Assessing the Nonlinear Dynamics of Climate - Rainfall Interactions using Machine Learning Models

Vashisth Ayush*, Sharif Mohammed and Shakeel Mohammad
Department of Civil Engineering, Jamia Millia Islamia (Central University), New Delhi-110025, INDIA
*ayushvashisth90@gmail.com

Abstract

The climate change modelling directly affects the hydrological water management systems. Accurate prediction of the rainfall is crucial for ensuring optimal water quality and supporting aquatic ecosystems. This study evaluates the predictive capability of machine learning models M5P tree and linear regression for estimating the monthly rainfall. Seven key climatic and metrological drivers including ENSO, IOD, NAO, PDO, AMO, T_{max} and T_{min}, were employed as input variables to analyse the global teleconnection patterns.

Among the models tested, the M5P tree model exhibited the best predictive performance, achieving correlation coefficient (CC) values for calibration and validation datasets respectively. In this study, the potential of climate data and advanced machine learning models is explored for accurate monthly rainfall prediction for irrigation planning and, management of water resources in the selected region.

Keywords: Rainfall Prediction, Climatic Indices, Machine Learning, Forecasting, Teleconnections.

Introduction

Prediction of rainfall is a non-linear, complex hydrological process for agriculture, water resource management and disaster preparedness[1]. To analyse, the insights of hybrid and data driven machine learning help to develop reliable and accurate prediction models and allow disaster management authorities to prepare for flood or drought events[3]. Traditional statistical models, particularly linear regression (LR), have long been employed for rainfall prediction because of their interpretability and computational simplicity. However, these methods use for nonlinear and regime-dependent relationships between climatic indices and rainfall. In machine learning approaches, M5P model tree can partition data into homogeneous subsets and fit localized linear models, making them better suited to represent climate interactions. This research study compares the performance of LR and M5P model in predicting monthly rainfall using climate indices, rainfall and temperature as predictors. The research closely aligns with sustainable development goals no. 13 for climate action. This work aims to provide a smart climate forecasting using machine learning approaches for climate services.

Climatic Drivers and Data Sources: Rainfall variability requires examining the influence of large-scale climatic oscillations that interact with atmospheric and oceanic systems worldwide. These oscillations, often referred to as climate teleconnections, describe recurrent and persistent patterns of climate variability and operate over time scales ranging from months to decades. In this study, climatic indices such as Niño 3.4 (ENSO), Indian Ocean Dipole (IOD), North Atlantic Oscillation (NAO), Pacific Decadal Oscillation (PDO) and Atlantic Multidecadal Oscillation (AMO) were selected as predictors to capture the influencing rainfall patterns. The climate indices have a distinct physical basis, spatial domain and temporal variability, yet their combined effect often determines regional precipitation anomalies.

Niño 3.4 (ENSO Indicator): The ENSO phenomenon is varying from 2 to 7 years based on the seasonal and monthly rainfall prediction The El Niño–Southern Oscillation (ENSO) is a phenomenon in the tropics, significantly influencing precipitation, temperature and extreme weather events. The Niño 3.4 index, which represents sea surface temperature anomalies region 5°N–5°S and 170°W–120°W in the central equatorial Pacific, serves as a key indicator of ENSO conditions. Positive anomalies in this index indicate El Niño conditions, typically associated with suppressed monsoon rainfall in South Asia, while negative anomalies (La Niña) often bring enhanced rainfall to the region[9].

ENSO influences rainfall patterns through alterations in the Walker circulation, atmospheric convection and the positioning of tropical rainfall belts. Its effects extend beyond tropical regions, with teleconnections influencing subtropical and mid-latitude areas, altering storm tracks, snowpack formation and seasonal drought patterns[6]. The time scale of ENSO variability ranges from 2 to 7 years, making it a critical input for seasonal and monthly rainfall forecasting models[6,9]. The Niño 3.4 index is used for identifying El Niño and La Niña events. It reflects anomalies in the east-central tropical Pacific region. Monthly Niño 3.4 index data, covering the period from 1870 to 2019, provide a long-term record for analyzing oceanic and climatic variability.

Indian Ocean Dipole (IOD): The Indian Ocean Dipole (IOD) is also known as Indian Nino which represents perturbation in the region of Indian Ocean. This region largely effects the areas like Indian, East Africa etc. The IOD played vital role for effecting the hydrological cycle especially for the case of Indian Monsson depending upon different ENSO phase. In this, there are two existing phases like positive and negative phase warmer and colder phase. The data sets of monthly anomalies of IOD which is

https://doi.org/10.25303/192da070076

spanning from 1984 to 2021 were obtained from National Oceanic and Atmospheric Administration (NOAA) database (https://www.esrl.noaa.gov/psd/gcos_wgsp/Timeseries/Data/dmi.long.data)

North Atlantic Oscillation (NAO): In the North Atlantic Oscillation, the fluctuation in sea level pressure is with positive phase generally in the parts of southern Europe and North Africa, whereas its negative phase tends to reverse these anomalies[4]. But for Northern Hemisphere, NAO-related shifts in atmospheric circulation patterns can influence rainfall even in remote regions. In the context of South Asia, the NAO has been linked to modulations in winter precipitation and certain aspects of the summer monsoon system through cross-hemispheric teleconnections[4,7].

Pacific Decadal Oscillation (PDO): The Pacific Decadal Oscillation (PDO) is a typical phase lasting 20 to 30 years[8]. PDO's influence on rainfall operates through modulations in ENSO teleconnections, atmospheric pressure fields and the position of the Pacific jet stream. For instance, during a positive PDO phase, the impacts of El Niño events on North American and Asian rainfall are often amplified. This makes PDO an important low-frequency background signal that interacts with higher-frequency oscillations like ENSO[8,10].

Atlantic Multidecadal Oscillation (AMO): In this, periodicity of roughly 60–80 years is positive in AMO phases (warmer SSTs). SST anomalies influence rainfall primarily through alterations in the Hadley circulation and changes in tropical cyclone activity. The persistence of AMO phases means that their influence can set decadal-scale backgrounds for regional hydrological variability, thereby impacting agricultural productivity and water resource planning[2,6].

Data Sources and Compilation: Monthly historical datasets of rainfall and five climatic indices were compiled for multiple decades to train and to evaluate the forecasting models. Rainfall data were obtained from national meteorological services and global gridded datasets. Climatic indices were sourced from recognized repositories: Niño 3.4 from NOAA's Climate Prediction Center, IOD from the Australian Bureau of Meteorology, NAO from NCAR's Climate Analysis Section, PDO from JISAO and AMO from NOAA's Earth System Research Laboratory. All datasets were standardized to a common monthly time step, missing values were interpolated linearly and anomalies were calculated relative to a baseline climatology to emphasize variability.

Rationale for Predictor Selection: The selection of these five indices was based on their proven statistical linkages to rainfall variability across multiple regions, their coverage of different ocean basins and their complementary time scales. ENSO and IOD capture interannual variability, NAO primarily influences seasonal-to-interannual fluctuations; PDO and AMO represent low-frequency, decadal to multidecadal signals. Combining these predictors provides a multi-scale representation of global climate drivers, enabling models to better resolve both short-term fluctuations and long-term tendencies in rainfall patterns.

Material and Methods

In this study, two distinct predictive modeling approaches were implemented to estimate monthly rainfall from large-scale climatic indices. The choice of models was driven by the need to compare the performance of a traditional statistical method with a more advanced machine learning approach, enabling an objective evaluation of their suitability for operational rainfall forecasting.

Linear Regression (LR): Linear regression (LR) serves as a baseline statistical model, assuming a linear relationship between predictors (selected climate indices) and the target variable (monthly rainfall)[13]. Each predictor contributes additively to the predicted outcome, with a corresponding weight representing its influence. Despite its simplicity, LR is a valuable benchmark due to its interpretability, computational efficiency and widespread use in hydrological and climatological studies.

M5P Model Tree: The M5P model tree is a hybrid machine learning technique that combines decision tree algorithms with localized linear regression models at the terminal nodes[3]. The algorithm recursively partitions the predictor space into smaller, more homogeneous regions, fitting a separate linear regression model within each terminal node. This structure enables the M5P model to capture both global trends and localized variations, making it particularly effective for modeling complex, nonlinear relationships such as those between climate indices and rainfall.

Rationale for Model Selection: The selection of LR and M5P was intentional. LR provides a well-established and interpretable baseline that can reveal the overall influence of each climate index on rainfall. However, rainfall processes are often influenced by non-linear and interacting effects among climate drivers. This complexity motivated the inclusion of the M5P model tree, which can accommodate non-linearities without the need for explicit feature engineering. The hybrid nature of M5P combining decision tree segmentation with linear models, offers the potential to capture patterns that LR might overlook, especially when predictor-response relationships vary across different climatic regimes.

Data Preparation and Preprocessing: The dataset comprised of monthly rainfall observations and contemporaneous values of the selected climatic indices include Niño 3.4, IOD, NAO, PDO and AMO, spanning multiple decades to capture a wide range of climate variability. Data quality checks were performed to address missing or erroneous entries, with gaps in climate indices filled via interpolation or alternative sources to maintain temporal continuity. Rainfall and climate index values were temporally aligned to ensure correspondence within the same month and year, preventing biases in model performance. Feature scaling was applied for linear regression to standardize predictor ranges and support numerical stability, while it was unnecessary for tree-based models like M5P.

Training and Testing Strategy: To simulate realistic forecasting conditions, the dataset was divided into training and testing subsets. The training set included earlier years of the record, while the testing set contained the most recent years. This split ensured that the model was evaluated on completely unseen data, reflecting the challenge of making predictions for future periods without knowledge of upcoming climate states.

A key consideration in this study was that no lagged features were included. This means that the models were trained solely on contemporaneous values of the climate indices for each month's rainfall prediction. The decision to exclude lagged predictors was guided by operational forecasting needs. Future rainfall predictions must rely only on climate index values known at the time of prediction, without assuming availability of future climate data.

Model Implementation: For LR, the implementation involved estimating the coefficients of each climate index through a least-squares optimization procedure. The simplicity of this approach allows for rapid computation, making it suitable for operational use in resource-limited contexts.

For M5P, the algorithm first constructed a decision tree by recursively splitting the data into subsets that minimized intra-node variance. At each terminal node, a local linear regression model was fitted to the observations within that node. This two-step approach enabled the model to capture piecewise linear patterns, adapting to variations in climate–rainfall relationships across different subgroups of the data. Pruning was employed to avoid overfitting and model performance was validated using the testing set.

Model Evaluation Metrics: To compare the predictive skill of LR and M5P, standard statistical performance metrics were employed such as:

- **Coefficient of Correlation (CC):** To measure the variance in rainfall model.
- **Root Mean Square Error (RMSE):** To assess the magnitude of prediction errors.
- **Mean Absolute Error (MAE):** To evaluate the average magnitude of errors in a less sensitive manner to outliers compared to RMSE.

The use of multiple metrics provided a comprehensive view of model performance, accounting for both accuracy and reliability.

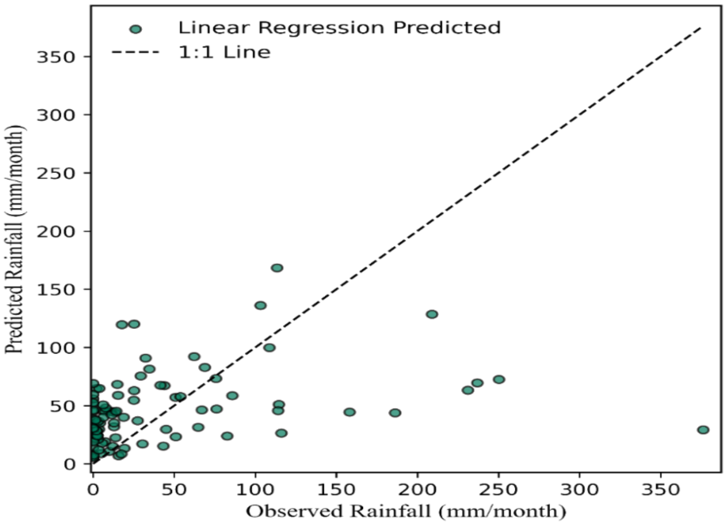

Figure 1: Scatter plots showing the predicted versus observed rainfall using the Linear Regression model for the testing datasets.

Table 1
Performance comparison of M5P and Linear Regression (LR) models in forecasting rainfall, showing training and testing metrics

	CC	MAE	RMSE	SI	NSE	CC	MAE	RMSE	SI	NSE
Models	Training					Testing				
M5P	0.7616	25.9738	50.6946	1.1817	0.5786	0.6515	28.2961	51.0526	1.2935	0.3919
LR	0.4375	42.9063	70.2221	1.6369	0.1914	0.3359	41.6945	62.4801	1.5831	0.0893

https://doi.org/10.25303/192da070076

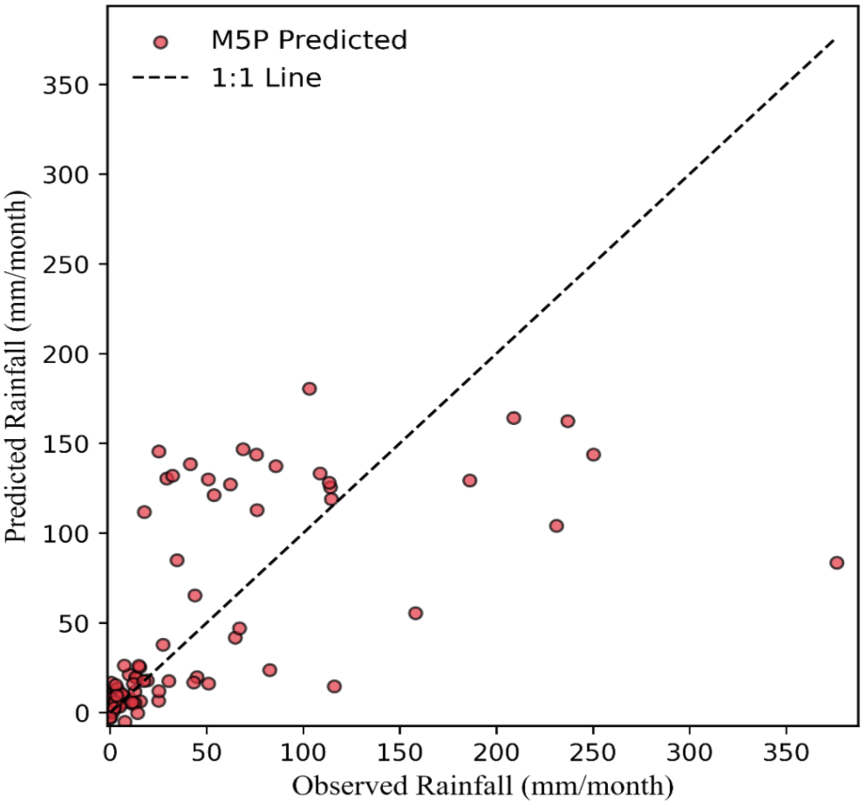

Figure 2 : Scattered plots showing the predicted versus observed rainfall using the M5P model for the testing datasets.

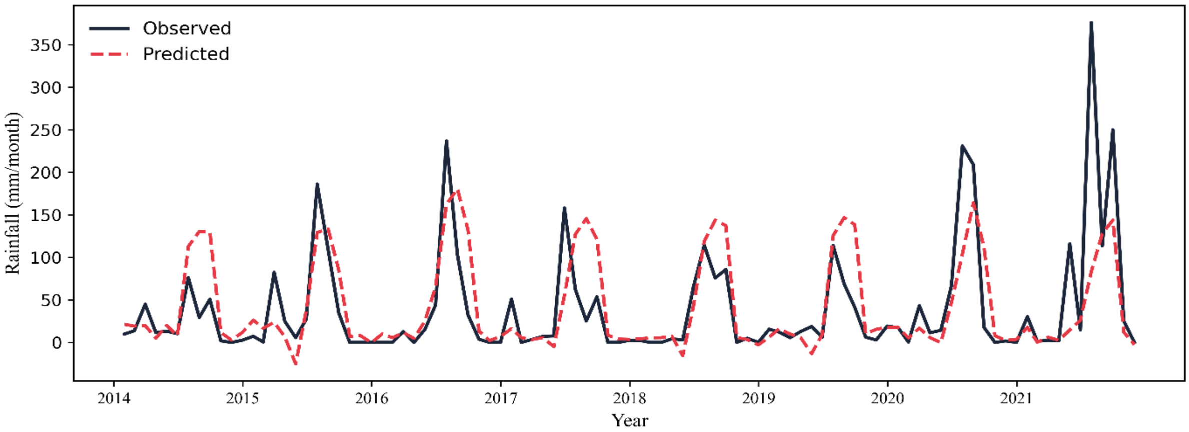

Figure 3: Times Series showing the predicted versus observed rainfall using the M5P model for the testing datasets.

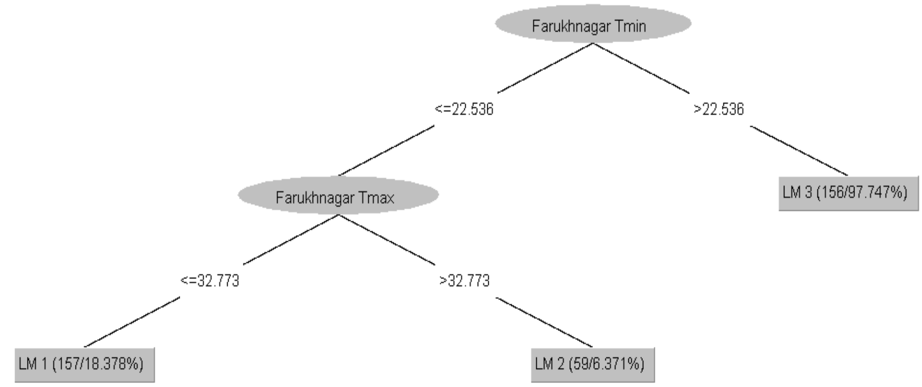

Figure 4: M5P model tree illustrating the partitioning of Farukhnagar minimum (Tmin) and maximum (Tmax) temperatures for predicting rainfall classes (LM 1, LM 2, LM 3) with corresponding sample counts and percentages.

https://doi.org/10.25303/192da070076

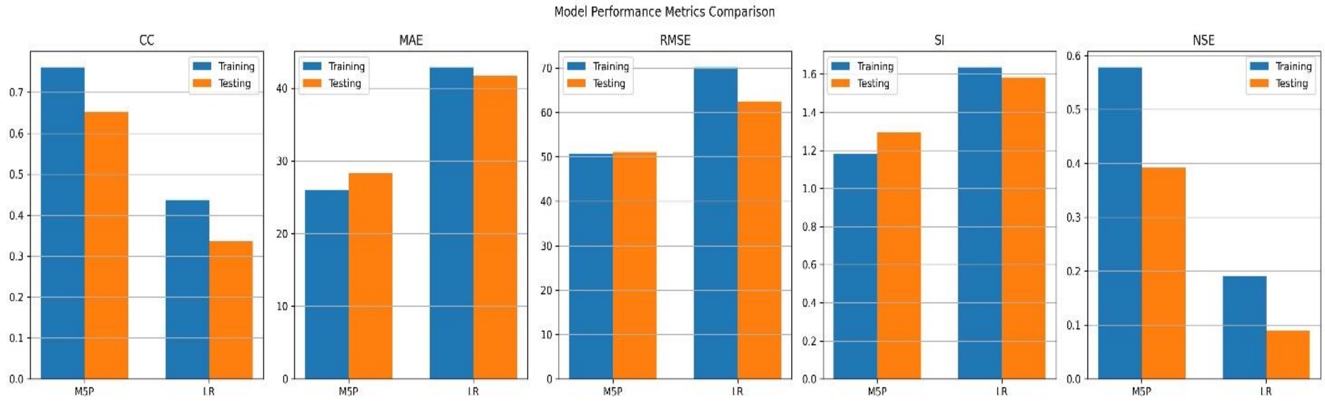

Figure 5: Performance evaluation of LR and M5P tree

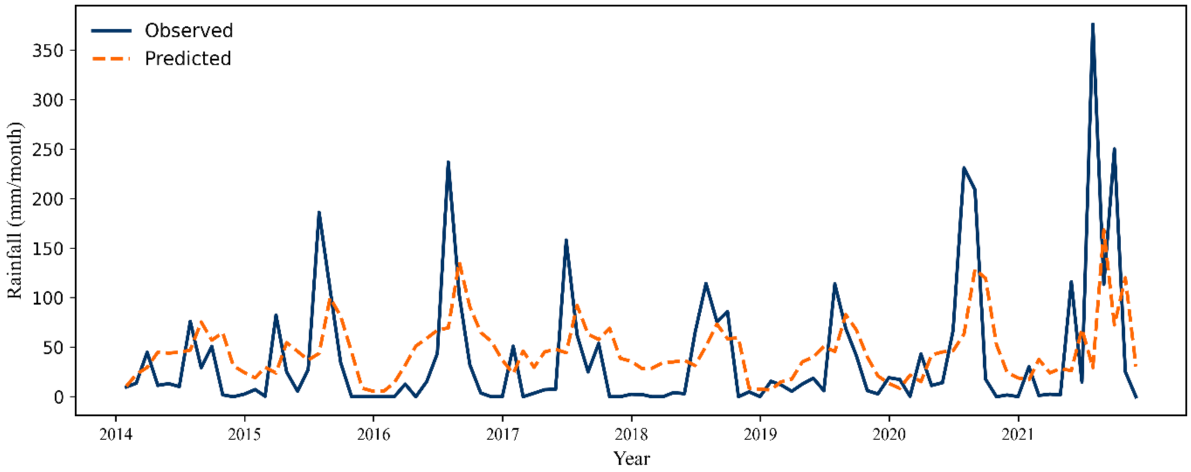

Figure 6: Time series plots of Best M5P tree

Results and Discussion

Table 1 presents the performance of M5P and linear regression (LR) models for monthly rainfall prediction, using performances of each models. During the training phase, the M5P model demonstrated superior performance with a CC of 0.7616, indicating a strong correlation between observed and predicted rainfall, compared to 0.4375 for LR. The MAE and RMSE values of M5P (25.97 mm and 50.69 mm) were substantially lower than those of LR (42.91 mm and 70.22 mm), reflecting higher predictive accuracy. Similarly, the NSE of 0.5786 for M5P was markedly higher than 0.1914 for LR, further confirming better model efficiency. In the testing phase, M5P maintained relatively robust performance with a CC of 0.6515 and NSE of 0.3919 whereas LR showed poor generalization (CC = 0.3359, NSE = 0.0893). The MAE and RMSE for M5P in testing (28.30 mm and 51.05 mm) remained lower than those of LR (41.69 mm and 62.48 mm), indicating more reliable predictions on unseen data. The slight increase in errors for M5P during testing suggests minor overfitting, yet overall performance remained superior.

These results highlight the advantage of tree-based models like M5P in capturing nonlinear relationships among climatic predictors whereas LR's linear framework limits its predictive capability. Overall, M5P is better suited for monthly rainfall forecasting in the study region as corroborated by the closer alignment of predicted and observed values in the corresponding scatter plots (Figure 2).

Conclusion

The comparative evaluation confirms that the M5P model consistently outperforms linear regression in monthly rainfall forecasting. M5P demonstrates higher correlation with observed rainfall, lower MAE and RMSE values and superior Nash–Sutcliffe Efficiency, reflecting more accurate and reliable predictions. Its ability to capture complex, nonlinear interactions among multiple climatic drivers allows it to generalize effectively to unseen data, unlike the linear LR model.

Although LR provides interpretability and simplicity, its predictive capability is limited by its linear framework. Slightly higher errors observed in testing for M5P indicate minor overfitting, but overall performance remains robust. These findings highlight the advantage of tree-based models in representing intricate teleconnections influencing regional rainfall. By integrating large-scale climate indices, M5P can serve as a dependable tool for operational rainfall prediction. Its robust performance supports the development of early-warning systems for climate-resilient water management and agricultural planning.

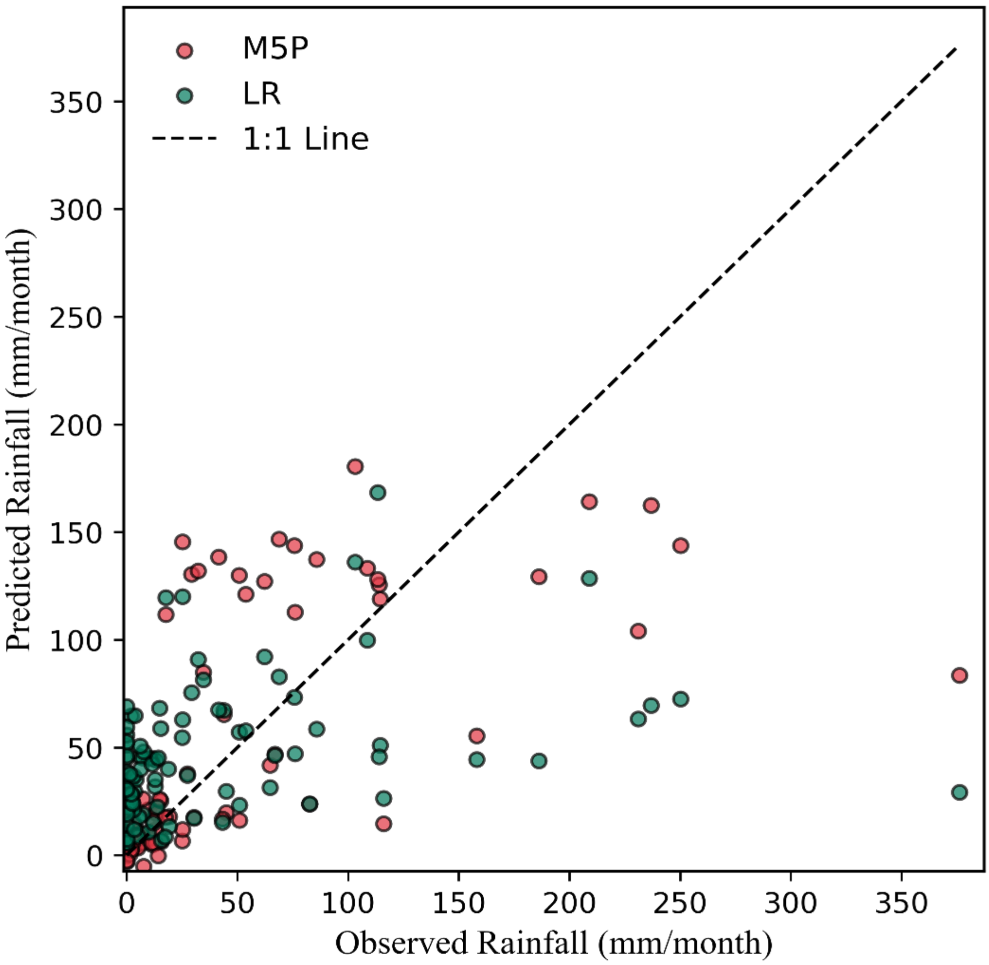

Figure 7: Performance evaluation of Best LR and M5P tree

Overall, the study underscores the potential of nonlinear machine learning frameworks in enhancing the precision of rainfall forecasts. M5P emerges as a highly effective approach for practical, high-accuracy monthly rainfall prediction in the study region.

A major innovation of this research is the utilization of satellite data both as input variables and as a reference for cross-validating extreme rainfall events. The strong correlations observed between satellite-derived predictors and ground-based rainfall data further reinforced this conclusion. Collectively, these results introduce a novel scientific framework for validating extreme rainfall events using remote sensing information, forming a two-tier validation approach not previously reported in existing studies.

In addition, the study offers a practical advancement by demonstrating that monthly rainfall can be reliably predicted without the need for physical rain gauge networks. Given the substantial expenses and logistical difficulties of installing and maintaining such systems, particularly in remote or conflict-prone regions, the proven accuracy of satellite-supported models has advantage like low-cost and scalable solution for hydrological assessment and rainfall monitoring.

References

1. Ashok K., Guan Z. and Yamagata T., Influence of the Indian Ocean Dipole on the Australian winter rainfall, *Geophysical Research Letters*, **30(15)**, https://doi.org/10.1029/2003GL017926, 1821 **(2003)**

2. Enfield D.B., Mestas-Nuñez A.M. and Trimble P.J., The Atlantic multidecadal oscillation and its relation to rainfall and river flows in the continental US, *Geophysical Research Letters*, **28(10)**, 2077–2080, https://doi.org/10.1029/2000GL012745 **(2001)**

3. Ghorbani K., Salarijazi M. and Ghahreman N., Developing stepwise M5 tree model to determine the influential factors on rainfall prediction and to overcome the greedy problem of its algorithm, *Water Resources Management*, **36(12)**, 3327–3348, https://doi.org/10.1007/s11269-022-03203-3 **(2022)**

4. Hurrell J.W., Decadal trends in the North Atlantic Oscillation: Regional temperatures and precipitation, *Science*, **269(5224)**, 676–679, https://doi.org/10.1126/science.269.5224.676 **(1995)**

5. Knight J.R., Folland C.K. and Scaife A.A., Climate impacts of the Atlantic Multidecadal Oscillation, *Geophysical Research Letters*, **33(1)**, L17706 **(2006)**

6. Kumar G.P. and Dwarakish G.S., Machine learning-based ensemble of global climate models and trend analysis for

projecting extreme precipitation indices under future climate scenarios, *Environmental Monitoring and Assessment*, **197(1)**, 1009, https://doi.org/10.1007/s10661-025-14469-6 **(2025)**

7. Kumar K.K., Pant G.B. and Parthasarathy B., On the weakening relationship between the Indian monsoon and ENSO, *Science*, **284(5423)**, https://doi.org/10.1126/science.284.5423.2156, 2156–2159 **(2001)**

8. Mantua N.J., Hare S.R., Zhang Y., Wallace J.M. and Francis R.C., A Pacific interdecadal climate oscillation with impacts on salmon production, *Bulletin of the American Meteorological Society*, **78(6)**, 1069–1079, https://doi.org/10.1175/1520-0477(1997)078<1069:APICOW>2.0.CO;2 **(1997)**

9. National Centers for Environmental Information, Equatorial Pacific sea surface temperatures (SST), https://www.ncei.noaa.gov/access/monitoring/enso/sst **(2025)**

10. Newman M., Alexander M.A. and Scott J.D., The Pacific Decadal Oscillation, revisited, *Journal of Climate*, **29(12)**, 4399–4427, https://doi.org/10.1175/JCLI-D-15-0508.1 **(2016)**

11. Saji N.H., Goswami B.N., Vinayachandran P.N. and Yamagata T., A dipole mode in the tropical Indian Ocean, *Nature*, **401(6751)**, 360–363, https://doi.org/10.1038/43854 **(1999)**

12. Shamim T. et al, Impact of El Nino Southern Oscillation, Indian Ocean Dipole and North Atlantic Oscillation on the drought scenarios in the Upper Indus Basin, *Theoretical and Applied Climatology*, **156(1)**, 234, https://doi.org/10.1007/s00704-025-05459-2 **(2025)**.

(Received 19th November 2025, accepted 15th December 2025)