Nature Environment and Pollution Technology
An International Quarterly Scientific Journal

p-ISSN: 0972-6268
(Print copies up to 2016)
e-ISSN: 2395-3454

Vol. 25 | No. 3 | Article ID B4408 | 2026

Original Research Paper

https://doi.org/10.46488/NEPT.2026.v25i03.B4408

Open Access Journal

Data-Driven Machine Learning Models for Predicting Monthly Rainfall Using Lagged Climate Indices

Ayush Vashisth†, Mohammed Sharif and Mohammed Shakeel
Department of Civil Engineering, Jamia Millia Islamia, New Delhi-110025, India
†Corresponding author: Ayush Vashisth; ayushvashisth90@gmail.com

Abbreviation: Nat. Env. & Poll. Technol.
Website: www.neptjournal.com

Received: 26-10-2025
Revised: 28-11-2025
Accepted: 01-12-2025

Key Words:

Machine learning
Rainfall prediction
Lagged climate indices
AAN-based model

Citation for the Paper:

Vashisth, A., Sharif, M. and Shakeel, M., 2026. Data-driven machine learning models for predicting monthly rainfall using lagged climate indices. *Nature Environment and Pollution Technology*, 25(3), p. B4408. https://doi.org/10.46488/NEPT.2026.v25i03.B4408

Note: From 2025, the journal has adopted the use of Article IDs in citations instead of traditional consecutive page numbers. Each article is now given individual page ranges starting from page 1.

ABSTRACT

Rainfall prediction is vital for several economic activities, including agriculture. The present study is focused on developing machine learning models to predict rainfall using different climate indices. Four machine learning techniques - RF, M5P tree, REP Tree algorithm, and ANN were used for predicting rainfall from climate indices. The practicality of the approach presented herein was demonstrated through application to rainfall data at the Safdarjung meteorological station in New Delhi, India. The machine learning models use climate indices-Indian Ocean Dipole, El Niño–Southern Oscillation, Pacific Decadal Oscillation, Atlantic Multidecadal Oscillation, and North Atlantic Oscillation as feature variables and rainfall as the target variable. The dataset was divided into a training set and a testing set. The ratio of the split was eighty to twenty. Among the four machine learning models evaluated, the random forest model demonstrates the best performance, with coefficients of correlation of 0.9853 and 0.6674, mean absolute errors of 15.1624 and 33.7026, and root mean square errors of 25.1173 and 44.0744 for training and testing phases, respectively. The Taylor diagram also showed that the RF-based model performed better at predicting rainfall.

1. INTRODUCTION

Rainfall is a crucial atmospheric parameter that plays a fundamental role in the water cycle, influencing the availability and distribution of freshwater resources globally. Excessive rainfall leads to severe weather events like catastrophic floods or prolonged drought. These extreme weather conditions disrupt daily life and cause substantial socioeconomic losses, including damage to infrastructure, reduced agricultural yields, and heightened food insecurity. There is a heavy reliance on rainfall, given that in an agricultural country like India, where 65% of India's agricultural land is, the industry is extremely susceptible to changes in rainfall. Accurate rainfall forecasting is needed for agricultural planning and water resource management. However, rainfall is still a significant scientific difficulty because of its stochastic character. Enhancing the precision of rainfall forecasts is crucial for maintaining food security, economic stability, and sustainable development, especially in regions where agriculture forms the backbone of the economy. (Abbot et al. 2014, Kashid et al.2012, Khan et al.2020).

Precipitation of large-scale climate predictors is essential for effective water resource management. ENSO indices with trade wind indicators strongly influence rainfall. SST predictors are often inversely related to seasonal precipitation, and basin-specific climate indices models improve forecast accuracy at longer lead times. (Dhillon et al. 2023, Asakereh 2020, Saji et al. 1999, Luo et al. 2017).

Numerical models use wind patterns, cloud formation, and moisture content. Although the models give accurate depictions of the climate system, they are frequently constrained by the intricacy of atmospheric dynamics, model structure, and parameter uncertainty. (Choubin et al. 2018). Alternatively, empirical models have become well known for their capacity to produce precise projections, especially in areas where uncertainties may make physical-based models unreliable. To determine correlations between weather patterns and projected precipitation, these models use statistical methods. Artificial neural networks (ANNs) and multiple linear regressions (MLRs) are two machine learning models that have demonstrated a great deal of promise in raising the accuracy of rainfall forecasting. The ANNs are especially good at managing complicated climatic systems because they can capture non-linear interactions (Haykin 2009). The precision of rainfall forecasts helped to improve resource management and agricultural planning. RF models often outperform LS-SVM and M5P in terms of efficiency and lower errors. (Pandhiani et al. 2020, Wani et al. 2024, Shankar 2024).

Naved et al. (2024) concluded that laboratory tests to predict the strength of concrete would be significantly reduced by optimal DNN models. Asim et al. (2022) highlight the significance of hyperparameter optimization. To improve predicted accuracy, scientists used a methodical methodology to adjust the DNN's architecture, neurons in each layer, activation functions, and learning rates.

Additionally, incorporating large-scale climatic indices like IOD, PDO, AMO, ENSO and NAO into forecasting systems has the potential to significantly improve prediction accuracy, as these teleconnection patterns. (Enfield et al. 2001, Feng et al. 2020, Mantua et al. 1997, Sham et al. 2025). The advanced forecasting methodologies are essential for supporting sustainable development and resilience to climate variability, particularly in regions where rainfall directly impacts agricultural productivity.

Most existing studies concentrate on correlations and climate indices, often overlooking the importance of lagged effects, nonlinear relationships, and interactions among multiple indices. Additionally, predictive models rarely incorporate a systematic examination of lag structures and temporal dynamics, which are essential for capturing delayed teleconnection patterns.

The present research aims to evaluate the performance of multiple machine learning models, including tree-based models, such as the RF, M5P tree, REPTree, and ANN, for multi-step rainfall prediction based on climate indices. This research is significant to improve the precision of rainfall prediction, which is critical for agricultural, water resource management, and disaster preparedness, especially in rainfall-dependent areas for agricultural activities.

2. MATERIALS AND METHODS

This study employs a two-step methodology. The first step involves the collection of rainfall data and large-scale climate indices. The Safdarjung is located in central Delhi, the capital of India. The station is operated by the Indian Meteorological Department (IMD) and serves as one of the key observatories for meteorological data collection in the region. The geographical coordinates of the station are approximately 28.585° N latitude and 77.221° E longitude. Monthly datasets of climatic indices and their respective data sources are provided in Table 1.

In the second step, machine learning models, including RF, M5P tree, REPTree and ANN, were trained using lagged climate indices as input variables and rainfall as the target variable. Once trained, these models were utilized to predict rainfall using the values of climate indices and performance indicators; the most efficient machine learning model was identified.

2.1. Performance Metrics

2.1.1. Mean Absolute Error

Mean Absolute Error (MAE) measures how well a regression model performs. This calculation does not consider the direction of the differences.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|PR_i - OR_i| \qquad ...(1)$$

2.1.2. Root Mean Square Error

RMSE shows how well regression models work - it calculates how much the prediction mistakes differ on average. But it weighs big mistakes more than MAE because it squares the differences. RMSE punishes big mistakes more than small ones.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(PR_i - OR_i)^2} \qquad ...(2)$$

2.1.3 Correlation Coefficient (R)

In the context of machine learning, it is used to evaluate how well it behaves.

$$R = \frac{\sum_{i=1}^{N}\left(PR_i - \underline{OR}\right)(OR_i - \underline{OR})}{\sqrt{\sum_{i=1}^{N}(PR_i - \underline{OR})^2 \sum_{i=1}^{N}(OR_i - \underline{OR})^2}} \qquad ...(3)$$

Scattering index (SI): $SI = \frac{RMSE}{\underline{OR}}$ …(4)

2.2. Machine Learning Models

2.2.1. M5P Model Tree

This Model was first introduced by Quinlan (1992), showing a large step forward in predictive modeling. This approach utilizes a binary decision tree structure in which the terminal

Table 1: Sources of climate indices and rainfall data.

S.No.	Feature Variable	Source	Reference
1.	Indian Ocean Dipole	https://psl.noaa.gov/gcos_wgsp/Timeseries/Data/dmi.had.long.data	Saji et al. (1999)
2.	El Niño-Southern Oscillation	https://www.cpc.ncep.noaa.gov/data/indices/sstoi.indices	Trenberth and Hoar (1997)
3.	Pacific Decadal Oscillation	https://psl.noaa.gov/data/correlation/pdo.data	Mantua et al. (1997)
4.	Atlantic Multidecadal Oscillation	https://psl.noaa.gov/data/correlation/amon.us.data	Enfield et al. (2001)
5.	North Atlantic Oscillation	https://psl.noaa.gov/data/correlation/nao.data	Hurrell (1995)
6.	Rainfall	https://imdpune.gov.in/cmpg/Griddata/Rainfall_25_Bin.html	Pai et al. (2014)

nodes, or leaves, contain linear regression functions. By integrating decision trees with linear regression, the M5P model effectively addresses nonlinear problems while maintaining interpretability and flexibility. The splitting of nodes is guided by error reduction, measured using the variance of class values at each node.

2.2.2. Artificial Neural Networks (ANN)

It is used to solve a specific set of closely interrelated functioning components called neurons that operate in parallel. After adequate calibration, the neural network can anticipate the output configuration and identify similarities when given a fresh input configuration. Even if a certain input has never been known before, neural networks may recognize connections in inputs. (Abebe et al.2023, Mokhtarzad et al. 2017, Sihag 2018)

2.2.3. Reduced Error Pruning Tree (REP Tree)

It is designed to build a decision tree model that is both effective and interpretable, addressing common challenges like overfitting that can arise in tree-based models. The process begins with the construction of a decision tree using a greedy algorithm. These iterative criteria are met, such as a minimum number of instances in a node or a maximum tree depth. The Reduced Error Pruning Tree model randomly selects features to enhance the efficiency of regression tree algorithms and reduce variance-related errors. It combines logistic regression with multiple tree constructions, selecting the simplest tree that maintains predictive accuracy. In the training dataset, a flexible yet straightforward modeling approach is applied to a post-pruning algorithm to remove unnecessary nodes, minimizing tree complexity while preserving accuracy. The model is intuitive, fast, and improves generalization, but it relies on a validation set, is sensitive to data noise, and uses a greedy algorithm that may lead to suboptimal decisions.

2.2.4. Random Forest (RF)

An ensemble of randomized regression trees is a widely used technique for both classification and regression tasks.

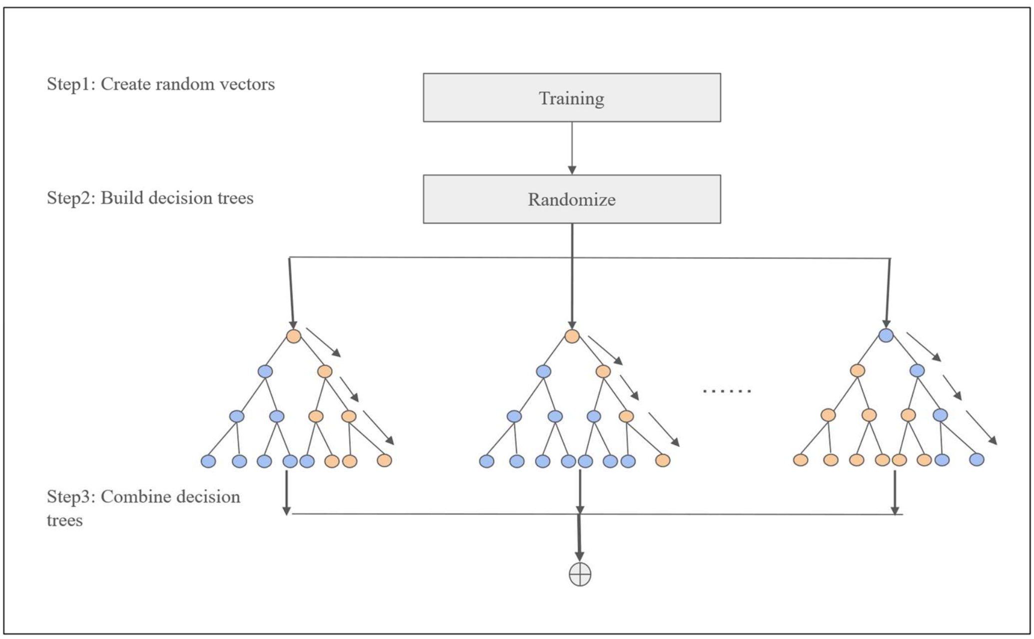


Fig. 1: RF model development steps.

It constructs an ensemble model that consists of decision trees, each trained on a different subset of data, to improve prediction accuracy. In this model, numerical features are randomly assigned to class labels, while in regression, each tree grows using a randomly selected subset of input parameters at every node (see Fig. 1). Random Forest overcomes this limitation, offering robust predictive performance, computational efficiency, and scalability. The aggregating individual tree predictions. In a single decision tree, sensitivity to small changes in training data often results in high variance and lower accuracy (Breiman 1999).

2.3. Model Runs

The correlation coefficients between different climate indices and rainfall provide valuable insights into their interrelationships and potential influences on precipitation patterns. AMO exhibits the highest positive correlation with rainfall (0.10), suggesting that while the relationship is weak, AMO may still have a subtle yet consistent influence on rainfall variability. This indicates its potential as a supporting predictor in long-term forecasting models. Similarly, PDO (0.07) and NAO (0.06) show modest positive correlations, highlighting their role in shaping regional precipitation trends. Although their individual impact may not be dominant, their combined effects could enhance predictive accuracy when incorporated into advanced models. The correlation between Niño 3.4 and rainfall (-0.04) appears negligible in Fig. 2. However, considering ENSO's well-established influence on global weather patterns, further investigations into lagged effects could provide deeper insights, as climate indices often exhibit delayed impacts on regional rainfall patterns due to ocean-atmosphere interactions. The near-zero correlation of IOD with rainfall implies minimal direct influence, but given its dynamic nature, future research incorporating time-lagged relationships could uncover hidden dependencies, capturing the delayed response of rainfall to changes in climate indices.

Although the correlation matrix indicates weak linear links between rainfall and climate indices, it reflects only direct pairwise relationships. The models can still achieve high predictive accuracy by capturing nonlinear patterns, lagged effects, and interactions among multiple indices that are not visible in simple correlations.

Thirteen models (N1-N13) were developed using different combinations of monthly rainfall, climate indices, and their lagged values to predict rainfall at Safdarjung, Delhi, India. The specific input variables incorporated in each model are summarized in Table 2.

Fig. 2: Correlation matrix plot among inputs and output variables.

Table 2: Input variables incorporated in each model of climate indices and rainfall.

Model	Input Combination
N1	NI(t-1), PD(t-1), NAO (t-1), AMO(t-1), IOD(t-1), R(t-1)
N2	NI(t-2), NI(t-1), NAO (t-2), NAO (t-1), PD(t-2), PD(t-1), AMO(t-2), AMO(t-1), IOD(t-2), IOD(t-1), R(t-2), R(t-1)
N3	NI(t-3),NI(t-2),NI(t-1),NAO (t-3),NAO (t-2),NAO (t-1),PD(t-3), PD(t-2),PD(t-1),AMO(t-3),AMO(t-2), AMO(t-1),IOD(t-3),IOD(t-2), IOD(t-1),R(t-3), R(t-2),R(t-1)
N4	NI(t-3),NI(t-2),NI(t-1),NAO (t-3),NAO (t-2),NAO (t-1),PD(t-3), PD(t-2),PD(t-1),AMO(t-3),AMO(t-2), AMO(t-1),IOD(t-3),IOD(t-2), IOD(t-1),R(t-3), R(t-2),R(t-1),R(t-4)
N5	NI(t-3),NI(t-2),NI(t-1),NAO (t-3),NAO (t-2),NAO (t-1),PD(t-3), PD(t-2),PD(t-1),AMO(t-3),AMO(t-2), AMO(t-1),IOD(t-3),IOD(t-2), IOD(t-1),R(t-3), R(t-2),R(t-1),R(t-4), R(t-5)
N6	NI(t-3),NI(t-2),NI(t-1),NAO (t-3),NAO (t-2),NAO (t-1),PD(t-3), PD(t-2),PD(t-1),AMO(t-3),AMO(t-2), AMO(t-1),IOD(t-3),IOD(t-2), IOD(t-1),R(t-3), R(t-2),R(t-1),R(t-4), R(t-5),R(t-6)
N7	NI(t-3),NI(t-2),NI(t-1),NAO (t-3),NAO (t-2),NAO (t-1),PD(t-3), PD(t-2),PD(t-1),AMO(t-3),AMO(t-2), AMO(t-1),IOD(t-3),IOD(t-2), IOD(t-1),R(t-3), R(t-2),R(t-1),R(t-4), R(t-5), R(t-6), R(t-7)
N8	NI(t-3),NI(t-2),NI(t-1),NAO (t-3),NAO (t-2),NAO (t-1),PD(t-3), PD(t-2),PD(t-1),AMO(t-3),AMO(t-2), AMO(t-1),IOD(t-3),IOD(t-2), IOD(t-1),R(t-3), R(t-2),R(t-1),R(t-4), R(t-5), R(t-6), R(t-7), R(t-8)
N9	NI(t-3),NI(t-2),NI(t-1),NAO (t-3),NAO (t-2),NAO (t-1),PD(t-3), PD(t-2),PD(t-1),AMO(t-3),AMO(t-2), AMO(t-1),IOD(t-3),IOD(t-2), IOD(t-1),R(t-3), R(t-2),R(t-1),R(t-4), R(t-5), R(t-6), R(t-7), R(t-8), R(t-9)
N10	NI(t-3),NI(t-2),NI(t-1),NAO (t-3),NAO (t-2),NAO (t-1),PD(t-3), PD(t-2),PD(t-1),AMO(t-3),AMO(t-2), AMO(t-1),IOD(t-3),IOD(t-2), IOD(t-1),R(t-3), R(t-2),R(t-1),R(t-4), R(t-5), R(t-6), R(t-7), R(t-8), R(t-9), R(t-10)
N11	NI(t-3),NI(t-2),NI(t-1),NAO (t-3),NAO (t-2),NAO (t-1),PD(t-3), PD(t-2),PD(t-1),AMO(t-3),AMO(t-2), AMO(t-1),IOD(t-3),IOD(t-2), IOD(t-1),R(t-3), R(t-2),R(t-1),R(t-4), R(t-5), R(t-6), R(t-7), R(t-8), R(t-9), R(t-10), R(t-11)
N12	NI(t-3),NI(t-2),NI(t-1),NAO (t-3),NAO (t-2),NAO (t-1),PD(t-3), PD(t-2),PD(t-1),AMO(t-3),AMO(t-2), AMO(t-1),IOD(t-3),IOD(t-2), IOD(t-1),R(t-3), R(t-2),R(t-1),R(t-4), R(t-5), R(t-6), R(t-7), R(t-8), R(t-9), R(t-10), R(t-11),R(t-12)
N13	NI(t), PD(t), NAO (t), AMO(t),IOD(t), R(t-1),R(t-2), R(t-3),R(t-4),R(t-5), R(t-6),R(t-7),R(t-8),R(t-9),R(t-10),R(t-11),R(t-12)

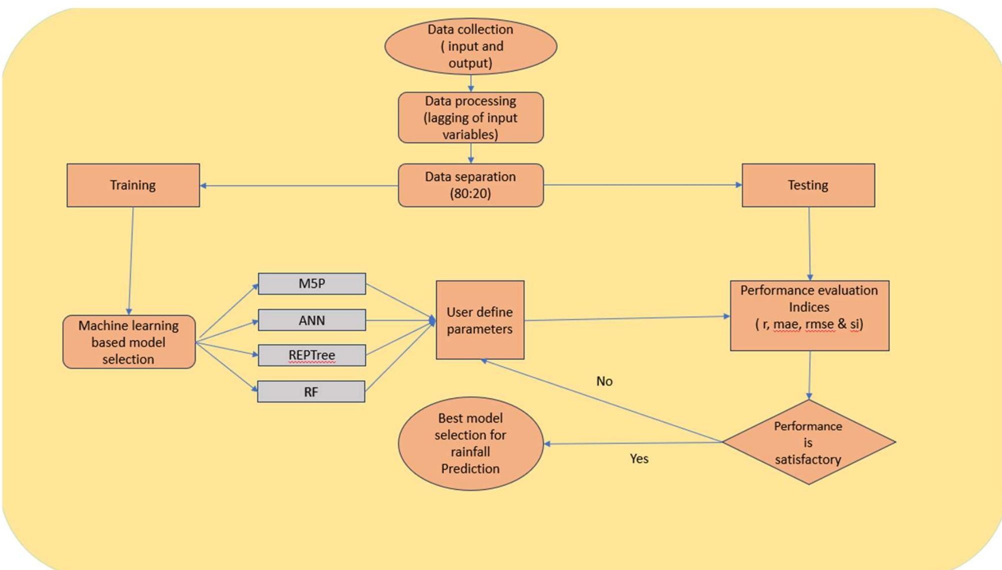


Fig. 3: Flowchart of the Methodology used for the prediction of rainfall.

2.4. Model Development

The optimal model parameters were determined through iterative trial-and-error, guided by established statistical criteria. Fig. 3 illustrates the methodological framework applied for rainfall prediction.

3. RESULTS AND DISCUSSION

3.1. M5P Results

The investigation comprised the creation of the M5P model via the use of the WEKA 3.8.6 software. It was decided

Table 3: Performance of developed M5P-based models.

Models	R	MAE	RMSE	SI	r	MAE	RMSE	SI
	Training				Testing			
M5P N1	0.5176	48.48	79.2594	1.5106	0.4059	45.4634	59.5555	1.4185
M5P N2	0.6058	45.1386	73.7042	1.4047	0.5139	43.2807	59.7516	1.4232
M5P N3	0.5908	48.8041	74.7343	1.4244	0.473	44.4037	55.9193	1.3319
M5P N4	0.5959	48.6935	74.3821	1.4176	0.4705	44.5282	55.8548	1.3304
M5P N5	0.6009	48.6046	74.0374	1.4111	0.4862	43.4679	55.1814	1.3143
M5P N6	0.7271	38.8033	63.6872	1.2138	0.4951	45.0456	65.5169	1.5605
M5P N7	0.8405	32.8633	51.1281	0.9745	0.4774	43.8179	65.3326	1.5561
M5P N8	0.7437	38.0853	62.0192	1.182	0.5137	42.628	63.916	1.5224
M5P N9	0.6349	47.6092	71.5632	1.3639	0.5342	43.3453	53.9078	1.284
M5P N10	0.6339	47.6002	71.6378	1.3653	0.5376	43.1718	54.0531	1.2874
M5P N11	0.7437	38.0978	61.9306	1.1803	0.5234	41.0272	59.2494	1.4112
M5P N12	0.8668	28.0312	46.8654	0.8932	0.5571	40.5989	60.74	1.4467
M5P N13	0.6737	41.367	68.4523	1.3046	0.6307	33.7512	45.5471	1.0848

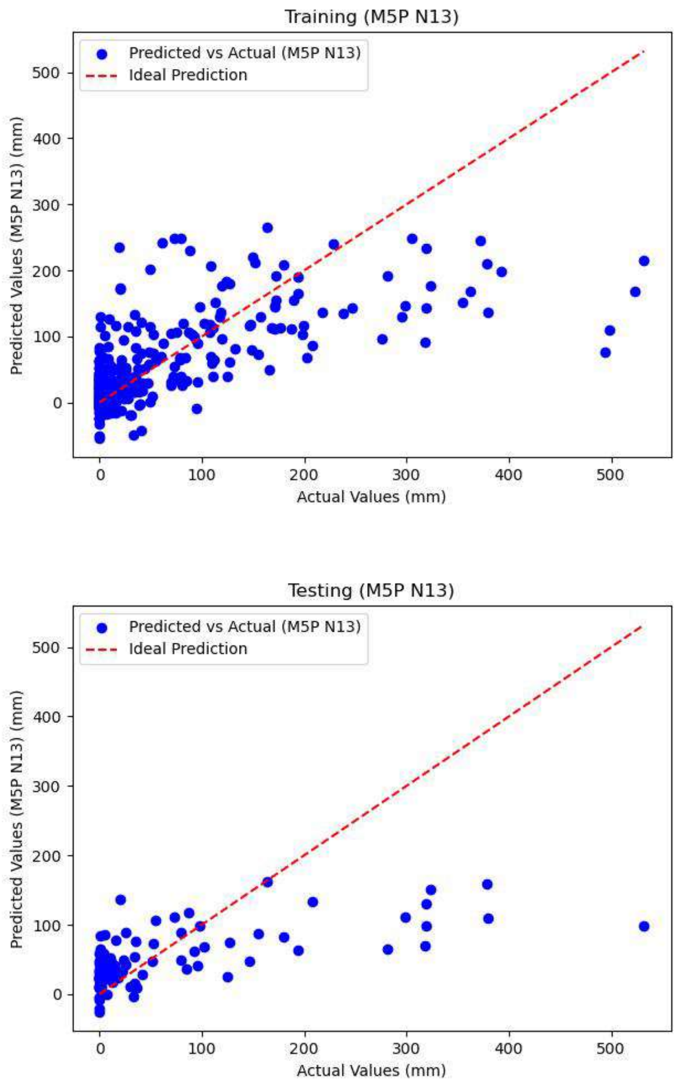


Fig. 4: Scatter plots of M5P N13 for training and testing, respectively.

to split the whole data set into two subgroups. In model building, the larger subset, which comprises eighty percent of the data, was employed, while the remaining twenty percent of the observations were utilized for model validation. The creation of the M5P model is accomplished via the use of a systematic methodology that involves trial and error. To facilitate a consistent assessment across a variety of M5P tree-based models, lagging months were included as a variable in thirteen different models that were developed for the research. The M5P N13 models have been shown to have greater capabilities in terms of forecasting rainfall levels, as shown in Table 3 by the performance assessment metrics. When it comes to testing, the values of R are 0.6307, whereas the values for training are 0.6737. In the training phase, the MAE values are 41.367, whereas in the testing phase, they are 33.7512. In the training phase, the RMSE values are 68.4523, whereas in the testing phase, they are 45.5471. For training purposes, the SI values are 1.3046, whereas for testing purposes, they are 1.0848. The M5P N13 model was used to create the agreement plot that is shown in Fig. 4.

3.2. ANN Model Results

When compared to M5P-based models, the creation of ANN models is similar to the process of M5P model development. ANN models, including hidden layers, neurons, learning rate, and momentum, critically influence model performance. Hidden layer = 1, number of neurons = 30, momentum = 0.2, and learning rate = 0.1 were found to be ideal for artificial neural networks (ANN), as measured by r, MAE, RMSE, and SI, and are shown in Table 4. When it comes to forecasting rainfall, the performance assessment metrics reveal that ANN N 13 models perform better than other models created. The R values for training are 0.7443, and the values for testing are 0.5758. After training, the MAE values are 33.8971, and after testing, they are 32.6226. In the training phase, the RMSE values are 66.2973, while in the testing phase, they are 51.7643. In the SI phase, the numbers are 1.2636 for training and 1.2329 for testing. The results indicate that a 12-month lag of rainfall, along with climate indices such as NI(t), PD(t), NAO(t), AMO(t), and IOD, exerts a stronger influence than other models. During the testing phase, the ANN N1 model yielded r values of 0.3398, which revealed that it had the lowest performance compared to all the other models that were built. This is shown by the data that is provided in Table 4. With the use of the ANN N13 model, Fig. 5 depicts the agreement plot that was created between the actual and anticipated values of rainfall throughout the training and testing stages. According to the statistics, the values that are the lowest overall are located quite near to the line that represents complete agreement (1:1).

3.3. REPTree Model Results

The performance assessment metrics demonstrate that the REPTree N12 models surpass the other models created for forecasting rainfall. The correlation coefficient (r) is 0.7113 for training and 0.6235 for testing. The MAE = 34.8753 for the training set and 30.3546 for the testing set. The RMSE = 65.1091 for the training set and 50.9935 for the testing set. The SI = 1.2409 for training and 1.2146 for testing. The findings demonstrate that a 12-month lag has a pronounced effect on delays. Table 5 indicates that the performance of REPTree N11 was the worst among all models assessed throughout the testing period, producing r = 0.3588. Fig. 6 illustrates the agreement plot juxtaposing actual and

Table 4: Performance of developed ANN-based models.

Models	r	MAE	RMSE	SI	r	MAE	RMSE	SI
	Training				Testing			
ANN N1	0.5046	47.0033	83.5359	1.5921	0.3398	44.5268	56.5845	1.3477
ANN N2	0.5912	45.4375	81.0723	1.5452	0.371	44.8411	60.9832	1.4525
ANN N3	0.6634	38.0748	73.5992	1.4027	0.3476	42.0743	64.2088	1.5293
ANN N4	0.6888	37.1285	71.6623	1.3658	0.3488	39.3225	62.8145	1.4961
ANN N5	0.6879	37.2417	71.967	1.3716	0.3448	39.6901	62.7206	1.4939
ANN N6	0.7089	36.5343	69.5002	1.3246	0.4705	36.1608	57.9243	1.3796
ANN N7	0.7088	37.3345	70.5309	1.3442	0.4106	38.548	61.5447	1.4659
ANN N8	0.7011	38.7884	72.5552	1.3828	0.3743	38.8323	63.8577	1.521
ANN N9	0.7654	36.7518	67.7697	1.2916	0.4281	38.1315	63.0389	1.5015
ANN N10	0.7659	36.1536	67.6965	1.2902	0.4289	37.4102	62.3641	1.4854
ANN N11	0.7888	31.9488	60.4988	1.153	0.4239	36.7848	61.4796	1.4643
ANN N12	0.7861	32.1256	60.6516	1.156	0.4135	36.7747	61.5798	1.4667
ANN N13	0.7443	33.8971	66.2973	1.2636	0.5758	32.6226	51.7643	1.2329

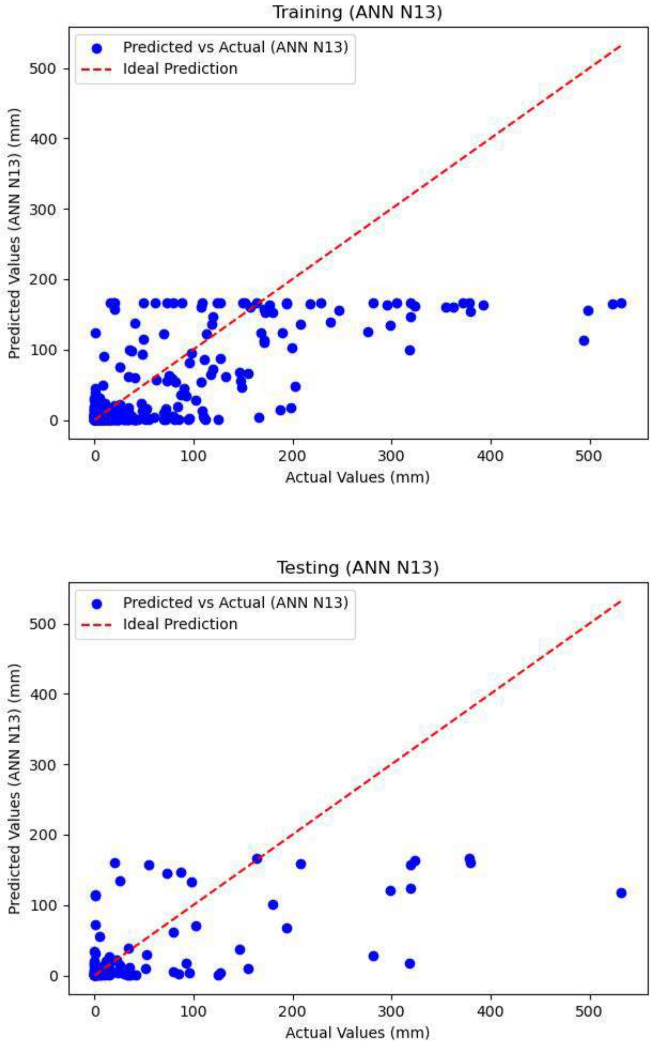


Fig. 5: Scatterplots of ANN N13 for training and testing, respectively.

Table 5: Performance of developed REPTree-based models.

Models	r	MAE	RMSE	SI	r	MAE	RMSE	SI
	Training				Testing			
REPTree N1	0.4444	52.0211	82.9797	1.5815	0.4133	45.723	57.8051	1.3768
REPTree N2	0.4909	48.6592	80.6974	1.538	0.423	43.4846	57.6648	1.3735
REPTree N3	0.5197	46.6973	79.1365	1.5083	0.4073	45.1586	58.5264	1.394
REPTree N4	0.5819	43.6694	75.3326	1.4358	0.4207	42.4785	59.286	1.4121
REPTree N5	0.5756	42.95	75.7424	1.4436	0.4213	42.4928	59.2866	1.4121
REPTree N6	0.5751	43.1905	75.7754	1.4442	0.4207	42.4785	59.286	1.4121
REPTree N7	0.6317	40.2517	71.8066	1.3686	0.4235	43.3999	62.848	1.4969
REPTree N8	0.6317	40.2517	71.8066	1.3686	0.4235	43.3999	62.848	1.4969
REPTree N9	0.6317	40.2517	71.8066	1.3686	0.4235	43.3999	62.848	1.4969
REPTree N10	0.613	41.9281	73.1802	1.3947	0.4526	41.8019	61.6975	1.4695
REPTree N11	0.5726	45.0772	75.942	1.4474	0.3588	46.1041	65.1423	1.5516
REPTree N12	0.7113	34.8753	65.1091	1.2409	0.6235	30.3546	50.9935	1.2146
REPTree N13	0.657	40.5763	69.8293	1.3309	0.6037	33.8162	50.7448	1.2086

anticipated rainfall levels throughout the training and testing periods, using the REPTree N12 model. The data suggest that the greatest lower values approximate the line of perfect concordance (1:1).

3.4. RF Model Results

In this investigation, an RF-based model is also developed. The process of model development is similar to other applied models using the WEKA 3.8.6 software. In the RF-based models, two user-defined parameters were set: the number of trees (k) and the number of variables per node (m). The optimal values were established as k = 10 and m = 1. The performance metrics of the constructed models for both calibration and validation phases are delineated in Table

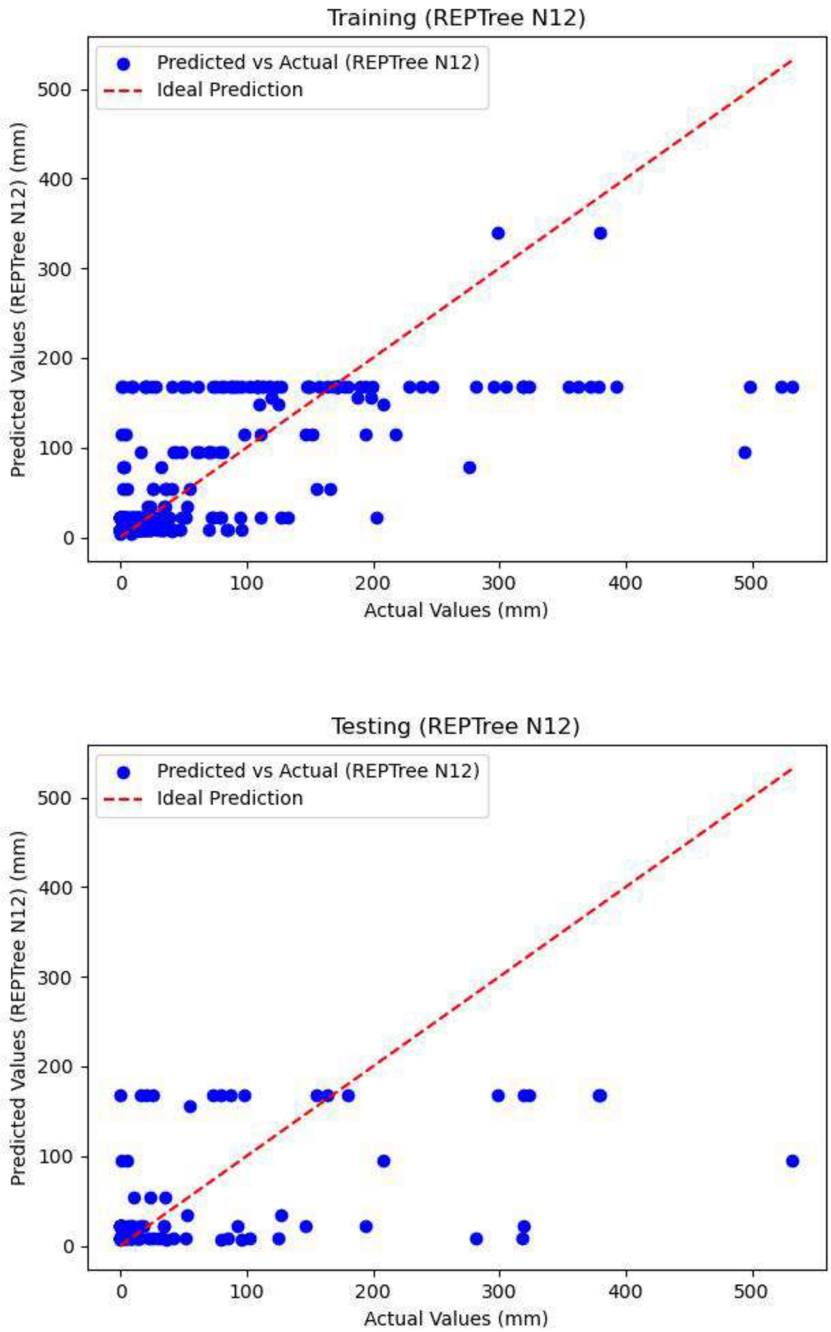
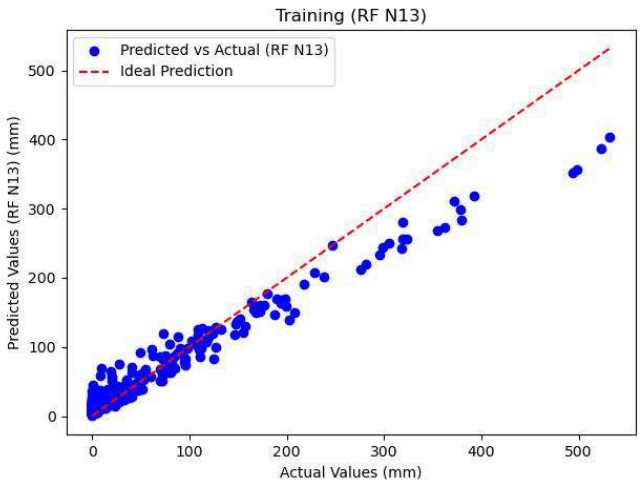


Fig. 6: Agreement diagrams of REPTree N12 for training and testing, respectively.

Table 6: Performance of developed RF-based models.

Models	r	MAE	RMSE	SI	r	MAE	RMSE	SI
	Training				Testing			
RF N1	0.9755	19.3346	31.4962	0.6003	0.2378	54.3335	66.1231	1.5749
RF N2	0.9816	19.3967	30.7187	0.5855	0.3276	50.0159	58.4158	1.3913
RF N3	0.9853	19.0794	30.032	0.5724	0.3823	49.6045	57.8882	1.3788
RF N4	0.9868	18.5794	29.7381	0.5668	0.3877	48.9038	57.7744	1.3761
RF N5	0.9835	18.7462	29.92	0.5702	0.4269	48.3531	56.0248	1.3344
RF N6	0.9852	19.1619	29.8753	0.5694	0.3526	48.0846	57.5738	1.3713
RF N7	0.9848	18.8918	29.862	0.5691	0.4802	45.0805	53.4902	1.274
RF N8	0.9867	18.2123	29.5814	0.5638	0.3994	46.4108	56.0134	1.3341
RF N9	0.9846	18.4241	29.3348	0.5591	0.5821	43.1603	50.6688	1.2068
RF N10	0.9877	17.5852	27.862	0.531	0.605	41.0999	48.454	1.1541
RF N11	0.9865	17.3586	28.148	0.5365	0.5525	41.7795	50.7573	1.2089
RF N12	0.9851	16.6593	27.4421	0.523	0.6594	36.7016	45.6076	1.0863
RF N13	0.9853	15.1624	25.1173	0.4787	0.6674	33.7026	44.0744	1.0498



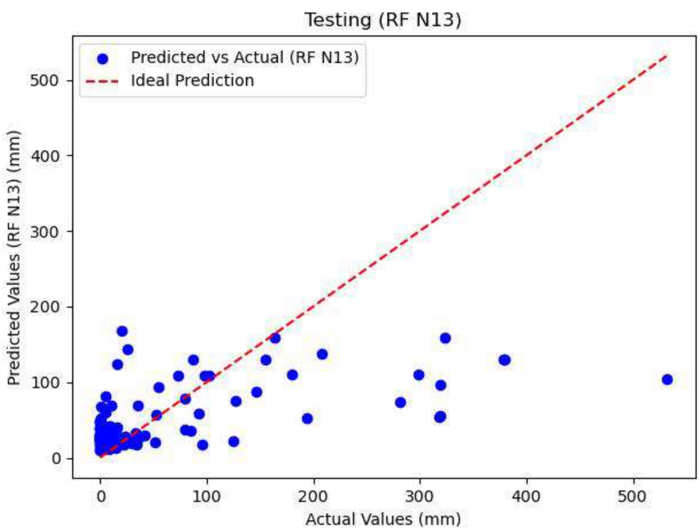


Fig. 7: Agreement diagrams of RF N13 for training and testing, respectively.

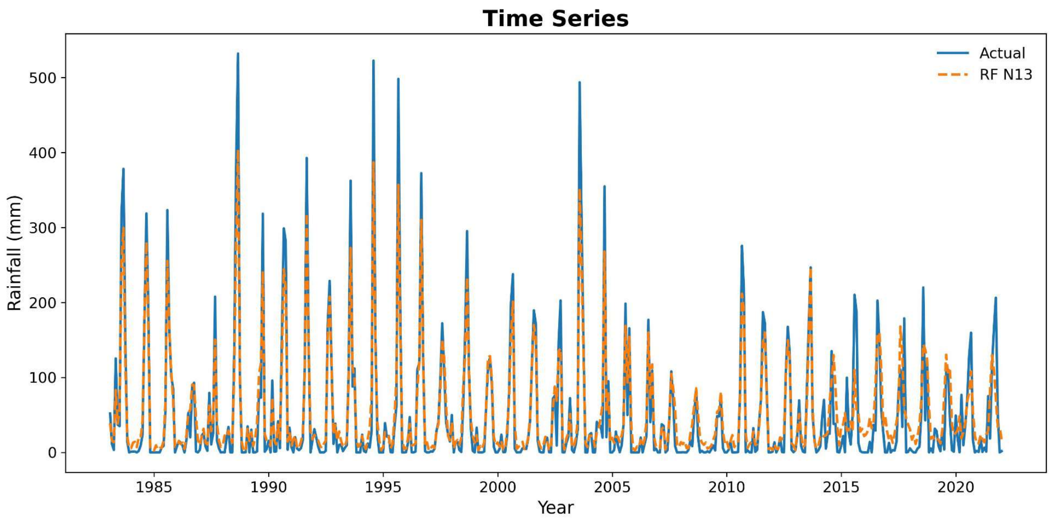


Fig. 8: Hydrograph of actual and predicted values using the RF N13 model for rainfall at Safdarjung Station.

6, measured by r, MAE, RMSE and SI. The performance assessment metrics demonstrate that the RF N13 model works superior to other input combination-based RF models for rainfall prediction. The r = 0.9853 and 0.6674, MAE =15.1624 and 33.7026, RMSE =25.1173 and 44.0744, SI = 0.4787 and 1.0498 for calibration and validation stages, respectively. Fig. 7 illustrates the agreement plot comparing real and anticipated rainfall levels throughout both stages (calibration and validation) with the RF N13 model. Fig. 8 illustrates the performance of the RF-N13 model during evaluation, where the predicted rainfall values align closely with the observed data along the 1:1 line of perfect agreement.

4. DISCUSSION

This research employed M5P, RF, ANN, and REPTree models to forecast rainfall in Safdarjung, Delhi, India. Fig. 9 presents the agreement plots for M5P N13, ANN N13, REPTree N12, and RF N13 models during the testing phase. The plots indicate that the predicted values from the RF N13 model are notably closer to the line of perfect agreement, reflecting superior performance.

The findings from Table 7 indicate that RF had strong performance throughout the training phase, shown by elevated r values and minimal SI, RMSE, and MAE values. The testing stage results indicate that the RF N13 model has superior performance in predicting rainfall for Safdarjung, Delhi, India, with r = 0.6674, MAE of 33.7026, RMSE of 44.0744, and SI of 1.0498. The ANN N13-based model exhibits the worst performance among all models evaluated, with r = 0.5758, MAE values of 32.6226, RMSE values of 51.7643, and SI values of 1.2329 throughout the testing phase. Fig. 10 illustrates the hydrograph of all implemented models with real data. The data demonstrates that the anticipated values from the RF N13 model align with the actual rainfall readings. Fig. 11 illustrates the Violin plots depicting the residual error between real and anticipated values generated by different models. This violin chart indicates that RF N13 is the highest-performing model with the smallest violin width. The Taylor diagram is also plotted for a fair comparison among all applied models. Fig. 12 indicates that RF N13 is the best performing model (blue circle), followed by the red circle (M5P N13 model). These circles lie near the hollow black circle. The Taylor diagram suggests that the ANN N13 model has the least performance (solid brown circle).

5. CONCLUSIONS

The present research concentrates on evaluating the efficacy

Table 7: Performance of developed M5P, ANN, REPTree and RF-based models.

Models	r	MAE	RMSE	SI	r	MAE	RMSE	SI
	Training				Testing			
ANN N13	0.7443	33.8971	66.2973	1.2636	0.5758	32.6226	51.7643	1.2329
M5P N13	0.6737	41.367	68.4523	1.3046	0.6307	33.7512	45.5471	1.0848
REPTree N12	0.7113	34.8753	65.1091	1.2409	0.6235	30.3546	50.9935	1.2146
RF N13	0.9853	15.1624	25.1173	0.4787	0.6674	33.7026	44.0744	1.0498

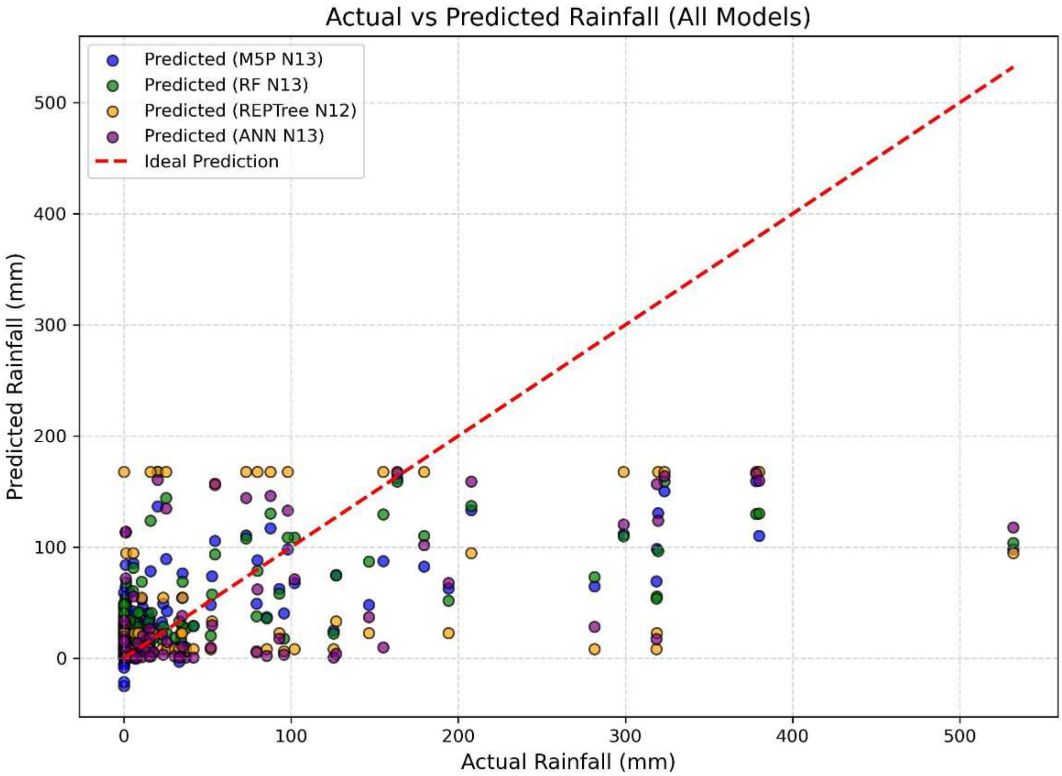


Fig. 9: Agreement diagrams of all applied models for the testing stage.

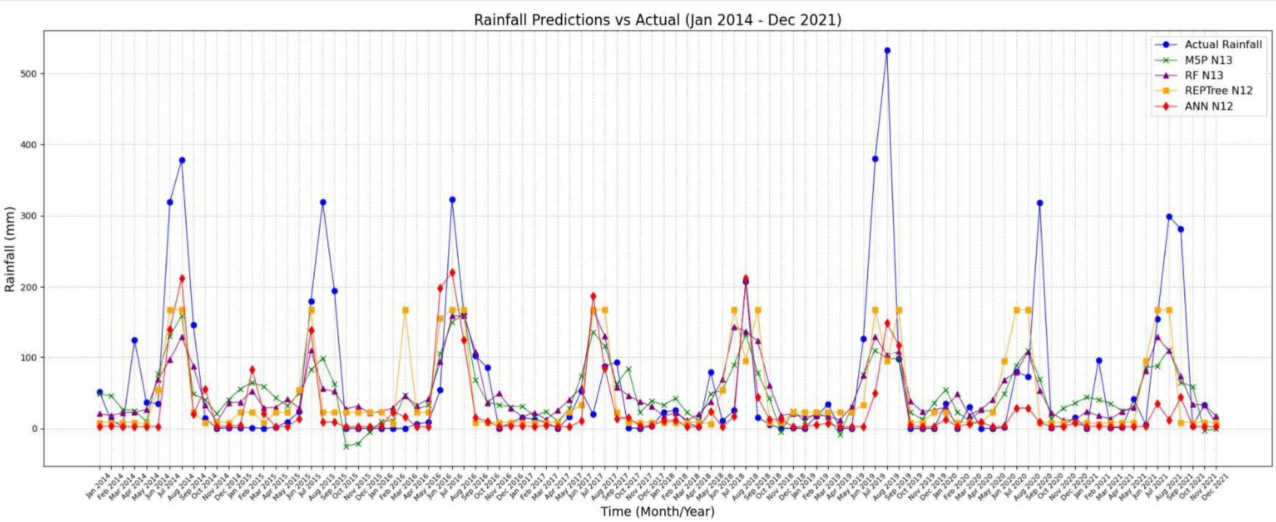


Fig. 10: Hydrograph of actual and forecasted values using all applied models for rainfall at Safdarjung Station.

of four machine learning models: RF, ANN, M5P, and REPTree, for high-accuracy rainfall forecasting based on climate indices. The findings indicated that Random Forest (RF) surpassed M5P, ANN, and REPTree in predicting accuracy and generalization. Random Forest consistently demonstrated better values of R, MAE, RMSE, and SI in both training and testing datasets. The model's strong performance across many configurations, particularly in reducing error measures, underscores its appropriateness for intricate, high-precision forecasting jobs. The RF N13

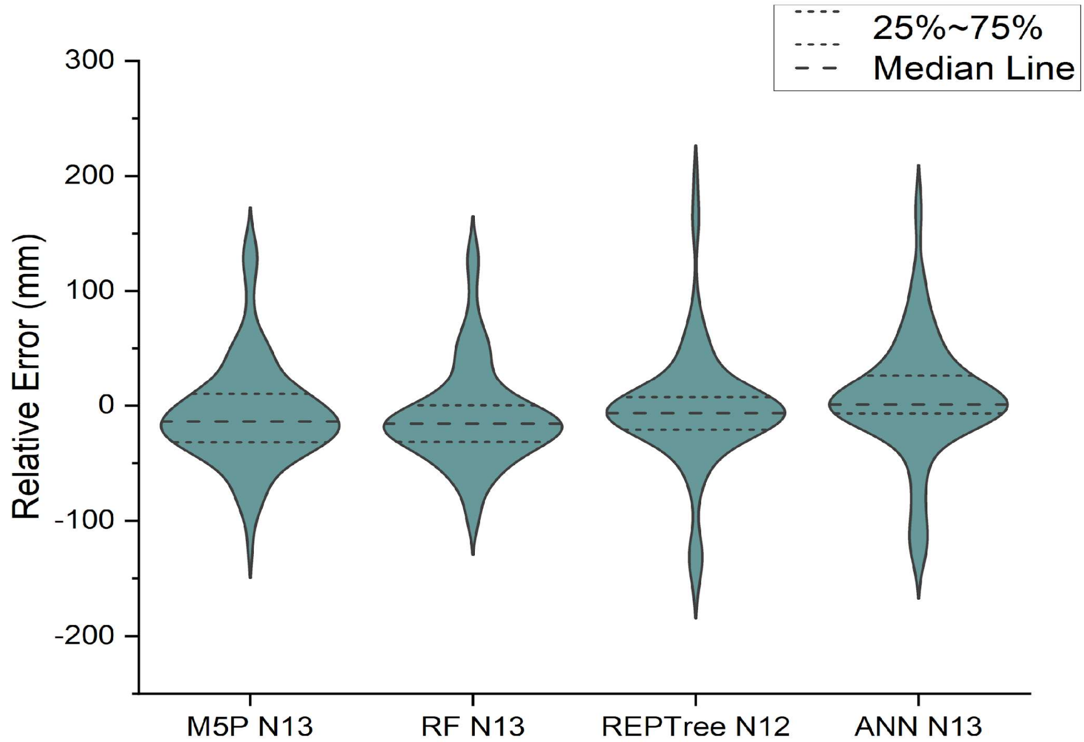


Fig. 11: Violin plots of residual value comparing the M5P N13, RF N13, REPTree N 12, and ANN N13 models.

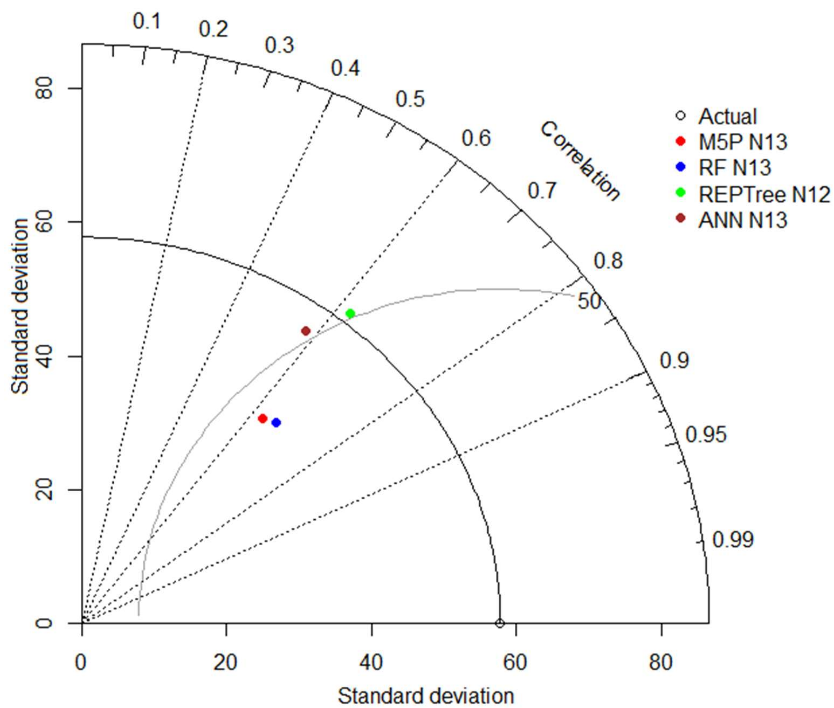


Fig. 12: Taylor plots for comparing the M5P N13, RF N13, REPTree N 12, and ANN N13model performance for the prediction of rainfall at Safdarjung, Delhi, India.

model for predicting rainfall at Safdarjung, Delhi, India, is the most effective model among those tested, exhibiting an R^2 of 0.6674, MAE of 33.7026, RMSE of 44.0744, and SI of 1.0498 during the testing phase. Conversely, the ANN N13 model is the lowest among all models used, exhibiting r = 0.5758, MAE = 32.6226, RMSE =51.7643, and SI =1.2329 throughout the testing phase. Future research may investigate the optimization of Random Forest and other sophisticated ensemble techniques to enhance the predictive accuracy for precipitation and other meteorological variables.

6. REFERENCES

Abbot, J. and Marohasy, J., 2014. Input selection and optimisation for monthly rainfall forecasting in Queensland, Australia, using artificial neural networks. *Atmospheric Research*, 138, pp.166-178. [DOI]

Abebe, W.T. and Endalie, D., 2023. Artificial intelligence models for prediction of monthly rainfall without climatic data for meteorological stations in Ethiopia. *Journal of Big Data*, 10, p.2. [DOI]

Asakereh, H., 2020. Decadal variation in precipitation regime in northwest of Iran. *Theoretical and Applied Climatology*, 139, pp.461-471. [DOI]

Asim, M., Rashid, A. and Ahmad, T., 2022. Scour modeling using deep neural networks based on hyperparameter optimization. *ICT Express*, 8(3), pp.357-362. [DOI]

Breiman, L., 1999. Random Forests-Random Features. Technical Report No. 567. University of California, Berkeley, pp.1-25.

Choubin, B., Zehtabian, G., Azareh, A., Rafiei-Sardooi, E., Sajedi-Hosseini, F. and Kişi, Ö., 2018. Precipitation forecasting using classification and regression trees (CART) model: a comparative study of different approaches. *Environmental Earth Sciences*, 77, p.314. [DOI]

Dhillon, M.S., Sharif, M., Madsen, H. and Jakobsen, F., 2023. Seasonal precipitation forecasting for water management in the Kosi Basin, India using large-scale climate predictors. *Journal of Water and Climate Change*, 14(6), pp.1868-1880. [DOI]

Enfield, D.B., Mestas-Nuñez, A.M. and Trimble, P.J., 2001. The Atlantic multidecadal oscillation and its relation to rainfall and river flows in the continental US. *Geophysical Research Letters*, 28(10), pp.2077-2080. [DOI]

Feng, P., Wang, B., Liu, D.L., Ji, F., Niu, X., Ruan, H. and Yu, Q., 2020. Machine learning-based integration of large-scale climate drivers can improve the forecast of seasonal rainfall probability in Australia. *Environmental Research Letters*, 15(8), p.084051. [DOI]

Haykin, S., 2009. Neural Networks and Learning Machines. 3rd ed. Pearson Education India, pp.1-950.

Hurrell, J.W., 1995. Decadal trends in the North Atlantic Oscillation: regional temperatures and precipitation. *Science*, 269(5224), pp.676-679. [DOI]

Kashid, S.S. and Maity, R., 2012. Prediction of monthly rainfall on homogeneous monsoon regions of India based on large-scale circulation patterns using genetic programming. *Journal of Hydrology*, 454, pp.26-41. [DOI]

Khan, M.I. and Maity, R., 2020. Hybrid deep learning approach for multi-step-ahead daily rainfall prediction using GCM simulations. *IEEE Access*, 8, pp.52774-52784. [DOI]

Luo, J.J., Liu, G., Hendon, H., Alves, O. and Yamagata, T., 2017. Inter-basin sources for two-year predictability of the multi-year La Niña event in 2010–2012. *Scientific Reports*, 7, p.2276. [DOI]

Mantua, N.J., Hare, S.R., Zhang, Y., Wallace, J.M. and Francis, R.C., 1997. A Pacific interdecadal climate oscillation with impacts on salmon production. *Bulletin of the American Meteorological Society*, 78(6), pp.1069-1080. [DOI]

Mokhtarzad, M., Eskandari, F., Jamshidi Vanjani, N. and Arabasadi, A., 2017. Drought forecasting by ANN, ANFIS, and SVM and comparison of the models. *Environmental Earth Sciences*, 76(21), p.729. [DOI]

Naved, M., Asim, M. and Ahmad, T., 2024. Prediction of concrete compressive strength using deep neural networks based on hyperparameter optimization. *Cogent Engineering*, 11(1), p.2297491. [DOI]

Pai, D.S., Rajeevan, M., Sreejith, O.P., Mukhopadhyay, B. and Satbha, N.S., 2014. Development of a new high spatial resolution (0.25° × 0.25°) long period (1901-2010) daily gridded rainfall data set over India and its comparison with existing data sets over the region. *Mausam*, 65(1), pp.1-18. [DOI]

Pandhiani, S.M., Sihag, P., Shabri, A.B., Singh, B. and Pham, Q.B., 2020. Time-series prediction of streamflows of Malaysian rivers using data-driven techniques. *Journal of Irrigation and Drainage Engineering*, 146(7), p.04020013. [DOI]

Quinlan, J.R., 1992. Learning with continuous classes. In: Proceedings of the 5th Australian Joint Conference on Artificial Intelligence, Hobart, Australia, Vol. 92, pp.343-348.

Saji, N.H., Goswami, B.N., Vinayachandran, P.N. and Yamagata, T., 1999. A dipole mode in the tropical Indian Ocean. *Nature*, 401(6751), pp.360-363. [DOI]

Sham, F.A.F., El-Shafie, A., Jaafar, W.Z.W., Sherif, M. and Ahmed, A.N., 2025. Advances in AI-based rainfall forecasting: a comprehensive review of past, present, and future directions with intelligent data fusion and climate change models. *Results in Engineering*, 27, p.105774. [DOI]

Shankar, A., 2024. Dynamic impact-based heavy rainfall warning with multi-classification machine learning approaches. *Nature Environment & Pollution Technology*, 23(4), pp.2205-2214. [DOI]

Sihag, P., 2018. Prediction of unsaturated hydraulic conductivity using fuzzy logic and artificial neural network. *Modeling Earth Systems and Environment*, 4(1), pp.189-198. [DOI]

Trenberth, K.E. and Hoar, T.J., 1997. El Niño and climate change. *Geophysical Research Letters*, 24(23), pp.3057-3060. [DOI]

Wani, O.A., Mahdi, S.S., Yeasin, M., Kumar, S.S., Gagnon, A.S., Danish, F., Al-Ansari, N., El Hendawy, S. and Mattar, M.A., 2024. Predicting rainfall using machine learning, deep learning, and time series models across an altitudinal gradient in the North-Western Himalayas. *Scientific Reports*, 14, p.27876. [DOI]