IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2024.Doi Number

Rainfall Prediction Using Integrated Machine Learning Models with K-Means Clustering: A Representative Case Study of Harirud Murghab Basin-Afghanistan

Ziaul Haq Doost[1], Ali Alsuwaiyan[2,3], Abdulazeez Abdulraheem[4], Nabil M. Al-Areeq5[*] and Zaher Mundher Yaseen[1,6]

[1]Department of Civil and Environmental Engineering, King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia, Email: ziaulhaq.doost@gmail.com & z.yaseen@kfupm.edu.sa
[2]Department of Computer Engineering, King Fahd University of Petroleum and Minerals, Dhahran, 31261, Saudi Arabia, Email: alisuwaiyan@kfupm.edu.sa
[3]Interdisciplinary Research Center for Intelligent Secure Systems, KFUPM, Dhahran, Saudi Arabia
[4]College of Petroleum Engineering & Geosciences, King Fahd University of Petroleum & Minerals, Dhahran, Saudi Arabia, Email: aazeez@kfupm.edu.sa
[5]Department of Geology and Water Resources, Center of Water and Climate Changes, Thamar University, Yemen, Email: alareeqnabil@gmail.com
[6]Interdisciplinary Research Centre for Membranes and Water Security, King Fahd University of Petroleum & Minerals, 31261, Dhahran, Saudi Arabia

[*]Corresponding author: Nabil M. Al-Areeq (alareeqnabil@gmail.com).

ABSTRACT Accurate rainfall prediction was essential for effective water resource management and disaster preparedness, especially in regions with limited observational data such as Afghanistan. This study objective was to develop a reliable rainfall prediction machine learning (ML) model by integrating satellite-derived meteorological data with ground-based observational rainfall data. Four ML models including Gradient Boosting Regressor (GBR), Hist Gradient Boosting Regressor (HGBR), Random Forest Regressor (RFR), and Xtreme Gradient Boosting Regressor (XGBR) were used. The models were evaluated at three stations (Nazdik-i Herat, Shinya, and Torghundi) using coefficient of determination (R^2), mean squared error (MSE), root mean squared error (RMSE), Mean absolute error (MAE), and median absolute error (MedAE) as evaluation metrics. Results showed at Nazdik-i Herat station, HGBR model achieved R^2 of 0.90 for training phase and 0.83 for testing phase, RMSE of 7.35 and 10.25, and MAE of 5.41 and 6.45 for training and testing phases respectively. At Shinya station, HGBR model obtained an R^2 of 0.76 and 0.76, RMSE of 10.91 and 9.51, and MAE of 6.25 and 7.44 for training and testing phases respectively. At Torghundi station, this model recorded R^2 of 0.92 and 0.80, RMSE of 5.51 and 8.54, and MAE of 3.97 and 5.97 for training and testing phases respectively. While other models showed good performance, HGBR was the only model that consistently maintained high accuracy and low error across both training and testing phases at all stations. Making it the best performing model for monthly rainfall prediction in the region. This study contributed scientifically by showing the effectiveness of satellite data and advanced ML models in monthly rainfall prediction for data-limited regions. Practically, it enables a cost-effective alternative to physical rain gauge installations by offering a reliable method to estimate monthly rainfall in order to support irrigation planning, water resource management, and disaster preparedness in climate-vulnerable regions.

INDEX TERMS Climate change, Machine Learning, Rainfall prediction, Time series analysis

I. INTRODUCTION

Accurate and precise forecasting of rainfall has historically been one of cornerstones of water resource management, disaster mitigation, and agricultural planning [1–3]. In areas like Afghanistan, with significant, subsistence agriculture communities, the need for accurate weather predictions, and cool-season storm forecasts is not just crucial, but lucrative [4,5]. The effect of these patterns is becoming increasingly unpredictable given that such patterns are on the trajectory due to climate change [6,7], which poses a significant risk to

the well-being, livelihoods, and food security of the settlers especially concerning if exacerbated. Further, the latest advancements in Artificial Intelligence (AI) and Machine Learning (ML) are showing new possibilities to make forecasting weather prediction models more trustable and accurate [6,8]. Unlike to traditional meteorological strategies, which usually dealing with the complicated dynamics of climate systems, ML models can be learnt from big datasets, revealing patterns and dependencies that by-pass human analysts and conventional computational methods [9,10].

A. RESEARCH BACKGROUND

The Harirud Murghab Basin (HMB) is located in the northwestern part of Afghanistan, and exposing a unique climatic condition reputable by its semi-arid and arid climate [11,12]. The variability of rainfall in this region puts crucial challenges to water resource management, agricultural productivity, and the entire sustainability of the local communities [11,13]. Agriculture, the foundation of Afghanistan's economy, depends heavily on the predictability of seasonal rainfalls, which are vital for crop's irrigation and thriving local water sources [14]. However, the intrinsic unpredictability synthesized with precipitation patterns, augmented by the impacts of climate alteration, necessitates the development of more accurate, and precise forecasting ways to ensure the flexibility of this natural society [14]. Furthermore, conventional methods of weather forecasting, while having been served their mission in the past, are progressively authenticating its deficiency in the face of complicated climate alteration and the requirement for precise and accurate predictions [15,16]. These techniques often rely on the past weather datasets and linear models that cannot secure adequately the non-linear engagements in the midst of the atmosphere [17]. Consequently, there is a progressing recognition of the drawbacks of traditional meteorological approaches in supplying the intricate and accurate predictions that are required to effectively and efficiently organize agricultural practices and water resources in Afghanistan [18].

Moreover, advancements in AI and ML technology in recent years opened new doors in the field of meteorology [19]. AI and ML models can process large datasets. As a result, ML models have the ability to learn and process from big datasets. Additionally, these models can easily be used to identify patterns and predict weather's conditions with higher efficiency and more precision [19]. Analyzing huge climate data, e.g., temperature, wind patterns, humidity, and atmospheric pressure, to generate accurate and reliable predictions is among the capabilities of these models [20]. For technological improvements, the utilization of ML in weather events prediction offer a significant change in our comprehension and provide easy prediction of climate conditions [21]. Further, increased prediction efficiency provides farmers for easier informed decisions about better farming, and irrigation strategies, minimizing the probability of crop failure and boosting food sustainability. Therefore, improved forecasting can also help in better management of water resources by relying on that the restricted water sources are organized effectively. Consequently, the study's research background focuses on the significant need for innovative solutions to Afghanistan's rainfall forecasting problems. This study uses ML models to assist the region evolve more flexible agricultural and water management techniques, supporting the communities' livelihoods and promoting sustainable development.

B. LITERATURE REVIEW

In the prospective of comprehensive literature review, a study that emphasized on rainfall prediction using a deep learning (DL) model was presented to advance rainfall anticipating technologies [22]. They used a large dataset that included past days' rainfall, temperature averages, relative humidity, changes in barometric pressure, and a few other meteorological features to forecast the total amount of rainfall that would fall on the following day [22]. They were motivated to unlock the potential of these cutting-edge AI approaches to improve rainfall forecast accuracy and precision by leveraging DL techniques, particularly multilayer perceptrons (MLP) and autoencoders [22]. For its application of DL structure, the research is important in indicating its effectiveness in achieving the intricate weather patterns dynamics in prediction of rainfall with unique precision and accuracy [22,23].

Moreover, another research was conducted on the prediction of rainfall's average over Udupi district in Karnataka by using Artificial Neuron Network (ANN) models [24]. A three-layered ANN with different configurations was employed using daily average humidity, and wind speed for rainfall forecasting over a period of 50 years [24]. Authors of the study used different algorithms, such as Learning Gradient Descent (LEARNGD), back propagation algorithm (BPA), Training Levenberg-Marquardt (TRAINLM), and Learning Gradient Descent with Momentum (LEARNGDM), to identify the models' performances in the network based on the number of hidden neurons [24]. The findings provided useful knowledge on configuring the ANN algorithms for rainfall prediction since this is necessary for accurate weather forecasts [24].

Another study tested several experiments Long Short-term Memory (LSTM), XGB, Bidirectional-LSTM, Stacked-LSTM, ensemble gradient boost regressor, linear support vector regression (LSVR), and Extra-tree Regressor (XTR) rainfall modeling [25]. The aim was to predict quartic mean volume of hourly rainfall by utilizing these ML algorithms on climate time-series data from critical cities in the United Kingdom (UK) [25]. The study compared the predictive quality of these algorithms and suggested the most reliable one for rainfall forecasts [25].

For deeper insights into rainfall prediction harnessing the potential of AI and ML, Table 1 outlined different studies that cover various geographical regions, data inputs, and AI strategies. These contributions show the advancements in the field of meteorological prediction, where AI and ML are at the heart of innovation, suggesting promising solutions to sustainable challenges in weather forecasting.

TABLE 1. Survey of different ML algorithms on predicting rainfall patterns

Reference	Problem	Data input	Data output	Data type	AI Methods used	Methodology
[70]	Rainfall Prediction	Rainfall time series data from 1871 to 2016 (8 input vectors used)	Predicted rainfall pattern	1871-2016	1. K-nearest neighbor (KNN) 2. Artificial neural network (ANN) 3. Extreme learning machine (ELM)	Empirical Correlation
[71]	Prediction of Daily Rainfall	maximum temperature, minimum temperature, wind speed, relative humidity and solar radiation	daily rainfall	3653	1. adaptive network based fuzzy inference system (ANFIS) 2. Particle Swarm Optimization (PSO) 3. Support Vector Machines (SVM)	Empirical Correlation
[72]	Rainfall Prediction	temperature, dew point, humidity, wind pressure, wind speed, and wind direction	Rainfall	2000-2014	1. Long Short-Term Memory 2. Recurrent Neural Networks and Tanh Activation 3. LSTM	Empirical Correlation
[73]	Using AI models to predict daily rainfall	Rainfall (Min, Max, Mean, St. D, and Mode)	Forecast daily rainfall (mm)	932	1. Artificial neural network (ANN) 2. Boundary-corrected maximal overlap discrete wavelet transformation (MODWT) 3. Long sort-term memory (LSTM)	Three methods of AI were modeled to predict daily rainfall
[77]	Spatial interpolation of daily rainfall data to predict the volume of rainfall at unknown locations within an area covered by existing observations	Rainfall (Min, Max, Mean, St. D, Median)	Predict volume of rainfall at unknown locations within area covered by existing observations.	367	1. self-organizing map (SOM), 2. backpropagation neural networks (BPNN) 3. fuzzy rule 4. Soft Computing 5. ANN	Soft computing and ANN models were utilized to predict the volume of rainfall
[78]	The study addresses the challenge of predicting rainfall in India for agriculture, Utilized ML algorithms, particularly Artificial Neural Network models to predict the rainfall.	Not Mentioned	Rainfall	N/A	1. ARIMA MODEL (Autoregressive Integrated Moving Average) 2. ANN 3. SVM 4. Self-Organizing Map	the study used four types of AI models to predict the rainfall and compare them for their performance
[79]	The problem addressed in the study is the challenging task of accurate rainfall prediction, particularly in the context of extreme climate variations, and the difficulty in selecting an appropriate classification technique for prediction.	Temperature, Visibility, Dew point temperature, atmospheric pressure (sea level) atmospheric pressure (weather station), relative humidity, pressure tendency, maximum temperature, minimum temperature, mean wind speed	Rainfall prediction system	25,919	1. Decision tree 2. Naïve Bayes 3. K-nearest neighbors 4. support vector machines.	The study used a ML fusion approach, integrating Decision Tree, Naïve Bayes, K-Nearest Neighbors, and Support Vector Machines with fuzzy logic, to create a real-time rainfall prediction system for smart cities. The model was trained and evaluated using 12 years of historical weather data for Lahore, including preprocessing steps like cleaning and normalization.
[80]	The study addresses the challenge of accurately predicting long-term rainfall, crucial for preventing disasters and impacting economies like India's agriculture. It explores using ML, specifically regression, to enhance prediction accuracy.	Rainfall data set	Accuracy/ Error of prediction of rainfall	Data consists of 19 attributes (individual months, annual, and combinations of 3 consecutive months) for 36 subdivisions.	1. Support Vector Regression 2. Kernel 3. Lasso Regression	The study used Support Vector Regression (SVR) with linear and RBF kernels for rainfall estimation, showing SVR's effectiveness in handling data challenges. Compared to Multiple Linear Regression (MLR), SVR proved superior in capturing non-linear relationships, and the tuned SVR model provided the most accurate predictions.
[81]	The problem covered in the study is the complexity of precise rainfall forecast in the edge of climate change, advancing the assessment of classification algorithms' accuracy for rainfall prediction in various ecological zones of Ghana.	maximum temperature, minimum temperature, rainfall, relative humidity 0600, relative humidity 1500, sunshine, wind speed.	Rainfall	5,746	1. ANN-MLP 2. KNN 3. DT 4. RF 5. XGB	The study predicted rainfall in Ghana using five algorithms on data from 1980 to 2019. Decision Tree was fastest; Random Forest, Extreme Gradient Boosting, and Multilayer Perceptron performed well. K-Nearest Neighbor showed poor performance, requiring further investigation.

C. RESEARCH MOTIVATION AND CONTRIBUTION

The pressing need for accurate and reliable rainfall prediction methodologies, especially in regions vulnerable to climate variability, water scarcity, and limited observational monitoring gauges formed the cornerstone of the motivation behind this research. The HMB in Afghanistan exemplifies such a region where agriculture, which is the lifeline of the local economy, depends heavily on predictable rainfall patterns. While useful, conventional prediction methods usually lead to low precision and reliability. A significant gap would be left in our capabilities for effectively managing water resources, and agricultural planning. The breakthrough of AI and ML algorithms offers a hope's beacon, promising to revolutionize the meteorology field through improved predictive capabilities that far surpass those of traditional methods.

The motivation of our study is inspired by the dual challenge of improving both rainfall prediction precision and accuracy, in addition to offering a feasible comprehension for water resource management and agricultural sustainability within HMB. By using the strength of ML algorithms, this study wants to bridge the gap between the potential of advanced AI technologies on one hand, and the practical requirements of regions facing severe challenges because of unforeseeable weather patterns on the other hand. The employment of algorithms in ML models in this context is not only a scientific experiment but also a focused attempt to respond to the significant susceptibility of agricultural societies to climate change.

The contributions of this study are threefold. (i) it introduced a novel and practical integration of satellite-obtained meteorological input data with rainfall observations from ground-based for monthly rainfall prediction in arid and data-scarce regions. This satellite-to-observational gauge data methodology not only enables accurate prediction but also reduces dependency on physical monitoring infrastructure. (ii) the study proposed a hybrid ML model that incorporates K-means clustering with advanced ML algorithms such as GBR, HGBR, RFR, and XGBR, to improve performance. (iii) it presented a comparative evaluation across multiple stations in Afghanistan and offered a scientifically grounded method for cross-validating extreme rainfall events using satellite data, thereby demonstrating the feasibility and reliability of this methodology for real-world applications.

D. RESEARCH OBJECTIVES

The primary objectives of this study were to address the urgent need for reliable rainfall prediction in data-scarce regions, to: (i) evaluate the effectiveness of integrating satellite-obtained meteorological data with ML models for monthly rainfall prediction, particularly for locations lacking extensive historical records. (ii) perform a comparative analysis of four advanced ML algorithms (GBR, HGBR, RFR, and XGBR), to identify the model that provides the best balance between accuracy and generalization. (iii) develop a hybrid prediction framework by integrating clustering (K-means) with ML models. (iv) assess the scientific validity of using satellite inputs to cross-validate extreme rainfall values, offering a robust and cost-effective alternative to installing physical gauges in similar arid and semi-arid regions globally.

II. MATERIALS AND METHODS

The study started with the collection of two distinct sets of data: observational gauge data capturing monthly rainfall amounts and satellite-derived meteorological data for the target study region. The data then underwent a meticulous preprocessing stage, which ensured the integration of various satellite-based inputs such as monthly sums of Earth Skin Temperature (TS), the ratio of Temperature at 2 Meters to Specific Humidity at 2 Meters (T2M/QV2M), Relative Humidity at 2 Meters (RH2M), Surface Pressure (PS), and Wind Speed at both 10 and 50 Meters (WS10M and WS50M), respectively. These input features were used as predictors in the subsequent ML modeling phase, where four different ML algorithms (GBR, HGBR, RFR, and XGBR) were thoughtfully trained using the training set (70% of dataset) and assessed for their performance to see if they could accurately and efficiently predict total monthly rainfall. The output feature from these models was then evaluated to check the models' effectiveness, with outcome representations showing the level of success obtained by each of the four ML models in predicting the total monthly rainfall (Fig. 1).

A. STUDY REGION

The study region, Harirud-Murghab Basin (HMB) (Fig. 2), is situated in the western part of Afghanistan and represents a key agricultural hub within the country [26]. This region is characterized by a semi-arid climate condition, where water shortage puts a remarkable challenge to farming communities [11]. The economy of Afghanistan fundamentally relies on the country's agriculture [27,28], with a considerable part of Afghanistan's population hinging on farming and livestock for their livelihoods [29]. This region has a wide range of climatic conditions, including hot summers and very cold winters, and extremely variable rainfall in terms of quantity, and distribution all over the year [30,31].

Unique geographical and climatic conditions of Afghanistan make it an ideal choice as case study for the employment of advanced ML models to predict meteorological data, especially rainfall patterns [32]. The change in the rainfall pattern, which is significant for the crops' irrigation, directly influences agricultural yield, and water resource management throughout the region [33]. The region's dependency on agriculture is well-known where efficient and accurate rainfall prediction could vitally increase the resilience of all local communities to the uncertainties of the climate [34]. Furthermore, the climate and environmental challenges of Afghanistan mirror those of other arid and semi-arid regions globally [11], making the results of this study potentially relevant on a world

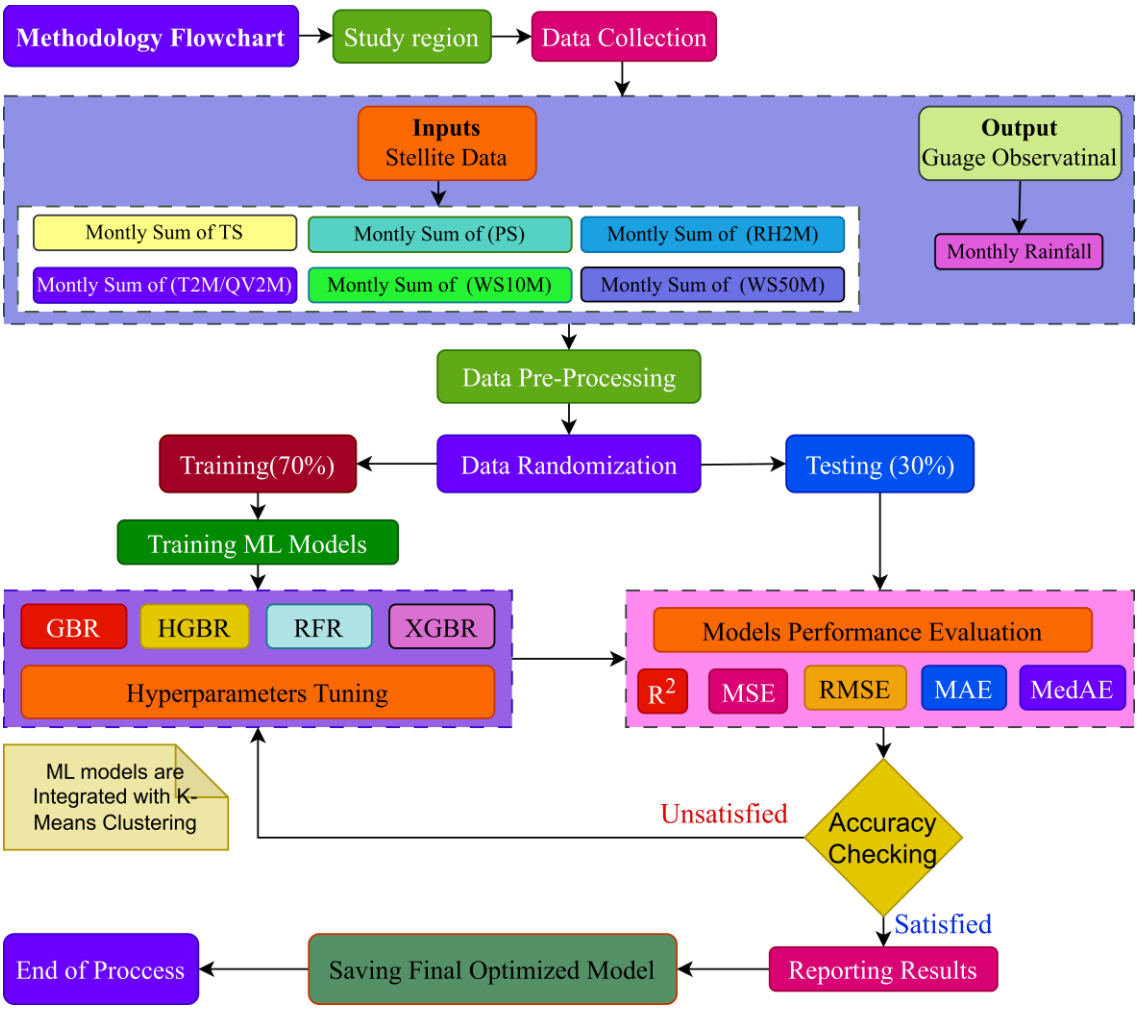


FIGURE 1. Adopted methodology flowchart.

scale. The selection of HMB as the study region was also inspired by a broader need for novel solutions to water management and agricultural planning in regions susceptible to climate variability. By concentrating on HMB, this vital research aims not only to handle local challenges but also to contribute to understanding how ML algorithms can be effectively and efficiently employed in a region with similar environmental conditions worldwide.

B. Data collection

The data for this study were thoughtfully collected from two primary sources to ensure a robust analysis of rainfall patterns within the HMB of Afghanistan.

1) Local Meteorological Data

The first portion of the data set was obtained officially from the Ministry of Energy and Water of Afghanistan. This compilation includes detailed records of daily rainfall measurements taken at various meteorological stations throughout the region (Fig. 2). The data span multiple years, with the specific time range varying from one station to another. Each station's data set provides a unique chronological insight into the local precipitation trends, crucial for the ML models. The geographical distribution and temporal range of these stations are systematically outlined in Table 2.

2) SATELLITE-DERIVED CLIMATIC DATA

The second data set was sourced from NASA's POWER (Prediction of Worldwide Energy Resources) Data Access Viewer [35]. This data encompasses a variety of climatic inputs necessary for the comprehensive modeling of rainfall predictions. It mirrors the stations used in the local data collection, thereby providing a direct comparison and augmentation of the ground-recorded meteorological data. Parameters from this data set include, but are not limited to, atmospheric temperature, humidity, solar radiation, and wind characteristics, which are crucial for understanding the micro and macro climatic variables influencing rainfall. The specifics of these inputs are detailed in Table 3, where each variable is outlined in terms of its relevance and contribution to the accuracy of the predictive

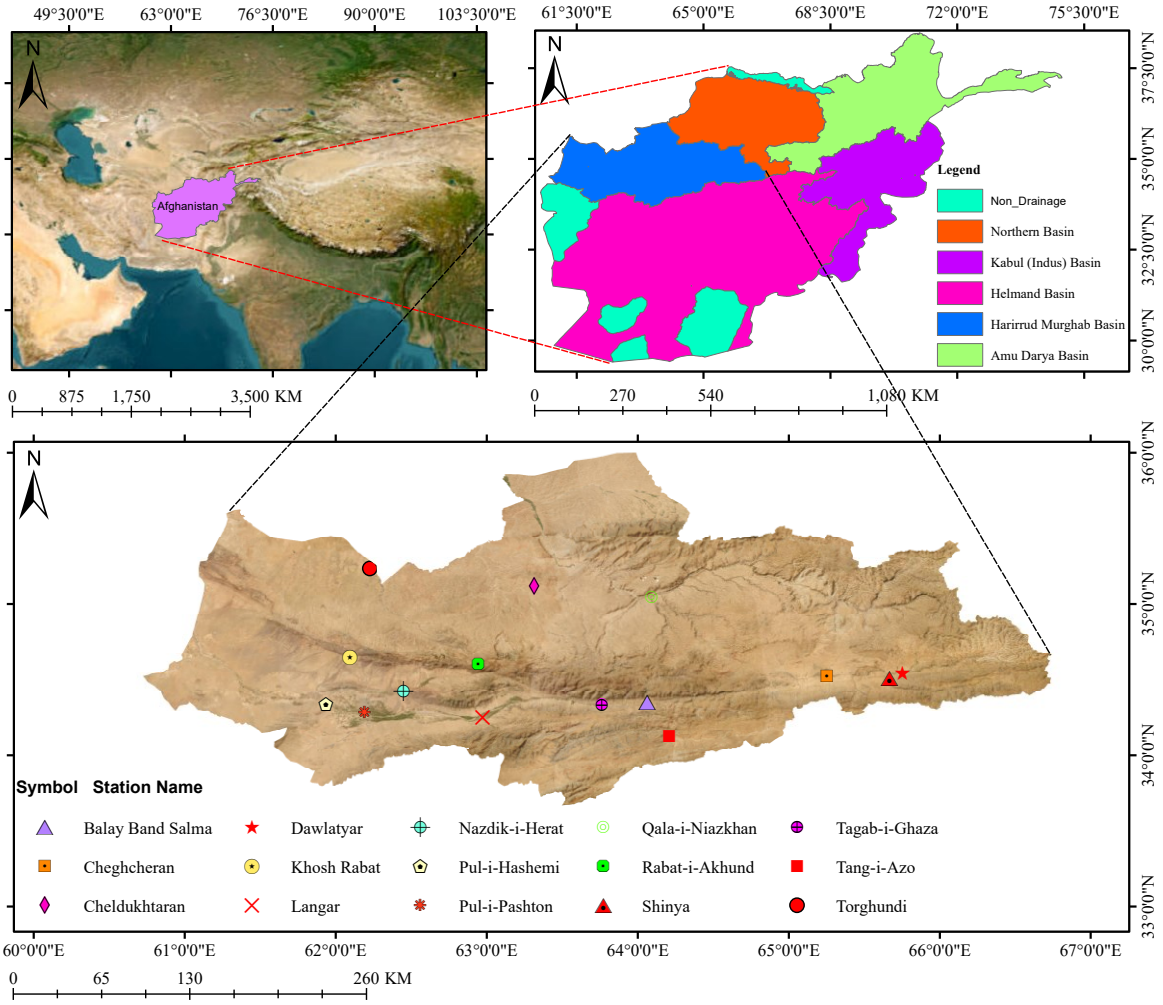


FIGURE 2. Study region map [Harirud Murghab Basin (HMB)] with installed gauge station

models developed in this study. By integrating these diverse data sources, the study leverages local and global meteorological insights to formulate a nuanced understanding of the rainfall dynamics in Afghanistan. This dual-source approach not only enriches the dataset but also increases the reliability

TABLE 2. Gauges' observational rainfall data for the Harirud Murghab Basin

Station Name	Available data collected for the period of		Latitude	Longitude
	From	To		
Balay Band Salma	June 7, 2020	February 21, 2024	34.35138	64.06421
Cheghcheran	January 1, 2009	March 3, 2024	34.522275	65.25357222
Cheldukhtaran	January 1, 2008	February 18, 2024	35.12134722	63.31555
Dawlatyar	January 1, 2008	February 28, 2024	34.54715278	65.75411944
Khosh Rabat	January 1, 2008	March 3, 2024	34.64427222	62.09452222
Langar	January 23, 2024	February 20, 2024	34.24761	62.9728233
Nazdik-i-Herat	January 1, 2008	January 31, 2023	34.4212467	62.44860556
Pul-i-Hashemi	January 1, 2008	February 20, 2024	34.34070278	61.93655556
Pul-i-Pashton	June 16, 2020	February 21, 2024	34.2873967	62.1917756
Qala-i-Niazkhan	October 05, 2020	December 27, 2023	35.04936	64.09158
Rabat-i-Akhund	January 1, 2008	February 20, 2024	34.60509444	62.94422778
Shinya	January 1, 2008	February 28, 2024	34.50804722	65.66881944
Tagab-i-Ghaza	January 1, 2008	February 15, 2024	34.33619167	63.76188056
Tang-i-Azo	January 1, 2008	April 9, 2018	34.12851111	64.20853005
Torghundi	January 1, 2008	February 18, 2024	35.252925	62.28284167

and validity of the predictive outcomes generated by our ML algorithms.

TABLE 3. Satellite data

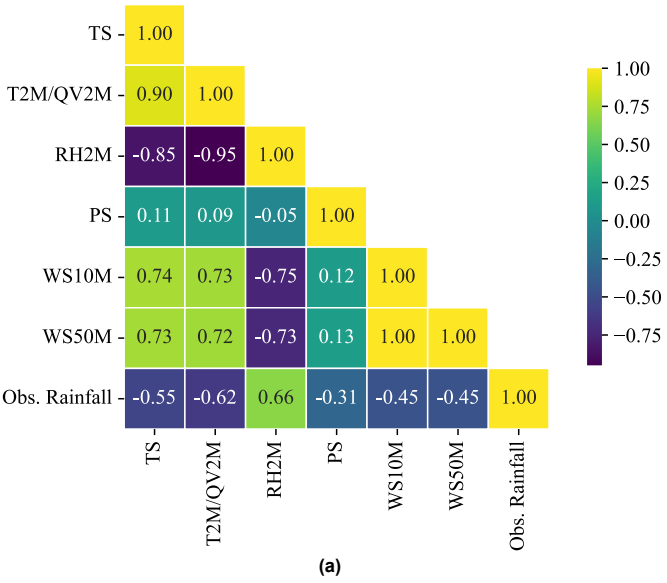
Short Term used the manuscript and models	Full name of the criteria
TS	Earth Skin Temperature (C°)
T2M	Temperature at 2 Meters (C°)
QV2M	Specific Humidity at 2 Meters (g/kg)
RH2M	Relative Humidity at 2 Meters (%)
PS	Surface Pressure (kPa)
WS10M	Wind Speed at 10 Meters (m/s)
WS50M	Wind Speed at 50 Meters (m/s)

C. DATA PRE-PROCESSING

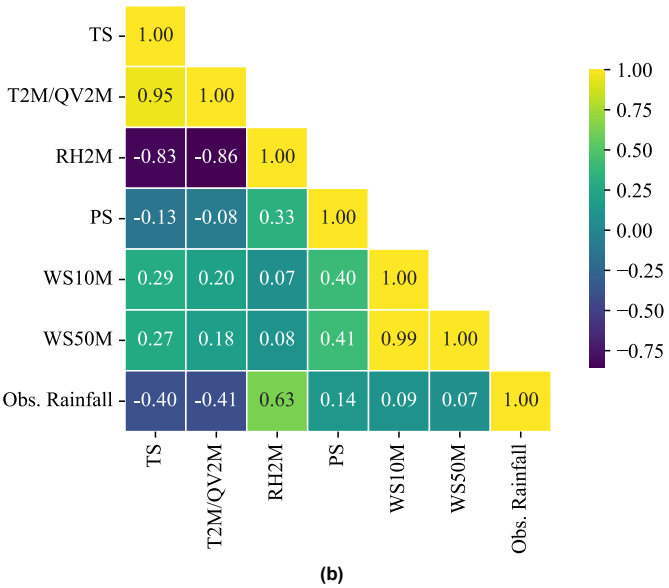
In this study, data were carefully obtained from gauged observations at selected meteorological stations and coincided with satellite data to ensure accuracy and completeness. Three stations within the HMB (Nazdik-i Herat, Shinya, and Torghundi stations) were randomly chosen for their well-maintained records spanning from January 1, 2008, to early 2024. The selection process was designed to avoid bias, including one station close to Herat City, believed to have reliable data due to its proximity to a major urban center, and another station further away to capture a more diverse range of data. Nazdik-i Herat station, with complete data up to January 31, 2024, and Shinya station, up to February 28, 2024, were included for their comprehensive monthly records, while Torghundi station was dataset covered up to January 31, 2024, excluding incomplete data for February. Satellite data corresponding to these stations' coordinates were collected to serve as predictive inputs. Data was initially processed on a daily basis, then aggregated into monthly totals through Excel pivot tables to align with the study's monthly prediction objectives. Subsequent preprocessing included transformations to improve input variable correlation, outlier removal for data quality assurance, K-means clustering for data stratification, and normalization to standardize the feature scales. These preparatory steps refined the datasets, ensuring that ML models were primed for robust training and validation.

1) DATA TRANSFORMATIONS

Data transformations are pivotal in ML to adjust skewness and enhance feature relationships, thereby improving model performance [36–38]. This study applied transformations like squaring, cubing, square roots, exponentials, and ratios between features themselves to uncover patterns that might not be evident in raw forms. These methods aim to increase the correlation coefficient (CC) with rainfall data. Notably, the ratio of Temperature at 2 Meters (T2M) to Specific Humidity at 2 Meters (QV2M) showed a consistent and significant improvement in CC across all stations, suggesting a strong predictive relationship with rainfall. Other transformations, despite potential at specific stations, did not universally enhance CC among all stations, leading to their exclusion from the final inclusion decision. The effective transformation, T2M/QV2M, was therefore adopted for all stations, ensuring a robust and consistent model performance. The results of these transformations and their impacts on CC are detailed in Fig. 3.



(a)



(b)

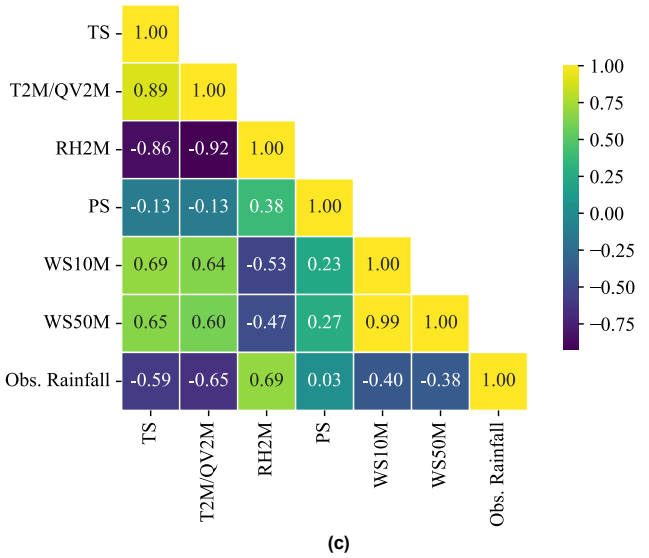


FIGURE 3. Correlation coefficient matrix heatmap; (a) for Nazdik-i Herat station, (b) for Shinya Station, (c) for Torghundi Station

2) HANDLING OUTLIERS

Outliers, which deviate significantly from other observations, can skew predictions and distort the accuracy of models, especially in meteorological applications like rainfall prediction [39,40]. These outliers often result from data collection errors, faulty measurements, or typical extreme weather events [41,42]. For linear regression models, which assume normally distributed data, outliers can be particularly problematic [43]. To enhance model reliability, outliers were identified using the Interquartile Range (IQR) method, defined as observations below Q1 - 1.5 × IQR or above Q3 + 1.5 × IQR [44]. This technique effectively handles skewed distributions common in environmental data, ensuring the models reflect more typical climatic behaviors [44].

3) DATA NORMALIZATION

Data normalization is very important in ML to adjust the scale of numerical features to a common range, ensuring no single feature disproportionately influences the model's outcome [45]. This study used Min-Max normalization, which scales features on a 0 to 1 range, maintaining the distribution without altering the relative relationships between values [46]. The formula used is:

$$X_{normalized} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

Here, X is the original value, with X_{min} and X_{max} representing the minimum and maximum values of the corresponding feature, respectively. While sensitive to outliers, this method is effective for distance-based and iterative algorithms [45,47]. Outlier removal prior to normalization mitigates potential distortions, enhancing model training efficiency and performance.

4) K-MEANS CLUSTERING

K-means clustering is a widely used unsupervised ML technique that partitions a dataset into K distinct, non-overlapping subsets or clusters [48]. The goal is to group the data points into clusters such that each point is assigned to the cluster with the nearest mean, which serves as the prototype of the cluster [48]. This method of clustering is very useful in revealing the intrinsic architecture in the data and is advantageous to identify the patterns or groupings in the data, where none are clearly labeled [49]. In this pivotal study, the K-means clustering method was used as a pre-processing phase to categorize and group the climatic variables within explicit clusters based on the similarities of each category [49]. Therefore, the aim was to uncover hidden patterns in the meteorological data that could be indicative of specific weather conditions leading to the rainfall. While applying K-means clustering to the dataset, we observed clear categories indicating distinct clusters in the input features for the Nazdik-i Herat, Shinya, and Torghundi stations. For better understanding of this important process clearly, a visualization of each input feature against Earth Skin Temperature (TS) represented by Fig. 4.

D. UTILIZED ML MODELS

Whenever there is a need for accurate rainfall prediction, the selection of suitable ML algorithms stands as a crucial approach, expanding the boundaries of conventional meteorological methods. These models, with their inherent capacity to identify complex patterns and relationships in a very large dataset, are specifically capable of tackling the details of climate variations and their effect on rainfall. The inspiration for the application of ML algorithms in this study came out with the objective of encompassing computational intelligence to integrate the different factors impacting weather patterns into reasonable, and feasible insights.

Furthermore, the selection of suitable and proper ML algorithms is vital in responding to the specific challenges presented by the rainfall prediction. To select a suitable model, careful consideration is very important for the models' harmony with the nature of data type (for this study climatology data), the predictive ability, and the capacity to generalize with different conditions, and scenarios while avoiding challenges of overfitting. In the current section, we delve deeper into the complexity of each model, with the inclusion of four advanced ML models (algorithms). Each model was chosen based on its tested abilities in handling the time-series data and forecasting assignments within the domain of atmospheric sciences. The models that were selected for this study include GBR, HGBR, RFR, and XGBR. These models were used for comparative analysis to determine models' performance, flexibility to data variability, accuracy, and efficiency.

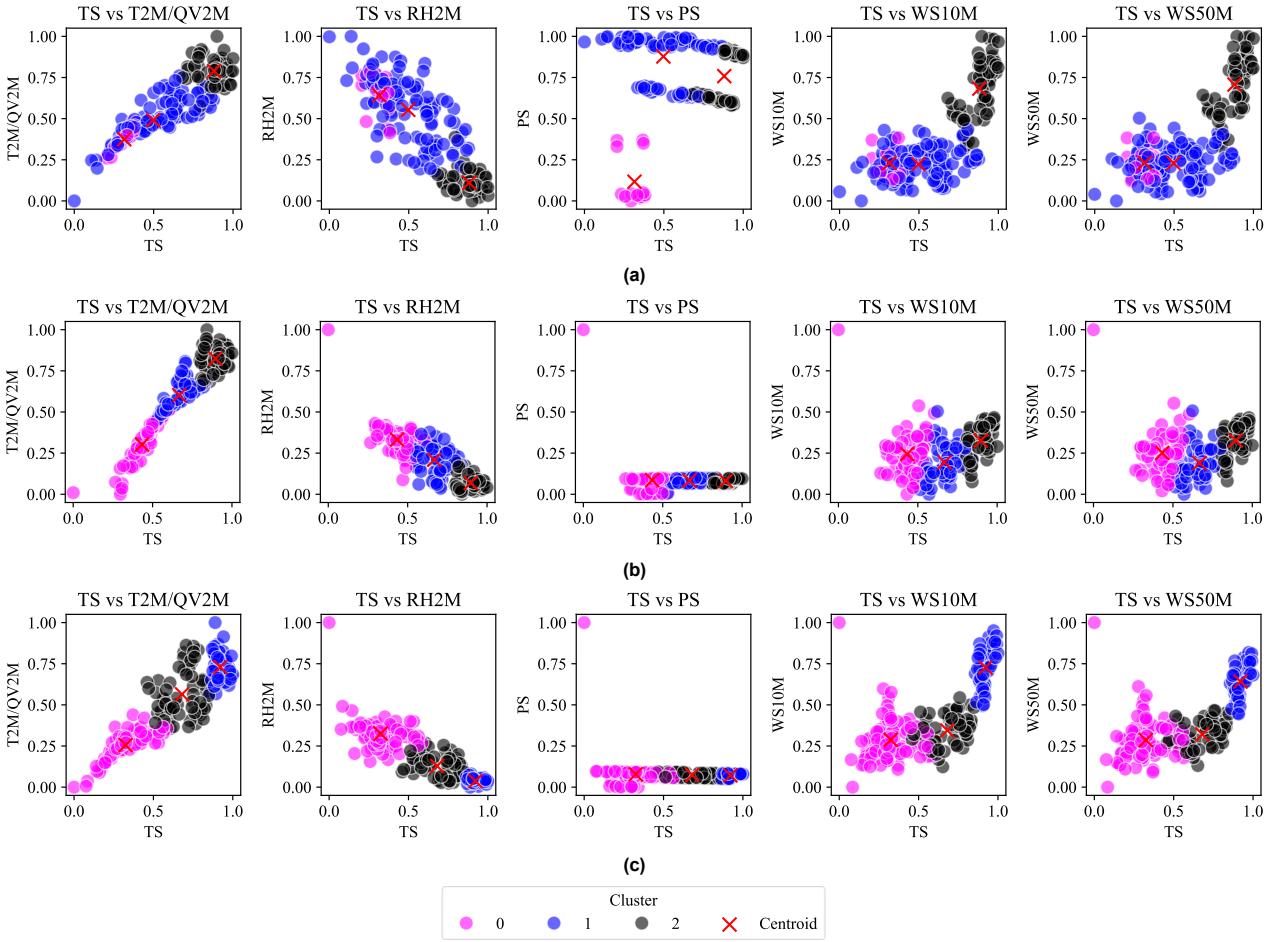


FIGURE 4. Input features clustering visualization (each input feature against TS) on: (a) Nazdik-i Herat station, (b) Shinya station, (c) Torghundi station

1) GRADIENT BOOSTING REGRESSOR (GBR)

The GBR ML algorithm is a very powerful and effective model that forms an ensemble of weak predictive models, usually such as decision trees, into a robust learner in a consecutive manner [50]. Errors correcting by tress of its predecessors, hence improving the accuracy and precision of model with iterations [51]. The main principle of GBR algorithm is to minimize a loss function, which quantifies the change between the actual and predicted values [52,53].

$$F(x) = \sum_{m=1}^{M} \gamma_m \, h_m(x) \tag{2}$$

In the above equation, M indicates the total number of boosting stages, $h_m(x)$ represents the weak learner at stage m, and γ_m is the corresponding learner weight. GBR algorithm scope is to minimize the squared error loss and increasing the accuracy by optimizing the loss within n data points [54]. The impact of each learner is managed by a learning rate η, which balances learning speed against the overfitting risk, which can be defined as [52]:

$$F_m(x) = F_{m-1}(x) + \eta \cdot h_m(x) \tag{3}$$

As loss function for most of the cases, especially for rainfall prediction, MSE is usually used. Where, GBR algorithm adjusts the predictions of the model based on the gradient of the loss function, correcting the model through addition of trees that forecast the residuals or errors of the last ones. The learning rate, which is a hyperparameter, monitors the contribution of trees with final model and is significant for overfitting prohibition. The final GBR model from the above iterative process is a strong predictor that has been shown to perform very well in nonlinear and complicated datasets.

2) HIST GRADIENT BOOSTING REGRESSOR (HGBR)

The HGBR algorithm is an extension of the traditional gradient boosting framework that focuses on speed and efficiency, particularly for large datasets [55]. It operates by constructing a histogram of the feature values which then allows for faster optimization of the loss function, especially when the data

contains continuous variables that would otherwise require more complex computations [55].

$$L(y, F(x)) = \sum_{i=1}^{n} (y_i - F(x_i))^2 \quad (4)$$

where *yi* is the actual values, and $F(x_i)$ are the predicted values. The model iteratively improves predictions over *M* boosting stages, with each stage attempting to correct the errors of the previous stages using the formula:

$$F_m(x) = F_{m-1}(x) + \eta \cdot h_m(x) \quad (5)$$

where $h_m(x)$ is the weak learner at stage *m*, and η is the learning rate (initially set to 0.05). Regularization is integrated through an L_2 penalty on the weights (λ=0.5) and by constraining the growth of the decision trees, allowing a maximum of 40 leaf nodes and requiring at least 15 samples to form a leaf. Further, the HGBR model partitions the continuous input features into discrete bins, transforming the dataset into a grid. This binning process significantly reduces the number of splitting points to consider, thus speeding up the computation. The gradient boosting then proceeds in a similar fashion to GBR, with the model sequentially fitting additional predictors to the residual errors made by the previous predictors. The algorithm optimizes the same loss function, typically MSE for regression problems.

3) RANDOM FOREST REGRESSOR (RFR)

The RFR algorithm uses a set of decision trees to generate a single endless prediction, making it a robust candidate for regression tasks, including rainfall prediction [56]. This algorithm generates a forest of un-correlated trees by utilizing of random selection of input features and data points, producing numerous models from different sub-sets of the data [56]. Moreover, decision trees in the forest are constructed on a different sample, with the overall prediction being the mean of the individual predictions of all these trees [57]. This method of averaging predictions can help in minimizing variance and helps avoid overfitting, which is a common issue with single decision trees. This approach can be mathematically represented as [50,53]:

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} f_t(x) \quad (6)$$

In the above equation, $\hat{y}$ shows the predicted value, *T* represents the number of trees, f_t express the prediction of the t^{th} tree, and *x* indicates the input feature vector. The RFR ML algorithm is well-known for its straightforwardness, easy application, and the capability to handle non-linear relationships without mandating feature scaling, making it a general tool in the field of ML as treasure for tasks of both regressions and classifications.

4) EXTREME GRADIENT BOOSTING REGRESSOR (XGBR)

The XGBR algorithm is an advanced, improved, and efficient implementation of the GB framework [58,59]. It is characterized by its power to address sparse data and its use of a more regularized model formalization to manage, and control the overfittings, which makes it strong, and accurate predictor [58,59]. The XGBR algorithm improves upon the GB technique by using second-order gradients information on the curvature of the loss function to optimize the model. This permits for more accurate and precise updates when adds up more trees and often leads in better performance with fewer number of trees [60]. The formula is given as:

$$\hat{y}^{(t)} = \hat{y}^{(t-1)} + \eta \cdot f_t(x_i) \quad (7)$$

Where, $\hat{y}^{(t)}$ indicates the prediction at each iteration of t, $\hat{y}^{(t-1)}$ represents the prediction from the last iteration, η stands for the learning rate, f_t illustrates the function represented by the tree added at iteration *t*, and x_i represents the feature vector for instance *i*. The performance and the speed of XGBR algorithm makes it a useful model for large-scale datasets, which are usual in climate and weather-related contexts [61].

E. MODELING DEVELOPMENT CONFIGURATION

In the modeling process, the selection of ML models was initially rooted in by using LazyPredict algorithm. Where, it revealed a primary ranked performance on numerous ML models on the dataset. The GBR, HGBR, RFR, and XGBR models were identified most efficient algorithms. The selected models were then considered for development and more fine tuning (Fig. **5**). Consequently, every model set for careful hyperparameter tuning under different settings crucial to increase the model's efficiency. Most common hyperparameters included tree depth, number of trees, learning rate adjustments, and number of iterations. After fine-tuned the models, a comprehensive accuracy evaluation was performed. This evaluation involved error metrics estimation such as mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), and median absolute error (MedAE). In addition, using R-squared (R^2) the goodness of fit was evaluated. The models in this study were not only designed to the specific dataset, but also thoughtfully assessed for the predictive precision, accuracy, efficiency, performance, and reliability.

F. MODELS' PERFORMANCE EVALUATION METRICS

Accuracy, precision, and efficacy of all ML models in rainfall prediction employed in this study is quantitatively evaluated through error metrics. These metrics provide a statistical measure of the performance of models. These metrics are mandatory as they display the predictions' performance in comparison to the observed data, directing the better tunning

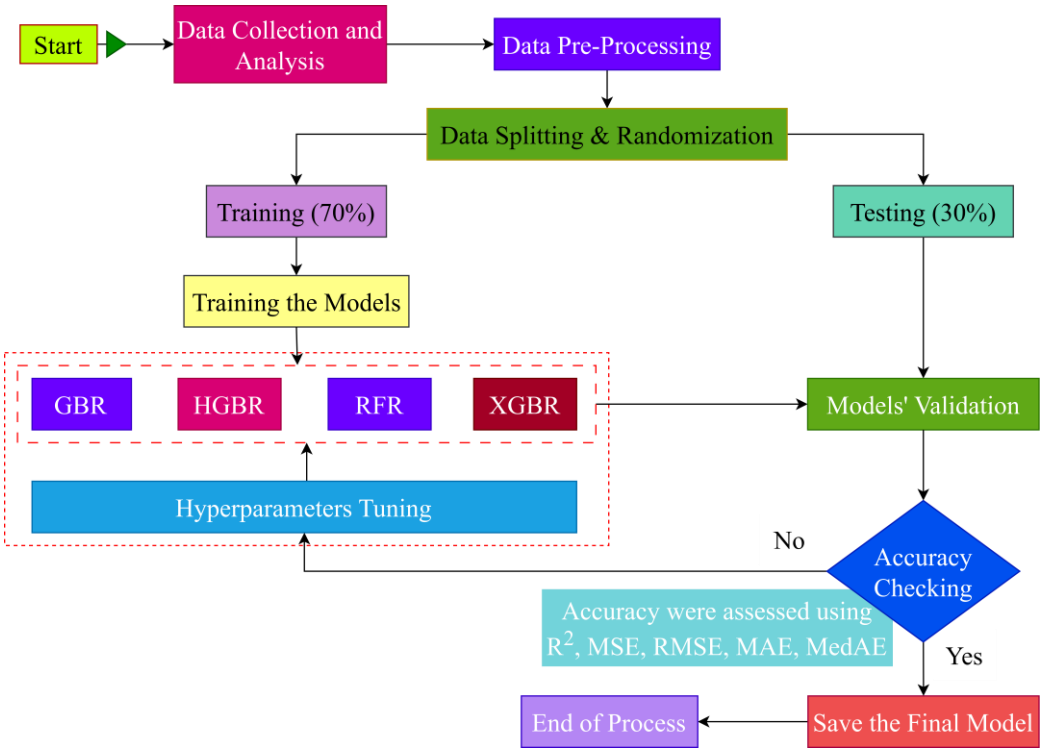


FIGURE 5. Modeling process framework adopted for ML algorithms

of models for increased precision. In this section the main error metrics used in this study were discussed. They are including R^2, MSE, RMSE, MAE, and MedAE. Each of them offers a specific perspective on the predictive ability of models.

(i) R-Squared (R^2)**:** Reflects the proportion of variance in the dependent variable explained by the independent variables, indicating the goodness of fit. A value closer to 1 suggests a model with minimal error in prediction [62,63].

$$R^2 = 1 - \frac{Sum\ of\ squares\ of\ residuals}{Total\ sum\ of\ squares} \tag{8}$$

$$R^2 = 1 - \frac{\sum(yi - \hat{y}\,i)^2}{\sum(yi - \overline{y})^2} \tag{9}$$

(ii) Mean Squared Error (MSE): Estimate the mean squared difference between the observed values and the predicted values. Emphasizing the errors' magnitude. Lower MSE values indicates more accurate predictions [62].

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(yi - \hat{y}i)^2 \tag{10}$$

(iii) Root Mean Squared Error (RMSE): Measures the square root of the average of the squares of the errors, providing insight into the typical size of the errors. Like MSE, lower RMSE values signify more accurate predictions [53,62,64].

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(yi - \hat{y}i)^2} \tag{11}$$

(iv) Mean Absolute Error (MAE): Computes the average of the absolute differences between predicted and actual observations, providing a straightforward measure of prediction error without emphasizing outliers. Lower MAE values reflect more accurate predictions [64].

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|yi - \hat{y}i| \tag{12}$$

(v) Median Absolute Error (MedAE): Determines all absolute differences of the median, providing a strong measure less influenced by skewed data, and outliers. It shows a central point of errors' prediction [65].

$$MedAE = median(|yi - \hat{y}i|) \tag{13}$$

In the above equations y_i is the observed values, $\hat{y}i$ is the predicted values, $\overline{y}$ is the mean of the observed values, and n represents the number of observations. These metrics collaboratively help to assess the accuracy, precision, and effectiveness of each model, helping in the improvement of model's performance.

III. RESULTS REPRESENTATIONS

A. DATA CATEGORIZATION AND RANDOMIZATION

For the analysis and model training using categorization of data, the dataset was split into two sets (training and testing). This division was vital to validate the ML models' predictive power and ensure that they can be generalized very well to new, and unseen data. This division was specifically 70 % of the data for training phase, which is used to train all the four selected ML models, while the rest 30 % of the dataset was allocated for the testing phase to evaluate the performance of each model. All model development and training were conducted in Python 3.13.2 using the Scikit-learn library [66], within JupyterLab v4.3.4 on an Anaconda environment, and visualizations were generated using Pandas and Matplotlib libraries. Further, to confirm that the dataset in both training and testing sets is representing the overall distribution, dataset was randomized (or shuffled) before the dataset division. Randomization or shuffling of data is a vital task to avoid any bias that could impact the outcomes of the models. This process grantees that the data allocated for both training and testing sets are samples of random from the whole dataset, thus providing integrity and randomness of the overall data distribution. Further, the shuffling was applied using a random-like number generator in the code environment, which is a feature generally available in many data processing and ML libraries (such as python that were used in this study). This technique ensures that the contribution of data points to the training and testing sets is sufficiently random and representative of overall dataset, where this approach reinforces the robustness of the analysis in the subsequent steps.

B. MODELS TUNNING

In the phase of tunning models in this study, the first calibration on the hyperparameters was carefully conducted using the datasets from each station. The tuning process of hyperparameters for each model required highly sensitive manual adjustments, with an emphasis on optimizing model parameters in line with main evaluation's metrics of R^2, MSE, RMSE, MAE, and MedAE. This foundational step was indispensable to save an optimal baseline performance for the models. For getting satisfactory tuning of ML models applied for the first station's dataset, these refined parameters were chosen to apply firmly across the rest of stations' datasets to thoughtfully evaluate the models' power, performance, and their ability to generalize.

Further, in this phase, an unexpected challenge was encountered during the process: the process of outlier removal, a standard data pre-processing step expected to increase the model's performance but unexpectedly led to reducing the accuracy of models. This phenomenon was an unusual issue, since outliers are commonly considered as noise that can distort model performance at each phase (training and testing). However, in this specific context, it was decided that the outliers might have inherent patterns that reflect the actual rainfall events (such as extreme conditions), which could have been mistaken as anomalies. Therefore, a careful decision was made to keep the data points (outliers) as they are. As outlined in Table 4, the removal of outliers resulted in a substantial decline in model performance across all stations, with R^2 values dropping as low as 0.17 and RMSE increasing by nearly double in some cases. This indicated that the removed points were not random noise but potentially represented actual extreme rainfall events. Further evidence of this is seen in the correlation heatmaps (Fig. 3), where strong correlations between the input variables and rainfall suggest that these extreme values were supported by coinciding anomalies in the meteorological predictors. Therefore, retaining these values ensured a more accurate reflection of real-world rainfall behavior, especially under extreme conditions. This issue indicates that, in specific cases, the outliers in the satellite data inputs could hold vital clues closely consistent with the observational data. Thereby, potentially validating the reliability of the satellite data, and concentrating their importance in regions where observational data may be sparse and incomplete due to lack of monitoring observational gauges. The decision not to eliminate the outliers emphasizes the complex and sometimes unusual nature of modeling the phenomena of real-world and proves the significance of an adopted method to data pre-processing in realm of ML.

C. MODELS PERFORMANCE FOR NAZDIK-I HERAT STATION

TABLE 4. Impact of outlier removal on the models' performance during the testing phase

Station	Model	R^2 (Before)	RMSE (Before)	R^2 (After)	RMSE (After)
Nazdik-i Herat	GBR	0.80	11.15	0.21	21.69
	HGBR	0.83	10.25	0.17	22.16
	RFR	0.74	12.61	0.41	17.20
	XGBR	0.75	11.14	0.37	17.40
Shinya	GBR	0.77	8.34	0.62	12.34
	HGBR	0.76	9.51	0.36	11.96
	RFR	0.79	8.13	0.52	12.58
	XGBR	0.72	11.26	0.20	19.30
Torghundi	GBR	0.82	8.17	0.22	12.37
	HGBR	0.80	8.54	0.52	11.40
	RFR	0.80	8.54	0.28	11.90
	XGBR	0.82	8.12	0.40	12.27

For the Nazdik-i Herat station, all models showed a decline in performance from training to testing phases, which is expected in real-world prediction tasks. Among the models, RFR model exhibited the smallest gap between training and testing phases with an R^2 of 0.88 to 0.74, indicated good generalization. However, despite this, its overall accuracy was lower compared to HGBR model, which obtained a higher R^2 values in both training (0.90) and testing (0.83) phases. This made the HGBR model not only consistent but also more accurate than RFR model. While GBR and XGBR models achieved high training R^2 values of 0.95 and 0.93 respectively, their testing

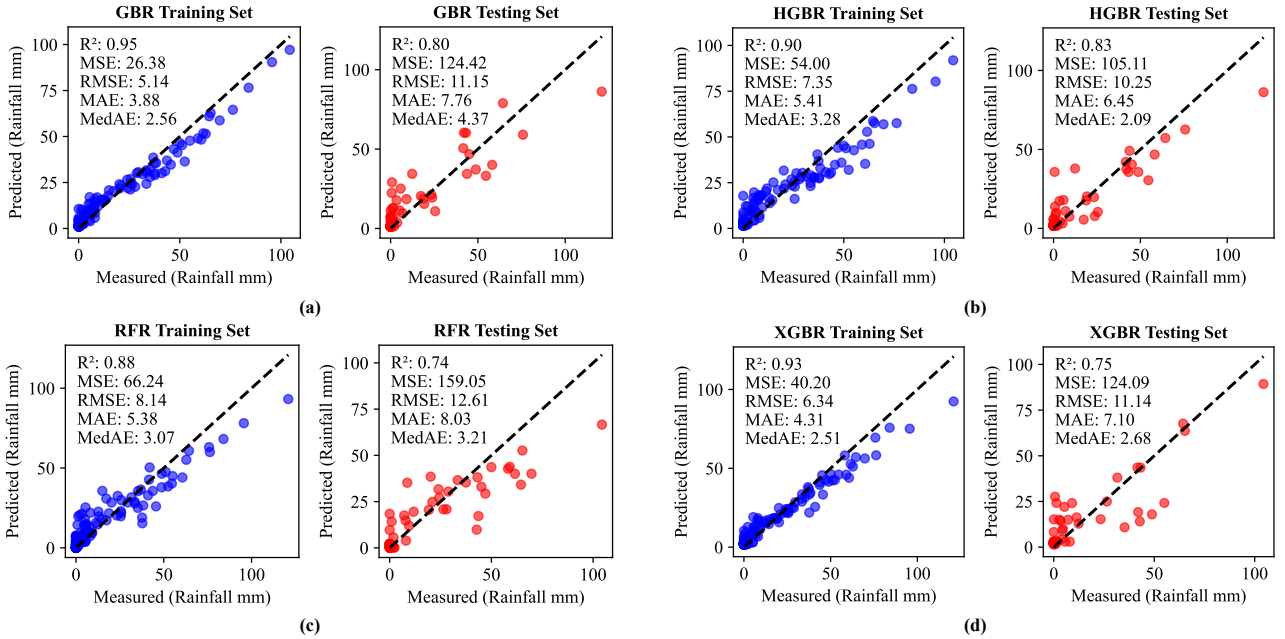


FIGURE 6. Scatter Plots for observational training and testing data sets vs Predicted for the station 'Nazdik-i Herat'. (a) GBR Model, (b) HGBR model, (c) RFR model, and (d) XGBR model

R^2 values dropped to 0.80 and 0.75, revealing a stronger tendency toward overfitting. The HGBR model, on the other hand, balanced generalization and accuracy effectively, with a relatively small drop between phases and lower RMSE and MAE values than RFR model. Therefore, HGBR model was considered the most reliable and high-performing model at this station (Fig. 6).

D. MODELS PERFORMANCE FOR SHINYA STATION

The assessment of all models in the Shinya station demonstrated that GBR and XGBR models showed the highest training R^2 values of 0.94 and 0.95, respectively, but both revealed notable drop in testing phase performance, with R^2 dropping to 0.77 for GBR model and 0.72 for XGBR model, where this flagged the concern of overfitting issues. In the other hand, the RFR model displayed the lower difference in R^2 between training (0.78) and testing (0.79) compared to GBR and XGBR models, suggested stable generalization. In contrast,

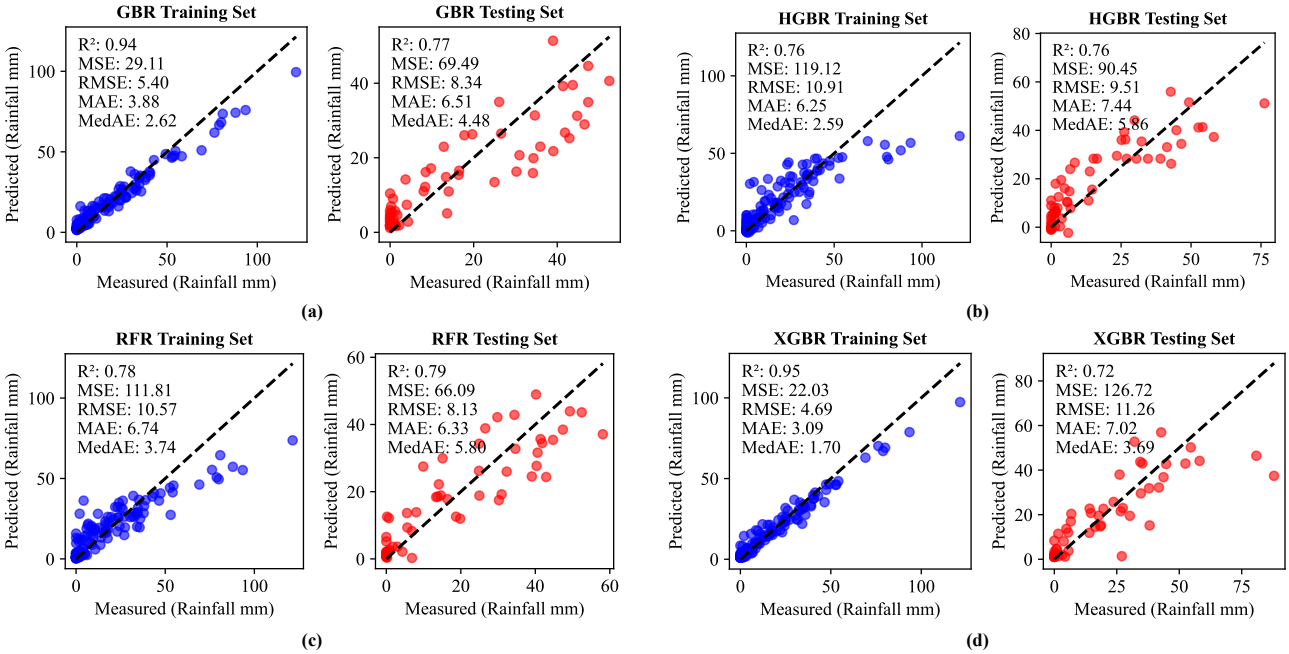


FIGURE 7. Scatter Plots for observational training and testing data sets vs Predicted for the station 'Shinya. (a) GBR Model, (b) HGBR model, (c) RF model, and (d) XGBR model

the HGBR model showed R^2 values of 0.76 for both training and testing phases respectively with zero differences. The HGBR model achieved significantly good accuracy, with RMSE of 10.91 and 9.51 and MAE of 6.25 and 7.44 for training and testing phases respectively (Fig. 7). Compared to other models, HGBR provided a more favorable balance between generalization and prediction quality. Therefore, the findings at this station were consistent with the performance at Nazdik-i Herat station.

E. MODELS PERFORMANCE FOR TORGHUNDI STATION

The results for model performance at Torghundi Station (Fig. 8), indicated that all models performed well, with different levels of accuracy and generalization. The XGBR model achieved the highest R^2 in training (0.95) and maintained a strong R^2 of 0.82 in testing. However, its RMSE increased from 4.35 in training to 8.12 in testing, and MAE from 2.89 to 5.37. GBR model followed a similar trend, with R^2 of 0.94 in the training phase and 0.82 in the testing phase, and RMSE rising from 4.86 to 8.17, and MAE from 3.66 to 6.11. The RFR model showed the smallest drop in R^2 (from 0.86 to 0.80), suggested stable generalization. However, its absolute accuracy was lower than HGBR model, as seen from higher RMSE values (7.13 and 8.54) and MAE values (4.70 and 5.69) for training and testing phases respectively. HGBR model showed strong predictive consistency, with R^2 of 0.92 and 0.80, RMSE of 5.51 and 8.54, and MAE of 3.97 and 5.97 for training and testing phases respectively. While RFR model performed well in terms of generalization, and XGBR model showed high training accuracy, HGBR model offered a better balance of generalization and accuracy. This performance was consistent with its results at other stations, reinforcing HGBR as the most reliable and robust model for rainfall prediction at Torghundi Station.

F. EVALUATION OF MODELS CONSISTENCY AND PRECISION

An overall evaluation of the four models GBR, HGBR, RFR, and XGBR across all three stations Nazdik-i Herat, Shinya, and Torghundi is presented in Fig. 9. The spider plots revealed that GBR and XGBR models consistently showed the best training accuracy but experienced a huge decline in testing performance, exposing the concern of overfitting. RFR model maintained relatively stable performance between training and testing phases but exhibited the highest RMSE and MAE values, making it the least accurate model overall. In contrast, the HGBR model consistently demonstrated a balanced performance across both phases (training and testing), maintained

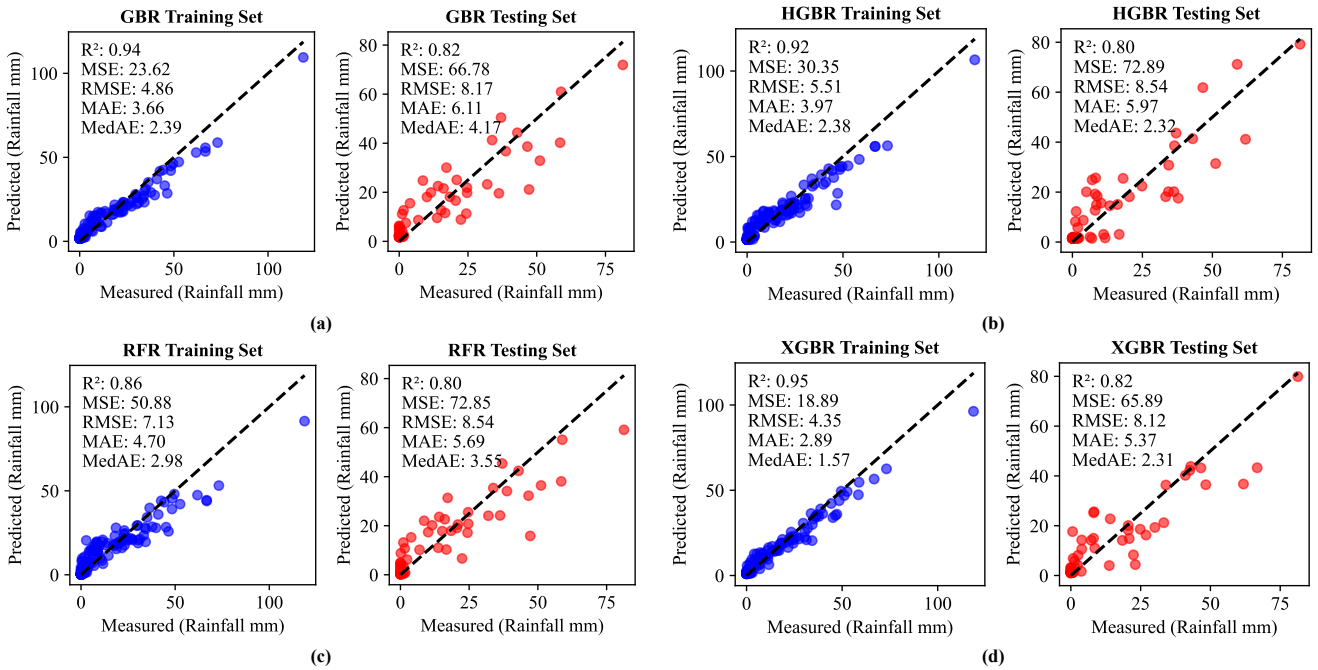


FIGURE 8. Scatter Plots for observational training and testing data sets vs Predicted for the station 'Torghundi'. (a) GBR Model, (b) HGBR model, (c) RF model, and (d) XGBR model

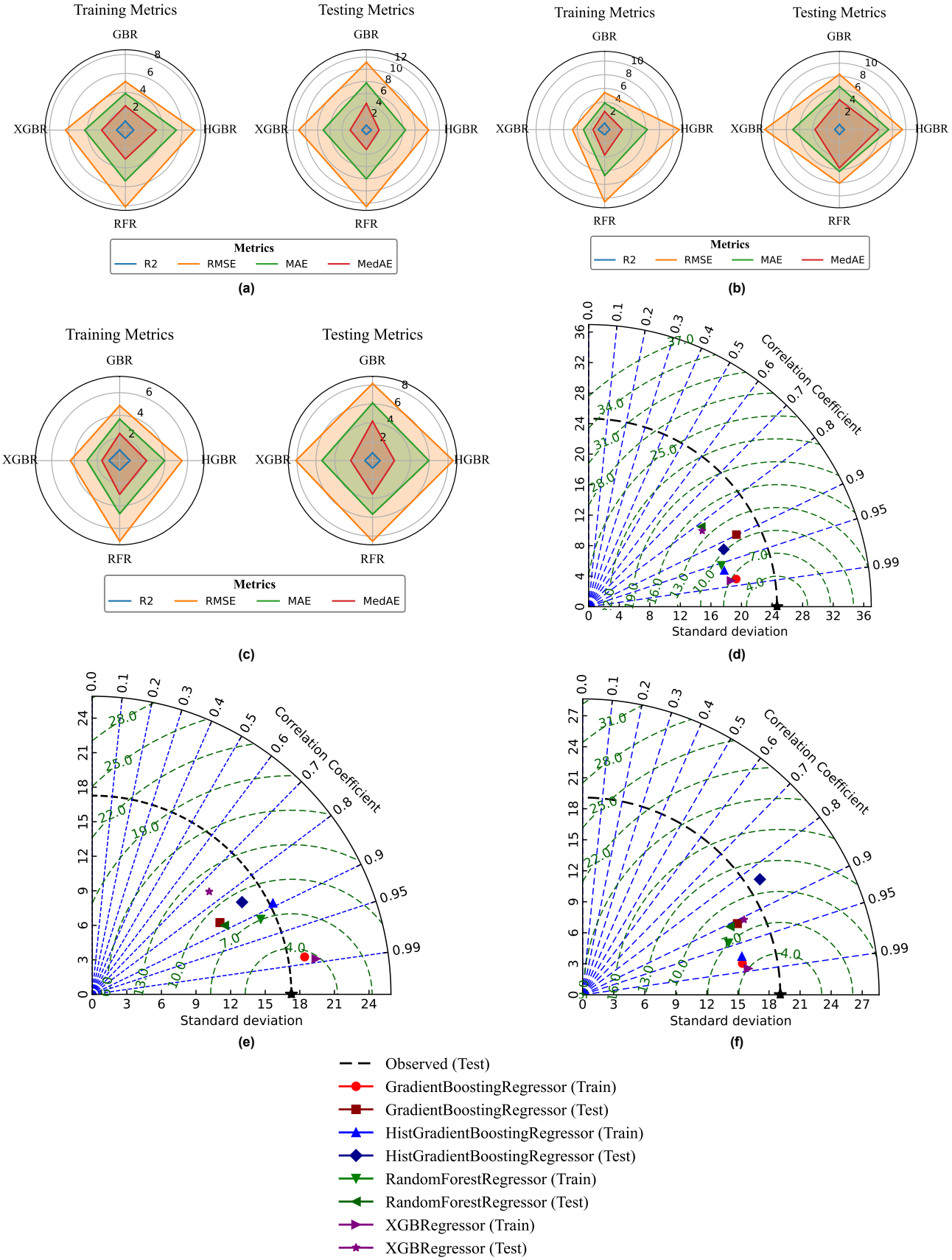


FIGURE 9. Comparative Performance Metrics of ML Models; (a) models performance using spider plot for Nazdik-i Herat station, (b) models' performance using spider plot for Shinya Station, (c) models' performance using spider plot for Torghundi Station, (d) models performance using Taylor diagram for Nazdik-i Herat station, (e) models performance using Taylor diagram for Shinya Station, (f) models performance using Taylor diagram for Torghundi Station.

moderate-to-high R^2 values with lower error metrics, reflect-

ing its strong generalization ability. The Taylor diagrams supported these findings: while GBR and XGBR models aligned closely with the observed standard deviation during training. However, their correlation sharply declined during the testing phase. RFR model maintained fair correlation but deviated in standard deviation, especially at Shinya and Torghundi stations. HGBR model stood out as the most consistent model, with minimal variation between training and testing in both correlation and standard deviation across all stations with respect to observed data. Collectively, these findings confirmed that HGBR model offered the best trade-off between accuracy and generalization, making it the most reliable and robust model for monthly rainfall prediction in data-scarce environments.

IV. DISCUSSION

This study addressed the challenge of monthly rainfall prediction in data-scarce regions by integrating satellite-obtained meteorological data inputs with observational gauge-based rainfall. Among four advanced ML models (GBR, HGBR, RFR, and XGBR), the HGBR model consistently showed the best performance. It achieved R^2 values of 0.90, 0.76, and 0.92 in training phase and 0.83, 0.76, and 0.80 in testing phase for three distinct stations Nazdik-i Herat, Shinya, and Torghundi, respectively. For this model the RMSE values ranged from 5.51 to 10.91 and MAE between 3.97 and 7.44.

Results of this study are aligned with those of study [67], which reported very high R^2 values (up to 0.9998) using Boosted Decision Tree Regression (BDTR) [67]. However, their model lacked multi-station validation and satellite integration. In contrast, our methodology combined satellite data input features with observational data output and extending the applicability of ML to real-world, and low-data environments. Study [68], applied GBR model for spatial downscaling (R^2 = 0.98, RMSE = 9.63 mm) [68], which is consistent with our model's strength in structured learning, but their work focused on spatial interpolation, not time-based forecasting.

Study [69], evaluated six models and found RF and ANN models yielded mean RMSE more than 40.0 [69]. These results are notably lower than ours, where our HGBR's RMSE remained below 11.0 across all stations. Our study further differs by introducing a clustering-enhanced ML method, where K-means clustering was integrated to optimize feature grouping.

In Study [70], ELM model yielded top performance for seasonal prediction with R^2 of 0.998 and RMSE ≈ 3.8% [70]. While these results are strong, the study focused only on two seasonal windows using long historical inputs. Our study offered monthly resolution with multi-year, multivariate satellite input features, extended the prediction potential to operational monitoring. Study [71], found that SVM as the best performed model with R of 0.863, and MAE of 2.728 on daily data in Vietnam [71], aligned with our findings on model stability. Yet, their work was limited to single-site testing, whereas ours was validated over three distinct stations.

Study [72], applied LSTM model with six meteorological features and reported 76% accuracy [72]. However, the study did not explain how this accuracy was calculated, leaving ambiguity around the evaluation method. This limits interpretability and comparison. Our results differ by offering transparent and reproducible metrics across stations. Study [73], proposed a Wavelet-coupled Multi-order Time Lagged Neural Network (WMTLNN). The study obtained R^2 of 0.95 and RMSE of 0.17 at Chillas station [73]. These values are aligned with our best HGBR results, but their study used univariate inputs only. In contrast, our integration of satellite-derived multivariate inputs improves real-world applicability.

A novel contribution of our study was the use of satellite data as inputs and a cross-validation reference for extreme values. When outliers were removed, R^2 dropped significantly (e.g., for HGBR from 0.83 to 0.17 at Nazdik-i Herat station), and RMSE increased sharply, confirmed that these values were not noise but actual extremes. This finding is supported by the strong correlation patterns between satellite input features and gauge rainfall as discussed in earlier sections. These results suggested a new scientific approach for validating extreme events in observational data using remote sensing data. This provides a dual-layer validation system that has not been reported previously.

In addition, this study presented a practical breakthrough: the ability to predict monthly rainfall without the need for installing and maintaining observational gauge stations. Given the high cost and logistical challenges of deploying such infrastructure in remote or conflict-prone areas. The shown accuracy of satellite-fed models such as HGBR model offered a scalable, low-cost alternative for hydrological monitoring.

V. PRACTICAL IMPLICATIONS

The practical implications of this study on rainfall prediction using ML models in the HMB have many aspects, particularly related to the water resource management, and agricultural sector within this arid to semi-arid region. Increased predictive accuracy, and models' efficiency for rainfall predictions has a direct impact for farming communities, where irrigation depends heavily on timely and accurate rainfall predictions. By promoting more reliable, and precise rainfall predictions, our models aim in optimizing irrigation schedules and minimizing water wastage, for that reason enhancing crop yields and supporting sustainable agricultural practices. This is vital in regions like Afghanistan, where agriculture forms the economic foundation and food security is tightly connected to the seasonal variability of rainfall.

Moreover, the models that demonstrated predictive capabilities can significantly help in risk management, and disaster preparedness. For early warning systems, and emergency response planning, especially in reducing the impacts of severe floods, and drought conditions accurate predictions of rainfall

are of utmost essential. Accurate predictions that maintained advance knowledge allows for better strategic planning and allocation of resources, which is important in mitigating the detrimental impacts of extreme weather events on both environment, and human populations. Further, policymaking, and infrastructure development have implications using the findings of this study. With more accurate, precise, and efficient predictions, policymakers can make informed decisions about water storage and distribution infrastructures. For example, dams and reservoirs construction, to manage the anticipated water flow more effectively. This strategic planning is very significant for providing water security in the face of climate change and enhancing water demand due to the increasing rate of population growth and industrial use.

In a more expanded landscape, the successful application of advanced ML models for meteorological context, specifically rainfall prediction in Afghanistan can serve as a method for similar arid and semi-arid regions globally. The methodology that was used in this study and outcomes can maintain a framework for maximizing advanced computational techniques to increase weather forecasting and climate analysis. Therefore, supporting worldwide efforts towards sustainable development, and climate resilience. Thus, the combination of technology and conventional meteorological approaches represents a vital step forward in our method to comprehend and managing the environmental challenges raised by the contemporary world.

VI. FUTURE RESEARCH DIRECTIONS

The results obtained in the current research paved the way for numerous future research directions. (i) the unification of real-time data into the ML models for the purpose of increasing their responsiveness in altering weather patterns. The integration of real-time data could allow for dynamic updates to the models. This can potentially enhance accuracy and make them stronger against the unpredictability posed by climate change. In addition, to include other weather-related variables, expanding the scope of the models could maintain a more comprehensive perspective of the climatic conditions. Therefore, improving the precision, accuracy, and efficiency of the rainfall predictions. (ii) investigating the application of integrated advanced ML models that synthesize the strengths of various algorithms could also be beneficial. For instance, integrating the DL approach with conventional ML models might grab complex non-linear relationships more effectively. In turn, increasing the overall predictive power. Furthermore, carrying out some comparative studies across various regions with same climatic conditions would aid in validating the generalizability of the models and improving them for broader applications. (iii) the development of models that can predict extreme weather events with high precision. Given the enhancing frequency of such events due to global warming, those models that can effectively predict extreme conditions are vital for effective response strategies, and disaster preparedness. (iv) interdisciplinary partnership among local government officials, meteorologists, and data scientists could lead to innovative solutions and ensure that the models are effectively synthesized into resource management systems, and local weather forecasting. This coordinative technique would not only increase the technical aspects of the models but also granted that they satisfy the practical requirements of the communities they are designed to serve. (v) finally, future research may also explore the integration of causal learning and spatio-temporal graph-based models, as recently revealed in the literature [74–76]. These methods could help uncover underlying causal relationships between satellite predictors and rainfall dynamics, offering additional interpretability and robustness in climate-sensitive prediction tasks.

VII. CONCLUSION

This study addressed the challenge of predicting monthly rainfall in data-scarce regions by evaluating four advanced ML models (GBR, HGBR, RFR, and XGBR), using satellite-derived meteorological inputs and gauge-based observational rainfall data from three stations in Afghanistan. The assessment showed clear differences in model behavior, particularly in accuracy and generalization between training and testing phases. Findings revealed that the HGBR model consistently achieved strong and stable performance across all stations. At Nazdik-i Herat, the HGBR model maintained R^2 of 0.90 (training) and 0.83 (testing), while at Shinya it preserved R^2 at 0.76 in both phases. At Torghundi, it reached 0.92 and 0.80 for training and testing, respectively. Corresponding RMSE and MAE values supported this reliability, remained moderate and closely aligned between training and testing phases. In contrast, GBR and XGBR models, although delivering high R^2 during training phase (up to 0.95), showed big drops in testing phase with an R^2 down to 0.72 at Shinya. Alongside rising RMSE and MAE values, raised the concern of overfitting. RFR model, despite having the smallest R^2 gaps (e.g., 0.78 to 0.79 at Shinya station), suffered from the highest error metrics across most stations and lower overall accuracy compared to HGBR model in both training and testing. These comparative results confirmed that HGBR model offered the best overall trade-off between precision and generalization, making it the most reliable model for monthly rainfall prediction in the region. Scientifically, the study demonstrated that satellite data combined with robust ML models can produce accurate rainfall estimates even in regions lacking observational monitoring gauges. Practically, the framework provided a cost-effective and scalable solution for rainfall prediction.

ACKNOWLEDGMENT

The authors would like to acknowledge the funding supported by KFUPM University for the PhD's degree scholarship. Additionally, to acknowledge the support from the Ministry of Energy and Water-Afghanistan for providing observational data.

REFERENCES

[1] E.A. Abioye, O. Hensel, T.J. Esau, O. Elijah, M.S.Z. Abidin, A.S. Ayobami, O. Yerima, A. Nasirahmadi, Precision Irrigation Management Using Machine Learning and Digital Farming Solutions, AgriEngineering 4 (2022) 70–103. https://doi.org/10.3390/agriengineering4010006.

[2] M. Sit, B.Z. Demiray, Z. Xiang, G.J. Ewing, Y. Sermet, I. Demir, A comprehensive review of deep learning applications in hydrology and water resources, Water Sci. Technol. 82 (2020) 2635–2670. https://doi.org/10.2166/wst.2020.369.

[3] S.Q. Salih, A. Sharafati, I. Ebtehaj, H. Sanikhani, R. Siddique, R.C. Deo, H. Bonakdari, S. Shahid, Z.M. Yaseen, Integrative stochastic model standardization with genetic algorithm for rainfall pattern forecasting in tropical and semi-arid environments, Hydrol. Sci. J. 65 (2020) 1145–1157. https://doi.org/10.1080/02626667.2020.1734813.

[4] P.J. Webster, J. Jian, Environmental prediction, risk assessment and extreme events: adaptation strategies for the developing world, Philos. Trans. A. Math. Phys. Eng. Sci. 369 (2011) 4768–4797. https://doi.org/10.1098/rsta.2011.0160.

[5] S. Ahmadzai, A. McKinna, Afghanistan electrical energy and trans-boundary water systems analyses: Challenges and opportunities, Energy Reports 4 (2018) 435–469. https://doi.org/10.1016/j.egyr.2018.06.003.

[6] A. Jones, J. Kuehnert, P. Fraccaro, O. Meuriot, T. Ishikawa, B. Edwards, N. Stoyanov, S.L. Remy, K. Weldemariam, S. Assefa, AI for climate impacts: applications in flood risk, Npj Clim. Atmos. Sci. 6 (2023). https://doi.org/10.1038/s41612-023-00388-1.

[7] W. Al, G. Orking, O. Clima, Climate change and food security: a framework document, FAO Rome (2008).

[8] M.A. Srivastava, A. Rastogi, P. Kaushik, A COMPREHENSIVE STUDY ON THE USE OF VARIOUS ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING TECHNIQUES IN THE FIELD OF WEATHER FORECASTING, Int. Res. J. Mod. Eng. Technol. Sci. 5 (2023) 10.

[9] L. Diop, S. Samadianfard, A. Bodian, Z.M. Yaseen, M.A. Ghorbani, H. Salimi, Annual Rainfall Forecasting Using Hybrid Artificial Intelligence Model: Integration of Multilayer Perceptron with Whale Optimization Algorithm, Water Resour. Manag. (2020). https://doi.org/10.1007/s11269-019-02473-8.

[10] S. Salcedo-Sanz, P. Ghamisi, M. Piles, M. Werner, L. Cuadra, A. Moreno-Martínez, E. Izquierdo-Verdiguier, J. Muñoz-Marí, A. Mosavi, G. Camps-Valls, Machine learning information fusion in Earth observation: A comprehensive review of methods, applications and data sources, Inf. Fusion 63 (2020) 256–272. https://doi.org/10.1016/j.inffus.2020.07.004.

[11] Z.H. Doost, Z.M. Yaseen, Allocation of reservoirs sites for runoff management towards sustainable water resources: Case study of Harirud River Basin, Afghanistan, J. Hydrol. 634 (2024) 131042. https://doi.org/10.1016/j.jhydrol.2024.131042.

[12] Z.H. Doost, Z.M. Yaseen, The impact of land use and land cover on groundwater fluctuations using remote sensing and geographical information system: Representative case study in Afghanistan, Environ. Dev. Sustain. (2023) 1–24. https://doi.org/10.1007/s10668-023-04253-2.

[13] Z.H. Doost, S. Chowdhury, A.M. Al-Areeq, I. Tabash, G. Hassan, H. Rahnaward, A.R. Qaderi, Development of intensity–duration–frequency curves for Herat, Afghanistan: enhancing flood risk management and implications for infrastructure and safety, Nat. Hazards (2024). https://doi.org/10.1007/s11069-024-06730-x.

[14] I.A. Nengroo, A. Shah, Irrigation Potential and Levels of Agricultural Development in Afghanistan, Univ. Kashmir Dr. Thesis (2012).

[15] R. KHALİLY, Evaluation of Climate Change on Agricultural Production in Afghanistan, Eurasian J. Agric. Res. 6 (2022) 91–100.

[16] B.M. Hashim, A.N.A. Alnaemi, B.A. Hussain, S.A. Abduljabbar, Z.H. Doost, Z.M. Yaseen, Statistical downscaling of future temperature and precipitation projections in Iraq under climate change scenarios, Phys. Chem. Earth, Parts A/B/C 135 (2024) 103647. https://doi.org/10.1016/j.pce.2024.103647.

[17] V. Aich, N. Akhundzadah, A. Knuerr, A. Khoshbeen, F. Hattermann, H. Paeth, A. Scanlon, E. Paton, Climate Change in Afghanistan Deduced from Reanalysis and Coordinated Regional Climate Downscaling Experiment (CORDEX)—South Asia Simulations, Climate 5 (2017) 38. https://doi.org/10.3390/cli5020038.

[18] M. Ghulami, Assessment of climate change impacts on water resources and agriculture in data-scarce Kabul basin, Afghanistan, (2017).

[19] C. Huntingford, E.S. Jeffers, M.B. Bonsall, H.M. Christensen, T. Lees, H. Yang, Machine learning and artificial intelligence to aid climate change research and preparedness, Environ. Res. Lett. 14 (2019) 124007. https://doi.org/10.1088/1748-9326/ab4e55.

[20] A.Y. Sun, B.R. Scanlon, How can Big Data and machine learning benefit environment and water management: a survey of methods, applications, and future directions, Environ. Res. Lett. 14 (2019) 073001. https://doi.org/10.1088/1748-9326/ab1b7d.

[21] M.G. Schultz, C. Betancourt, B. Gong, F. Kleinert, M. Langguth, L.H. Leufen, A. Mozaffari, S. Stadtler, Can deep learning beat numerical weather prediction?, Philos. Trans. R. Soc. A Math. Phys. Eng. Sci. 379 (2021) 20200097. https://doi.org/10.1098/rsta.2020.0097.

[22] E. Hernández, V. Sanchez-Anguix, V. Julian, J. Palanca, N. Duque, Rainfall Prediction: A Deep Learning Approach, in: Hybrid Artif. Intell. Syst. 11th Int. Conf. HAIS 2016, Seville, Spain, April 18-20, 2016, Proc. 11, Springer, 2016: pp. 151–162. https://doi.org/10.1007/978-3-319-32034-2_13.

[23] M. Jamei, M. Ali, A. Malik, M. Karbasi, P. Rai, Z.M. Yaseen, Development of a TVF-EMD-based multi-decomposition technique integrated with Encoder-Decoder-Bidirectional-LSTM for monthly rainfall forecasting, J. Hydrol. 617 (2023) 129105. https://doi.org/10.1016/j.jhydrol.2023.129105.

[24] K. Abhishek, A. Kumar, R. Ranjan, S. Kumar, A rainfall prediction model using artificial neural network, in: 2012 IEEE Control Syst. Grad. Res. Colloq., IEEE, 2012: pp. 82–87. https://doi.org/10.1109/ICSGRC.2012.6287140.

[25] A.Y. Barrera-Animas, L.O. Oyedele, M. Bilal, T.D. Akinosho, J.M.D. Delgado, L.A. Akanbi, Rainfall prediction: A comparative analysis of modern machine learning algorithms for time-series forecasting, Mach. Learn. with Appl. 7 (2022) 100204. https://doi.org/10.1016/j.mlwa.2021.100204.

[26] M. King, B. Sturtewagen, Making the most of Afghanistan's river basins, Oppor. Reg. Coop. (2010).

[27] I. Leao, M. Ahmed, A. Kar, Jobs from Agriculture in Afghanistan, Washington, DC: World Bank, 2018. https://doi.org/10.1596/978-1-4648-1265-1.

[28] R. Jurenas, Agriculture in Afghanistan and neighboring Asian countries, Asian Econ. Polit. Issues 8 (2003) 191.

[29] A.J. Muradi, I. Boz, The contribution of Agriculture Sector in the Economy of Afghanistan, Int. J. Sci. Res. Manag. 6 (2018) 750–755. https://doi.org/10.18535/ijsrm/v6i10.em04.

[30] Q. Aliyar, S. Dhungana, S. Shrestha, Spatio-temporal trend mapping of precipitation and its extremes across Afghanistan (1951–2010), Theor. Appl. Climatol. 147 (2022) 605–626. https://doi.org/10.1007/s00704-021-03851-2.

[31] I. Qutbudin, M.S. Shiru, A. Sharafati, K. Ahmed, N. Al-Ansari, Z.M. Yaseen, S. Shahid, X. Wang, Seasonal Drought Pattern Changes Due to Climate Variability: Case Study in Afghanistan, Water 11 (2019) 1096. https://doi.org/10.3390/w11051096.

[32] A. McNally, J. Jacob, K. Arsenault, K. Slinski, D.P. Sarmiento, A. Hoell, S. Pervez, J. Rowland, M. Budde, S. Kumar, C. Peters-Lidard, J.P. Verdin, A Central Asia hydrologic monitoring dataset for food and water security applications in Afghanistan, Earth Syst. Sci. Data 14 (2022) 3115–3135. https://doi.org/10.5194/essd-14-3115-2022.

[33] A. Azzam, W. Zhang, F. Akhtar, Z. Shaheen, A. Elbeltagi, Estimation of green and blue water evapotranspiration using machine learning algorithms with limited meteorological data: A case study in Amu Darya River Basin, Central Asia, Comput.

Electron. Agric. 202 (2022) 107403. https://doi.org/10.1016/j.compag.2022.107403.

[34] N. Mohammad, Integrated Assessment of Climate Change Impact on Drought Severity, Water Sustainability, and Agricultural Potential in Afghanistan, (2023). https://ci.nii.ac.jp/naid/500002559821/.

[35] NASA POWER Project, POWER | Data Access Viewer, (2023). https://power.larc.nasa.gov/data-access-viewer/.

[36] M.S. Ali, M.K. Islam, A.A. Das, D.U.S. Duranta, M.F. Haque, M.H. Rahman, A Novel Approach for Best Parameters Selection and Feature Engineering to Analyze and Detect Diabetes: Machine Learning Insights, Biomed Res. Int. 2023 (2023). https://doi.org/10.1155/2023/8583210.

[37] A. Zhang, L. Xing, J. Zou, J.C. Wu, Shifting machine learning for healthcare from development to deployment and from models to data, Nat. Biomed. Eng. 6 (2022) 1330–1345. https://doi.org/10.1038/s41551-022-00898-y.

[38] V. Gudivada, A. Apon, J. Ding, Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations, Int. J. Adv. Softw. 10 (2017) 1–20.

[39] S. Seo, A review and comparison of methods for detecting outliers in univariate data sets, (2006). http://d-scholarship.pitt.edu/id/eprint/7948.

[40] H. Aguinis, R.K. Gottfredson, H. Joo, Best-Practice Recommendations for Defining, Identifying, and Handling Outliers, Organ. Res. Methods 16 (2013) 270–301. https://doi.org/10.1177/1094428112470848.

[41] K. Singh, S. Upadhyaya, Outlier detection: applications and techniques, Int. J. Comput. Sci. Issues 9 (2012) 307.

[42] V. Barnett, PRINCIPLES AND METHODS FOR HANDLING OUTLIERS IN DATA SETS, in: Stat. Methods Improv. Data Qual., Elsevier, 1983: pp. 131–166. https://doi.org/10.1016/B978-0-12-765480-5.50012-6.

[43] G. Bogale Begashaw, Y. Berihun Yohannes, Review of Outlier Detection and Identifying Using Robust Regression Model, Int. J. Syst. Sci. Appl. Math. 5 (2020) 4. https://doi.org/10.11648/j.ijssam.20200501.12.

[44] M. Hubert, E. Vandervieren, An adjusted boxplot for skewed distributions, Comput. Stat. Data Anal. 52 (2008) 5186–5201. https://doi.org/10.1016/j.csda.2007.11.008.

[45] D. Singh, B. Singh, Investigating the impact of data normalization on classification performance, Appl. Soft Comput. 97 (2020) 105524. https://doi.org/10.1016/j.asoc.2019.105524.

[46] H. Henderi, Comparison of Min-Max normalization and Z-Score Normalization in the K-nearest neighbor (kNN) Algorithm to Test the Accuracy of Types of Breast Cancer, IJIIS Int. J. Informatics Inf. Syst. 4 (2021) 13–20. https://doi.org/10.47738/ijiis.v4i1.73.

[47] R. Madurai Elavarasan, R. Pugazhendhi, M. Irfan, L. Mihet-Popa, P.E. Campana, I.A. Khan, A novel Sustainable Development Goal 7 composite index as the paradigm for energy sustainability assessment: A case study from Europe, Appl. Energy 307 (2022) 118173. https://doi.org/10.1016/j.apenergy.2021.118173.

[48] K.N. Neeraj, V. Maurya, A review on machine learning (feature selection, classification and clustering) approaches of big data mining in different area of research, J. Crit. Rev. 7 (2020) 2610–2626. https://doi.org/http://dx.doi.org/10.31838/jcr.07.19.322.

[49] X. Shu, Y. Ye, Knowledge Discovery: Methods from data mining and machine learning, Soc. Sci. Res. 110 (2023) 102817. https://doi.org/10.1016/j.ssresearch.2022.102817.

[50] R. Gayathri, S.U. Rani, L. Čepová, M. Rajesh, K. Kalita, A Comparative Analysis of Machine Learning Models in Prediction of Mortar Compressive Strength, Processes 10 (2022) 1387. https://doi.org/10.3390/pr10071387.

[51] N. Bagalkot, A. Keprate, R. Orderløkken, Combining Computational Fluid Dynamics and Gradient Boosting Regressor for Predicting Force Distribution on Horizontal Axis Wind Turbine, Vibration 4 (2021) 248–262. https://doi.org/10.3390/vibration4010017.

[52] A. Keprate, R.M.C. Ratnayake, Using gradient boosting regressor to predict stress intensity factor of a crack propagating in small bore piping, in: 2017 IEEE Int. Conf. Ind. Eng. Eng. Manag., IEEE, 2017: pp. 1331–1336. https://doi.org/10.1109/IEEM.2017.8290109.

[53] Z.H. Doost, L. Goliatt, M.S. Aldlemy, M. Ali, B. da S. Macêdo, Enhancing Predictive Accuracy of Compressive Strength in Recycled Concrete Using Advanced Machine Learning Techniques with K-means Clustering, AUIQ Tech. Eng. Sci. 1 (2024) 10. https://doi.org/10.70645/3078-3437.1009.

[54] D.A. Otchere, T.O.A. Ganat, J.O. Ojero, B.N. Tackie-Otoo, M.Y. Taki, Application of gradient boosting regression model for the evaluation of feature selection techniques in improving reservoir characterisation predictions, J. Pet. Sci. Eng. 208 (2022) 109244. https://doi.org/10.1016/j.petrol.2021.109244.

[55] O.A. Alawi, H.M. Kamar, S.Q. Salih, S.I. Abba, W. Ahmed, R.Z. Homod, M. Jamei, S.S. Shafik, Z.M. Yaseen, Development of optimized machine learning models for predicting flat plate solar collectors thermal efficiency associated with Al2O3-water nanofluids, Eng. Appl. Artif. Intell. 133 (2024) 108158. https://doi.org/10.1016/j.engappai.2024.108158.

[56] G. N., P. Jain, A. Choudhury, P. Dutta, K. Kalita, P. Barsocchi, Random Forest Regression-Based Machine Learning Model for Accurate Estimation of Fluid Flow in Curved Pipes, Processes 9 (2021) 2095. https://doi.org/10.3390/pr9112095.

[57] Z.S. Khozani, K. Khosravi, B.T. Pham, B. Kløve, W.H.M.W. Mohtar, Z.M. Yaseen, Determination of compound channel apparent shear stress: Application of novel data mining models, J. Hydroinformatics (2019). https://doi.org/10.2166/hydro.2019.037.

[58] D.A. Tran, M. Tsujimura, N.T. Ha, V.T. Nguyen, D. Van Binh, T.D. Dang, Q.-V. Doan, D.T. Bui, T. Anh Ngoc, L.V. Phu, P.T.B. Thuc, T.D. Pham, Evaluating the predictive power of different machine learning algorithms for groundwater salinity prediction of multi-layer coastal aquifers in the Mekong Delta, Vietnam, Ecol. Indic. 127 (2021) 107790. https://doi.org/10.1016/j.ecolind.2021.107790.

[59] R. Piraei, M. Niazkar, S.H. Afzali, A. Menapace, Application of Machine Learning Models to Bridge Afflux Estimation, Water 15 (2023) 2187. https://doi.org/10.3390/w15122187.

[60] J. Ge, L. Zhao, Z. Yu, H. Liu, L. Zhang, X. Gong, H. Sun, Prediction of Greenhouse Tomato Crop Evapotranspiration Using XGBoost Machine Learning Model, Plants 11 (2022) 1923. https://doi.org/10.3390/plants11151923.

[61] T. Tiyasha, T.M. Tung, S.K. Bhagat, M.L. Tan, A.H. Jawad, W.H.M.W. Mohtar, Z.M. Yaseen, Functionalization of remote sensing and on-site data for simulating surface water dissolved oxygen: Development of hybrid tree-based artificial intelligence models, Mar. Pollut. Bull. 170 (2021) 112639. https://doi.org/10.1016/j.marpolbul.2021.112639.

[62] D. Chicco, M.J. Warrens, G. Jurman, The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation, PeerJ Comput. Sci. 7 (2021) e623. https://doi.org/10.7717/peerj-cs.623.

[63] A. Colin Cameron, F.A.G. Windmeijer, An R-squared measure of goodness of fit for some common nonlinear regression models, J. Econom. 77 (1997) 329–342. https://doi.org/10.1016/S0304-4076(96)01818-0.

[64] T.O. Hodson, Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not, Geosci. Model Dev. 15 (2022) 5481–5487. https://doi.org/10.5194/gmd-15-5481-2022.

[65] J.-M. Sánchez-González, C. Rocha-de-Lossada, D. Flikier, Median absolute error and interquartile range as criteria of success against the percentage of eyes within a refractive target in IOL surgery, J. Cataract Refract. Surg. 46 (2020) 1441–1441. https://doi.org/10.1097/j.jcrs.0000000000000248.

[66] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, Scikit-learn: Machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830. https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf.

[67] W.M. Ridwan, M. Sapitang, A. Aziz, K.F. Kushiar, A.N. Ahmed, A. El-Shafie, Rainfall forecasting model using machine learning methods: Case study Terengganu, Malaysia, Ain Shams Eng. J.

12 (2021) 1651–1663. https://doi.org/10.1016/j.asej.2020.09.011.
[68] Y. Wu, Z. Zhang, M.J.C. Crabbe, L. Chandra Das, Statistical Learning-Based Spatial Downscaling Models for Precipitation Distribution, Adv. Meteorol. 2022 (2022) 1–12. https://doi.org/10.1155/2022/3140872.
[69] Z. Zhou, J. Ren, X. He, S. Liu, 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 (2021) e14424. https://doi.org/10.1002/hyp.14424.
[70] Y. Dash, S.K. Mishra, B.K. Panigrahi, Rainfall prediction for the Kerala state of India using artificial intelligence approaches, Comput. Electr. Eng. 70 (2018) 66–73. https://doi.org/10.1016/j.compeleceng.2018.06.004.
[71] B.T. Pham, L.M. Le, T.-T. Le, K.-T.T. Bui, V.M. Le, H.-B. Ly, I. Prakash, Development of advanced artificial intelligence models for daily rainfall prediction, Atmos. Res. 237 (2020) 104845. https://doi.org/10.1016/j.atmosres.2020.104845.
[72] I. Salehin, I.M. Talha, M. Mehedi Hasan, S.T. Dip, M. Saifuzzaman, N.N. Moon, An Artificial Intelligence Based Rainfall Prediction Using LSTM and Neural Network, in: 2020 IEEE Int. Women Eng. Conf. Electr. Comput. Eng., IEEE, 2020: pp. 5–8. https://doi.org/10.1109/WIECON-ECE52138.2020.9398022.
[73] M. Hammad, M. Shoaib, H. Salahudin, M.A.I. Baig, M.M. Khan, M.K. Ullah, Rainfall forecasting in upper Indus basin using various artificial intelligence techniques, Stoch. Environ. Res. Risk Assess. 35 (2021) 2213–2235. https://doi.org/10.1007/s00477-021-02013-0.
[74] S. He, Q. Luo, X. Fu, L. Zhao, R. Du, H. Li, CAT: A causal graph attention network for trimming heterophilic graphs, Inf. Sci. (Ny). 677 (2024) 120916. https://doi.org/10.1016/j.ins.2024.120916.
[75] Y. Wang, S. He, Q. Luo, H. Yuan, L. Zhao, J. Zhu, H. Li, Causal invariant geographic network representations with feature and structural distribution shifts, Futur. Gener. Comput. Syst. 169 (2025) 107814. https://doi.org/10.1016/j.future.2025.107814.
[76] S. He, Q. Luo, R. Du, L. Zhao, G. He, H. Fu, H. Li, STGC-GNNs: A GNN-based traffic prediction framework with a spatial–temporal Granger causality graph, Phys. A Stat. Mech. Its Appl. 623 (2023) 128913. https://doi.org/10.1016/j.physa.2023.128913.
[77] K.W. Wong, P.M. Wong, T.D. Gedeon, C.C. Fung, Rainfall prediction model using soft computing technique, Soft Comput. - A Fusion Found. Methodol. Appl. 7 (2003) 434–438. https://doi.org/10.1007/s00500-002-0232-4.
[78] C.Z. Basha, N. Bhavana, P. Bhavya, S. V, Rainfall Prediction using Machine Learning & Deep Learning Techniques, in: 2020 Int. Conf. Electron. Sustain. Commun. Syst., IEEE, 2020: pp. 92–97. https://doi.org/10.1109/ICESC48915.2020.9155896.
[79] A. Rahman, S. Abbas, M. Gollapalli, R. Ahmed, S. Aftab, M. Ahmad, M.A. Khan, A. Mosavi, Rainfall Prediction System Using Machine Learning Fusion for Smart Cities, Sensors 22 (2022) 3504. https://doi.org/10.3390/s22093504.
[80] M. Mohammed, R. Kolapalli, N. Golla, S.S. Maturi, Prediction of rainfall using machine learning techniques, Int. J. Sci. Technol. Res. 9 (2020) 3236–3240.
[81] N.K.A. Appiah-Badu, Y.M. Missah, L.K. Amekudzi, N. Ussiph, T. Frimpong, E. Ahene, Rainfall Prediction Using Machine Learning Algorithms for the Various Ecological Zones of Ghana, IEEE Access 10 (2022) 5069–5082. https://doi.org/10.1109/ACCESS.2021.3139312.

Ziaul Haq Doost received his Bachelor of Science degree in Civil Engineering from Herat University, Afghanistan, in December 2015, and a Master of Science in Civil and Environmental Engineering from King Fahd University of Petroleum and Minerals (KFUPM), Saudi Arabia, where the degree awarded on January 4th, 2024. He is currently pursuing his PhD in the same field at KFUPM, where he also serves as a Teaching Assistant. His academic and professional journey is distinguished by extensive research and practical experience in water resource management and environmental engineering. Doost's work includes significant publications in peer-reviewed journals addressing critical issues such as land use impacts on groundwater fluctuations. His research contributes to sustainable environmental practices and water resource management.

Ali Alsuwaiyan received his B.Sc. and M.Sc. degrees in Computer Engineering from King Fahd University of Petroleum and Minerals (KFUPM) in 1998 and 2001, respectively. He then worked as a system analyst at Saudi Aramco for eight years before earning his Ph.D. degree in electrical engineering from the University of Pittsburgh in 2017. Currently, he is an assistant professor in the Department of Computer Engineering at KFUPM and a member of the interdisciplinary research center for intelligent and secure systems at KFUPM. His research interests include emerging nonvolatile memories, computer architecture, and digital systems design, security, and testing.

Dr. Abdulazeez Abdulraheem is a Professor in the Department of Petroleum Engineering at King Fahd University of Petroleum & Minerals (KFUPM), Dhahran, Saudi Arabia. He holds a Ph.D. in Geotechnical Engineering with a perfect GPA of 4.0/4.0 from the University of Oklahoma, USA (1994), an M.E. in Civil Engineering from the Indian Institute of Science, Bangalore (1985), and a B.E. in Civil Engineering from Osmania University, Hyderabad (1983). From 1994 to 2011, he served as a Research Engineer at KFUPM's Research Institute, focusing on wellbore instability, sand production, and experimental rock mechanics. His research interests include geomechanics and the application of artificial intelligence in petroleum engineering. He has published over 300 research papers and has received multiple awards, including KFUPM's Excellence in Teaching and recognition from the International Association of Drilling Contractors.

Prof Nabil M. AL-Areeq is a full professor affiliated with the Center of Water and Climate Changes, Thamar University, Yemen. His major expertise in hydrology and geology. He has published several scientific research articles in different international journals. His research mainly contributing the sustainable development watersheds and geo-science engineering. He is also interested in modeling hydrological applications using the application of computer aid models.

Dr. Zaher Mundher Yaseen is an Assistant professor and research scientist in the field of civil and environmental engineering. Currently, he is working at King Fahd University of Petroleum and Minerals, Saudi Arabia. The scope of his research is quite abroad, covering water resources engineering, environmental engineering, knowledge-based system development, climate and the implementation of data analytic and artificial intelligence. He has published over 440 research articles within international journals and total number of citations over 18000 (Google Scholar H-Index = 76). He has collaborated with over 50 international countries and more than 900 researchers. He has served as a reviewer for more than 140 international journals and academic editor in 8 Clarivate journals.