An MRA Based MLR Model for Forecasting Indian Annual Rainfall Using Large Scale Climate Indices

Abstract

A novel method for rainfall forecasting has been proposed using Multi Resolution Analysis (MRA). This approach decomposes annual rainfall series and long-term climate indices into component sub-series at different temporal scales, allowing for a more detailed analysis of the factors influencing annual rainfall. Multiple Linear Regression (MLR) is then used to predict annual rainfall, with climate indices sub-series as predictive variables, using a step-wise linear regression algorithm. The proposed model has been tested on Indian annual rainfall data and compared with the traditional MLR model. Results show that the MRA-based model outperforms the traditional model in terms of relative absolute error and correlation coefficient metrics. The proposed method offers several advantages over traditional methods as it can identify underlying factors affecting annual rainfall at different temporal scales, providing more accurate and reliable rainfall forecasts for better water resource management and agricultural planning. In conclusion, the MRA-based approach is a promising tool for improving the accuracy of annual rainfall predictions, and its implementation can lead to better water resource management and agricultural planning.

Key Words: *climate indices; forecasting; MLR; MRA; rainfall; time series, India*.

Abbreviations

CC, Correlation Coefficient; DMI, Dipole Mode Index; DWT, Discrete Wavelet Transforms; ENSO, El Niño-Southern Oscillation; IOD, Indian Ocean Dipole; ML, Machine Learning; MLCC, Maximum Lag Correlation Coefficient, MLR, Multiple Linear Regression; MRA, Multi Resolution Analysis; PDO, Pacific Decadal Oscillation; RAE, Relative Absolute Error; SST, Sea Surface Temperature; TMLR, Traditional Multiple Linear Regression; WMLR, Wavelet-Based Multiple Linear Regression

1. Introduction

Timely and abundant amount of rainfall increase agricultural productivity which ensures food security for the citizens of a country. Agricultural productivity and quality water supply can be safeguarded by efficient rainfall prediction mechanism. However, scarcity of rainfall has a

negative impact also to aquatic ecosystem. Agriculture, water quality and aquatic ecosystem are highly correlated with daily and annual rainfall amount (Kusiak et al. 2012).

MRA refers to the processing of any image or signal where signal is processed at multiple resolutions. An MRA or Multiscale Approximation (MSA) technique indicates the practically useful Discrete Wavelet Transforms (DWT). Some of the most commonly used transformations include the Daubechies family, Symlet, B-Splines, Gabor, Coiflet, Meyer etc. (Daubechies 1992). However, wavelets was firstly suggested by Stephane Mallat (1989). It is mainly based on the pyramid methods of signal processing as introduced by Burt and Adelson (1983, 1987).

On the other hand, linear regression methodology can be conceptualized in multivariate level (Makridakis et al. 1982). This is considered as supervised learning algorithm to predict or forecast unknown dependent variable based on the known features or independent variables (Draper and Smith 1998; Shah and Sands 2021; Sandberg and Sands 2022). We can decompose the series of values of response and explanatory variables using MRA and use this information in the MLR model to get accurate prediction for response variable.

Accuracy of forecasts of precipitation for periods of large anomalies is more important than that of a normal precipitation period (Schneider and Garbrecht 2003). Change in rainfall pattern is likely relevant to large-scale climate variabilities, and maybe to global warming as well (Ummenhofer et al. 2009). However, Past methods of forecasting rainfall can be broadly classified into two categories: empirical and dynamical. The dynamical models such as general circulation models (GCMs) are conformed with the laws of physics, which have been used to forecast climate variables (Lim et al. 2009; DelSole and Shukla 2012; Schepen et al. 2012). Whereas, the empirical models are based on observational relationships of the predictand variable with various predictors. Unknown parameters are to be determined by regression or other optimization methods from the data (Sahai et al. 2000). The second one is more accurate and thus are used in agricultural planning (Meinke and Stone 2005). The empirical methods include statistical models (Mutai et al. 1998; Prasad et al. 2010) and Machine Learning (ML) algorithms (Sahai et al. 2000).

Generally, most rainfall forecasting models use only a set of climate-related variables or historical rainfall data as input. Rather in the rainfall forecast models a combination of historical rainfall data and other climatic attributes can be used simultaneously. Furthermore,

some studies have shown that the variability of Indian rainfall has been linked to several dominant large-scale climate signals. They include El Niño Southern Oscillation (ENSO) (Agilan and Umamahesh 2018), Indian Ocean Dipole (IOD) (Karumuri and Saji 2007), Pacific Decadal Oscillation (PDO) (Krishnamurthy and Krishnamurthy 2014). However, Precipitation often operate under a large range of temporal scales varying from one day to several decades (Tessier et al. 1996).

Several authors identified various atmospheric features like air pressure, air temperature, wind speed, relative humidity etc. to accurately model rainfall using distinctive ML algorithms including MLR (Diez-Sierra and del Jesus 2017; Basha et al. 2020; Vijayan et al. 2020). Ghosh et al. (2010) utilized DWT and MRA approaches to find the trend in rainfall data by using Daubechies (D4) and Haar filters at various scales. Paul and Birthal (2016) employed wavelet approach to find rainfall trend over India and its agro-climatic zones. They concluded that there was no overall trend in Indian rainfall but there were changes in rainfall pattern in certain agroclimatic zones during 1901-2002. Liyew and Melese (2021) utilized MLR algorithm to predict daily rainfall amount utilizing various environmental components. Be that as it may, production of an agricultural crop is dependent on the total cultivated area under the crop, climatic factors like rainfall, temperature, price of different agricultural inputs etc. This information can be utilized to forecast the time series of production of the crop after decomposition into multiple time scale. This additional information will improve the accuracy of forecasting future values of key variable (Paul and Garai 2021, 2022; Garai and Paul 2023) than the traditional model of forecasting where the lag relationship of the time series is utilized only. Therefore, rainfall prediction is also considered an important study area and indirectly related to agricultural produces and prices. Hence, Quilty and Adamowski (2021) integrated wavelet data driven framework with MODWT and Maximum Overlap Discrete Wavelet Packet Transform (MODWPT) to forecast streamflow, precipitation etc. which change over multiple timescales.

In order to optimally utilize the information contained in the data, an MRA based on the WT is employed. An effective rainfall forecasting model from the historical rainfall data and climate signals by incorporating the MRA and MLR model has been developed and presented in this article. It will be examined whether historical precipitation and large-scale climate signals are useful in annual rainfall forecasting in an area of a range of climate types and a big precipitation gradient.

2. Materials and methods

2.1.Study area

India is a country with diverse rainfall patterns that can be broadly categorized into four main regions, namely, the Himalayan region, the Indo-Gangetic plain, the Deccan plateau, and the coastal areas which are further classified into 36 rainfall subdivisions (Figure 1). The country's location in the tropics and its proximity to the Arabian Sea, Indian Ocean, and Bay of Bengal heavily influences its climate.

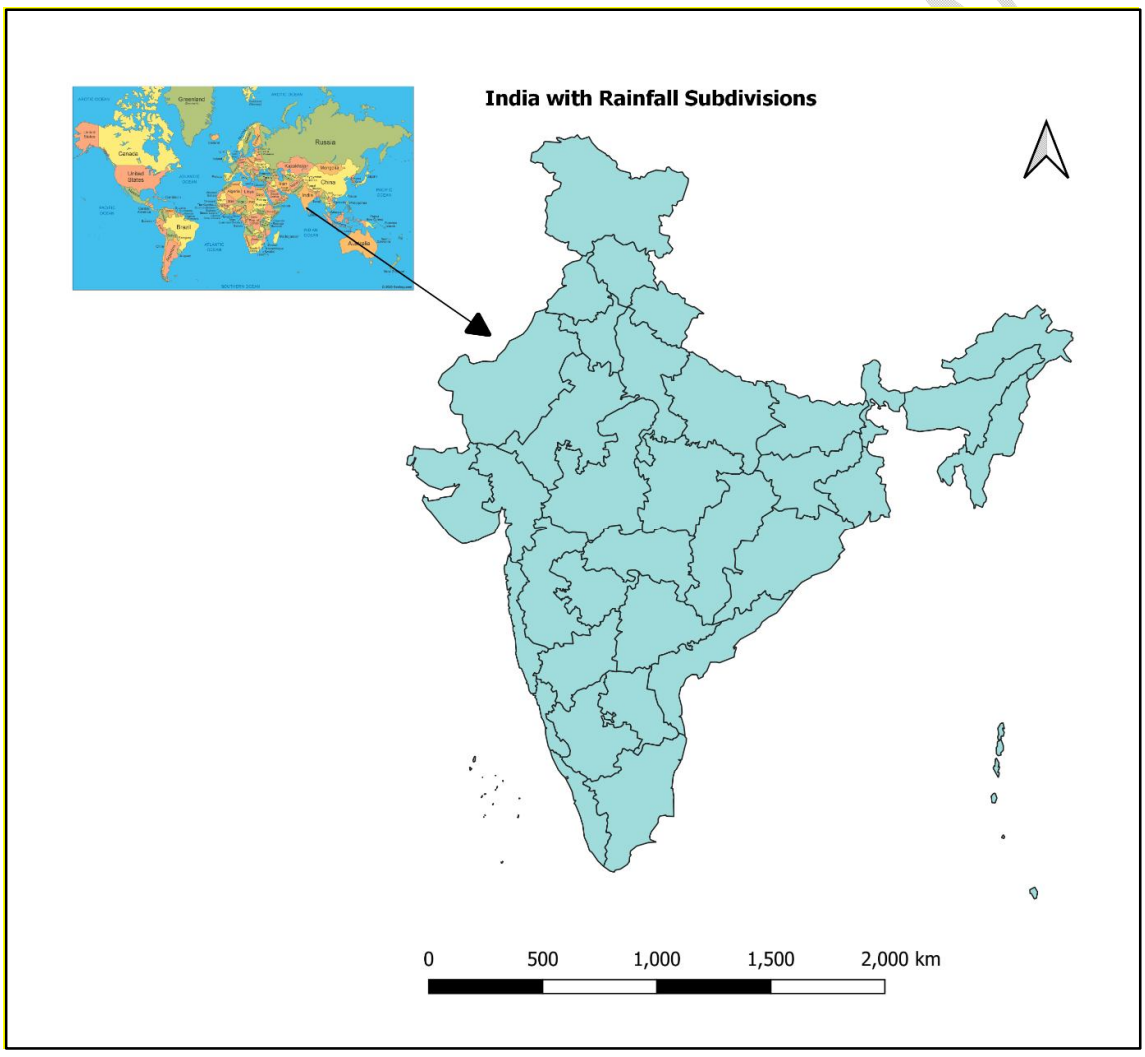


Figure 1. Indian rainfall subdivisions

The monsoon season, which typically lasts from June to September, brings heavy rainfall to most parts of the country. However, the Himalayan region receives most of its rainfall from the western disturbance system, while the Indo-Gangetic plain and Deccan plateau regions

rely heavily on the monsoon rains. The coastal areas of India, which include the west coast, east coast, and northeast region, receive rainfall from both the Arabian Sea and the Bay of Bengal. The country's agriculture heavily relies on the monsoon rains, making rainfall prediction and management critical for the country's food security.

2.2.Empirical Illustration

2.2.1. Rainfall data

Indian annual rainfall ranges from less than 1,000 millimeters (mm) in the west to over 2,500 mm in parts of the northeast. All India long term area weighted monthly, seasonal and annual rainfall data from the year 1901 to 2014 is obtained from the website (https://data.gov.in/sites/default/files/datafile/All_India_Area_Weighted_Monthly_Seasonal_And_Annual_Rainfall.xls) for this study.

2.2.2. Climate indices

Selected large-scale climate signals which are related to Indian rainfall for this study are: IOD and PDO. The possible relation between the Indian summer monsoon and the PDO observed in the Sea Surface Temperature (SST) of the North Pacific Ocean. Using long records of observations and coupled model simulation, it has been found that the warm (cold) phase of the PDO is associated with deficit (excess) rainfall over India. The PDO extends its influence to the tropical Pacific and modifies the relation between the monsoon rainfall and ENSO. During the warm PDO period, the impact of El Niño (cold period- La Niña) on the monsoon rainfall is enhanced (reduced). Hadley circulation in the monsoon region determines the impact of PDO on the monsoon rainfall. Knowing the phase of PDO may lead to better long-term prediction of the seasonal monsoon rainfall and also the impact of ENSO on monsoon can be known. PDO which is used as a common index for ENSO, was chosen as a potential predictor of the presence of Indian rainfall. PDO index data is obtained from the website-(http://jisao.washington.edu/pdo/PDO.latest).

Variability in the Indian Ocean is associated with variability of rainfall in India, This Variability can be described by IOD index which is represented by anomalous SST gradient between the western equatorial Indian Ocean (50◦E–70◦E and 10◦S–10◦N) and the south-eastern equatorial Indian Ocean (90◦E–110◦E and 10◦S–0◦N), outgoing longwave radiation, and sea surface height anomalies. This gradient is named as Dipole Mode Index (DMI). When the DMI is positive then, the phenomenon is refereed as the positive IOD and when it is negative, it is refereed as negative IOD. In this study, the DMI index derived from

HadISST dataset is also selected as a predictor of the Indian rainfall because it is frequently updated and has a relatively long period of record. The data is obtained from (https://www.esrl.noaa.gov/psd/gcos_wgsp/Timeseries/DMI).

2.3.Methodology

2.3.1. MLR

Generally, any particular response variable can be described by many influential variables. Moreover, MLR attempts to find the underlying relationship between the explanatory variables (X) and a response variable (Y) to fit a linear equation to the observed data. Every single value of X is associated with a value of Y. An MLR model with k predictor variables $X_1, X_2, \ldots, X_k$ and a response Y, can be written as:

$$y = \beta_0 + \beta_1 x1 + \beta_2 x_2 + \cdots + \beta_k x_k \quad (1)$$

$\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ mentioned in the model are regression coefficients, need to be estimated (de Andrade Lima Neto et al. 2021).

2.3.2. Wavelet-based approach

A wavelet is a mathematical function useful in digital signal processing and image compression. Although the theory is not new, the principles are similar to those of Fourier analysis. Wavelet is a 'small wave'. A small wave grows and decays essentially in a limited time period. The contrasting notion is obviously a 'big wave'. An example of big wave is the sine function, which keeps on oscillating up and down on a plot of $\sin(u)$ versus $u \in (-\infty, +\infty)$. To begin to quantify the notion of a wavelet, let us consider a real-valued function $\psi(.)$ defined over the real axis $(-\infty, +\infty)$ satisfying two basic properties:

(i) The integral of $\psi(.)$ is zero:

$$\int_{-\infty}^{+\infty} \psi(u)du = 0 \quad (2)$$

(ii) The square of $\psi(.)$ integrates to unity:

$$\int_{-\infty}^{+\infty} \psi^2(u)\, du = 1 \quad (3)$$

If Equation (3) holds then for any $\in$ satisfying $0 < \in < 1$, there must be an interval $[-T, T]$ of finite length such that

$$\int_{-T}^{+T} \psi^2(u)\, du > 1-\epsilon.$$

(4)

If we think of ϵ as being very close to zero, then $\psi(.)$ can only deviate insignificantly from zero outside of [-T, T] (equation 4). Since the length of the interval [-T, T] is vanishingly small compared to the infinite length of the entire real axis$(-\infty, +\infty)$, the non-zero activity of $\psi(.)$ can be considered as limited to relatively small interval of time. While Equation (3) tells us that $\psi(.)$ has to make some excursions away from zero, Equation (2) says that any excursions it makes above zero must be cleared away from zero, so $\psi(.)$ must resemble a wave. Hence Equations (2), (3), and (4) lead to a 'small wave' or wavelet. One important and common additional condition, namely, the so-called admissibility condition (Equation 6) should also be satisfied along with the above-mentioned ones. A wavelet $\psi(.)$ is said to be admissible if its Fourier transform (Equation 5), namely,

$$\psi(f) \equiv \int_0^{\infty} \psi(u)\, e^{-i2\pi fu}\, du$$

(5)

is such that

$$C_\psi \equiv \int_0^{\infty} \frac{|\psi(f)|^2}{f} df;\ 0 < C_\psi < \infty$$

(6)

2.3.2.1.WT

The WT is a mathematical technique introduced in signal analysis in early 1980s (Goupillaud et al. 1984). It is a method based on expressing signals as sums of little waves. The basic idea of the WT is the decomposition of a signal at different time scales onto a set of basic functions. The set of basic functions $\{\psi_{a,b}(t)\}$ (Equation 7) can be generated by translating and scaling the wavelet function $\psi(t)$, called the mother wavelet. According to Daubechies (1992):

$$\psi_{a,b}(t) = (1/\sqrt{a})[\psi(t-b)/a],\ a > 0,\ -\infty < b < \infty \qquad (7)$$

In Equation (7), a is the scale parameter which adjusts the dilation of the wavelet and b determines the location of the wavelet. The mother wavelet $\psi(t)$ satisfies two basic properties of wavelet mentioned above in Equations (2), and (3). However, there have been two mainstreams of wavelets. The first one is known as the Continuous Wavelet Transform (CWT), which produces redundant amount of subseries; and the second one is the DWT, which generates finite number of subseries. If $\{\psi(t)\}$ satisfies Equation (1), for a finite time

series (usually $t = 0,1,2,\dots,N-1$; where N denotes the number of observations in the series) or energy signal $f(t)$, CWT (Equation 8) is defined as

$$W_\psi(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} \bar{\psi}\left(\frac{t-b}{a}\right) f(t)\, dt \quad (8)$$

In Equation (8), $\bar{\psi}$ is the mother wavelet complex conjugate. For each scale a, the WT result is a set of coefficients associated with different locations b (Lindsay et al. 1996). The DWT can be thought as dyadic (concepts of two parts) sampling of $W_\psi(a,b)$, in which the mother wavelet is scaled by powers of two, $a = 2^j$ and, within a given scale, translated by integers, $b = k2^j$, where k is a location index and j is referred to as the decomposition level. Thus, from Equation (7), a discretely scaled and translated wavelet (Equation 9) is expressed as

$$\psi_{j,k}(t) = 2^{-j/2}\psi\left(2^{-j}t - k\right) \quad (9)$$

and the the DWT of $f(t)$ can be written as in Equation (10)

$$W_{\psi_{j,k}(t)} = 2^{-j/2} \int_{-\infty}^{+\infty} \bar{\psi}\left(2^{-j}t - k\right) f(t)\, dt\,; j = 0,1,2,\dots; k \in \mathbf{Z} \quad (10)$$

The characteristics of the original time series $f(t)$ at the decomposition level j and time location index k at the same time are reflected by $W_{\psi_{j,k}(t)}$. When the time domain resolution of WT is high, j becomes small. As the level and the scale decrease, the time domain resolution increases and the smaller and finer components of the signal can be accessed.

2.3.2.2.MRA

MRA based on DWT is to decompose the signal with different frequencies into a certain number of component time series at different temporal scales. In order to perform the MRA, the DWT is implemented in a hierarchical algorithm, well known as the pyramid algorithm (Figure 2) (Mallat 1989). The MRA decomposes the signal into different scales by successively translating and convolving the elements of a high-pass filter (which passes signals with a frequency higher than a certain cutoff frequency and attenuates signals with frequencies lower than the cutoff frequency) and low-pass scaling filter (which rejects all unwanted higher frequencies of signals) associated with the mother wavelet (Primer et al. 1998; Percival and Walden 2000). These filters retain the small-and large-scale components of the signals respectively (Mart'inez and Gilabert 2009), also known as detail (D) (Equation 11) and approximation (A) (Equation 12) subseries. For a particular decomposition level j,

the sum of the products of the wavelets and their coefficients for the signal $f(t)$ over all locations (but for one value of the scale parameter 2^j) results in the detail component D_j, i.e.,

$$D_j(t) = \sum_{k=-\infty}^{\infty} W_{\psi_{j,k}} \psi_{j,k}(t)$$

(11)

Beside the detail component, it also results in a smoothed representation of the signal for scale 2^j, also known as approximation component A_j and described as (Percival and Walden 2000):

$$A_j(t) = \sum_{k=-\infty}^{\infty} V_{\phi_{j,k}} \phi_{j,k}(t) \quad (12)$$

Where $\phi_{j,k}(t)$ is a scaled and translated basis function, called the scaling function, which is given together with the wavelet function when a wavelet is chosen. $V_{\phi_{j,k}}$ is the scaling coefficient calculated from $\phi_{j,k}(t)$ in a similar way for the wavelet coefficient $W_{\psi_{j,k}}$ from $\psi_{j,k}(t)$. The signal $f(t)$ can be reconstructed from the approximation and detail components as in Equation (13):

$$f(t) = D_1(t) + D_2(t) + \ldots + D_J(t) + A_J(t)$$

(13)

In Equation (13), J is the highest decomposition level considered. In the first level of the decomposition, $f(t) = D_1(t) + A_1(t)$, the signal has a low-pass filtered component A_1, and a high-pass filtered component D_1. The same procedure is performed on A_1 in order to obtain a decomposition at coarser scales, $A_1 = D_2 + A_2$. The process is continued in such a way that $A_j = D_{j+1} + A_{j+1}$ for $j = 2, \ldots, J-1$. This is known as pyramid algorithm depicted in the Figure 2.

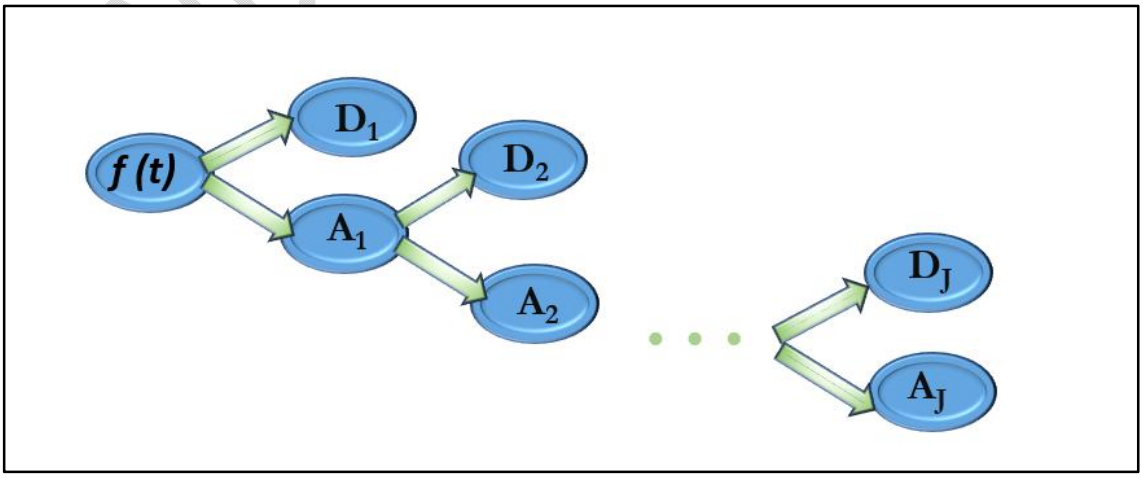


Figure 2. Pyramid Algorithm

2.3.3. Lag time estimates:

We may assume that rainfall responds to large-scale climate signals with a time lag in this study. Cross correlation between the two sets of signals are calculated for the lag relationship. The lag time estimation is determined by finding the time shift resulting in a maximum cross correlation. All monthly time series are decomposed using the MRA into a certain number of subseries components under different temporal scales. The lag correlation coefficients (LCC) are measured between rainfall subseries versus each potential predictor subseries (Equation 14). Let $y_s(t_n)$ denote the rainfall subseries, and $x_s(t_n - \tau)$) denote the lagged rainfall or largescale climate index subseries, where $n = 0, 1, \ldots, N; t_n = t_0 + n\,\Delta t$; t_0 is the offset, and Δt is the sampling interval. The lag correlation coefficient between $y_s(t_n)$ and $x_s(t_n)$ is defined as

$$LCC(\tau) = \frac{\left|\sum_{n=\tau}^{N}[y_s(t_n)-\bar{y_s}][x_s(t_{n-\tau})-\bar{x_s}]\right|}{\sqrt{\sum_{n=\tau}^{N}[y_s(t_n)-\bar{y_s}]^2\sum_{n=\tau}^{N}[x_s(t_{n-\tau})-\bar{x_s}]^2}}$$

(14)

where $\bar{y_s}$ and $\bar{x_s}$ stand for the mean of $y_s(t_n)$ and $x_s(t_{n-\tau})$, respectively, and $\tau = 1, 2, \ldots$. Here, in order to forecast rainfall in advance, we take $\tau \geq 1$. The lag time between the subseries $y_s(t_n)$ and $x_s(t_n)$ is found from the peak of $LCC(\tau)$. The lag time between the original rainfall series $y(t_n)$ and predictor series $x(t_n)$ can be found similarly. By doing this, an optimal time lag is determined for each potential predictor variable for further developing rainfall forecasting models. This is different from some other methods in which time series of different lags are included in the model optimisation stage, as discussed in Beriro et al. (2012).

2.3.4. Proposed models for rainfall forecasting

After the lag relation between rainfall and each candidate predictor is identified, two types of MLR models can be constructed which are discussed below.

2.3.4.1.Traditional Multiple Linear Regression (TMLR) models

Model based on original rainfall series and climate signal series is the TMLR model (Equation 15) given as

$$\hat{y}(t_n) = a_0 + a_1x_1(t_{n-\tau_1}) + \cdots + a_ix_i(t_{n-\tau_i}) + \cdots + a_Ix_I(t_{n-\tau_I}) \quad (15)$$

where $\widehat{y}(t_n)$ is the forecasted rainfall value, x_i $(i = 1, \ldots, I)$ are the potential predictors including HRA and large-scale climate indices, $\tau_i \geq 1$ is the lag time between rainfall and the i^{th} predictor, and a_i is the model regression parameter. The a_i values are estimated from the training period data, i.e., (1901-1995). In order to prevent overfitting, and to find the optimal (final) regression model, a stepwise regression algorithm (Draper and Smith 1998) is employed to select significant predictors from all the candidate variables in this study. Finally, the forecasted rainfall values are computed.

2.3.4.2.Wavelet-based Multiple Linear Regression (WMLR) models

Model based on wavelet-decomposed subseries of the predictor series is WMLR model. The WMLR is constructed by incorporating two methods, MLR and MRA with DWT. For the WMLR model inputs, each of the original rainfall and climate index time series are decomposed into a certain number of subseries components A_J and D's by the MRA. Then the forecasted value of the approximation component of the rainfall in the highest decomposition level J, i.e., $\widehat{A_J^y}(t_n)$ can be obtained by Equation (16):

$$\widehat{A_J^y}(t_n) = a_{J,1} + a_{J,2}A_J^{x_1}\left(t_{n-\varsigma_{J,1}}\right) + \cdots + a_{J,i}A_J^{x_i}\left(t_{n-\varsigma_{J,i}}\right) + \cdots + a_{J,I}A_J^{x_I}\left(t_{n-\varsigma_{J,I}}\right) \quad (16)$$

Where $A_J^{x_i}$ $(i = 1, 2, 3 \ldots I)$ are the approximation components of the potential predictor variables x_i $(i = 1, \ldots, I)$ at the highest decomposition level J, $a_{J,i}$ is the regression parameter which can be estimated from the training period data, and $\varsigma_{J,i}, i \geq 1$ is the lag time between the rainfall subseries A_J^y and predictor subseries $A_J^{x_i}$. Similarly, the forecasted value of the detail components of the rainfall series at each decomposition level j, i.e., $D_j^y(t_n)$ $(j = 1, \ldots, J)$ can be expressed as Equation (17):

$$\widehat{D_j^y}(t_n) = d_{j,1} + d_{j,2}D_j^{x_1}\left(t_{n-\tau_{j,1}}\right) + \cdots + d_{j,i}D_J^{x_i}\left(t_{n-\tau_{j,i}}\right) + \cdots + d_{j,I}D_j^{x_I}\left(t_{n-\tau_{j,I}}\right)$$

(17)

where $D_j^{x_i}$ $(i = 1, \ldots, I)$ is the detail components of the potential predictor x_i at the decomposition level j, and $d_{j,i}$ and $\tau_{j,i} \geq 1$ are the regression parameter and the time lag, respectively. From Equations (13), (16) and (17), we can get the forecasted value of the rainfall, i.e., $\widehat{y}(t_n)$ as in Equation (18):

$$\widehat{y}(t_n) = \widehat{A_J^y}(t_n) + \sum_{j=1}^{J}\widehat{D_j^y}(t_n) = c_0 + \sum_{j=1}^{J}a_{J,i}A_J^{x_i}\left(t_{n-\varsigma_{J,i}}\right) + \sum_{j=1}^{J}\sum_{i=1}^{I}d_{j,i}D_J^{x_i}\left(t_{n-\tau_{j,i}}\right) \quad (18)$$

Where $c_0 = a_{J,1} + d_{j,1}$. Similar to TMLR, the stepwise regression algorithm is performed to select significant predictors from all candidate component variables for WMLR.

2.3.5. Implementation and evaluation of the WMLR and TMLR

SAS codes are written for finding the correlation coefficient matrix of Indian annual rainfall and monthly PDO and DMI indices. The procedure begins with the finding of maximum significant correlation (at 5% level of significance) between annual rainfall and monthly indices. Max significant correlated months for each of the indices are selected as potential predictor for fitting the model to predict the annual rainfall. MODWT is carried out on the basis of 'Haar' wavelet filter at level 6 to annual rainfall series and the predictor series using R environment. After that each decomposed subseries of rainfall and corresponding different predictors are fitted into linear regression model. Prediction for those fitted models are obtained and combined for getting the final prediction for rainfall. This whole process involves the WMLR model.

In traditional regression model the monthly indices which are having high and significant correlation (at 5% level of significance) with annual rainfall are selected as predictor variables. Then linear regression model is fitted using R codes. 'Stepwise' algorithm is used for inclusion of those predictors which are significant predictors from all the candidate variables. Prediction on the basis of this model is done and result is obtained.

The Relative Absolute Error (RAE) and correlation coefficient (CC) statistics are used to assess performance of WMLR, in comparison to that of TMLR. The CC shows the degree to which two variables are linearly related. The information about the predictive capability of the models is measured through RAE. The RAE is defined as in Equation (19):

$$RAE = \sum_{i=1}^{M} \frac{|Y_{for,i} - Y_{obs,i}|}{Y_{obs,i}} \tag{19}$$

and the correlation coefficient as in Equation (20):

$$CC = \frac{\sum_{i=1}^{M} (Y_{obs,i} - \overline{Y}_{obs})(Y_{for,i} - \overline{Y}_{for})}{\sqrt{\sum_{i=1}^{M} (Y_{obs,i} - \overline{Y}_{obs})^2 \sum_{i=1}^{M} (Y_{for,i} - \overline{Y}_{for})^2}} \tag{20}$$

Where $Y_{obs,i}$ and $Y_{for,i}$ stand for observed and forecasted annual rainfall respectively for the i^{th} time step; M is the number of time steps in the test period. $\overline{Y}_{obs}$ and $\overline{Y}_{for}$ stand for mean of observed and forecasted rainfall values, respectively. Both WMLR and TMLR models were trained with the data of 94 -year period (1901-1994). The trained models were used to forecast annual rainfall in test period from 1995 to 2014. Last 20 years data was taken for testing and validation of the newly developed models.

3. Results and discussion

The descriptive statistics of annual rainfall and monthly PDO indices are presented in the Table 1. Same kind of information is obtained for annual rainfall and monthly DMI, which is provided in Table 2. Number of observations is 114 for each of the cases. 'Std Dev' in both the tables indicates standard deviation. 'ANN' indicates Annual rainfall series and other variables are monthly representations from January to December of the particular indices (Tables 1, and 2).

Table 1. Descriptive Statistics of all India annual rainfall and monthly PDO indices

Variable	Mean	Std Dev	Minimum	Maximum
ANN	1176	106.65	947.1	1464
JAN	-0.04	1.05	-2.48	2.14
FEB	-0.005	1.06	-3.6	2.07
MAR	0.06	0.97	-2.56	2.41
APR	0.19	0.98	-2.17	2.37
MAY	0.23	0.97	-2.23	2.32
JUN	0.15	1.07	-2.44	3.01
JUL	0.07	1.09	-2.93	3.51
AUG	-0.09	1.05	-2.25	3.31
SEP	-0.17	0.95	-2.28	2.44
OCT	-0.14	0.99	-2.8	2.1
NOV	-0.13	1.05	-3.08	2.65
DEC	0.0007	1.06	-2.75	2.51

Table 2. Descriptive Statistics of Indian annual rainfall and monthly DMI

Variable	Mean	Std Dev	Minimum	Maximum
ANN	1176	106.65	947.1	1464
JAN	0.04	0.25	-0.53	0.69
FEB	0.03	0.26	-0.67	0.72
MAR	0.03	0.25	-0.58	0.62
APR	0.01	0.25	-0.58	0.67
MAY	0.006	0.26	-0.57	0.74
JUN	0.009	0.3	-0.72	0.93
JUL	0.017	0.37	-0.76	1.01
AUG	0.006	0.4	-0.95	1.09
SEP	0.02	0.44	-1.23	1.14
OCT	0.02	0.4	-0.72	1.24
NOV	0.03	0.32	-0.61	1.52

DEC	0.04	0.25	-0.57	1.08

3.1.Significant predictor variables in the WMLR and TMLR models

Correlation coefficient analysis was carried out to obtain indices which were significantly related to annual rainfall. predictors Correlation coefficient of annual rainfall and monthly PDO index is given Table 3.

Table 3. Pearson Correlation Coefficients of ANN and monthly PDO indices

	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NEV	DEC
ANN	0.17	0.11	0.11	0.07	0.15	0.07	-0.07	-0.11	-0.23	-0.23	-0.19	-0.24
p-value	0.06	0.23	0.22	0.46	0.10	0.42	0.44	0.22	0.01	0.01	0.03	0.008

From Table 3 it can be found that correlation coefficient of PDO index for September, October and November December months with annual rainfall is significant. So, we can choose indices of any or all of these months as significant predictor variable for forecasting annual rainfall for both the WMLR and TMLR models.

Table 4. Pearson Correlation Coefficients of ANN and monthly DMI, N = 114

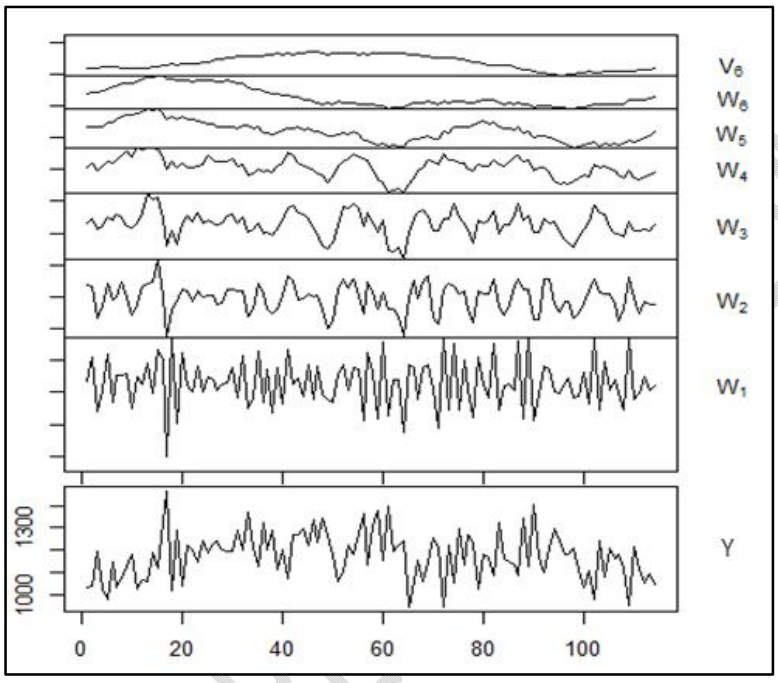
	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NEV	DEC
ANN	0.04	-0.01	-0.08	-0.05	-0.007	-0.13	0.06	-0.10	-0.27	-0.28	-0.19	-0.11
p-value	0.61	0.90	0.36	0.99	0.53	0.10	0.51	0.26	0.001	0.002	0.04	0.24

From Table 4 it is clear that September, October, November months' IOD index i.e. DMI are significantly correlated with annual rainfall. All the 3 months can be selected as explanatory variable. Important thing is that if you include every month's data for the two predictors discussed above in the model, through stepwise algorithm only significant months for the prediction purpose will be kept.

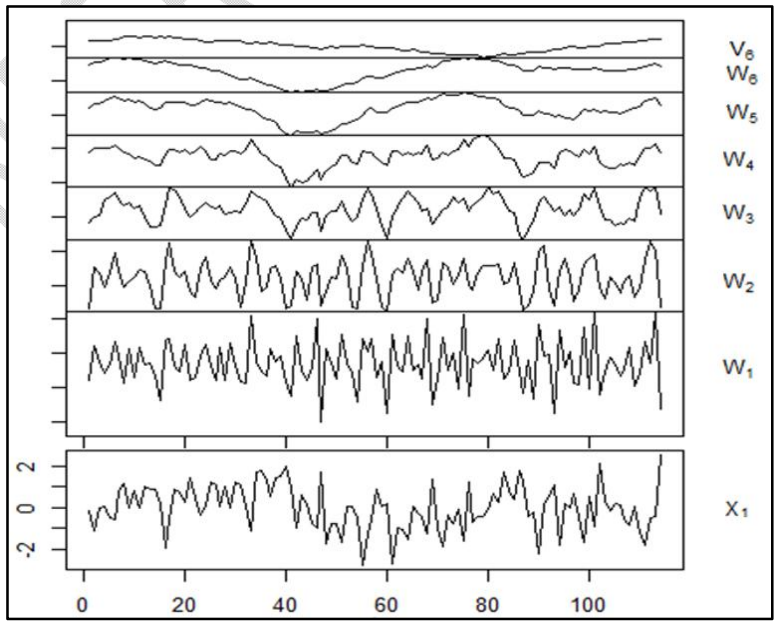
3.2.Multi Resolution decomposition of rainfall and predictor series

MODWT is carried out on the basis of 'Haar' wavelet filter at level 6. Figure 3(a) represents the 6 wavelet coefficients of annual rainfall in India over 114 years. The most significant PDO index, i.e., the month of December is decomposed through MRA process and presented in Figure 3(b). Similarly, the high correlation of DMI index with the Indian annual rainfall was found for the month of September. September month's DMI index time series data has been decomposed and coefficients are presented in Figure 3(c). The coefficients, i.e., W_1, W_2,

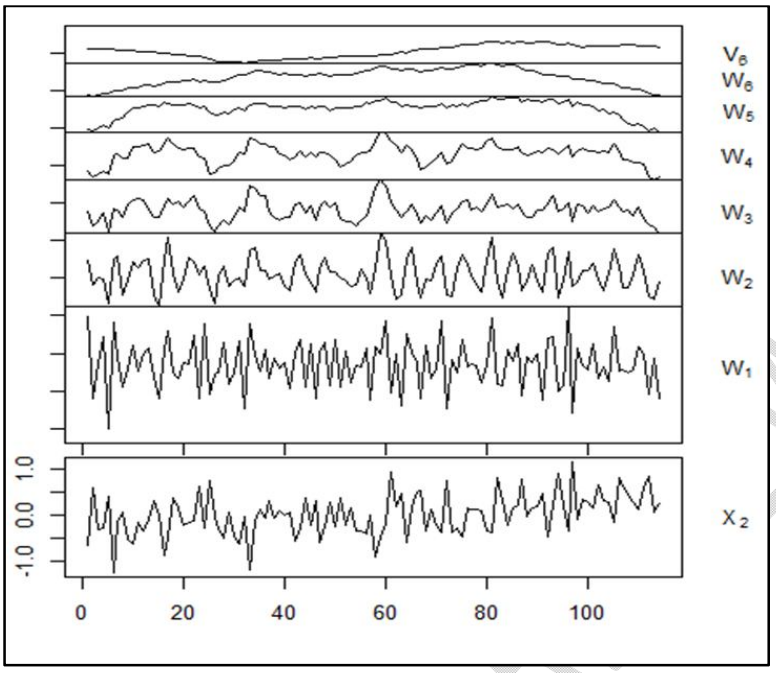
W_3, W_4, W_5, W_6 and V_6 are obtained using 'Haar' filter. The graph of W_2 is much smoother than the W_1. Similarly, smoothness increases as we are going to top of the graphs with upper coefficient. In the graph V_6 scaling coefficient shows the smooth lot and others are detailed coefficients. This smooth coefficient (V_6) is actual the trend component of signal, hidden in the noisy time series data.



(a)



(b)



(c)

Figure 3. MODWT plot for (a) annual rainfall, (b) Dec month PDO indices, and (c) Sep month DMI

3.3.Model fitting for WMLR and TMLR model

After decomposing the time series into detail (W) and smooth or approximation (V) component, Maximum LCC (MLCC) are found between predictand and predictor subseries using Equation (14). This reveals how strongly one predictor subseries is related to the corresponding rainfall subseries. It appears that MRA decomposed series are able to capture more details of rainfall correlations than the original time series. The question is whether they are also useful to improve the rainfall forecasts relative to the original series. This is discussed in subsequent sections. Correlation coefficients between response and predictor variables are given in below tables.

Table 5. Pearson Correlation Coefficients

	PV	DV		DEC (PDO)	SEP (DMI)		PD	DD
RV	-0.21	-0.71	ANN	-0.25	-0.28	RD	-0.36	-0.33
p-Value	0.03	<0.0001	p-Value	0.008	0.003	p-Value	<0.0001	0.0003

(N.B.: RV, PV, DV are indicating smooth component for rainfall, PDO index and DMI; RD, PD, DD are indicating D_1 component for rainfall, PDO index and DMI respectively)

In this study, current year predictor data for the prediction of annual rainfall have been used. Two indices have been used for the prediction purpose. By observing the correlation measures from Table 5, it is clear that decomposition of data increases the correlation between predictor and predictand variables. In WMLR model, after the multi resolution decomposition of rainfall and predictor series predicted value $\widehat{A_j^y}(t_n)$ of the approximation component of rainfall in the highest decomposition level J=6 is obtained through using Equation (16). Likewise, predicted value $\widehat{D_j^y}(t_n)$ $(j = 1,2 \dots 6)$ of the detail components of the rainfall series at each decomposition level j is obtained using Equation (17). And finally, combining these forecasted values as in Equation (18), we get the forecasted value $\hat{y}(t_n)$ of the rainfall. Similarly, in case of TMLR, original predictor series are used to predict the rainfall. Firstly, predictor variables may be forecasted and those values can be used in the models to get the forecasted rainfall. Forecasted rainfall values from TMLR and WMLR models for the test periods are obtained from Equations (15) and (18) respectively. Forecasted values by these models have been presented in Figure 4 for visual comparison with the observed rainfall data.

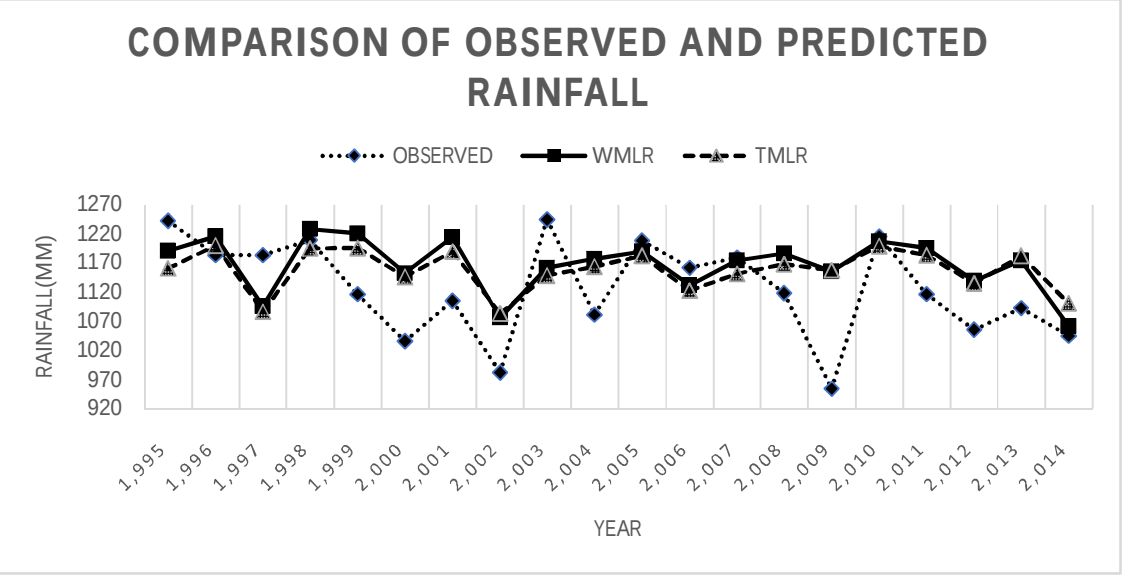


Figure 4. Line chart Comparison between observed and forecasted annual rainfall through WMLR and TMLR models

3.4.Evaluation of the WMLR and TMLR performance

The performance statistics of WMLR and TMLR models are calculated for the test data (1995-2014). RAE values for WMLR and TMLR models are 1.08 and 1.30 respectively (Table 6). Similarly, CC for these two models are 0.46 and 0.34 respectively (Table 6).

Table 6. RAE and CC values for two models

	WMLR	TMLR
RAE	1.08	1.30
CC	0.46	0.34

The WMLR model for rainfall forecasting in India has shown significantly improved accuracy compared to the traditional TMLR model. The CC metric from the WMLR model is 0.112 points higher than that of the TMLR model, and the RAE is reduced by 1%, indicating that the WMLR model outperforms the traditional model in forecasting annual rainfall in India. The superior performance of the WMLR model may be attributed to its ability to capture the impacts of predictor variables on rainfall at different time scales, which the traditional model cannot do. Although the number of significant predictor variables in the regression based on the decomposition subseries is greater than that based on the original time series, the input data for both models is the same. By utilizing MRA to decompose the original predictor series, the WMLR model optimally utilizes the information contained in the original input data, resulting in improved forecasting skill. The sixth level of decomposition was used in this study, assuming that the most useful information between annual rainfall and potential predictor variables is included within these time scales.

4. Conclusion

This study presents an MRA-based MLR model for forecasting annual rainfall in India. The proposed WMLR model combines MRA for both predictand and candidate predictor variables, and was trained on 94 years of data and tested on the remaining 20 years. The WMLR model outperforms the traditional TMLR model based on original time series, reducing relative absolute errors and increasing the correlation coefficient. The WMLR model also benefits from using DWT, which eliminates multicollinearity issues in the regression problem. This study used only two indices variables, but future research could incorporate additional variables like atmospheric temperature, water vapor, number of sunny days, and relative humidity to further improve forecasting accuracy. Additionally, exploring

other wavelet filters such as Mexican hat, other Daubechies family of filters (D4, D6, etc.), Symlet, Morlet, B-Splines, Meyer, and multiple levels of decompositions could also enhance the forecasting performance. Overall, this study provides promising results for the potential of MRA-based forecasting models for rainfall prediction in India.

Highlights:

- Large-scale climate indices have been included to improve the rainfall forecasting performance.
- Suitable indices have been chosen using MLCC.
- A new forecasting model based on wavelet decomposition and MLR algorithm (WMLR) has been proposed.
- Empirical study of the newly proposed method has been carried out on Indian annual rainfall.

Software availability

The open-source R-package 'wavelets' developed by Eric Aldrich has been used for present analysis and the package is available for download from the following website https://CRAN.R-project.org/package=wavelets.

COMPETING INTERESTS:

Authors have declared that they have no known competing financial interests OR non-financial interests OR personal relationships that could have appeared to influence the work reported in this paper.

Availability of data and materials: available on request

References

Agilan V, Umamahesh N (2018) Covariate and parameter uncertainty in non□stationary rainfall IDF curve. Int J Climatol 38(1):365-383

Basha CZ, Bhavana N, Bhavya P, Sowmya V (2020) Rainfall prediction using machine learning \& deep learning techniques. In: 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC). pp 92–97

Beriro DJ, Abrahart RJ, Mount NJ, Nathanail CP (2012) Letter to the Editor on "Precipitation Forecasting Using Wavelet-Genetic Programming and Wavelet-Neuro-Fuzzy

Conjunction Models" by Ozgur Kisi & Jalal Shiri. Water Resour Manag 26(12)3653–3662

Burt P, Adelson E (1983) A multiresolution spline with application to image mosaics. ACM Trans Graph 2:217-236

Burt PJ, Adelson EH (1987) The Laplacian pyramid as a compact image code. In: Readings in computer vision. Elsevier, pp 671–679

Daubechies I (1992) Ten Lectures on Wavelets SIAM Philadelphia

de Andrade Lima Neto E, Pinheiro A, de Oliveira Ferreira A (2021) On wavelet to select the parametric form of a regression model. Commun Stat Comput 50:2619–2642

Diez-Sierra J, del Jesus M (2017) A rainfall analysis and forecasting tool. Environ Model \& Softw 97:243–258

Draper NR, Smith H (1998) Applied Regression Analysis Wiley- Inter-science: Hoboken de Andrade Lima Neto E, Pinheiro A, Gomes de Oliveira Ferreira A (2021) On wavelet to select the parametric form of a regression model. Comm Statist Simulation Comput 50(9):2619-2642

Delsole T, Shukla J (2012) Climate models produce skilful predictions of Indian summer monsoon rainfall. Geophys Res Lett 39:L09703

Garai S, Paul RK (2023) Development of MCS Based-Ensemble Models Using CEEMDAN Decomposition and Machine Intelligence. Intell Syst with Appl 18:200202

Ghosh H, Paul RK, Prajneshu (2010) Wavelet Frequency Domain Approach for Statistical Modeling of Rainfall Time-Series Data. J Stat Theory Pract 4(4):813-825

Goupillaud P, Grossmann A, Morlet J (1984) Cycle-octave and related transforms in seismic signal analysis. Geoexploration 23(1):85–102

Karumuri A, Saji NH (2007) On the impacts of ENSO and Indian Ocean dipole events on sub-regional Indian summer monsoon rainfall. Nat Hazards 42:273–285

Krishnamurthy L, Krishnamurthy V (2014) Influence of PDO on South Asian summer monsoon and monsoon--ENSO relation. Clim Dyn 42:2397–2410

Kusiak A, Verma AP, Roz E (2013) Modeling and prediction of rainfall using radar refectivity data: a data-mining approach. IEEE Trans Geosci Remote Sens 51:2337–2342

Lim E, Hendon H, Hudson D, Wang G, Alves O (2009) Dynamical forecast of inter-El Ni˜no variations of tropical SST and Australian spring rainfall. Mon Weather Rev 137(11):3796–3810

Lindsay RW, Percival DB, Rothrock DA (1996) The discrete wavelet transforms and the scale analysis of the surface properties of sea ice. IEEE Trans Geosci Remote Sens 34:771–787

Liyew CM, Melese HA (2021) Machine learning techniques to predict daily rainfall amount. J Big Data 8:1–11

Mallat S (1989) A theory for multiresolution signal decomposition: the wavelet representation IEEE Trans Pattern Anal Mach Intell 11:674–693

Makridakis S, Anderson A, Carbone R, Fildes R, Hibdon M, Lewandowski NJ, Parzen E, Winkler R (1982) The accuracy of extrapolation (time series) methods: results of a forecasting competition. J Forecast 1:111-153

Mart´ınez B, Gilabert M (2009) Vegetation dynamics from NDVI time series analysis using the wavelet transform. Remote Sens Environ 113:1823–1842

Meinke H, Stone R (2005) Seasonal and interannual climate forecasting: the new tool for increasing preparedness to climate variability and change in agricultural planning and operations. Clim Change 70:221–253

Mutai C, Ward M, Colman A (1998) Towards the prediction to the East Africa short rains based on the Sea Surface Temperature – Atmospheric Coupling. Int J Climatol 18:975–997

Paul RK, Birthal PS (2016) Investigating rainfall trend over India using the wavelet technique J Water Clim Chang 7(2):353-364

Paul RK, Garai S (2021) Performance comparison of wavelet-based machine learning technique for forecasting agricultural commodity prices. Soft Comput 25:12857–12873

Paul RK, Garai S (2022) Wavelets Based Artificial Neural Network Technique for Forecasting Agricultural Prices. J Ind S Prob Stat 23:47-61

Percival DB, Walden AT (2000) Wavelet methods for time series analysis Cambridge University Press. Cambridge, UK

Prasad K, Dash SK, Mohanty UC (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

Primer A, Burrus CS, Gopinath RA (1998) Introduction to wavelets and wavelet transforms. In: Proceedings of International Conference

Quilty J, Adamowski J (2021) A maximal overlap discrete wavelet packet transform integrated approach for rainfall forecasting–A case study in the Awash River Basin (Ethiopia). Environ Model Softw 144:105119

Sahai AK, Soman MK, Satyan V (2000) All India summer monsoon rainfall prediction using an artificial neural network. Clim Dyn 16:291–302

Sandberg A, Sands T (2022) Autonomous Trajectory Generation Algorithms for Spacecraft Slew Maneuvers. Aerospace 9:1–22

Schepen A, Wang Q, Robertson D (2012) Combining the strengths of statistical and dynamical modeling approaches for forecasting Australian seasonal rainfall. J Geophys Res 117:D20107

Schneider JM, Garbrecht JD (2003) A measure of the usefulness of seasonal precipitation forecasts for agricultural applications. Trans ASAE 46(2):257–267

Shah R, Sands T (2021) Comparing methods of DC motor control for UUVs. Appl Sci 11:1–16

Tessier Y, Lovejoy S, Hubert P, Schertzer D, Pecknold S (1996) Multifractal analysis and modelling of rainfall and river flows and scaling causal transfer functions. J Geophys Res 101(D21):26427–26440

Ummenhofer C, England M, McIntosh P, Meyers G, Pook M, Risbey J, Gupta A, Taschetto A (2009) What causes southeast Australia's worst droughts? Geophys Res Lett 36:L0

Vijayan R, Mareeswari V, Kumar PM, et al (2020) Estimating Rainfall prediction using machine learning techniques on a dataset. Int J Sci Technol Res 1:440–445