Proc. Indian Acad. Sci. (Earth Planet. Sci.), Vol. 101, No. 1, March 1992, pp. 27–34.

The chaotic time series of Indian monsoon rainfall and its prediction

SUJIT BASU and H I ANDHARIA

Meteorology and Oceanography Division, Space Applications Centre, Ahmedabad 380 053, India

MS received 4 July 1991; revised 11 November 1991

Abstract. The time series of Indian summer monsoon rainfall for the period 1871–1989 has been analysed using the method of deterministic chaos. It is found that a strange attractor underlies the time series implying the existence of a prediction function. This function has been approximated by a second-degree polynomial, involving the rainfalls of the past seven years and the coefficients have been estimated by least squares fit. The interannual variations of actual and computed rainfalls have been presented for a comparative study.

Keywords. Chaos; strange attractor; Indian monsoon rainfall.

1. Introduction

Summer monsoon rainfall is crucial to India because of its impact on agriculture and consequently on the economy. In view of this impact, a reasonable forecast of this rainfall, as far advance as possible, is of vital importance for the planners. Presently available statistical techniques using limited data set (Das 1987; Shukla and Mooley 1987; Hastenrath 1988) treat the time series of monsoon rainfall essentially as the realization of a stochastic process and the resulting forecast equations are regression relations with one or more predictor variables, the physical connections of which with the monsoon are still poorly understood. Apart from this, some of the predictors, e.g., April latitude position of the 500 mb ridge at 75°E (Mooley *et al* 1986) are available only a few months before the monsoon onset making the economic planning rather difficult.

In this note, we present an alternative approach based on the theory of chaos (Moon 1987) which treats the time series of monsoon rainfall (hereafter monsoon rainfall will be used to mean summer monsoon rainfall) as deterministic but, possibly, chaotic. Although this method has its own limitations, it is free from the disadvantages mentioned above. The resulting forecast formula, although an approximate one, uses only the rainfalls of past seven years as predictors, making a forecast possible eight months in advance.

2. Method

The method followed by us basically consists of reconstruction of the attractor underlying the time series of monsoon rainfall, estimation of the attractor dimension and the corresponding embedding dimension (Packard *et al* 1980; Grassberger and

Procaccia 1983) followed by approximation of the prediction function (Henderson and Wells 1988). The method is well-known and has been recently applied to meteorological data (Nicolis and Nicolis 1984; Fraedrich 1986; Satyan 1988). The work of Satyan (1988), although somewhat similar in nature to the present one, did not attempt to predict the monsoon rainfall of a particular year using the values of past several years as predictors. Below, we briefly sketch the outlines of the method adopted by us.

First, we embed the attractor into an m-dimensional pseudophase-space spanned by the time series and its time-shifted values so that a point in the space is described by

$$X(t) = \{x(t), x(t+\tau), \ldots, x[t+(m-1)\tau]\}, \tag{1}$$

where $x(t)$ is the time series and τ is the fixed time delay (one year, in our case).

The next step involves the calculation of the cumulative correlation function (Henderson and Wells 1988):

$$C(l) = \frac{1}{N^2} \sum_{i,j=1}^{N} H(l - r_{ij}), \tag{2}$$

where r_{ij} is the Euclidean distance between the ith and jth points, N is the total number of points, l is a distance variable and H is the Heaviside function with $H(x) = 0$ if $x \leqslant 0$ and $H(x) = 1$ if $x > 0$. Next we calculate the slope $d(m)$ of the linear part of the ln $C(1)$ vs. ln 1 curve by fitting a least squares line. By repeating the process for increasing m ($m = 1, 2, 3, \ldots$) we obtain successive estimates of the attractor dimension. If the slopes converge to a limiting value $d_\infty = d(M) = d(M+1) = \ldots$, then d_∞ is the true correlation dimension of the attractor and the corresponding embedding dimension M is a measure of the number of variables sufficient to model the dynamics (Fraedrich 1986). A noninteger value of d_∞ indicates the presence of a strange attractor, a term coined by Ruelle and Takens (1971), signifying deterministic but chaotic dynamics. It is interesting to note that there will be no saturation of slopes for a purely random time series.

The final step is the construction of a prediction function f, existing because of the deterministic dynamics (provided a convergence of the slopes is obtained), such that any entry in the time series is predicted by the previous ones in a fixed manner (Henderson and Wells 1988):

$$x_{k+M+1} = f(x_{k+1}, x_{k+2}, \ldots, x_{k+M}). \tag{3}$$

The prediction function f is necessarily of a nonlinear nature, and of course, unknown. In practice, one has to approximate it by the standard least squares technique.

3. Results

We applied the method sketched in § 2 to the 119-year time series of Indian monsoon rainfall starting from 1871. The data from 1871 to 1987 were kindly provided to us by B Parthasarathy and the data for the years 1988 and 1989 were obtained from the India Meteorological Department.

In figure 1 we present the ln $C(1)$ vs ln 1 curves for increasing embedding dimensions. Figure 2 shows the initial increase of estimated dimension followed by a saturation

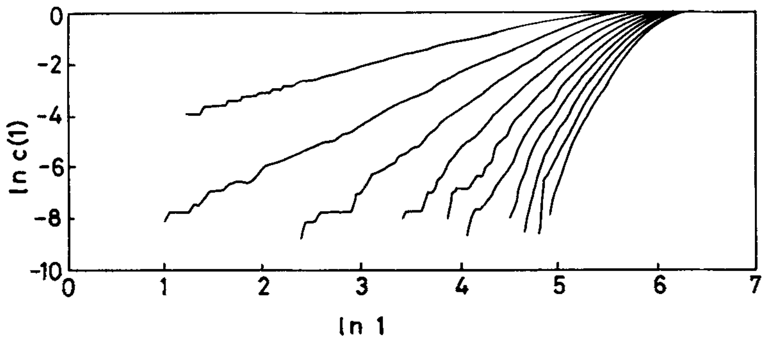

Figure 1. Cumulative correlation function of distances of the 119-year time series of Indian monsoon rainfall in m-dimensional pseudophase-spaces ($m = 1$ to 10) of time-delayed ($\tau = 1$ year) coordinates. m increases from left to right.

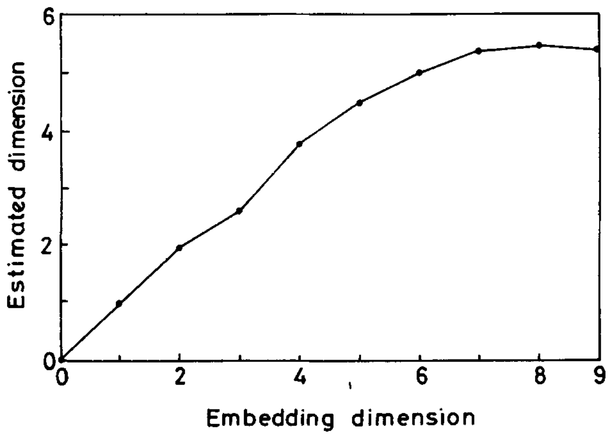

Figure 2. Estimated dimension (d) of the monsoon rainfall attractor as a function of the embedding dimension (m).

at $M = 7$. The fractal dimension 5·4 of the rainfall attractor indicates chaotic dynamics. However, the actual value of this fractal dimension was not used by us in the prediction analysis. This noninteger value served just as a pointer of chaos. We have used rather the value $M = 7$ of the embedding dimension to fix the number of predictors, because this gives the number of variables sufficient to model the dynamics. If so desired, this value may be looked upon as empirically derived.

The prediction function of equation (3) which is inevitably nonlinear in some or all of the independent variables, is unknown. Since there is no general method of finding it, we resorted to the simplest approximation, i.e. approximation by a general second-degree polynomial in seven variables. The number of coefficients is 36 (since there are $7(7 + 1)/2 = 28$ quadratic terms, 7 linear terms and one constant term). These were estimated by least squares fit.

To see the stability of the performance of the prediction function, we first used only 108 years of data starting from 1871 to estimate the coefficients. Next we computed rainfall of 113 years starting from 1878 using the prediction polynomial and calculated root mean square deviation between the actual and the computed rainfalls. This was repeated by including successively one more year of data while estimating the coefficients. We found that root mean square error was rather insensitive to inclusion of more data. The individual coefficients also changed very little. This shows that the performance of the regression was quite stable. Finally we opted for the regression, where 119 years of available data were used for estimating the coefficients. A statistical F-test was also performed to test the overall performance of the regression. The result of the test showed that the performance was satisfactory.

We provide below the prediction formula. The rainfall of ith year in centimetres is given by

$$R_i = \sum_{\substack{j,k=1 \\ j \leq k}}^{7} A_{jk} R_{i-8+j} R_{i-8+k} + \sum_{j=1}^{7} B_j R_{i-8+j} + C. \quad (4)$$

The coefficients A_{jk} are given by the following (7×7) upper triangular matrix (all the values are to be multiplied by 10^{-3} before use):

−22·3086	−54·4345	−41·8242	−13·7864	5·5128	17·1401	30·2722
—	−15·9128	−18·5298	−5·3283	−0·9656	1·0525	−24·2074
—	—	19·2183	32·9514	−12·3363	−15·9538	−3·5932
—	—	—	−12·2469	13·1628	3·6420	−10·7390
—	—	—	—	−11·9669	7·8672	7·5383
—	—	—	—	—	−6·2117	−0·0266
—	—	—	—	—	—	13·3680

The vector B is given by:

8·5597 11·4801 1·9006 0·2087 0·3528 0·0605 −2·23240

and constant $C = -782{\cdot}939$.

The RMS error of the regression is 5·84 cm which is much less than the standard deviation of the data which is about 8·1 cm (for data from 1871 to 1990). We also found the correlation between the predicted and the observed series. The coefficient of correlation was found to be 0·69. This coefficient was tested for statistical significance by performing student's t-test. It was seen that the correlation was significant at 1% level i.e. the level of confidence was 99%.

In figure 3 we present the interannual variation of the actual rainfall and rainfall computed by equation (4) for a comparative study. In this figure we have also given the result of computing rainfall using a seventh order linear autoregression. This formula also uses the rainfalls of the past seven years as predictors. The RMS error of this regression was 7·93 cm. By looking at the figure one can see the significant improvement achieved by using a second-degree polynomial.

It can be seen that the general trend of the data series is well-represented by the estimated rainfalls, although there are significant departures for a few individual years. Incidentally, the prediction for the year 1991 is 90 cm.

It would be quite interesting to compare our results with those of other investigators using the same data set as we have used here. Although we have used no other information apart from the rainfall time series, other researchers have used auxiliary information. Shukla and Mooley (1987), e.g., used 46 years of data to obtain multiple regression equations involving the following predictors: the April latitude position

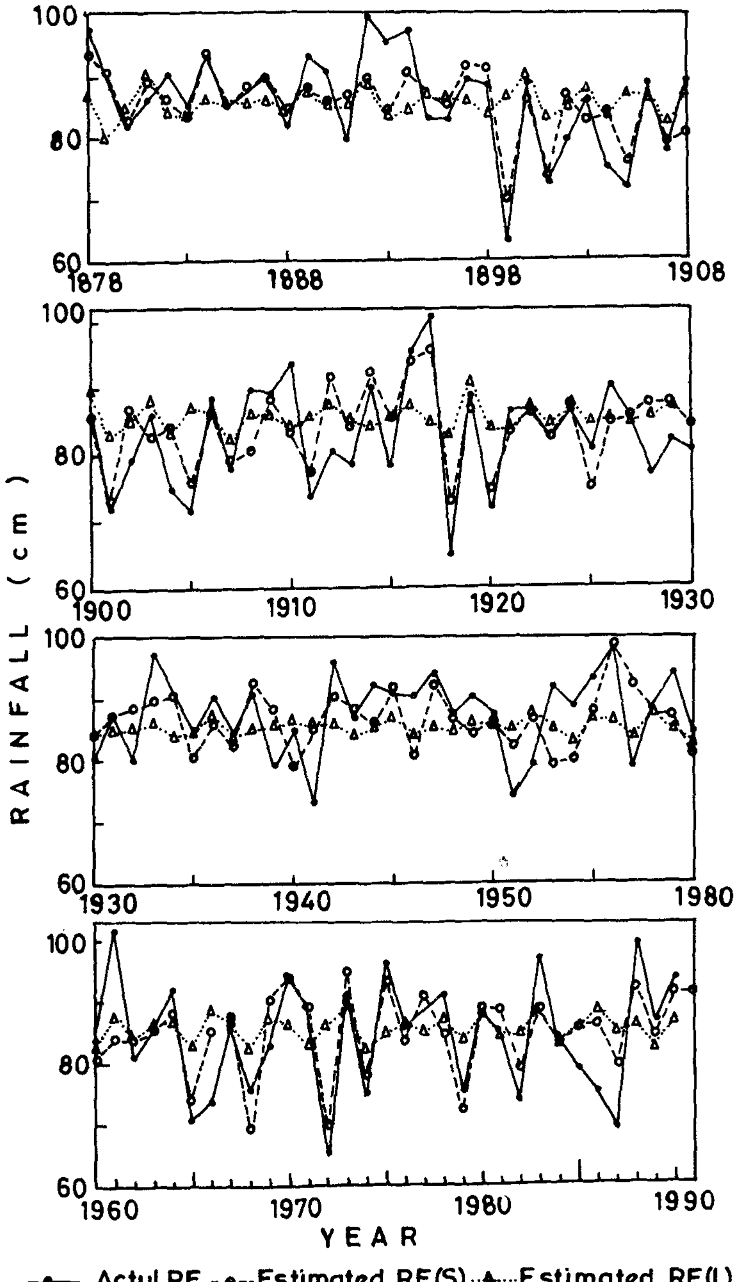

Figure 3. Interannual variations of the actual rainfall and the rainfalls estimated by the second degree polynomial approach as also by linear autoregression. The continuous line with dot denotes actual rainfall, the dashed line with circles denotes rainfall estimated by second-degree polynomial and the dotted line with triangles denotes rainfall estimated by linear autoregression. The value for the year 1991 is a predicted value.

of the 500 mb ridge at 75°E and the tendency of southern oscillation index (TSOI). The regressions were obtained for sliding 30-year periods from 1939–1984 and they were then used for obtaining forecasts for the rainfalls of the years 1939–1954 and 1969–1984. The coefficient of correlation (CC) between the observed and forecast rainfalls was found to be about 0·8 and the corresponding root mean square error (RMSE) was only 3·6 cm. In contrast to this, our computation for these same years shows a CC of only 0·65 with RMSE of 5·8 cm. However, for better comparison we should have resorted to the method of sliding periods. This is impossible because we need to estimate 36 coefficients for which data of 30 years are insufficient. More importantly, as observed by Hastenrath (1988), the excellent result obtained by Shukla and Mooley (1987) is basically because of a high degree of correlation of rainfall with the 500 mb ridge location in the earlier and latter parts of the period 1939–1984. Mooley and Paolino (1988) have also reported the result of applying multiple regression analysis using various sets of predictors. In the opinion of these authors the ridge location and mean May minimum temperature of a particular area of Indian landmass can serve as two good predictors. Using regression equations for sliding 25-year periods they have made forecasts for 24 years, 1939–1950 and 1964–1975. In table 1a we present their result as well as ours (for the same years) for a comparative study. The table gives RMSE, CC between observed and forecast rainfall and variance explained (ratio of variance of forecast rainfall and variance of observed rainfall expressed as percentage). In table 1b we present similar results for the 19 forecasts of rainfalls for the years 1939–1950 and 1969–1975.

Table 1a. Results of 24 forecasts of monsoon rainfall for the years 1939–1950 and 1964–1975.

Study	RMSE (cm)	CC	Variance explained (%)
Mooley and Paolino (1988)	4·17	0·88	85·0
Present	5·67	0·74	80·9

Table 1b. Results of 19 forecasts of monsoon rainfall for the years 1939–1950 and 1969–1975.

Study	RMSE (cm)	CC	Variance explained (%)
Mooley and Paolino (1988)	3·85	0·88	73
Present	5·48	0·72	76

Table 1c. Results of 24 forecasts of monsoon rainfall for the years 1916–1927 and 1964–1975.

Study	RMSE (cm)	CC	Variance explained (%)
Present	4·48	0·88	83

If there is a feeling from these tables that the CC between forecast and observed rainfall for our case is never very high and the RMSE is always large then that can be possibly removed by looking at table 1c where we present results of 24 forecasts but for a different group of years. The years involved are 1916–1927 and 1964–1975. One can see a significant improvement compared to table 1a. In fact there is a common subgroup of 12 years (1964–1975). This was just to show how one can arrive at quite different results by selecting different groups of years. In the language of Hastenrath (1988) one can say that one group of years is intrinsically more predictable than the other group.

So far we have compared our results with those of regressions using sliding periods. To be more reasonable one should compare our result with regression using a fixed set of data. Hastenrath did just that using three different predictors: the ridge location, TSOI and the May surface wind speed in an area in western equatorial Indian Ocean. Using a regression based on 20-years of data (1939–1958) he has computed rainfalls of 25 years (1959–1983). The CC between observed and forecast rainfall is 0·78 while the RMSE is 5·71 cm. For the same years our formula gives a CC of 0·76 with an RMSE of 5·76 cm. If one bears in mind that, apart from the rainfall data set, other auxiliary data are not used by us, the advantage becomes quite obvious. One can issue a prediction as soon as the rainfall value of the current year becomes available. One does not have to wait for the values of parameters like latitude of April 500-mb ridge. Again Hastenrath (1988) claims that the most representative is the prediction for the latter-15-year period (1969–1983). He obtained a CC of 0·89 and RMSE of 4·73 cm. Our corresponding results are 0·87 and 4·16 cm respectively.

Summarizing we can say the following. Our forecast has performed rather poorly compared to the forecast using sliding periods while its accuracy is comparable to that using a fixed set of rainfall data (and a variety of other auxiliary informations). Since our prediction formula is nothing but a crude approximation of a surface in $(7 + 1)$ dimensions by a second-degree polynomial it is not surprising that the surface closely approaches a second-degree surface only at a few selective regions of space. Rainfalls of the corresponding years are intrinsically more predictable. We can conclude by saying that our prediction formula, although not quite accurate, can still serve the purpose of providing a rough estimate of the rainfall of coming year immediately after the end of the monsoon of the current year.

4. Discussion

We are well aware of the limitations of our method. The length of the time series is, regrettably, too short and thus the attractor may not have been sufficiently uniformly covered by the data points. For a truly reliable estimate of the attractor dimension the length of the time series should have been sufficiently large. However, the actual value of the fractal dimension was not used in our prediction formula. We have only used the value of the corresponding embedding dimension ($M = 7$) to fix the number of predictor variables. If so desired this value may be alternatively looked upon as empirically derived.

The approximation of the prediction function by a second degree polynomial is not very accurate as has been found in the previous section. The somewhat smooth behaviour of the predicted rainfall compared to the actual one may be an attribute of this approximation. Higher order approximations are formidable because of the prohibitively large number of coefficients involved. In our case they are truly impossible because of the simple reason that the number of coefficients to be estimated exceeds the number of available data points. The number of coefficients in a general nth degree polynomial in m-dimensions is $(n+m)!/(n!m!)$. In our case $m=7$. Thus even for $n=3$ we have to estimate as many as 120 coefficients. The situation only worsens as n becomes higher and higher. Our study is thus more of a demonstrative nature. The aim was to demonstrate the role played by inherent nonlinearity of the problem. This should be clear from an inspection of figure 3 where our prediction formula has been compared with a linear autoregression.

However, it is quite possible that a cleverly constructed nonlinear function (not necessarily a polynomial) of the predictor variables may provide a better approximation of the prediction function than our simple second degree polynomial. We will be most happy if the meteorological community pays serious attention to this exciting problem.

Acknowledgements

The authors would like to express their deep gratitude to Dr P S Desai for many helpful discussions and to two anonymous referees for their valuable comments and suggestions.

References

Das P K 1987 Short and long-range monsoon prediction in India; in *Monsoons* (eds) J S Fein and P L Stephens (New York: John Wiley) pp. 549–578

Fraedrich K 1986 Estimating the dimensions of weather and climate attractors; *J. Atmos. Sci.* **43** 419–432

Grassberger P and Procaccia J 1983 Measuring the strangeness of strange attractors; *Physica* **D9** 189–208

Hastenrath S 1988 Prediction of Indian monsoon rainfall: Further experiments; *J. Climate* **1** 298–304

Henderson H W and Wells R 1988 Obtaining attractor dimensions from meteorological data; *Adv. Geophys.* **30** 205–237

Mooley D A, Parthasarathy B and Pant G B 1986 Relationship between Indian summer monsoon rainfall and location of the ridge at the 500 mb along 75°E; *J. Climate Appl. Meteorol.* **25** 633–640

Mooley D A and Paolino D A 1988 A predictive monsoon signal in the surface level thermal field over India; *Mon. Weath. Rev.* **116** 256–264

Moon F 1987 *Chaotic vibrations: An introduction for applied scientists and engineers* (New York: John Wiley)

Nicolis C and Nicolis G 1984 Is there a climatic attractor? *Nature (London)* **311** 529–532

Packard N H, Crutchfield J P, Farmer J D and Shaw R S 1980 Geometry from a time series; *Phys. Rev. Lett.* **45** 712–716

Ruelle D and Takens F 1971 On the nature of turbulence; *Commun. Math. Phys.* **20** 167–192

Satyan V 1988 Is there an attractor for the Indian summer monsoon? *Proc. Indian Acad. Sci. (Earth Planet. Sci.)* **97** 49–52

Shukla J and Mooley D A 1987 Empirical prediction of the summer monsoon rainfall over India; *Mon. Weath. Rev.* **115** 695–703