CrossMark

ORIGINAL ARTICLE

Analysing Predictability in Indian Monsoon Rainfall: A Data Analytic Approach

Sarita Azad[1] · **Shoubhik Debnath**[1] · **M. Rajeevan**[2]

Abstract This paper examines monthly and annual data to analyse predictability in the Indian monsoon rainfall. The periodic structure in the time series data is extracted using wavelets and the residual random part is separately modeled using artificial neural networks (ANN). Although wavelet and neural network based hybrid techniques have been widely applied in the recent years, the present approach has not been investigated so far. Our results show that the estimated periodic and random components comprise 30 and 15 %, respectively, variance of the total rainfall in case of annual data, whereas the model explains 93 % of variance in case of monthly data. It is shown that the prediction is more accurate when periodic and random parts are treated separately.

Keywords Indian monsoon · Prediction · Time series · Wavelets · Artificial neural networks

1 Introduction

The prediction of Indian monsoon has been a challenge for scientists around the globe. In this undertaking, statistical and empirical models seek to establish significant relationships among various climate parameters (Rajeevan et al. 2000, 2004, 2007; Narasimha and Bhattacharyya 2010). It is well known that the presence of periodic signals can make the dynamic system predictable with empirical models (Shukla and Mooley 1987; Vautard et al. 1992). These models make use of historical data to make predictions; however, they have been proved generally unsuccessful for operational forecasts. There have been many attempts to find the most appropriate method for rainfall prediction, and several new methods have been extensively employed in the recent past, for example: wavelet transform (Sahay and Srivastava

✉ Sarita Azad
  sarita@iitmandi.ac.in

[1] Indian Institute of Technology Mandi, Mandi 75001 Himachal Pradesh, India

[2] Indian Institute of Tropical Meteorology, Dr Homi Bhabha Road, Pashan, Pune 411 008, India

 Springer

2014; Maheswaran and Khosa 2014; Sehgal et al. 2014); coupled wavelet-neural network models (Ramana et al. 2013); genetic algorithms (Kishtawal et al. 2003); empirical mode decomposition (Iyengar and Kanth 2004); and uncertainty analysis (Narsimlu et al. 2015), among others. Although, in meteorology, data analytic studies of historical data sets have been traditionally very useful, the most obvious reason for the failure of empirical prediction is the stochastic nature of the rainfall data. Although the presence of quasi-periodicities indicate certain potential predictability, it is difficult to distinguish a periodic signal from the fluctuations that would be inherent in the relatively small sample size available (of order 10^2). To detect periodicities in the data, various methods have been employed, and a wide variety of periodicities have been extracted from the time series of rainfall data. However, lack of statistically significant periodicities could be the cause of failure of empirical prediction (Rangarajan 1994; Narasimha and Kailas 2001).

In this connection, the most popular background spectrum to distinguish noise from signal was introduced in Gilman et al. (1963). In Azad and Narasimha (2008), this background spectrum is used as a coloured spectrum and has revealed many more significant periodicities. Based on the significant periodicities, Annamalai (1995) pointed that about 34 % of the rainfall is deterministically predictable but the most dominant modes of All-India summer monsoon rainfall (AISMR) and many of its predictors are associated with the stochastic components. For long range forecasting, methods of fitting linear autoregressive models are suggested in Yao (1983).

Rajkumar and Kumar (2004) developed the stochastic modeling of daily rainfall for southwest monsoon season of Baptala, Andhra Pradesh (India) for the period 1988–2000. The periodic component in the time series was determined by Fourier analysis, and the stochastic component was handled using autoregressive model of order two. Bhakar et al. (2006) carried out a study on stochastic modeling of wind speed at Udaipur (India) and its periodic component is represented by second harmonic expression and the stochastic component followed a second order Markov model. Most recently, stochastic modeling of rainfall in humid region of Northeast India was performed in Dabral et al. (2008). Keppene and Ghil (1992) have shown that an empirical prediction can be improved by separating the deterministic oscillations from noise in the original time series data.

Kishtawal et al. (2003) evoked the feasibility of a nonlinear technique based on genetic algorithm (Artificial Intelligence) for the prediction of summer rainfall over India. Guhathakurta (2006) introduced ANN to forecast the summer-monsoon rainfall over the Kerala state in India. Various workers utilized ANN as a forecasting tool involving atmosphere related phenomena. Michaelides et al. (2001) compared the performance of ANN with multiple linear regressions in estimating missing rainfall data over Cyprus. By implementing ANN, reconstruction of the rainfall time series over Cyprus has been carried out. Moreover, the method is used in rainfall prediction by splitting available data into homogeneous subpopulation.

2 Data and Method

Looking at the present scenario, we attempted a novel forecasting strategy which separately handles periodic and stochastic component in the time series data. The Indian time series data used in many of the prediction studies are defined over homogeneous regions (Parthasarathy et al. 1995). However, Azad et al. (2010) argued that though used popularly, these regions are
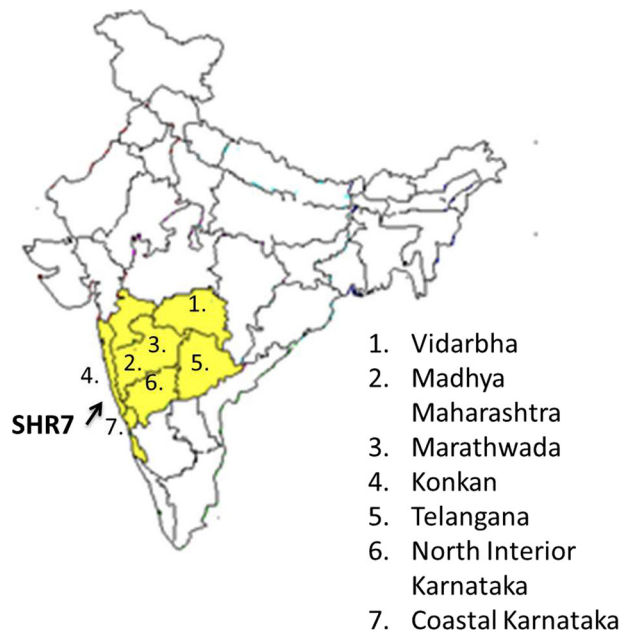
not the good choice. The data samples here are mixed up all over the heterogeneous zones. Hence, authors suggested an alternative approach to determine homogeneous rainfall zones. The most significant of such spectrally homogeneous regions (SHRs) is SHR7, a region having characteristics of southwest monsoon. Even though magnitude of rainfall varies across SHR7 between 600 and 3200 mm, this region is homogenous in terms of spectral similarity not in terms of rainfall. The observed monthly and annual rainfall data have been taken from the website of the Indian Institute of Tropical Meteorology (http://www.tropmet.res.in). We have used sub-divisional data which is in mm units. The time series of SHR7 rainfall is obtained from the area weighted average of seven sub-divisions namely: Coastal Karnataka; Konkan; Madhya Maharashtra; North interior Karnataka; Marathwada; Telangana; and Vidarbha. The SHR7 region is shown in Fig. 1. The autocorrelation coefficient of SHR7 rainfall is 0.1 at lag 1, whereas for Homogeneous Indian monsoon (HIM) rainfall (most commonly used homogeneous region) it is −0.007, which shows HIM data is random. The SHR7 data does not follow Gaussian distribution unlike HIM. Therefore, we envisaged that there is more predictability in SHR7 rainfall than HIM (which contains 14 meteorological sub-divisions) or any other rainfall indices like AISMR (autocorrelation −0.001 at lag 1). Hence, we performed our analysis on this zone over the time period 1871–2005. This work provides insight into long-range forecasting of Indian monsoon as this new region has not been analyzed for prediction.
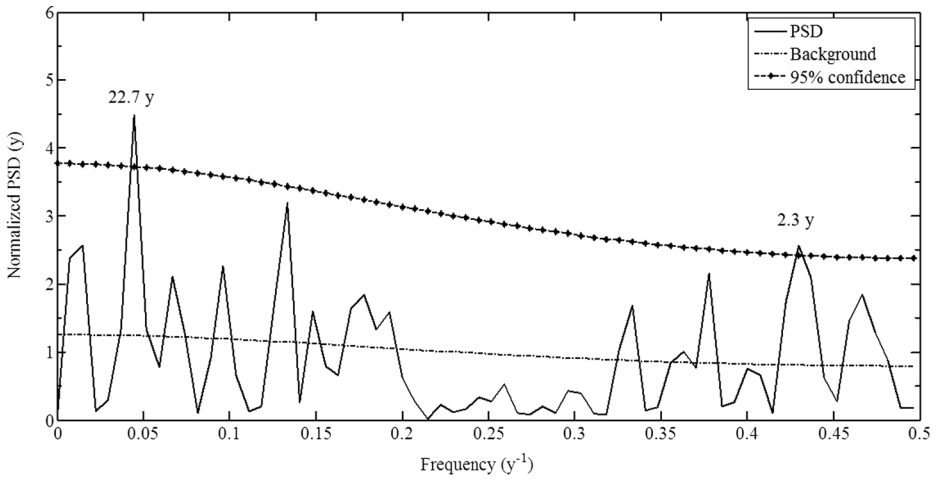
In the present study, the periodic part of the signal is extracted using wavelet based multi-resolution analysis (MRA). The remaining stochastic part is modelled using artificial neural networks as discussed below.

2.1 Wavelet Method to Separate the Periodic Part

We analyse the periodic structure in the data with wavelets which is based on MRA decomposition of the signal into a range of frequency scales (Azad and Narasimha 2007,

Fig. 1 The shaded area shows the seven sub-divisions constituting SHR7 region

1. Vidarbha
2. Madhya Maharashtra
3. Marathwada
4. Konkan
5. Telangana
6. North Interior Karnataka
7. Coastal Karnataka

Fig. 2 The estimated spectrum of SHR7 rainfall obtained from the Welch technique compared with the background spectrum at 95 % confidence levels

2008). The prediction uses the reconstructed time series at each scale. A redundant transform based on an N-length input series has an N-length resolution scale for each of the resolution levels that we consider. The MRA method described is applied to the normalized time series of SHR7 rainfall x(t), t =1871,...,2005. Using this technique the time series is decomposed at eight levels using discrete Meyer wavelet. However, the first four details and the fourth approximation are sufficient for the present analysis, as the original time series can be reconstructed by decomposition up to four levels with an error of only 10^{-6}. The power spectral density (PSD) of the partially reconstructed time series obtained from the MRA decomposition at four scales are estimated using Welch technique (Stoica and Mosses 1997), as described in Azad and Narasimha (2008).

Therefore, at each level of MRA decomposition, reconstructed time series is represented with periodic component of cosine functions as

$$\hat{x}(t) = \sum_{i=1}^{5} a_i \cos(2\pi f_i t + \varphi_i) \tag{1}$$

where $1/f_i$ are the periodicities obtained from PSD on MRA at each scale. The amplitudes a_i and phases ϕ_i are then estimated using least square estimation at each scale.

	95 % significant periods (y) Using MRA+PSD	95 % significant periods (y) direct PSD
d1	2.6, 3.0	2.3, 22.7
d2	7.5, 6.7, 5.6	
d3	13.5, 10.4	
d4	23	
a4	66.6, 33.3	

Table 1 The significant periods at different scales of MRA decomposition of SHR7 rainfall for the period 1871–2005. The wavelet details are represented by d1–d4, and a4 is the fourth approximation

Table 2 Performance of different methods in SHR7 rainfall prediction for the time period 1871–2005

|	Error in Prediction Using MRA+PSD	Error in Prediction Using direct PSD
EV	0.3	0.05
CC	0.52	0.11
RMSE	0.03	0.9

2.2 Artificial Neural Networks

Modeling the stochastic process requires a technique like ANN which is capable of modeling highly non-linear relationships without any prior assumption and can be trained to generalize accurately when presented with a new data set.

There are several neural network training algorithms. We employ the Back-Propagation (BP) multilayer feedforward neural network algorithm in our study. BP has three major steps: Feedforward of input patterns, backpropagation of calculated error and weights adjustment. The model parameters are listed as follows: for annual data from 1871 to 2005, the training phase is from 1871 to 1980 which mainly works on adjusting the weights of the neural network, so that network error, as computed from the difference between network output and the required output, is minimized. And the Back-propagation algorithm does it through gradient descent minimization of square of network error. The testing phase is from 1981 to
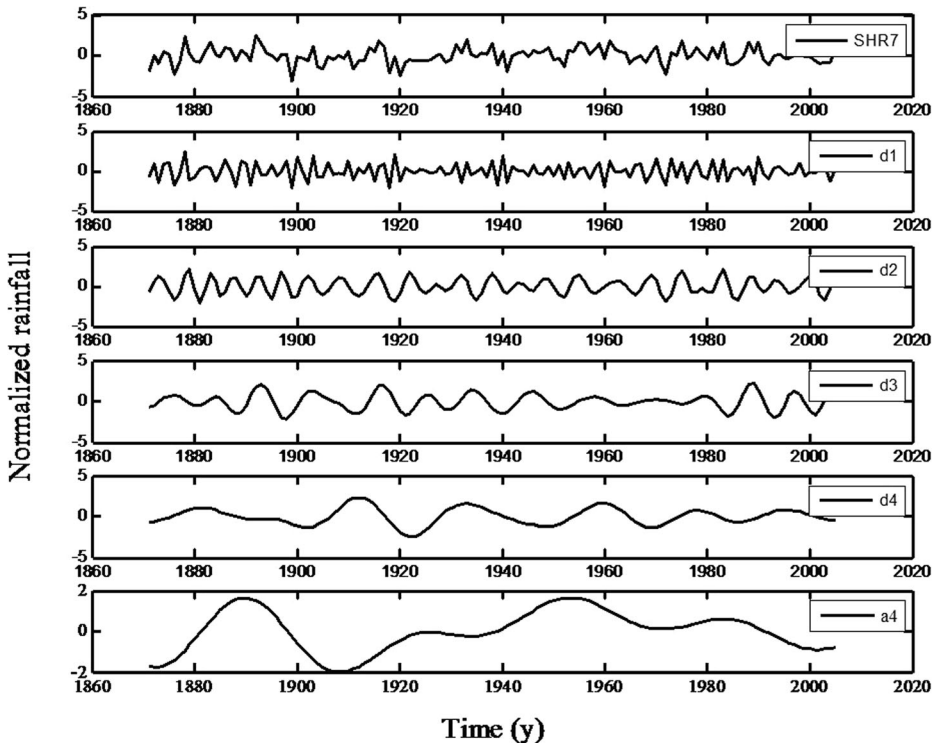
Fig. 3 MRA of SHR7 rainfall: **a** first reconstructed detail $d1$; **b** second reconstructed detail $d2$; **c** third reconstructed detail $d3$; **d** fourth reconstructed detail $d4$; **e** fourth reconstructed approximation $a4$
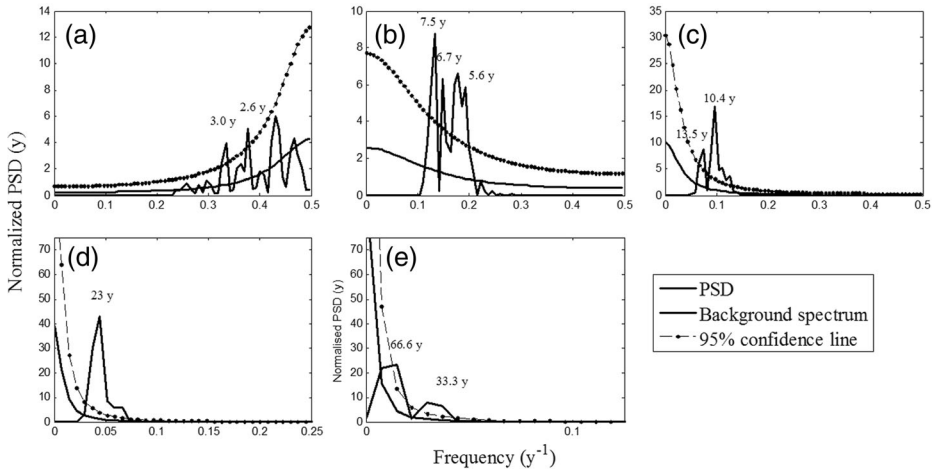
Fig. 4 Significance test on MRA of SHR7 rainfall: **a** PSD on its 1st reconstructed detail $d1$; **b** PSD on its 2nd reconstructed detail $d2$; **c** PSD on its 3rd reconstructed detail $d3$; **d** PSD on its 4th reconstructed detail $d4$; **e** PSD on its fourth reconstructed approximation $a4$. Periods (y) are *marked* at each peak

2005 which mainly validates the neural network model obtained from training phase on the given SHR7 data. Finally, the Multi-Layer Perceptron used contained a single hidden layer having 9 input neurons, 4 hidden neurons and 1 output neuron. Since it is a time-series data, the nine input neurons contain the data for the past 9 years. The 1 output neuron represents the yearly data for the tenth year. For example, if the input neurons use the yearly data for 9 years from 1991 to 1999 for training of the model, then the output neuron return the data for the year 2000.

With the help of NNSYSID toolbox in *Matlab*, the neural network is implemented on the monthly data from 1871 to 2011. The Neural Network AR model is identified using the *'nnarx'* function in the toolbox to train the network from January 1871–December 1983. For determining the non-linear AR model, the *Levenberg-Marquardt* method is used for implementation. The testing phase, which involved the built-in function *'nnvalid'*, validated the identified Neural Network AR model on the dataset from January 1984–December 2011. The

Fig. 5 Observed and fitted periodic component for the time period 1871–2005 and predicted for the time period 2006–2011
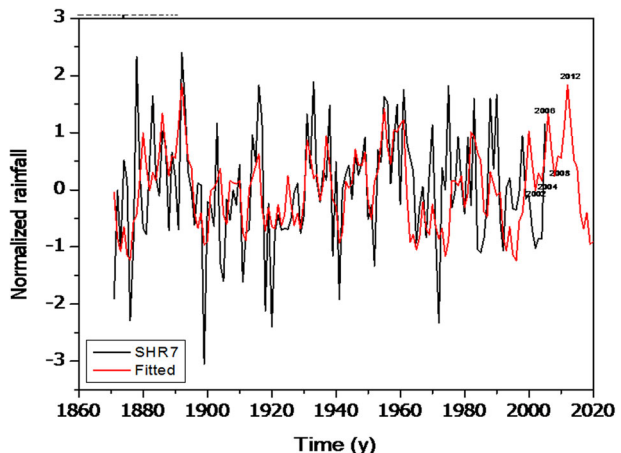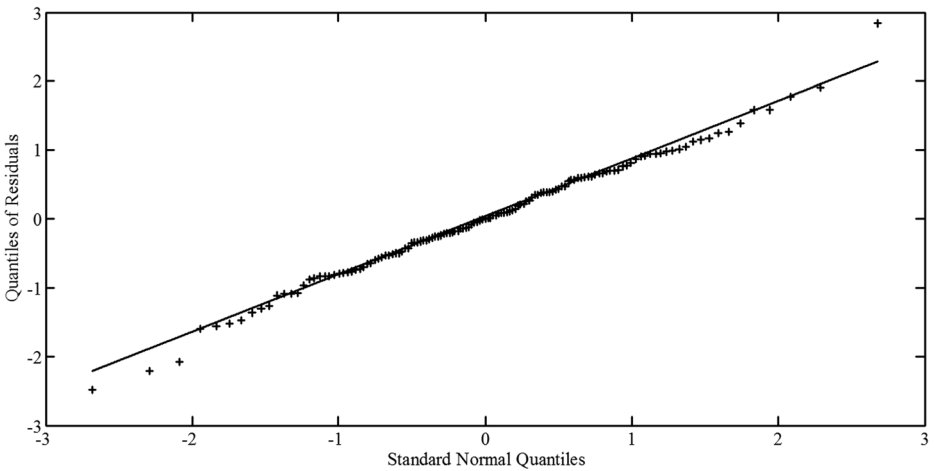
Fig. 6 Probability plot of the stochastic component

Neural Network AR model had *9* hyperbolic tangent units in the hidden layer with a lag value of *5* and *1 output* neuron.

2.3 Evaluation Criteria

To determine model fitness, we use three criteria: Root Mean Squared Error (RMSE), Correlation Coefficient (CC), and Explained Variance (EV). These are commonly used statistics to evaluate the performance of models. RMSE is defined as the square root of the mean squared difference between actual and predicted values. RMSE is often used in conjunction with Standard Deviation. RMSE value less than Standard Deviation shows good model fitness. CC is a measure of closeness between two variables. CC values which are closer to +1 and −1 show strong positive and negative correlation, respectively, while values closer to 0 show weak or no correlation. EV is defined as the ratio of variances of forecast and observed rainfall. There is considerable variability in the Indian monsoons and EV quantifies the percentage of variability explained by the model. All the statistics are standardized to measure the percentage departures.

3 Results and Discussion

In the present work we perform the periodicities search in SHR7 rainfall data. We estimate the PSD function of SHR7 annual rainfall data for the time period 1871–2005 using Welch technique. The significance testing of periodicities is performed against white noise spectrum, as explained in Gilman et al. (1963). It is found that there are only two significant periods of 2.3 y

Table 3 Performance of two methods in the training phase

Method	RMSE	CC	EV
ANN on SHR7 data	0.31	0.35	0.3875
Periodic (wavelet)+ANN on random part	0.11	0.56	0.5749

Table 4 Performance of two methods in the training phase

Method	RMSE	CC	EV
ANN on SHR7 data	0.09	0.91	0.89
ANN on wavelet MRA	0.02	0.93	0.95

and 22.7 y above 95 % confidence level, as shown in Fig. 2. These two periods (listed in Table 1) are fitted in the sum of cosine series (Eq. 1). The results on evaluating parameters are listed in Table 2. The EV between SHR7 rainfall and the fitted periodic component is very low as 0.05, while CC is 0.11. The RMSE is also high as 0.9. Therefore, we observe that performing prediction on the direct rainfall data is an inadequate choice.

We then perform the periodic modeling at each scale of MRA decomposition. First, the rainfall time series during 1871–2005 (N=135) is decomposed at dyadic scales (2^j, j=1,2,..,J= $\log_2 N$), using discrete wavelet transforms (Fig. 3). By its mathematical definition, the MRA constructs a hierarchy of approximations to a time series x(t), t=0,…N-1, in various subspaces of a linear vector space, such that d1, for example, is the amount of "detail" lost in going to the first approximation x_{N-1}. The process is repeated until all the details are extracted which represent the high frequency components in the signal. Since this decomposition is done at the dyadic scale, and the length of our data is 135, we consider the first four details (represented as d1, d2, d3, d4) and fourth approximation (a4) which fully reconstructs our rainfall series. Once the MRA decomposition is achieved, we have an idea of frequencies (or periods) occurring in a particular range of scales. For example, at first scale, d1 will have periodicities in the 2–4 years scale, d2 in the 4–8 years and so on. These periodicities can be then extracted using spectral analysis Welch technique. The significance levels of periodicities obtained from the PSD are tested against the appropriate colored reference spectra in each scale band as described in Azad and Narasimha (2008). The procedure automatically takes account of the general nature of the rainfall spectrum, including in particular the dip at around $0.25y^{-1}$. The significant periodicities above 95 % confidence levels are shown in Fig. 4 and also listed in Table 1.

It is worth noting that the classical significance test (against white noise) on the PSD of SHR7 rainfall time series (without MRA decomposition) showed only two periodicities of 2.3 y and 22.7 y above 95 % confidence level. However, we found there are ten periodicities significant above 95 % confidence level using the method of PSD on the reconstructed time series from MRA. This is the strength of our proposed technique. It takes into account the different noise present in the signal because the data is cut down at various scales and at each scale a different background spectrum is considered. For example, in Fig. 4a, at first level of MRA decomposition (i.e., d1) periodicities are tested against blue background spectrum, whereas at second level of MRA decomposition (i.e., d2), in Fig. 4b, periodicities are tested against red background spectrum.

The significant periods at each scale are fitted in a sum of cosine series (Eq. 1). The parameters are estimated using least square and a periodic component is obtained which

Table 5 Performance of the two methods in testing phase (1981–2005)

Method	RMSE	CC	EV
ANN on SHR7 data	0.43	0.16	0.1937
Periodic (wavelet)+ANN on random part	0.18	0.38	0.4524

Table 6 Performance of different methods in the testing phase (1984–2011)

Method	RMSE	CC	EV
ANN on SHR7 data	0.12	0.85	0.83
ANN on wavelet MRA	0.04	0.90	0.93

accounts for 30 % of the variance of the total rainfall (Table 2). The calculated RMSE (standardised) value between observed and fitted periodic part is 0.03 and CC is 0.52, which demonstrate the efficacy of the wavelet decomposition approach. The fitted periodic part is plotted in Fig. 5 along with the observed data of SHR7 rainfall. In Fig. 5, we also show the predicted time series. It is to be noted that we have fitted our function for the time period 1871–2005 and it is found that the MRA technique is able to capture the two important drought years at 2002 and 2004. We have validated our prediction on the time period 2006–2011 and found that it is able to capture the trend in the rainfall.

The stochastic or random part is then obtained by removing the periodic component from the given data. A Q-Q plot for the stochastic data is shown in Fig. 6. The plot reveals that the distribution does not deviate greatly from normality. It is to be noted that we are not modelling the probability distribution of the random part but the stochastic part is modelled using ANN. The details of the method are given in Section 2.2. The results from periodic and random part are combined and listed in Tables 3, 4, 5 and 6. Table 3 contains the results obtained in training phase of ANN for both methods, when ANN is applied on MRA and when it is applied directly on the annual data.

For a 25-year period from 1981 to 2005 the overall performance of the testing phase is listed in Table 5. Results show that when ANN is applied directly to the SHR7 annual data, it explains 19 % of the total variance of data. While if the data is treated separately as periodic (using wavelets) and random part (using ANN), our model explains 45 % of the variance. It is to be noted that we have used the same input parameters for ANN as described in Section 2.2. The CC between observed data and fitted is calculated as 0.16 in case of direct method, whereas 0.38 in case of treated separately. The root mean square is also calculated and listed in tables.

We apply the same procedure on monthly data for the time period 1871–2011. It is difficult to find periodicities in monthly data using the same procedure. Hence, we have decomposed the data into different scales using the above technique of MRA and applied ANN on each scale time series. The results obtained are shown in Tables 4 and 6. For the time period 1984 to 2011, the overall performance of the testing phase is listed in Table 6. It shows that the ANN on wavelet decomposition at each scale delivers good results. The root mean square error in case of ANN on MRA is 0.04, whereas in case of ANN on direct monthly data it is 0.12. Also, the explained variance in case of ANN on MRA is 93 % and the correlation coefficient between the regressed and observed series reached 0.90. Therefore, such methodology can be useful to predict time series up to large extent.

4 Conclusions

An empirical technique has been developed using wavelet-neural network model that allows the prediction of Indian monsoon rainfall time series. We have analysed the monthly and annual rainfall data over a spectrally homogenous zone defined as SHR7 constituting the seven

monsoon sub-divisions of Indian subcontinent. The skill of the prediction of annual rainfall in terms of earlier results is improved because we have investigated the periodic component in the data using wavelet decomposition against appropriate background spectrum. Also, it is shown that the ANN, when applied directly on the SHR7 annual data, explains 19 % of the total variance of data. While if the data is treated separately as periodic (using wavelets) and random part (using ANN), our model explains 45 % of the variance. Also, the proposed method is able to extract a larger number of periodic components compared to a simple power-spectrum analysis. In fact it has been reported in the literature that only 34 % of the Indian rainfall is predictable (Annamalai 1995) using data analysis techniques. This is because Indian monsoon rainfall data is highly random, and none of the empirical prediction models have proved to be successful in predicting the rainfall. However, we have shown that using MRA+ ANN hybrid technique, 45 % of the rainfall is captured, hence the present approach can be useful in this direction. Therefore, we conclude that our results are more accurate when the data is treated separately in the proposed framework. Also, spectrally homogenous region for prediction is a useful choice.

References

Annamalai H (1995) Intrinsic problems in the seasonal prediction of the Indian summer monsoon rainfall. Meteorol Atmos Phys 55:61–76

Azad S, Narasimha R, Sett SK (2007) Multiresolution analysis for separating closely spaced frequencies with an application to Indian monsoon rainfall data. Int J Wavelets Multiresolution Inf Process 5: 735–752

Azad S, Narasimha R, Sett SK (2008) A wavelet based significance test for periodicities in Indian monsoon rainfall. Int J Wavelets Multiresolution Inf Process 6:291–304

Azad S, Vignesh T, Narasimha R (2010) Periodicities in Indian monsoon rainfall over spectrally homogeneous regions. Int J Climatol 30:2289–2298

Bhakar SR, Singh R, Hari R (2006) Stochastic modeling of wind speed at Udaipur. Indian J Agril Eng 43(1):1–7

Dabral PP, Pandey A, Baithuri N, Mal BC (2008) Stochastic modeling of rainfall in humid region of northeast India. Water Resour Manag 22:1395–1407

Gilman DL, Fuglister FG, Mitchell JJ (1963) On the power spectrum of "Red noise". J Atmos Sci 20:182–184

Guhathakurta P (2006) Long-Range Monsoon Rainfall Prediction of 2005 for the Districts and Sub-Division Kerala with Artificial Neural Network. Curr Sci 90(6):773–779

Iyengar RN, Kanth STG (2004) Intrinsic mode functions and a strategy for forecasting Indian monsoon rainfall. Meteorol Atmos Phys 90:17–36

Keppenne CL, Ghil M (1992) Adaptive spectral analysis and prediction of the Southern Oscillation Index. J Geophys Res 97:20449–20454

Kishtawal CM, Basu S, Patadia F, Thapliyal PK (2003) Forecasting summer monsoon rainfall over India using genetic algorithm. Geophys Res Lett 30(23):2203–2207

Maheswaran R, Khosa R (2014) A wavelet-based second order nonlinear model for forecasting monthly rainfall. Water Resour Manag 28(15):5411–5431

Michaelides SC, Pattichis CS, Kleovoulou G (2001) Classification of rainfall variability by using artificial neural networks. Int J Climatol 21:1401–1414

Narasimha R, Bhattacharyya S (2010) A wavelet cross-spectral analysis of solar–ENSO– rainfall connections in the Indian monsoons. Appl Comput Harmon Anal 28:285–295

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

Narsimlu B, Gosain AK, Chahar BR, Singh SK, Srivastava PK (2015) SWAT Model Calibration and Uncertainty Analysis for Streamflow Prediction in the Kunwari River Basin, India, Using Sequential Uncertainty Fitting. Environ Process 2:64. doi:10.1007/s40710-015-0064-8

Parthasarathy B, Rupa Kumar K, Munot AA (1995) Homogeneous regional summer monsoon rainfall over India: Interannual variability and teleconnections. RR No: 070

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

Rajeevan M, Pai DS, Dikshit SK, Kelkar RR (2004) IMD's new operational models for long range forecast of south-west monsoon rainfall over India and their verification for 2003. Curr Sci 86:422–431

Rajeevan M, Pai DS, Anil Kumar R, Lal B (2007) New statistical models for long range forecasting of southwest monsoon rainfall over India. Clim Dyn 28:813–828

Rajkumar KN, Kumar D (2004) Stochastic modeling of daily rainfall for south west monsoon season of Baptla, Andhra Pradesh. Indian J Agril Engg 41(3):41–45

Ramana RV, Krishna B, Kumar SR, Pandey NG (2013) Monthly rainfall prediction using wavelet neural network analysis. Water Resour Manag 27:3697–3711

Rangarajan GK (1994) Singular spectral analysis of homogeneous Indian monsoon (HIM) rainfall. Proc Indian Acad Sci (Earth Planet Sci) 103(4):439–448

Sahay RR, Srivastava A (2014) Predicting monsoon floods in rivers embedding wavelet transform, genetic algorithm and neural network. Water Resour Manag 28:301–317

Sehgal V, Sahay RR, Chatterjee C (2014) Effect of utilization of discrete wavelet components on flood forecasting performance of wavelet based ANFIS Models. Water Resour Manag 28(6):1733–1749

Shukla J, Mooley DA (1987) Empirical prediction of the summer monsoon rainfall over India. Monthly Weather Rev 117:695–703

Stoica P and Moses RL (1997) Introduction to Spectral Analysis, Prentice Hall, Upper Saddle River, NJ

Vautard R, Yiou P, Ghil M (1992) Singular spectrum analysis: A toolkit for short, noisy and chaotic series. Physica D 58:95–126

Yao CS (1983) Fitting a linear autoregressive model for long-range forecasting. Monthly Weather Rev 111:692–700