Advances in Water Resources 33 (2010) 327–347

Contents lists available at ScienceDirect

Advances in Water Resources

journal homepage: www.elsevier.com/locate/advwatres

Nonlinear ensemble prediction of chaotic daily rainfall

C.T. Dhanya, D. Nagesh Kumar *

Department of Civil Engineering, Indian Institute of Science, Bangalore 560 012, India

ARTICLE INFO

Article history:
Received 25 June 2009
Received in revised form 28 December 2009
Accepted 9 January 2010
Available online 15 January 2010

Keywords:
Chaotic nature of rainfall
Nonlinear prediction
Ensembles
Uncertainty

ABSTRACT

The significance of treating rainfall as a chaotic system instead of a stochastic system for a better understanding of the underlying dynamics has been taken up by various studies recently. However, an important limitation of all these approaches is the dependence on a single method for identifying the chaotic nature and the parameters involved. Many of these approaches aim at only analyzing the chaotic nature and not its prediction. In the present study, an attempt is made to identify chaos using various techniques and prediction is also done by generating ensembles in order to quantify the uncertainty involved. Daily rainfall data of three regions with contrasting characteristics (mainly in the spatial area covered), Malaprabha, Mahanadi and All-India for the period 1955–2000 are used for the study. Auto-correlation and mutual information methods are used to determine the delay time for the phase space reconstruction. Optimum embedding dimension is determined using correlation dimension, false nearest neighbour algorithm and also nonlinear prediction methods. The low embedding dimensions obtained from these methods indicate the existence of low dimensional chaos in the three rainfall series. Correlation dimension method is done on the phase randomized and first derivative of the data series to check whether the saturation of the dimension is due to the inherent linear correlation structure or due to low dimensional dynamics. Positive Lyapunov exponents obtained prove the exponential divergence of the trajectories and hence the unpredictability. Surrogate data test is also done to further confirm the nonlinear structure of the rainfall series. A range of plausible parameters is used for generating an ensemble of predictions of rainfall for each year separately for the period 1996–2000 using the data till the preceding year. For analyzing the sensitiveness to initial conditions, predictions are done from two different months in a year viz., from the beginning of January and June. The reasonably good predictions obtained indicate the efficiency of the nonlinear prediction method for predicting the rainfall series. Also, the rank probability skill score and the rank histograms show that the ensembles generated are reliable with a good spread and skill. A comparison of results of the three regions indicates that although they are chaotic in nature, the spatial averaging over a large area can increase the dimension and improve the predictability, thus destroying the chaotic nature.

1. Introduction

The theory of chaos which deals with unpredictable complex nonlinear systems had its breakthrough in the late 1800s, when Poincaré addressed the stability of the solar system and the position of planets. Later in 1963 Lorenz's study on the convectional rolls in the atmosphere lead to the rediscovery of chaotic motion of a strange attractor. He found that the solutions to his equations continued to oscillate in an irregular, aperiodic fashion instead of settling to an equilibrium condition. He also noted the system's sensitivity to initial conditions and hence the unpredictability [43]. Thereafter, the theory of chaos, in which a deterministic system exhibits aperiodic long term behaviour and depends sensitively on the initial conditions, gained popularity with the scientific community.

The features or qualities of a chaotic system can be summarized as: (i) they are deterministic, i.e., there are some determining equations ruling their behavior; (ii) they are sensitive to initial conditions (a slight change in the starting point can lead to significantly different outcomes); (iii) they are neither random nor disorderly. Chaotic systems do have a sense of order and pattern, even though they do not repeat.

Literature shows numerous applications of deterministic chaos in hydrology, particularly rainfall and runoff dynamics. Most of these studies investigate the existence of chaos in rainfall [12,25,27,34,35,37,38], runoff series [12,16,17,21,22,42,47], lake volume [29] and also rainfall disaggregation [40]. Chaotic dynamics of joint rainfall–runoff process was investigated by Sivakumar

* Corresponding author. Tel.: +91 80 22932666; fax: +91 80 2360040.
E-mail addresses: dhanya@civil.iisc.ernet.in (C.T. Dhanya), nagesh@civil.iisc.ernet.in (D. Nagesh Kumar).

0309-1708/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.advwatres.2010.01.001

et al. [36] considering the runoff coefficient as a parameter connecting rainfall and runoff. A few studies have also tried the nonlinear prediction of rainfall and runoff series by treating them as univariate series [10,12,21,22,38] and also as multivariate series [23] taking into consideration information from other time series. Much debate has occurred on the effects of the data size [18,26,41] and noise [3,11,14,31,39] in the estimation of the dimensions of a chaotic series.

A dynamical system is any system that evolves in time from some known initial state and can be described by a set of equations. This is usually described in terms of trajectories in the state space (which is a mathematically constructed abstract space in which each dimension represents one variable of the system). Since the set of equations is not known a priori in time series analysis, the state space is represented by a phase space which can be reconstructed from the time series itself. Each trajectory in the phase space represents the evolution of the system from one initial condition. An attractor can be defined as a subset of trajectories into which all trajectories, originating from different initial conditions, will eventually converge. Since this subset of trajectories attracts all other trajectories in the phase space, it is called the attractor of the system. For a time series of a regular system, the attractor will be of integer dimension. But for a chaotic system, which is irregular and is sensitive to the initial conditions, the attractor may be characterized by a non-integer dimension.

Since the dynamics of a chaotic time series are not known, as also the original theoretical attractor, the state space of a scalar time series is approximated to a phase space where the attractor is reconstructed using the scalar series. Thus, phase space reconstruction provides a simplified, multi-dimensional representation of a single-dimensional nonlinear time series. Packard et al. [20] proposed a method of delays for reconstructing the phase space which was introduced and mathematically demonstrated by Takens [44]. According to this approach, for a scalar time series X_i where $i = 1,2,\ldots,N$, the dynamics can be fully embedded in m-dimensional phase space represented by the vector,

$$Y_j = (X_j, X_{j+\tau}, X_{j+2\tau}, \ldots, X_{j+(m-1)\tau}), \quad (1)$$

where $j = 1,2,\ldots,N-(m-1)\tau/\Delta t$; m is called the embedding dimension ($m \geqslant d$, where d is the dimension of the attractor); τ is the delay time and Δt is the sampling time. The dimension m can be considered as the minimum number of state variables required to describe the system. The popular methods used for estimating the embedding dimension are the Grassberger–Procaccia approach (GPA) [6,7], and the False Nearest Neighbour (FNN) method [15].

The delay time τ is the average length of memory of the system. An appropriate delay time is to be chosen for the best representation of a phase space. If τ is too small, the phase space coordinates would not be independent, resulting in some loss of information about the characteristics of the attractor structure. On the other hand, if τ is too large, then there would be no dynamic correlation between the state vectors since the neighbouring trajectories diverge, thus resulting in some loss of information about the original system. The optimum τ is usually determined using either autocorrelation function or the mutual information method [5]. For study of hydrological time series, the most popularly used method is the autocorrelation function. Several recommendations are available for the selection of τ from the autocorrelation function. Tsonis and Elsner [46] recommended that if the autocorrelation function is approximately exponential, then the delay time can be chosen as the lag time at which the autocorrelation falls below the threshold value e^{-1}. Another method is to take τ as the first lag time at which the autocorrelation crosses the zero line [9].

The estimation of these parameters depends on the methods employed and also on the noise of the time series. Noise in the series can be removed by smoothing [22,30], but it may alter the underlying dynamics of the series itself [3,39]. Considering all these uncertainties, while prediction, instead of a single forecast, an ensemble of forecasts has to be generated from a plausible combination of parameters. Such an ensemble approach provides an estimate of the forecast uncertainty and also the probability density function of the response variable.

The aim of this paper is to analyse the chaotic behaviour of a time series employing various techniques. The set of plausible parameters thus obtained are used to generate an ensemble of forecasts of the time series. The various methods usually employed are described first, followed by delineating the methodology used in this study. The methodology developed is demonstrated by applying it to the Malaprabha, Mahanadi and All-India daily rainfall series and finally the results are discussed.

2. Methods employed

A variety of techniques have emerged for the identification of chaos which include correlation dimension method [6], false nearest neighbour algorithm [15], nonlinear prediction method [4], Lyapunov exponent method [13], Kolmogorov entropy [7], surrogate data method [45], etc. In this study, correlation dimension, false nearest neighbour method and Lyapunov exponent are employed to analyse the chaotic nature of the time series. Nonlinear prediction method is used as an inverse method for chaos identification in addition to prediction. The nonlinearity of the time series is analysed by surrogate data method.

2.1. Correlation dimension method

In correlation dimension method also known as correlation integral analysis, the correlation integral $C(r)$ is estimated using the Grassberger–Procaccia algorithm [6]. This algorithm uses the reconstructed phase space of the time series in Eq. (1). According to the algorithm, for an m-dimensional phase space, the correlation integral $C(r)$ is given by

$$C(r) = \lim_{N\to\infty} \frac{2}{N(N-1)} \sum_{\substack{ij \\ (1\leqslant i<j\leqslant N)}} H(r - |Y_i - Y_j|), \quad (2)$$

where H is the Heaviside function, with $H(u) = 1$ for $u > 0$ and $H(u) = 0$ for $u \leqslant 0$, where $u = (r - |Y_i - Y_j|)$, r is the radius of the sphere centered on Y_i or Y_j and N is the number of data. For small values of r, the correlation integral holds a power law relation on r, $C(r) \sim r^d$, where d is the correlation dimension of the attractor. The correlation exponent or the dimension, d can be calculated from the slope of the plot of $\log C(r)$ versus $\log r$.

If the correlation exponent saturates to a constant value even on increase in embedding dimension m, then the series is generally considered to be chaotic. The nearest integer above that saturation value indicates the number of variables necessary to describe the evolution in time. On the other hand, if the correlation exponent increases without reaching a constant value on increase in the embedding dimension, the system under investigation is generally considered as stochastic. This is because, contrary to the low dimensional chaotic systems, stochastic systems acquire large dimensional subsets of the system phase space, leading to an infinite dimension value.

However, Osborne and Provenzale [19] opposed the traditional view that stochastic processes lead to a non-convergence of the correction dimension by demonstrating that "colored random noises" characterized by a power law power spectrum exhibit a finite and predictable value of the correlation dimension. Thus the sole presence of finite, non-integer dimension value is not suffi-

cient to indicate the presence of a strange attractor. While for the low dimensional dynamic systems the saturation of correlation dimension is due to the phase correlations, for the above mentioned stochastic systems it is mainly due to the shape of the power spectrum (power law). Hence, it would be worthwhile to carry out some additional tests to distinguish low dimensional dynamics and randomness [24].

2.1.1. Phase randomization

Stochastic surrogate data of the same Fourier spectra as that of the original data are generated. The Fourier phases are randomized and are uniformly distributed. Upon performing the correlation dimension method on this surrogate data, correlation dimension estimate will be invariant, if the convergence of the dimension is forced only by the shape of the power spectrum and not due to any low dimensional dynamics.

2.1.2. Signal differentiation

Another method is to take the first (numerical) derivative of the signal and examine its correlation integral. The correlation dimension of the differentiated signal will be much larger than that of the original signal in the case of stochastic systems. This is attributed to the change in the spectral slope on differentiation. For low dimensional dynamic systems, correlation dimension will be almost invariant.

2.2. False nearest neighbour method

The concept of false nearest neighbour was introduced by Kennel et al. [15]. This method is based on the concept that if the dynamics in phase space can be represented by a smooth vector field, then the neighbouring states would be subject to almost the same time evolution [14]. Hence, after a short time into the future, any two close neighbouring trajectories emerging from them

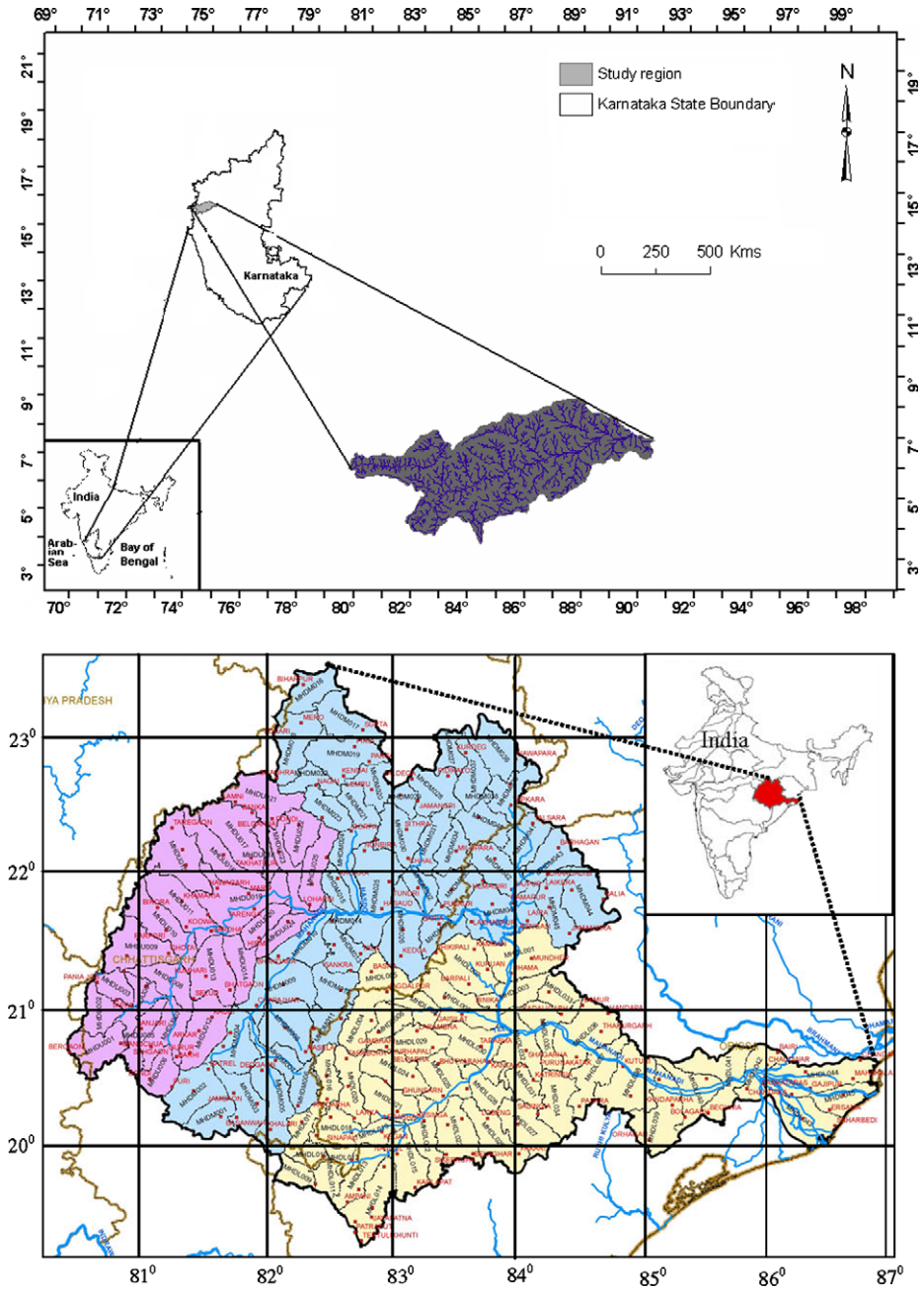

Fig. 1. (a) Location map of the Malaprabha basin. The latitude, longitude and scale of the map refer to the Karnataka state. (b) Location map of the Mahanadi basin.

should still be close neighbours. Hegger and Kantz [8] had modified the original algorithm of Kennel et al. [15] to avoid any spurious results due to noise. In the present study, this modified algorithm in which the fraction of false nearest neighbours are computed in a probabilistic way has been used.

The basic idea is to search for all the data points which are neighbours in a particular embedding dimension m and which do not remain so, upon increasing the embedding dimension to $m + 1$. Considering a particular data point, determine its nearest neighbour in the mth dimension. Compute the ratio of the distances between these two points in the $(m + 1)$th and mth dimensions. If this ratio is larger than a particular threshold f, then the neighbour is false. When the percentage of false nearest neighbours falls to zero (or a minimum value), the corresponding embedding dimension is considered high enough to represent the dynamics of the series.

According to Hegger and Kantz [8], for the fraction of false nearest neighbours to touch zero, the threshold value mentioned above should be sufficiently large. They suggested a minimal reasonable threshold determined from the local deterministic expansion rate as $e^{\lambda_{\max}\tau}$, where $\lambda_{\max}$ is the maximal Lyapunov exponent (explained in the next section) and τ is the time lag. A distortion in FNN fraction can also be caused by too low or too large time lag. Hence, the fraction of false nearest neighbours depends upon the threshold and also the time lag.

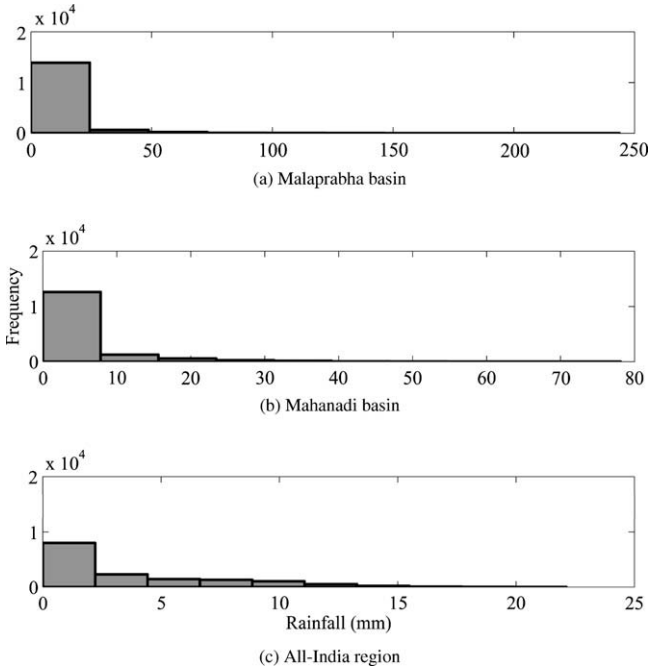

Fig. 2. Daily rainfall histograms for the period 1955–2000 of (a) Malaprabha basin; (b) Mahanadi basin and (c) All-India region.

2.3. Lyapunov exponent

The most striking feature of a chaos system is the unpredictability due to the sensitive dependence on initial conditions. The very small deviations in initial conditions of all the trajectories are blown up after a few time steps. In the case of chaotic systems, this divergence will be exponentially fast. Lyapunov exponent gives the averaged information of this divergence and thus the unpredictability of the system. It characterizes the rate of separation of infinitesimally close trajectories. Let s_{t1} and s_{t2} be two points in two trajectories in state space such that the distance between them is $\|s_{t1} - s_{t2}\| = \partial_0 \ll 1$. After Δt time steps ahead, the distance between these two trajectories will be $\partial_{\Delta t} \cong \|s_{t1+\Delta t} - s_{t2+\Delta t}\|$, $\partial_{\Delta t} \ll 1,\ \Delta t \gg 1$. Hence, trajectories with initial separation ∂_0 diverge in the form of an exponential function, $\partial_{\Delta t} \cong e^{\lambda \Delta t}\partial_0$, where λ is the Lyapunov exponent [13]. Since the rate of separation is different for various orientations of initial separation vector, the total number of Lyapunov exponents is equal to the number of dimensions of the phase space defined, i.e., a spectrum of exponents will be available. Among them, the highest (global) Lyapunov exponent need only be considered, as it determines the total predictability of the system.

A positive λ indicates an exponential divergence of the nearby trajectories, and thus chaos. The orbit is unstable and chaotic. Negative Lyapunov exponents are characteristic of dissipative or non-conservative systems. Their orbits attract to a stable fixed point or periodic orbit. The stability is directly proportional to the negativeness of the exponent. Conservative systems exhibit a zero Lyapunov exponent. The orbit is a neutral fixed point.

Since a positive Lyapunov exponent is a strong signature of chaos, many algorithms have been developed to calculate the maximal Lyapunov exponent. Wolf's algorithm [49] is one of the first kinds developed for this, although it requires much care as the algorithm does not allow testing the presence of exponential divergence and can lead to wrong results. However, exponential divergence can be examined using algorithms introduced by Rosenstein et al. [28] and Kantz [13].

For calculating the maximum Lyapunov exponent, one has to compute

$$S(\Delta t) = \frac{1}{N}\sum_{t_o=1}^{N} \ln\left(\frac{1}{|U(s_{t_0})|}\sum_{s_t \in U(s_{t_0})} |s_{t_0+\Delta t} - s_{t+\Delta t}|\right), \tag{3}$$

where s_{t_0} are reference points or embedding vectors, $U(s_{t_0})$ is the neighbourhood of s_{t_0} with diameter ξ. For a reasonable range of ξ

Table 1
Mean and standard deviation of monthly rainfall of three regions for the period 1955–2000.

Month	Malaprabha		Mahanadi		All-India	
	Mean (mm)	Standard deviation (mm)	Mean (mm)	Standard deviation (mm)	Mean (mm)	Standard deviation (mm)
January	1.0	3.9	13.6	13.7	22.9	9.3
February	0.8	2.1	15.1	14.2	26.1	10.8
March	6.0	11.6	15.5	19.3	33.8	13.8
April	29.0	27.1	13.2	8.6	43.5	11.8
May	83.7	78.0	21.0	18.3	74.6	14.5
June	417.4	168.5	193.8	87.3	183.2	29.4
July	770.0	299.8	398.2	86.1	311.0	35.0
August	441.3	195.4	390.0	81.6	271.3	32.1
September	167.1	74.1	215.7	83.1	180.9	36.0
October	138.4	93.4	59.8	47.5	84.5	32.1
November	39.5	49.1	9.5	14.0	33.7	16.0
December	5.2	11.4	7.8	14.2	19.5	10.7

and for all embedding dimensions m which is larger than some minimum dimension m_0, if $S(\Delta t)$ exhibits a linear increase, then its slope can be taken as an estimate of the maximal Lyapunov exponent λ.

2.4. Nonlinear prediction method

The nonlinear prediction method is used to investigate the presence of chaos in the time series, in addition to obtain forecasts. It is also used to counter-check the correlation dimension obtained from the correlation integral method.

The procedure of nonlinear prediction can be explained as follows: As a first step, the phase space reconstruction of the scalar series X_i, where $i = 1,2,\ldots,N$ is done, using the method of delays as per Eq. (1). Once the reconstruction of the attractor is successfully achieved in an embedding dimension m, the dynamics can be interpreted in the form of an m-dimensional map f_T such that

$$Y_{j+T} = f_T(Y_j), \tag{4}$$

where Y_j and Y_{j+T} are vectors of dimension m, Y_j being the state at current time j and Y_{j+T} being the state at future time $j + T$. Now the problem is to find a good approximation of f_T using the current data.

The selection of a nonlinear model for f_T can be made either globally or locally. The global approach approximates the map by working on the entire phase space of the attractor and seeking a form, valid for all points. Neural networks and radial basis functions adopt the global approach. In the second approach which works on local approximation [4] the dynamics are modeled locally piecewise in the embedding space. The domain is broken up into many local neighbourhoods and modeling is done for each neighbourhood separately, i.e., there will be a separate f_T valid for each neighbourhood. The complexity in modeling f_T is thus considerably reduced without affecting the accuracy of prediction.

The prediction of Y_{j+T} is done based on values of Y_j and k nearest neighbours of Y_j. These k nearest neighbours are selected based on the minimum values of $\|Y_j - Y_{j'}\|$ where $j' < j$. If only one nearest neighbour is considered then Y_{j+T} will be $Y_{j'+T}$. Since normally $k > 1$, the prediction of Y_{j+T} is taken as a weighted average of the k values. In the present study, the prediction of Y_{j+T} is done by averaging for k neighbours in the form $\widehat{Y}_{j+T} = \frac{1}{k}\sum_{i=1}^{k} Y_{i'+T}$. The optimum number of nearest neighbours is decided by trial and error. The prediction accuracy is estimated using the correlation coefficient,

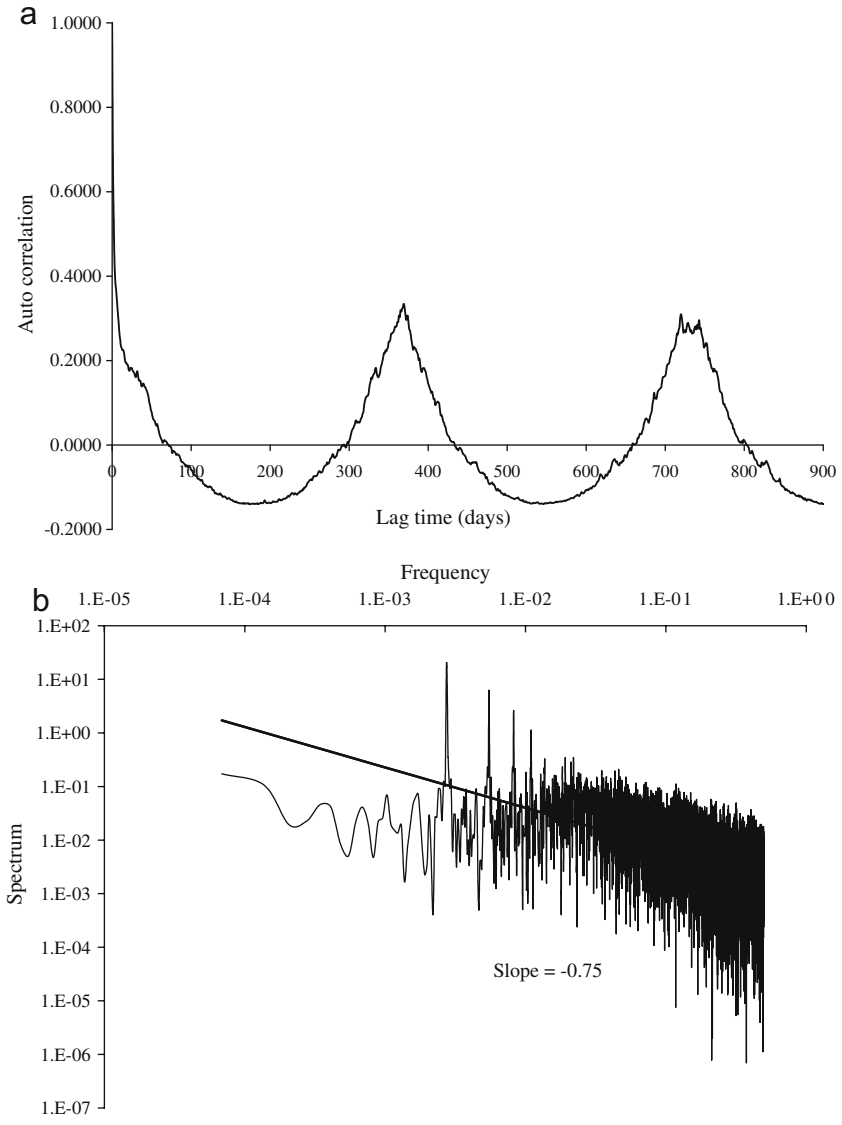

Fig. 3. (a) Auto-correlation function of Malaprabha daily rainfall; (b) power spectrum of Malaprabha daily rainfall; (c) autocorrelation function of Mahanadi daily rainfall; (d) power spectrum of Mahanadi daily rainfall; (e) autocorrelation function of All-India daily rainfall and (f) power spectrum of All-India daily rainfall.

Nash efficiency coefficient and also normalized mean square error between the predicted series and the corresponding observed series.

2.5. Surrogate data method

The method of surrogate data [45] generates substitute data with the same probabilistic structure as of the original data. The surrogate data is generated according to a null hypothesis (for example the data has been created by a stationary Gaussian linear process); but possess some of the statistical properties of the original data, such as the mean, the standard deviation, the cumulative distribution function, the power spectrum, etc. In Amplitude Adjusted Fourier Transform algorithm (AAFT) introduced by Theiler et al. [45], an ensemble of synthetic sequences are generated stochastically retaining the probability density function (pdf) and the linear correlation structure (power spectrum) of the original series. However the surrogates generated from a linear Gaussian function have a Gaussian pdf. Since very few real time series have a Gaussian structure, the original pdf is reconstructed by using an invertible nonlinear transform from the ensemble of surrogates. Finally, the original series is compared with the ensemble of the surrogate under the null hypothesis that the data has been generated by a stationary Gaussian linear stochastic process (equivalently, an *autoregressive moving average* or ARMA process) that is observed through an invertible, static, but possible nonlinear observation function represented as: $s_n = s(x_n)$, $\{x_n\}$: $ARMA(M,N)$.

Here without modeling the parameters (the orders M and N, the *ARMA* coefficients and the function $s(.)$) a prior, it is known that the above process would show characteristic linear correlations (reflecting the ARMA structure) and a characteristic single time probability distribution (reflecting the action of $s(.)$ on the original Gaussian distribution) [33].

In the present study, the iterative amplitude adjusted Fourier transform (IAAFT) method proposed by Schreiber and Schmitz [32] is used, in which the probability density function and correlation structure (and hence power spectrum) of the original data are maintained by iteratively minimizing the deviation. The algorithm proceeds as follows:

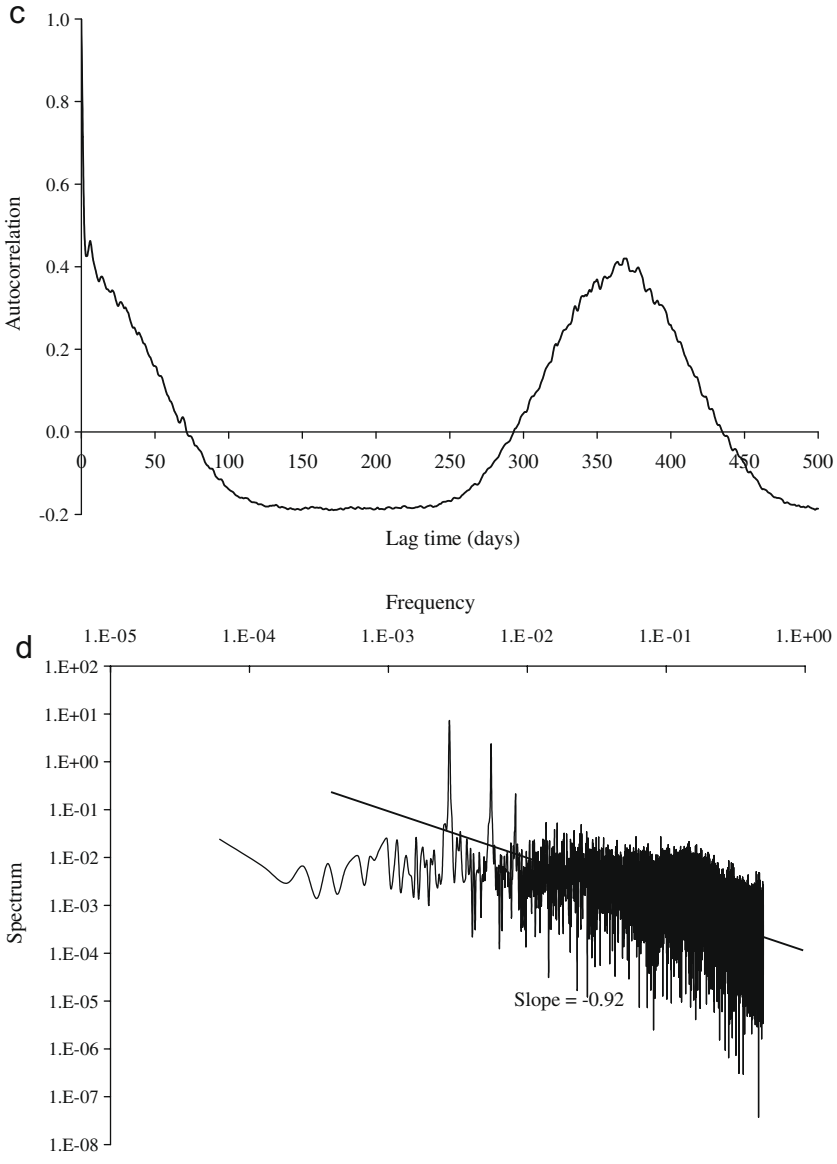

Fig. 3 (*continued*)

1. A sorted list of the original series $\{s_n\}$ and the squared amplitudes of its Fourier transform, $S_k^2 = \left|\sum_{n=0}^{N-1} s_n e^{i2\pi kn/N}\right|^2$ are stored.
2. Random shuffle the data (without replacement) $\left\{s_n^{(0)}\right\}$ to destroy any nonlinear relationships and correlations.
3. Now take the Fourier transform of $\left\{s_n^{(i)}\right\}$, replace its squared amplitude by $\left\{S_k^2\right\}$ and transform back.
4. To correct the pdf, rank order the resulting series and replace each value with the original series value with the same rank. This will modify the power spectrum again.
5. Repeat steps 3 and 4 until a given accuracy is reached.

A hypothesis testing is done by comparing a test statistic of the original data with those of the ensemble of surrogates. Any nonlinearity measure such as correlation dimension, the Lyapunov exponent, the Kolmogorov entropy, the prediction accuracy, etc. can be used as a test statistic. If the test statistic values, obtained from the ensemble of surrogate data, are significantly different from that of the original time series, then the null hypothesis that the original time series emanated from a linear process is rejected. On the other hand, if the test statistics from the surrogates and the original series are not significantly different, then the original time series is considered to be a linear stochastic process. In the present study, the nonlinear prediction error is used as the test statistic.

3. Methodology

The approach used for generating the ensemble prediction is described below:

i. Choose a suitable range of values of delay time τ using the autocorrelation method and mutual information method.
ii. Analyse the chaotic nature of the time series using correlation dimension, false nearest neighbour, Lyapunov exponent and nonlinear prediction methods.
iii. Repeat the correlation dimension method on phase randomized and also on first derivative of the original signal, for examining the presence of any pseudo-low dimensional chaos.
iv. Compute a suitable range of values of embedding dimension m using the above methods.

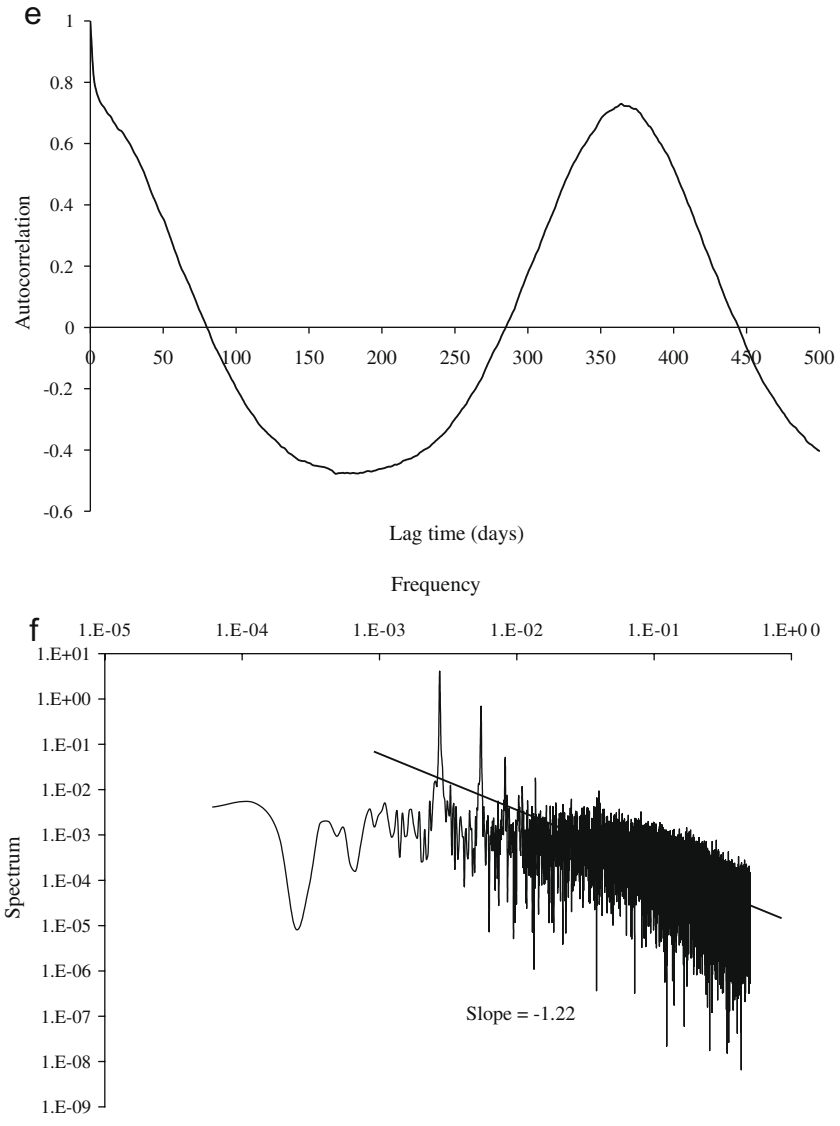

Fig. 3 (*continued*)

v. Select a range of neighbourhood size (radius) from which the nearest neighbours searching can be done, using nonlinear prediction method. The neighbourhood radius is expressed as a fraction α of the standard deviation.
vi. Examine the nonlinearity of the time series using the surrogate data method.
vii. Reconstruct the phase space for all the available combinations of the parameters m, τ and α.
viii. Compute the generalized cross validation (GCV) value for all the possible combinations.

$$\mathrm{GCV}(m,\tau,\alpha) = \frac{\sum_{i=1}^{n} \frac{e_i^2}{n}}{\left(1-\frac{p}{n}\right)^2}, \tag{5}$$

where e_i is the error, n is the number of data points, p is the number of parameters to be determined.
ix. Select a set of best parameter combinations falling under 10% of the lowest GCV value.
x. The best parameter combinations are used to produce an ensemble of forecasts. The quality of the ensembles is analysed using two performance measures: rank probability skill score (RPSS) and rank histogram [48].
xi. Repeat all these above steps on different time series with different characteristics in order to analyse the change in the chaotic behaviour due to the change in characteristics.

For determining RPSS, the dataset is divided into n number of categories. The rank probability score (RPS) is calculated as the sum of the squares of the difference of the cumulative probabilities of each of the predicted – observed data pair. RPS is given by

$$\mathrm{RPS} = \sum_{i=1}^{n} (P_i - O_i)^2, \tag{6}$$

where P_i is the cumulative probability of the forecast for category i and O_i is the cumulative probability of the observation for category i. Cumulative probability of the forecast for each category is based on adding all the ensemble values of that category. The cumulative probability of the observation is determined by assigning a value of zero for all categories less than the observation's category and a value of 1 for all categories equal to and greater than the observation's category. Finally, RPSS is given by

$$\mathrm{RPSS} = 1 - \frac{\overline{\mathrm{RPS}}}{\overline{\mathrm{RPS}}_{c\,\mathrm{lim}}}, \tag{7}$$

where $\overline{\mathrm{RPS}}$ is the mean rank probability score of all observation – forecast pairs and $\overline{\mathrm{RPS}}_{c\,\mathrm{lim}}$ is the mean rank probability score of climatological forecast. An RPSS value of 1.0 indicates a perfect forecast and a negative value indicates an output worse than climatology. An RPSS of 0.0 implies no improvement in skill over the reference climatological forecast, $\mathrm{RPS}_{c\,\mathrm{lim}}$.

Rank histogram is a graphical method to evaluate the reliability and probable predictability of the targeted parameter by the ensembles. Suppose there are n observation forecast pairs and n_{ens} ensemble forecasts corresponding to each observation. Then, assuming that for each of these n data sets, all the ensembles and also the observations are having the same probability distribution, the rank of the observation is likely to take any of the values $i = 1,2,3,\ldots,n_{ens}+1$. The rank of the observation is determined for each of the n data points. These ranks are plotted in the form of a histogram to produce the rank histogram. While an ideal rank histogram should be a flat one, ensemble members from a less variable distribution results in a U-shaped rank histogram. A U-shaped histogram indicates that the spread is too small that many observations are falling outside the extremes of the ensembles; whereas a dome shape indicates that ensemble spread is too large that too many observations are falling in the middle range. An ensemble bias (positive or negative) excessively populates the low and high ranks.

4. Data used

The daily rainfall data of three regions in India: Malaprabha basin, Mahanadi basin and also All-India for the period 1955–2000 are considered for the present study. The location map of the Malaprabha and Mahanadi basins are shown in Fig. 1. The regions chosen vary widely in the spatial area coverage and also in the rainfall intensity. Malaprabha basin is the smallest with only around

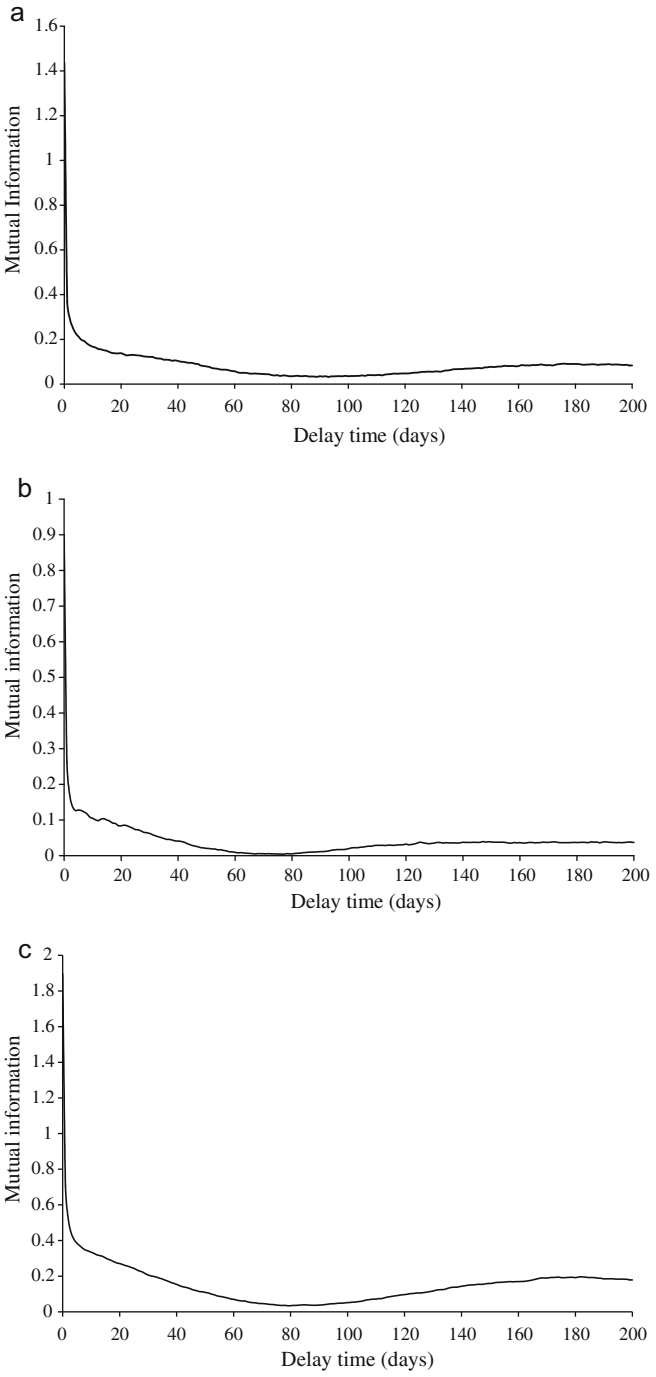

Fig. 4. Variation of mutual information with lag time: (a) Malaprabha daily rainfall; (b) Mahanadi daily rainfall and (c) All-India daily rainfall.

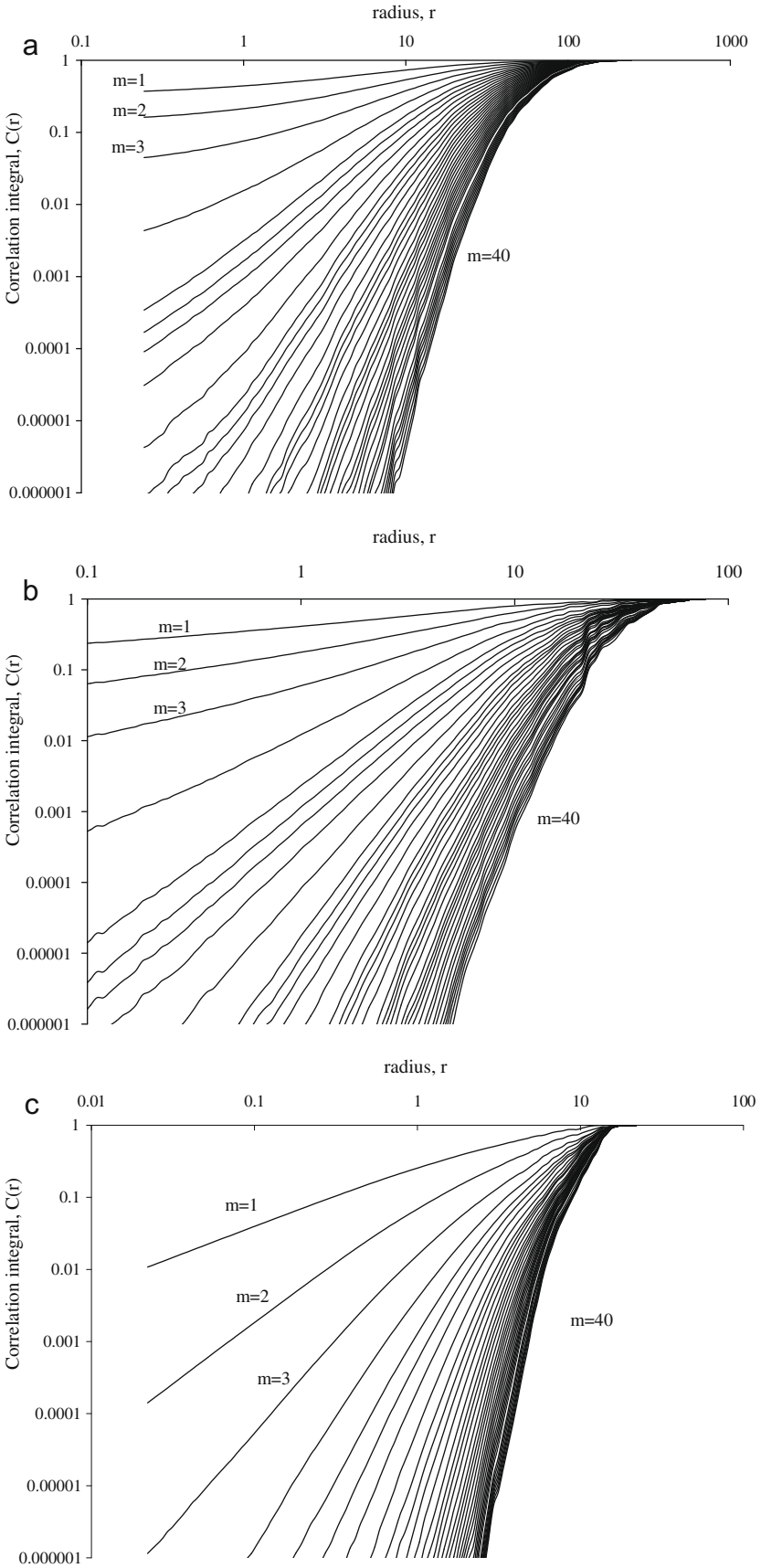

Fig. 5. Variation of correlation integral with radius on a log–log scale for embedding dimensions from 1 to 40: (a) Malaprabha daily rainfall; (b) Mahanadi daily rainfall and (c) All-India daily rainfall.

2500 km^2 area; yet receiving the highest rainfall of 1800 mm in the monsoon months. While Mahanadi and All-India rainfall differ only 200 mm in monsoon months (with Mahanadi receiving 1200 mm and All-India receiving 945 mm), their areas vary widely, 1.4×10^5 km^2 and 32.8×10^5 km^2, respectively. While Malaprabha and Mahanadi daily rainfall series have 58% and 38% zeroes, All-India has only 2% since it is averaged over a large area.

The frequency histograms of the daily rainfall series for the period 1955–2000 are shown in Fig. 2. The mean and standard deviation of monthly rainfall of the regions are presented in Table 1. Major portion of the annual rainfall is received in the monsoon months of June, July, August and September.

5. Results and discussion

The daily rainfall data from 1955 to 1995 is used for chaotic nature analysis and for determining the embedding dimension and delay time. As a preliminary investigation, the autocorrelation function and Fourier spectrum of the three time series are plotted and are shown in Fig. 3. The initial exponential decay of autocorrelation functions indicates that the rainfall series may be of chaotic nature. The periodic behaviour of the autocorrelation function for higher lags is due to the seasonal periodicity of the rainfall. Also, the broad band form of the power spectrum and its power law

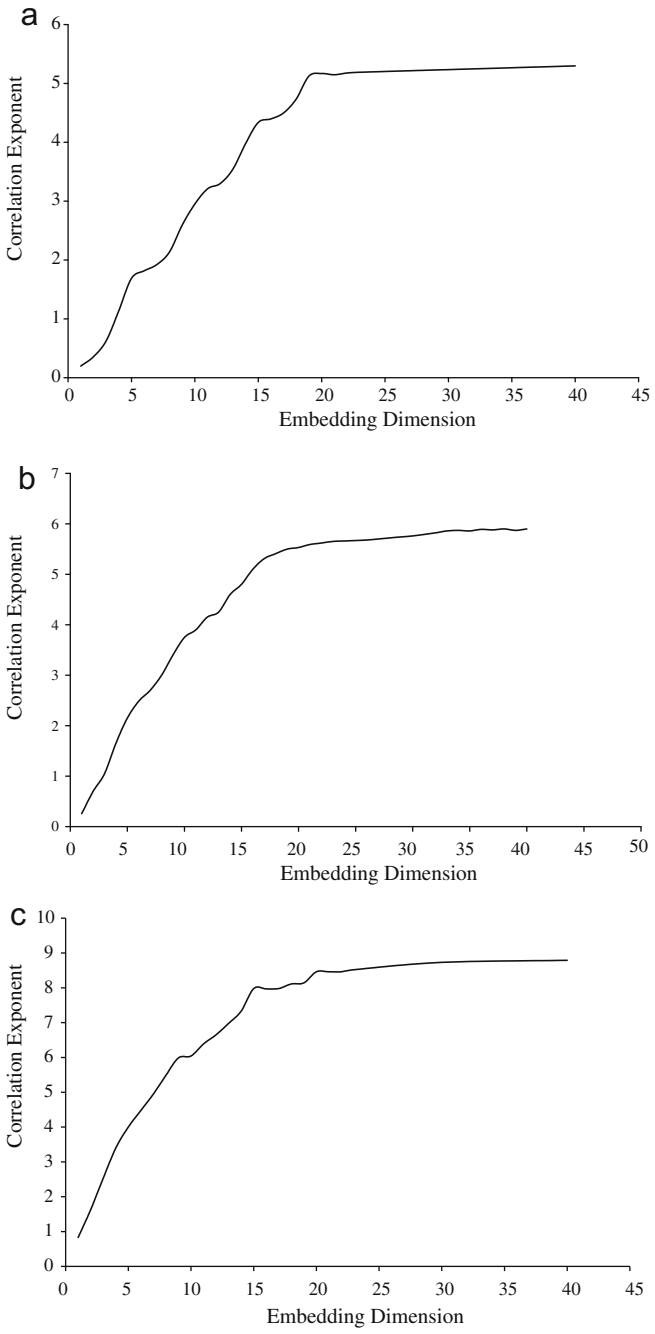

Fig. 6. Variation of correlation exponent with embedding dimension: (a) Malaprabha daily rainfall; (b) Mahanadi daily rainfall and (c) All-India daily rainfall.

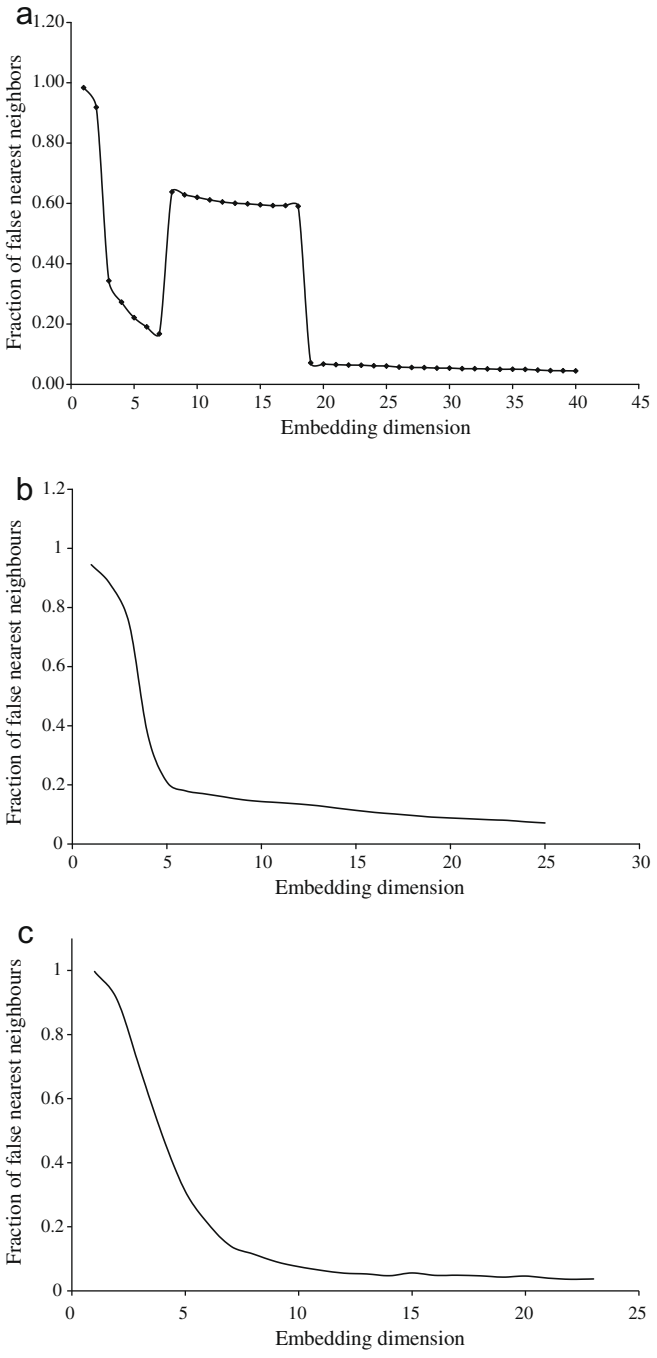

Fig. 7. Variation of fraction of false nearest neighbours with embedding dimension: (a) Malaprabha daily rainfall; (b) Mahanadi daily rainfall and (c) All-India daily rainfall.

shape, i.e, $P(f) \propto f^{-\alpha}$, with $\alpha \approx 1.0$ is clearly visible for a large frequency range for all the three time series.

5.1. Determination of delay time

The choice of the delay time τ is made using the autocorrelation method and the mutual information method. In autocorrelation method, the delay time is determined as the lag time at which the autocorrelation function attains a zero value. Hence, the delay times for the three series from the autocorrelation plot are 71, 72 and 81 days respectively. The mutual information obtained for various lag times are shown in Fig. 4. The delay time for the phase space reconstruction is the first minimum value, which is at 93, 74 and 84 days. Hence, the ranges of delay time for the ensemble prediction are chosen as 60–100 days, allowing a little extra spread, for all the three time series.

5.2. Determination of embedding dimension

5.2.1. Correlation dimension method

The correlation integral $C(r)$ according to Grassberger–Procaccia algorithm is calculated for embedding dimensions 1–40. Fig. 5 shows a plot of correlation integral $C(r)$ versus radius r on a log–log scale for embedding dimension $m = 1$–40. In this figure, clear scaling regions are visible between $C(r)$ values of 10^{-2} and 10^{-5}, for the calculation of correlation exponents. The correlation exponent is determined by the slope of the plot of $C(r)$ versus r on a log–log scale. Fig. 6 shows the variation of the correlation exponent with the embedding dimension for the three regions. It can be noticed that for all the three regions, the correlation exponent is increasing with embedding dimension and reaching a constant value at embedding dimension $m \geqslant 19$.

The saturation of the correlation exponent beyond a certain embedding dimension is an indication of the existence of chaos in rainfall series. However, the saturation value is slightly different for different regions. The saturation values of the correlation exponent for Malaprabha and Mahanadi basins are found to be 5.12 and 5.79, respectively, which means that the number of variables dominantly influencing the rainfall dynamics of these basins is ≈6. The correlation exponent of All-India region is slightly higher, with a value of 8.14, necessitating the total number of influencing variables ≈9. The high correlation dimension for All-India daily rainfall

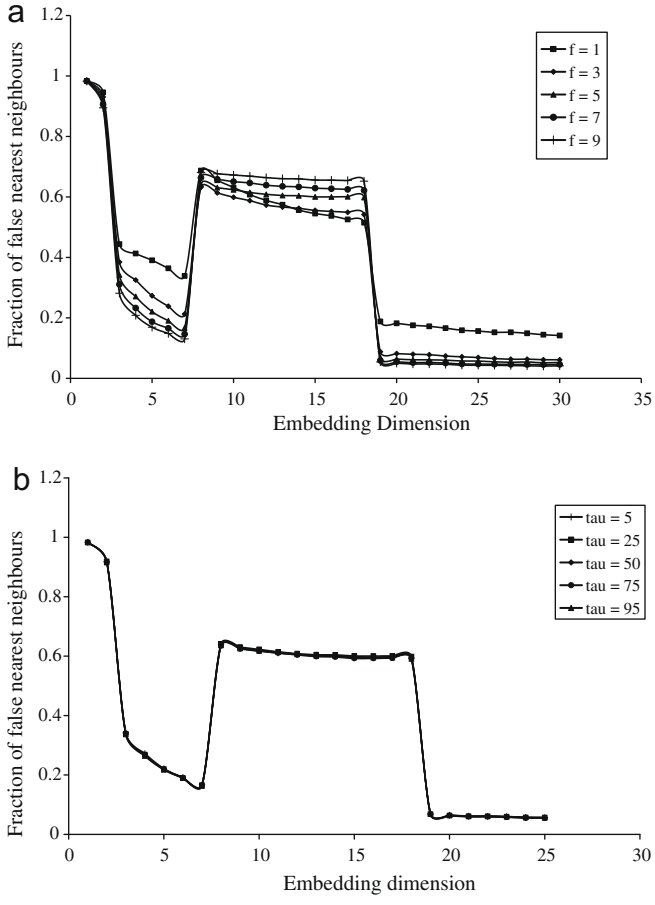

Fig. 8. (a) Sensitivity analysis of fraction of false nearest neighbours for different threshold values and (b) sensitivity analysis of fraction of false nearest neighbours for different delay times.

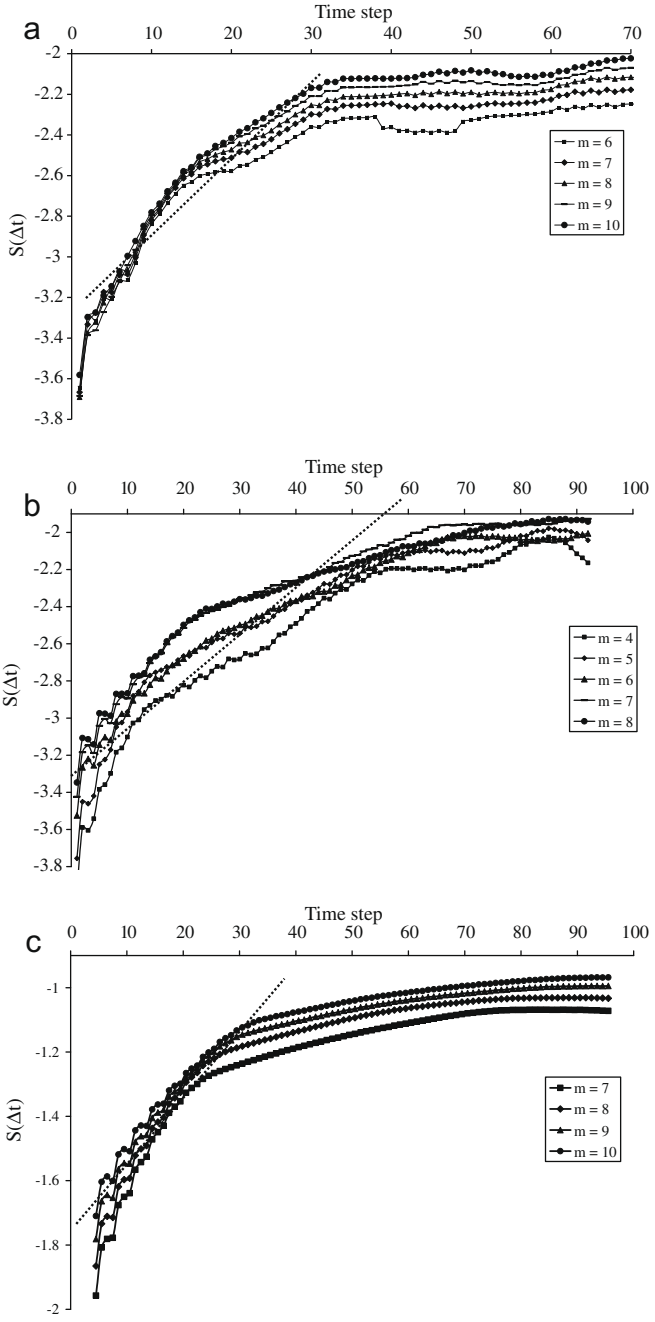

Fig. 9. Variation of $S(\Delta t)$ with time for various embedding dimensions: (a) Malaprabha daily rainfall; (b) Mahanadi daily rainfall and (c) All-India daily rainfall.

can be attributed to the large spatial area contributing to its rainfall and hence the requirement of more variables to explain its dynamics. Nevertheless, the low correlation dimensions obtained in three cases suggest the possible presence of low dimensional chaotic behavior.

5.2.2. False nearest neighbour method

The FNN algorithm is applied on the rainfall series of three regions. The threshold value *f* is fixed at 5. The variation of the fraction of false nearest neighbours for different embedding dimensions is shown in Fig. 7. It can be seen that for Malaprabha basin, the fraction of nearest neighbours is falling to a minimum value at an embedding dimension of 7. This indicates that an embedding dimension of 7 is sufficient to explain the dynamics

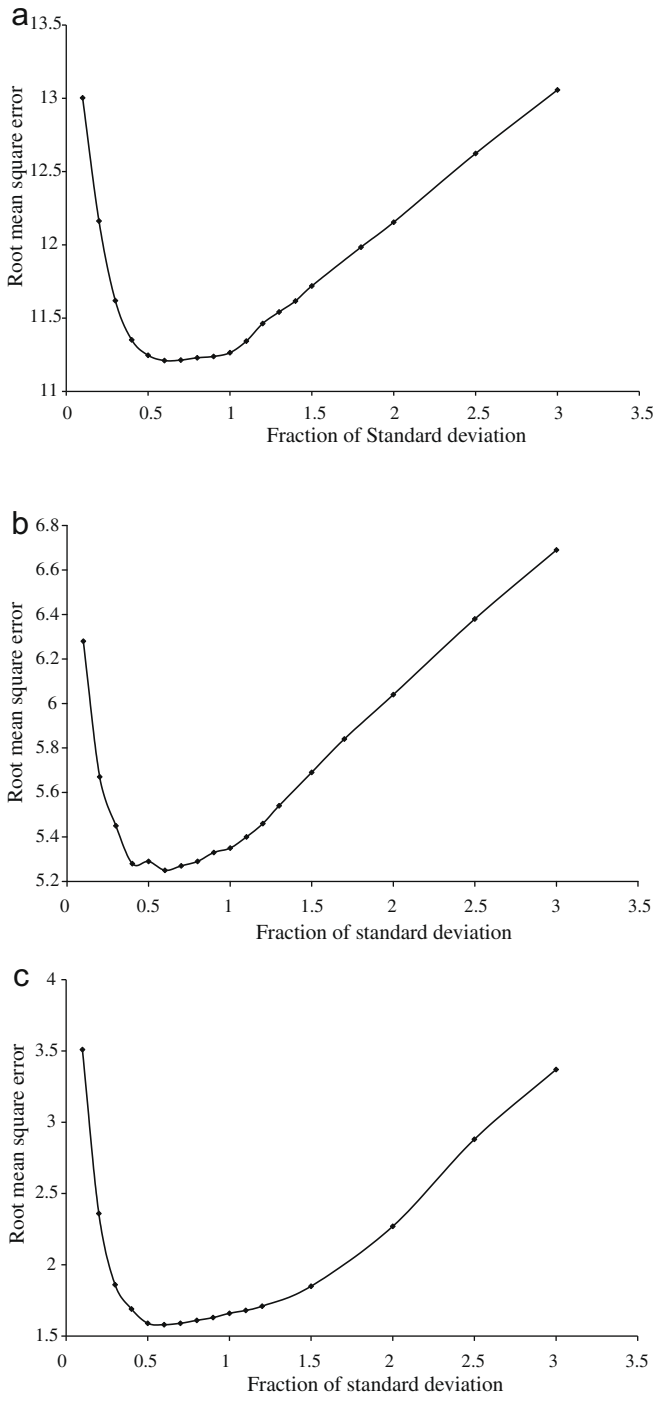

Fig. 10. Variation of prediction error with the neighbourhood size: (a) Malaprabha daily rainfall; (b) Mahanadi daily rainfall and (c) All-India daily rainfall.

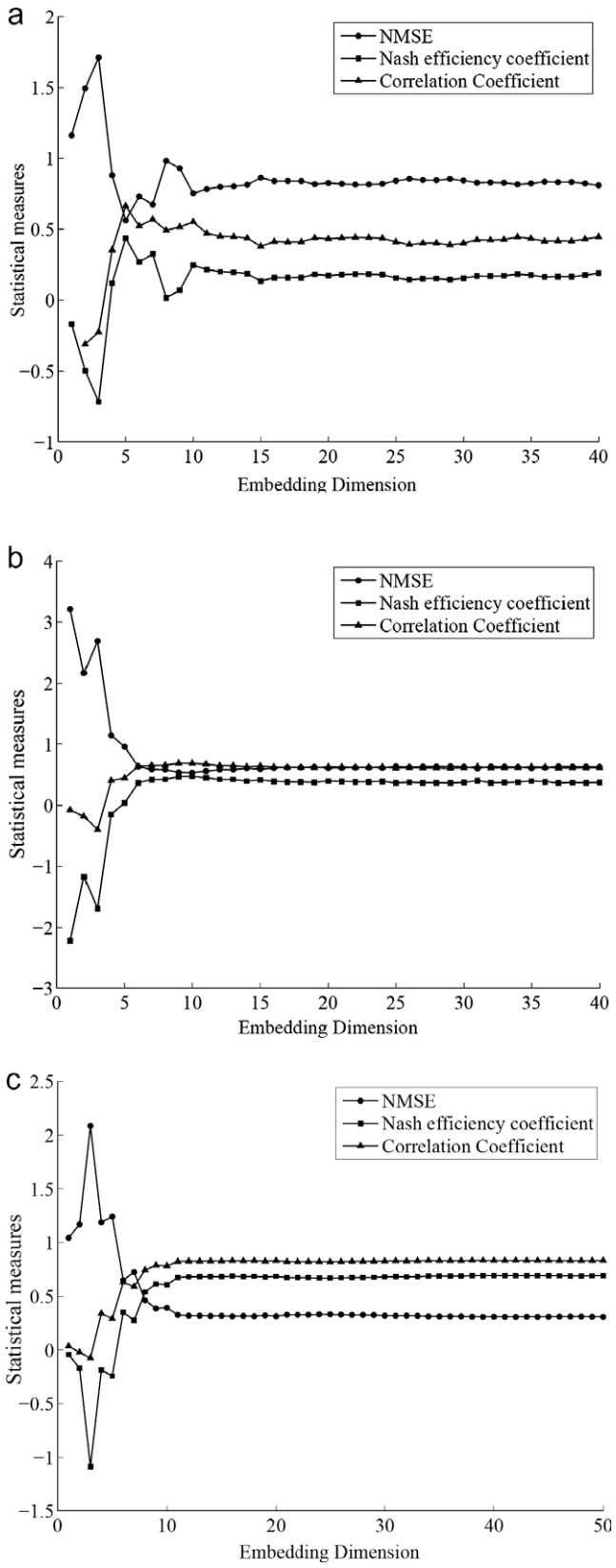

Fig. 11. Variation of statistical measures with embedding dimension: (a) Malaprabha daily rainfall; (b) Mahanadi daily rainfall and (c) All-India daily rainfall.

of the rainfall series. However, as can be seen from the figure, the percentage of nearest neighbours is not attaining a zero or minimum value as desired. It is steeply increasing after 7th dimension and thereafter remains almost constant up to $m = 18$. At $m = 19$, it again falls to a minimum value and thereafter it remains almost constant. Despite this unexpected behaviour, it can be concluded that the embedding dimension is ≈ 7 and it is in close agreement with the value obtained by the correlation dimension method.

For Mahanadi basin and All-India region, the FNN fractions are falling steeply till embedding dimensions of 6 and 8, respectively. The decrease of FNN fraction afterwards is comparatively insignificant. The values obtained in both cases are in close agreement with that obtained from correlation dimension method.

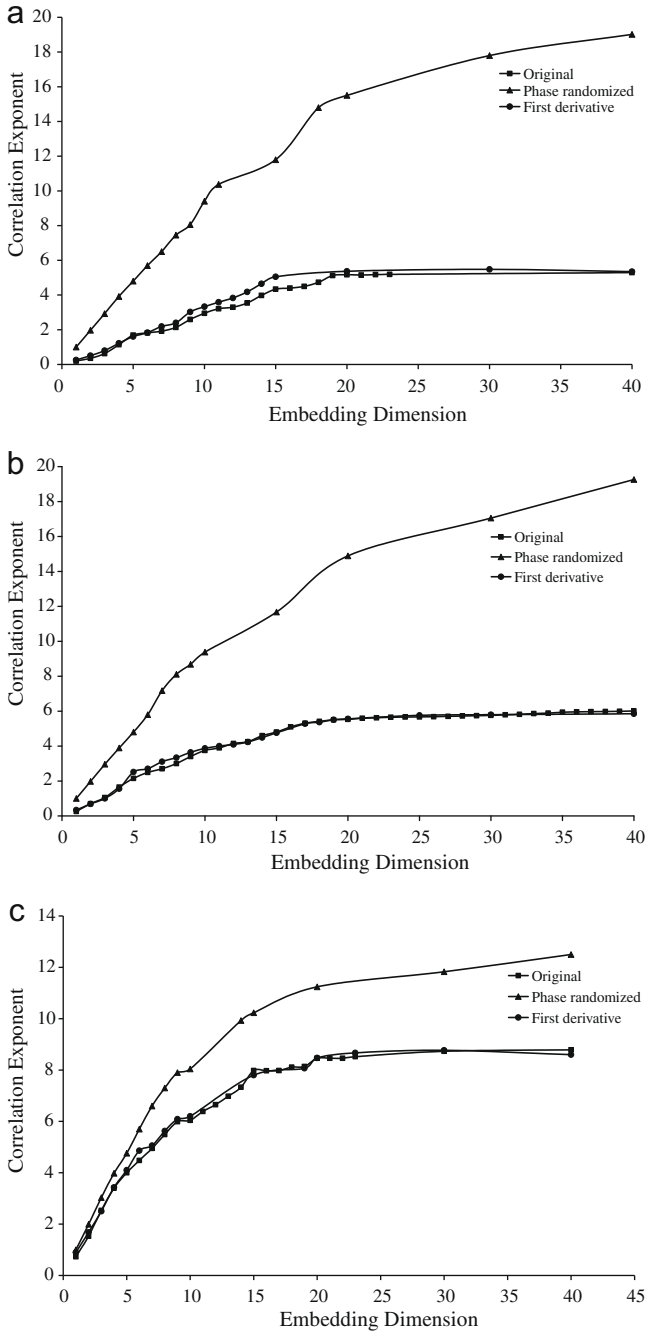

Fig. 12. Variation of correlation exponent with embedding dimension for original data, phase randomized data and first derivative of data: (a) Malaprabha daily rainfall; (b) Mahanadi daily rainfall and (c) All-India daily rainfall.

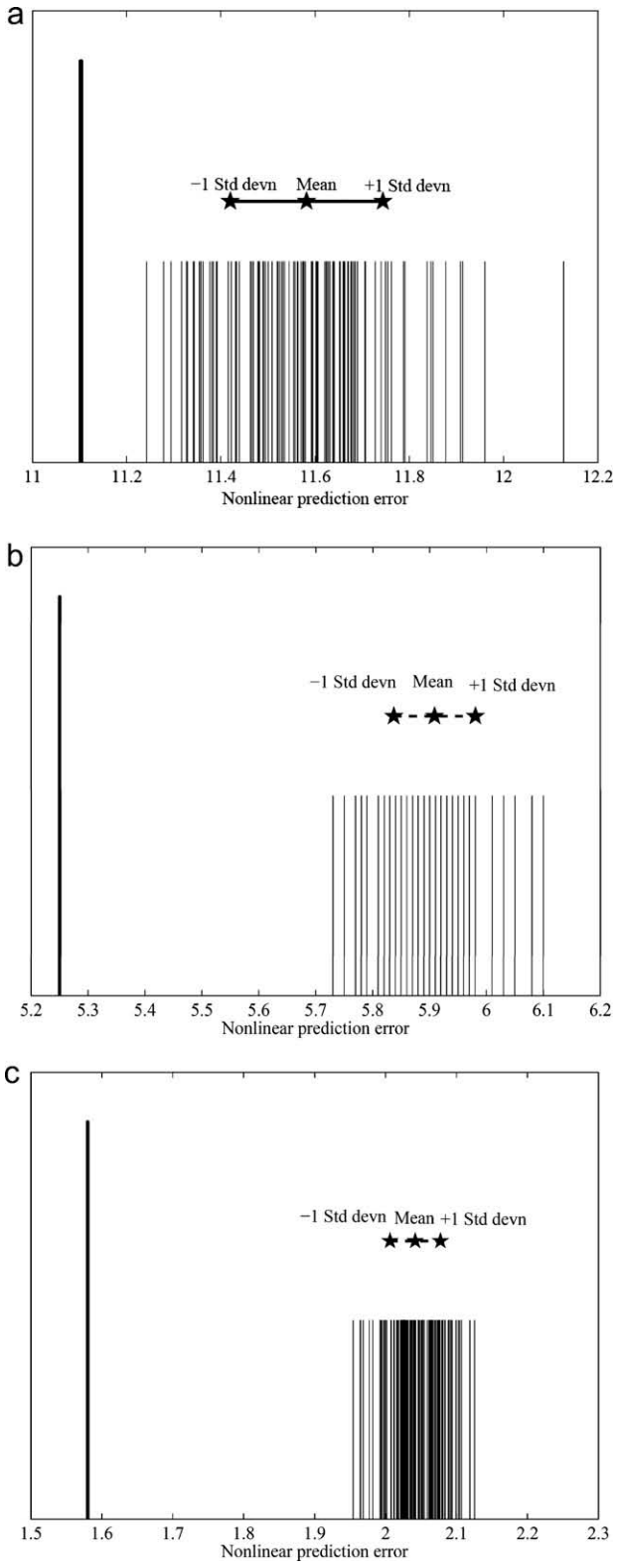

Fig. 13. Nonlinear prediction error for original data and 99 surrogates. The thick long line indicates the nonlinear prediction error of original data and thin short lines indicate the error from the surrogates. The mean and ±1 standard deviation of the prediction errors of surrogates are also shown. (a) Malaprabha daily rainfall; (b) Mahanadi daily rainfall and (c) All-India daily rainfall.

The unusual rise of FNN fraction at embedding dimension 7 and a plateau till embedding dimension 18 for Malaprabha basin may be due to the presence of additive noise in the data series. The presence of additive noise leads to high space dimensionality at smaller scales. Since the selection of a suitable noise reduction method needs further investigation, it is not dealt in the present study.

Also, a comparison of FNN diagrams of latter regions with the former one suggests that another possible reason for the unexpected behaviour in the former case may be due to the presence of a large amount of zeros (about 57%) in the time series. The percentage of zeros (single values) are comparatively less for the Mahanadi and All-India rainfall series.

The sensitivity of FNN fraction on threshold, f for Malaprabha basin is shown in Fig. 8(a). The unusual behaviour is seen for all the thresholds taken. Even though the fraction of FNN is lesser for larger thresholds, the dimension at which the FNN fraction drops considerably remains the same for all the thresholds. Hence, it can be concluded that for the mere estimation of embedding dimension, the value of threshold taken does not matter. Also, the sensitivity of FNN on delay time ranging from 5 to 95 days is shown in Fig. 8(b). It is clear from the figure that contrary to Hegger and Kantz [8] the FNN fraction does not depend on delay time.

5.2.3. Lyapunov exponent

The maximal Lyapunov exponent is calculated employing the algorithm by Rosenstein et al. [28]. The variation of $S(\Delta t)$ with time, t for Malaprabha basin at dimensions $m = 6–10$ is shown in Fig. 9(a). The maximum Lyapunov exponent which is given by the slope of the linear part of the curve is around 0.0317. Similarly, from Fig. 9(b) and (c), the maximum Lyapunov exponent for Mahanadi and All-India regions are calculated as 0.025 and 0.020, respectively. Positive values of Lyapunov exponent for the three regions confirm the exponential divergence of trajectories and hence the chaotic nature of the daily rainfall. The inverse of the Lyapunov exponent defines the predictability of the system, which is around 31 days for Malaprabha, 40 days for Mahanadi and 50 days for All-India. It is significant to note here that an increase in spatial area leads to an increase in predictability of the system.

Table 2
Optimum parameter combinations with minimum GCV.

Parameter	Region		
	Malaprabha	Mahanadi	All-India
Dimension	4–7	4–7	7–11
Delay time (days)	65–85	65–85	75–90
Neighbourhood size (% standard deviation)	0.4–0.6	0.4–0.7	0.4–0.7

5.2.4. Nonlinear prediction method

Nonlinear prediction method is used here as an inverse method for identifying the chaotic nature and also for determining the embedding dimension necessary for revealing the underlying dynamics of the rainfall series. The rainfall series from 1955 to 1999 is used to predict the rainfall for the year 2000, using the local constant method. As a first step, the optimum neighbourhood sizes are determined for the three regions by plotting the variation of the prediction error (root mean square error, RMSE) with the neighbourhood size (which is a fraction of standard deviation) for an optimum embedding dimension (m obtained from correlation dimension method) and are shown in Fig. 10.

It can be seen that the prediction error decreases for a neighbourhood size of around 0.5 $\times$ standard deviation and thereafter it starts increasing with further increase in neighbourhood size. This is in agreement with the Casdagli's test for nonlinearity [2] which states that if the prediction accuracy increases up to a certain number of nearest neighbours and decreases for higher number of nearest neighbours, it shows the evidence of chaos in the data series. If the prediction accuracy is the maximum for a large number of nearest neighbours, then the process can be better explained through a stochastic process. The ranges of neighbourhood size for the ensemble prediction for the three regions are fixed as 0.3–1.3 of the standard deviation, since the RMSE at 0.3 and 1.3 standard deviations are almost equal, as can be noticed from Fig. 10.

Further, the prediction accuracy is measured in terms of normalized mean square error (NMSE), Nash efficiency coefficient and also correlation coefficient. The variation of these performance measures for various embedding dimensions using the optimum nearest neighbours for the three regions are shown in Fig. 11.

For a chaotic time series, the prediction efficiency is expected to increase to a value close to 1 with an increase in embedding dimension m up to an optimal m and remain constant afterwards.

Table 3
Correlations between the observed and mean ensemble daily rainfall values for predictions beginning from January and June.

Year	Malaprabha prediction beginning from		Mahanadi prediction beginning from		All-India prediction beginning from	
	January	June	January	June	January	June
1996	0.56	0.57	0.73	0.73	0.87	0.87
1997	0.73	0.74	0.655	0.66	0.84	0.84
1998	0.65	0.63	0.625	0.63	0.87	0.87
1999	0.71	0.69	0.68	0.68	0.86	0.865
2000	0.70	0.68	0.665	0.66	0.82	0.82

Table 4
Correlations between the observed and the mean ensemble average daily rainfall values for predictions beginning from January and June.

Year	Malaprabha prediction beginning from		Mahanadi prediction beginning from		All-India prediction beginning from	
	January	June	January	June	January	June
1996	0.89	0.90	0.98	0.98	0.99	0.99
1997	0.94	0.95	0.97	0.98	0.98	0.98
1998	0.96	0.97	0.96	0.965	0.99	0.99
1999	0.94	0.94	0.97	0.97	0.97	0.97
2000	0.98	0.98	0.95	0.965	0.97	0.975

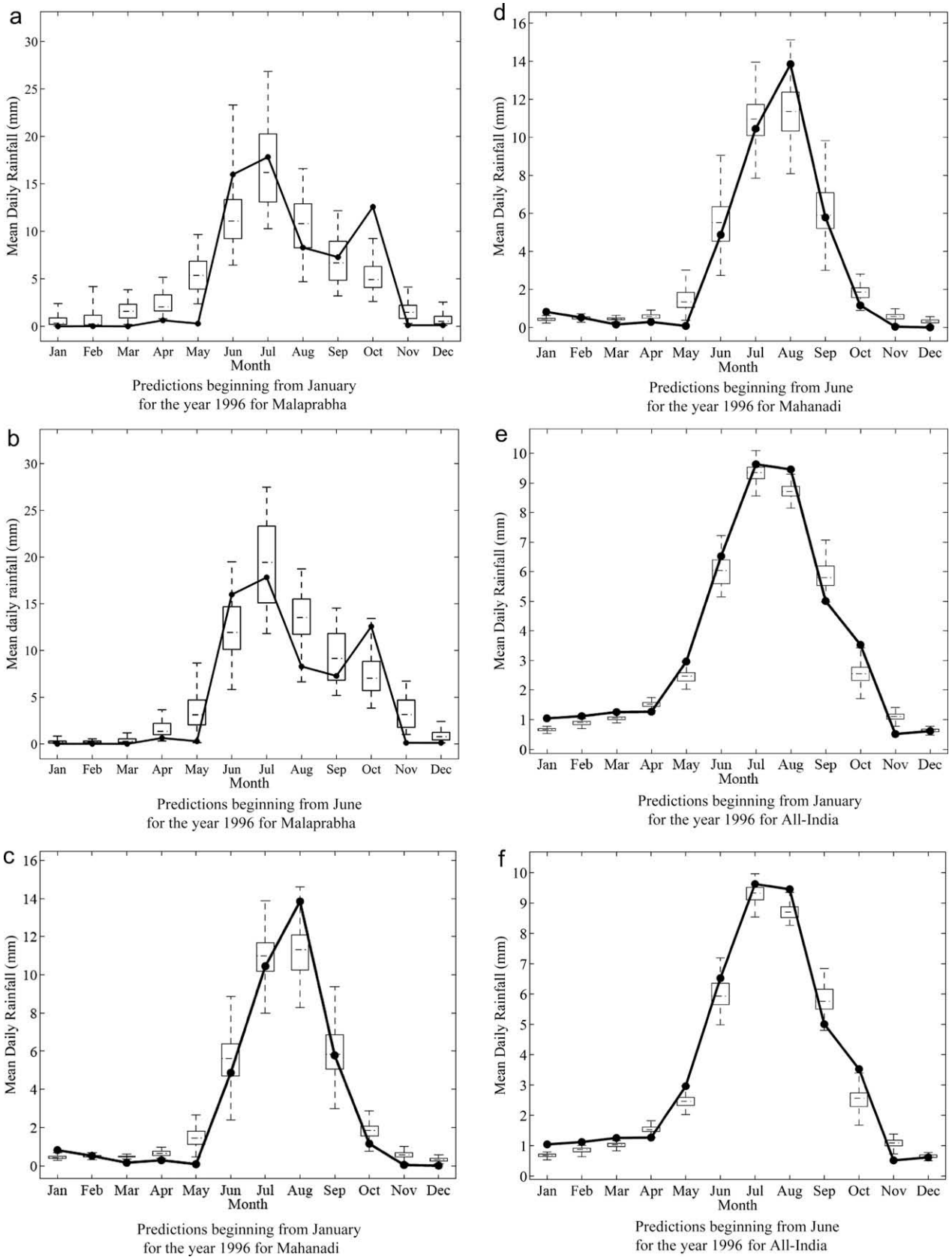

Fig. 14. Box plots of the mean daily rainfall values of the ensembles for the year 1996. The observed mean daily rainfall is shown as a continuous solid line. The median ensemble values are shown as a dashed line within each box. Predictions beginning from (a) January (b) June for Malaprabha region (c) January (d) June for Mahanadi region (e) January (f) June for All-India region the year 1996 are shown.

Table 5
Rank probability skill score values for each month over the period 1996–2000.

Month	Malaprabha prediction beginning from		Mahanadi prediction beginning from		All-India prediction beginning from	
	January	June	January	June	January	June
January	−0.96	−0.87	0.65	0.80	0.95	0.94
February	−5.59	−5.06	0.38	0.38	−0.53	−0.54
March	−0.94	−0.70	−0.59	−0.59	0.25	0.26
April	0.05	−0.21	0.03	0.02	0.43	0.48
May	0.32	0.44	0.02	0.19	0.44	0.49
June	0.58	0.64	0.86	0.87	0.36	0.36
July	0.51	0.64	0.83	0.88	0.96	0.97
August	0.44	0.50	0.63	0.66	0.97	0.96
September	0.45	0.44	0.68	0.67	0.34	0.32
October	0.16	0.27	0.35	0.36	0.04	0.03
November	−3.72	−4.03	−0.14	−0.12	−0.27	−0.30
December	−0.07	−0.12	−1.34	−1.34	−0.24	−0.24

On the other hand, for a stochastic time series, there would not be any increase in the prediction accuracy with an increase in the embedding dimension [1]. As can be seen from Fig. 11(a), the maximum prediction accuracy for Malaprabha region is for an embedding dimension of 5. Similarly, as inferred from Fig. 11(b) and (c), for Mahanadi and All-India regions, maximum prediction accuracy is for 6th and 8th dimensions, respectively. Also, the prediction accuracy remains a constant after attaining a maximum for all the three cases, which again supports the presence of chaos in the rainfall series. Hence, the optimum embedding dimensions from the nonlinear prediction method are 5, 6 and 8 for the three regions.

Considering the dimension values obtained by the various methods, the ranges of embedding dimension for ensemble prediction are fixed as 3–10 for Malaprabha and Mahanadi regions and 5–12 for All-India region.

5.3. Check for pseudo-low dimensional chaos

Since the power spectrums are showing a power law behaviour as shown in Fig. 3(b), (d) and (f), which could be the reason for the convergence of the correlation dimension as pointed by Osborne and Provenzale [19], it is recommended to carry out the correlation dimension on phase randomized data and also on first derivative of the original signals.

A comparison of the variations of correlation exponent with embedding dimension for the original data, phase randomized data and the first derivative of data of the three regions are shown in Fig. 12. For the Malaprabha and Mahanadi daily rainfall for which the spectral slopes are less than −1.0, the correlation dimensions of the phase randomized data sets are not converging, thus confirming the presence of a low dimensional strange attractor. But, in the case of All-India daily rainfall (Fig. 12(c)), even though correlation exponent is not converging for phase randomized data, the deviation from the original data is much smaller when compared with that of the former regions. Nevertheless, the non-convergence points out the chance for the further increase of correlation exponent and hence is a strong indication of low dimensional dynamics.

Likewise, in all the three cases, the variation of correlation exponent of first derivative adheres very well to that of the original data. The saturation values are the same as that obtained for the original rainfall. This eliminates the possibility of linear correlations forcing the saturation of correlation exponent and thereby confirms the presence of a low dimensional strange attractor in all the three rainfall series.

5.4. Nonlinearity test using surrogate data method

An ensemble of surrogates, assumed to be generated from a process of the form

$$s_n = S(x_n), \quad x_n = \sum_{i=1}^{M} a_i x_{n-i} + \sum_{i=0}^{N} b_i \eta_{n-i} \tag{8}$$

are generated, where S could be any invertible nonlinear function, $\{x_n\}$ is the underlying linear process, $\{a_i\}$ and $\{b_i\}$ are coefficient constants and $\{\eta_n\}$ is white Gaussian noise. M and N are the orders of an autoregressive (first term) and moving average (second term) model, respectively.

For testing the null hypothesis that the original data is also from a linear process of the form given by the above equation, at 1% significance level, a collection of surrogates are generated. The number of surrogates required for a one sided hypothesis test at 1% significance level is equal to $\frac{1}{\alpha} - 1 = 99$, where α, the significance level = 0.01. The nonlinear prediction error is used as the test statistic. Comparison of nonlinear prediction errors of 99 surrogates and of the original data is shown in Fig. 13. It can be seen that for all the three regions, the prediction error of observed data is much less than that of surrogates, thus rejecting the null hypothesis that the data comes from a linear stochastic process. This further confirms that the convergence of the correlation dimension is not due the linear stochastic nature; but is due to the low dimensional dynamics which is dominant in these systems.

5.5. Ensemble prediction

The aforementioned methods have confirmed that the daily rainfall series of three regions are nonlinear and low dimensional chaotic. The embedding dimensions and delay times obtained slightly vary for different methods. Hence, an ensemble of predic-

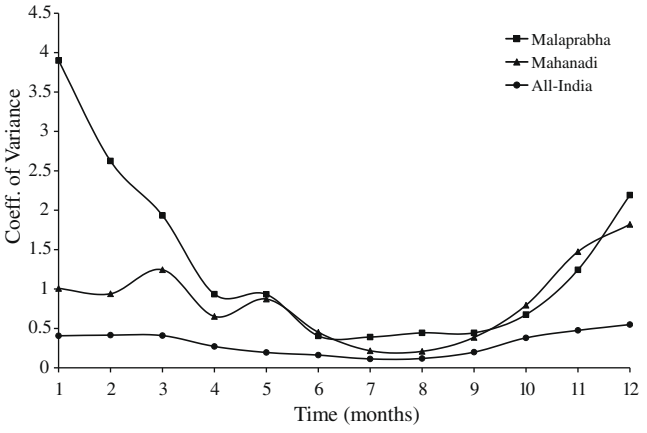

Fig. 15. Variation of coefficient of variance over the year for the period 1955–1995 for the three regions.

tions is produced from appropriate range of embedding dimension, delay time and neighbourhood size. The ranges of values used for the ensemble prediction are: (i) embedding dimension: 3 to 10 for Malaprabha and Mahanadi basins; 5–12 for All-India region; (ii) delay time: 60–100 for all the three regions; (iii) neighbourhood size: 0.3–1.3 of standard deviation for all the three regions.

As described in the methodology, the optimum parameter values are selected based on the minimum GCV value. It is found that for all the three cases, sufficient numbers of ensembles (about 150–250) are obtained even when the GCV threshold is set to 10%. Hence, the parameter combinations falling under 10% of the lowest GCV value are selected as the optimum ones. In order to generate a constant number of ensembles for prediction, the num-

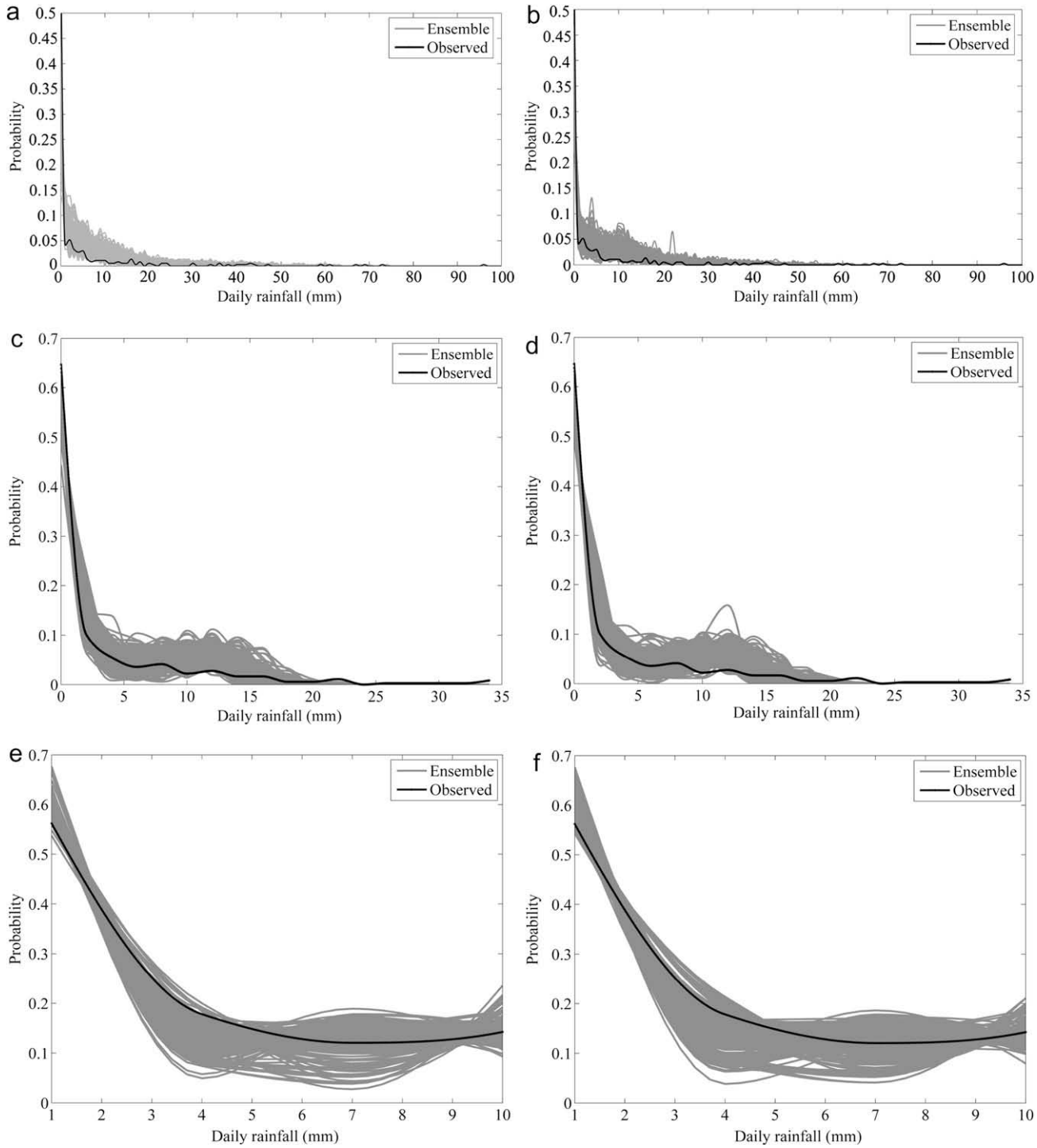

Fig. 16. Probability density functions of daily rainfall for the year 1996. The ensemble PDFs and also the observed rainfall PDF are shown. PDFs of predictions beginning from January for the year 1996 and June for all the three regions for the year 1996 are shown. (a) Probability density function of predictions beginning from January for Malaprabha for the year 1996; (b) probability density function of predictions beginning from June for Malaprabha for the year 1996; (c) probability density function of predictions beginning from January for Mahanadi for the year 1996; (d) probability density function of predictions beginning from June for Mahanadi for the year 1996; (e) probability density function of predictions beginning from January for All-India for the year 1996; and (f) probability density function of predictions beginning from June for All-India for the year 1996.

ber of optimum parameter combinations are fixed as 150 for Malaprabha and 200 for Mahanadi and All-India.

Prediction is done using local approximation method with these selected parameter combinations for a particular year using the data till the preceding year. This is done for five years from 1996 to 2000. Also, predictions are done from two different starting points, first one from the beginning of January till end of December of the corresponding year and the second prediction is done from the beginning of June till the end of May next year. Predictions beginning from June are started from June month of the year 1995 and prediction is done for an entire year till May month of next year 1996. However, for comparing both predictions (beginning from January and June), predicted values of one calendar year from January to December are taken into account instead from June to May. The optimum parameter combinations which give minimum GCV values for the three regions are given in Table 2.

The correlations of the observed daily rainfall for each year with the mean ensemble daily rainfall values for the predictions beginning from January and June are shown in Table 3. For all the three regions, there is no appreciable difference between the correlation values of January and June predictions. However, when comparing the correlations for observed average daily rainfall values for a month and mean ensemble average daily rainfall values, it can be noticed that for Malaprabha region, predictions beginning from June are comparatively better than those beginning from January. However, for the other two regions, the correlation values are almost equal. The corresponding correlation values are shown in Table 4.

A detailed analysis is done by constructing the box plots of mean daily rainfall values of the ensembles and comparing them with the observed series. As an illustration the box plots of the mean daily rainfall ensemble values of predictions beginning from January and June for the year 1996 for the three regions are shown in Fig. 14(a) to (f). The box plots give the range of values of the ensembles generated for each month. A box in the box plots indicates the interquartile range of the mean daily rainfall ensemble values and the horizontal dashed line within the box indicates the median ensemble mean daily rainfall value. The upper and lower whiskers of the box plots indicate the 95th and 5th percentile value and thus show the extent of the rest of the data. The observed mean daily rainfall values are also shown as a solid continuous line in the figure.

Considering Malaprabha basin, even though both the predictions (January and June) are able to capture the original rainfall values (with monsoon months values well within the range and non-monsoon months values marginally), it can be noticed that predictions beginning from June are relatively better for capturing the unusual variations of the annual cycle of rainfall. This may be due to the nearness of the prediction starting point (i.e., June month) to the summer monsoon rainfall season (June to September), during which the basin gets maximum rainfall as shown in Table 1. This is also indicated by the positive Lyapunov exponent value, which also indicates the inefficiency of long term prediction. But, such a distinction is not evident for the January and June predictions of Mahanadi and All-India regions.

Hence, it can be concluded that June prediction is effective only for Malaprabha region. For the rest two regions, both predictions are almost giving equal performances. The evident reasons for this behaviour are (i) the difference in spatial areas and (ii) the difference in Lyapunov exponents (Mahanadi and All-India have low Lyapunov expoenents indicating a high predictability when compared to Malaprabha basin). This behaviour can also be due to the difference in the coefficients of variance (Cv) of the three regions as shown in Fig. 15. The deviation of the Cv values for Malaprabha basin over the year is much higher when compared to the other two regions. The better performance of June prediction for Malaprabha may be due to starting of the prediction at a comparatively low Cv time period. Such a behaviour is not seen for Maha-

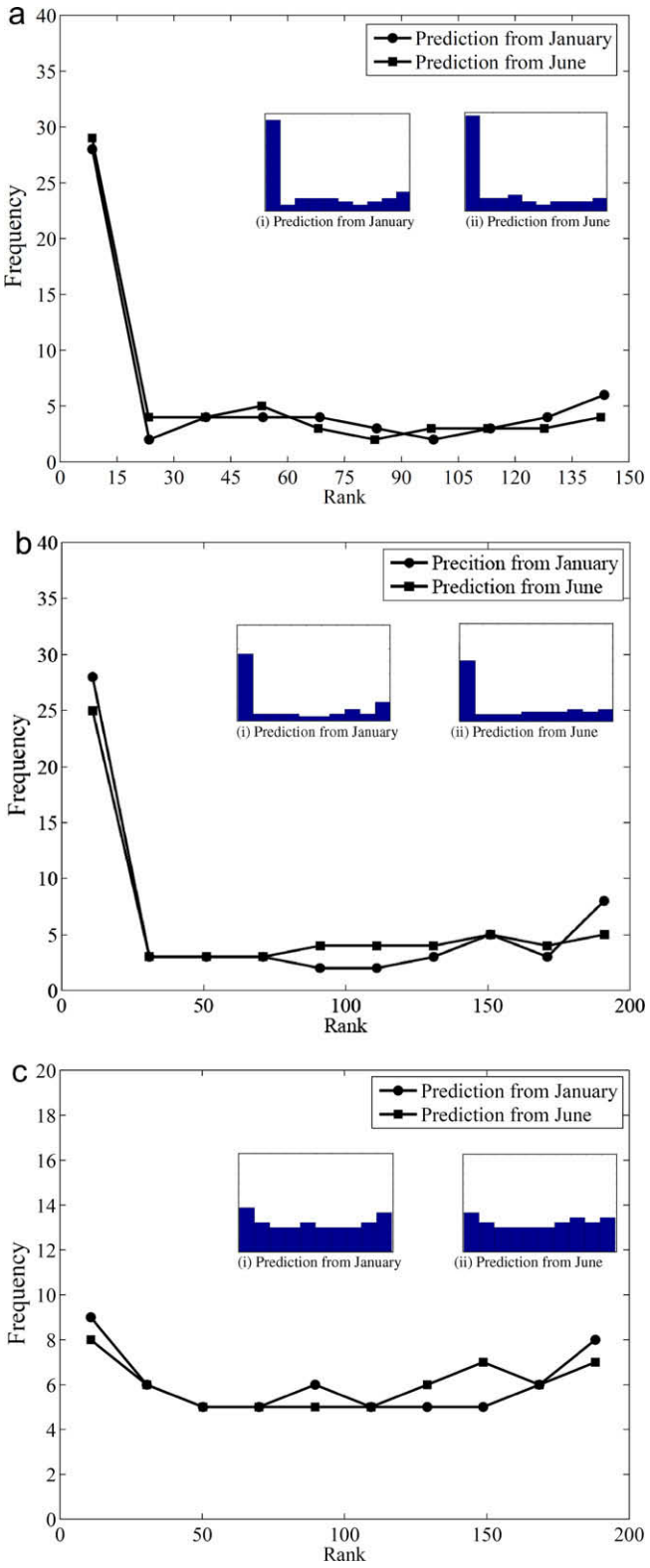

Fig. 17. Rank plots for ensembles generated considering all months of a calendar year. Rank plot for predictions beginning from January and June for (a) Malaprabha daily rainfall; (b) Mahanadi daily rainfall and (c) All-India daily rainfall are shown. The original rank histograms are shown in insets (the axes labels are same as those of the main figure).

nadi and All-India regions since their *Cv* values are more or less the same throughout the year.

A comparison of the probability density functions (PDFs) of all the ensembles and the corresponding observed series for the daily rainfall of the year 1996 are shown in Fig. 16(a)–(f). As can be seen from the figures, the ensembles and the observed series are following same distribution function. For Malaprabha region the original pdf is towards the lower end of the ensemble spread for the particular year shown. But for Mahanadi and All-India regions the ensemble pdfs for predictions beginning from January and June are able to reasonably catch the observed series PDF within its spread.

The quality of the ensembles generated is ascertained using two measures: rank probability skill score (RPSS) and rank histogram. The dataset is divided into three categories based on the 33rd and 66th percentile values derived from the observed dataset from 1955 to 1995. These categories are determined for each month separately. The RPSS values for the January and June predictions for each month for the five years are presented in Table 5. Excluding the months of January, February, March and December, the RPSS values for the other months are positive, which indicate a better forecast than the climatological forecast. The negative RPSS values for the first three months and for the December month may be due to the low rainfall received during these months. The ensembles are not able to reproduce a value closer to zero as can be seen from the box plots of monthly rainfall values of these months, even though the spreads of the concerned box plots are less (Fig. 14).

The rank histograms for evaluating the mean and spread of the ensembles (150 for Malaprabha and 200 for the other two regions) generated for each month and each year (a total of $12 \times 5 = 60$ data points) are prepared. The rank plots of predictions beginning from January and June are compared in Fig. 17. The rank histograms for both cases are given as insets in Fig. 16. It can be seen that for Malaprabha and Mahanadi (Fig. 17(a) and (b)) there is an ensemble bias which causes more population towards the lower ranks, in both predictions. The rest of the histogram is almost flat. Such a bias again may be due to too low rainfall values in the first three months and the last month (around 58% of Malaprabha data and 38% of Mahanadi data are zeros). The rank histograms of All-India rainfall (Fig. 17(c)) which contains only 2% zeros are almost flat in both prediction cases. To analyse the effect of the bias in Malaprabha and Mahanadi basins, the data points from the months of January, February, March, November and December are excluded. Rank histograms are now constructed for a total of $7 \times 5 = 35$ data points for both predictions. The corresponding rank plots are shown in Fig. 18, in which rank histograms are shown as insets. It can be seen that these rank histograms are almost flat in both cases implying ensembles of reliable spread.

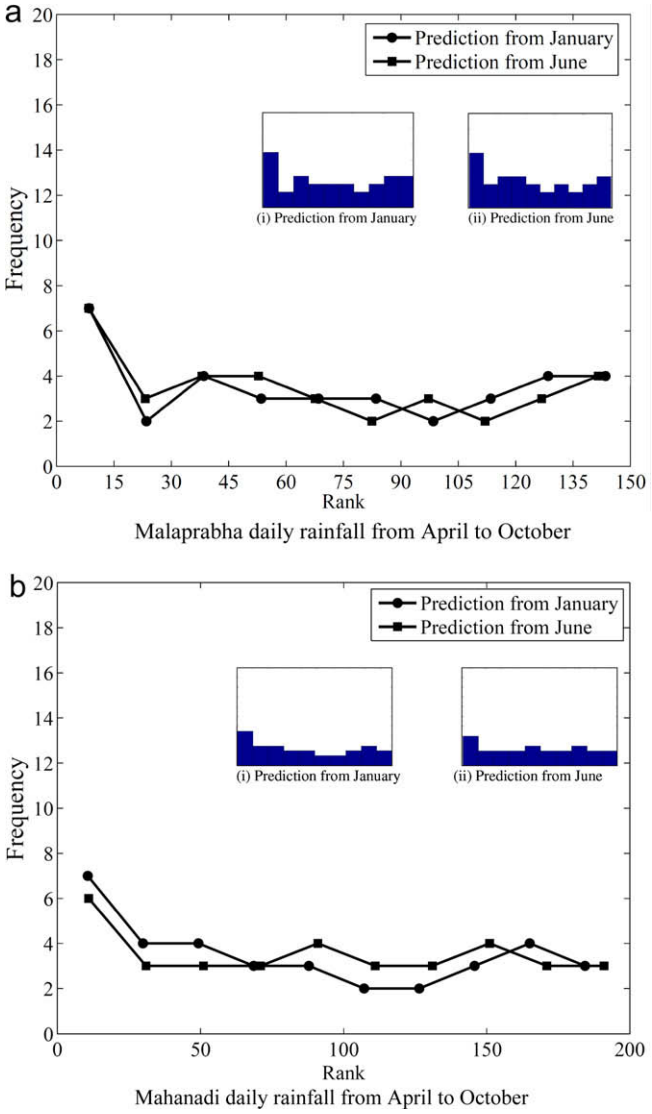

Fig. 18. Rank histograms for the ensembles generated considering only the months from April to October. Rank plot for predictions beginning from January and June for (a) Malaprabha and (b) Mahanadi are shown. The original rank histograms for both the predictions are shown in insets (the axes labels are same as those of the main figure).

6. Conclusions

The task of modeling rainfall is quite difficult because of its complexity and also due to large variability in both space and time. Over decades, the processes connecting rainfall have been treated as stochastic. The recent interest in nonlinear dynamics and also chaos theory has drawn attention towards considering rainfall as a chaotic system which is much sensitive to initial conditions and is short term predictable. The present study was aimed at analyzing the chaotic nature of rainfall series using different techniques and finally employing the nonlinear prediction technique for generating an ensemble of predictions. Daily rainfall data for the period 1955–2000 of Malaprabha, Mahanadi and All-India regions exhibiting distinct areal behaviours were considered for the study. These regions having different characteristics were selected to analyse the association of chaotic behaviour on the coverage area of basin.

The behaviour of rainfall dynamics was investigated using correlation dimension method with Grassberger–Procaccia algorithm (GPA). The clear scaling region in the $C(r)$ versus r plots on a log–log scale and also correlation exponent saturation values indicate low dimensional chaotic behaviour of the three rainfall series. The correlation dimension (minimum number of variables required to describe the system) is increasing with an increase in the coverage area. It is also notable that the embedding dimension (maximum number of variables required to describe the system) remains the same i.e., 19 for all the three regions.

Since colored random noises also exhibit a finite correlation dimension value, the above method is repeated on phase randomized data and on first derivative of the three rainfall series. The correlation dimensions of phase randomized data are not converging, while those of first derivative are almost same as of the original data. This elucidates that the saturation of correlation dimension is not due to the inherent linear correlation in the data; but because of the low dimensional chaotic dynamics present in the data. However, since one should not confirm the chaotic nature based on

the correlation dimension method alone, two other methods namely False nearest neighbour (FNN) algorithm and nonlinear prediction method are employed.

The fraction of false nearest neighbours is falling to a minimum value at an embedding dimension of 7 for Malaprabha basin, which indicates that the optimum embedding dimension of the rainfall series is 7. However, there is steep increase of fraction of FNN after embedding dimension 7. Such behaviour is not encountered in the other two cases. Therefore, the steep increase can be attributed to the presence of either noise or too much of singular values in the Malaprabha rainfall series. The inverse approach using nonlinear prediction method also supported the low dimensional chaotic nature of the three rainfall series. Hence, the minimum number of variables essential to model the dynamics of the rainfall was in the range between 5 and 7 for Malaprabha, 6 for Mahanadi and 8–9 for All-India.

The positive Lyapunov exponents of the three regions confirm the unpredictability of the systems. However, it can be seen that the predictability increases with coverage area with All-India region having the highest predictability. Even though the methods employed support the low dimensional chaotic nature of the rainfall series, a surrogate data test was done for three cases to confirm the nonlinearity of the data set. The much lower nonlinear prediction error of the observed data set when compared to the 99 surrogates rejects the null hypothesis that the original data is from a linear stochastic process at 1% significance level. These results suggest that the seemingly irregular behavior of rainfall process can be better explained though a chaotic framework for three of the rainfall series taken for the study.

Since different methods are giving slightly different values for the optimum embedding dimension and delay time, an ensemble of predictions was generated using a range of parameters (embedding dimension, delay time and neighbourhood size). Such an ensemble of predictions will be able to capture the uncertainty in the complex rainfall process. Predictions were done from the starting of January and also from the starting of June for an entire year in each case. Results had shown that for Malarabha basin, predictions beginning from June are comparatively better due to their closeness to the summer monsoon months (June–September) and also due to the low coefficient of variance in those months. The other two regions are showing almost same outcomes for January and June predictions due to their constant coefficients of variance throughout the year and also due to the high predictability (low Lyapunov exponent) when compared to that of Malaprabha basin. The rank histograms and RPSS values for the three regions indicate reasonably good spread ensembles from the nonlinear prediction method.

The methods employed support the short term predictability nature of a chaotic series. Also, the nonlinear prediction method, i.e., local approximation method was able to create quality ensemble predictions with good spread and skill. The reasonably good predictions obtained using a chaos theory based nonlinear prediction method affirm the suitability of a chaotic approach for modeling and understanding the underlying dynamics of the complex rainfall process. It is worthwhile to note that an increase in coverage area causes an improvement in predictability and an increase in dimension which further point towards a shift from chaotic nature to stochastic nature. It is well known that climate which is the spatial and temporal average of weather is more predictable than weather itself. However, it is evident from the above results that spatial averaging over a large area alone can also increase the predictability, hence shifting the system to a non-chaotic one.

Acknowledgement

We sincerely acknowledge the three anonymous reviewers for their valuable suggestions and comments that helped us considerably to refine the conceptual aspects of the subject and hence enabled us to improve presentation.

References

[1] Casdagli M. Nonlinear prediction of chaotic time series. Physica D 1989;35:335–56.
[2] Casdagli M. Chaos and deterministic versus stochastic nonlinear modeling. J Roy Stat Soc B 1991;54:303–24.
[3] Elshorbagy A, Simonovic SP, Panu US. Noise reduction in chaotic hydrologic time series: facts and doubts. J Hydrol 2002;256(3/4):845–8.
[4] Farmer JD, Sidorowich JJ. Predicting chaotic time series. Phys Rev Lett 1987;59:845–8.
[5] Frazer AM, Swinney HL. Independent coordinates for strange attractors from mutual information. Phys Rev A 1986;33(2):1134–40.
[6] Grassberger P, Procaccia I. Measuring the strangeness of strange attractors. Physica D 1983;9:189–208.
[7] Grassberger P, Procaccia I. Estimation of the Kolmogorov entropy from a chaotic signal. Phys Rev A 1983;28:2591–3.
[8] Hegger R, Kantz H. Improved false nearest neighbor method to detect determinism in time series data. Phys Rev E 1999;60:4970–3.
[9] Holzfuss J, Mayer-Kress G. An approach to error-estimation in the application of dimension algorithms. In: Mayer-Kress G, editor. Dimensions and entropies in chaotic systems. New York: Springer; 1986. p. 114–22.
[10] Islam MN, Sivakumar B. Characterization and prediction of runoff dynamics: a nonlinear dynamical view. Adv Water Resour 2002;25:179–90.
[11] Jayawardena AW, Gurung AB. Noise reduction and prediction of hydrometeorological time series: dynamical systems approach vs. stochastic approach. J Hydrol 2000;228:242–64.
[12] Jayawardena AW, Lai F. Analysis and prediction of chaos in rainfall and stream flow time series. J Hydrol 1994;153:23–52.
[13] Kantz H. A robust method to estimate the maximal Lyapunov exponent of a time series. Phys Lett A 1994;185:77–87.
[14] Kantz H, Schreiber T. Nonlinear time series analysis. 2nd ed. Cambridge, UK: Cambridge University Press; 2004.
[15] Kennel MB, Brown R, Abarbanel HDI. Determining embedding dimension for phase space reconstruction using a geometric method. Phys Rev A 1992;45:3403–11.
[16] Krasovskaia I, Gottschalk L, Kundzewicz ZW. Dimensionality of Scandinavian river flow regimes. Hydrol Sci J 1999;44(5):705–23.
[17] Liu Q, Islam S, Rodriguez-Iturbe I, Le Y. Phase-space analysis of daily streamflow: characterization and prediction. Adv Water Resour 1998;21:463–75.
[18] Nerenberg MAH, Essex C. Correlation dimension and systematic geometric effects. Phys Rev A 1990;42(12):7065–74.
[19] Osborne AR, Provenzale A. finite correlation dimension for stochastic systems with power law spectra. Physica D 1989;35:357–81.
[20] Packard NH, Crutchfield JP, Farmer JD, Shaw RS. Geometry from a time series. Phys Rev Lett 1980;45(9):712–6.
[21] Porporato A, Ridolfi L. Clues to the existence of deterministic chaos in river flow. Int J Mod Phys B 1996;10(15):1821–62.
[22] Porporato A, Ridolfi L. Nonlinear analysis of river flow time sequences. Water Resour Res 1997;33(6):1353–67.
[23] Porporato A, Ridolfi L. Multivariate nonlinear prediction of riverflow. J Hydrol 2001;248:109–22.
[24] Provenzale A, Smith LA, Vio R, Murante G. Distinguishing between low-dimensional dynamics and randomness in measured time series. Physica D 1992;58:31–49.
[25] Puente CE, Obregon N. A deterministic geometric representation of temporal rainfall: results for a storm in Boston. Water Resour Res 1996;32(9):2825–39.
[26] Ramsey JB, Yuan HJ. The statistical properties of dimension calculations using small data sets. Nonlinearity 1990;3:155–76.
[27] Rodriguez-Iturbe I, De Power FB, Sharifi MB, Georgakakos KP. Chaos in rainfall. Water Resour Res 1989;25(7):1667–75.
[28] Rosenstein MT, Collins JJ, De Luca CJ. A practical method for calculating largest Lyapunov exponents from small data sets. Physica D 1993;65:117–34.
[29] Sangoyomi T, Lall U, Abarbanel HDJ. Nonlinear dynamics of the Great Salt Lake: dimension estimation. Water Resour Res 1996;32(1):149–59.
[30] Schreiber T, Grassberger P. A simple noise reduction method for real data. Phys Lett A 1991;160:411–8.
[31] Schreiber T, Kantz H. Observing and predicting chaotic signals: Is 2% noise too much? In: Kravtsov YuA, Kadtke JB, editors. Predictability of complex dynamical systems. Springer series in synergetics. Berlin, Germany: Springer; 1996. p. 43–65.
[32] Schreiber T, Schmitz A. Improved surrogate data for nonlinearity tests. Phys Rev Lett 1996;77:635–8.
[33] Schreiber T, Schmitz A. Surrogate time series. Physica D 2000;142:346–82.
[34] Sharifi MB, Georgakakos KP, Rodriguez-Iturbe I. Evidence of deterministic chaos in the pulse of storm rainfall. J Atmos Sci 1990;47:888–93.
[35] Sivakumar B. Rainfall dynamics at different temporal scales: a chaotic perspective. Hydrol Earth Syst Sci 2001;5(4):645–51.
[36] Sivakumar B, Berndtsson R, Olsson J, Jinn K. Evidence of chaos in the rainfall–runoff process. Hydrol Sci J 2001;46(1):131–45.
[37] Sivakumar B, Liong SY, Liaw CY. Evidence of chaotic behavior in Singapore rainfall. J Am Water Resour Assoc 1998;34(2):301–10.

[38] Sivakumar B, Liong SY, Liaw CY, Phoon KK. Singapore rainfall behavior: chaotic? J Hydrol Eng 1999;4(1):38–48.
[39] Sivakumar B, Phoon KK, Liong SY, Liaw CY. A systematic approach to noise reduction in chaotic hydrological time series. J Hydrol 1999;219(3/4):103–35.
[40] Sivakumar B, Sorooshian S, Gupta HV, Gao X. A chaotic approach to rainfall disaggregation. Water Resour Res 2001;37(1):61–72.
[41] Smith LA. Intrinsic limits on dimension calculations. Phys Lett A 1988;133(6):283–8.
[42] Stehlik J. Deterministic chaos in runoff series. J Hydraul Hydromech 1999;47(4):271–87.
[43] Strogatz SH. Nonlinear dynamics and chaos: with applications to physics, biology, chemistry and engineering. Cambridge: Westview Press, Perseus Books Group; 1994.
[44] Takens F. Detecting strange attractors in turbulence. In: Rand DA, Young LS, editors. Lectures notes in mathematics, vol. 898. Berlin, Germany: Springer-Verlag; 1981. p. 366–81.
[45] Theiler J, Eubank S, Longtin A, Galdikian B, Farmer JD. Testing for nonlinearity in time series: the method of surrogate data. Physica D 1992;58:77–94.
[46] Tsonis AA, Elsner JB. The weather attractor over very short timescales. Nature 1988;333:545–7.
[47] Wang Q, Gan TY. Biases of correlation dimension estimates of streamflow data in the Canadian prairies. Water Resour Res 1998;34(9):2329–39.
[48] Wilks DS. Statistical methods in the atmospheric sciences: an introduction. 2nd ed. New York: Elsevier; 2005.
[49] Wolf A, Swift JB, Swinney HL, Vastano A. Determining Lyapunov exponents from a time serie. Physica D 1985;16:285–317.