	0.1663	0.6081	-1.2352	1.1312	-0.6857	42.4480	3.4102	0.8351
Peninsular	0.0239	0.6536	-1.5749	1.3011	-0.6771	83.7672	2.7413	0.8965

the database. The regression coefficients are found from the data series of 1871–1991, such that IMF_1 and y_j are available for the period 1872 to 1990. The regression coefficients and the resulting standard deviation of the error $\sigma_y(\varepsilon)$ are presented in Table 4. In each case, the correlation coefficient (CC) between the actual data and fitted value as per the above equation is also presented in the table. In all the cases, the correlation is highly significant, indicating the appropriateness of identifying y_j as the linear part of monsoon rainfall.

7. Artificial Neural Network (ANN) model for IMF_1

It has been pointed out that the first IMF, which accounts for most of IAV of monsoon rainfall is non-Gaussian and is perhaps the outcome of a complex nonlinear process. It is not apparent from the time series of IMF_1 what type of nonlinear model would be appropriate. In such unstructured problems, it has been pointed out by Eisner and Tsonis (1992) that ANN approach can provide efficient working models. These authors have showed that ANN works for modeling and extending the chaotic trajectories of the Lorenz equation. Hsieh and Tang (1998) have highlighted the fact that ANN provides variational data assimilation models, which can be viewed as extensions of linear statistical models. Here, after several trials an ANN model with one hidden layer as shown in Fig. 7, depending on past five steps of data is chosen for IMF_1 . There are totally 36 parameters to be found in this model, which are found using the MATLAB toolbox on ANN algorithms, with 1872–1990

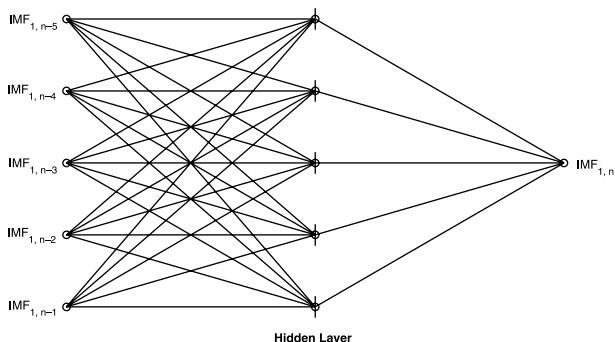


Fig. 7. ANN architecture for modeling IMF_1

Table 5. Statistics of ANN model for IMF. Training period (1872–1990)

Region	$\sigma_a(\varepsilon)$	CC
All India	2.8545	0.9255
Homogeneous	4.8702	0.9237
Core monsoon	4.1812	0.9260
WCIND	3.2231	0.9010
CNEIN	3.7801	0.9112
NEIND	4.8502	0.8847
NWIND	5.0850	0.9413
Peninsular	4.0511	0.9012

as the training period. For verifying the use of the model in forecasting, in any year n , the IMF_1 value is taken as the difference between the observed R_n and y_n of Eq. (1). With the help of five antecedent IMF_1 values, the ANN model is capable of predicting IMF_1 for the year $n + 1$. In Table 5, the standard deviation (σ_a) of the error in hind casting, conducted on the training period data, is shown along with the correlation coefficient (CC) between the actual IMF_1 and the ANN results. It is observed that the ANN methodology is quite versatile in capturing the embedded nonlinear structure as evidenced by the high correlation between the actual data and simulated values. An advantage of this approach is that the error in the model can also be characterized statistically.

8. Forecasting

The successful modeling of IMF_1 and y_j can be exploited to extend the data by one year, to make a forecast of the next year rainfall value as a random variable with a standard deviation much less than the climatic deviation. This will have to be done in two steps, following the procedure described above, first for y_{n+1} and then for $IMF_{1,n+1}$. The sum of the two values produces a forecast for R_{n+1} , with a definable sample probability distribution. Here, the robustness of the forecast strategy is investigated by considering all the eight data sets for the period 1991–2002, which was deliberately kept out of the modeling exercise. The quality of modeling R_j in the training period (1872–1990) and the efficiency of one-step-ahead forecasting in the testing period (1991–2002) are presented in Table 6. This table indicates that uniformly, in all cases

Table 6. Performance of the modeling and forecasting strategy

Region	Modeling period (1872–1990)			Forecasting period (1991–2002)		
	$\sigma_m(\varepsilon)$	CC _m	PP _m	$\sigma_f(\varepsilon)$	CC _f	PP _f
All India	3.29	0.89	0.83	3.04	0.91	0.82
Homogeneous	6.04	0.87	0.74	4.44	0.89	0.67
Core Monsoon	5.54	0.86	0.79	6.58	0.85	0.64
WCIND	4.09	0.95	0.87	5.83	0.86	0.65
CNEIN	4.78	0.90	0.81	4.74	0.90	0.74
NEIND	5.88	0.88	0.76	12.07	0.78	0.56
NWIND	5.61	0.91	0.82	4.48	0.92	0.87
Peninsular	4.05	0.91	0.82	4.83	0.89	0.80

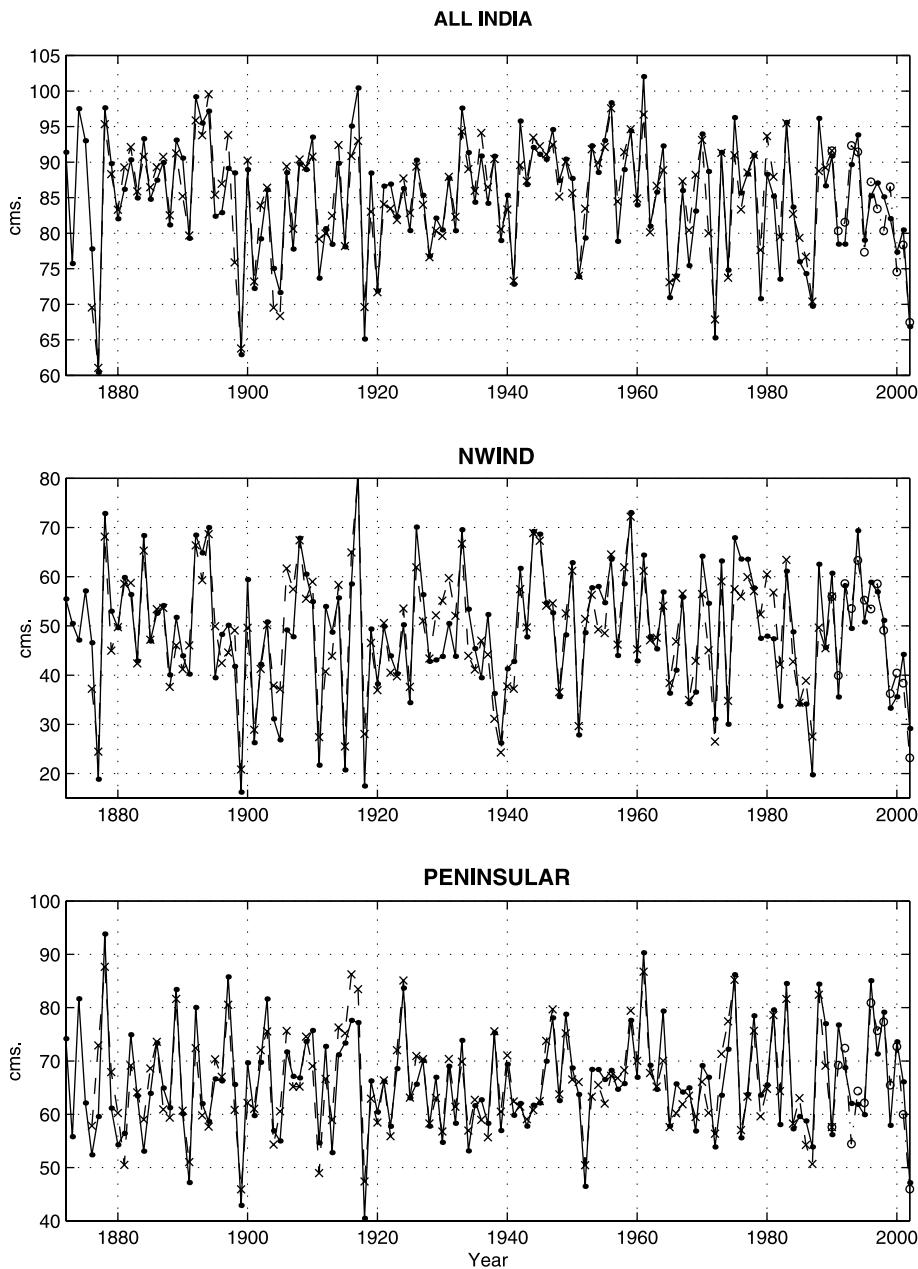


Fig. 8. Comparison between actual data and predictions; — Actual data, \times Modeling period, \circ Testing period

Table 7a. Independent test forecasting (NA: Not available)

Year	All India		Homogeneous		Core monsoon	
	Actual cms	Forecast cms	Actual cms	Forecast cms	Actual cms	Forecast cms
1991	78.47	80.32	64.50	73.69	74.60	77.24
1992	78.47	81.53	73.36	76.43	81.82	100.87
1993	89.66	92.33	79.95	76.23	90.15	96.53
1994	93.82	91.43	91.74	87.34	103.82	103.73
1995	79.03	77.34	64.74	70.78	71.60	68.89
1996	85.24	87.23	73.61	78.74	83.49	76.83
1997	87.06	83.43	71.94	74.88	81.64	81.18
1998	85.12	80.34	74.67	78.96	86.13	81.46
1999	82.05	86.54	67.90	64.67	78.05	76.03
2000	77.33	74.56	63.30	69.93	67.82	70.68
2001	80.46	78.34	67.80	71.13	69.90	70.78
2002	66.88	67.47	56.23	62.56	66.19	67.44
2003	87.06	85.65	NA	67.83	NA	78.19
2004	NA	80.34	NA	NA	NA	NA

Table 7b. Test forecasting with independent data

Year	WCIND		CNEIN		NEIND	
	Actual cms	Forecast cms	Actual cms	Forecast cms	Actual cms	Forecast cms
1991	83.56	83.37	90.67	84.78	138.40	131.59
1992	83.39	98.68	83.84	82.73	113.72	131.03
1993	100.04	108.27	98.60	104.32	174.01	165.09
1994	106.44	105.94	111.35	115.34	121.74	143.99
1995	73.94	73.92	93.32	96.36	165.36	149.67
1996	83.30	85.04	96.83	103.12	130.23	143.69
1997	81.85	74.01	109.80	101.56	154.43	147.26
1998	90.17	89.11	96.43	94.38	132.96	139.93
1999	90.70	87.65	110.91	105.45	149.42	141.58
2000	81.56	81.19	96.96	98.78	133.22	144.71
2001	83.35	81.59	107.30	117.58	131.27	137.65
2002	74.08	77.25	NA	90.78	151.32	150.75
2003	NA	84.08	NA	—	NA	140.74

the present strategy for forecasting SWM rainfall one year ahead, works well within certain limits. For three regions the modeling and forecasting results are shown graphically also in Fig. 8. The sample forecast is an expected value and hence need not precisely match with the actual observation. In Table 7a–c, detailed numerical results on the independent forecasts are presented. The low rainfall value of 2002, which lead to drought conditions in large parts of the country, has been very well foreshadowed by the present model. In Table 7a, the forecast for 2004 has been shown only for AIRF, since observed

data for other regions for year 2003 are not yet available.

9. Discussion

A novel approach for investigating interannual variability of monsoon rainfall of India has been presented here. The empirical decomposition of the time series data of eight different spatial regimes, in terms of IMF's brings out some interesting features of SWM rainfall. The first is the possibility of interpreting the rainfall in a given year as the sum of six independent modes.

Table 7c. Test forecasting with independent data

Year	NWIND		Peninsular	
	Actual cms	Forecast cms	Actual cms	Forecast cms
1991	35.58	39.92	76.75	69.22
1992	58.17	58.64	68.74	72.41
1993	49.50	53.58	62.03	54.43
1994	69.37	63.33	61.85	64.36
1995	50.80	55.24	59.93	62.14
1996	58.89	53.46	85.07	80.94
1997	56.93	58.57	71.33	75.68
1998	51.14	49.12	79.15	77.34
1999	33.30	36.23	57.93	65.45
2000	35.61	40.45	73.59	72.69
2001	44.21	38.36	66.08	59.96
2002	29.17	23.18	47.16	45.95
2003	NA	42.45	NA	62.99

Secondly, from the strong similarity of the first four IMF's in different regions, it is conjectured that these are associated with physical factors such as QBO, ENSO, sunspot and tidal forcing. Further, it emerges that the first IMF exhibits bimodality, which is characteristic of a nonlinear dynamical system. The contribution of this mode is of the order of 60% to the total IAV of the rainfall. The part of the data after removing IMF_1 is a stationary process represented in terms of a linear model. Since it is not obvious what type of nonlinear model has to be used for IMF_1 , after several trials, an ANN model with five input nodes and a single hidden layer with five nodes has been found to be suitable. To verify the robustness of the model proposed, three statistical parameters have been chosen. The first two are the r.m.s error $\sigma_m(\varepsilon)$ and the correlation coefficient CC_m between the given data and the values simulated out of the model. A third statistic called performance parameter (Sahai et al, 2003) $PP_m = 1 - \sigma_m^2 / \sigma_d^2$, where, σ_m^2 is the mean square error and σ_d^2 is the actual data variance, has also been found. In a perfect model, $\sigma_m(\varepsilon)$ will be zero and both CC_m and PP_m will tend towards unity. Table 6 demonstrates that the efficiency of the present model is excellent in all the cases. For verifying the ability of the model to independently forecast SWM rainfall one year ahead, 1991–2002 has been used as the testing period. The model parameters are held constant all through the twelve years, which represents

conditions more stringent than necessary. This is so since in a real forecasting exercise, the model parameters can be updated, every year before giving a forecast. It is observed that even under this less than ideal condition, the forecasts produced by the model are very good. For a sample size of $N = 12$, the correlation coefficient (CC_f) in the test period has to be at least 0.6 to be taken as significant. It is found from Table 6 that the forecasting skill is well beyond this threshold level, in all the eight cases. In Fig. 9, all the twelve-year predictions for the eight data series are combined to demonstrate that the proposed strategy for forecasting is highly efficient. A relevant question in this connection is the usefulness of the above statistical parameters in understanding the error of forecasting in a given year. With this in view, for all the eight regions, the actual and predicted values are compared in Table 7a–c, year by year. An empirical forecast for R_n is by its very nature statistical. Hence any forecast in Table 7 is to be treated as a random variable with $\sigma_m(\varepsilon)$ as its standard deviation. The probability distribution function of this variable can also be found in terms of the sample distribution of the model error ε . Thus, when a forecast is given, it is only appropriate to interpret it in terms of a cumulative distribution function as shown in Fig. 10. This figure presents the forecast of AIRF for the monsoon season of the year 2004. In the same figure, the climatic distribution of the data is also shown to highlight the reduction in variance obtained by the present forecast methodology. The drought of 2002 has been discussed previously by Gadgil et al (2002). Their conclusion that the drought was part of the ambient variability of the monsoon is corroborated by the present investigation, since the deficit has been foreshadowed in terms of only the past data. It has been pointed out that the last IMF is nearly the climatic average and the sum of the six IMF's gives back the actual rainfall. This property points to the fact that whenever the total rainfall is significantly below the climatic normal, the sum of the other IMF's would be negative. Among the five variable IMF's, it has been shown that the first three contribute nearly 85% to the variability. Hence, if these are simultaneously negative, the chances of a drought are high. For flood like conditions, the first three IMF's are likely to be strongly positive. This effect can be observed in Table 8, wherein for droughts and

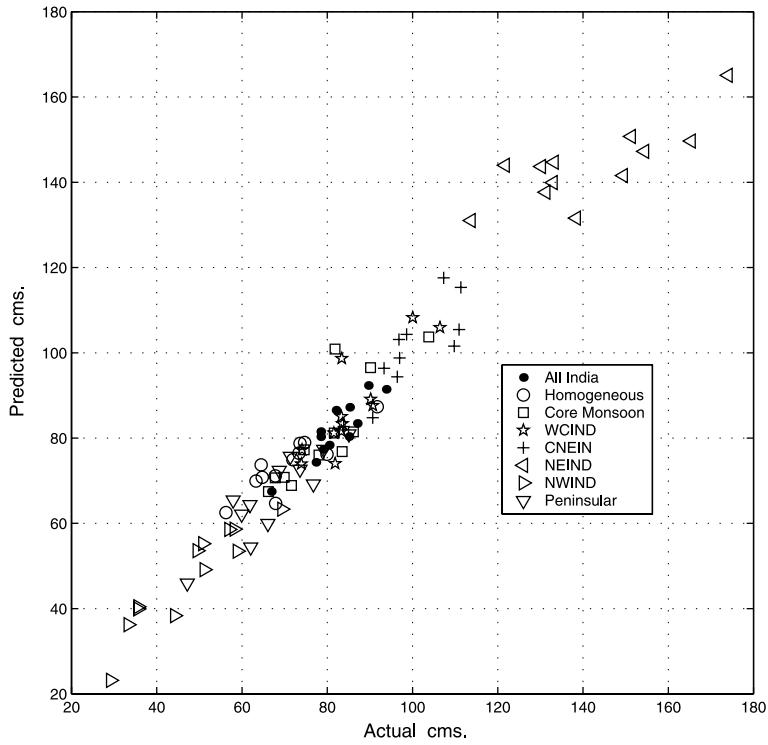


Fig. 9. Comparison between actual rainfall and one year ahead predictions during testing period (1991–2002). Correlation coefficient = 0.97

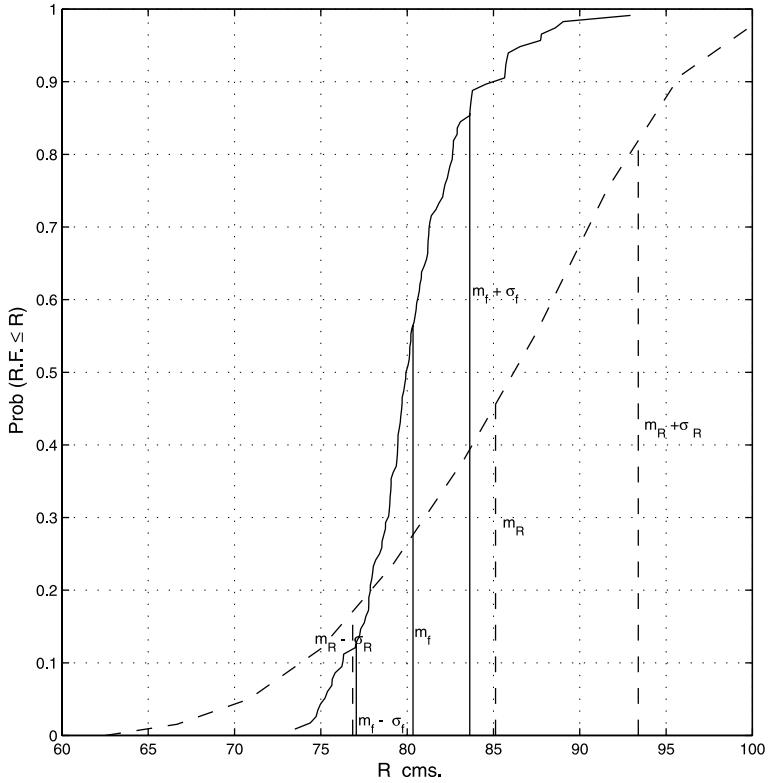


Fig. 10. Statistical forecast of AIRF for 2004 monsoon season. Expected value: $m_f = 80.34$ cms., Standard deviation: $\sigma_f = 3.3$ cms. Climatic distribution (---); Forecast distribution (—)

floods during 1901–2002, the actual All India rainfall and IMF_1 , IMF_2 and IMF_3 are shown. Here, droughts and floods are, respectively,

defined in terms of rainfall being less than $(m_R - \sigma_R)$ or greater than $(m_R + \sigma_R)$. The values of the IMF's shown in boldface refer to them

Table 8a. AIRF drought years and IMF's

Year	SWM $R < 76.78$	IMF_1 $\sigma = 6.52$	IMF_2 $\sigma = 3.31$	IMF_3 $\sigma = 2.76$
1901	72.23	-8.17	0.98	2.61
1904	75.06	-1.77	-0.60	-1.65
1905	71.66	-7.25	0.71	-2.56
1911	73.67	-4.56	-6.09	-0.99
1918	65.11	-15.16	-2.38	0.98
1920	71.91	-7.38	-2.39	-0.79
1941	72.84	-7.84	-2.39	-4.26
1951	73.89	-5.55	-4.47	-4.36
1965	70.94	-7.43	-3.62	-4.45
1966	74.02	-2.34	-5.07	-4.92
1968	75.46	-7.64	1.99	-4.99
1972	65.29	-12.60	-4.28	-3.08
1974	74.81	-9.57	0.61	-0.88
1979	70.79	-10.03	-6.49	4.58
1982	73.54	-10.11	1.29	0.83
1985	76.00	-1.54	-0.72	-3.16
1986	74.32	4.46	-9.12	-2.85
1987	69.73	-6.12	-5.15	-1.43
2002	66.88	-6.99	-1.94	-1.33

Table 8b. AIRF flood years and IMF's

Year	SWM $R > 93.71$	IMF_1 $\sigma = 6.52$	IMF_2 $\sigma = 3.31$	IMF_3 $\sigma = 2.76$
1916	95.06	4.43	7.72	1.16
1917	100.43	12.24	5.22	1.42
1933	97.59	6.65	1.10	4.25
1942	95.78	7.47	2.25	-1.41
1947	94.59	2.12	2.55	1.77
1956	98.35	6.80	0.39	3.75
1959	94.41	5.62	-1.19	3.06
1961	102.03	9.46	5.79	0.04
1970	93.97	9.09	3.48	-4.34
1975	96.27	7.35	4.12	0.47
1983	95.58	7.25	8.07	-1.06
1988	96.15	6.58	5.93	0.44
1994	93.82	5.93	3.08	-1.48

being outside their respective standard deviation range. Out of the nineteen droughts, of Table 8a, eighteen exhibit IMF_1 to be negative. Among these eighteen cases, in twelve years, IMF_1 was less than its negative σ value. In 1986, the drought was due to IMF_2 and IMF_3 being strongly negative. Similarly out of the thirteen flood years, nine were due to IMF_1 exceeding its positive σ level. These observations hint at the importance of understanding the physics of the first few IMF's in a better fashion.

10. Conclusions

Interannual variability of Indian monsoon rainfall has been investigated in this paper from a novel perspective. It is demonstrated that the seasonal SWM rainfall time series sampled annually can be decomposed into six statistically uncorrelated modes, the sum of which gives back the original data. The sixth mode is easily associated with the climatic variation over the whole period of the data base. The remaining five empirical modes are narrowband random processes, with well defined central periods, connected with other meteorological parameters. The first IMF which accounts for the highest variability, is also strongly non-Gaussian. It is shown that this can be modeled successfully using ANN techniques. The remaining part of the rainfall is amenable for a linear autoregressive representation. With the help of the above two representations, a methodology has been developed to forecast rainfall, incorporating only the interannual variability. This clearly does not account for all the variability present in the monsoon rainfall. However the approach is general enough, in that it may be possible to include interseasonal and intraseasonal variabilities also along the lines presented here. The forecast All India SWM rainfall for year 2004 is an expected value of $m_f = 80.34$ cms, which is 94.25% of the longterm normal. The standard deviation of the forecast distribution is $\sigma_f = \sigma_m = 3.3$ cms which is 39% of the climatic variation.

Acknowledgements

Authors are thankful to Profs. R. Narasimha, J. Srinivasan, Sulochana Gadgil, B. N. Goswami for useful discussions. Thanks are due to Dr. Rupakumar and Dr. Rajeevan Nair for several clarifications on the database of IITM and IMD.

References

- Bhalme HN, Jadhav SK (1984) The double (Hale) sunspot cycle and floods and droughts in India. *Weather* 39: 112
- Cadet DL, Daniel P (1988) Long-range forecast of the break and active summer monsoons. *Tellus* 40A: 133–150
- Campbell WH, Blechman JB, Bryson RA (1983) Long-period tidal forcing of Indian monsoon rainfall: an hypothesis. *J Clim Appl Meteor* 22: 287–296
- Eisner JB, Tsonis AA (1992) Nonlinear prediction, chaos, and noise. *Bull Amer Meteor Soc* 73: 49–60

Gadgil S et al (2002) On forecasting the Indian summer monsoon: the intriguing season of 2002. *Curr Sci* 83(4): 394–403

Gowariker V et al (1989) Parametric and power regression models: New approach to long range forecasting of monsoon rainfall in India. *Mausam* 40: 115–122

Hartmann DL, Michelsen ML (1989) Intraseasonal periodicities in Indian rainfall. *J Atmos Sci* 46(18): 2838–2862

Hastenrath S, Greischar L (1993) Changing predictability of Indian monsoon rainfall anomalies? *Proc Indian Acad Sci (Earth Planet Sci)* 102(1): 35–47

Hsieh WW, Tang B (1998) Applying neural network models to prediction and data analysis in meteorology and oceanography. *Bull Amer Meteor Soc* 79(9): 1855–1870

Huang NE et al (1998) The empirical mode decomposition and the Hilbert spectrum for nonlinear and nonstationary time series analysis. *Proc Royal Soc London A* 454: 903–995

Iyengar RN, Raghu Kanth STG (2003) Empirical modeling and forecasting of Indian monsoon rainfall. *Curr Sci* 85(8): 1189–1201

Narasimha R, Kailas SV (2001) A wavelet map of monsoon variability. *Proc Indian Natl Sci Acad* 67(3): 327–341

Parthasarathy B, Pant GB (1984) The spatial and temporal relationships between the Indian summer monsoon rainfall and the Southern Oscillation. *Tellus* 36A: 269–277

Raja Rao KS, Lakhole NJ (1978) Quasi-biennial oscillation and summer southwest monsoon. *Indian J Meteor Hydrol Geophys* 29: 403–411

Rajeevan M (2001) Prediction of Indian summer monsoon: Status, problems and prospects. *Curr Sci* 81(11): 1451–1457

Rajeevan M, Guhathakurta P, Thapliyal V (2000) New models for long range forecasts of summer monsoon rainfall over Northwest and Peninsular India. *Meteorol Atmos Phys* 73: 211–225

Sahai AK, Soman MK, Satyan V (2000) All India summer monsoon rainfall prediction using an artificial neural network. *Climate Dyn* 16: 291–302

Sahai AK, Grimm AM, Satyan V, Pant GB (2003) Long-lead prediction of Indian summer monsoon rainfall from global SST evolution. *Climate Dyn* 20: 855–863

Shukla J, Paolino DA (1983) The southern oscillation and long-range forecasting of the summer monsoon rainfall over India. *Mon Wea Rev* 111: 1830–1837

Thapliyal V (1990) Large-scale prediction of summer monsoon rainfall over India: Evolution and development of new models. *Mausam* 41: 339–346

Walker GT (1923) Correlations in seasonal variations of weather, VIII. *Mem India Meteor Dept* 24: 75–131

Wu Z, Huang NE (2003) A study of the characteristics of white noise using the Empirical Mode Decomposition Method. *Proc Royal Soc London A* (accepted)

Authors' addresses: R. N. Iyengar, Indian Institute of Science, Center for Atmospheric and Oceanic Sciences, Bangalore 560012, India (E-mail: rni@civil.iisc.ernet.in); S. T. G. Raghu Kanth, Indian Institute of Science, Department of Civil Engineering, Bangalore 560012, India