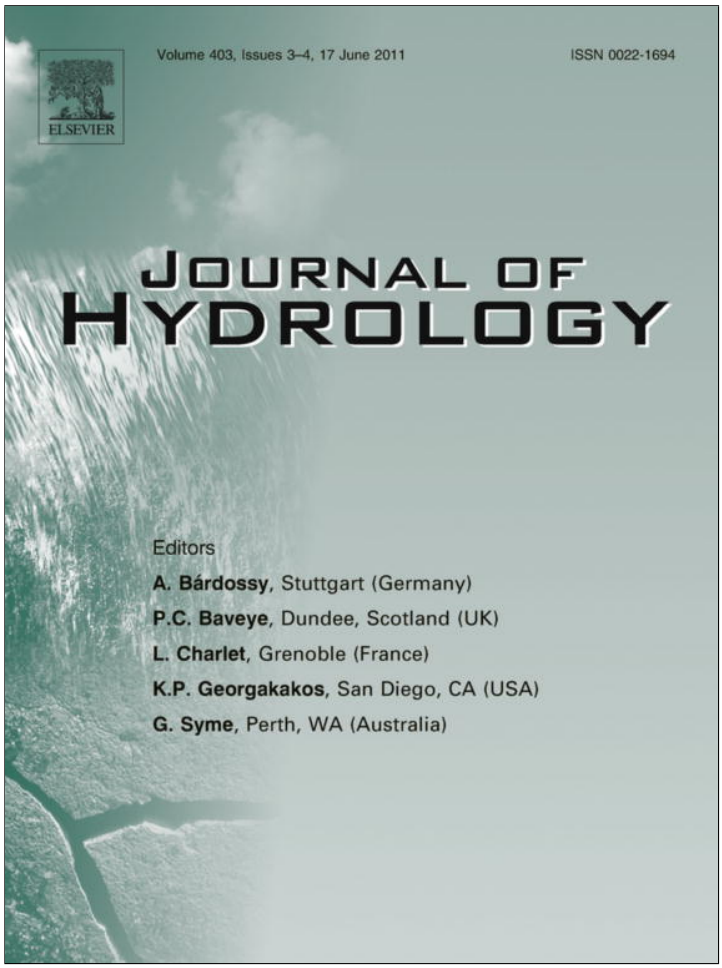


Journal of Hydrology 403 (2011) 292–306

Contents lists available at ScienceDirect

Journal of Hydrology

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

Multivariate nonlinear ensemble prediction of daily chaotic rainfall with climate inputs

C.T. Dhanya, D. Nagesh Kumar *

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

ARTICLE INFO

Article history:
Received 28 June 2010
Received in revised form 18 March 2011
Accepted 4 April 2011
Available online 9 April 2011
This manuscript was handled by
A. Bardossy, Editor-in-Chief, with the
assistance of Alin Andrei Carsteanu,
Associate Editor

Keywords:
Chaotic nature
Nonlinearity
Ensemble prediction
Uncertainty
Predictability
Multivariate phase space

SUMMARY

The basic characteristic of a chaotic system is its sensitivity to the infinitesimal changes in its initial conditions. A limit to predictability in chaotic system arises mainly due to this sensitivity and also due to the ineffectiveness of the model to reveal the underlying dynamics of the system. In the present study, an attempt is made to quantify these uncertainties involved and thereby improve the predictability by adopting a multivariate nonlinear ensemble prediction. Daily rainfall data of Malaprabha basin, India for the period 1955–2000 is used for the study. It is found to exhibit a low dimensional chaotic nature with the dimension varying from 5 to 7. A multivariate phase space is generated, considering a climate data set of 16 variables. The chaotic nature of each of these variables is confirmed using false nearest neighbor method. The redundancy, if any, of this atmospheric data set is further removed by employing principal component analysis (PCA) method and thereby reducing it to eight principal components (PCs). This multivariate series (rainfall along with eight PCs) is found to exhibit a low dimensional chaotic nature with dimension 10. Nonlinear prediction employing local approximation method is done using univariate series (rainfall alone) and multivariate series for different combinations of embedding dimensions and delay times. The uncertainty in initial conditions is thus addressed by reconstructing the phase space using different combinations of parameters. The ensembles generated from multivariate predictions are found to be better than those from univariate predictions. The uncertainty in predictions is decreased or in other words predictability is increased by adopting multivariate nonlinear ensemble prediction. The restriction on predictability of a chaotic series can thus be altered by quantifying the uncertainty in the initial conditions and also by including other possible variables, which may influence the system.

1. Introduction

A major break through in the routine weather prediction occurred in the mid 20th century with the development of various climate models that numerically integrate an adequate set of mathematical equations which explain the physical laws governing the climatic processes. However, these mathematical equations form a nonlinear dynamical system in which an infinitesimally small uncertainty in the initial conditions will grow exponentially even under a perfect model, leading to a chaotic behavior (Smith et al., 1998). The sensitivity of any deterministic system to a slight change in the initial conditions leading to a vast change in the final solution is often known as "butterfly effect" in the field of weather forecasting (Lorenz, 1972). Hence, earth's weather can be treated as a chaotic system with a finite limit in the predictability, arising mainly due to the indefiniteness of the initial conditions.

An infinitesimal initial uncertainty ∂_0 grows exponentially with time at a rate of separation given by the highest Lyapunov exponent λ (Wolf et al., 1985; Rosenstein et al., 1993). Thus, the separation or uncertainty after Δt time steps ahead is $\partial_{\Delta t} \cong e^{\lambda \Delta t} \partial_0$. Variations in the reliability of any individual forecast due to this uncertainty can be quantified by generating an ensemble of forecasts with slightly varying initial conditions. Any uncertainty in the initial conditions is reflected in the evolution of the ensemble, and hence the nonlinearity. Also, an estimate of the stability of the forecasts can be obtained by observing how quickly the ensemble spreads out (or shrinks). Many operational centers now adopt the ensemble approach replacing the traditional best guess initial condition approach. Smith (2000) stated that since the ensembles can accurately reflect the likelihood of occurring of various future conditions, given a perfect model, chaos places no a priori limits on predictability. Hence, the predictability of a chaotic system is limited primarily (i) due to the indefiniteness in the initial conditions (given a perfect model) and also (ii) due to the imperfection of the model.

Many of the weather phenomena have been modeled so far employing the concept of stochastic systems. However, a large

* 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).

0022-1694/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.jhydrol.2011.04.009

number of studies employing the science of chaos to model and predict various hydrological phenomena have emerged only in the past 20 years (Elshorbagy et al., 2002; Islam and Sivakumar, 2002; Jayawardena and Lai, 1994; Porporato and Ridolfi, 1996, 1997; Puente and Obregon, 1996; Rodriguez-Iturbe et al., 1989; Liu et al., 1998; Sangoyomi et al., 1996; Sivakumar et al., 1999, 2001; Sivakumar, 2001; Shang et al., 2009; Wang and Gan, 1998). Most of these studies dealt with scalar time series data of various hydrological phenomena like rainfall, runoff, sediment transport, lake volume, etc. In these cases, since neither the mathematical relations nor the influencing variables are known, the state space in which the variable is lying is reconstructed from the time series itself using phase space reconstruction method by Takens (1981). The outcomes of these studies affirm the existence of low-dimensional chaos, thus indicating the possibility of only short-term predictions. Better predictions can be obtained using the chaotic approach since it takes into account the dynamics of the irregular hydrological phenomena from a chaotic deterministic view, thereby reducing the model uncertainty. Also, the dynamic approach employing chaotic theory outperforms the traditional stochastic approach in prediction (Jayawardena and Gurung, 2000). Studies attempting prediction have dealt with mainly univariate prediction by reconstructing the phase space from a scalar time series.

Most of these studies rely only on the low correlation dimension as a measure of the chaotic nature of the time series and as an estimate of the embedding dimension. Osborne and Provenzale (1989) claimed that a low correlation dimension can also be observed for a linear stochastic process. Hence, it is advised to assess the chaotic nature and to determine the embedding dimension and delay time by employing a variety of methods (Islam and Sivakumar, 2002; Dhanya and Kumar, 2010). Since different methods give slightly different embedding dimensions and delay times for a single series, one should opt for an ensemble of predictions with a set of these parameters in order to capture the uncertainty in parameter estimation (Dhanya and Kumar, 2010, 2011). Thus, while dealing with time series data, the uncertainty in initial conditions can be quantified by generating a set of predictions with different combinations of parameters.

Although according to embedding theorem by Takens (1981), a scalar series is sufficient to reconstruct the dynamics of the system, its application in practical problems is not very fruitful. For example, while considering the three – variable Lorenz system (Lorenz, 1963), the measurements of z co-ordinate alone cannot resolve the dynamics of the Lorenz system (Cao et al., 1998). Hence, inclusion of other influencing variables in the time series strengthen the model, thus substantially improving the prediction. A very small number of studies (Cao et al., 1998; Jin et al., 2005; Porporato and Ridolfi, 2001) have attempted multivariate prediction by utilizing the information from other time series.

In view of the above, the present study attempts to improve the predictability of a chaotic daily rainfall series by employing a multivariate phase space prediction method. The information from various atmospheric variables is incorporated. Also, the uncertainty in initial conditions is quantified by generating an ensemble of predictions with a variety of plausible parameters. The predictability of a multivariate case is then compared with that of the univariate case.

2. Data used

The daily rainfall data of Malaprabha basin in India for the period 1955–2000 is considered for the present study. The daily rainfall data for the basin is extracted from the daily gridded rainfall data at $1^\circ \times 1^\circ$ resolution from the Indian Meteorological Department (IMD) (Rajeevan et al., 2006). The location map of Malaprabha basin is shown in Fig. 1. It has a basin area of 2500 km^2 area situated between 15°30N and 15°56N latitudes and 74°12E and 75°15E longitudes. The catchment receives an average monsoon (June–September) rainfall of around 1800 mm. The mean and standard deviations of monthly rainfall of the region are presented in Table 1.

Anandhi et al. (2008) had selected 15 variables as the probable predictors for downscaling precipitation in Malaprabha basin. The daily variables extracted from the National Centre for Environmental Prediction (NCEP) reanalysis data sets (Kalnay et al., 1996) are the air temperature at 925, 700, 500 and 200 mb pressure levels (at925, at700 at500 and at200), geopotential height at 925, 500 and 200 mb pressure levels (gpt925, gpt500 and gpt200), specific humidity at 925 and 850 mb pressure levels (sh925 and sh850), zonal (u-wind) and meridional (v-wind) wind velocities at 925 and 200 mb pressure levels (uwnd925, uwnd200, vwnd925 and vwnd200), precipitable water (pwr) and surface pressure (press). In addition to these variables, daily surface air temperature (atsurf) for the period 1955–2000 is also considered for multivariate nonlinear prediction, since it is directly related to the daily rainfall. These climate variables are extracted for the grid point at 15°N and 75°E.

The correlation of daily rainfall with these atmospheric variables for different lags is shown in Fig. 2. In most cases, the maximum correlation (either positive or negative) is obtained for lags zero or one.

3. Methodology

The methods employed and the stepwise procedure adopted for generating the ensemble prediction are described below.

3.1. Phase space reconstruction

The prediction algorithms on nonlinear dynamics are based upon the theory of dynamic reconstruction of a scalar series, which is done by reconstructing the phase space using the method of delays by Takens (1981). The phase space reconstruction provides a simplified, multi-dimensional representation of a single-dimensional nonlinear time series. 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) (Grassberger and Procaccia, 1983), and the False Nearest Neighbor (FNN) method (Kennel et al., 1992). 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. The phase space coordinates would not be independent if τ is too small, thus resulting in loss of information about the characteristics of the attractor structure. On the other hand, if τ is too large, there would be no dynamic correlation between the state vectors since the neighboring trajectories diverge, thus resulting in loss of information about the original system. The optimum τ which allows a reasonable spread of state space data points is usually determined using either autocorrelation function or the mutual information method (Frazer and Swinney, 1986).

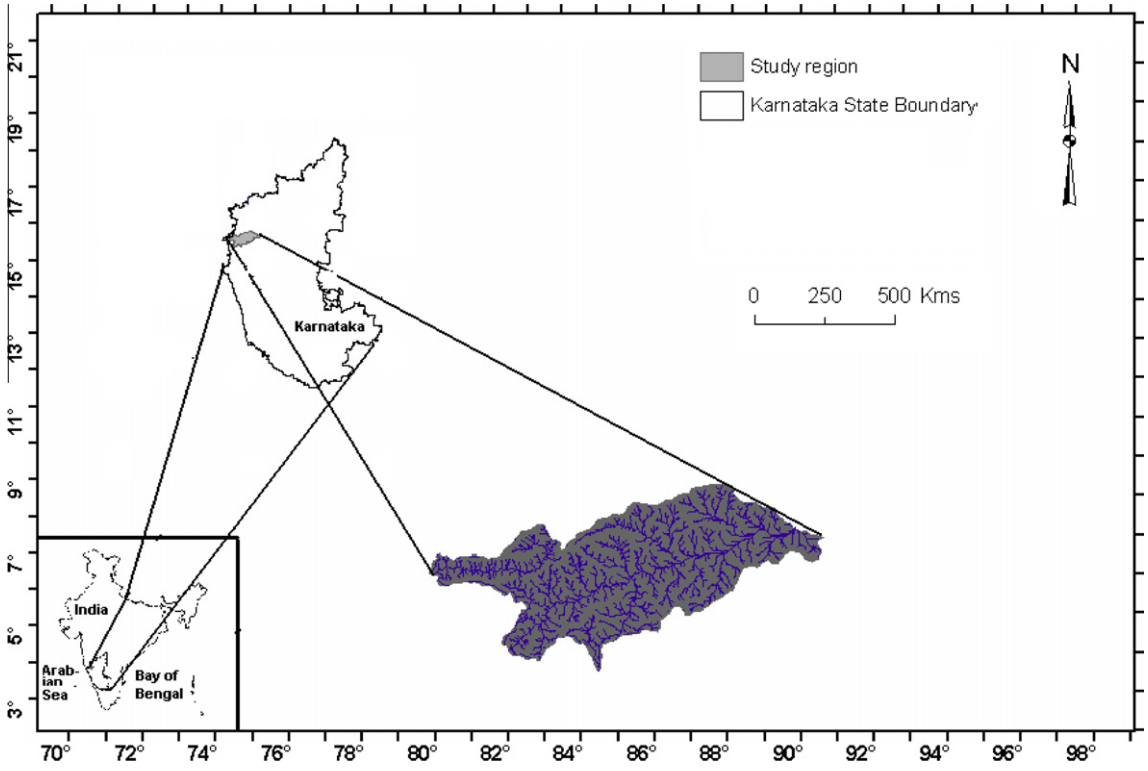


Fig. 1. Location map of Malaprabha basin. The latitude, longitude and scale of the map refer to the Karnataka state.

Table 1
Mean and standard deviations of monthly rainfall for the period 1955–2000.

Month	Mean (mm)	Standard deviation (mm)
January	1.0	3.9
February	0.8	2.1
March	6.0	11.6
April	29.0	27.1
May	83.7	78.0
June	417.4	168.5
July	770.0	299.8
August	441.3	195.4
September	167.1	74.1
October	138.4	93.4
November	39.5	49.1
December	5.2	11.4

The dynamics can be interpreted in the form of an m-dimensional map f_T such that

$$Y_{j+T} = F_T(Y_j) \tag{2}$$

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 $j+T$.

In multivariate case, instead of a single scalar series ($p = 1$), the time series of a number of independent variables are considered ($p > 1$). Consider a p-dimensional time series $X_1, X_2, \ldots, X_N$, where $X_i = (x_{1,i}, x_{2,i}, \ldots, x_{p,i})$, $i = 1, 2, \ldots, N$. The time delay vectors can be reconstructed as in the case of scalar time series in Eq. (1):

$$Y_j = \begin{pmatrix} x_{1,j}, x_{1,j-\tau_1}, & x_{1,j-2\tau_1}, \ldots, & x_{1,j-(m_1-1)\tau_1}, \\ x_{2,j}, x_{2,j-\tau_2}, & x_{2,j-2\tau_2}, \ldots, & x_{2,j-(m_2-1)\tau_2}, \\ \cdots & \cdots & \cdots \\ x_{p,j}, x_{p,j-\tau_M}, & x_{p,j-2\tau_M}, \ldots, & x_{p,j-(m_p-1)\tau_p} \end{pmatrix} \tag{3}$$

where τ_i, m_i, $i = 1, 2, \ldots, p$ are the delay times and the embedding dimensions of the p variables respectively. The total embedding dimension M is the sum of the individual embedding dimensions for each time series $M = \sum_{i=1}^{p} m_i$.

Similar to Eq. (2) there exists a function $F : \Re^d \to \Re^d (M = \sum_{i=1}^{p} m_i)$. Thus, the future value is based on the relation $Y_{j+T} = F_T(Y_j)$.

This is also equivalent to

$$\begin{aligned} x_{1,j+T} &= F_{1,T}(Y_j) \\ x_{2,j+T} &= F_{2,T}(Y_j) \\ &\vdots \qquad\qquad \vdots \\ x_{p,j+T} &= F_{p,T}(Y_j) \end{aligned} \tag{4}$$

provided m or m_i is sufficiently large.

The time delays τ_i are determined separately for each time series using auto correlation function or mutual information method. The complexity lies in the determination of embedding dimension of the multivariate data. All possible combinations with different m_i's need to be tried out to determine the optimal combination of embedding dimensions.

3.2. Correlation dimension method

Correlation dimension method also known as correlation integral analysis, is employed to analyse the chaotic nature of the daily rainfall time series. In this method, the correlation integral $C(r)$ is estimated using the Grassberger–Procaccia algorithm (Grassberger and Procaccia, 1983). According to the algorithm, for an m-dimensional reconstructed phase space (as given in Eq. (1)), the correlation integral $C(r)$ is given by

$$C(r) = \lim_{N\to\infty} \frac{2}{N(N-1)} \sum_{\substack{i,j \\ (1\leqslant i<j\leqslant N)}} H(r - |Y_i - Y_j|) \tag{5}$$

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

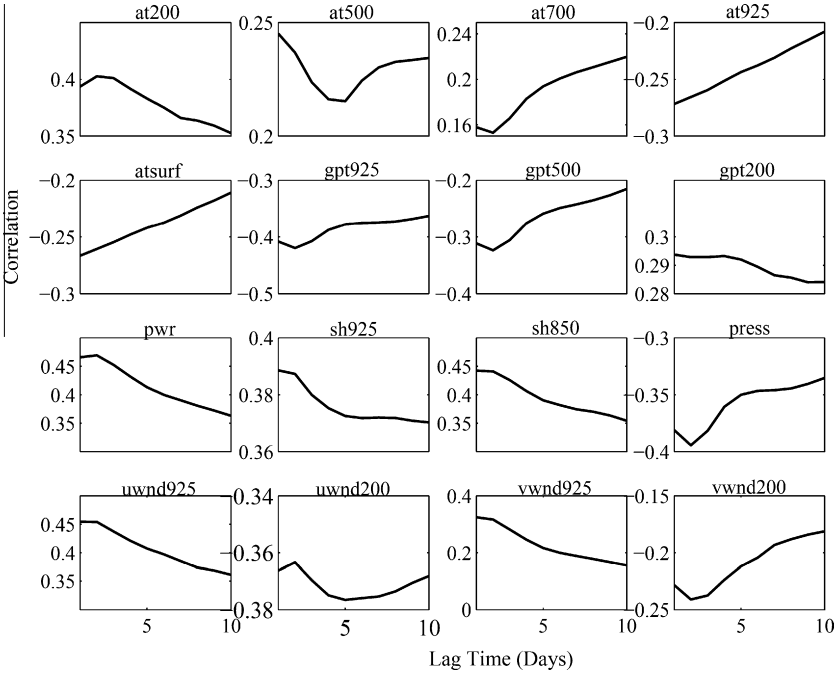


Fig. 2. Correlation of daily rainfall with all atmospheric variables for different lag times.

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 $\log r$.

The series is generally considered to be chaotic, if the correlation exponent saturates to a constant value even on increase in embedding dimension m. 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.

3.3. False nearest neighbor (FNN) method

The false nearest neighbor method (Kennel et al., 1992) is based on the concept that if the dynamics in phase space can be represented by a smooth vector field, then the neighboring states would be subject to almost the same time evolution (Kantz and Schreiber, 2004). Hence, any two close neighboring trajectories emerging from them should still be close neighbors, after a short time into the future. The original algorithm of Kennel et al. (1992) is modified by Hegger and Kantz (1999) to avoid any spurious results due to noise. This modified algorithm in which the fraction of false nearest neighbors is computed in a probabilistic way has been used in the present study.

In this modified algorithm, the basic idea is to search for all the data points which are neighbors in a particular embedding dimension m and which do not remain so, upon increasing the embedding dimension to $m+1$. To do this, consider a particular data point and determine its nearest neighbor 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, the neighbor is false. When the percentage of false nearest neighbors falls to zero (or a minimum value), the corresponding embedding dimension is considered high enough to represent the dynamics of the series.

While dealing with multivariate time series, the total dimension M is determined from the individual embedding dimensions m_i's for each component. The exact combination of individual embedding dimensions m_i's can be computed by either the prediction error minimization method by Cao et al. (1998) or the dimension reduction method by Velichov (2004). However, in the present study, since the approach is to generate an ensemble of predictions with different combinations of embedding dimensions, an exact combination of dimensions taken by each variable is not required. The following approach employing FNN algorithm is used to determine an approximate combination of the embedding dimensions and hence the total embedding dimension required for the data set.

The basic idea of FNN algorithm for multivariate case is the same as that for the univariate case. The change in the fraction of FNN is calculated over an increase in the embedding dimension from m to $m+1$. For multivariate data set of p components, as mentioned in Section 3.1, total dimension M is the sum of the individual embedding dimensions for each component i.e., $M = \sum_{i=1}^{p} m_i$.

For calculating the particular embedding dimension $M = \sum_{i=1}^{p} m_i$, all the combinations of the dimension vector $\boldsymbol{m} = (m_1, m_2, \ldots, m_p)$ need to be considered. The fraction of FNN is calculated for each such $\boldsymbol{m} = (m_1, m_2, \ldots, m_p)$ with all possible combinations in which M is increased to $M+1$ i.e., $(m_1+1, m_2, \ldots, m_p)$, $(m_1, m_2+1, \ldots, m_p), \ldots, (m_1, m_2, \ldots, m_p+1)$. The procedure is illustrated up to $M=3$ for a bivariate series in Fig. 3. This is done until the FNN fraction drops to zero for at least one of the embedding dimension vectors $\boldsymbol{m} = (m_1, m_2, \ldots, m_p)$ (Vlachos and Kugiumtzis, 2008).

3.4. Nonlinear prediction

To obtain contrasting predictions, nonlinear prediction method is used to counter-check the embedding dimensions obtained from

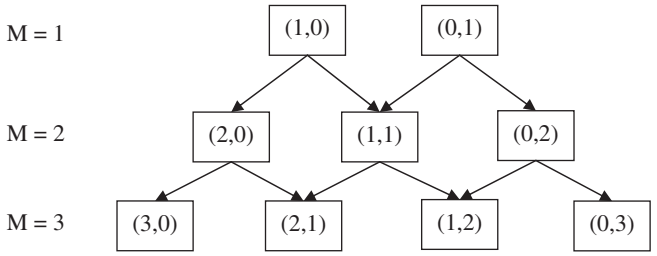


Fig. 3. Various combinations of embedding dimensions considered for a bivariate series.

the correlation integral and false nearest neighbor methods. 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

$$Y_{j+T} = f_T(Y_j) \tag{6}$$

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 local approach which works on local approximation (Farmer and Sidorowich, 1987), the dynamics are modeled locally piecewise in the embedding space. The domain is broken up into many local neighborhoods and modeling is done for each neighborhood separately, i.e., there will be a separate f_T valid for each neighborhood. The complexity in modeling f_T is thus considerably reduced without affecting the accuracy of prediction. Because of these advantages, local approximation method is employed in this study.

In local approximation method, the prediction of Y_{j+T} is done based on values of Y_j and k nearest neighbors of Y_j. These k nearest neighbors are selected based on the minimum values of $\|Y_j - Y_{j'}\|$ where $j' < j$. If only one nearest neighbor is considered then Y_{j+T} will be $Y_{j'+T}$. Since normally $k > 1$, the prediction of Y_{j+T} is taken as the weighted average of the k values. In the present study, the prediction of Y_{j+T} is done by averaging for k neighbors in the form $\widehat{Y}_{j+T} = \frac{1}{k}\sum_{i=1}^{k} Y_{i'+T}$. The optimum number of nearest neighbors is decided by trial and error. The prediction accuracy is estimated using the correlation coefficient, Nash efficiency coefficient and also normalized mean square error between the predicted series and the corresponding observed series.

Nonlinear prediction of multivariate case is similar to that for univariate case except in the phase space reconstruction. Once the reconstruction of the attractor is successfully achieved, the approximation of F_T in Eqs. (2) and (4) can be done using local approximation approach.

3.5. Principal component analysis (PCA)

Principal component analysis method generates a new set of variables which are linear combinations of the original variables. These uncorrelated new variables called principal components (PCs) contain no redundant information since they are orthogonal to each other. The principal components as a whole form an orthogonal basis for the space of the data. The following steps are used to derive the principal components:

i. The time series is standardized by dividing it with the corresponding standard deviation of each variable.
ii. The covariance matrix of the standardized time series is computed.
iii. The eigenvalues λ and the eigenvectors are calculated.
iv. These eigenvectors will be orthogonal to each other. The eigenvalues and the corresponding vectors are arranged according to the descending value of the eigenvalues. Hence, kth eigenvector corresponds to kth largest eigenvalue λ_k.
v. The multivariate series is projected on the space spanned by the eigenvectors to obtain the principal components (PCs). This is done by multiplying the eigenvector matrix and the original standardized time series.
vi. The percentage total variance ω_k explained by the kth principal component is given by: $\omega_k = \frac{\lambda_k}{\sum_{i=1}^{p}\lambda_i} \times 100$, where p is the dimensionality of the original data set.

3.6. Stepwise procedure

Univariate prediction:

(1) The range of delay time of the daily rainfall series is estimated using auto correlation and mutual information methods.

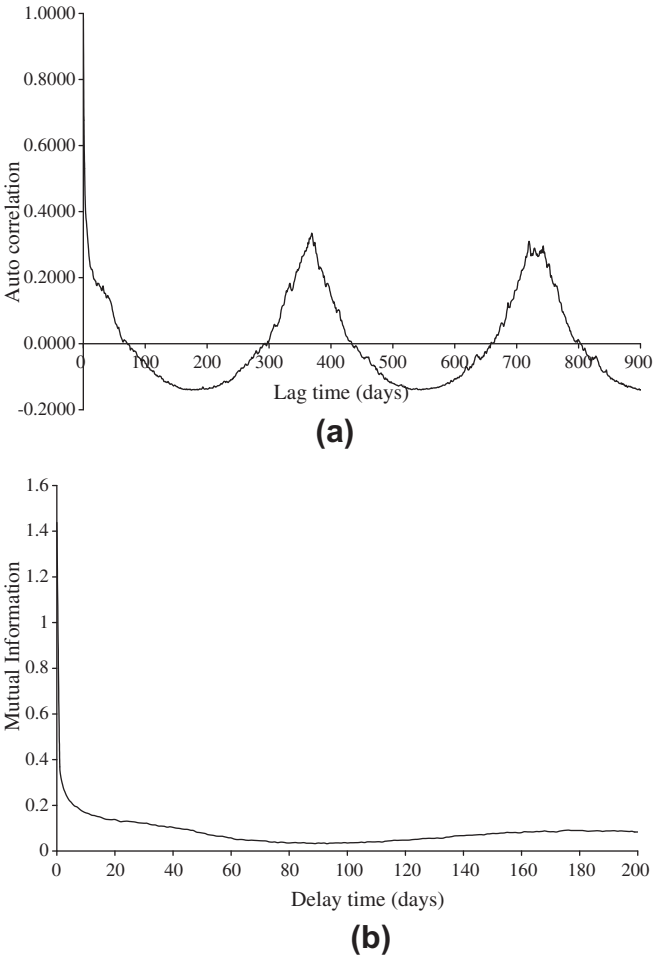


Fig. 4. (a) Autocorrelation function of Malaprabha daily rainfall. (b) Variation of mutual information with lag time.

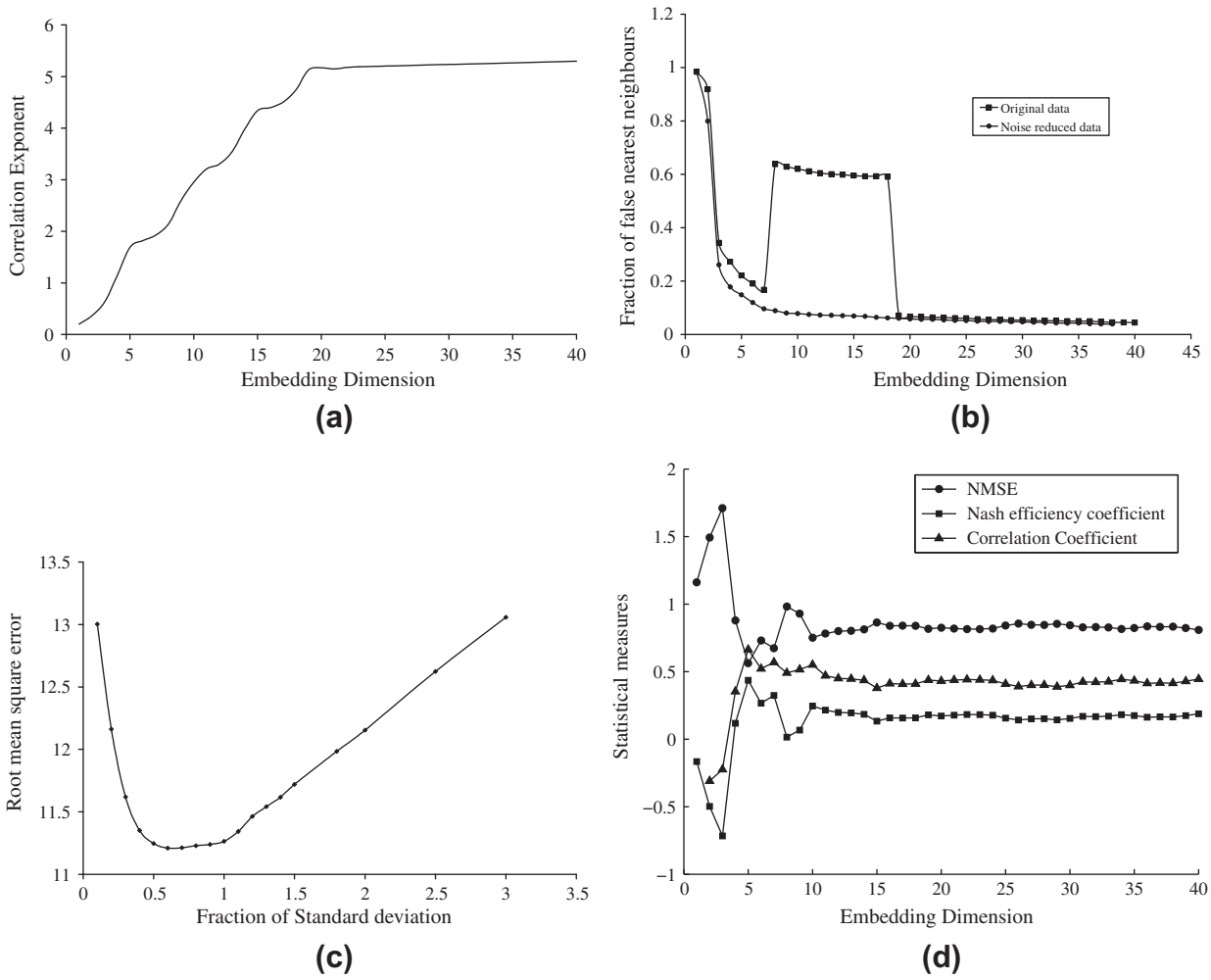


Fig. 5. (a) Variation of correlation exponent with embedding dimension. (b) Variation of fraction of false nearest neighbors with embedding dimension. (c) Variation of prediction error with neighborhood size. (d) Variation of statistical measures with embedding dimension.

(2) The range of embedding dimension of the daily rainfall series is estimated through correlation dimension, FNN and nonlinear prediction (local approximation) methods. The optimum neighborhood size (radius) from which the nearest neighbors are to be searched is determined using nonlinear prediction method. The neighborhood radius is expressed as a fraction α of the standard deviation.
(3) The phase space is reconstructed according to Eq. (1) for all the available combinations of the parameters m and τ. These parameter combinations are used to produce an ensemble of forecasts.

Multivariate prediction:

(4) The range of neighborhood is kept the same to allow a comparison with the univariate case.
(5) Analyze the chaotic nature of each time series using false nearest neighbor method.
(6) Reduce the dimension of the predictor time series (=16) by principal component analysis (PCA) method, retaining '*n*' number of principal components (PCs) which explain more than 90% of the variance of the original time series.
(7) The multivariate time series is generated by combining rainfall with *n* PCs.
(8) Determine the delay time τ_i for *n* PCs using the autocorrelation method and mutual information method.
(9) Check the embedding dimension of this multivariate series using false nearest neighbor method.
(10) Reconstruct the phase space according to Eq. (3) for dimensions from $n+1$ to $2(n+1)$ with various possible combinations of m_i. Ensembles are generated with all these combinations.
(11) The quality of the ensembles of univariate and multivariate cases is analyzed using rank histogram which is a graphical method to evaluate the reliability and probable predictability of the targeted parameter by the ensembles. If 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 is a flat one, a U-shaped rank histogram indicates ensemble members from a less variable distribution. U-shaped histogram indicates that the spread is too small and that many observations are falling outside the extremes of the ensembles; whereas a dome shape indicates that ensemble spread is too large and that too many observations are falling in the middle range. An ensemble bias (positive or negative) excessively populates the low and high ranks.

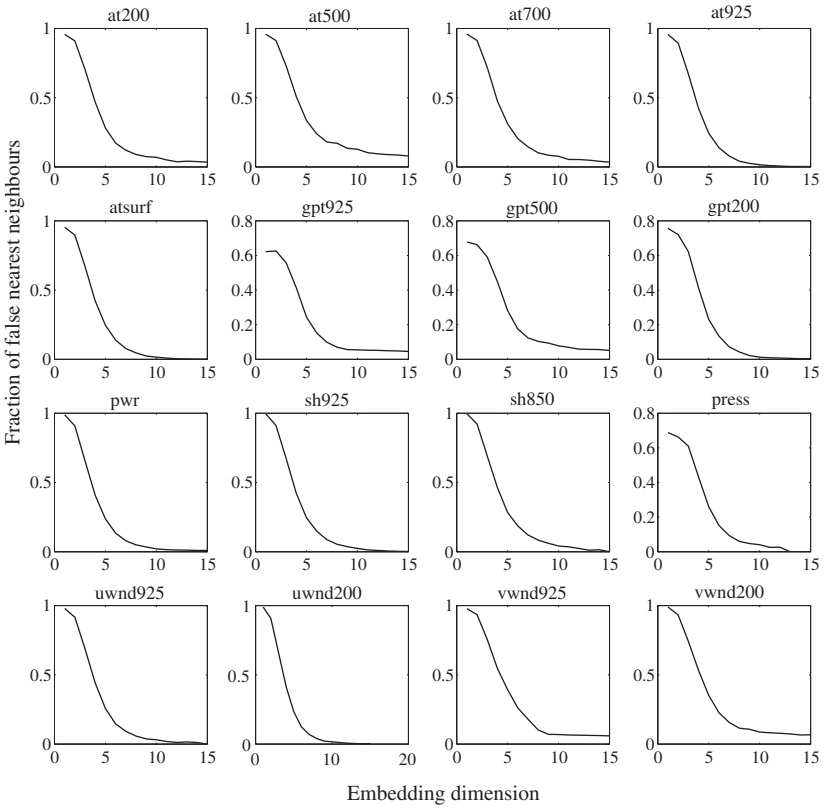


Fig. 6. Variation of fraction of FNN over embedding dimension for atmospheric variables.

4. Results and discussion

4.1. Univariate series

4.1.1. 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. In mutual information method, the delay time is chosen as the first minimum value. The autocorrelation and mutual information plots for various lag times are shown in Fig. 4a and b respectively. The delay times obtained from these two methods are 71 and 93 respectively. The range of delay time for the ensemble prediction is adopted as 60–100 days, allowing a little extra spread.

4.1.2. Determination of embedding dimension

The embedding dimension of the daily rainfall series of Malaprabha is determined using correlation dimension, false nearest neighbor and nonlinear prediction methods. The chaotic nature and the nonlinearity structure of the Malaprabha daily rainfall are analyzed in detail in Dhanya and Kumar (2010). Some of those results are reproduced here for ready reference.

The correlation integral $C(r)$ is determined according to Grassberger–Procaccia algorithm for embedding dimensions 1–40. The variation of the correlation exponent with the embedding dimension is shown in Fig. 5a. It can be noticed that the correlation exponent is increasing with embedding dimension and reaching a constant value at embedding dimension $m \geqslant 18$, which indicates the chaotic nature of the rainfall.

The variation of fraction of FNN with embedding dimension is shown in Fig. 5b. The fraction of nearest neighbors is falling to a minimum value at an embedding dimension of 7. The steep increase of FNN after 7th dimension up to $m = 18$ can be attributed to the presence of additive noise in the data series or to the presence of a large amount of zeros (about 57%) in the time series (Dhanya and Kumar, 2010). 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. Still, the application of a simple nonlinear noise reduction method (Schreiber, 1993) reveals that the unusual rise of FNN fraction does not exist in the noise reduced data (Fig. 5b). At higher dimensions, the FNN fraction of noise reduced data and original data are merging. However, in the noise reduced data also, the decrease of FNN fraction after an embedding dimension of 7 is minimal. Thus, the optimum embedding dimension is adopted as 7.

Nonlinear local constant prediction method is used as inverse method to determine the embedding dimension. For this, daily rainfall for the year 2000 is predicted using the daily rainfall series from 1955 to 1999. The optimum neighborhood size is determined by plotting the variation of the prediction error (root mean square error, RMSE) with the neighborhood size (which is a fraction of standard deviation) for $m = 6$ and is shown in Fig. 5c. The optimum neighborhood size is adopted as $0.5\times$ standard deviation for which the prediction error is the least.

In order to determine the embedding dimension, the prediction accuracy (in terms of normalized mean square error (NMSE), Nash efficiency coefficient and correlation coefficient) is calculated for various embedding dimensions using the optimum nearest neighbors and its variation is shown in Fig. 5d. The maximum prediction accuracy is found for an embedding dimension of 5. Since the embedding dimensions obtained by the various methods differ slightly, the range of embedding dimension for ensemble prediction is adopted as 3–10.

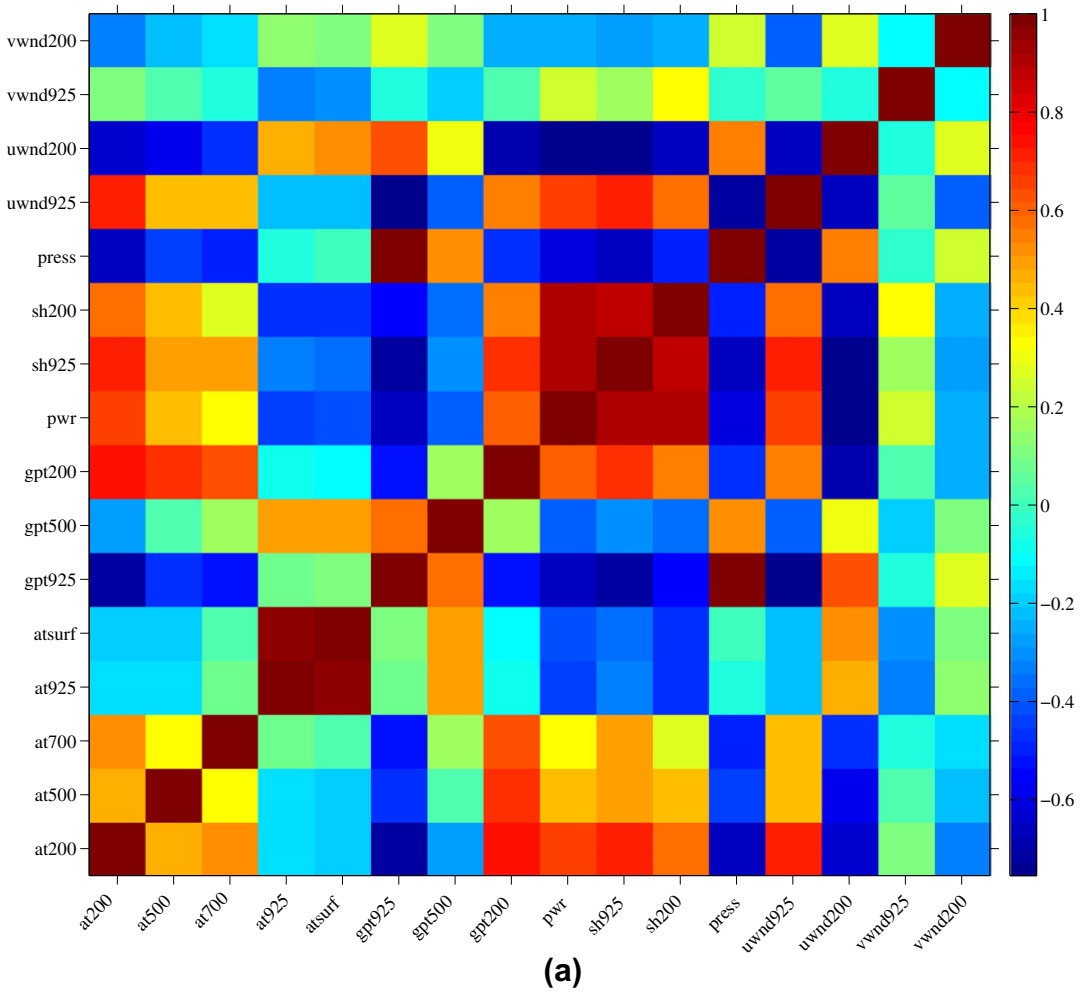


Fig. 7. (a) Correlation map of atmospheric variable data set. (b) Percent variance explained by 16 principal components.

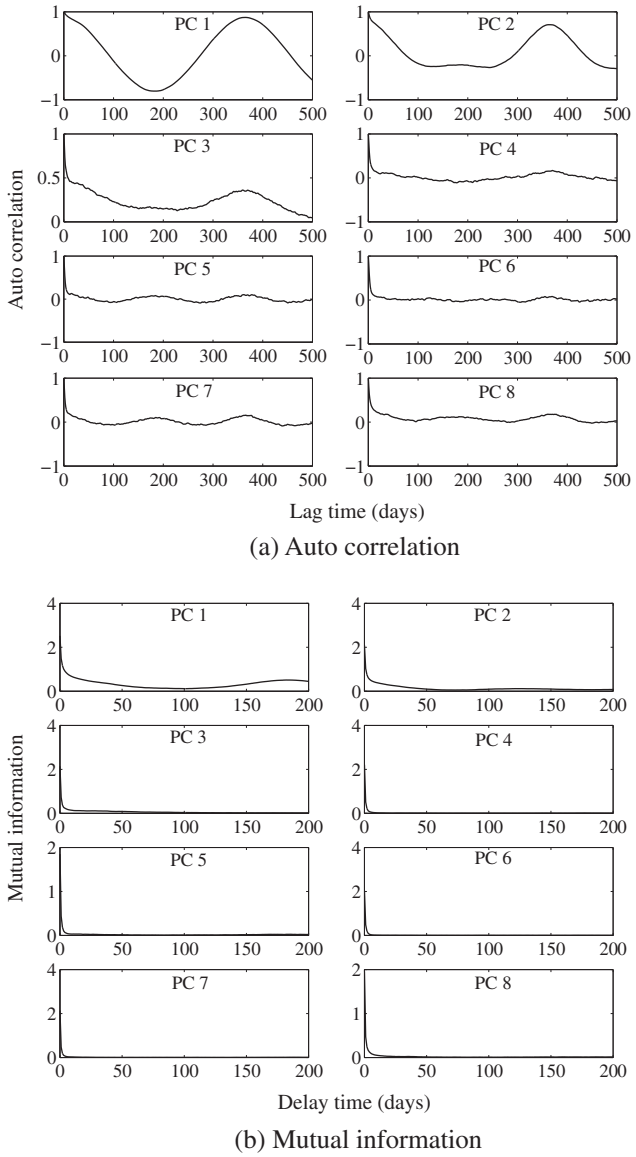


Fig. 8. Variation of (a) autocorrelation and (b) mutual information with lag time for all PCs.

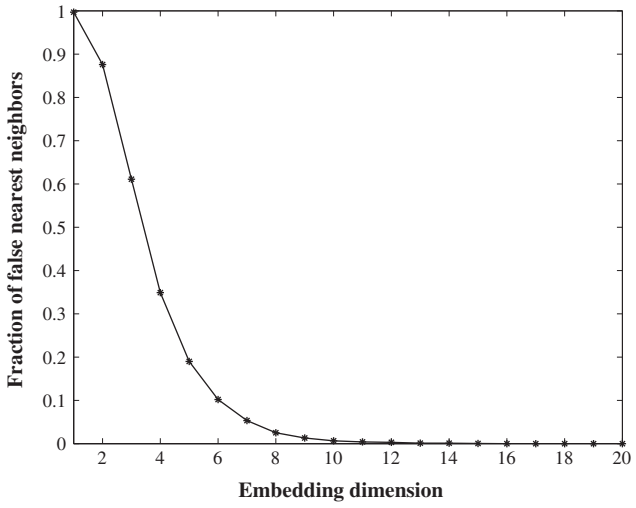


Fig. 9. Variation of fraction of FNN with total embedding dimension.

Table 2
A sample of various combinations of individual embedding dimensions taken for multivariate phase space reconstruction.

Total dimension, *M*	Rainfall	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6	PC 7	PC 8
9	1	1	1	1	1	1	1	1	1
10	2	1	1	1	1	1	1	1	1
11	3	1	1	1	1	1	1	1	1
	2	2	1	1	1	1	1	1	1
12	4	1	1	1	1	1	1	1	1
	3	2	1	1	1	1	1	1	1
	2	2	2	1	1	1	1	1	1

Table 3
Comparison of daily rainfall correlations, average daily rainfall correlations and daily rainfall RMS errors of univariate and multivariate predictions.

Year	Corrln. daily flow		Corrln. mean daily flow		Mean RMSE (mm)	
	Uni	Multi	Uni	Multi	Uni	Multi
1996	0.3539	0.4200	0.7718	0.8451	8.9535	5.9529
1997	0.5729	0.6275	0.8565	0.9169	9.6578	6.8830
1998	0.4752	0.4793	0.9317	0.9498	7.9928	5.2425
1999	0.5129	0.5338	0.8533	0.8600	8.9069	6.5211
2000	0.5032	0.5090	0.9249	0.9683	8.2302	5.7646

4.2. Multivariate series

4.2.1. Analysis of chaotic nature of atmospheric variables

The chaotic dynamics of atmospheric variables is analyzed by false nearest neighbor (FNN) method. The threshold value *f* is fixed at 5. The fall of fraction of FNN with embedding dimension for all variables is shown in Fig. 6. The FNN fraction decreases with an increase in embedding dimension, finally achieving a minimum value (or even zero). The embedding dimension at which the fraction of nearest neighbors is reaching a minimum value is in the range of 8–10 for all variables. This indicates the presence of a low dimensional strange attractor (thus chaos) in all these atmospheric time series. Therefore, an embedding dimension of 8–10 is considered sufficient to explain the dynamics of these variables.

4.2.2. Principal component analysis

The initial dimension of the multivariate time series is 17 (rainfall and 16 atmospheric variables), when all the atmospheric variables are included. This may cause an over embedding while performing phase space reconstruction of the multivariate time series. The 16 atmospheric variables chosen as predictors are also dependent on each other as can be noticed from the correlation map of the variables shown in Fig. 7a. Some of the variables exhibit a high positive correlation, while a few others exhibit a medium negative correlation. Hence, it would be appropriate to reduce the dimension of this correlated atmospheric time series (dimension = 16) to a smaller number of uncorrelated variables called principal components. The principal components of the atmospheric time series of dimension 16 are extracted.

Fig. 7b shows the percentage of variance explained by the principal components. Around 95.1% of the total variability of the data set is explained by the first 8 principal components. Therefore, along with the daily rainfall series, these first 8 PCs are selected for multivariate phase space reconstruction. Thus, the dimension of the multivariate data set is reduced to 9.

4.2.3. Determination of delay time of eight PCs

The delay time of the first variable in the multivariate data set i.e., daily rainfall is determined using autocorrelation and mutual

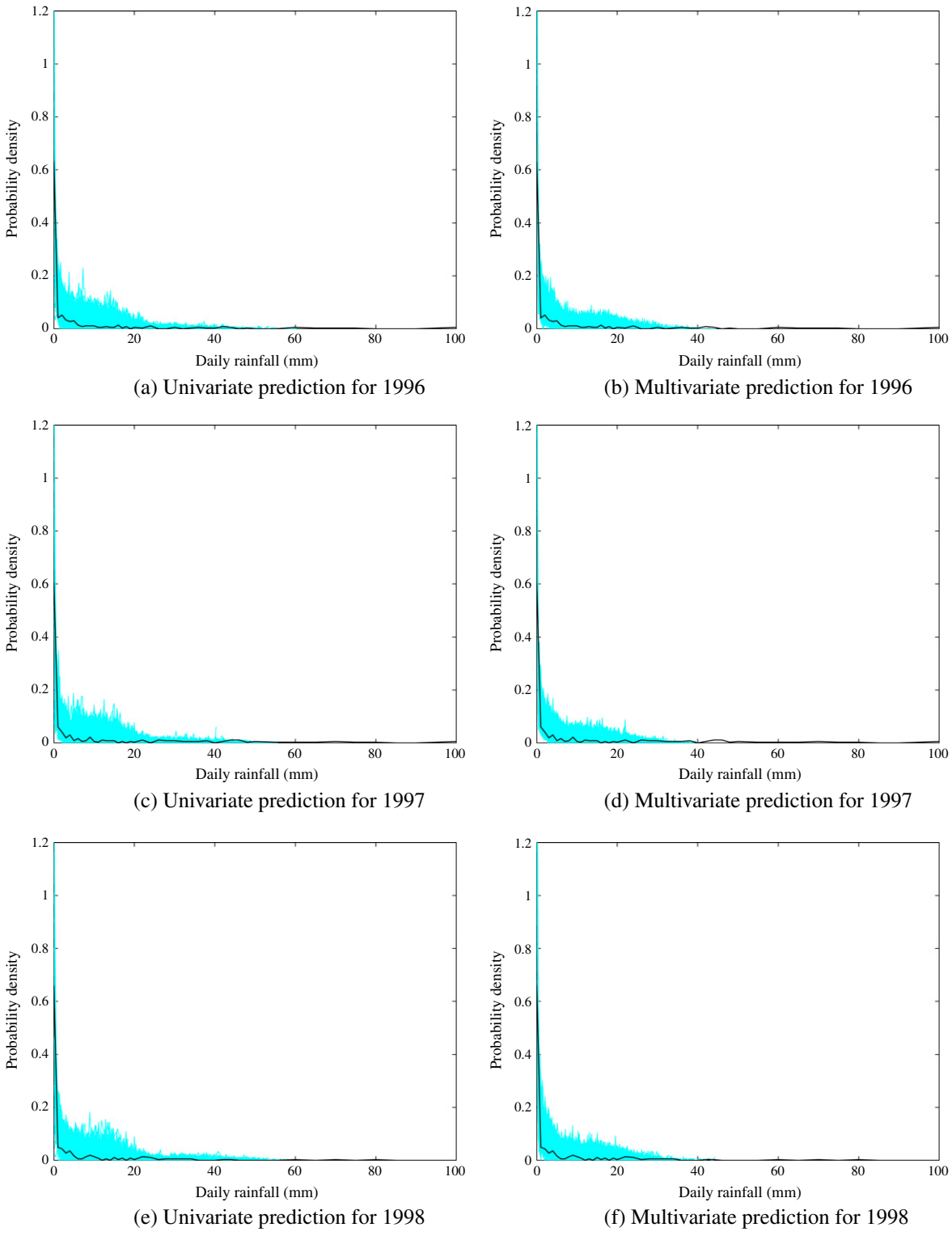
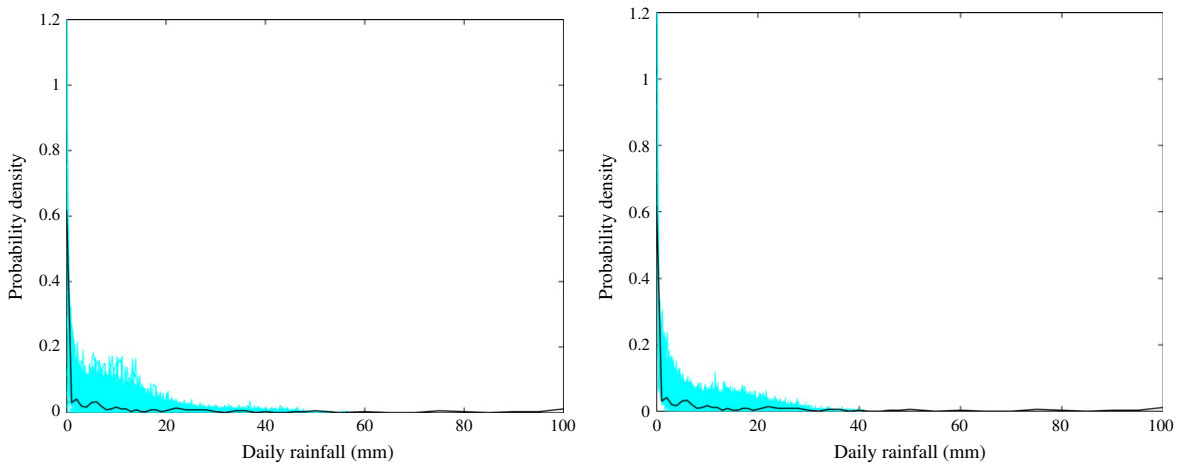


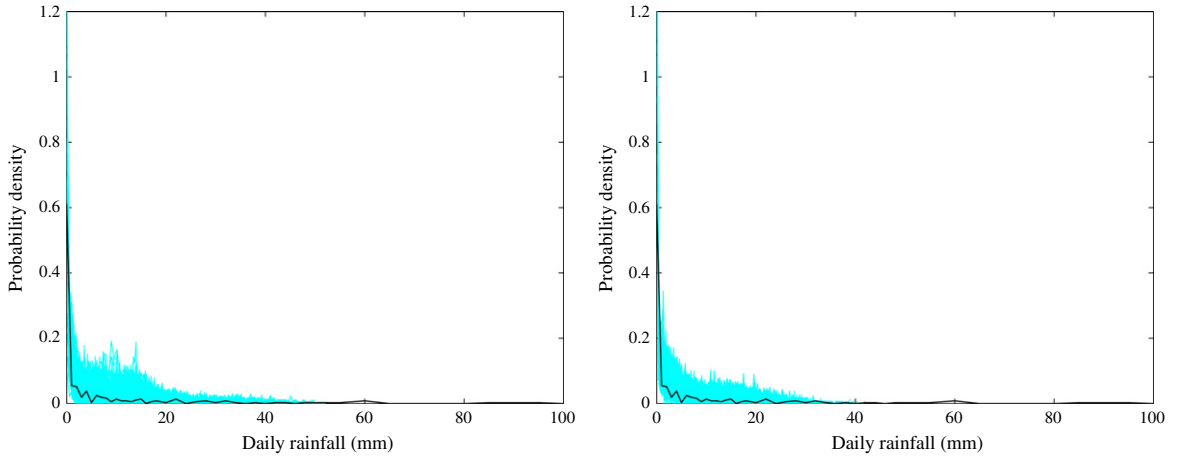
Fig. 10. Comparison of probability density functions of ensembles and the observations for univariate and multivariate cases for all the five years. (a) Univariate prediction for 1996. (b) Multivariate prediction for 1996. (c) Univariate prediction for 1997. (d) Multivariate prediction for 1997. (e) Univariate prediction for 1998. (f) Multivariate prediction for 1998. (g) Univariate prediction for 1999. (h) Multivariate prediction for 1999. (i) Univariate prediction for 2000. (j) Multivariate prediction for 2000.

information methods as explained in Section 3.1. In order to reconstruct the multivariate phase space as given in Eq. (3), the delay times τ_i of rest of the eight variables (eight PCs) need to be determined. The variation of autocorrelation and mutual information with lag time for each series is shown in Fig. 8a and b respectively. While the variation of mutual information over lag time is evident for the first few PCs, the values are either zero or very low for the rest of the PCs after 1–2 days. This is obvious since much of the variance of the data set is captured by first few PCs themselves.

The zero auto correlation and the first minimum mutual information for the first three PCs are in the range of 75–95 days. To avoid more combinations of delay time and also since the delay



(g) Univariate prediction for 1999 (h) Multivariate prediction for 1999



(i) Univariate prediction for 2000 (j) Multivariate prediction for 2000

Fig. 10 (*continued*)

time of rainfall series is similar to that of PCs, daily rainfall and also all the eight PCs are delayed at the same rate for phase space reconstruction. Hence, the range of delay time for the multivariate ensemble prediction is adopted as 75–95 days (with an increment of 5 days).

4.2.4. Determination of dimension of multivariate series (rainfall and 8 PCs)

As explained in Section 3.3, all possible combinations in which M is increased to $M+1$ are considered. The variation of fraction of FNN with total embedding dimension M is shown in Fig. 9. The embedding dimension at which the FNN fraction falls to a zero value is 10. Hence, the minimum total dimension required for the multivariate series can be fixed at 10.

However, the range of total embedding dimension for multivariate ensemble prediction is varied in the range of 9–18. This range is selected since the initial dimension of the data set itself is nine (rainfall and eight PCs) and also from correlation dimension analysis of univariate daily rainfall series (see Fig. 5a) the maximum number of variables required to explain the system dynamics is ≈18. Since the main objective is the prediction of rainfall series only (1st component), instead of trying out all the combinations of individual m_i's which result in the variation of M from 9 to 18, those combinations which satisfy the criteria of $m_1 \geqslant m_2 \geqslant \cdots \geqslant m_p; m_i \geqslant 1 \ \forall i = 1, 2, \ldots p$ and $9 \leqslant \sum_{i=1}^{p} m_p \geqslant 18$ only are considered. Hence, rainfall and the first eight PCs are embedded into a higher dimension (therefore more components in the phase space) when compared to the rest of the PCs. Also, the criterion $m_i \geqslant 1$ ensures the existence of all components in the phase space. Such a priority to the rainfall series and the first eight PCs may help in better capturing the dynamics and hence the unusual variations in the future daily rainfall. A sample of the combinations considered is given in Table 2 for $M = 9$ to $M = 12$. Thus, a total of 95 combinations are considered up to $M = 18$.

4.3. Ensemble prediction

The ranges of parameters considered for univariate and multivariate cases are given below:

Univariate case: (i) Embedding dimension: 3–12 (10 values); (ii) Delay time: 60–100 (41 values). Phase space is reconstructed for all the available parameter combinations, thus generating a total of 410 ensembles.

Multivariate case: (i) Total embedding dimension, M: 9–18. This will lead to 95 combinations of m_i's as sampled in Table 2. (ii) Delay time: 75–95. In order to reduce the total number of ensembles, in this case, the delay times are considered in increments of 5 days only. Hence, the total number of parameter combinations for which phase space is to be reconstructed is 475.

Predictions are done with local approximation method using the above selected parameter combinations for both univariate case and multivariate case. Ensembles are generated for both cases

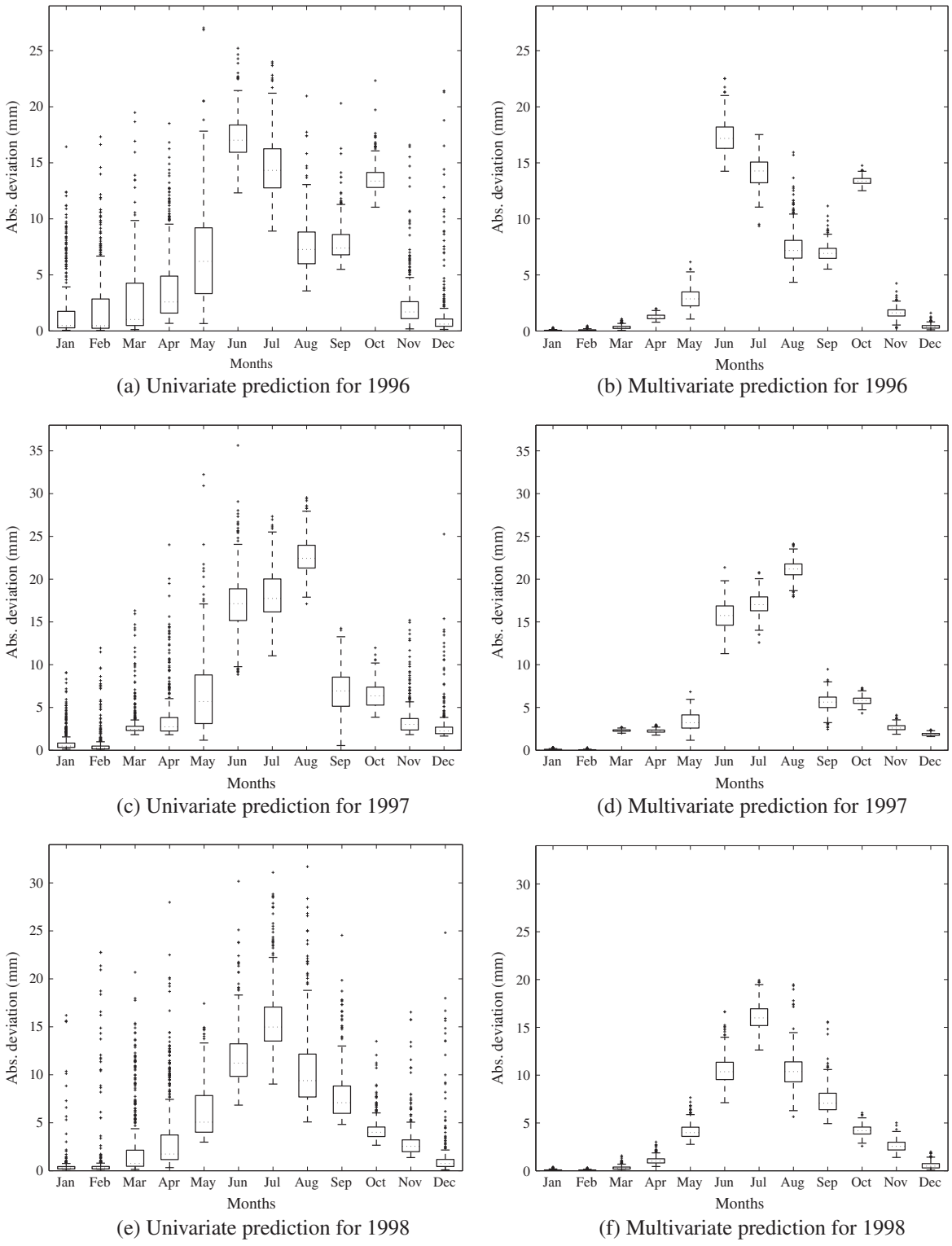
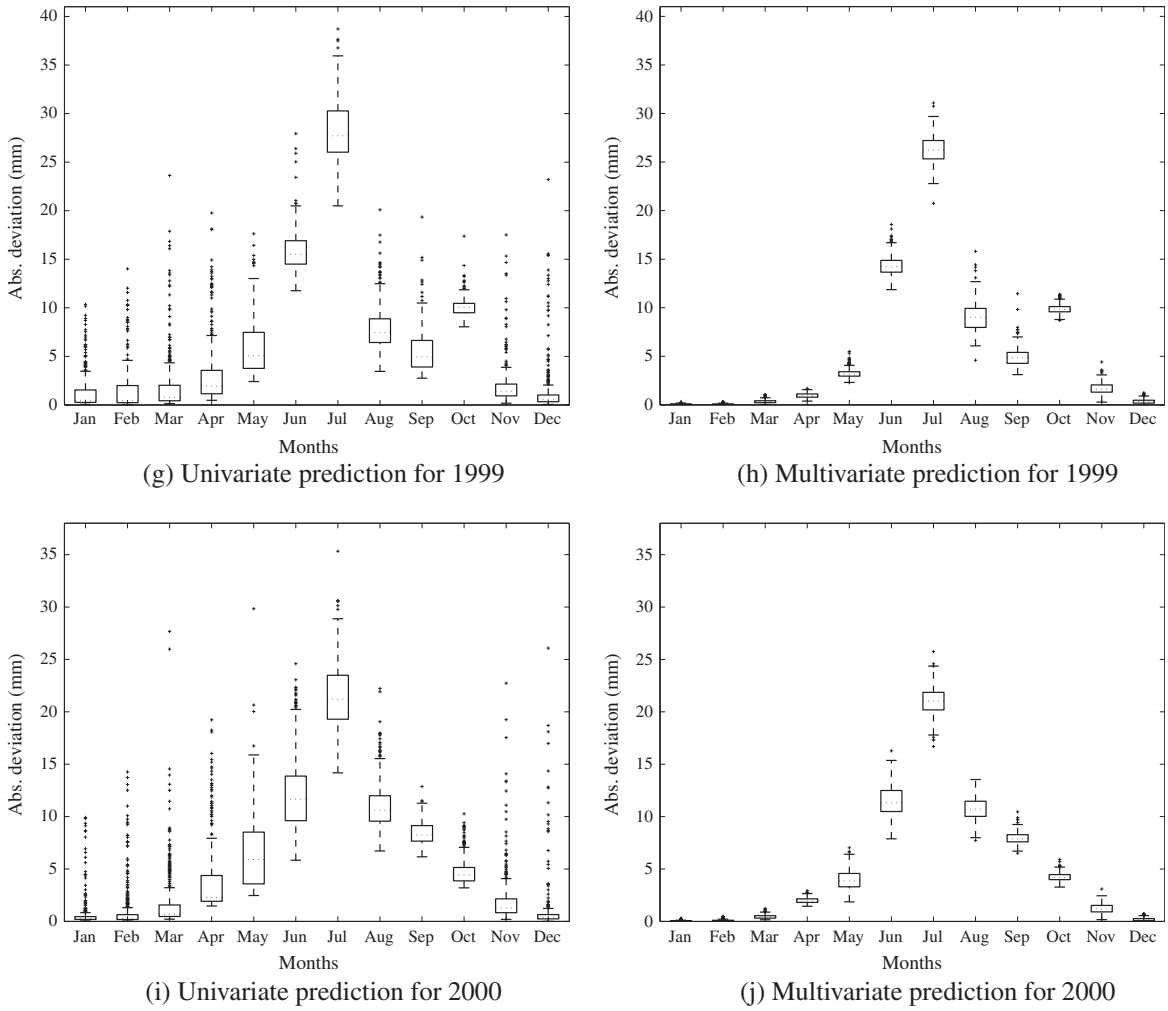


Fig. 11. Box plots of the mean daily rainfall values of the ensembles for both cases. The median ensemble values are shown as a dashed line within each box. (a) Univariate prediction for 1996. (b) Multivariate prediction for 1996. (c) Univariate prediction for 1997. (d) Multivariate prediction for 1997. (e) Univariate prediction for 1998. (f) Multivariate prediction for 1998. (g) Univariate prediction for 1999. (h) Multivariate prediction for 1999. (i) Univariate prediction for 2000. (j) Multivariate prediction for 2000.

for a particular year using the data till the corresponding preceding year. This is done for 5 years from 1996 to 2000. A comparison of the correlations of the observed daily rainfall for each year with the mean ensemble daily rainfall values, correlations with observed average daily rainfall values for a month and mean ensemble average daily rainfall values and root mean square error (RMSE) of the daily rainfall for both univariate and multivariate predictions are shown in Table 3. Multivariate prediction ensembles show higher correlations for both daily rainfall and mean daily rainfall of all the five years. Similarly, RMS error is always less



(g) Univariate prediction for 1999

(h) Multivariate prediction for 1999

(i) Univariate prediction for 2000

(j) Multivariate prediction for 2000

Fig. 11 (*continued*)

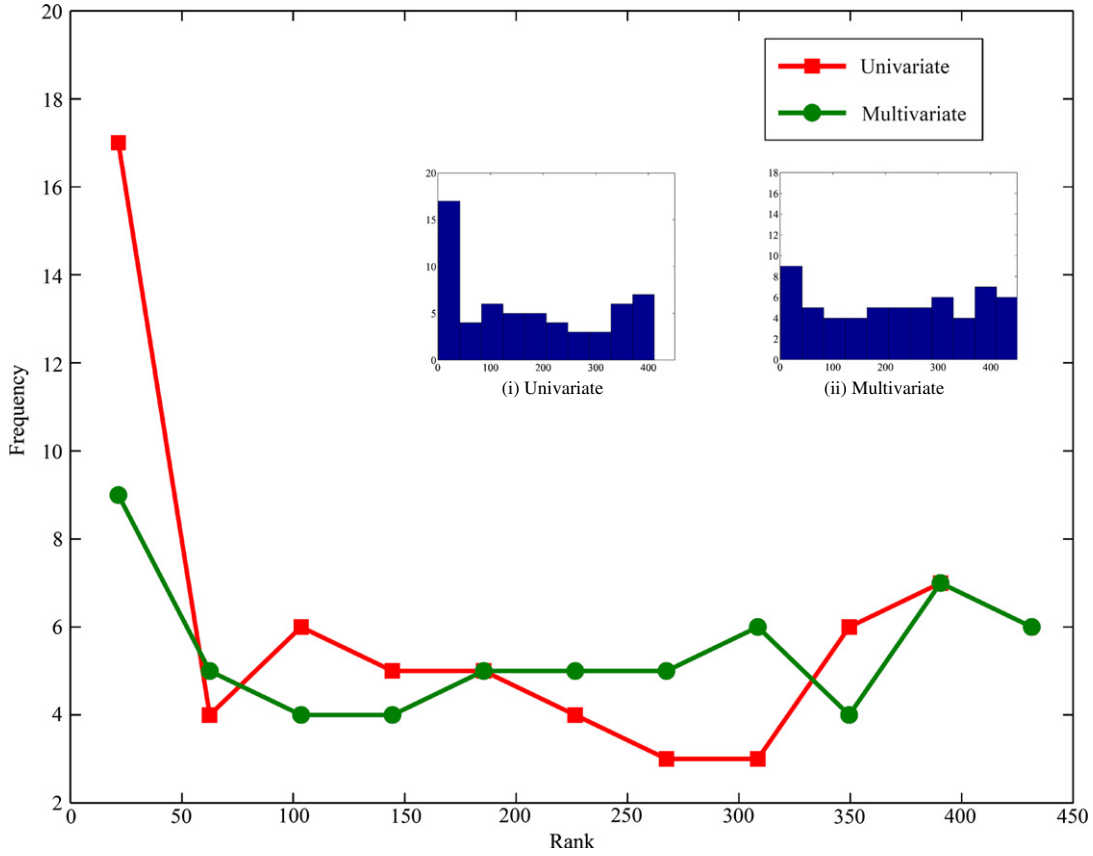
for multivariate case. This indicates that multivariate phase space reconstruction is able to effectively reveal the dynamics of the daily rainfall system.

The probability density functions (PDF) of the ensembles (in blue) and the corresponding observed series (solid black line) for both univariate and multivariate cases for all the 5 years are shown in Fig. 10a–j. Even though both the predictions are able to catch the observed series PDF within its spread, the width (probability density) of the PDF is less for the multivariate case for all the years considered. This in turn indicates a reduction in the uncertainty caused due to the indefiniteness of the reconstructed parameters.

A detailed analysis is done by constructing the box plots of absolute deviations of average daily ensemble rainfall from the corresponding average daily observed rainfall values. The box plots of these absolute deviations of average daily rainfall values for the 5 years are shown in Fig. 11a–j. The box plots give the range of deviations of the ensembles from the observed. A box in the box plots indicates the inter-quartile range of the absolute deviations and the horizontal dashed line within the box indicates the median of the absolute deviations. 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 box plots are shown for both univariate and multivariate ensembles. The box plots (or the range of the absolute deviation) of univariate predictions are much wider than those of multivariate predictions. This further confirms the reduction in uncertainty by multivariate representation and hence an increase in the predictability of the series. It can be seen that the predictions obtained from multivariate case are significantly better than those obtained from univariate series.

The quality of the ensembles generated using both approaches are further compared using rank histograms. The rank histograms for evaluating the mean and spread of the ensembles (410 for Univariate and 475 for Multivariate) generated for mean daily rainfall over a month for each year (a total of $12 \times 5 = 60$ data points) are prepared and are compared in Fig. 12a. The rank histograms for both cases are given as insets. There is an ensemble bias which causes more population towards the lower ranks in univariate prediction. Such a bias is due to the inability of univariate representation to capture too low rainfall values. The rank histograms of multivariate ensembles are almost flat implying ensembles of reliable spread.

Further, the time evolution of average root mean square error (RMSE) of all ensembles for 5 years is computed and compared for both cases in Fig. 12b. The mean RMSE for multivariate case is always lesser than the corresponding univariate value. Such reduction in RMSE establishes the increase in predictability due to the multivariate approach. This further proves that even though daily rainfall series is chaotic and hence limitedly predictable, the predictability can be significantly improved by utilizing information from the influencing atmospheric variables.



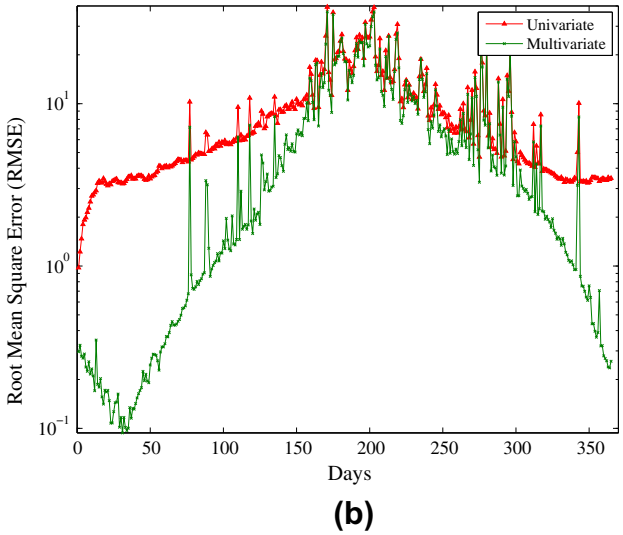


Fig. 12. (a) Rank plots for mean daily rainfall ensembles. Rank plot for univariate and multivariate predictions are shown. The original rank histograms are shown in insets (the axes labels are same as those of the main figure). (b) Evolution of mean RMSE over days for univariate and multivariate case.

5. Conclusions

The unexpected exponential growth due to the infinitesimal uncertainty in the initial conditions limits the predictability of any chaotic series. Many atmospheric variables, wind patterns, rainfall etc have been proven to exhibit such a sensitive dependence on initial conditions. In a chaotic system, predictability is limited due to the uncertainty in initial conditions and also due to the uncertainty in the model dynamics. The present study was aimed at quantifying these uncertainties by adopting a multivariate nonlinear ensemble prediction. The uncertainty in initial conditions was quantified by reconstructing the phase space for different combinations of parameters (embedding dimension and delay time). This will ultimately generate a set of attractors, which in turn will lead to a change in the initial condition. The model dynamics was further modified by exploiting information from other influencing variables and reconstructing a multivariate phase space, since it may contain more information about the dynamic system.

Daily rainfall data for the period 1955–2000 of Malaprabha was considered for the study. This series was found to exhibit a low dimensional chaotic behavior with an embedding dimension 5–7. For multivariate prediction, 16 atmospheric variables were considered. The chaotic nature of these atmospheric variables was confirmed by employing false nearest neighbor method. Since these variables may be interdependent and hence may contain redundant information, the dimension of the atmospheric data set was reduced to 8 by retaining the first eight uncorrelated principal components (PCs) that capture around 95% variance of the original data set. The daily rainfall series along with these eight PCs were considered for multivariate phase space reconstruction and prediction.

An ensemble of predictions was generated using a range of parameters (embedding dimension and delay time) for both univariate series and multivariate series. For univariate series, the embedding dimension was varied from 3 to 12 and delay time was varied from 60 to 100. For multivariate series (initial dimension = 9), the total embedding dimension was varied from 9 to 18 and delay time was varied from 75 to 95. Selected combinations of parameters were only considered for multivariate case, so that priority can be given to rainfall series followed by the PCs catching large variance. Ensembles were generated for 5 years from 1996, considering data till the corresponding previous year.

A comparison of the ensembles generated by both approaches revealed that the uncertainty in the multivariate ensembles is significantly small compared to that in the univariate ensembles. The probability density functions (PDFs) from both approaches were able to capture the observed PDF in its spread. The reduction in the uncertainty using multivariate approach was evident from the reduced width of absolute deviation boxes. The reliability of multivariate ensembles was also noticeable from the rank histograms. A comparison of the evolution of root mean square error also revealed the superiority of the multivariate ensembles and its increased predictability.

The uncertainty in the initial conditions of a chaotic system was hence quantified by making predictions from different initial conditions (different parameter combinations). Better predictions obtained from multivariate nonlinear prediction method confirm the suitability of modeling and understanding the underlying dynamics of the complex rainfall process by utilizing information from the causative atmospheric variables. It is worthwhile to note that even for a chaotic system, predictability can be improved by efficaciously quantifying the underlying uncertainties.

References

Anandhi, A., Srinivas, V.V., Nanjudaiah, R., Kumar, D.N., 2008. Downscaling precipitation to river basin in India for IPCC SRES scenarios using support vector machine. Int. J. Climatol. 28, 401–420.

Cao, L., Mees, A., Judd, K., 1998. Dynamics from multivariate time series. Physica D 121, 75–88.

Dhanya, C.T., Kumar, D.N., 2010. Nonlinear ensemble prediction of chaotic daily rainfall. Adv. Water Res. 33, 327–347.

Dhanya, C.T., Kumar, D.N., 2011. Predictive uncertainty of chaotic daily streamflow using ensemble wavelet networks approach, Water Resour. Res. doi:10.1029/2010WR010173.

Elshorbagy, A., Simonovic, S.P., Panu, U.S., 2002. Noise reduction in chaotic hydrologic time series: facts and doubts. J. Hydrol. 256 (3/4), 147–165.

Farmer, J.D., Sidorowich, J.J., 1987. Predicting chaotic time series. Phys. Rev. Lett. 59, 845–848.

Frazer, A.M., Swinney, H.L., 1986. Independent coordinates for strange attractors from mutual information. Phys. Rev. A 33 (2), 1134–1140.

Grassberger, P., Procaccia, I., 1983. Measuring the strangeness of strange attractors. Physica D 9, 189–208.

Hegger, R., Kantz, H., 1999. Improved false nearest neighbor method to detect determinism in time series data. Phys. Rev. E 60, 4970–4973.

Islam, M.N., Sivakumar, B., 2002. Characterization and prediction of runoff dynamics: a nonlinear dynamical view. Adv. Water Resour. 25, 179–190.

Jayawardena, A.W., Gurung, A.B., 2000. Noise reduction and prediction of hydrometeorological time series: dynamical systems approach vs. stochastic approach. J. Hydrol. 228, 242–264.

Jayawardena, A.W., Lai, F., 1994. Analysis and prediction of chaos in rainfall and stream flow time series. J. Hydrol. 153, 23–52.

Jin, Y.H., Kawamura, A., Jinno, K., Berndtsson, R., 2005. Nonlinear multivariable analysis of SOI and local precipitation and temperature. Nonlinear Process. Geophys. 12, 67–74.

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., Zhu, Y., Leetmaa, A., Reynolds, R., Chelliah, M., Ebisuzaki, W., Higgins, W., Janowiak, J., Mo, K.C., Ropelewski, C., Wang, J., Jenne, R., Joseph, D., 1996. The NCEP/NCAR 40-year reanalysis project. Bull. Am. Meteor. Soc. 77, 437–470.

Kantz, H., Schreiber, T., 2004. Nonlinear Time Series Analysis, second ed. Cambridge University Press, Cambridge, UK.

Kennel, M.B., Brown, R., Abarbanel, H.D.I., 1992. Determining embedding dimension for phase space reconstruction using a geometric method. Phys. Rev. A 45, 3403–3411.

Liu, Q., Islam, S., Rodriguez-lturbe, I., Le, Y., 1998. Phase-space analysis of daily streamflow: characterization and prediction. Adv. Water Resour. 21, 463–475.

Lorenz, E.N., 1963. Deterministic nonperiodic flow. J. Atmos. Sci. 20, 130–141.

Lorenz, E.N., 1972. Predictability: does the Flap of a Butterfly's Wings in Brazil set off a Tornado in Texas? AAAS Section on Environmental Sciences New Approached to Global Weather: GARP (The Global Atmospheric research Program), 139th Meeting, American Association for the Advancement of Science.

Osborne, A.R., Provenzale, A., 1989. Finite correlation dimension for stochastic systems with power law spectra. Physica D 35, 357–381.

Porporato, A., Ridolfi, L., 1996. Clues to the existence of deterministic chaos in river flow. Int. J. Mod. Phys. B 10 (15), 1821–1862.

Porporato, A., Ridolfi, L., 1997. Nonlinear analysis of river flow time sequences. Water Resour. Res. 33 (6), 1353–1367.

Porporato, A., Ridolfi, L., 2001. Multivariate nonlinear prediction of riverflow. J. Hydrol. 248, 109–122.

Puente, C.E., Obregon, N., 1996. A deterministic geometric representation of temporal rainfall: results for a storm in Boston. Water Resour. Res. 32 (9), 2825–2839.

Rajeevan, M., Bhate, J., Kale, J.D., Lal, B., 2006. High resolution daily gridded rainfall data for the Indian region: analysis of break and active monsoon spells. Curr. Sci. 91, 296–306.

Rodriguez-lturbe, I., De Power, F.B., Sharifi, M.B., Georgakakos, K.P., 1989. Chaos in rainfall. Water Resour. Res. 25 (7), 1667–1675.

Rosenstein, M.T., Collins, J.J., De Luca, C.J., 1993. A practical method for calculating largest Lyapunov exponents from small data sets. Physica D 65, 117–134.

Sangoyomi, T., Lall, U., Abarbanel, H.D.J., 1996. Nonlinear dynamics of the Great Salt Lake: dimension estimation. Water Resour. Res. 32 (1), 149–159.

Schreiber, T., 1993. Extremely simple nonlinear noise reduction method. Phys. Rev. E. 47 (4), 2401–2404.

Shang, P., Xu, N., Kamae, S., 2009. Chaotic analysis of time series in the sediment transport phenomenon. Chaos, Solitons & Fractals 41 (1), 368–379. doi:10.1016/j.chaos.2008.01.014.

Sivakumar, B., 2001. Rainfall dynamics at different temporal scales: a chaotic perspective. Hydrol. Earth System Sci. 5 (4), 645–651.

Sivakumar, B., Liong, S.Y., Liaw, C.Y., Phoon, K.K., 1999. Singapore rainfall behavior: chaotic? J. Hydrol. Eng. 4 (1), 38–48.

Sivakumar, B., Berndtsson, R., Olsson, J., Jinn, K., 2001. Evidence of chaos in the rainfall–runoff process. Hydrol. Sci. J. 46 (1), 131–145.

Smith, L.A., 2000. Disentangling uncertainty and error: on the predictability of nonlinear systems. In: Mees, A. (Ed.), Nonlinear Dynamics and Statistics. Birkhauser, pp. 31–64.

Smith, L.A., Ziehmann, C., Fraedrich, K., 1998. Uncertainty dynamics and predictability in chaotic systems. Q.J.R. Meteorol. Soc. 125, 2855–2886.

Takens, F., 1981. Detecting strange attractors in turbulence. In: Rand, D.A., Young, L.S. (Eds.), Lectures Notes in Mathematics, vol. 898. Springer-Verlag, Berlin, Germany.

Velichov, S., 2004. Nonlinear Dynamics and Chaos: with Applications to Hydrodynamic and Hydrological Modeling. Taylor & Francis.

Vlachos, I., Kugiumtzis, D., 2008. State space reconstruction for multivariate time series prediction. Nonlinear phenomena in complex systems 11, 241–249.

Wang, Q., Gan, T.Y., 1998. Biases of correlation dimension estimates of streamflow data in the Canadian prairies. Water Resour. Res. 34 (9), 2329–2339.

Wolf, A., Swift, J.B., Swinney, H.L., Vastano, A., 1985. Determining Lyapunov exponents from a time series. Physica D 16, 285–317.