ºURSI - RCRS 2022, IIT (Indore), India, 1 - 4 December, 2022

A Discussion on Machine Learning Approach of Rainfall Prediction

Ranjana Ray (1), Swastika Chakraborty* (2)

(1) JIS College of Engineering, Kalyani, West Bengal, India
(2) Narula Institute of Technology, Kolkata, West Bengal, India

Abstract

A comprehensive correlation analysis of extreme rainfall (>204.5 mm/day) and very heavy rainfall (>115.6 mm/day) has been done over Indian subcontinent considering 40 years (1981-2021) of data of meteorological parameters having a causal relation with rainfall. Seasonal analysis has been done after locating the place having maximum rainfall over the forty years for both the two categories of extreme rainfall and very heavy rainfall. For each of the categories best correlated parameters are taken for prediction of rainfall through linear regression, a machine learning based approach. The performance analysis of the prediction has been done. In the next step another machine learning approach, multivariate linear regression analysis has been done connecting rainfall with the associated parameters. Performance analysis of multivariate regression also has been done for rainfall prediction. Finally, comparison of performances for single variable analysis and multivariate analysis has been done to come to a conclusion about the prediction of rainfall (Precip) over Indian landmass and it has been found that best correlation performance has been obtained at the month March-April-May at Acharkund, Madhya Pradesh (22°30'00.0"N 79°00'00.0"E) taking the wind speed(U_2), relative humidity(RH), specific humidity(SH), surface temperature(T), dew point temperature(T_{dew}) and surface pressure(P) as input variable for prediction.

.

1. Introduction

Rainfall extreme situation has a severe impact over the societal life for a densely populated country like India. Rainfall caused land slide at hill stations, over flooded condition at plane land and a threat of thunderstorm and lightning has a severe societal impact over Indian landmass. Spatial analysis of extreme rainfall over Indian landmass has been analyzed thoroughly [1] and quantified using long term data and multimodal ensemble method. An attempt has been made to find out the time delay in the occurrence of rainfall and causes of rainfall like convective growth, specific humidity, and relative humidity [2] for a specific location using the input from satellite observation, reanalysis data and forecast model. Besides traditional approaches of numerical weather prediction methods with its consequent limitations, there is increasing popularity of machine learning methods in the prediction of extreme rainfall [3]. Quantification of daily rainfall data and exploring most relevant causes of rainfall by Pearson correlation analysis has been found successful in literature when prediction of rainfall has been done with the conclusion from correlation analysis [4] successfully. In addition to machine learning, deep learning [5] has gained popularity also for the prediction of the extreme event of rainfall at an early stage and necessary measure may be made possible to safeguard societal life. Besides all the discussions and the proven fact for some cases where machine learning is found to provide a good result in prediction of rainfall, particularly deep learning faces challenge [6] also without the availability of high-power computer (HPC). This is the first time we are reporting the detailed analysis of correlation coefficients and performance of machine learning based prediction models for the prediction of extremely heavy and very heavy rainfall considering long term data over Indian locations which will help in nowcasting as compared to medium range forecasting of rainfall over Indian location. Moreover, for the complex atmospheric phenomenon like rainfall if machine learning generated result can be combined with forecast model predicted result may be improved prediction can be obtained in future.

2. Methodology

2.1. Data Sources

One of the data sources used for this work is the repository of Indian Meteorological Department for rainfall data over Indian sub-continent region. The rainfall data of a different flavor has been collected from Climate Hazards Group InfraRed Precipitation with Station data (CHIRPS) which produces data combining data from real-time observing meteorological stations with infra-red data. Another data sources used is the reanalysis data sets from numerical weather prediction models of gridded ERA-5 dataset. Native Resolution Daily Data is used for rainfall for this work (https://power.larc.nasa.gov/data-access-viewer/).

2.2. Data Pre-processing

After the collection of data of forty years initial processing has been done to remove the fill value of the data from the

dataset used and sufficient care has been taken to address the problem of missing data. Use of Python software has been done for the said purposes as and when required. Data categorization has been done using proper filtering process to identify extreme rainfall and very heavy rainfall locations over the Indian peninsula.

2.3. Methodology

Rainfall is one of the most complex atmospheric processes which needs to be quantified with the help of the some other atmospheric parameters which have a direct or indirect relation with rainfall. Here the prediction of rainfall is explored with an attempt of exploring the variation of associated parameters with the rainfall. For this work in the initial stage Pearson correlation has been attempted to find the comparatively strong relation of rainfall with other climatic parameter and in the later stage a multivariate regression process is attempted considering rainfall as dependent variable and all other parameters discussed above as independent variable.

2.3.1 Pearson Correlation Analysis

Here Pearson Correlation is chosen to find the correlation between the variables as it is a most common statistical method of finding a correlation between the two variables.

2.3.2. Prediction Methods

Simple linear regression (SLR) has been attempted as the prediction method for the prediction of rainfall with a knowledge of variation of individual dependent variable, considering 70% of the total available data for the training purpose and 30% of the However Simple linear regression method for this case cannot provide an optimistic result.

In the next step multivariate linear regression (MLR) has been attempted with wind speed, relative humidity, specific humidity, pressure, surface temperature and dew point Temperature as the input variable to predict and estimate extreme rainfall and very heavy rainfall over the locations having maximum of extreme rainfall considering forty years statistics of rainfall and maximum of very heavy rainfall considering forty years statistics of rainfall over that particular location.

R-squared value, Adj. R-squared value, Mean Absolute Error (MAE), Root Mean Square Error (RMSE) as performance parameter has been calculated.

3 Results

Fig 1(a), 1(b) and 1(c) shows a spatial variation of twenty five years of rainfall time series from the experimental data source of Indian Meteorological Department(https://imdpune.gov.in/), from rain gauge and satellite observations CHIRP(https://data.chc.ucsb.edu/products/CHIRPS-2.0/global_daily/netcdf/p25/) and from reanalysis data (https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels?tab=overview) respectively. From the figures it is observed that rainfall extreme is found at Acharkund, Madhya Pradesh (22°50'00.0"N 79°00'00.0"E) on 6th October 2001 and the amount of rainfall is 822.06mm on that particular day. From the figures it is observed that very heavy rainfall is found at Ratnagiri, Maharashtra (17°50'00.0"N 73°50'00.0"E) on 24th June 2007 and the amount of rainfall is 204.4 mm on that particular day.

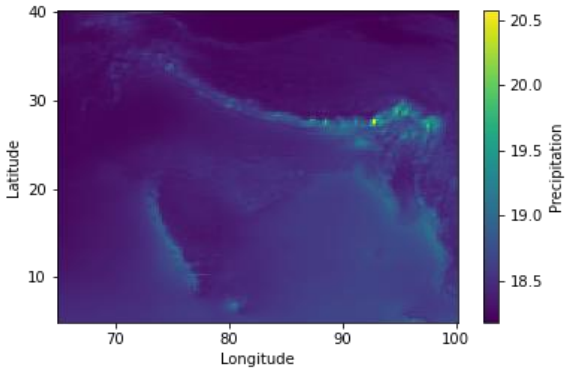

Figure 1(a). Spatio-temporal Variation of 25years of rainfall data from ERA5 data source.

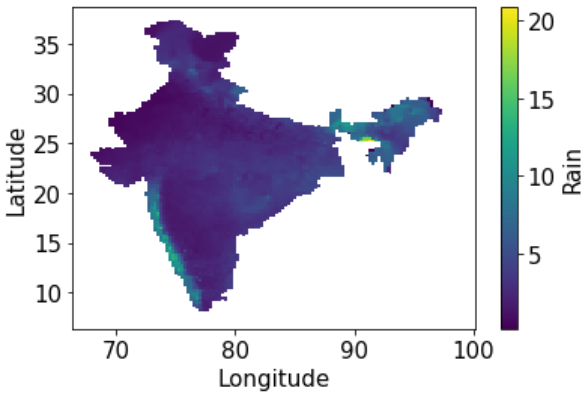

Figure 1(b). Spatio-temporal Variation of 25years of rainfall data from IMD data source.

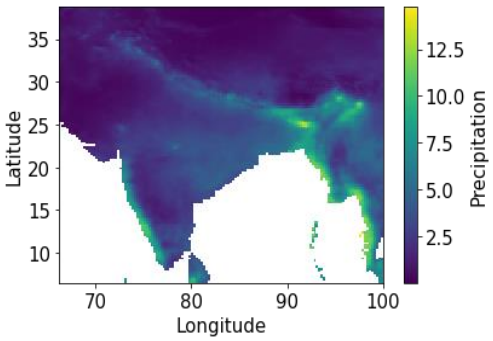

Figure 1(c). Spatio-temporal Variation of 25years of rainfall data from CHIRPS data source

Fig. 2(a-f) shows time series plots of wind speed, relative humidity, specific humidity, surface temperature, dew point temperature and surface pressure extreme rainfall location and very heavy rainfall location considering complete years from 1981-2021.

3.1 Correlation Analysis

Pearson correlation between some atmospheric parameters wind speed, relative humidity, specific humidity, pressure, surface temperature and dew point temperature and precipitation is listed in Table I for extreme rainfall and very heavy rainfall respectively considering forty years of rainfall data. Seasonal analysis of rainfall has been done after clustering of month of the years in four categories like pre-monsoon (March-April-May), monsoon (June, July, August and September), post-monsoon (October, November and December) and winter (January and February) for extreme rainfall in the Table II (a) and the same is done for very heavy rainfall in the Table II (b) respectively considering forty years of rainfall data.

Table I. Correlation of atmospheric parameters with rainfall for extreme rainfall and very heavy rainfall over 40years

Atmospheric Parameters	extreme rainfall	very heavy rainfall
T	0.03	-0.08
T_{dew}	0.42	0.42
SH	0.48	0.47
RH	0.45	0.46
Precip	1.00	1.00
P	-0.45	-0.30
U_2	0.34	0.54

Table II (a) Seasonal analysis of correlation between precipitation and various atmospheric parameters for extreme rainfall

	T(°C)	T_{dew}(°C)	SH (g/kg)	RH (%)	P (kPa)	U_2 (m/s)
Pre Monsoon	-0.13	0.62	0.69	0.81	-0.03	0.06
Monsoon	-0.41	0.59	0.64	0.54	-0.42	0.29
Post Monsoon	0.34	0.58	0.64	0.62	-0.48	0.18
Winter	-0.18	0.64	0.7	0.66	0.08	-0.11

Table II (b) Seasonal analysis of correlation between precipitation and various atmospheric parameters for very heavy rainfall

	T(°C)	T_{dew}(°C)	SH (g/kg)	RH (%)	P (kPa)	U_2 (m/s)
Pre Monsoon	-0.05	0.45	0.48	0.54	-0.42	0.28
Monsoon	0.13	0.67	0.69	0.26	-0.59	0.57
Post Monsoon	0.57	0.67	0.71	0.64	-0.68	-0.34
Winter	-0.05	0.16	0.15	0.13	-0.08	0.15

3.2 Performance Analysis

SLR has been done for the individual dependent variable of rainfall considering the forty years of rainfall for the extreme rainfall location and very heavy rainfall location and R-squared value is less than 0.2 for each analysis. Therefore, the model developed model developed is not considered for prediction.
MLR has also been done considering yearly data as well as seasonal data over 40 years as 1981-2021.Performance of MLR model over yearly data is listed in Table III. Performance of MLR model over seasonal data is listed in Table IV (a) and Table IV (b). Prediction of coefficients of all variables is given in equations (1) and (2).

4. Equations

After finding coefficients of all variables using MLR:

a) Predicted best fitted line for seasonal prediction of extreme rainfall location is:

$$Precip = ((1.49 \times \mathrm{SH}) + (1.978 \times \mathrm{RH}) - (2.364 \times \mathrm{T_{dew}}) - (0.107 \times \mathrm{U_{2)}} + (0.73 \times \mathrm{T}) - (0.003 \times \mathrm{P})) \quad (1)$$

b) Predicted best fitted line for seasonal prediction of very heavy rainfall location is:

$$Precip = ((3.59 \times \mathrm{SH}) + (0.844 \times \mathrm{RH}) - (4.03 \times \mathrm{T_{dew}}) - (0.008 \times \mathrm{U_{2)}} - (0.192 \times \mathrm{T}) - (0.0499 \times \mathrm{P})) \quad (2)$$

5. Figures and Tables

5.1 Figure

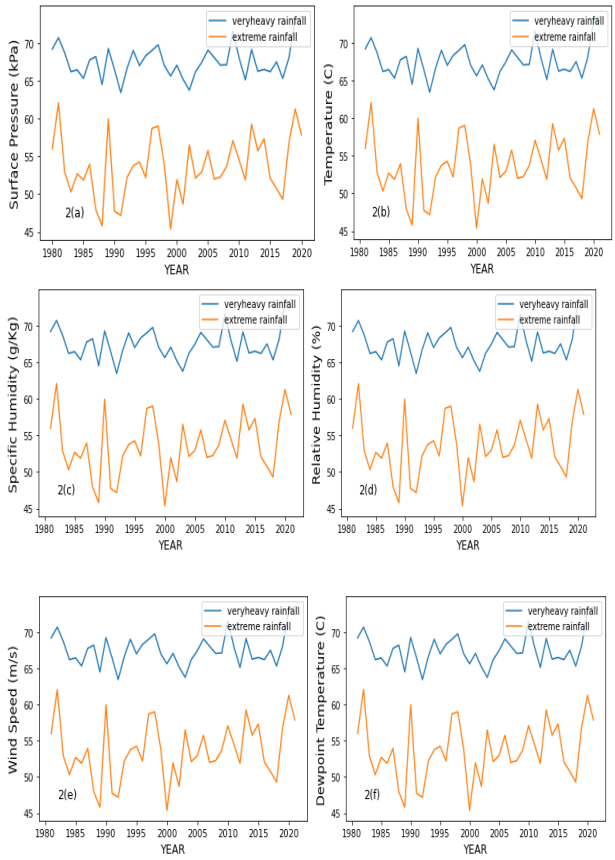

Figure 2(a-f).Time Series Plots of various atmospheric parameters of extreme rainfall location and very heavy rainfall location over 40 vears.

5.2 Table

Table III Performance Analysis of MLR model based on 40years rainfall data

	Extreme Rainfall Location(Latitude 22.5°N, Longitude 79°E)	Very heavy Rainfall Location(Latitude 17.5°N, Longitude 73.5°E)
R-squared:	0.353	0.426
Adj. R-squared:	0.352	0.426
MAE	12.3	15.27
RMSE	8.07	11.05

Table IV (a) Performance Analysis of MLR model for seasonal extreme rainfall prediction over 40years

Extreme Rainfall Location(Latitude 22.5°N, Longitude 79°E)				
Season	Pre Monsoon	Monsoon	Post Monsoon	Winter
R-squared:	0.823	0.668	0.631	0.581
Adj. R-squared:	0.823	0.667	0.631	0.580
MAE	36.8	48.12	42.8	10.66
RMSE	5.56	9.07	8.33	1.44

Table IV (b) Performance Analysis of MLR model for seasonal extreme rainfall prediction over 40 years

Very heavy Rainfall Location(Latitude 17.5°N, Longitude 73.5°E)				
Season	Pre Monsoon	Monsoon	Post Monsoon	Winter
R-squared:	0.434	0.657	0.685	0.145
Adj. R-squared:	0.434	0.656	0.685	0.145
MAE	26.33	36.03	38.4	46.1
RMSE	2.75	2.83	3.14	9.11

6. Acknowledgements

Authors thankfully acknowledge the scientists of India Meteorological Department Pune, Climate Hazards Center InfraRed Precipitation with Station data (CHIRPS).Authors also acknowledge scientists of ECMWF, MERRA2.

7. References

1. U. Saha, M. Sateesh, “Rainfall extremes on the rise: Observations during 1951–2020 and bias-corrected CMIP6 projections for near- and late 21st century over Indian landmass,” Journal of Hydrology, 608, 127682, May 2022, doi: 10.1016/j.jhydrol.2022.127682.

2. U. Saha**,** T. Singh, P. Sharma, M. Das Gupta,V.S.Prasad, “Deciphering the extreme rainfall scenario over Indian landmass using satellite observations, reanalysis and model forecast: Case studies,” Atmospheric Research,August,2020,doi:10.1016/j.atmosres.2020.104943.

3. K. V. Subrahmanyam, C. Rosenthal, A. G. Imran, A. Chakravorty, R. Sreedhar, E Ezhilrajan, D. B. Subrahamanyam, R. Ramachandran, K. K. Kumar, M. Rajasekhar ,C. S. Jha, “Prediction of heavy rainfall days over a peninsular Indian station using the machine learning algorithms,” Journal of Earth System Science,30,November,2021, doi: 10.1007/s12040-021-01725-9.

4. C. M. Liyew, H. A, Melese, “Machine learning techniques to predict daily rainfall amount,” Journal of Big Data, 07, December, 2021, doi: 10.1186/s40537-021-00545-4.

5. S. Gope, S. Sarkar, P. Mitra, S. Ghosh, “Early Prediction of Extreme Rainfall Events: A Deep Learning Approach,” Advances in Data Mining. Applications and Theoretical Aspects, 2016, pp. 154–167, doi: 10.1007/978-3-319-41561-1 12.

6. M. Chantry, H. Christensen, P. Dueben, T. Palmer, “Opportunities and challenges for machine learning in weather and climate modelling: hard, medium and soft AI,” Philosophical Transactions of the Royal Society A, 15, February, 2021, doi:10.1098/rsta.2020.0083.