

A Wavelet-Based Second Order Nonlinear Model for Forecasting Monthly Rainfall

R Maheswaran · Rakesh Khosa

Received: 24 April 2014 / Accepted: 30 September 2014 /

Published online: 6 November 2014

© Springer Science+Business Media Dordrecht 2014

Abstract In this article, a rainfall forecasting model using monthly historical rainfall data and climate indices is developed by incorporating wavelet analysis (WA) and second order volterra nonlinear model. The monthly rainfall time series and large-scale climate index time series are decomposed using wavelets into a certain number of component subseries at different temporal scales. The lag relationship between the rainfall anomaly and each potential predictor is identified by cross correlation analysis with a lag time of at least 1 month at different temporal scales. The components of predictor variables with known lag times are then integrated using a second order Volterra model. Further, orthogonal least squares method is used to reduce the redundant variables and select the significant variables to be included into the final forecast model. The proposed multivariate wavelet nonlinear rainfall forecasting method is examined with over three places in India, and compared to the traditional ANN model based on the original time series and linear wavelet regression model. The models are trained with data from the 1916 to 1968 period and then tested in the 1968–1989 period. The results show that the proposed wavelet nonlinear model provides considerably more accurate monthly rainfall forecasts for the three selected places in India than the traditional regression model, neural networks model and the wavelet based linear model. It was seen that for the proposed models and other models also, both the past rainfall and the large-scale climate signals were useful in forecasting the future.

Keywords Climate indices · Monthly rainfall forecasting · Wavelet transform · Second order nonlinear model · India

1 Introduction

Forecast of rainfall is essential for planning and management of water resources especially in an agriculture based country like India. About 65 % of the total cultivated land in India is under the influence of rainfed agriculture system (Swaminathan and 1998). Especially, monthly and seasonal rainfall forecasts provide useful information for water resource management, agricultural planning, and associated crop insurance application (Garbrecht et al.

R. Maheswaran (✉) · R. Khosa

Department of Civil Engineering, Indian Institute of Technology, New Delhi, Delhi, India
e-mail: maheswaran27@yahoo.co.in

2010) in regions which are completely depend on rainfall for agriculture. Reliable forecast of rainfall help the farmers in deciding the type of crops to be cultivated and manage the resources. Therefore, forecasting the monsoon temporally is a major scientific issue in the field of hydrology. Understanding the importance of the rainfall forecasting, there has been intense research in this area which has resulted in numerous approaches for the same. In the past decades, various approaches have been used to study and forecast monthly and seasonal rainfall. These methods can be broadly classified into two categories: numerical and empirical. The numerical models such as general circulation models (GCMs) are based on the laws of physics, which have been used to forecast climate. On the other hand, the empirical models are based on observational relationships of the predictand variable with various predictors. Using the numerical models by Bustamante et al. (1999) and (Olsson et al. 2004) and physical models by Georgakakos and Bras (1984), studies of rainfall quantitative prediction have been carried out. However, they are not successful enough in forecasting rainfall (Olson et al. 1995) due to inaccurate initial conditions, parameterization schemes of subscale phenomena, and limited spatial resolution (Valverde Ram rez et al. 2005).

Owing to numerical models' low forecasting skill and/or their complexity (Meinke et al. 2007), empirical methods are still the most widely used approaches for seasonal precipitation forecasts when they are used in agricultural planning (Meinke et al. 2007). The empirical methods include statistical models (Mutai and Ward 2000; Immerzeel et al. 2010; Prasad et al. 2010) and artificial neural networks (French et al. 1992; Navone and Ceccatto 1994; Sivapragasam et al. 2001; Freiwan and Cigizoglu 2005; Marzano et al. 2006; Moustris et al. 2011; Abbot and Marohasy 2012; Jeong et al. 2012). Most of these studies use only a set of climate-related variables or historical rainfall data as input. Studies in the past (Ropelewski and Halpert, 1986; Kurtzman and Scanlon, 2007; Ishak et al., 2013; Khedun et al., 2014) have shown that the presence of tele-connections between rainfall and large scale climate signals such as Southern Oscillations Index, Pacific Decadal Oscillation (PDO), Indian Ocean Dipole (IOD), SST (Sea Surface Temperature). There have been relatively few investigations (Khedun et al., (2014)) where the rainfall forecast models have used a combination of historical rainfall data and other climatic attributes (Abbot and Marohasy 2012). Even though many studies such as Murphy and Timbal (2008), Shukla (Shukla and Paolino 1983), and (Chattopadhyay et al. 2010) have tried to establish the relationship with climate modes, but there has been few studies to use these climate indices for rainfall prediction. Even though, the reasonable correlations between these large scale climate indices and rainfall have motivated scientists for using them for rainfall forecasting, however, it is not certain about their usefulness in rainfall forecasting. This may be attributed because the signals are highly non-stationary and these processes related to rainfall operate across varying range of temporal scales. In order to exhumate the underlying relationships across different scales scientists have been using the recent technique, wavelet analysis. Wavelet analysis is a useful mathematical tool that provides a time–frequency representation of an analysed signal in the time domain (Daubechies 1992; Percival and Walden 2000). Wavelet analysis is a multi resolution decomposition in time and frequency domains. Anctil and Coulibaly (2004) proposed a wavelet-based approach to describe local interannual variabilities in streamflow, and to identify plausible climatic tele-connections that could explain these local variations. (Rivera et al. 2012) presented a self-organizing map approach using sea surface temperature (SST) filtered by WT for forecasting monthly precipitation in central Chile. More recently, (He et al. 2013) have used wavelet based linear model for utilizing the different climate indices for forecasting 1 month ahead rainfall. Basically, (He et al. 2013) have used linear regression model for forecasting rainfall. However, it is a general notion that the physical processes which induce rainfall is usually nonlinear (Sivakumar 2001; Dhanya and Nagesh Kumar 2011) and it is not possible for linear models to capture the underlying nonlinear dynamics.

Therefore, as an improvement over the linear approach, in this article, a more effective rainfall forecast model from the past rainfall data and climate signals by incorporating the wavelet analysis and multiple second order nonlinear model is proposed. Recent literature review suggest that there has been no any work reported so far on wavelet based multivariate nonlinear model using climate indices for rainfall forecasting. The main contributions of this work are as follows.

1. Development of wavelet based multivariate nonlinear model for rainfall forecasting using the climate indices.
2. Comparison of the proposed approach with other methods such as Artificial Neural Networks model and the wavelet based multiple linear regression approach.

The rest of this article is organized as follows. In Section 2, we briefly describe the wavelet analysis. Study area and the details on climate indices are provided in Section 3. Section 4 is focused on some necessary mathematical methods and the wavelet based multivariate nonlinear model. In Section 5, the proposed method is applied to monthly rainfall forecasting at 3 rainfall stations across India and compared with a linear regression model based on data with and without WT. Some conclusions are made in Section 6.

2 Wavelet Analysis

Wavelet analysis, initially formalized by Grossmann and Morlet (1984), is the most recent solution to overcome the main shortcoming of the Fourier transforms that identifies the frequencies present in a signal but not their moment. Wavelet analysis results in a time frequency (or time-scale) representation of a signal. Wavelet analysis transforms a signal into scaled and translated versions of an original (mother) wavelet, instead of decomposing a signal into constituent harmonic functions as in Fourier analysis. The wavelet transform as defined by Eq. (1) (Daubechies 1992) is called the continuous wavelet transform (abbreviated CWT) because the scale and time parameters, a and τ assume continuous values.

$$W(a, \tau) = \frac{1}{\sqrt{|a|}} \int_{-\infty}^{\infty} f(t) \psi\left(\frac{t-\tau}{a}\right) dt \quad (1)$$

It provides a redundant representation of a signal as the CWT of a function at scale ' a ' and location ' τ ' can be obtained from the continuous wavelet transform of the same function at other scales and locations. Since the CWT behaves like an orthonormal basis decomposition, it can be shown that it is also isometric as it preserves the overall energy content of the signal and, thereby, allows for the recovery of the function $f(t)$ from its transform by using the following reconstruction formula as provided by (Daubechies 1992) in Eq. (2)

$$f(t) = \frac{1}{C_{\psi}} \int_{-\infty}^{\infty} \int_0^{\infty} a^{-2} W(a, \tau) \psi_{a,\tau}(t) da d\tau \quad (2)$$

where C_{ψ} is a constant and depends on the choice of the wavelet. Clearly, the above equation suggests that the function $f(t)$ may be seen as a superposition of signals at different scales and obtained by varying the scale parameter ' a '.

Further, the energy of the signal $f(t)$ can be represented scale wise as given by (Daubechies 1992) in Eq. (3)

$$\int_{-\infty}^{\infty} f^2(t)dt = \frac{1}{C_{\psi}} \int_0^{\infty} \left[\int_{-\infty}^{\infty} |W(a, \tau)|^2 d\tau \right] \frac{da}{a^2} \tag{3}$$

The left-hand side of Eq. (3) is called the ‘energy’ in the signal $f(t)$ (it is, however, not energy in the physical sense unless $f(t)$ has the proper units). We can thus interpret $[W(\alpha, \tau)]^2 d\tau$ as being proportional to an energy density function that decomposes the energy in $f(t)$ across different scales and times. Again, if $[W(\alpha, \tau)]^2 d\tau$ is large (small), we can say that there is an important (insignificant) contribution to the energy in $f(t)$ at scale τ and time t .

Flandrin (1988) proposed calling the function $|W(\alpha, \tau)|^2$ a scalogram and, in general, for two different functions $f(t)$ and $g(t)$, the product $W_f(\alpha, \tau)$ and $W_g(\alpha, \tau)$ may be called a cross scalogram. While, in general, a scalogram provides an unfolding of the characteristics of a process in the scale-space plane, a cross scalogram, on the other hand, provides a similar unfolding of possible interactions of two processes, and this measure can be quite revealing about the structure of a particular process or about the interaction between different processes.

As can be seen, the CWT offers a promising platform for understanding a given dynamic process and facilitates its objective characterization, in terms of the time series of observations available on the process and, particularly in the area of hydrology, there are several applications wherein wavelet analysis has already been shown to be a credible analysis technique designed to foster understanding of these natural processes.

For practical applications, the hydrologist has access only to a discrete time signal, rather than to a continuous time signal. A discretization of the Continuous Wavelet Transform (CWT) produces N^2 coefficients from a data set of length N ; hence redundant information is locked up within the coefficients, which may or may not be a desirable property (Addison 2002). To overcome this redundancy, logarithmic uniform spacing can be used for the a scale discretization with a correspondingly coarser resolution of the b locations, which allows for N transform coefficients to completely describe a signal of length N . Such a discrete wavelet has the form: [Mallat, 1989]

$$f_{m,n}(t) = \frac{1}{\sqrt{a_0^m}} f\left(\frac{t-nb_0a_0^m}{a_0^m}\right) \tag{4}$$

where m and n =integers that control the wavelet dilation and translation, respectively; b_0 = location parameter and must be greater than zero; a_0 =a specified fined dilation step greater than 1. The most common and simplest choice for parameters are $a_0=2$ and $b_0=1$. This power of two logarithmic scaling of the translation and dilation is known as the dyadic grid arrangement (Szilagyai et al. 1996). The dyadic wavelet can be written in a more compact notation as:

$$f_{m,n}(t) = 2^{-m/2} f(2^{-m}t-n) \tag{5}$$

Discrete dyadic wavelets of this form are usually chosen to be orthonormal. This allows for the complete regeneration of the original signal as an expansion of a linear combination of translating and dilating orthonormal wavelets. For a discrete time series x_i , the dyadic wavelet transform becomes:

$$T_{m,n} = 2^{-m/2} \sum_{i=0}^{N-1} f(2^{-m}i-n)x_i \tag{6}$$

where $T_{m,n}$ =wavelet coefficient for the discrete wavelet of scale $a=2^m$ and location. Eq. (7) considers a finite time series, $x_i, i=0, 1, 2, \dots, N-1$, where N is an integer power of 2: $N=2^M$.

This gives the ranges of m and n as b_0 and $1 < m < M$, respectively. At the largest wavelet scale (i.e., 2^m where $m=M$), just one wavelet is needed to cover the time interval and only one coefficient is created. At the next scale (2^{m-1}), two wavelets cover the time interval, therefore two coefficients are created, and so on down to $m=1$. At $m=1$, the a scale is 2^1 , i.e., 2^{M-1} or $N/2$ coefficients are needed to describe the signal at this scale. The total number of wavelet coefficients for a discrete time series of length $N=2^M$ is then $1+2+4+8+\dots+2^{M-1}=N-1$. A signal smoothed component, $\bar{T}(t)$, is left, which is the signal mean. Therefore, a time series of length N is broken into N components, i.e., with zero redundancy. The inverse discrete transform is given by:

$$f(t) = \bar{T}(t) + \sum_{m=1}^M W_m(t) \quad (7)$$

where $\bar{T}(t)$ is called the approximation subsignal at level M ; and $W_m(t)$ =detailed sub signals at levels $m=1,2,\dots,M$. The wavelet coefficients, $W_m(t)$, $m=1,2,\dots,M$, provide the detail signals, which can capture small features of interpretational value in the data; the residual term, $\bar{T}(t)$, represents the data background information.

In this study, wavelet decomposition is performed using the db2 orthogonal discrete wavelet function, as suggested by Maheswaran and Khosa (2012b) taking care of the boundary distortion. Four levels of decomposition is implemented. For a sampling period of 1 month, the time scales of the wavelet decomposition are 2-, 4-, 8-, 16- and 32-months, respectively, for the resolution levels $j = 1, 2, 3, 4$ and 5. These decomposition levels allow examining usefulness of a range of time-scale signals in rainfall forecasting.

3 Study Area and Data

To test the proposed method, monthly rainfall data from two different sub basins of Cauvery Basin, India and one from Gurgaon district in Delhi NCT(National Capital Territory),India were considered.

Cauvery basin receives an annual average rainfall of 1,129 mm and of which, about 50 % is received during the south-west monsoon (June-September), 33 % in the northeast monsoon (October – January) and the rest in the summer months (February – March). The mean daily maximum temperature ranges from 19.5 to 33.7 °C, whereas the mean daily minimum varies from 9.1 to 25.2 °C.

The climate of the Gurgaon district can be classified as tropical steppe, semi-arid and hot which is mainly characterized by the extreme dryness of the air except during monsoon months, intensely hot summers and cold winters. During 3 months of south west monsoon from last week of June to September, the moist air of oceanic origin penetrates into the district and causes high humidity, cloudiness and monsoon rainfall. The normal annual rainfall in Gurgaon district is about 596 mm spread over 28 days. The south west monsoon sets in the last week of June and withdraws towards the end of September and contributes about 85 % of the annual rainfall.

3.1 Rainfall Data

Long-term monthly rainfall data were obtained from the IMD monthly rainfall dataset. The data is then interpolated to obtain the basin wise average rainfall. The resulted basin wise average rainfall data was available for the period between 1916 and 1989. The 70 % of the data set was used for calibration and the remaining data was used for validating purposes.

3.2 Climate Data

To forecast monthly rainfall, in this study we choose different large-scale climate signals, which have been identified to be influencing Indian rainfall by different researches (Polaski et al., (2013) and (He et al. 2013)). Following are the selected large-scale climate indices which have been found to have significant influence on rainfall over Indian subcontinent,

- a) Southern Oscillation Index (SOI), which is used as a common index for ENSO, was chosen as a potential predictor of rainfall because it has the longest period of record (1876 to present), and it was successfully used in previous rainfall forecasting studies (Meinke and Stone 2005; Abbot and Marohasy 2012). Monthly SOI values were obtained from the Australian Bureau of Meteorology at <http://www.bom.gov.au/climate/glossary/soi.shtml>.
- b) The IOD has widespread effects on rainfall in East Africa, India, Indonesia, and the western a southern Australia (Webster *et al.* 1999; Ummenhofer et al., 2009). Variability in the Indian Ocean is associated with rainfall variability in the Indian subcontinent, and Indian Ocean variability is reported to be the key driver climate in India. The IOD index is represented by anomalous SST gradient between the western equatorial Indian Ocean (50°E–70°E and 10°S–10°N) and the southeastern equatorial Indian Ocean (90°E–110°E and 10°S–0°N).
- c) The PDO is a pattern of climate variability with a similar expression to El Niño, but acting on a longer timescale, and with a pattern most clearly expressed in the North Pacific (Mantua *et al.* 1997; Mantua and Hare 2002). The PDO index is based on a projection of SST anomalies onto a pattern defined by the leading principal component of monthly SST anomalies in the North Pacific poleward of 20°N. PDO index is available from <http://jisao.washington.edu/pdo/PDO.latest>

Therefore, in our study these three climate indices were taken as the input variables for the nonlinear regression model.

3.3 Identification of Significant Components

It is assumed that the rainfall responds to large scale signals with a time lag (He et al. 2013). Cross correlation function (CCF) is a common method generally used for evaluation the lag relationship between two variables. In this study, all monthly time series are decomposed into a certain number of subseries components under different temporal scales using a specific mother wavelet. The mother wavelet and the depth of decomposition are chosen based on the previous studies by (Maheswaran and Khosa 2012b). The cross correlation function is implemented to identify lag relationships between rainfall subseries versus each potential predictor subseries. The lag correlation coefficient between the two sets of subseries is used for this purpose.

4 Methods Used

4.1 Multiple Input Wavelet Volterra Coupled (MWVC) Nonlinear Model

The decomposed time series of the various climate indices and the past rainfall data form the input variables for the model. From these input variables, those which have a significant lagged cross correlation with the rainfall time series were identified and were then integrated using the second order Multiple Input Single Output (MISO) Volterra model to provide the forecast at next step. The Fig. 1 shows the model scheme for the propose scheme.

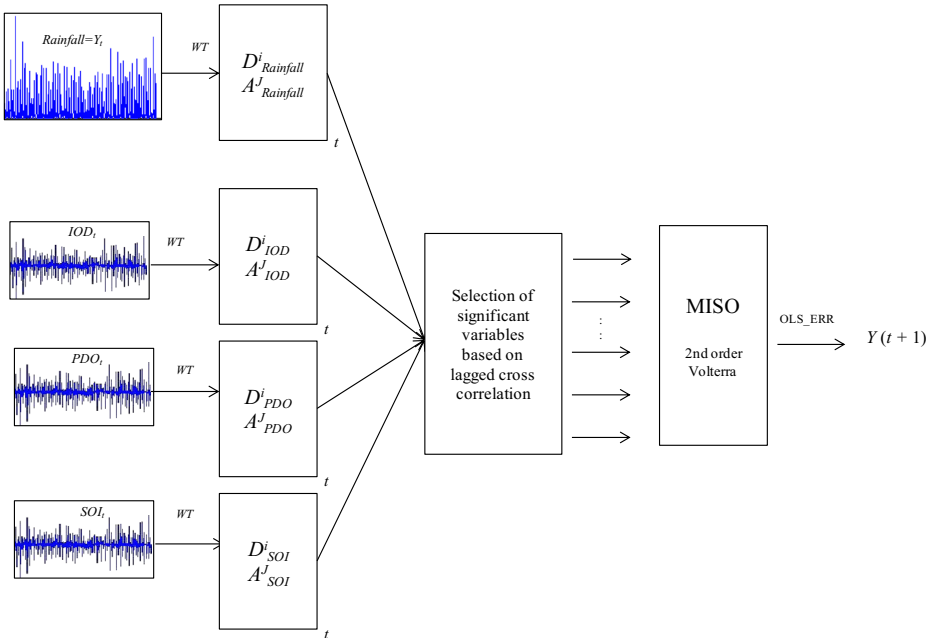


Fig. 1 Wavelet Nonlinear Model –Scheme I

Here, D^i_X $i=1,2,...J$ denotes the detail component of the wavelet decomposition of a certain input variable X and A^J_X denotes the approximation component of the wavelet decomposition of the same input variable X .

From the wavelet coefficients of the different input time series, the significant wavelet coefficients are selected based on the lag correlation with the observed rainfall time series. Let for example, some of these may be denoted by $D^{i_{rainfall}}(t), D^{i_{IOD}}(t), \dots, D^{i_{PDO}}(t)$ and similarly the significant scaling coefficients at the decomposition level J of the different input series be denoted by $A^J_{Rainfall}(t), A^J_{IOD}(t), \dots, A^J_{PDO}(t)$, where i denotes the depth of decomposition which varies from 1 to J .

Now, the significant wavelet coefficients and scaling coefficients of the different input series are nonlinearly convolved using the second order Volterra representation within a multiple inputs-single output frame work. For simplicity of notation, let these different series be denoted by $u_1, u_2 \dots u_L$ where L is total number of inputs.

If $y(t)$ denotes the rainfall time series, L denotes the number of input variables, N is the length of the time series, m denotes the memory of each input variable up to which there is a significant lag relationship with the rainfall time series and ξ_t represents the model noise including modelling errors and the unobservable disturbances, the multiscale nonlinear relationship may be written as

$$\left. \begin{aligned}
 y(t) = & \sum_{n=1}^L \sum_{\tau=1}^m h_1^{(n)}(\tau) u_n(t-\tau) + \\
 & \sum_{n=1}^L \sum_{\tau_1=1}^m \sum_{\tau_2=1}^m h_{2s}^{(n)}(\tau_1, \tau_2) u_n(t-\tau_1) u_n(t-\tau_2) + \\
 & \sum_{n_1=1}^L \sum_{n_2=1}^m \sum_{\tau_1=1}^m \sum_{\tau_2=1}^m h_{2x}^{(n_1, n_2)}(\tau_1, \tau_2) u_{n_1}(t-\tau_1) u_{n_2}(t-\tau_2) + \xi_t
 \end{aligned} \right\} \tag{8}$$

First order kernels $h_1^{(n)}$ describe the linear relationship between the n th input u_n and y , the second order self-kernels $h_{2s}^{(n)}$ describe the 2nd order nonlinear relation between the n th input u_n and y respectively and the second order cross-kernels $h_{2x}^{(n_1, n_2)}$ describe the 2nd order nonlinear interactions between each unique pair of inputs (u_{n_1} and u_{n_2}) as they affect y .

Eq. (8) can be simplified by combining the last two terms to yield Eq. (8) and it now remains to estimate kernels h_1 and h_2 .

$$y(t) = \sum_{n=1}^L \sum_{\tau=1}^m h_1^{(n)}(\tau) u_n(t-\tau) + \sum_{n_1=1}^L \sum_{n_2=1}^L \sum_{\tau_1=1}^m \sum_{\tau_2=1}^m h_2^{(n_1, n_2)}(\tau_1, \tau_2) u_{n_1}(t-\tau_1) u_{n_2}(t-\tau_2) + \xi_t \tag{9}$$

The representation of Eq. (9) can be further simplified by considering each of the lagged variables $u_1(t-1), u_1(t-\tau), \dots, u_2(t-1), u_2(t-\tau), \dots$ as separate variables $d_1(t), d_2(t), d_3(t), \dots, d_{N_l}(t)$ then, Eq. (8) can be written as

$$y(t) = \sum_{l=1}^{N_l} h_1(l) d_l(t) + \sum_{l_1=1}^{N_l} \sum_{l_2=1}^{N_l} h_2(l_1, l_2) d_{l_1}(t) d_{l_2}(t) \tag{10}$$

More clearly,

$$\begin{aligned} d_l(t) &= \{x_k(t) \mid 1 \leq k \leq L; 1 \leq l \leq L\} \\ d_l(t) &= \{x_k(t-\tau) \mid 1 \leq k \leq L; L < l \leq N_l; \tau = 1, 2, 3, \dots, m\} \\ \tau &= \tau^{\text{th}} \text{ lagged value.} \\ L &= \text{total predictor time series.} \\ N_l &= \text{total number of lagged variables} \end{aligned}$$

Using the Orthogonal Least Squares- Error Reduction Ratio (OLS-ERR) method of Chen and Billings (1989), the significant regressor terms were selected and correspondingly kernels were estimated. The complete mathematical derivation of the Wavelet Volterra coupled model can be found in (Maheswaran and Khosa 2012a). The programs were coded and executed in the MATLAB 7.6.0.

4.2 Multivariate Wavelet Linear Regression Model

The Multivariate wavelet-based Linear Regression (MWLR) is constructed by incorporating two methods: Linear Regression and Wavelet Transform (Kisi 2009 and He et al. 2013). The details and approximations of the different input variables are combined using the linear regression to predict the future rainfall. For the MWLR model inputs, each of the original rainfall and climate index time series is decomposed into a certain number of subseries components A^j and D^j s using the wavelet transform.

Then the forecasted value $y(t)$, the rainfall at time t can be obtained using the multiple linear regression formulation as given below,

$$y(t) = \sum_{n=1}^L \sum_{\tau=1}^m a_n(\tau) u_n(t-\tau) \tag{11}$$

Where, $u_n(t)$ denotes the wavelet decompositions of the different input predictor variables and a_n denotes the regression parameter which can be obtained from the calibration period data and τ denotes the lag time between the rainfall time series and the predictor time series. The programs were coded and executed in the MATLAB 7.6.0.

4.3 Single Scale Neural Network Model

In the past, neural networks have been extensively applied for forecasting purposes and the results have been very encouraging. In this study, the neural networks model has been used for the basic comparison with the proposed wavelet based approach. The basics of neural networks are available extensively in literatures ((Thirumalaiah and Deo 1998), (Adamowski and Chan 2011)). The choice of the ‘most appropriate’ network training algorithm is usually resolved by means of a trial and error based judgement and, it is understandable that there would be diverse opinions on the use of a specific network algorithm. In the present study, ANN models have been implemented with various training algorithms and their performance assessed through a comparative evaluation and the corresponding ‘best’ results are presented for clarity. The input variables for the ANN model are selected based on the cross correlation between the predictor variables and prectictand. The programs were run using the Neural Network Toolbox and executed in the MATLAB 7.6.0.

4.4 Model Performance Measures

To evaluate the performance of predictions, the following statistical measures of error are considered

1. Mean absolute error (MAE)

$$MAE = \frac{\sum |Observed(i) - Forecast(i)|}{N} \quad (12)$$

2. Root Mean Square Error (RMSE).

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (Observed(i) - Forecast(i))^2}{N-1}} \quad (13)$$

3. Nash Sutcliffe Criteria (NSC)

$$NSC = E = 1 - \frac{\sum_{i=1}^N (Actual(i) - Forecast(i))^2}{\sum_{i=1}^N (Actual(i) - Actual)^2} \quad (14)$$

To compare the performances of the different models, the present study has used measures such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE). Karunanithi et al. (1994) suggested that RMSE is a good measure for indicating goodness of fit. In general, $RMSE \geq MAE$, and the degree to which RMSE exceeds MAE is an indicator of the extent to which large outliers (Variance between the observed and the forecasted values) exist in the evaluation set. Lower the value of RMSE better is the model performance.

Equations (12), (13) and (14) as given above have been used to estimate these performance measures.

5 Model Application

5.1 Selection of input components

As a first step, the wavelet decompositions of the input predictand variables were performed. Figure 2 shows an example of wavelet decomposition using the ‘db2’ wavelet with four resolution levels corresponding to the monthly rainfall series for MH Halli. Similarly Fig. 3 shows wavelet decomposition for the SOI indices. These figures clearly show how the wavelet transform decomposes original series into its detail (D 's) and approximation $A4$ subseries.

The selection of the input wavelet components was done by estimating the cross lag correlation coefficients. For each potential predictor subseries, a monthly lag which lies within the interval $[1, 30]$ is identified and the corresponding maximum lag correlation coefficient (MLCC) is found for this subseries and the rainfall time series. Some of the sample cross correlation plots are shown in Fig. 4. Table 2 reports the summary of the MLCC for the various subseries and the rainfall time series for all the three station under investigation. It can be seen that the components D1, D2, D3, D5 and A5 of the IOD are having significant correlation with the rainfall anomalies. Apart from this, the D4 and A4 component of the SOI time series and the IOD time series are having significant cross lag correlation with the original rainfall time series. Also, it is seen that the D4 and A4 of PDO is having good correlation with the rainfall time series.

5.2 MWVC Model

In line with the assumption that a given time series constituting observations on a natural process is a result of an amalgamation of various sub-processes or phenomena that individually

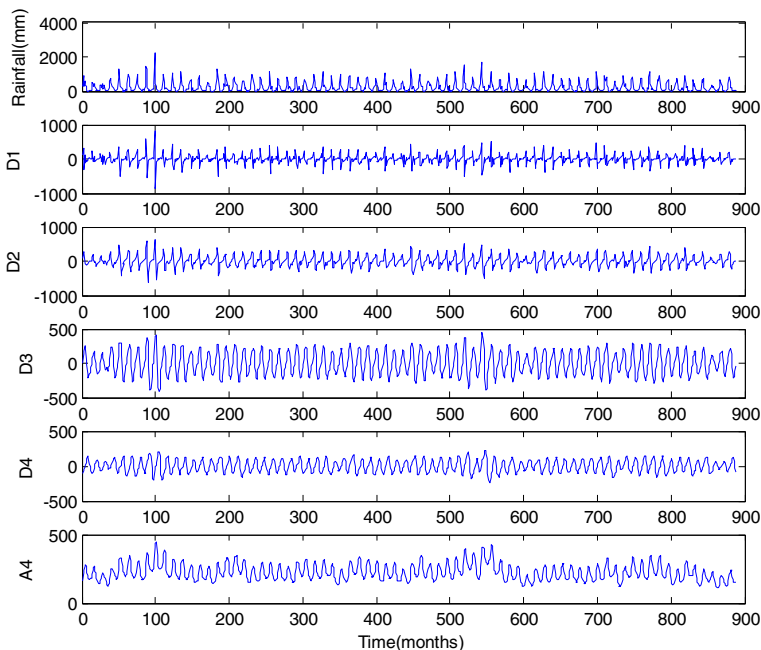


Fig. 2 Wavelet decomposition of the observed rainfall time series at M H Halli using db2 wavelet

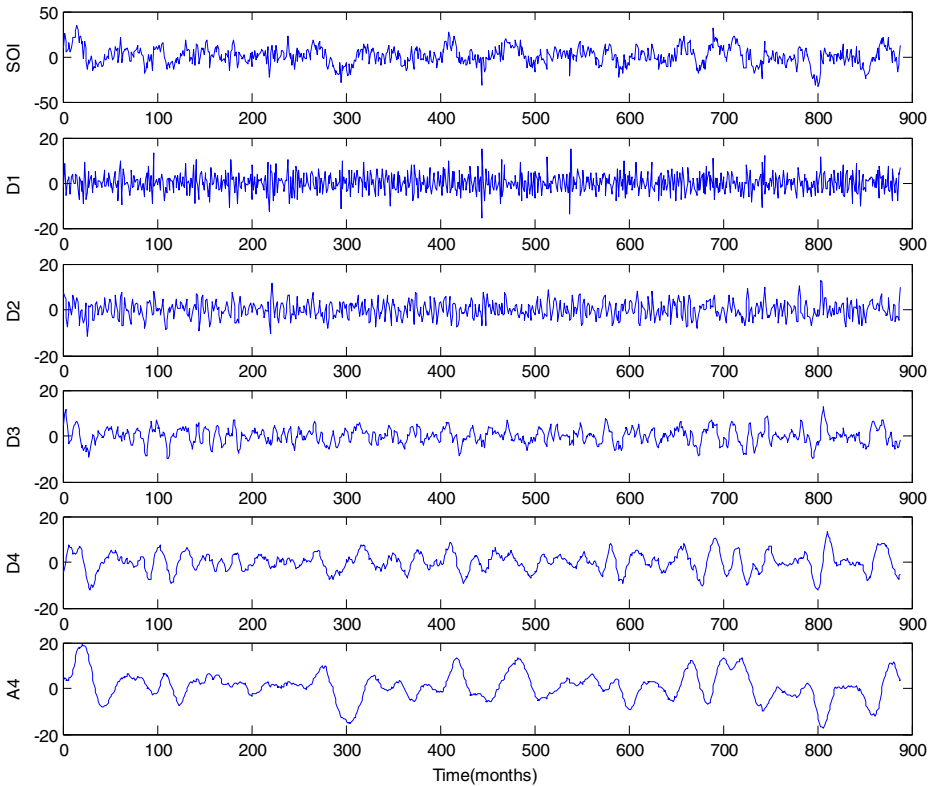


Fig. 3 Wavelet decomposition of the SOI time series using db2 wavelet

operate at various scales and accordingly each of these sub-process has the attribute of memory albeit of different spans. For the rainfall time series of MH Halli, Table 1 indicates that the optimal components that contribute significantly to variability in the observed rainfall process, $X(t)$ are $D1_{IOD}$, $(D4, A4)_{SOI}$, $(D1, D3, A4)_{PDO}$, $(D1, D2, D4, A4)_{rainf}$. The lags at which these variables are having significant correlation at 95 % confidence limits were also estimated. From this group of variables a total of 24 significant input variables were selected based on the higher lagged cross correlation analysis. These input variables were regressed using the 2nd order volterra model. This resulted in an estimation of a total of 356 volterra kernels. However, using the OLS-ERR algorithm, it was seen that the 40 kernels were seen to be significant. Figure 5 shows the plot of the NSC vs. No. of significant kernels selected and it is observed that there is no significant change in the NSC values beyond 40 terms leading to the logical inference that the MWVC model would comprise of just these 40 terms. Significantly, it was seen that the model has only 5 linear terms while the remaining 35 happen to be nonlinear terms. Figure 6 shows the second order kernel values of the derived MWVC model for MH Halli station and the model validation results are compared with the test data on observed rainfall time series and the comparison is presented in Fig. 7. Several input combinations were tried and the results were tabulated in Table 2.

Similar approach was followed for the remaining two stations and the results are presented in the later section.

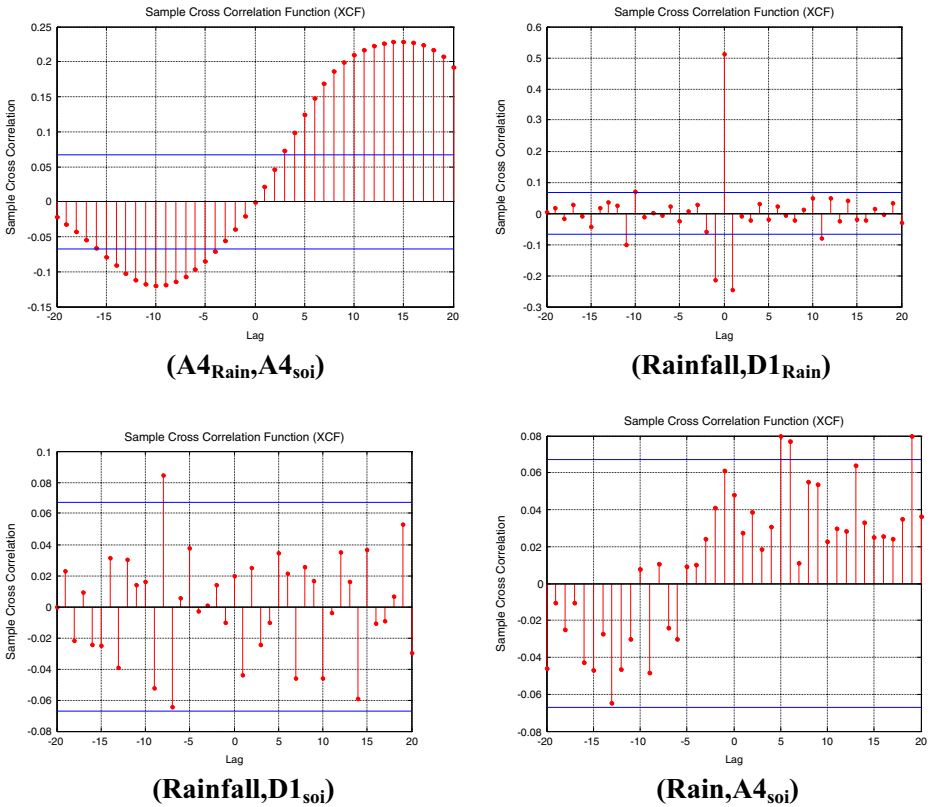


Fig. 4 Cross correlation plots between the rainfall time series and the wavelet decomposition of the SOI time series

5.3 Wavelet Multivariate linear Regression Model

Selection of input variables was based on the strength of cross correlation between individual wavelet decompositions of rainfall and different climate indices. The significant input variables were numbering to 26. Figure 8 shows the plot of the NSC vs. No. of significant terms selected and it is observed that there is no significant change in the NSC values beyond 7 terms leading to the logical inference that the MWVC model would comprise of just these 7 terms. It was seen that out of these 26 inputs only 7 were significantly contributing to the rainfall. The scatter plot of the results for the MHHalli is shown in Fig. 9. The results of the WMLR for all other stations are summarised in the later part of the section. The explicit forecasting equation for the WMLR is given by Eq. (15)

$$R_{t+1} = 0.21D3_t^R + 1.06D4_t^R + 0.29D2_t^{SOI} + 0.43D3_{t-6}^{SOI} + 5.57A4_t^R - 5.10A4_{t-1}^R + 0.52A4_{t-12}^R \tag{15}$$

5.4 Neural networks Model

Recently, Meknaik et al., (2013) have used neural network for forecasting rainfall using the climate indices such as SST and IOD. They have used the ENSO, IOD, SOI and Nino 3.4 as

Table 1 Summary of the Cross Correlation analysis between the wavelet decomposition of the input variables and the observed rainfall time series

Input time series	Kudige		MHHalli		Gurgoan	
	Maximum correlation coefficient	Maximum memory in months	Maximum correlation coefficient	Maximum memory in months	Maximum correlation coefficient	Maximum memory in months
IOD						
D1	0.08	2	0.067	2	-0.13	3
D2	0.15	3	0.16	3	0.12	6
D3	0.002	0	0.001	0	0.04	0
D4	0.002	0	0.002	0	0.13	4
A4	0.08	3	0.09	3	0.15	6
SOI						
D1	0.09	9	0.09	9	0.17	3
D2	0.03	0	0.045	0	0.12	3
D3	0.02	0	0.01	0	0.11	4
D4	0.2	1	0.20	1	0.19	6
A4	0.28	6	0.24	6	0.25	6
PDO						
D1	0.15	20	0.14	20	0.07	1
D2	0.10	12	0.14	12	0.16	8
D3	0.15	10	0.15	10	0.14	10
D4	0.1	6	0.12	6	0.05	0
A4	0.15	6	0.17	6	0.21	12
Historical Rainfall						
D1	0.5	3	0.54	3	0.56	12
D2	0.28	2	0.34	2	0.67	14
D3	0.07	0	0.06	0	0.51	10
D4	0.15	3	0.14	3	-0.1	1
A4	0.3	5	0.33	5	0.36	20

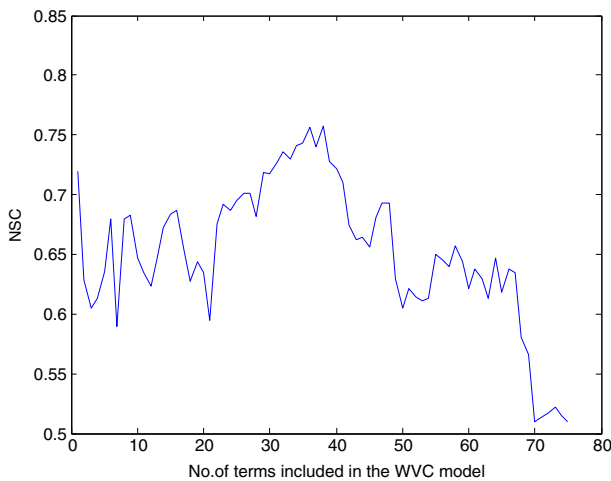


Fig. 5 NSC vs. No. of significant terms used – MH Halli

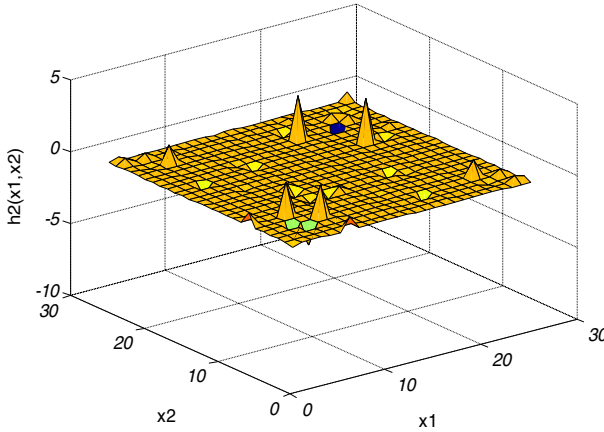


Fig. 6 Second order kernel for the MWVModel – MH Halli

input for the ANN models for forecasting the future rainfall. In this work, a similar approach has been taken for forecasting the rainfall based on the past rainfall values and the past climate indices. Table 3 summarises the prediction skills of the different models that were tested in the study. The scatter plot of the results from the best model for the MHHalli is shown in Fig. 10.

Similar approach was used to obtain the best model for the other two sub-basins viz. Kudige and Gurgoan. The best results from each of the category of the model are reported in Table 4.

6 Results and Discussions

The performance statistics of the Wavelet nonlinear model, wavelet linear models and the neural networks are shown in Table 4 for the test period 1970–1989 for all the three places. NSC for the Wavelet nonlinear Model ranges from 0.74 to 0.78, while that from the WMLR varies from 0.50 to 0.62. As a result, the Wavelet nonlinear model increases the forecast NSC by almost 50 %. This clearly indicates that Wavelet nonlinear provides significantly improved accuracy relative to WMLR for monthly rainfall forecasts.

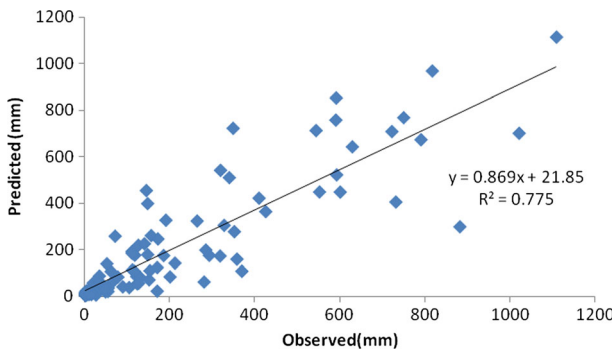


Fig. 7 Observed and forecast rainfall values for MH Halli station using MWVC models

Table 2 Selection of Best MWVCmodel for one month ahead forecasting at MHalli

S.No	Model Inputs	WMLR Results				
		RMSE (m)	MARE (%)	MAE (m)	NSC	CC
(i).	$D1_{t-1}, IOD_{D2_{t-1}}, SOI_{D4_{t-1}}, SOI_{A4_{t-1}}, SOI_{A4_{t-6}},$ $PDO_{D1_{t-1}}, PDO_{A4_{t-1}}, R_{D1_{t-1}}, R_{D1_{t-2}}, R_{D2_{t-1}},$ $R_{D2_{t-3}}, R_{D4_{t-1}}, R_{D4_{t-3}}, R_{A4_{t-1}}, \dots, R_{A4_{t-5}},$	174.56	1.26	83.56	0.58	0.79
(ii)	$D1_{t-1}, IOD_{D2_{t-1}}, SOI_{D4_{t-1}}, SOI_{A4_{t-1}}, SOI_{A4_{t-6}},$ $PDO_{D1_{t-1}}, PDO_{A4_{t-1}}, R_{D1_{t-1}}, R_{D1_{t-2}}, R_{D2_{t-1}},$ $R_{D2_{t-3}},$ $R_{D4_{t-1}}, R_{D4_{t-3}}, R_{A4_{t-1}}, \dots, R_{A4_{t-5}},$	156.25	1.050	70.67	0.65	0.83
(iii)	$D1_{t-1}, IOD_{D2_{t-1}}, SOI_{D4_{t-1}}, SOI_{A4_{t-1}}, SOI_{A4_{t-6}},$ $PDO_{D1_{t-1}}, PDO_{A4_{t-1}}, R_{D1_{t-1}}, R_{D1_{t-2}}, R_{D2_{t-1}},$ $R_{D2_{t-3}},$ $R_{D4_{t-1}}, R_{D4_{t-3}}, R_{A4_{t-1}}, \dots, R_{A4_{t-5}},$	121.27	0.69	73.45	0.77	0.89
(iv)	$D1_{t-1}, IOD_{D2_{t-1}}, SOI_{D4_{t-1}}, SOI_{A4_{t-1}}, SOI_{A4_{t-6}},$ $PDO_{D1_{t-1}}, PDO_{A4_{t-1}}, R_{D1_{t-1}}, R_{D1_{t-2}}, R_{D2_{t-1}},$ $R_{D2_{t-3}},$ $R_{D4_{t-1}}, R_{D4_{t-3}}, R_{A4_{t-1}}, \dots, R_{A4_{t-5}},$	130.76	0.74	79.29	0.69	0.84

Similarly, in terms of RMSE the multivariate Wavelet Volterra model performs better than the other models (WMLR, ANN) by 16 % and 25 %. Therefore, from the RMSE and NSC viewpoints, the proposed wavelet nonlinear model performs better than the WMLR model for monthly rainfall forecasting. This may be because the proposed nonlinear wavelet model has capability to capture the nonlinear relationships of the predictor variables on the rainfall at different time scales, while WMLR doesn't. Even though the number of parameters to be estimated may be more for the wavelet based nonlinear model than the WMLR model (Eq. 15), however, it is to be noted that the original number of input predictor variables is same.

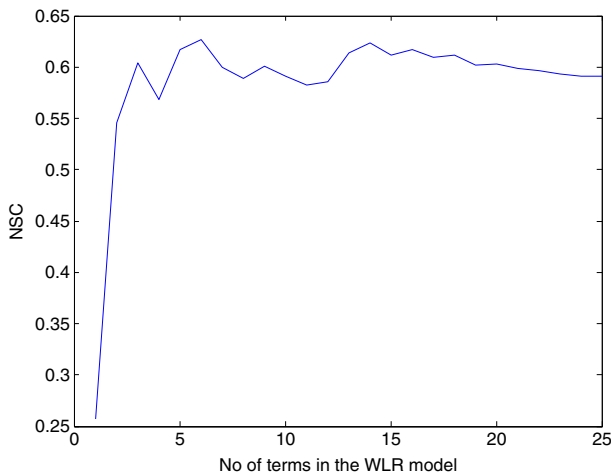


Fig. 8 NSC vs. No. of significant terms used for the MWLR model – MH Halli

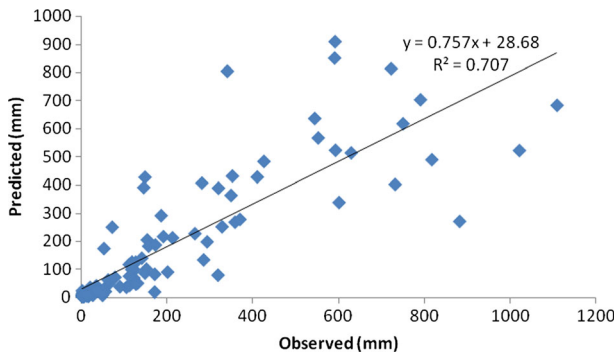


Fig. 9 Observed and forecast rainfall values for MH Halli station using MWLR models

On the other hand, it has been seen that the WMLR performs better than the NN model because of the wavelet decomposition has the capability to unravel the hidden relationships between the predictor variables and the rainfall time series. To inspect how the forecast models perform for dry and wet months, the forecasted results are plotted for MHHalli in Fig. 11, for the test period from January 1985 to Dec 1987. At MHHalli both MWVC and MWLR perform similarly for months with normal rainfall, but for some extreme months, MWVC model provides much better prediction than MWLR. It can be observed that for 1985 and 1986, the MWVC models produced very close forecasts, whereas the MWLR model and ANN model underestimates the rainfall. However, for 1987 both MWVC and MWLR model overestimates the wet months rainfall whereas there is an underestimation by ANN model. For one extreme wet month at the MWVC forecasts 960.23 mm in comparison to the observed 1020.2 mm, with an underestimation of 5.88 %, whereas the MWLR forecasts 521.26 mm, with an underestimation of 48.91 %.

In the case of the summer rainfall also, the MWVC yields better results whereas the other two models either over estimate or underestimates the rainfall. For example, in the year 1987 for June, the MWVC gives a value of 108.34 mm in comparison with the observed rainfall of 123.56 mm. For the same month the MWLR and ANN model provide forecasts of 97.67 mm and 88.56 mm respectively.

Table 3 Selection of Best ANN model for one month ahead forecasting at MHalli. (After normalizing)

S.No	Model Inputs	ANN Results			
		RMSE (mm)	MARE (%)	MAE (mm)	NSC
(i).	Lagged Rainfall anomalies(t-1.....t-15), IOD(t-3), IOD(t-7), SOI(t-1), SOI(t-6), PDO(t-1),PDO(t-3)	2.95	6.8	1.568	-0.99
(ii)	Lagged Rainfall anomalies(t-1.....t-15), IOD(t-3), IOD(t-7), SOI(t-1), SOI(t-6), SOI(t-12), PDO(t-1),PDO(t-3)	1.93	7.6	1.28	-0.35
(iii)	Lagged Rainfall anomalies(t-1.....t-15)	1.52	0.80	1.17	0.55
(iv)	Lagged Rainfall anomalies(t-1.....t-15), IOD(t-3), IOD(t-7), SOI(t-1), SOI(t-6), SOI(t-12), PDO(t-1),PDO(t-3), PDO(t-12)	1.59	0.81	1.23	0.51

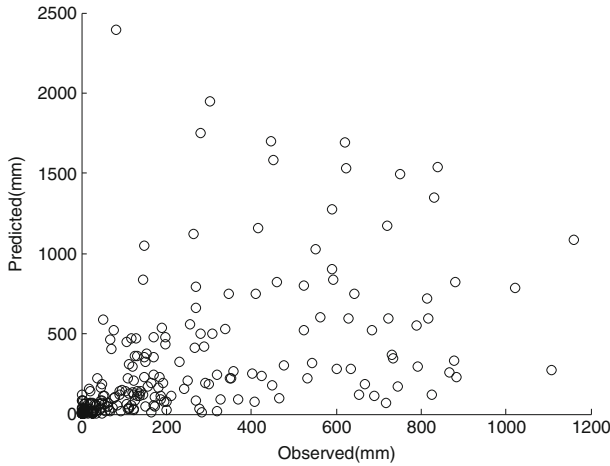


Fig. 10 Observed and forecast rainfall values for MH Halli station using ANN models

The models were compared for their overall capability for forecasting the rainfall. Figure 12 shows the comparison of the forecasts values with the observed rainfall during the months of June, July, Aug and September for the period of 5 years. The analysis of the results indicate that the MWVC performs well in picking up the extreme rainfall events on a monthly time scale.

Table 4 Result Statistics for one month ahead forecasting of one month ahead total rainfall at MHalli and Gurgoan using different models

Performance Measures	Wavelet 2nd order Nonlinear Model	WMLR	ANN (3 hidden layers)	Wavelet Nonlinear Model without climate indices
MHalli				
RMSE (mm)	121.27	149.43	170.52	136.90
MAE (mm)	73.45	87.57	95.78	84.84
Correlation Coefficient	0.89	0.78	0.65	0.82
NSC	0.78	0.62	0.51	0.69
Gurgoan				
RMSE (mm)	72.56	82.47	97.56	77.58
MAE (mm)	36.45	40.89	55.75	38.21
Correlation Coefficient	0.87	0.77	0.60	0.81
NSC	0.79	0.63	0.45	0.69
Kudige				
RMSE (mm)	126.78	146.67	165.88	138.23
MAE (mm)	75.98	90.09	94.35	83.55
Correlation Coefficient	0.91	0.76	0.67	0.81
NSC	0.77	0.60	0.53	0.67

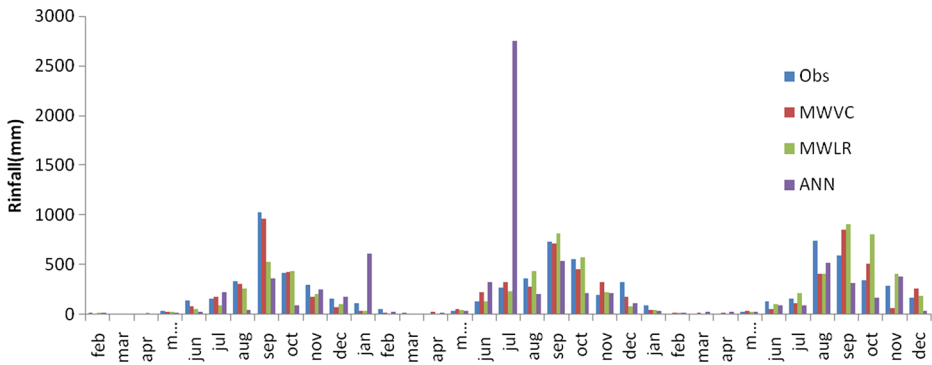


Fig. 11 Model Results for the three year period (1985–1987)

Figure 13 shows error percent for each of the model with reference to the observed values for the monsoon periods. The analysis of the plot show that for these periods indicates that the MWVC had a better performance during the extreme months. Similar results were obtained for Gurgoan also.

6.1 Analysis of the Influence of the Climate Indices in Driving the Rainfall

In order to evaluate the influence of the climate indices on the rainfall, a WVC model was developed without climate indices and compared with the MWVC model. The comparison of the results of the WVC model without climate indices is tabulated in Table 4. It was seen that in all the three cases, including the climate indices as the model input variables drastically improved the model performance. In the case of Kudige, including the climate indices increased the NSC from 0.71 to 0.77. These figures imply that about 15 % of the total variance is explained by the climate indices.

From the results obtained from the orthogonal least squares analysis, it was observed that the SOI index and IOD index are the major regressors when compared with the PDO indices. More research has to be pursued to bring out the reason for this kind of behaviour.

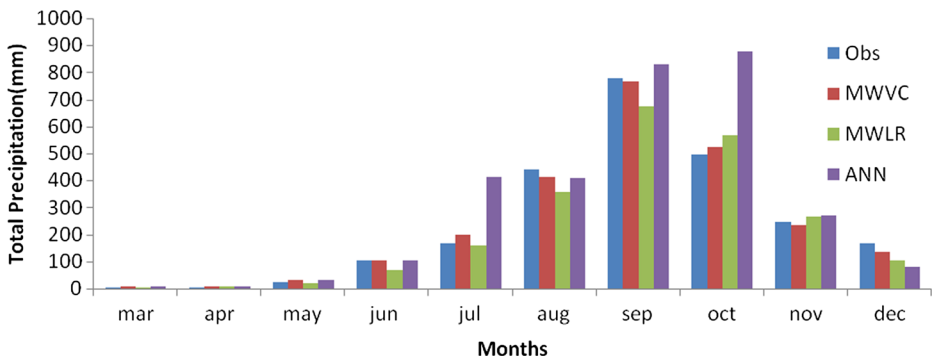


Fig. 12 Average forecast performance for MHHalli

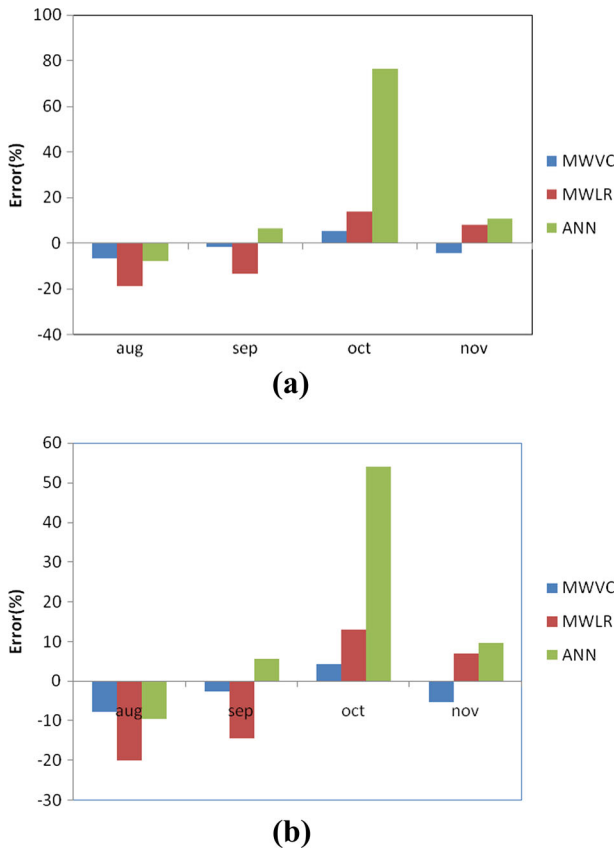


Fig. 13 Percentage of error in the different models during the monsoon seasons for **a** MHHalli and **b** Gurgoan

7 Conclusions

A wavelet based nonlinear method is presented, tested and discussed for forecasting monthly rainfall. The proposed wavelet based multivariate nonlinear model combines the wavelet decompositions of the candidate predictor variables (historical rainfall and climate indices using the MISO 2nd order Volterra model). The proposed method is compared with two other competent models such as MWLR and ANN models. The analysis of the results reveal that the MWVC models have better performance than the other contesting models. In fact, the linear nature of MWLR model estimators makes it inadequate to provide good prognostics for a variable characterized by a highly nonlinear physics. On the other hand the ANN model even though equipped with the ability of picking up the nonlinear features, but it performed poorly because of their inability to pick up the nonlinearity in the rainfall dynamics.

The proposed model was trained with the 53-year data, and tested with the remaining 20-year data, and compared to the traditional ANN model based on original time series. The WMVC forecast skill appears to be significantly better than the ANN and MWLR model. The WMVC model reduced mean absolute error by 16 % and increased the NSC by 26 %, respectively, in comparison to the WMLR model. The improvement is more significant for the extreme wet months, and for the dry months there is no significant improvement over the

MWLR and ANN models. These results indicate that the WMVC can capture the nonlinear impacts of predictor variables on rainfall series under different time scales. Also, the analysis clearly shows that there is a significant improvement in the model performance (15 % increase in model performance) by the inclusion of the climate variables in the forecasting model.

Acknowledgments The authors are grateful to the Department of Science and Technology for providing funds for the execution of the work through the INSPIRE FACULTY AWARD of the first author.

References

- Abbot J, Marohasy J (2012) Application of artificial neural networks to rainfall forecasting in Queensland, Australia. *Adv Atmos Sci* 29(4):717–730
- Adamowski J, Chan HF (2011) A wavelet neural network conjunction model for groundwater level forecasting. *J Hydrol* 407(1):28–40
- Addison P.S. (2002) *The illustrated wavelet transform handbook: introductory theory and applications in science, engineering, medicine and finance*. CRC Press
- Anctil, F. and Coulibaly, P., (2004). Wavelet analysis of the interannual variability in southern Quebec streamflow. *Journal of Climate*, 17(1)
- Bustamante, J., Gomes, J., Chou, S. and Rozante, J. (1999) Evaluation of April 1999 rainfall forecasts over South America using the Eta Model. INPE, Cachoeira Paulista, Sp Brasil
- Chattopadhyay G, Chattopadhyay S, Jain R (2010) Multivariate forecast of winter monsoon rainfall in India using SST anomaly as a predictor: neurocomputing and statistical approaches. *Compt Rendus Geosci* 342(10):755–765
- Chen S, Billings S (1989) Representations of non-linear systems: the NARMAX model. *Int J Control* 49(3): 1013–1032
- Daubechies, I., 1992. *Ten lectures on wavelets*, 61. SIAM
- Dhanya C, Nagesh Kumar D (2011) Multivariate nonlinear ensemble prediction of daily chaotic rainfall with climate inputs. *J Hydrol* 403(3):292–306
- Flandrin P (1988) A time-frequency formulation of optimum detection. *Acoust Speech Signal Process IEEE Trans* 36(9):1377–1384
- Freiwan M, Cigizoglu HK (2005) Prediction of total monthly rainfall in Jordan using feed forward backpropagation method. *Fresenius Environ Bull* 14(2):142–151
- French MN, Krajewski WF, Cuykendall RR (1992) Rainfall forecasting in space and time using a neural network. *J Hydrol* 137(1):1–31
- Garbrecht JD, Zhang XC, Schneider JM, Steiner JL (2010) Utility of seasonal climate forecasts in management of winter-wheat grazing. *Applied Eng Agric* 26(5):855–866
- Georgakakos KP, Bras RL (1984) A hydrologically useful station precipitation model: 1. Formulation *Water Resour Res* 20(11):1585–1596
- Grossmann A, Morlet J (1984) Decomposition of hardy functions into square integrable wavelets of constant shape. *SIAM J Math Anal* 15(4):723–736
- He, X., Guan, H., Zhang, X. and Simmons, C.T., A (2013) wavelet-based multiple linear regression model for forecasting monthly rainfall. *International Journal of Climatology*
- Immerzeel WW, Van Beek LP, Bierkens MF (2010) Climate change will affect the Asian water towers. *Science* 328(5984):1382–1385
- Jeong C, Shin J-Y, Kim T, Heo J-H (2012) Monthly precipitation forecasting with a neuro-fuzzy model. *Water Resour Manag* 26(15):4467–4483
- Karunanithi N, Grenney WJ, Whitley D, Bovee K (1994) Neural networks for river flow prediction. *J Comput Civ Eng* 8(2):201–220
- Kisi O (2009) Neural networks and wavelet conjunction model for intermittent streamflow forecasting. *J Hydrol Eng* 14(8):773–782
- Maheswaran R, Khosa R (2012a) Comparative study of different wavelets for hydrologic forecasting. *Comput Geosci* 46:284–295
- Maheswaran R, Khosa R (2012b) Wavelet Volterra coupled model for monthly stream flow forecasting. *J Hydrol* 450:320–335
- Mantua NJ, Hare SR (2002) The Pacific decadal oscillation. *J Oceanogr* 58(1):35–44

- Mantua NJ, Hare SR, Zhang Y, Wallace JM, Francis RC (1997) A Pacific interdecadal climate oscillation with impacts on salmon production. *Bull Am Meteorol Soc* 78(6):1069–1079
- Marzano FS, Fionda E, Ciotti P (2006) Neural-network approach to ground-based passive microwave estimation of precipitation intensity and extinction. *J Hydrol* 328(1):121–131
- Meinke H, Stone RC (2005) Seasonal and inter-annual climate forecasting: the new tool for increasing preparedness to climate variability and change in agricultural planning and operations. *Clim Chang* 70(1–2):221–253
- Meinke H, Sivakumar M, Motha RP, Nelson R (2007) Preface: climate predictions for better agricultural risk management. *Crop and Pasture Science* 58(10):935–938
- Moustris KP, Larissi IK, Nastos PT, Paliatatos AG (2011) Precipitation forecast using artificial neural networks in specific regions of Greece. *Water Resour Manag* 25(8):1979–1993
- Murphy BF, Timbal B (2008) A review of recent climate variability and climate change in southeastern Australia. *Int J Climatol* 28(7):859–879
- Mutai CC, Ward MN (2000) East African rainfall and the tropical circulation/convection on intraseasonal to interannual timescales. *J Clim* 13(22)
- Navone H, Ceccatto H (1994) Predicting Indian monsoon rainfall: a neural network approach. *Clim Dyn* 10(6–7):305–312
- Olson DA, Junker NW, Korty B (1995) Evaluation of 33 years of quantitative precipitation forecasting at the NMC. *Weather Forecast* 10(3):498–511
- Olsson J et al (2004) Neural networks for rainfall forecasting by atmospheric downscaling. *J Hydrol Eng* 9(1):1–12
- Percival DB, Walden AT (2000) *Wavelet methods for time series analysis* (Cambridge series in statistical and probabilistic mathematics)
- Prasad K, Dash S, Mohanty U (2010) A logistic regression approach for monthly rainfall forecasts in meteorological subdivisions of India based on DEMETER retrospective forecasts. *Int J Climatol* 30(10):1577–1588
- Rivera D, Lillo M, Uvo CB, Billib M, Arum  JL (2012) Forecasting monthly precipitation in central Chile: a self-organizing map approach using filtered sea surface temperature. *Theor Appl Climatol* 107(1–2):1–13
- Shukla J, Paolino DA (1983) The southern oscillation and long-range forecasting of the summer monsoon rainfall over India. *Mon Weather Rev* 111(9):1830–1837
- Sivakumar B (2001) Is a chaotic multi-  fractal approach for rainfall possible? *Hydrol Process* 15(6):943–955
- Sivapragasam C, Liong S, Pasha M (2001) Rainfall and runoff forecasting with SSA-SVM approach. *J Hydroinf* 3:141–152
- Swaminathan MS (1998) Padma Bhusan Prof. P. Koteswaram first memorial lecture, 23rd March 1998. In: *Climate and sustainable food security*, vol 28. Vayu Mandal, p 3–10
- Szilagyi J, Katul GG, Parlange MB, Albertson JD, Cahill AT (1996) The local effect of intermittency on the inertial subrange energy spectrum of the atmospheric surface layer. *Bound-Layer Meteorol* 79(1–2):35–50
- Thirumalaiah K, Deo M (1998) River stage forecasting using artificial neural networks. *J Hydrol Eng* 3(1):26–32
- Valverde Ram rez MC, De Campos Velho HF, Ferreira NJ (2005) Artificial neural network technique for rainfall forecasting applied to the Sao Paulo region. *J Hydrol* 301(1):146–162
- Webster PJ, Moore AM, Loschnigg JP, Leben RR (1999) Coupled ocean atmosphere dynamics in the Indian Ocean during 1997–98. *Nature* 401(6751):356–360