Turkish Journal of Computer and Mathematics Education Vol.12 No.13 (2021), 2014-2020

Research Article

A stepwise Principal Component Regression Model to predict Seasonal Rainfall over Idukki district of Kerala

Suvarna J[a] and Archana Nair[b]

[a,b] Department of Mathematics, Amrita School of Arts and Sciences, Kochi,Amrita Vishwa Vidyapeetham

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 4 June 2021

Abstract: The study comprises of developing a mathematical model based on principal component regression and step-wise multiple linear regression over Idukki district of Kerala. Different parameters such as sea surface temperature, ocean heat content and wind are utilized for the development of model. 63 different models are constructed for this purpose. The observed rainfall data for the district has been collected from the Kerala Government. Results suggest that the rainfall in the region has become asymmetric with more rainfall in theAdarsh-Suvarna J and Archana Nair month of September. The efficiency of model is judged by root mean square error (RMSE) and it has been found that the model by taking only OHC in the region (0-50N,500E-750E) as parameter gives, the least RMSE.

Keywords: Rainfall, prediction, principal compnent regression, idukki, kerala

1. Introduction

Rainfall is a key physical process that transports water from the atmosphere back to Earth's surface and links weather, climate, and hydrological cycle (Kharol et al. 2013). In India, main monsoon starts from June to September. Two major monsoons in India are south- west monsoon, which arrives in June and North-east monsoon arrives in the month October. Some regions in India receive more rainfall during monsoon whereas some regions receive very less rainfall. Mumbai receives a lot of rain during south –west monsoon and Delhi, Hyderabad, Bangalore receives very less.

Kerala lies more closer to the equator when compared to other states in India. Kerala is blessed with a pleasant climate throughout the year. Kerala receives highest rainfall in the year 2018 and 2019 (Ashrit et al. 2020). People depend more on rainfall for agriculture purposes. The annual rainfall of the state varies from 3,800 mm over the north to 1,800 mm in the extreme south. The potential rainy season for Kerala is the southwest monsoon period, which contributes more than 80% of the annual rainfall (Nathan 2000). The state of Kerala experienced the worst disaster in its history in 2018. The disaster affected around 5.4 million people and 433 lives were lost (Nathan. 2000).

Kerala's one of the most nature rich area is Idukki and it is known as spice garden. Idukki lies in the western ghats (see Figure 1) and places are full of greenery, waterfalls and varied vegetation. Place receives rainfall in the beginning of June, due to north –east monsoon it also receives rain in October, November, December. Agriculture is the main occupation of the people. It is suitable for the cultivation of plantation crops like tea, coffee, rubber etc. Rain is very necessary for agriculture. Higher or Lower rainfall, or changes in its spatial and seasonal distribution would influence the spatial and seasonal distribution would influence the spatial and temporal distribution of runoff, soil moisture and ground water reserves, and would affect the frequency of droughts and floods (Kharol et al. 2013). People in this place suffers from many problems due to heavy rain. In the state of Kerala, Idukki was the worst-hit district during 2018 disaster, with 143 major landslides in the state government record (Nathan , 2000) .The southern Indian state of Kerala experienced exceptionally high rainfall during August 2018, which led to devastating floods in many parts of the state (Ashrit et al. 2020). So rainfall prediction is very necessary, since such predictions are very rare in Kerala. According to the Intergovernmental Panel on climate change (IPCC 2007), future climate change is likely to affect agriculture, increase in the risk of hunger and water scarcity (Kharol et al. 2013).

Three Artificial Intelligence approaches like K-nearest neighbor(KNN), Artificial neural network(ANN), Extreme learning machine(ELM),are used for predicting seasonal monsoon and post monsoon rainfall from 2011 to 2016. Study also shows that these approaches will predict with low prediction error (Dash et al. 2018) . Researchers developed Climate Predictiability Tool using global SST for predicting seasonal rainfall in rainy season from 1975 to 2008 and also tried to evaluate starting month of rainy season compare to one month before rainy season (Hossain et.al. 2019). A multilayered feedforward neural networks trained with error-back-propogation algorithm is used to predict seasonal rainfall over India (C. Venkatesan et.al. 1997). India faced a major drought in 2009, a study explains features of this drought examines real-time seasonal prediction has made by six GCM(General Circulation Model)(Acharya et al. 2011) .To predict rainfall in Bangladesh , Multiple linear regression model is used (Navid and Niloy. 2018). A logistic approach has developed , based on DEMETER to

predict rainfall in metrological subdivisions of India (Prasad et al. 2010). Mann-Kendall (MK) test and Sen's innovative were performed to analyse the rainfall trend (Praveen et al. 2020). An artificial Neural Network (ANN) is used for rainfall prediction and it also reports ANN technique is more good than traditional statistical and numerical methods (Nayak et al. 2013).

Many attempts were carried out to predict rainfall, but those techniques were not perfect. Principal Component Regression (PCR) is an approach to predict rainfall. It involves a mathematical procedure that transforms a number of possibly correlated variables into a smaller number of uncorrelated variables called principal components (Mohanty et al. 2019). The number of principal components will be less than the number of features. PCA actually reduce the number of features. PCA is also called data reduction method. It also reduces the problem of overfitting. The manuscript is arranged in the following way so as to explain the methodology in Section2, Results and discussion in Section 3 and Conclusions in Section 4.

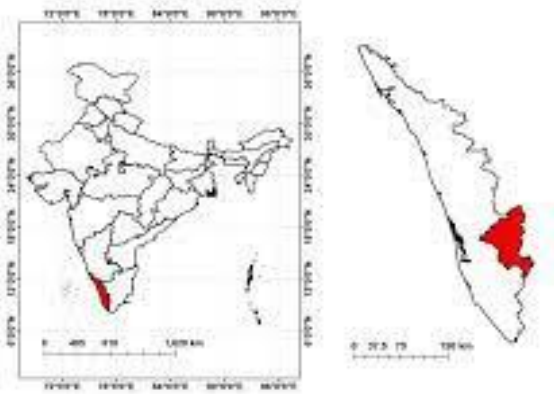

Figure 1: The area of study

2. Data and Methodology

2.1.Sea Surface Temperature

About 71% of earth surface is covered by ocean. Scientists record sea surface temperature (SST) to know how the ocean communicates with earth's surface. This is an essential parameter for weather prediction and for the study of marine ecosystems.Ocean temperature influence rainfall, so when ocean. This component has a major influence in exchanges of energy between ocean and earth atmosphere (Huang et al. 2014, 2015 Liu et al. 2014). Atmospheric water vapour over the ocean increases due to increase in sea surface temperature which causes heavy rain and snow. Here we have SST1 is sea surface temperature over 0-5N 50E-75E, SST2 sea surface temperature over 5-10N 50E-75E.

2.2.Ocean heat content

Ocean plays a major role in Earth's climate system. It has the ability to store and release heat over long periods of time. Rising of greenhouse gases have caused ocean to absorb excess heat and warm up (Levitus et al. 2017). Heat absorbed by ocean leads to increase in ocean temperature and rise in sea level. A assessment report of IPCC in 2013 shows that more than 93% of greenhouse gases is absorbed by ocean . Here we have OHC1 is ocean heat content over 0-5N 50E-75E,OHC2 is ocean heat content over 5-10N 50E-75E.

2.3.Wind

Wind is the one of the significant factor which influences the weather.Wind transports moisture and temperature from one place to other. So there will be a change in weather. Wind blows from sea causes rain in the coast. Here we have WND1 is wind magnitude over 0-5N 50E-75E and WND2 is wind magnitude over 5-10N 50E-75E.

2.4.Rainfall

Rainfall in India varies from one region to the other. Average rainfall of India is about 118cm .Kerala lies in the southern most tip of sub-continent receives more rain than other states in the country. Kerala received heavy rainfall in the year 2018 and almost all resovoirs were full upto its limit. But after this flood, rivers started drying up. Kerala witnessed heavy drought in the year 2016, which affected agriculture and hydrology. Frequency of drought year is increasing . Excessive rainfall can also cause landslide. Rainfall is essential for agriculture purposes. People depend on rainfall for agricultural activities.

3.Methodology

In this paper we use the method of multiple linear regression using principal components (PCR). Rainfall data from 1991 to 2014 has been used for the prediction of rainfall in Idukki.

3.1.Correlation

Correlation shows the relationship between two variables. The value ranges from -1 to +1.Positive value indicates that when one increases other also increases, wheras negative shows that when one increases other decreases.

Covariance matrix is a matrix which contains covariances of pair of variables.

3.2.Regression

Regression is a statistical method which helps to find the relationship between one dependent variable and a series of independent variable. There are two types of regression simple loinear regression and multiple linear regression.

Linear regression is commonly used predictive analysis method. This model uses straight line to predict the relationship between one dependent and one independent variable.

Simple linear regression is represented as:

$Y = a_0 + a_1 X_1 + e_i$

This equation represents how Y is related to X.

Multiple linear regression is estimating the relationship between one dependent variable and two or more independent variable. A Multiple linear regression analysis with dependent variable Y and independent variables $X_1, X_2, X_3, \ldots\ldots..., X_p$ is

$Y = a_0 + a_1X_1 + a_2X_2 + a_3X_3 + a_4X_4 + a_5X_5 + a_6X_6 + e_i \ldots\ldots\ldots\ldots$ (1)

where *Y* is the predictand a_0 , a_1, a_2, a_3, a_4, a_5, a_6 are the predictor coefficients, X are the predictors (predictor variables such as sea surface temperature , ocean heat, wind) and ei is the error.

Equation (1) can also be written as:

$$Y = \begin{pmatrix} 1 & X_{11} & X_{12} \ldots. & X_{1p} \\ \vdots & & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} \ldots\ldots & X_{np} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_p \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_p \end{pmatrix}$$

Using equation (1) model will predict rainfall by using predictor variables such as SST1 is sea surface temperature over 0-5N 50E-75E, SST2 sea surface temperature over 5-10N 50E-75E, OHC1 is ocean heat content over 0-5N 50E-75E, OHC2 is ocean heat content over 5-10N 50E-75E, WND1 is wind magnitude over 0-5N 50E-75E and WND2 is wind magnitude over 5-10N 50E-75E.

Pincipal component regression is dimensionality reduction technique which help to identify correlation in data. PCA highlights the differences and similarities of data. When the number of predictor variables is high PCA can be used. It used for reducing dimension of datasets thereby making the data preserved. Data of high dimension is complex for processing, since it consume more time and data processing is more difficult. Principal components are the output of PCA , whose number will be less than or equal to number of original variables. PCs are orthogonal in nature.

PCA is a statistical procedure that convert a set of correlated variables into set of uncorrelated variables and principal components are independent to one another. But in linear regression we are finding a straight line that fits the data.

PCA are used in facial recognition, image compression and computer vision, data mining, psychology. Principal components are used to analyse near- infrared spectra (Ian et al.1988). PC is also used in forecasting a single time series when there is a large number of predictors (Stock & Watson. 2002).

Steps in PCA

1.First standardize the data. Standarization of data means scaling data in such a way that all variables and their values will lie in range.

Z= (value-mean)/ Standard deviation.

2. Next step is to find covariance matrix. Covariance matrix will help to find highly correlated variables. If the covariance is positive, it shows that two variables increases or decreases together. If it is negative,it shows that one increases and other decreases.

3. Computing the eigenvectors and eigenvalues of the covariance matrix.

In general if A is a square matrix , v is a vector and λ a scalar then λ is called eigenvalue if it satisfies $Av= \lambda v$. Eigen values are roots of the equation $\det(A- \lambda I) = 0$.

4. Computing the eigenvectors and eigenvalues of the covariance matrix is to identify the Principal Components. PCs are new variables formed as a linear combination of initial variables. New variables are uncorrelated ones. These components provide a new angle to see and evaluate data. The k th PC is the eigen vector of matrix corresponding to k th largest eigenvalue.

5.The last step is to rearrange the original data. Principal components with most significant information is given priority.

4. Results and Discussions

In the upcoming sections, different characteristics of rainfall over Idukki district is reported before the development of mathematical model.

4.1. Seasonal and monthly rainfall over Idukki

The seasonal rainfall over Idukki district of Kerala is depicted in Figure 2. Figure 2a shows the inter-annual variation of rainfall from 1991 to 2014 in the summer monsoon season. The rainfall in the season ranges from 474 mm to 937 mm. The mean rainfall in the region is found to be 680mm and the standard deviation is 123 mm. The 75^{th} percentile is recorded at 805mm. The trend analysis suggests a decrease of 2.06mm/year which is not significant. Figure 2b represents the standardised anomaly of rainfall. The blue bars show the year which are above normal and red bars below normal. The black bars show normal monsoon years. Five years are found to be excess (1992,1998,2005,2007,2013,2014) while four instances of below normal rainfall are there (1999,2002,2003,2012). Rest 15 years were normal rainfall years. The highest rainfall year is 2013 which is two standard deviations more than the mean rainfall. The minimum rainfall is recorded in the year 2012 where 1.6 standard deviation less rainfall has happened. From the discussion, it is clear that the frequency of wet years is more than the drought years.

The monthly rainfall over Idukki district in the summer monsoon season is shown in Figure 3a-d for June, July, August and September respectively. The mean rainfall in the month of June 715.4 mm and the standard deviation is 245.14mm and hence the coefficient of variation is 34%. The maximum rainfall that happened in June is 1202 mm and the minimum is 381 mm. Six instances of above normal (1991, 1992, 1994,2000, 2001 and 2013) and four instances of below normal rainfall years are recorded (1997,2003,2009 and 2012). The analysis suggests that more droughts have happened after 2000. The increase of drought years leads to a decrease of 9.8mm/year of rainfall per year.

In comparison to June, July rainfall is found abundant. The mean rainfall is 922mm and the standard deviation is 213mm. There are four above normal (1997,2005,2007 and 2013) and below normal years (2000,2004,2008 and 2012). An interesting point to ponder here is the periodicity of below normal years which happens in every four years. A noticeable decrease of 4mm/year is recorded for the July rainfall.

In the month of August, 695 mm of rainfall occurs with standard deviation of 162 mm. The rainfall in the month of August, is much less than the June and July rainfall. Three instances of above normal (1998,2000 and 2014) and below normal years (1999, 2006 and 2009) are noticed. A cyclic nature is observed for the August rainfall where two epochs of above normal and below normal rainfall are seen. For example, from 1991 to 1997 above normal phase is noticed and from 2001 to 2013 below normal epoch is observed. A small increase of 0.042mm/day is found. In the month of September, a mean of 388mm and a standard deviation of 197 mm is recorded. A maximum rainfall of 829.7mm and a minimum of 106mm is found. It is noticed that the coefficient of variation in the month of September is the highest among the other months (50%). It shows that the rainfall variability is high in this month. Three instances of above normal (1998, 2005, 2007) and four instances of below normal (1999,2001, 2002 and 2003) is recorded. No cyclic nature is found in the September month. The trend analysis of approximately 6mm/day is noted here. This implies that the rainfall in the Idukki district is becoming asymmetric with more increasing rainfall in the September rainfall.

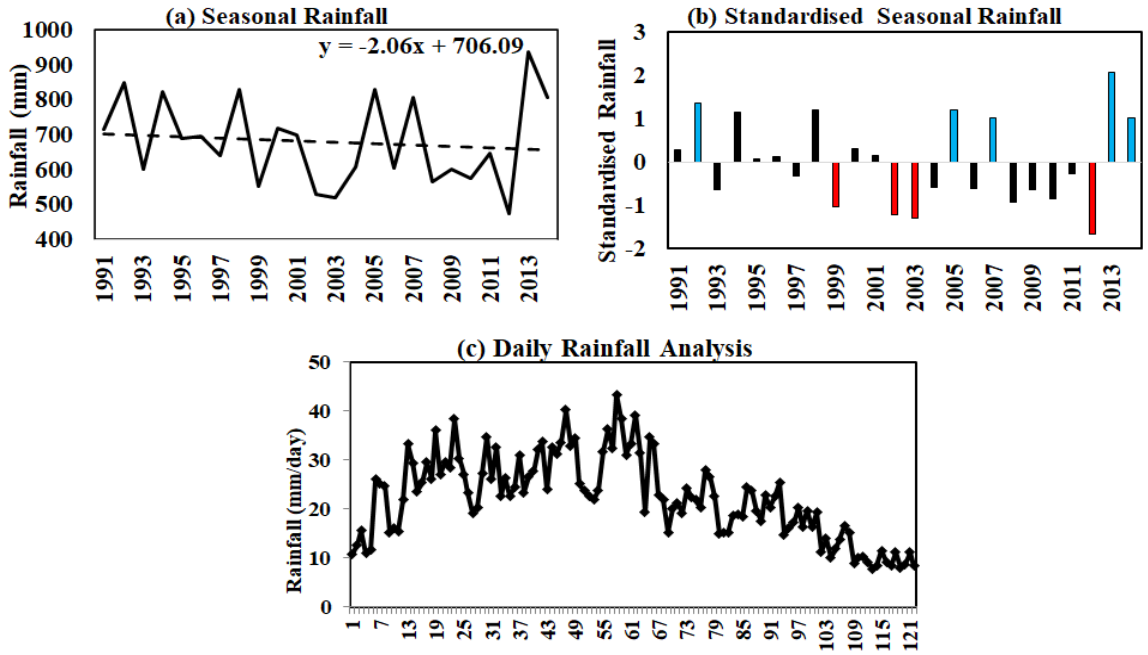

Figure 2 (a) The seasonal variation of rainfall (b) The standardised seasonal rainfall (c) The daily rainfall in the summer monsoon season over Idukki district of Kerala.

4.2. Daily rainfall scenario over India

The daily rainfall over Idukki district is presented in the Figure 2c for 122 days. It shows that the daily rainfall follows gamma distribution. The daily rainfall shows the decrease of 0.12 mm/day is noticed. The rainfall ranges from 10mm/day to 45 mm/day and the standard deviation is 8.47 mm/day. The maximum rainfall that occurred in a day is 43.3 mm/day and the minimum is 7.6 mm/day.

4.3.Mathematical model to predict rainfall

By considering different parameters at a time, 63 different models were constructed on a leave-one-out cross-validated mode. Figure 4 shows the RMSE of different models. By selecting one parameter i.e., SST1, SST2, OHC1, OHC2, WND1, WND2, the least RMSE is found for the ocean heat content (0.92mm/year) whereas the highest is found as 1.10mm/year. In the two-parameter model, 15 models are developed using leave one out cross-validation. The range of RMSE is between 0.94 to 1.15mm/year. The RMSE is found the least for the pair OHC1 and WND2 i.e OHC in the 0-5, 50-75E and wind at 5-10, 50-75E.

In the three-parameter model, 20 models are constructed and the least RMSE is found for the combination of OHC1, OHC2 and WND2 with 0.96 mm/day of RMSE. In rest of the models, the RMSE is greater than 1mm/day. For the four-parameter model, the least RMSE is found as 1.02mm/day with SST1, OHC1, OHC2 and WND2. The highest RMSE is 1.26mm/day. In the five-parameter model, the least RMSE is found to be 1.09 mm/day and the highest is 1.11 mm/day. In the six-parameter model, an RMSE of 1.17mm/day is found.

From the above discussion, it is clear that the OHC1 region gives the best prediction of rainfall with very less RMSE.

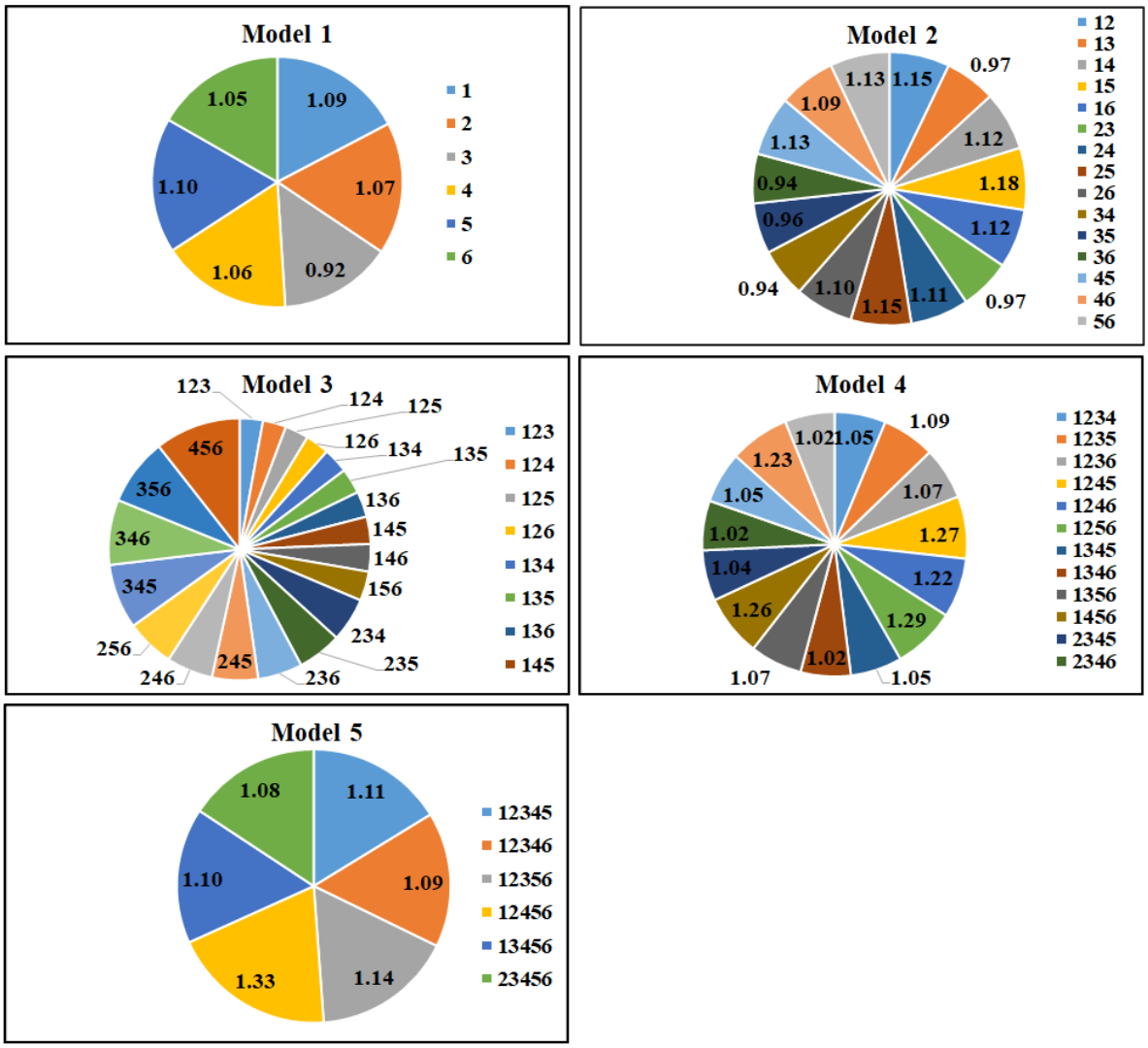

Figure 4 The RMSE of the five mathematical model described as a pie chart (mm/year).[1 stands for SST1, 2 stands for SST2,3 stands for OHC1, 4 stands for OHC2, 5 stands for WND1 and 6 stands for WND2]

5. Conclusions

Nowadays, the climate has been very erratic, be it rainfall or temperature, from torrential rains to extreme heat, everywhere and every time, the chaotic nature can be visualised. In this study, a mathematical model based on principal component regression has been employed. It has been found that the rainfall over the Idukki district has become vey anomalous with more rainfall values shifting to the month of September where the rainfall is found to be increasing at the rate of 6mm/year. In all the other months, the rainfall is said to be decreasing with the highest decrease in the month of June with 9.8mm.day of rainfall. The seasonal rainfall also shows a decrease of 2.06mm/day of rainfall. Results also suggest a greater number of wet years than dry years. 63 different models are developed using different parameters such as SST1, SST2, OHC1, OHC2, WND1 and WND2. The efficacy of model is judged by evaluating root mean square error (RMSE). It is found that the model by taking only OHC in the region (0-5^{0}N,50^{0}E-75^{0}E) as parameter gives the least RMSE. So, in view of above, it can be inferred that OHC is the important factor in controlling the rainfall values of the region

References

1. Acharya, Nachiketa & Kar, Sarat & Mohanty, U C & Kulkarni, Makarand & Dash, Sushil. 2011: Performance of GCMs for seasonal prediction over India—A case study for 2009 monsoon. Theor. Appl. Climato, volume 105 ,Issue 3.
2. Ashrit R, Sharma K, Kumar S, et al, 2020 : Prediction of the August 2018heavy rainfall events over Kerala with high resolution NWP models. Meteorol Appl. 2020;27:e1906. https://doi.org/10.1002/met.190614 of 14 ASHRIT ET AL.

3. C. Venkatesan, S. D. Raskar, S. S. Tambe, B. D. Kulkarni & R. N. Keshavamurty 1997: Prediction of all India summer monsoon rainfall using error-back-propagation neural networks, ,Meteorology and Atmospheric Physics volume 62.
4. Dash, Yajnaseni & Mishra, Saroj & Panigrahi, Bijaya 2018: Rainfall prediction for the Kerala state of India using artificial intelligence approaches. Computers & Electrical Engineering. 70. 66-73. 10.1016/j.compeleceng.2018.06.004, volume 70
5. Hossain, Md & Abul, Md & Karmakar, Samarendra & Nazrul, Islam & Mondal, & Das, Mohan & Rahman, Md & Haque, Md 2019 : Assessment of Better Prediction of Seasonal Rainfall by Climate Predictability Tool Using Global Sea Surface Temperature in Bangladesh. Asian Journal of Advanced Research and Reports, volume 4 , Issue 4.
6. Huang, B., P. Thorne, T. Smith, W. Liu, J. Lawrimore, V. Banzon, H. Zhang, T. Peterson, and M. Menne, 2015: Further Exploring and Quantifying Uncertainties for Extended Reconstructed Sea Surface Temperature (ERSST) Version 4 (v4). Journal of Climate, 29, 3119–3142, doi:10.1175/JCLI-D-15-0430.1 (link is external).
7. Huang, B., V.F. Banzon, E. Freeman, J. Lawrimore, W. Liu, T.C. Peterson, T.M. Smith, P.W. Thorne, S.D. Woodruff, and H.-M. Zhang, 2014: Extended Reconstructed Sea Surface Temperature version 4 (ERSST.v4): Part I. Upgrades and intercomparisons. Journal of Climate, 28, 911–930, doi:10.1175/JCLI-D-14-00006.1
8. Ian A. Cowe, Sigrid Koester, Christian Paul, James W. McNicol, D.Clifford Cuthbertson, 1988: Principal component analysis of near infrared spectra of whole and ground oilseed rape (Brassica napus L.) samples, Chemometrics and Intelligent Laboratory Systems, Volume 3, Issue 3.
9. Liu, W., B. Huang, P.W. Thorne, V.F. Banzon, H.-M. Zhang, E. Freeman, J. Lawrimore, T.C. Peterson, T.M. Smith, and S.D. Woodruff, 2014: Extended Reconstructed Sea Surface Temperature version 4 (ERSST.v4): Part II. Parametric and structural uncertainty estimations. Journal of Climate, 28, 931–951, doi:10.1175/JCLI-D-14-00007.1
10. J.H, Stock & M.W, Watson, 2002: Forecasting Using Principal Components from a Large Number of Predictors. Journal of the American Statistical Association, Volume 97, Issue 460.
11. Nathan, 2000: Drought Network News (1994-2001), Characteristics of Drought in Kerala, India, Water Technology Centre, Indian Agricultural Research Institute, New Delhi 110 012, India, Drought Network News Vol. 12, No. 1, Published by the International Drought Information Center and the National Drought Mitigation Center, School of Natural Resources, University of Nebraska – Lincoln.
12. K. Prasad, S. K. Dash, U. C. Mohanty,2010: A logistic regression approach for monthly rainfall forecasts in meteorological subdivisions of India based on DEMETER retrospective forecasts Volume 30, Issue 10.
13. Kharol, Shailesh & Kaskaoutis, Dimitris & Sharma, Anu & Singh, 2013: Long-Term (1951–2007) Rainfall Trends around Six Indian Cities Current State,Meteorological, and Urban Dynamics. Advances in Meteorology, Volume 2013, https://doi.org/10.1155/2013/572954
14. MAI Navid, NH Niloy, 2018: Multiple Linear Regressions for Predicting Rainfall for Bangladesh, Communications. Vol. 6, No.1
15. Mohanty, U C & Sinha, Palash & Mohanty, Manas & Maurya, R. & Malasala, Murali & Pattanaik, D 2019: A review on the monthly and seasonal forecast of the Indian summer monsoon. Mausam, Volume 2.
16. Nayak, Deepak & Mahapatra, Amitav & Mishra, Pranati, 2013: A Survey on Rainfall Prediction using Artificial Neural Network. International Journal of Computer Applications, Volume 72 - No 16.
17. Praveen, Bushra & Talukdar, Swapan & Shahfahad & Mahato, Susanta & Mondal, Jayanta & Sharma, Pritee & Islam, Abu Reza Md & Rahman, Atiqur 2020: Analyzing trend and forecasting of rainfall changes in India using non-parametrical and machine learning approaches. Scientific Reports, doi: 10.1038/s41598-020-67228-7.