



A Comparative Study of Machine Learning Models for Daily and Weekly Rainfall Forecasting

Vijendra Kumar¹ · Naresh Kedam² · Ozgur Kisi^{3,4} · Saleh Alsulamy⁵ · Khaled Mohamed Khedher⁶ · Mohamed Abdelaziz Salem⁷

Received: 6 June 2024 / Accepted: 29 August 2024
© The Author(s) 2024

Abstract

Accurate rainfall forecasting is crucial for various sectors across diverse geographical regions, including Uttarakhand, Uttar Pradesh, Haryana, Punjab, Himachal Pradesh, Madhya Pradesh, Rajasthan, and the Union Territory of Delhi. This study addresses the need for precise rainfall predictions by bridging the gap between localized meteorological data and broader regional influences. It explores how rainfall patterns in neighboring states affect Delhi's precipitation, aiming to improve forecasting accuracy. Historical rainfall data from neighboring states over four decades (1980–2021) were collected and analyzed. The study employs a dual-model approach: a daily model for immediate rainfall triggers and a weekly model for longer-term trends. Several machine learning algorithms, including CatBoost, XGBoost, ElasticNet, Lasso, LGBM, Random Forest, Multilayer Perceptron, Ridge, Stochastic Gradient Descent, and Linear Regression, were used in the modeling process. These models were rigorously assessed based on performance metrics from training, validation, and testing datasets. For daily rainfall forecasting, CatBoost, XGBoost, and Random Forest emerged as top performers, showcasing exceptional accuracy and pattern-capturing capabilities. In weekly rainfall forecasting, XGBoost consistently achieved near-perfect accuracy with an R^2 value of 0.99, with Random Forest and CatBoost also demonstrating strong performance. The study provides valuable insights into how climate patterns in neighboring states influence Delhi's weather, leading to more reliable and timely rainfall predictions.

Keywords Rainfall forecasting · Water resource management · Precipitation patterns · Machine Learning · CatBoost · XGBoost

1 Introduction

Accurate rainfall forecasting is crucial for applications such as disaster preparedness, agricultural planning, water resource management, and urban development. However, predicting future rainfall patterns remains challenging due to the dynamic and complex nature of weather systems (Chen et al. 2023). Identifying underlying patterns and linkages in historical rainfall data is difficult, as it requires models capable of integrating diverse climatic

factors and addressing the intrinsic variability and unpredictability of weather systems (Zhao et al. 2024). Two essential techniques in rainfall forecasting have emerged: conceptual modeling and system theoretical modeling. Conceptual modeling is widely used in hydrological forecasting due to its focus on understanding physical principles governing hydrological processes. This approach relies on specific catchment characteristics but faces challenges in rainfall prediction, such as data calibration difficulties and computational demands (Xu et al. 2022). In contrast, system theoretical methods, like the ARMAX model, focus on mapping input–output relationships without detailed physical insights. While useful in time series forecasting, ARMAX models struggle with predicting nonlinear rainfall variations (Zhao et al. 2023).

1.1 Literature Review

Traditional forecasting methods often fall short, particularly in regions like Delhi, where rainfall patterns are influenced by multiple neighboring states. These methods struggle with integrating complex inter-regional climatic data and tend to rely on simplistic, linear models that cannot adequately capture both rapid daily changes and longer-term weekly trends. Existing models also face difficulties in managing noisy data and distinguishing significant climatic signals from transient anomalies (Xie et al. 2021). An alternative approach that has gained prominence in improving rainfall forecasting is the use of machine learning algorithms. Machine learning provides a data-driven method that can complement traditional forecasting techniques by uncovering complex patterns and relationships within rainfall data (Truong et al. 2023). Techniques such as Support Vector Machines (SVM), Artificial Neural Networks (ANN), Gradient Boosting, and Random Forests are examples of algorithms that can be trained on historical rainfall data to detect hidden patterns and trends. These methods excel at capturing nonlinear relationships and adapting to dynamic conditions, which is particularly beneficial for rainfall forecasting, where abrupt shifts and complex interactions are common. By inputting historical meteorological data into these algorithms, machine learning models can learn to identify correlations between various factors (such as temperature, humidity, and wind speed) and rainfall (Zhou et al. 2021). Once trained, these models can predict future rainfall based on real-time or forecasted meteorological data. A key advantage of machine learning is its capacity to simultaneously consider numerous variables and their interactions, leading to more precise predictions (Yin et al. 2023). Moreover, as machine learning models encounter new data, they can continually refine their predictions through ongoing learning.

Rainfall forecasting has made significant strides with the use of machine learning and data-driven methods. The unpredictable nature of rainfall poses unique challenges for accurate prediction (Kumar and Yadav 2020). Machine learning algorithms have been shown to outperform traditional methods in forecasting. For instance, (Abbot and Marohasy 2014) use ANN to forecast continuous rainfall based on climate indices, demonstrating improved accuracy. Chen et al., (2013) find that a neural network model enhances runoff prediction by managing complex spatial rainfall distributions in Taiwan. Mekanik et al., (2013) compare ANN and multiple regression for spring rainfall forecasting in Victoria, Australia, showing that ANN provides better generalization. Yu et al., (2017) compare Random Forests (RF) and SVM for real-time radar-derived rainfall forecasting, noting that single-mode RF and SVM models perform better than their multiple-mode counterparts for 1-h ahead predictions. Feng et al., (2015) introduce a wavelet analysis-support vector machine (WA-SVM) model for arid regions, which forecasts monthly rainfall accurately over various

lead times. This approach is especially useful for urban areas needing short-term rainfall forecasts.

Abbasi et al., (2021) developed a hybrid model that integrates RF, Deep Auto-Encoder, and support vector regression (SVR) methods for streamflow prediction, improving long-term accuracy and reducing uncertainty. Tan et al., (2021) introduced a novel method combining RF and inverse distance weighting (RF-IDW) for generating precipitation and temperature climate surfaces, which improves accuracy by accounting for spatial complexities and environmental factors. Rahman et al., (2022) proposed a real-time rainfall prediction system for smart cities, integrating fuzzy logic with four supervised machine learning techniques. Using historical weather data, this fusion-based framework outperforms other models. Diez-Sierra and Del-Jesus, (2020) evaluated the performance of eight statistical and machine learning methods driven by atmospheric synoptic patterns for long-term daily rainfall prediction in Tenerife, Spain. They found that neural networks excel in predicting rainfall occurrence and intensity. (Nunno et al., (2022) developed a reliable precipitation prediction model using machine learning algorithms for the northern region of Bangladesh. A hybrid model, based on M5P and SVR, performs exceptionally well for precipitation prediction. Exploring machine learning's potential, (Zhang et al. 2022) focus on spatial patterns in precipitation forecasts, investigating convolutional neural network models to improve the skill of predicting precipitation occurrence while balancing the trade-off between false positives and negatives. Additionally, studies by (Endalie et al. 2022) demonstrate the application of various machine learning techniques, to enhance rainfall prediction accuracy. Collectively, these studies underscore the growing role of machine learning in addressing the complexities of rainfall prediction across different contexts, highlighting its potential for improved accuracy and practical application.

The novelty of this study lies in its innovative approach of integrating meteorological data from multiple states to enhance rainfall forecasting accuracy for a specific target region, Delhi, India. By utilizing diverse climatic data inputs from Uttarakhand, Haryana, Punjab, Uttar Pradesh, Himachal Pradesh, Madhya Pradesh, and Rajasthan States, this research pioneers a cross-regional data integration methodology. This approach not only improves predictive performance but also provides a robust framework for regional weather forecasting, demonstrating the practical application of advanced machine learning techniques in enhancing meteorological predictions. The study also explores and compares the performance of various machine learning models CatBoost, ElasticNet, Multilayer Perceptron (MLP), Lasso, Random Forest (RF), LGBM, Linear Regression (LR), Ridge, XGBoost, and Stochastic Gradient Descent (SGD), in capturing intricate patterns and delivering high accuracy in both daily and weekly forecasts. This cross-state data integration strategy represents a significant advancement in the field of rainfall prediction, offering a novel perspective on leveraging regional climatic variations for more accurate and reliable forecasts.

1.2 Objective of the study

The primary objective of this study is to enhance the accuracy of rainfall forecasting for Delhi by leveraging historical rainfall data from neighboring states, including Uttarakhand, Uttar Pradesh, Haryana, Punjab, Himachal Pradesh, Madhya Pradesh, and Rajasthan. The research examines how the climate and geography of these surrounding regions influence the rainfall patterns in Delhi. To achieve this, the study employs a range of algorithms, including CatBoost, ElasticNet, MLP, Lasso, RF, LGBM, LR, Ridge, XGBoost, and SGD.

2 Study Area and Data Preparation

2.1 Study Area

The research region encompasses Uttarakhand, Uttar Pradesh, Haryana, Punjab, Himachal Pradesh, Madhya Pradesh, Rajasthan, and the Union Territory of Delhi. This area features a diverse range of geographic and climatic conditions, including the Himalayan foothills, the Gangetic plains, deserts, and the metropolitan agglomeration of Delhi. Due to this geographical diversity, the region experiences a wide range of climatic conditions. Uttarakhand and Himachal Pradesh have a moderate climate with robust summer monsoons. The Gangetic plains, covering Uttar Pradesh, Haryana, and Punjab, experience a subtropical climate characterized by hot summers and milder winters. Rajasthan is largely desert or semi-arid, with extreme temperatures and scant rainfall. Delhi, part of the National Capital Territory, has a subtropical climate with additional urban heat island effects. Rainfall in the region is significantly influenced by the Indian Monsoon, which provides most of the annual precipitation. The monsoon season generally spans from June to September, with varying start and end dates across the region. The Himalayan foothills and parts of Himachal Pradesh receive substantial rainfall due to orographic lifting, while Rajasthan's arid areas experience minimal monsoon precipitation. The variability in rainfall patterns results from the intricate interaction between monsoon winds, topography, and local climatic factors. Figure 1(a) illustrates the location of the research area.

Accurate rainfall forecasting is essential for various sectors in this region. Since a significant portion of the population relies on agriculture, timely and precise rainfall predictions are crucial for optimizing irrigation and crop management. Reliable forecasts also support reservoir management, water allocation, and flood control decisions. Given the region's vulnerability to both droughts and floods, accurate rainfall predictions enable proactive disaster management measures. In urban areas like Delhi, forecasts play a key role in planning drainage and flood control infrastructure. The unique aspect of this study lies in its cross-state analysis aimed at improving rainfall forecasting for Delhi. By utilizing historical rainfall data from neighboring states, the research seeks to enhance the accuracy of predictions for Delhi. This approach recognizes the interconnections between the precipitation patterns of these states, driven by the monsoon system and regional climatic influences. The study aims to bridge the gap between localized meteorological data and broader regional factors, investigating how the rainfall patterns of Uttarakhand, Uttar Pradesh, Haryana, Punjab, Himachal Pradesh, Madhya Pradesh, and Rajasthan affect Delhi. Additionally, it explores how these interconnections can be leveraged to refine the accuracy of rainfall forecasting for Delhi.

2.2 Data Collection

The approach for collecting data for this comprehensive research region, which includes Delhi and the states of Uttarakhand, Uttar Pradesh, Haryana, Punjab, Himachal Pradesh, Madhya Pradesh, and Rajasthan, involves gathering daily rainfall data. This data was sourced from WRIS (India's Water Resources Information System), a comprehensive database managed by the Ministry of Jal Shakti of the Indian Government. WRIS provides a repository of hydrological and meteorological information, offering access to essential

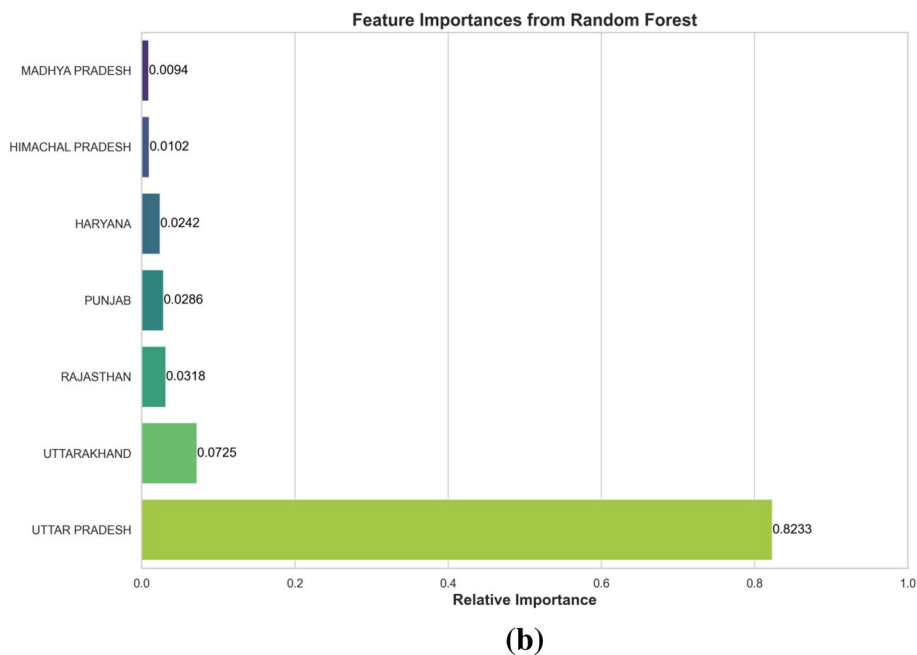
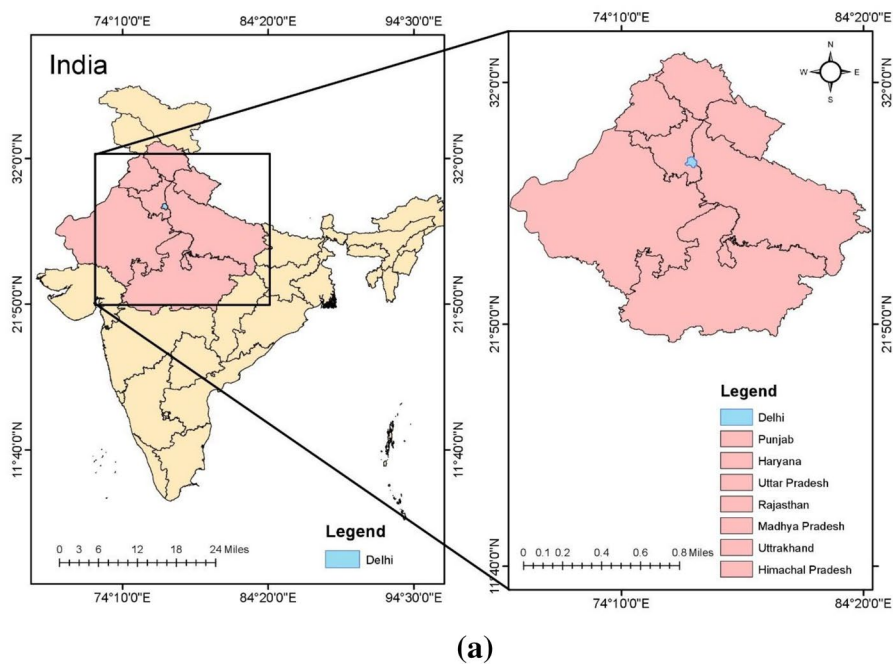
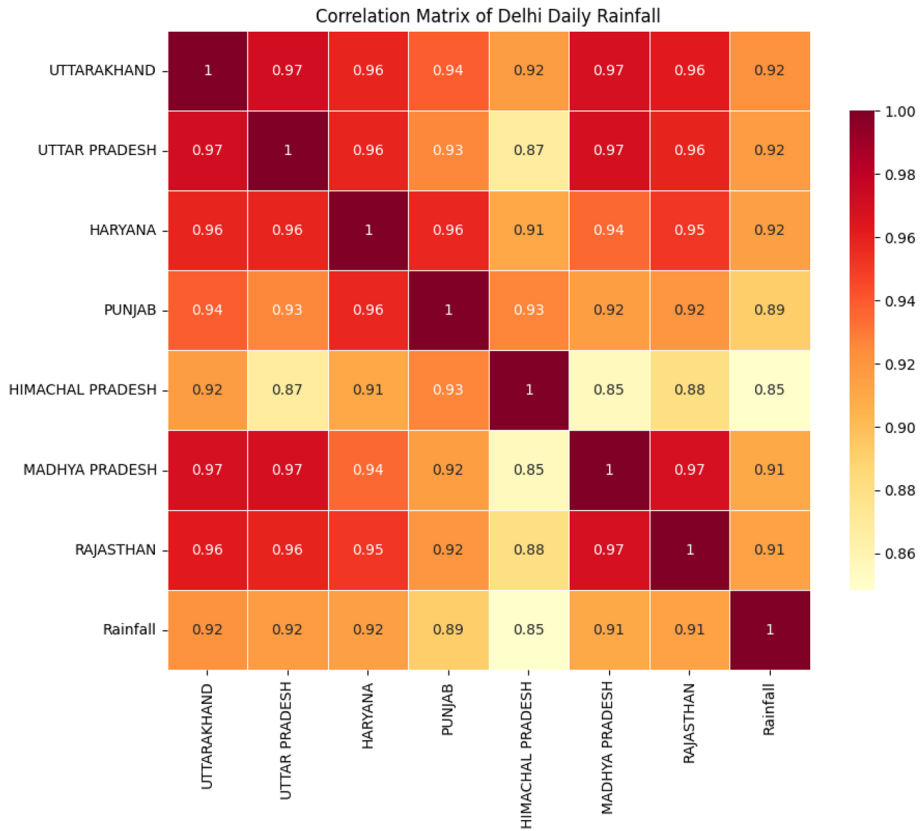


Fig. 1 (a) Study area map. (b) Feature importance results using random forest. (c) Presents the correlation matrix for daily rainfall in Delhi



(c)

Fig. 1 (continued)

historical rainfall observations. The dataset includes daily rainfall records spanning four decades, from 1980 to 2021.

2.3 Data Preprocessing

Temporal adjustments were crucial in processing the dataset. The approach involved using rainfall metrics from neighboring states from the previous day to predict Delhi's rainfall for the current day. This shift is based on the meteorological insight that weather patterns in adjacent areas can influence conditions in a central location like Delhi. For example, rainfall in Uttar Pradesh today might indicate potential rainfall in Delhi tomorrow. This alignment is essential for accurate predictive modeling. Given the complexity of meteorological events, the study employed a dual model approach. One model was designed to capture the day-to-day variations in rainfall, while the other focused on longer, weekly trends. This bifurcation provided a comprehensive perspective, allowing for precise detection of immediate triggers and broader, week-long patterns.

2.4 Feature and Correlation Analysis

In developing the machine learning model for rainfall prediction, various features were incorporated. The RandomForestRegressor from the `sklearn.ensemble` library was employed to train the model using seven features: 'Uttarakhand', 'Uttar Pradesh', 'Haryana', 'Punjab', 'Himachal Pradesh', 'Madhya Pradesh', and 'Rajasthan'. Feature importance was assessed using the trained Random Forest model, where the importance of each feature is determined by the total reduction in the criterion attributed to that feature, also known as Gini importance. As shown in Fig. 1(b), 'Uttar Pradesh' had the highest importance score of 0.823254, indicating its substantial impact on the model's predictions. This was followed by 'Uttarakhand' with a score of 0.072523, 'Rajasthan' at 0.031806, 'Punjab' at 0.028555, 'Haryana' at 0.024239, 'Himachal Pradesh' at 0.010216, and 'Madhya Pradesh' at 0.009406. These results highlight that rainfall in 'Uttar Pradesh' plays a significant role in the model's predictive accuracy.

A correlation matrix was also computed to examine the relationships between the different features. This matrix provides correlation coefficients that measure the linear relationship between pairs of features, with values ranging from -1 to 1. A value closer to 1 indicates a strong positive correlation, while a value closer to -1 signifies a strong negative correlation. As illustrated in Fig. 1(c), the correlation matrix reveals that all features exhibit a strong positive correlation. This suggests that similar weather patterns influence rainfall across these regions. Analyzing the features in this manner offers valuable insights into their importance and relationships, which is essential for understanding the model's predictions and improving its performance.

3 Methodology

Rainfall forecasting using machine learning follows a structured process. It begins with the collection of historical weather data, including rainfall observations. The data is then preprocessed to ensure quality and consistency by addressing missing values, outliers, and normalization. Key climatic factors are identified through feature selection and engineering. The dataset is divided into training, validation, and testing sets. The model is trained on the training set and its performance is evaluated on the validation dataset. Once validated, the model predicts future rainfall for different time frames. It is subsequently deployed for operational use, with ongoing monitoring to ensure accuracy as weather patterns evolve. Figure 2 illustrates the flowchart of the approach used in this study for the daily model, with a similar flowchart being developed for the weekly model.

3.1 Model Preparation

The study employed time series segmentation to partition the data into training, validation, and test sets, reflecting the importance of maintaining temporal order in time series data. This method segments the data across chronological intervals: the training set includes the earliest data, the validation set consists of sequential data, and the test set covers the most recent observations. Unlike random splits, this approach preserves chronological accuracy and the inherent temporal dependencies of the data. By using time series splitting, the study maintained the temporal structure of the data, ensuring that future predictions are based on a

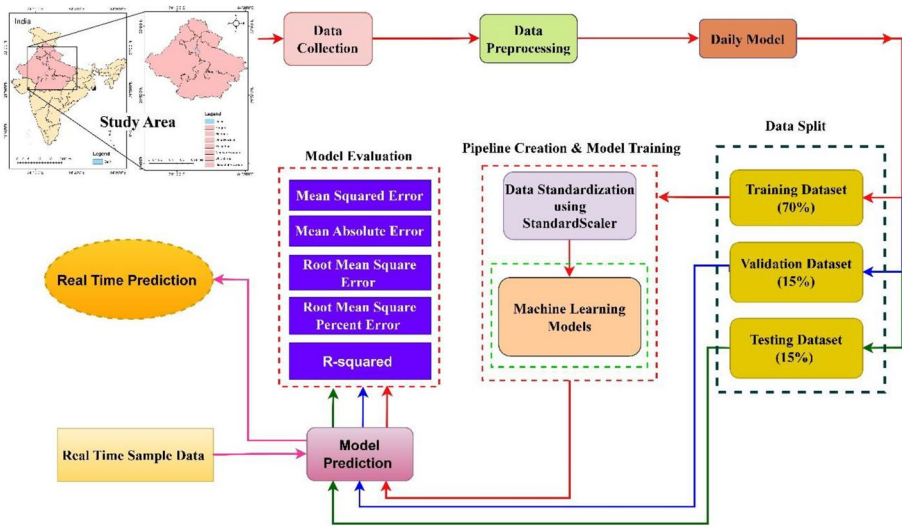


Fig. 2 Illustrates the methodology flowchart for the daily forecasting model

realistic sequence of events. Based on this concept, two models are developed, as presented in Table SI-18 (supplementary information), a Daily Model and a Weekly Model. Both models use data from Uttarakhand, Haryana, Punjab, Uttar Pradesh, Himachal Pradesh, Madhya Pradesh, and Rajasthan as inputs to predict rainfall for Delhi. The Daily Rainfall Model must quickly detect significant shifts, such as predicting rain after a sunny day, while maintaining robustness against brief anomalies and focusing on daily trends. The Weekly Rainfall Model analyzes seven-day patterns to identify trends like increasing rainfall or alternating dry and wet days, aiming to capture broader trends despite occasional data anomalies. Both models must balance sensitivity to temporal changes with resilience to noise in meteorological data.

3.2 Pipelines Creation and Model Training

In machine learning, feature scaling is crucial for maintaining the effectiveness of models, particularly those that are sensitive to the magnitudes of data points or rely on gradient descent optimization techniques, such as K-Nearest Neighbors (KNN) and Neural Networks. To address the issue of diverse units and scales in the dataset, normalization is essential. For this purpose, the Scikit-learn StandardScaler was employed. This scaler standardizes features by adjusting the mean to zero and scaling each feature to have unit variance. This process ensures that no single feature disproportionately influences the model, leading to more balanced and accurate predictions. Scikit-learn's pipelines facilitated this process by integrating preprocessing and modeling steps, simplifying the workflow and ensuring consistent preprocessing across datasets, which enhances reproducibility.

3.3 CatBoost Regressor Algorithm

CatBoost, like other gradient boosting algorithms, builds a strong predictive model by combining multiple weak learners, typically decision trees (Karbasi et al. 2022). The final

prediction is generated by aggregating the predictions from these individual trees, each weighted according to their importance. The overall process can be described as follows:

For a binary classification problem, the formula for CatBoost's prediction can be expressed as:

$$F(x) = \sigma \left(\sum_{m=1}^M \alpha_m \cdot f_m(x) \right) \quad (1)$$

where, $F(x)$ represents the estimated probability of the positive class (class 1) for input M is the number of trees in the ensemble. The prediction of the m_{th} decision tree is $f_m(x)$. The weight associated with the m_{th} tree is denoted by α_m . σ is the sigmoid function that converts the total of tree predictions into a probability between 0 and 1.

CatBoost employs several techniques to effectively manage categorical data and enhance the boosting process. It utilizes ordered boosting, which adjusts the learning rate for each tree according to its depth and sequence position. Additionally, CatBoost uses a method that optimizes the order of categorical feature levels during splits to minimize overfitting. While the provided formula offers a high-level overview of CatBoost's functionality, the actual implementation includes additional complexities and refinements that further boost its performance and accuracy.

Other methodologies, including Lasso, Ridge, ElasticNet Regression, Light Gradient-Boosting Machine Regressor (LGBM), Linear Regression (LR), Multilayer Perceptron (MLP), Random Forest (RF), Stochastic Gradient Descent (SGD), and XGBoost, have been added to the supplementary information as SI-1 to SI-9.

3.4 Model Evaluation

Several assessment metrics are employed to compare against the actual observations as follows:

Mean squared error (MSE) is calculated as the average of the squared differences between actual (y_i) and predicted (\hat{y}_i) values:

$$MSE = (1/n) * \sum [(y_i - \hat{y}_i)^2] \quad (2)$$

Mean absolute error (MAE) is the average of the absolute differences between actual and predicted values:

$$MAE = (1/n) * \sum |y_i - \hat{y}_i| \quad (3)$$

Root means square error (RMSE) is the square root of MSE:

$$RMSE = \sqrt{MSE} = \sqrt{\left[\left(\frac{1}{n} \right) * \sum [(y_i - \hat{y}_i)^2] \right]} \quad (4)$$

Root means square percent error (RMSPE) calculates the root mean square percentage error:

$$RMSPE = \sqrt{[(1/n) * \sum ((y_i - \hat{y}_i)/y_i)^2]} \quad (5)$$

R-squared (R^2) is computed as 1 minus the ratio of the sum of squared differences between actual and predicted values to the sum of squared differences between actual values and the mean (\bar{y}) of the dependent variable:

$$R^2 = 1 - \left(\frac{\sum [(y_i - \hat{y}_i)^2]}{\sum [(y_i - \bar{y})^2]} \right) \quad (6)$$

Here, n represents the total number of data points, y_i is the observed value for the i_{th} data point, and \hat{y}_i is the predicted value for the i_{th} data point. The \bar{y} the mean of the dependent variable and Σ reflects the sum of squared differences across data points.

4 Results and Discussion

Table SI-19 in the supplementary information presents a detailed summary of key parameters for various machine learning models. These parameters are crucial for optimizing model performance and are adjusted based on the dataset and desired outcomes. This table illustrates the adaptability and customization options available in modern regression algorithms.

4.1 Daily Model Results and Discussion

This section presents the results and discussion of the daily forecasting model. The performance of various models is evaluated using critical assessment criteria, including MSE, MAE, RMSE, RMSPE, and R^2 . These metrics offer insights into each model's accuracy, precision, and ability to generalize.

4.1.1 Training Performance Metrics for the Daily Models

Table 1(a) summarizes the training results of different models used for daily forecasting, with the best-performing models highlighted in bold. CatBoost, XGBoost, and RF stand out, showcasing excellent accuracy with perfect R^2 values. These models effectively capture underlying data patterns, resulting in minimal errors across various metrics. LGBM also performs well, though its slightly higher RMSPE indicates some room for improvement in capturing percentage errors. The Neural Network MLP model, built with multiple layers, achieves low error scores and a high R^2 value, demonstrating its capability to understand complex relationships within the data. In contrast, the LR and Ridge models deliver moderate results, indicating that they may not effectively grasp complex relationships compared to more advanced models. ElasticNet and Lasso models exhibit relatively higher errors and lower R^2 values, suggesting that their regularization components might hinder optimal data fitting. The SGD model performs similarly to LR and Ridge but could benefit from hyperparameter fine-tuning for better results. CatBoost, XGBoost, and RF emerge as top contenders for the daily forecasting task due to their remarkable accuracy and ability to capture underlying patterns. LGBM and MLP also offer strong alternatives, especially if computational efficiency is a priority. While simpler linear models like LR, Ridge, ElasticNet, and Lasso yield satisfactory results, they may struggle with intricate data relationships. Evaluating these models on validation and test datasets is essential to ensure their generalizability and ultimately determine the most suitable model, considering both accuracy and interpretability.

Table 1 (a) Training results, (b) Validation results, and (c) Testing results of models in daily rainfall forecasting

Model	MAE (mm)	MSE (mm ²)	RMSE (mm)	RMSPE	R ²
(a) Training results of the models in daily rainfall forecasting					
CatBoost	0.0011	1.94×10^{-6}	0.0014	0.49	0.99
ElasticNet	0.86	1.73	1.31	253.38	0.81
Lasso	1.03	2.29	1.51	325.30	0.75
LGBM	0.03	0.001	0.03	13.28	0.99
LR	0.67	1.25	1.11	130.20	0.86
MLP	0.26	0.17	0.41	84.23	0.98
RF	0.0017	2.85×10^{-10}	0.0022	0.66	0.99
Ridge	0.67	1.25	1.11	130.20	0.86
SGD	0.67	1.25	1.12	131.16	0.86
XGBoost	0.0017	4.78×10^{-6}	0.0022	0.66	0.99
(b) Validation results of the models in daily rainfall forecasting					
CatBoost	0.0012	2.1×10^{-6}	0.0015	0.47	0.99
ElasticNet	0.89	1.88	1.37	246.06	0.81
Lasso	1.07	2.49	1.57	316.47	0.75
LGBM	0.03	0.001	0.03	12.89	0.99
LR	0.71	1.36	1.16	126.80	0.86
MLP	0.27	0.18	0.43	81.92	0.98
RF	0.0017	3.17×10^{-10}	0.0022	0.64	0.99
Ridge	0.71	1.36	1.16	126.81	0.86
SGD	0.71	1.36	1.16	127.72	0.86
XGBoost	0.0017	4.80×10^{-6}	0.0022	0.64	0.99
(c) Testing results of the models in daily rainfall forecasting					
CatBoost	0.0011	1.91×10^{-6}	0.0014	0.486	0.99
ElasticNet	0.85	1.70	1.30	252.78	0.81
Lasso	1.02	2.25	1.50	327.44	0.75
LGBM	0.03	0.0016	0.03	13.21	0.99
LR	0.67	1.23	1.11	128.63	0.86
MLP	0.26	0.17	0.41	84.38	0.98
RF	0.0017	2.79×10^{-10}	0.0022	0.6583	0.99
Ridge	0.67	1.23	1.11	128.63	0.86
SGD	0.67	1.23	1.11	129.60	0.86
XGBoost	0.0017	4.73×10^{-6}	0.0022	0.6583	0.99

Bold value shows the better solution

4.1.2 Validation Performance Metrics for Daily Models

Table 1(b) presents the validation results for various models used in the daily forecasting task. Comparing these results with the training dataset in Table 1(a) reveals several key observations. CatBoost maintains its exceptional performance on the validation dataset, reaffirming its strong generalization ability. It achieves a perfect R² score and minimal errors across all metrics, consistently capturing the data's underlying patterns with high accuracy. ElasticNet, Lasso, LR, Ridge, and SGD show similar trends in both datasets, with

only marginal differences in their error scores and R^2 values. While these models may not achieve the same level of accuracy as more complex ones, their consistency across datasets suggests they capture core trends reasonably well. LGBM and MLP, which demonstrated strong performance in Table 1(a), continue to excel on the validation dataset. Their low error scores and high R^2 values highlight their robustness and ability to generalize beyond the training data. XGBoost and RF maintain their accuracy, mirroring their training dataset performance. These models exhibit a perfect R^2 score, indicating their capacity to consistently capture and reproduce the data's patterns. The validation results reaffirm the strengths identified in the training dataset. CatBoost, XGBoost, and RF consistently deliver remarkable accuracy across both datasets. LGBM and MLP also maintain strong performance, indicating their reliability and potential for practical application.

4.1.3 Test Performance Metrics for the Daily Models

Table 1(c) provides an overview of the testing results for various models employed in the daily forecasting task. Comparing these testing results with the training and validation datasets in Tables 1(a) and 1(b), several trends can be discerned. CatBoost demonstrates exceptional consistency, maintaining a perfect R^2 value and exhibiting minimal errors across all metrics, underscoring its robust generalizability and ability to consistently capture underlying patterns. ElasticNet, Lasso, LR, Ridge, and SGD continue to show stable performance across the datasets, with slight variations in their error scores and R^2 values. LGBM and MLP, which exhibited strong results in the training and validation datasets, retain their robust performance on the test dataset, with low error scores and high R^2 values, suggesting their capacity to generalize beyond the training data. XGBoost and RF again demonstrate impressive accuracy on the test dataset, reflected by perfect R^2 values. These results highlight the dependability and generalizability of several models, namely CatBoost, XGBoost, RF, LGBM, and MLP.

The best model, which predicts rainfall using CatBoost, RF, and LGBM for daily data, is displayed in Figures SI-10 to SI-12 of the supplementary information. The worst model, however, is depicted in Figure SI-13, which uses Lasso for daily data to predict rainfall. Figure SI-10 shows how well the CatBoost model matches actual rainfall data with its predictions, demonstrating striking accuracy. Similarly, the RF model performs well, successfully capturing daily fluctuations, as shown in Figure SI-11. The LGBM model also precisely identifies rainfall peaks and troughs, maintaining close agreement with actual data (Figure SI-12). Conversely, the Lasso model has significant difficulties, exhibiting noticeable discrepancies from actual rainfall data, as illustrated in Figure SI-13.

4.1.4 Comparison of Predicted Error Plots for Daily Models

To gain a deeper understanding of the performance of various machine learning models (CatBoost, ElasticNet, Lasso, LGBM, LR, MLP, RF, Ridge, SGD, and XGBoost) in predicting rainfall, error scatter plots shown in Fig. 3(a)-(j) were analyzed to illustrate the relationship between predicted and actual rainfall values. The figures reveal that CatBoost and RF exhibit data point errors of less than 0 to 2% for rainfall predictions, as shown in Fig. 3(a) and (g). XGBoost demonstrates similar data point errors, ranging from 2 to 5%, as illustrated in Fig. 3(j). In contrast, LGBM displays a broader range of data point errors, spanning from 2 to 15%, as shown in Fig. 3(c). ElasticNet, Lasso, LR, MLP, Ridge, and SGD consistently show higher percentages of data point errors,

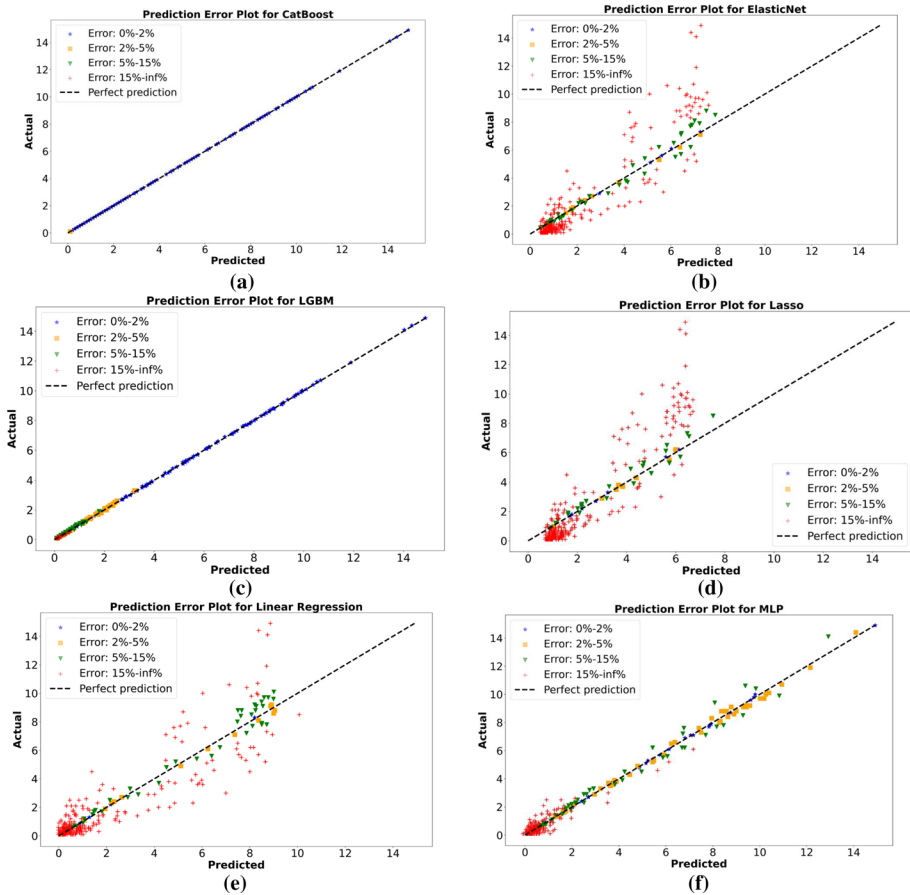


Fig. 3 (a) Scatter plots for the CatBoost daily model. (b) Scatter plots for the ElasticNet daily model. (c) Scatter plots for the LGBM daily model. (d) Scatter plots for the Lasso daily model. (e) Scatter plots for the LR daily model. (f) Scatter plots for the MLP daily model. (g) Scatter plots for the RF daily model. (h) Scatter plots for the Ridge daily model. (i) Scatter plots for the SGD daily model. (j) Scatter plots for the XGBoost daily model

exceeding 15%, as depicted in Fig. 3(b), (d), (e), (f), (h), and (i), respectively. Based on these findings, it is reasonable to conclude that CatBoost and RF models outperform the others in terms of robustness and reliability for predicting daily rainfall.

4.2 Results and Discussion for Weekly Models

This section presents the results and discusses the weekly forecasting models. The performance of various models is evaluated with respect to MSE, MAE, RMSE, RMSPE, and R^2 .

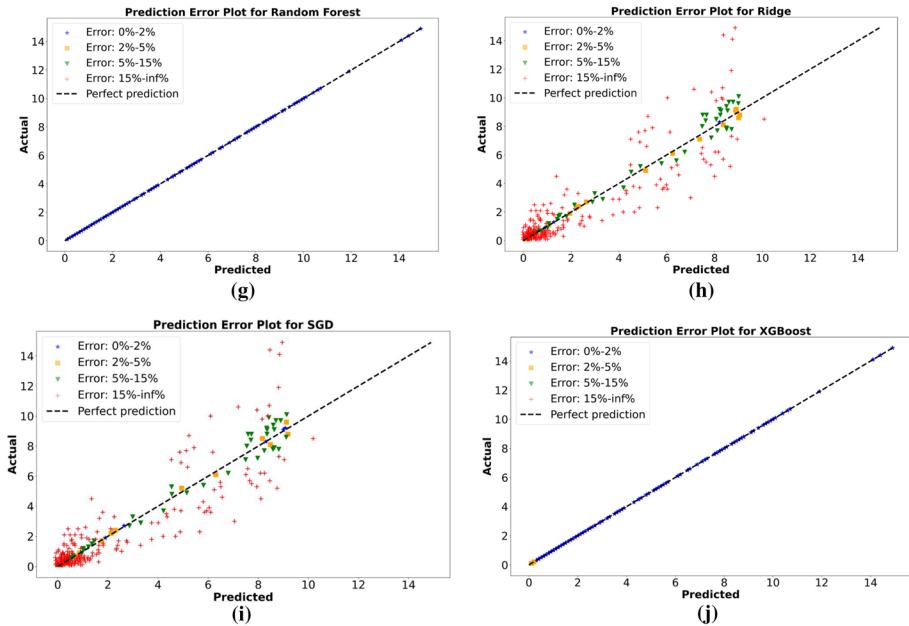


Fig. 3 (continued)

4.2.1 Training Performance Metrics for the Weekly Models

Table 2(a) presents the training results for various models used in a weekly forecasting. The analysis reveals the following insights into model performance: XGBoost emerges as the most accurate, exhibiting minimal error scores across all metrics and an impressive R^2 value of 0.99. This indicates that XGBoost effectively captures underlying patterns within the data, leading to highly accurate forecasts. RF and CatBoost also show near-perfect accuracy with R^2 values of 0.99, benefiting from ensemble approaches that leverage multiple decision trees to achieve accurate forecasts and strong performance across all metrics. ElasticNet, Lasso, LR, Ridge, and SGD deliver reasonably good performance with moderate error scores and R^2 values ranging from 0.92 to 0.96. LGBM and MLP further demonstrate strong performance, with LGBM achieving a high R^2 value of 0.99 and MLP a value of 0.98, showcasing their proficiency in capturing complex relationships and modeling weekly trends. The top performers, based on consistent and accurate forecasts, are XGBoost, RF, CatBoost, LGBM, and MLP. To ensure their generalizability and suitability for real-world applications, further validation on independent validation and test datasets are considered.

4.2.2 Validation Performance Metrics for the Weekly Models

Table 2(b) provides a comprehensive overview of the validation results for various models used in the weekly forecasting task. The insights gained from these results allow for a thorough assessment of each model's performance. XGBoost continues to stand out as the most

Table 2 (a) Training results, (b) Validation results, and (c) Testing results of models in weekly rainfall forecasting

Model	MAE (mm)	MSE (mm ²)	RMSE (mm)	RMSPE	R ²
(a) Training results of the models in weekly rainfall forecasting					
CatBoost	0.055	0.0067	0.082	2.92	0.99
ElasticNet	3.56	29.42	5.42	108.02	0.92
Lasso	3.06	23.23	4.81	87.39	0.94
LGBM	0.12	0.031	0.17	10.77	0.99
LR	2.81	15.15	3.89	80.99	0.96
MLP	1.60	5.09	2.25	83.63	0.98
RF	0.026	0.0049	0.070	1.36	0.99
Ridge	2.77	15.23	3.90	78.54	0.96
SGD	2.78	16.57	4.07	74.28	0.95
XGBoost	0.014	0.0026	0.051	1.21	0.99
(b) Validation results of the models in weekly rainfall forecasting					
CatBoost	0.055	0.0063	0.079	2.44	0.99
ElasticNet	3.73	30.82	5.55	106.07	0.92
Lasso	3.20	24.07	4.90	85.47	0.94
LGBM	0.12	0.034	0.18	10.75	0.99
LR	2.88	15.49	3.93	80.09	0.96
MLP	1.62	5.112	2.26	82.95	0.98
RF	0.026	0.0042	0.065	1.13	0.99
Ridge	2.86	15.63	3.95	77.58	0.96
SGD	2.87	17.18	4.14	72.97	0.96
XGBoost	0.013	0.0022	0.047	1.01	0.99
(c) Testing results of the models in weekly rainfall forecasting					
CatBoost	0.062	0.013	0.118	4.44	0.99
ElasticNet	3.50	28.79	5.36	110.18	0.92
Lasso	3.03	22.83	4.77	89.63	0.94
LGBM	0.12	0.038	0.197	11.91	0.99
LR	2.77	15.19	3.897	81.89	0.96
MLP	1.62	5.15	2.27	86.86	0.98
RF	0.032	0.012	0.112	3.55	0.99
Ridge	2.738	15.24	3.90	79.14	0.96
SGD	2.740	16.44	4.05	74.03	0.95
XGBoost	0.019	0.0101	0.10	3.50	0.99

Bold value shows the better solution

accurate model, showing consistent performance across all metrics with an R^2 value of 0.99. This indicates that XGBoost effectively captures the underlying trends in the data, leading to precise weekly forecasts on the validation dataset. RF and CatBoost also maintain exceptional accuracy, each with an R^2 value of 0.99. These ensemble methods successfully utilize multiple decision trees to produce accurate predictions, resulting in strong performance across all metrics. These ensemble methods continue to leverage multiple decision trees effectively, resulting in high performance across all metrics. ElasticNet, Lasso, LR, Ridge, and SGD demonstrate stable and respectable performance, with R^2 values ranging from 0.92 to 0.96. This

range highlights their ability to consistently capture the fundamental trends in the data. LGBM and MLP also perform strongly, with LGBM achieving an R^2 value of 0.99 and MLP a value of 0.98, showcasing their effectiveness in modeling complex relationships and reinforcing their value for weekly trend forecasting. Again, XGBoost, RF, CatBoost, LGBM, and MLP emerge as the most consistent and accurate models, providing precise predictions across the validation dataset.

4.2.3 Test Performance Metrics for the Weekly Models

Table 2(c) provides an overview of the testing results for various models used in the weekly forecasting. The analysis of these results offers valuable insights into each model's performance. XGBoost remains a standout performer, maintaining consistent accuracy across all metrics. ElasticNet, Lasso, LR, Ridge, and SGD demonstrate steady performance, with moderate error scores and R^2 values ranging from 0.92 to 0.96, indicating their ability to consistently capture the fundamental trends in the data. LGBM and MLP also continue to excel, showcasing their strengths with low error scores and high R^2 values of 0.99 and 0.98, respectively. RF and CatBoost maintain their impressive accuracy with an R^2 value of 0.99. XGBoost, RF, CatBoost, LGBM, and MLP consistently deliver accurate predictions across the testing dataset.

The best models for predicting rainfall using CatBoost, RF, and LGBM for weekly data are illustrated in Figures SI-14 to SI-16 of the supplementary information. Figure SI-14 demonstrates the CatBoost model's ability to match actual rainfall data with its predictions, showcasing striking accuracy. Figure SI-15 highlights the RF model's effective performance in capturing weekly fluctuations. Figure SI-16 shows how well the LGBM model identifies rainfall peaks and troughs, maintaining close agreement with actual data. Conversely, Figure SI-17 depicts the Lasso model, which faces significant challenges and exhibits noticeable discrepancies from the actual rainfall data.

4.2.4 Comparison of Predicted Error Plots for Weekly Models

To gain a deeper understanding of the performance of various machine learning models (CatBoost, ElasticNet, Lasso, LGBM, LR, MLP, RF, Ridge, SGD, and XGBoost) in predicting rainfall, error scatter plots shown in Fig. 4(a)–(j) were analyzed to illustrate the relationship between predicted and actual rainfall values. The analysis reveals that XGBoost exhibits data point errors of less than 0 to 2% for rainfall predictions, as shown in Fig. 4(j). CatBoost and RF demonstrate data point errors ranging from 2 to 5%, with some errors extending between 5 and 15%, as depicted in Fig. 4(a) and (g). In contrast, LGBM displays a broader error range, spanning from 2 to 15% for rainfall data, as illustrated in Fig. 4(c). ElasticNet, Lasso, LR, MLP, Ridge, and SGD consistently show higher error percentages, exceeding 15% for rainfall predictions, as shown in Fig. 4(b), (d), (e), (f), (h), and (i), respectively. Based on these findings, it is reasonable to conclude that XGBoost outperforms the other models in terms of robustness and reliability for predicting weekly rainfall.

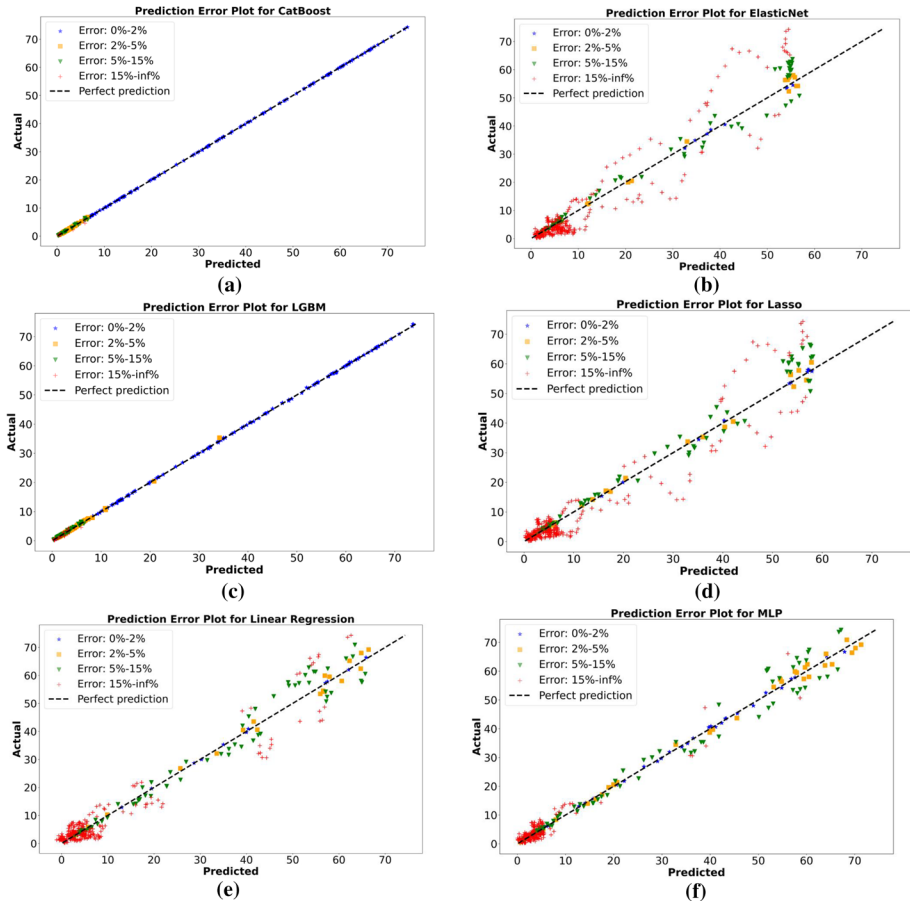


Fig. 4 (a) Scatter plots for the CatBoost weekly model. (b) Scatter plots for the ElasticNet weekly model. (c) Scatter plots for the LGBM weekly model. (d) Scatter plots for the Lasso weekly model. (e) Scatter plots for the LR weekly model. (f) Scatter plots for the MLP weekly model. (g) Scatter plots for the RF weekly model. (h) Scatter plots for the Ridge weekly model. (i) Scatter plots for the SGD weekly model. (j) Scatter plots for the XGBoost weekly model

5 Conclusion

This study provides an in-depth analysis of rainfall forecasting models using historical data from Uttarakhand, Uttar Pradesh, Haryana, Punjab, Himachal Pradesh, Madhya Pradesh, Rajasthan, and Delhi. The evaluation of both daily and weekly forecasting approaches revealed significant insights into model performance and accuracy. For daily forecasting, the CatBoost model demonstrated exceptional accuracy with an R^2 of 0.99, RMSE of 0.0014, and MAE of 0.0011, effectively capturing intricate daily rainfall patterns. XGBoost and RF followed closely with an R^2 of 0.99, RMSE of 0.0022, and MAE of 0.0017, also showing strong predictive capabilities. In contrast, Lasso Regression exhibited lower accuracy with an R^2 of 0.75, RMSE of 1.50, and MAE of 1.02, highlighting its limitations in modeling detailed rainfall variability. For weekly forecasting, XGBoost emerged as the top performer with an R^2 of 0.99, RMSE of 0.10, and MAE of 0.019, demonstrating its robustness in capturing weekly rainfall trends. RF achieved an R^2 of 0.99,

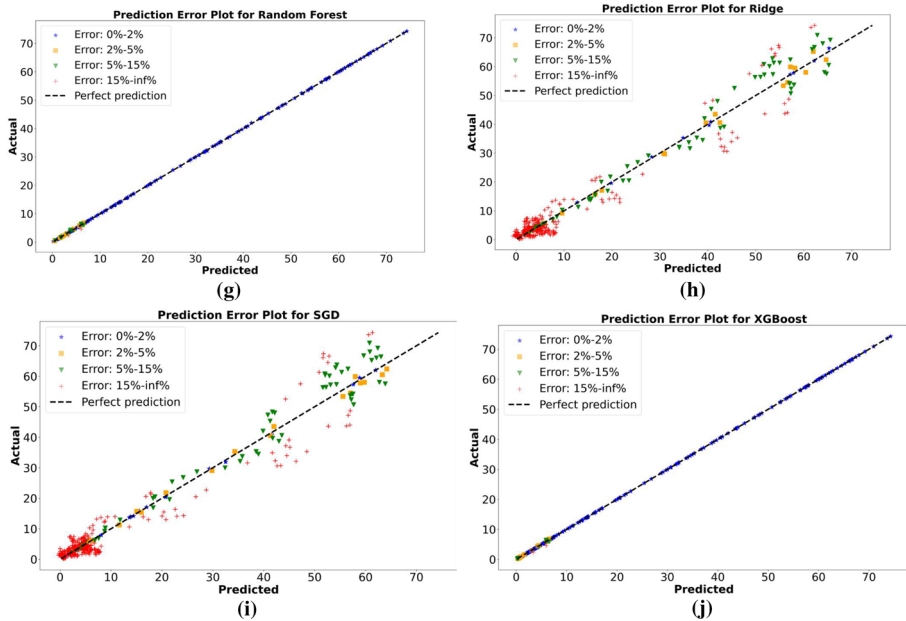


Fig. 4 (continued)

RMSE of 0.112, and MAE of 0.032, providing consistent and reliable predictions. CatBoost also performed effectively with an R^2 of 0.99, RMSE of 0.118, and MAE of 0.062. These findings underscore the effectiveness of advanced machine learning techniques in enhancing rainfall forecasting accuracy. The study highlights the critical role of precise rainfall predictions in informed decision-making across agriculture, water resource management, flood control, and urban planning. The successful application of these models promises to improve our understanding of regional climate patterns and their implications for weather forecasting in northern India.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s11269-024-03969-8>.

Acknowledgements The Authors extend their appreciation to the Deanship of Scientific Research at King Khalid University for funding this work through a large research group. Project group number RGP2/279/45.

Authors' Contributions Vijendra Kumar: Writing – original draft, Writing – review & editing, Data curation, Visualization, Formal analysis. Naresh Kedam: Investigation, Methodology. Ozgur Kisi: Writing – review & editing, Supervision, Conceptualization. Khaled Mohamed Khedher: Writing – review & editing, Investigation, Supervision. Saleh Alsulamy: Conceptualization, Writing – original draft. Mohamed Abdelaziz Salem: Methodology, Formal analysis.

Funding Open Access funding enabled and organized by Projekt DEAL. This research work was supported by the Deanship of Scientific Research at King Khalid University under grant number research group RGP2/279/45.

Data availability Data will be made available on request.

Declarations

Conflict of Interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.


References

- Abbasi M, Farokhnia A, Bahreinimotlagh M, Roozbahani R (2021) A hybrid of Random Forest and Deep Auto-Encoder with support vector regression methods for accuracy improvement and uncertainty reduction of long-term streamflow prediction. *J Hydrol (Amst)* 597:125717. <https://doi.org/10.1016/j.jhydrol.2020.125717>
- Abbot J, Marohasy J (2014) Input selection and optimisation for monthly rainfall forecasting in Queensland, Australia, using artificial neural networks. *Atmos Res* 138:166–178. <https://doi.org/10.1016/j.atmosres.2013.11.002>
- Bang Truong H, Cuong Nguyen X, Hur J (2023) Recent advances in g-C₃N₄-based photocatalysis for water treatment: Magnetic and floating photocatalysts, and applications of machine-learning techniques. *J Environ Manage* 345:118895. <https://doi.org/10.1016/j.jenvman.2023.118895>
- Chen SM, Wang YM, Tsou I (2013) Using artificial neural network approach for modelling rainfall-runoff due to typhoon. *J Earth Syst Sci* 122:399–405. <https://doi.org/10.1007/s12040-013-0289-8>
- Chen G, Zhang K, Wang S et al (2023) iHydroSlide3D v1.0: an advanced hydrological-geotechnical model for hydrological simulation and three-dimensional landslide prediction. *Geosci Model Dev* 16:2915–2937. <https://doi.org/10.5194/gmd-16-2915-2023>
- Diez-Sierra J, Del-Jesus M (2020) Long-term rainfall prediction using atmospheric synoptic patterns in semi-arid climates with statistical and machine learning methods. *J Hydrol (Amst)* 586:124789. <https://doi.org/10.1016/j.jhydrol.2020.124789>
- Endalie D, Haile G, Taye W (2022) Deep learning model for daily rainfall prediction: case study of Jimma, Ethiopia. *Water Supply* 22:3448–3461. <https://doi.org/10.2166/ws.2021.391>
- Feng Q, Wen X, Li J (2015) Wavelet Analysis-Support Vector Machine Coupled Models for Monthly Rainfall Forecasting in Arid Regions. *Water Resour Manage* 29:1049–1065. <https://doi.org/10.1007/s11269-014-0860-3>
- Karbasi M, Jamei M, Ali M et al (2022) Developing a novel hybrid Auto Encoder Decoder Bidirectional Gated Recurrent Unit model enhanced with empirical wavelet transform and Boruta-Catboost to forecast significant wave height. *J Clean Prod* 379:134820. <https://doi.org/10.1016/j.jclepro.2022.134820>
- Mekanik F, Imteaz MA, Gato-Trinidad S, Elmahdi A (2013) Multiple regression and Artificial Neural Network for long-term rainfall forecasting using large scale climate modes. *J Hydrol (Amst)* 503:11–21. <https://doi.org/10.1016/j.jhydrol.2013.08.035>
- Nunno F, Granata F, Pham QB, de Marinis G (2022) Precipitation Forecasting in Northern Bangladesh Using a Hybrid Machine Learning Model. *Sustainability* 14:2663. <https://doi.org/10.3390/su14052663>
- Rahman A, Abbas S, Gollapalli M et al (2022) Rainfall Prediction System Using Machine Learning Fusion for Smart Cities. *Sensors* 22:3504. <https://doi.org/10.3390/s22093504>
- Tan J, Xie X, Zuo J et al (2021) Coupling random forest and inverse distance weighting to generate climate surfaces of precipitation and temperature with Multiple-Covariates. *J Hydrol (Amst)* 598:126270. <https://doi.org/10.1016/j.jhydrol.2021.126270>
- Xie X, Xie B, Cheng J et al (2021) A simple Monte Carlo method for estimating the chance of a cyclone impact. *Nat Hazards* 107:2573–2582. <https://doi.org/10.1007/s11069-021-04505-2>
- Xu J, Zhou G, Su S et al (2022) The Development of A Rigorous Model for Bathymetric Mapping from Multispectral Satellite-Images. *Remote Sens (Basel)* 14:2495. <https://doi.org/10.3390/rs14102495>
- Yin L, Wang L, Keim BD et al (2023) Spatial and wavelet analysis of precipitation and river discharge during operation of the Three Gorges Dam. *China Ecol Indic* 154:110837. <https://doi.org/10.1016/j.ecoliind.2023.110837>
- Yu P-S, Yang T-C, Chen S-Y et al (2017) Comparison of random forests and support vector machine for real-time radar-derived rainfall forecasting. *J Hydrol (Amst)* 552:92–104. <https://doi.org/10.1016/j.jhydrol.2017.06.020>

- Zhao Y, Li J, Zhang L et al (2023) Diurnal cycles of cloud cover and its vertical distribution over the Tibetan Plateau revealed by satellite observations, reanalysis datasets, and CMIP6 outputs. *Atmos Chem Phys* 23:743–769. <https://doi.org/10.5194/acp-23-743-2023>
- Kumar V, Yadav SM (2020) Optimization of Water Releases from Ukai Reservoir Using Jaya Algorithm. Springer Singapore
- Zhang C, Brodeur ZP, Steinschneider S, Herman JD (2022) Leveraging Spatial Patterns in Precipitation Forecasts Using Deep Learning to Support Regional Water Management. *Water Resour Res* 58:. <https://doi.org/10.1029/2021WR031910>
- Zhao Y, Li J, Wang Y, et al (2024) Warming Climate-Induced Changes in Cloud Vertical Distribution Possibly Exacerbate Intra-Atmospheric Heating Over the Tibetan Plateau. *Geophys Res Lett* 51:. <https://doi.org/10.1029/2023GL107713>
- Zhou Z, Ren J, He X, Liu S (2021) A comparative study of extensive machine learning models for predicting long-term monthly rainfall with an ensemble of climatic and meteorological predictors. *Hydrol Process* 35: <https://doi.org/10.1002/hyp.14424>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Vijendra Kumar¹ · Naresh Kedam² · Ozgur Kisi^{3,4}  · Saleh Alsulamy⁵ · Khaled Mohamed Khedher⁶ · Mohamed Abdelaziz Salem⁷

✉ Ozgur Kisi
ozgur.kisi@th-luebeck.de

Vijendra Kumar
vijendra.kumar@mitwpu.edu.in

Naresh Kedam
naresh.kedam@ssau.ru

Saleh Alsulamy
s.alsulamy@kku.edu.sa

Khaled Mohamed Khedher
kkhedher@kku.edu.sa

Mohamed Abdelaziz Salem
moabdulaziz@kku.edu.sa

¹ Department of Civil Engineering, Dr. Vishwanath Karad MIT World Peace University, Kothrud, Pune, Maharashtra 411038, India

² Department of Thermal Engineering and Thermal Engines, Samara National Research University, Mos-Kovskoye Shosse, 34, Samara 443086, Russia

³ Department of Civil Engineering, Luebeck University of Applied Sciences, 23562 Lübeck, Germany

⁴ Department of Civil Engineering, Ilia State University, 0162 Tbilisi, Georgia

⁵ Department of Architecture & Planning, College of Engineering, King Khalid University, 61421 Abha, Saudi Arabia

⁶ Department of Civil Engineering, College of Engineering, King Khalid University, 61421 Abha, Saudi Arabia

⁷ Department of Industrial Engineering, College of Engineering, King Khalid University, 61421 Abha, Saudi Arabia