Water Practice & Technology

IWA PUBLISHING

Water Practice & Technology Vol 19 No 4, 1315 doi: 10.2166/wpt.2024.079

Machine learning approaches for improving precipitation forecasting in the Ambica River basin of Navsari District, Gujarat

Divyang Baudhanwala[a], Darshan Mehta [a,*] and Vijendra Kumar[b]

[a] Civil Engineering Department, Dr. S. & S. S. Ghandhy Government Engineering College, Surat, Gujarat, India
[b] Department of Civil Engineering, Dr Vishwanath Karad MIT World Peace University, Pune, Maharashtra 411038, India
*Corresponding author. E-mail: ap_darshan_mehta@gtu.edu.in

DM, 0000-0001-8418-0026

ABSTRACT

The article examines machine learning models for precipitation forecasting in the Ambica River basin, addressing the important requirement for accurate hydrological forecasts in water resource management. Using a comprehensive collection of meteorological variables such as temperature, humidity, wind speed, and precipitation, four separate models are used: Support Vector Regression (SVR), Random Forest (RF), Decision Tree (DT), and Multiple Linear Regression (MLR). These models' performance is rigorously evaluated using various assessment indicators. The cross-correlation function (XCF) is used in this study to evaluate the correlations between climatic variables and precipitation. The XCF analysis reveals several noteworthy trends, such as a high link between maximum temperature and precipitation, with maxima consistently found at months across all four sites. Furthermore, relative humidity and wind speed have significant connections with precipitation. The findings highlight the value of machine learning approaches in improving precipitation forecast accuracy. The RF and SVR models typically outperform, with values ranging from 0.74 to 0.91. This impressive accuracy underlines their effectiveness in precipitation forecasting, beating competing models in both the training and testing stages. These findings have significant consequences for hydrological processes, notably in the Ambica River basin, where accurate precipitation forecasting is critical for sustainable water resource management.

Key words: Ambica River, forecasting, machine learning, precipitation, supervised learning, SVM

HIGHLIGHTS

- To create precise precipitation forecasting models for the Ambica River basin and evaluating the efficiency of various machine learning techniques.
- To estimate precipitation in the Ambica River basin most accurately, the study will determine the most accurate ML model.

1. INTRODUCTION

Models for precipitation forecasting are essential tools for predicting and managing water resources, lessening the effects of extreme weather, and guaranteeing community safety and security (Ludwig *et al.* 2014; Mehta & Kumar 2022; Mehta *et al.* 2022a, 2022b). Agriculture, energy production, water management, emergency services, and disaster management organizations all depend on precise precipitation forecasts (Wang & Xie 2018). Rainfall is crucial to agriculture, and variations in rainfall patterns may have a big influence on agricultural output and food security. The productivity of farmers can be increased and crop losses from droughts, floods, and other extreme weather events can be decreased by using accurate precipitation forecasting models to guide planting, irrigation, and other agricultural operations (Ali *et al.* 2017). Another area where precise precipitation forecasting is crucial is water resource management. Water availability is directly impacted by precipitation patterns, which also have an influence on reservoir levels, river flows, and groundwater recharge rates (Kumar & Yadav 2022). Water managers can limit the danger of flooding, maintain water availability during droughts and other occurrences that cause water scarcity, and improve water allocation and distribution with the use of accurate precipitation forecasts (Gong *et al.* 2016; Mehta & Kumar 2021).

Precipitation patterns have an impact on the energy industry as well because hydroelectric power generation is highly dependent on rainfall. Energy firms may limit the risk of power outages caused by extreme weather by planning for changes in water supply, maximizing power output, and using precipitation forecasting models (Sharma *et al.* 2023). Precipitation forecasting models are also used by emergency services and disaster management organizations to plan for and respond to severe weather occurrences. Planning evacuations, allocating resources, and preparing for probable flooding, landslides, or other dangerous events can all be facilitated by accurate precipitation forecasts. Recent variations in rainfall patterns brought on by global climate change have increased the frequency and severity of severe precipitation events. Therefore, it is essential to manage the impact of these events on communities and infrastructure through accurate precipitation forecasting. To lessen the effects of severe occurrences, it is crucial to create trustworthy and accurate precipitation forecasting models (Schumann *et al.* 2016; Mehta & Yadav 2021).

Models for predicting precipitation may be created using a variety of modelling approaches. Statistical models use historical data to find patterns and relationships between various variables, including temperature, humidity, pressure, and precipitation, they are frequently used in precipitation forecasting (Dastorani *et al.* 2016). These models forecast future precipitation patterns by using the relationships between these factors. Statistical models have the benefits of being straightforward, simple to use, and inexpensive to compute. The autoregressive model, which analyses past precipitation data to produce forecasts for the future, is one of the most often used statistical models in precipitation forecasting (Pham *et al.* 2019). Since past precipitation patterns are thought to be a good indicator of future precipitation patterns, autoregressive models are excellent for making short-term forecasts (Pérez-Alarcón *et al.* 2022). It is also excellent for recognizing trends and cycles in precipitation patterns, which makes it appropriate for long-term predictions. The multiple regression models, which take into account numerous independent factors to forecast precipitation patterns, are another frequently used statistical model (Jobson 1991). When a number of factors may affect precipitation patterns and their relationships with precipitation may not be linear, this model can be helpful (Themeßl *et al.* 2011).

Statistical models have several drawbacks despite their benefits. The accuracy of statistical models can be constrained by changes in the underlying connections between variables over time, which is one of their most important limitations. For instance, changes in land use or urbanization may affect how temperature and precipitation relate to one another, making historical data less relevant for generating forecasts about the future. Furthermore, because extreme precipitation events may not exhibit the same patterns as more typical precipitation patterns, statistical models may have difficulty capturing these events (Benyahya *et al.* 2007; Patel *et al.* 2023).

The most sophisticated and reliable precipitation forecasting models now in use are numerical weather prediction models (Markovics & Mayer 2022). To mimic atmospheric conditions and forecast weather patterns, they employ complex mathematical and physical equations (Moosavi *et al.* 2021). To provide very accurate forecasts at various geographical and temporal dimensions, these models may include data from several sources, including weather stations, satellites, and radar. Meteorological agencies frequently utilize numerical weather prediction models to provide official weather predictions. High precision is one benefit of numerical weather prediction models (Weyn *et al.* 2020). These projections may be produced by these models up to a week in advance, which can be crucial for planning and decision-making in a variety of industries. They are adaptable tools for a variety of uses because they can create forecasts at many geographical and temporal dimensions, from global to regional to local. Additionally, decision-makers can prepare for uncertain events by using probabilistic forecasts that can be provided by numerical weather prediction models. However, there are a number of drawbacks to numerical weather prediction models. The requirement for large computing resources and skill to construct and maintain these models is one of the key limitations. They need to be regularly updated and calibrated based on the most recent data and scientific developments, and they need powerful computers and sophisticated software to operate. Furthermore, these models are sensitive to input data errors, such as inaccurate temperature or pressure readings, which can seriously impair forecast accuracy (Yoon 2019).

Artificial intelligence models known as 'machine learning (ML)' have grown in popularity recently. They handle complicated and varied data associated with precipitation forecasting and utilize algorithms to find patterns and correlations in huge databases (Akkem *et al.* 2023). In comparison to other models, such as statistical and numerical weather prediction models, ML models have a number of advantages (Cho *et al.* 2020). The capacity of ML models to handle vast and complicated datasets is one of its benefits (Dhal & Azad 2022). Precipitation forecasting takes into account a lot of different factors, including terrain, pressure, humidity,

temperature, and wind speed and direction (Espeholt *et al.* 2022). Traditional statistical models struggle to find relationships and trends because these variables can be mixed in various ways. Large and complicated datasets may be handled by ML models, which can also spot patterns and connections that statistical models could miss out on (Choudhury *et al.* 2021).

The capacity of ML models to learn from data and adjust to changing circumstances is another advantage. As new data becomes available, ML models can update their predictions by using historical data to find patterns and relationships between various variables (Reichstein *et al.* 2019). For forecasting precipitation, where weather patterns can change quickly and unpredictably, the capacity to learn from data and adapt to changing conditions is particularly crucial (Basha *et al.* 2020). Additionally, ML models can be used to create ensemble models, which combine the output of various models to increase accuracy. When forecasting precipitation, ensemble models are especially helpful because different weather models may be required to capture various aspects of weather patterns. As an illustration, one model could perform better at forecasting precipitation in hilly areas than another model would perform better at forecasting precipitation in coastal areas. Ensemble models are able to provide forecasts that are more reliable and accurate by integrating the output of numerous models. Finally, probabilistic predictions may be generated using ML models, which can be helpful for decision-making in a variety of industries. Probabilistic predictions give information on the possibility of various events, such as the likelihood that a specified amount of rain will fall over a particular time frame. Making informed decisions regarding water management, agriculture, and other industries sensitive to precipitation patterns may be done using this knowledge. Various researchers have used different ML methods for the precipitation forecasting such as deep learning (Salman *et al.* 2015), artificial neural network(ANN) (Shah *et al.* 2018; Basha *et al.* 2020; Kumar & Yadav 2021), support vector regression (SVR) (Cramer *et al.* 2017; Singh *et al.* 2023), recurrent neural network (RNN) (Tang *et al.* 2022), long short-term memory (LSTM) network (Barrera-Animas *et al.* 2022), decision tree (Rahman *et al.* 2022).

1.1. Need of the study

The region of South Asia, including the Ambica River basin in South Gujarat, India, is seeing an increase in the frequency and severity of severe precipitation events, which has prompted the necessity for this study. The management of water resources, agriculture, infrastructure, and human livelihoods are all significantly impacted by these events. For successfully managing water resources and reducing the effects of severe catastrophes, accurate precipitation forecasting models are essential. However, there are drawbacks to the current precipitation forecasting models, such as their inability to account for the intricate non-linear relationships between meteorological variables and precipitation. This study develops and evaluates the precision of various machine learning models for precipitation forecasting in the Ambica River basin in an effort to overcome these limitations. In order to improve water resource management strategies and lessen the effects of extreme precipitation events, the study will add to the body of knowledge by selecting the best precise model for precipitation forecasting in the area. The findings of the research will be especially useful to decision-makers, managers of water resources, and other interested parties in vulnerable areas who must manage water resources in the face of changing climate conditions.

1.2. Research gap

Recent catastrophic flooding occurrences in the area have brought to light the necessity for precise precipitation forecasting models in order to enhance water resource management and lessen the effects of extreme events. Despite the significance of the problem, little study has been conducted especially on the Ambica River watershed. Additionally, there are no studies that compare how well various machine learning models perform in predicting the region's monthly precipitation. Finding the most precise ML model for precipitation forecasting in the Ambica River basin is crucial because it may help with the creation of efficient mitigation plans to lessen the effects of floods on the neighbourhood's infrastructure and inhabitants.

1.3. Research objective and novelty

The objective of this research is to create precise precipitation forecasting models for the Ambica River basin and evaluating the efficiency of various machine learning techniques. In order to estimate precipitation in the Ambica River basin most accurately, the study will determine the most accurate ML model. This will aid in reducing the effects of severe precipitation occurrences and enhancing the management of water resources. Support Vector Regression, Random Forest, and Decision Tree are three ML models that will be used to assess the precision

of predictions. The ML models are also compared to the statistical model Multiple Linear Regression (MLR) method. For areas like South Asia that are susceptible to severe precipitation occurrences, this study will be helpful. It will aid in the decision-making process and enable managers of water resources to take the required steps to lessen the effects of severe occurrences.

2. STUDY AREA AND DATASET

The study area for the current research is the Ambica river basin, located in the south Gujarat region of western India. The Ambica river, which flows westward, originates from the Sahyadri mountain ranges near Kotambi village in the Nashik district of Maharashtra, at an elevation of 1,050 meters above mean sea level. The entire river spans 136 kilometres, and the catchment area of the Ambica basin is approximately 2,715 km^2, of which 97.24% is in Gujarat and 3.76% in Maharashtra State of India. The basin is situated between latitudes 20°31′–20°57′N and longitudes 72°48′–73°52′E, covering parts of Valsad, Dang, and Navsari districts in Gujarat State, as well as Nashik in Maharashtra State. The major tributaries of the Ambica river are Olan, Khapri, Kharera, and Kaveri. Figure 1 shows the Index map of Ambica river basin and a map of rain gauge and weather station locations are shows in Figure 2 which were created using ARCGIS 10.8 software.

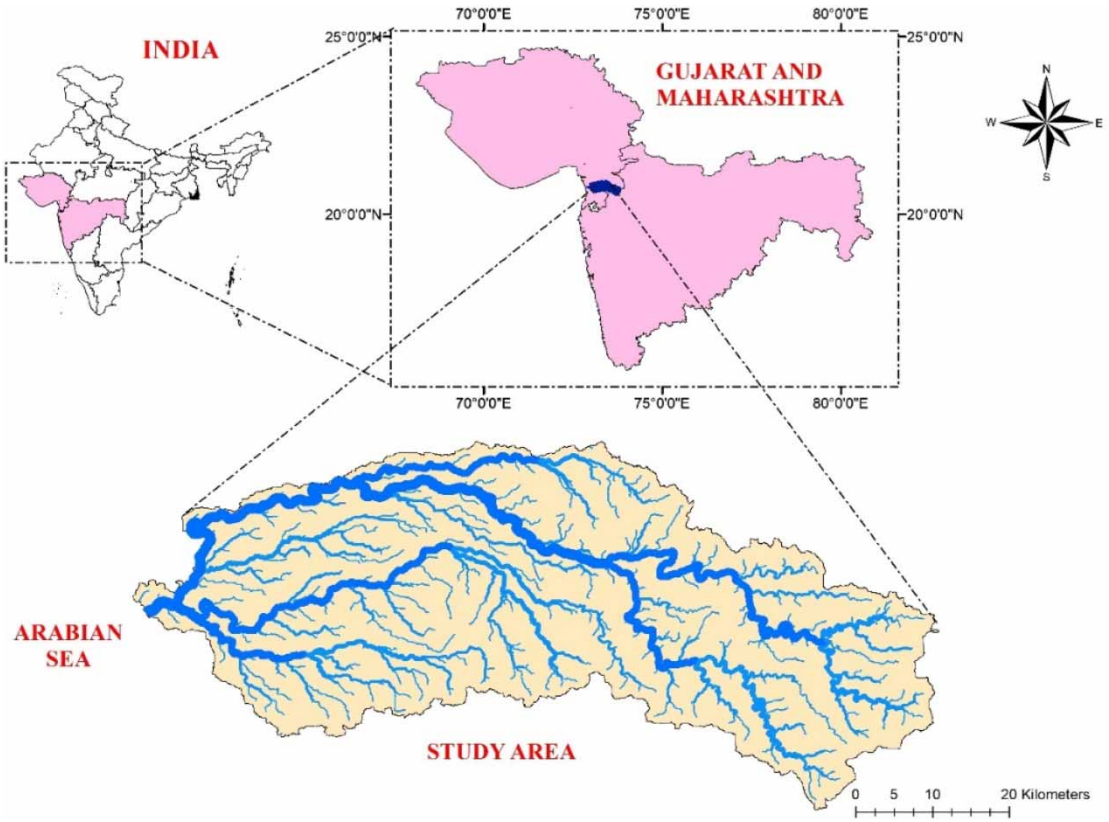

Figure 1 | Index map of Ambica River Basin (study area).

Except for the southwest monsoon season, the basin experiences a hot and relatively dry climate during summers. The temperature varies from 32 to 40 °C during the day and 25 to 8 °C during the night. The southwest monsoon, which lasts from June to September, accounts for nearly 98% of the annual rainfall in the basin. Outside of the monsoon season, there is very little rainfall.

Table 1 shows the rain gauge stations details of Ahwa, Borkhal, Gandevi, and Mankunia in the Navsari and Dang districts includes latitude, longitude, and elevation information. Table 2 displays the data that was used, which included parameters such as mean monthly precipitation, wind speed, minimum and maximum monthly temperature, and relative humidity. The mean monthly precipitation data, which is available from 1981 to 2021, was obtained from NASA Power Access' Data Access Viewer. These parameters provide information about the

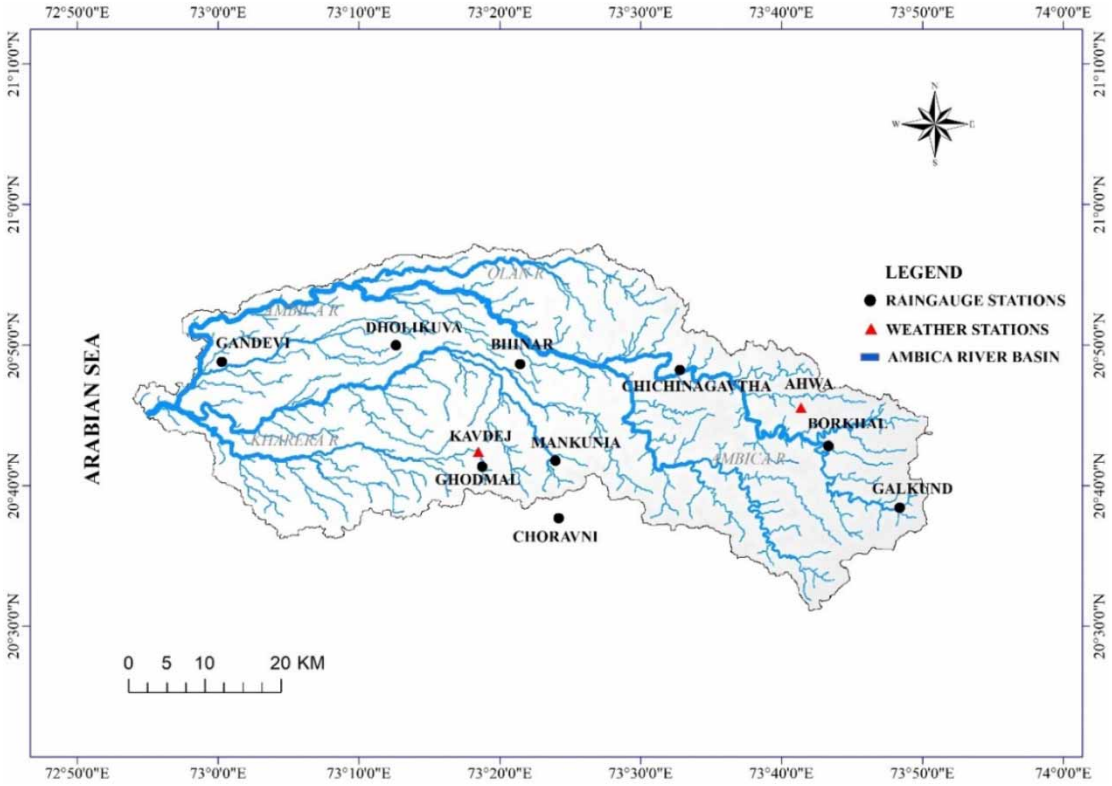

Figure 2 | Location of rain gauge and weather stations in Ambica River Basin.

Table 1 | Presents information on rain gauge stations

Station Name	Districts	Taluka	Latitude	Longitude	Elevations (m)
Ahwa	Dang	Ahwa	20°45′33″	73°41′23″	468
Borkhal	Dang	Ahwa	20°42′33″	73°43′23″	318
Gandevi	Navsari	Navsari	20°48′49″	73°00′16″	15
Mankunia	Navsari	Vansda	20°41′48″	73°23′55″	98

Table 2 | 1981–2021 Climate data from NASA for Ambica River

Data	Period	Source
Mean Monthly Precipitation data (mm)	1981–2021	From Data access Viewer- NASA Power access (https://power.larc.nasa.gov/data-access-viewer/)
Wind speed(m/s)	1981–2021	
Minimum Monthly Temperature data (°C)	1981–2021	
Maximum Monthly Temperature data (°C)	1981–2021	
Relative Humidity (%)	1981–2021	

climatic conditions in the study area, which aids in understanding the impact of climate on the Ambica River Basin's hydrological regime.

The study by (Nunno *et al.* 2022) reported a time series of precipitation data for one station, which was split into training and testing sets with a 70–30% ratio. Accordingly, 70% of the data were used to train the model,

and the remaining 30% were used to evaluate how well it performed. Data analysis and ML often divide the data into training and testing sets (Sun *et al.* 2023). The training set is used to develop the model, while the testing set is used to assess how well the model works with untested data (Wang *et al.* 2020). (Nunno *et al.* 2022) presumably attempted to create a model that could precisely forecast future precipitation patterns at that station using a time series of precipitation data. Analyzing data in time series includes spotting trends and patterns over a period of time. Informed judgments may be made as a result, which can assist forecast future trends. A common split ratio in ML is 70–30%, which provides for an adequate quantity of data to be utilized for both training and testing the model (Gadze *et al.* 2021). Depending on the size of the dataset and the particular requirements of the study, a different precise ratio may be employed. Figure 3 provides a visual depiction of the dataset utilized for both the training and testing phases.

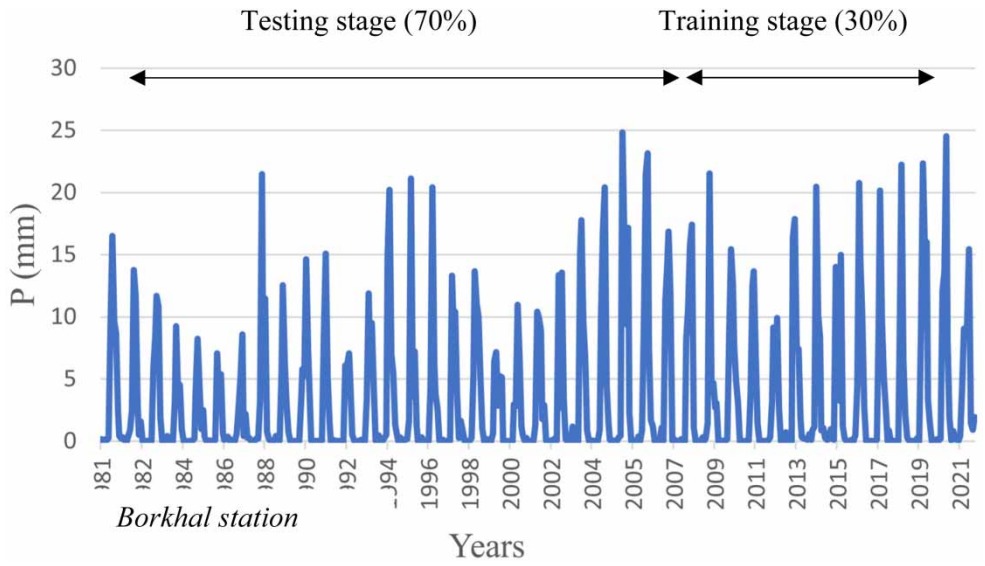

Figure 3 | Data used for training and testing stage.

3. METHODOLOGY

The research methodology is shown in Figure 4, which contains the problem description, input data collection, model training, assessment, error checking, and output. Defining the issue statement and determining the study's

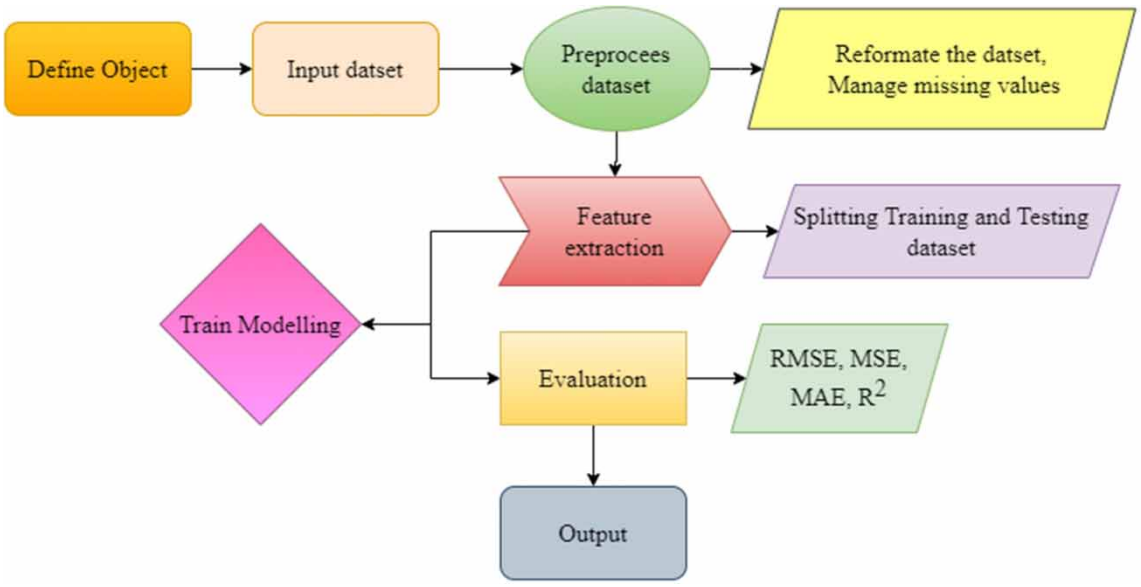

Figure 4 | The graphical representation of research methodology.

goal are the initial steps. After the goal has been established, the necessary input data is gathered and prepared for analysis. The model is trained using the input dataset in the next phase. Here, a machine learning algorithm like SVR, RF, or DT was used. After the model has been trained, its accuracy and generalization performance are assessed on a different test dataset. Finally, the model's output is discovered. Once conclusions from the study have been reached, this output is used to guide decision-making.

3.1. Support vector machine

Support Vector Machine (SVM) techniques are a subset of supervised learning algorithms used for classification and regression applications. Support Vector Regression (SVR), one of several prediction algorithms, is renowned for being trustworthy and highly efficient for regression applications (Shams *et al.* 2023; Zhu *et al.* 2023). SVR is an SVM variation, uses an integrated method for working with continuous data. SVRs are capable of both linear and nonlinear regression, they are a flexible choice for various dataset types. In comparison to other local models and algorithms that rely on traditional chaotic techniques, SVRs are more resilient and robust for datasets that have a high level of noise. They are also more reliable for datasets that have mixed noise compared to other models that use conventional chaotic techniques. The goal of SVR is to maximize the number of instances that fit within the 'street' while minimizing the number of margin violations (Pan *et al.* (2023). The width of the street is determined by a hyper-parameter known as Epsilon.

In SVR, a kernel is a function used to transform low-dimensional data into higher-dimensional data. The hyperplane is the separation line between data classes, and in SVR, it is the line used to predict the continuous or target value. The boundary line is the plane in SVR that separates two classes, and the support vectors are the data points closest to the boundary with the minimum distance (Méndez *et al.* 2023). SVR is an algorithm used for regression problems. In regression problems, the objective of SVR is to find a linear function $f(x)$ that predicts the target values yi with a deviation less than a certain threshold ε. The training dataset is represented as Equation (1):

$$\{(xi\ ,\ yi\ ),\ i\ =\ 1,\ \ldots,\ l\}\quad \subset\ X\ \times\ R, \tag{1}$$

where xi belongs to the space of input arrays X and yi is a real number. The function $f(x)$ is represented as $f(x) = hw, xi + b$, where w is a weight vector in X and b is a bias term in R. The objective is to minimize the Euclidean norm $||w||^2$ subject to the constraints defined by the slack variables ζ_i and ζ_i^*. The optimization problem is given in Equations (2)–(5):

$$\text{minimize: } \frac{1}{2}||w||^2 + C\ \sum_{i=1}^{l}(\zeta_i + \zeta_i^*) \tag{2}$$

$$\text{subject to: } y_i - (w, x_i\ ) - b\ \leq\ \varepsilon +\ \zeta_i \tag{3}$$

$$(w, x_i\ ) + b -\ y_i\ \leq\ \varepsilon +\ \zeta_i^* \tag{4}$$

$$\zeta_i, \zeta_i^* \geq 0, i = 1,\ \ldots, n \tag{5}$$

Here, C is a hyperparameter that controls the tradeoff between maximizing the margin and minimizing the errors, and ζ_i and ζ_i^* are the slack variables that allow deviations from ε. The objective function penalizes samples whose predictions are far from their true targets, with the penalty depending on whether the prediction lies above or below the ε tube.

The effectiveness of SVR depends on the selection of the kernel function and its parameters. The kernel function maps the low-dimensional input space to a high-dimensional feature space, and the Radial Basis Function (RBF) kernel is a commonly used kernel function in SVR. The RBF kernel between two points x_1 and x_2 is given in Equation (6):

$$K(x,\ y)\ = \exp\left(-\frac{||x_1 - x_2||^2}{2\sigma^2}\right) \tag{6}$$

where, σ is the variance, which is a hyper-parameter that controls the smoothness of the decision boundary, and $||x_1 - x_2||$ is the Euclidean distance between the two points x_1 and x_2. Appendix 6 shows the flowchart of the SVR model.

Pseudocode for the SVM

1. Start
2. Import libraries i.e. numpy, pandas, seaborn, matplotlib.pyplot
3. Input-
   Dataset with such variables
4. Define X and Y variables.
5. Preprocessing of the dataset-find missing values, normalization of data
6. from sklearn.model_selection import train_test_split
7. X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.3, random_state = 1
8. Splitting the dataset into train-test with 70:30 ratio

$$\{(xi\ ,\ yi\ ),\ i\ =\ 1,\ \ldots,\ l\}$$

9. Import SVR model with kernel from sklearn library.

$$\text{minimize: } \frac{1}{2}\left|\left|w\right|\right|^2 + C\ \sum_{i=1}^{l}(\zeta_i + \zeta_i^*)$$
$$\text{subject to: } y_i - (w, x_i\ ) - b\ \leq\ \varepsilon + \ \zeta_i$$

10. Apply RBF Kernal:

$$K(x,\ y)\ = \exp\left(-\frac{||x_1 - x_2||^2}{2\sigma^2}\right)$$

11. Find the accuracy from the evaluation metrices.
12. End

3.2. Decision tree

Decision trees (DTs) are a supervised learning technique used for classification and regression problems. They are non-parametric in nature and aim to develop a model that predicts the value of a target variable by learning simple decision rules based on the data attributes. A decision tree can be thought of as an approximate piecewise constant function. Bagging is a parallelized approach that is not dependent on the base learners (Méndez *et al.* 2023). Examples of such models include random forest and additional trees regression. Random forest creates multiple decision tree models by using training data and feature bootstrapping, and then averages several base learners to arrive at the final prediction. On the other hand, extra-tree's regression trains various base learners using all the data, and the node splitting is more randomly distributed (Cojbasic *et al.* 2023).

Given training vectors $x_i \in R^n$, 1, $\ldots$, l and a label vector $y \in R^l$, the algorithm splits the data at each node m into left and right subsets based on a candidate split $\theta = (j, t_m)$ consisting of feature j and threshold t_m. $Q_m^{left}(\theta)$ Represents the subset of data where feature j is less than or equal to t_m, while $Q_m^{\text{right}}(\theta)$ is the subset of data where feature j is greater than t_m, which is represented in Equations (7)–(8).

$$Q_m^{\text{left}}(\theta) = \{(x, y)|x_j \leq t_m\} \tag{7}$$

$$Q_m^{\text{right}}(\theta) = Q_m \backslash Q_m^{\text{left}}(\theta) \tag{8}$$

The quality of a candidate split is evaluated using an impurity function or loss function $H()$ which depends on the task being solved (classification or regression) is given in Equation (9). The impurity of a candidate split θ is given in Equation (10).

$$G(Q_m,\ \theta) = \frac{n_m^{\text{left}}}{n_m} H(Q_m^{\text{left}}(\theta)) + \frac{n_m^{\text{right}}}{n_m} H(Q_m^{\text{right}}(\theta)) \tag{9}$$

$$\theta^* = \text{argmin}_\theta G(Q_m,\ \theta) \tag{10}$$

The algorithm selects the candidate split θ^* that minimizes the impurity measure $G(Q_m,\ \theta)$, and recursively applies the same process to the left and right subsets $Q_m^{\text{left}}(\theta)$ and $Q_m^{\text{right}}(\theta)$ until a stopping criterion is met. The

stopping criterion can be either reaching the maximum allowable depth, where no further splits are allowed, or when the number of samples in a node n_m is less than a minimum threshold ($\min_{\text{samples}}$) or when $n_m = 1$. Appendix 7 shows the flowchart of the DT model.

Pseudocode for the DT

1. Import libraries i.e. numpy, pandas, seaborn, matplotlib.pyplot
2. Input-
   Train Dataset n observations
   Test Dataset m predictors.
3. Splitting datasets for the best partitions into individual classes.
4. Import Decision Tree model.
5. Create the root node for the tree.
6. Apply best condition for tree i.e. max depth, leaf node, decision node.
   Root node: splitting of tree starts.
   Decision node: where decisions are made.
   Child node: this node can't be split for further.
7. Check the accuracy if it is not good then try repeating with above condition until best accuracy.
8. Output: Decision Tree

3.3. Multiple linear regression

Regression is a crucial supervised learning that deals with a set of records having X and Y values. These values are utilized to learn a function that can predict Y for an unknown X. In regression, the aim is to find the value of Y, given that XY is continuous. Here, Y is referred to as the criterion variable, and X is called the predictor variable. Different types of functions or models can be employed for regression, where a linear function is the simplest one. In this case, X can be a single or multiple features that represent the problem.

$$Y = C_1 + C_2 \,.\, X \tag{11}$$

where, $X =$ Input Training data, $Y =$ predicted value of Y for a given X, $C_1 =$ intercept, $C_2 =$ Coefficient of X. Once the optimal values of C_1 and C_2 are determined, the best fit line can be obtained. Appendix 8 shows the flowchart of the MLR model.

Pseudocode for the MLR

1. Start
2. Import libraries i.e. numpy, pandas, seaborn, matplotlib.pyplot
3. Input-
   Train Dataset n observations
   Test Dataset m predictors.
4. Read multiple variable numbers of data.
5. Locate the F-Statistics by OLS, check is it significant or not.
6. If P value < 0.05 if yes the find (β_1 X_1, β_2 $X_{2\ldots}$ β_n $X_{n)}$ and if not so it is not significant.
7. Interpret the slopes values (β_1, β_2, β_3, β_4)
8. Generate MLR model.
9. Find the accuracy from the evaluation metrices.
10. End

3.4. Random forest regression

Random forest is a powerful and accurate regression model that can handle a variety of problems, including those with non-linear relationships. It is an ensemble learning method used for regression in supervised machine learning (Zhang *et al.* 2022). During the training phase, multiple decision trees are built, and each tree predicts the mean of the classes. The steps involved in the random forest algorithm are as follows: first, a random set of p data points are chosen from the training set. Then, a decision tree is constructed using these p data points. This process is repeated for a total of N trees. Finally, for each of the N trees, the value of y is predicted for a

new data point, and the average of all the predicted y values is assigned to the new data point. For predicting rainfall data using environmental input variables, the random forest algorithm is selected as the predictive model. The algorithm builds a large number of decision trees during the training phase, and the mode of mean prediction or regression of each tree determines the resulting class. According to (Kusiak *et al.* 2013), the random forest technique is effective in handling large datasets and can produce positive experimental results even with a significant amount of missing data. Appendix 9 shows the flowchart of the random forest model.

Pseudocode for the RFR

1. Import libraries i.e. numpy, pandas, seaborn, matplotlib.pyplot
2. Input-
   a. Train Dataset n observations
   b. Test Dataset m predictors.
3. Splitting datasets for the best partitions into individual classes.
4. Import random forest model.
5. Select randomly n observation from the observation set.
6. Generate tree from the randomly chosen data.
7. Split the node into sub-node.
8. Model is trained on each bootstrapped independently.
9. Random forest model is created using maximum values.
10. Check the accuracy if it is not good then try repeating with above condition until best accuracy.

3.5. Developed model and measuring performance

In this study, three different machine learning models were developed to predict precipitation using different combinations of exogenous inputs. The models were evaluated based on four metrics: coefficient of determination (R^2), mean absolute error (MAE), root mean square error (RMSE), and mean squared error (MSE). The R^2 metric assesses how well the model can replicate measured values and predict future values. The MAE is the average magnitude of the predicted and measured values, while the RMSE is the square root of the average squared difference between predicted and measured values. The MSE is the average of the set of errors from predicted and measured values, and the Explained Variance score (EVS) is equal to the dispersion of errors between measured and predicted values. Table 3 summarizes the three different models developed based on different combinations of input variables. Model A used four inputs, namely maximum temperature (Tmax), minimum temperature (Tmin), relative humidity (RH), and wind speed (WS). Model B utilized RH and WS as two inputs, and Model C only used RH.

Table 3 | Models developed based on different input

Model	Inputs
A	Tmax, Tmin, RH, WS
B	RH, WS
C	RH

4. RESULTS AND DISCUSSIONS

4.1. Time series analysis

The statistics of the monthly data for the four stations Ahwa, Borkhal, Gandevi, and Mankunia are shown in Appendix 1. The variables included in the table are relative humidity, wind speed, maximum temperature, minimum temperature, and precipitation. The statistics reported for each variable include the mean, standard deviation (σ), and coefficient of variation (CV). The maximum and minimum values for each variable are also reported. The CV is the ratio between the standard deviation and the mean, and it provides a measure of the variability of the data. The table shows that there are variations in the monthly statistics of the four stations, indicating that the climate conditions differ across the studied area.

Various methods exist for selecting input variables, such as average mutual information, alkaline selection technique (AIC), and autocorrelation function. However, in this study, the cross-correlation function (XCF) was used

to determine the delay between input variables and precipitation, as shown in Appendix 10–13. The XCF formula is expressed in Equation (12).

$$XCF = \int_0^S I(t) \cdot [P(t+\tau)]\, d\tau \tag{12}$$

where, I represents the exogenous input variable, s is the duration of the time series, and τ represents the delay.

The patterns observed in the four stations were very similar. The Appendix 10 shows the Tmin showed *XCF* peaks of 0.63 for Borkhal, 0.6 for Ahwa and Gandevi, and 0.62 for Mankunia. On the other hand, the Appendix 11 shows the Tmax showed XCF peaks of 0.65 for Ahwa and Gandevi, 0.7 for Borkhal, and 0.65 for Mankunia. This indicates a stronger correlation between maximum temperature and precipitation. In all four stations, the peaks were observed at $\tau = 9$ months. The cross-correlation between relative humidity and precipitation is shows in Appendix 12, all the stations showed an XCF value close to 0.7, indicating a good relationship between humidity and precipitation. Similarly, for the cross-correlation between wind speed and precipitation is shows in Appendix 13, peaks were observed at $\tau = 11$ months, and an XCF value close to 0.7 was obtained, indicating a good relationship between wind speed and precipitation.

4.2. Ahwa station

This section discusses the predictions made for the Ahwa stations, and presents the computed evolution metrics for both the training and testing stages in an Appendix 2. The best performance during the training stage was achieved by SVR Model A, which included all the monitored inputs, with an R^2 value of 0.85 and a low MAE of 0.80 mm, RMSE of 2.01 mm, MSE of 2.96 mm, and EVS of 0.85. Random Forest Model A also performed well, with an R^2 value of 0.95 and a low MAE of 0.87 mm, RMSE of 1.94 mm, MSE of 3.75 mm, and EVS of 0.95. However, the performance reduced significantly for SVR Model B, which did not include Tmax and Tmin as input, with an R^2 value of 0.71, MAE of 1.18 mm, RMSE of 2.73 mm, MSE of 7.48 mm, and EVS of 0.73. Model C, which simply took relative humidity as input, performed even worse, with an R^2 value of 0.60, MAE of 1.46 mm, RMSE of 3.25 mm, MSE of 10.54 mm, and EVS of 0.62. MLR Model C had the lowest performance, with an R^2 value of 0.54, MAE of 0.51 mm, RMSE of 3.56 mm, MSE of 12.69 mm, and EVS of 0.54. However, there was a significant difference in the model's performance during the testing stage, with SVR Model A still performing the best, with an R^2 value of 0.87, MAE of 1.18 mm, RMSE of 1.98 mm, MSE of 1.98 mm, and EVS of 0.87. Random Forest Model A performed admirably as well, with an R^2 of 0.71, MAE of 3.16 mm, RMSE of 5.92 mm, MSE of 3.11 mm, and EVS of 0.72.

Appendix 14(a) depicts a scatter plot with an R2 value of 0.87 of the association between precipitations measured and predicted by the SVR Model A for the Ahwa station. Appendix 14(b) depicts the predicted precipitation (in mm) vs. the measured precipitation (in mm) for the SVR Model A at the Ahwa station. The figure depicts a perfect match between predicted and measured precipitation. Appendix 14(c) depicts the predicted precipitation (in mm) vs. the measured precipitation (in mm) using the Decision Tree Model A for the Ahwa station. The graphic also shows the coefficient of determination ($R^2 = 0.51$), which indicates how well the trend line fits the data.

4.3. Borkhal station

In this section, the predictions for the Borkhal stations are discussed, and the evaluation metrics are presented for both the training and testing stages in a Appendix 3. For the training stage, the best performance was obtained with SVR Model A, which included all monitored inputs ($R^2 = 0.84$, MAE (mm) = 0.89, RMSE (mm) = 2.17, MSE (mm) = 4.71, EVS = 0.84), and Random Forest Model A ($R^2 = 0.97$, MAE (mm) = 0.48, RMSE (mm) = 0.98, MSE (mm) = 0.97, EVS = 0.97), as shown in the Appendix 3. However, the performance decreased for SVR Model B, which did not include Tmax and Tmin as inputs ($R^2 = 0.76$, MAE (mm) = 1.26, RMSE (mm) = 2.68, MSE (mm) = 7.18, EVS = 0.76) and for Model C, which only included relative humidity ($R^2 = 0.64$, MAE (mm) = 1.58, RMSE (mm) = 3.25, MSE (mm) = 10.57, EVS = 0.65). The worst performance was achieved by Multiple Linear Regression Model C for the testing stage ($R^2 = 0.51$, MAE (mm) = 2.82, RMSE (mm) = 3.88, MSE (mm) = 13.64, EVS = 0.51). However, there was a marked difference in performance between the models for the testing stage, with better performance for SVR Model A ($R^2 = 0.87$, MAE (mm) = 1.26,

RMSE (mm) = 2.04, MSE (mm) = 4.17, EVS = 0.84) and Random Forest Model A (R^2 = 0.84, MAE (mm) = 1.17, RMSE (mm) = 2.24, MSE (mm) = 5.00, EVS = 0.84).

Appendix 15(a) depicts a scatter plot with an R2 value of 0.87 of the association between precipitations measured and predicted by the SVR Model A for the Borkhal Station. Appendix 15(b) depicts the predicted precipitation (in mm) vs. the measured precipitation (in mm) for the SVR Model A at the Borkhal Station. The figure depicts a perfect match between predicted and measured precipitation. Appendix 15(c) depicts the predicted precipitation (in mm) vs. the measured precipitation (in mm) using the Decision Tree Model A for the Borkhal Station. The graphic also shows the coefficient of determination (R^2 = 0.66), which indicates how well the trend line fits the data.

4.4. Gandevi station

In this section, the predictions made for the Gandevi stations and present the evaluation metrics computed for the training and testing stages in the Appendix 4. Among the models tested during the training stage, the SVR Model A with all the monitored inputs (R^2 = 0.83, MAE = 0.98 mm, RMSE = 2.65 mm, MSE = 7.05 mm, EVS = 0.83) and Random Forest Model A (R^2 = 0.96, MAE = 0.63 mm, RMSE = 1.39 mm, MSE = 1.94 mm, EVS = 0.96) performed the best, as shown in the Appendix 4. The performance reduced when passing to SVR Model B, which excluded Tmax and Tmin as inputs (R^2 = 0.69, MAE = 1.46 mm, RMSE = 3.55 mm, MSE = 12.59 mm, EVS = 0.70). Furthermore, the inclusion of relative humidity in Model C resulted in a further reduction in performance (R^2 = 0.59, MAE = 1.76 mm, RMSE = 4.11 mm, MSE = 16.92 mm, EVS = 0.60). The worst performance was observed for Multiple Linear Regression Model C during the testing stage (R^2 = 0.51, MAE = 3.29 mm, RMSE = 4.69 mm, MSE = 21.99 mm, EVS = 0.51). However, a marked improvement in performance was observed for the testing stage, with better performance for SVR Model A (R^2 = 0.85, MAE = 1.49 mm, RMSE = 2.75 mm, MSE = 7.59 mm, EVS = 0.86) and Random Forest Model A (R^2 = 0.84, MAE = 1.35 mm, RMSE = 2.65 mm, MSE = 7.02 mm, EVS = 0.84).

Appendix 16(a) depicts a scatter plot with an R2 value of 0.85 of the association between precipitations measured and predicted by the SVR Model A for the Gandevi Station. Appendix 16(b) depicts the predicted precipitation (in mm) vs. the measured precipitation (in mm) for the SVR Model A at the Gandevi Station. The figure depicts a perfect match between predicted and measured precipitation. Appendix 16(c) depicts the predicted precipitation (in mm) vs. the measured precipitation (in mm) using the Decision Tree Model A for the Gandevi Station. The graphic also shows the coefficient of determination (R^2 = 0.74), which indicates how well the trend line fits the data.

4.5. Mankunia station

In this section, the predictions for the Mankunia stations are discussed, and the evaluation metrics for both training and testing stages are presented in a Appendix 5. For the training stage, the best performance was achieved by SVR Model A, which included all the monitored inputs (R^2 = 0.84, MAE = 1.16 mm, RMSE = 3.00 mm, MSE = 9.21 mm, EVS = 0.84) and Random Forest Model A (R^2 = 0.97, MAE = 0.66 mm, RMSE = 1.38 mm, MSE = 1.92 mm, EVS = 0.97), as shown in the Appendix 5. The performance decreased for SVR-Model B, which did not include Tmax and Tmin as input (R^2 = 0.74, MAE = 1.63 mm, RMSE = 3.81 mm, MSE = 14.52 mm, EVS = 0.74) and for Model C, which included relative humidity (R^2 = 0.63, MAE = 2.05 mm, RMSE = 4.55 mm, MSE = 20.72 mm, EVS = 0.64). The worst performance was observed for Multiple Linear Regression Model C in the testing stage (R^2 = 0.54, MAE = 3.80 mm, RMSE = 5.29 mm, MSE = 27.99 mm, EVS = 0.54). However, during the testing stage, a significant difference was observed between the models, with SVR Model A achieving the best performance (R^2 = 0.87, MAE = 1.58 mm, RMSE = 3.00 mm, MSE = 9.00 mm, EVS = 0.87) and Random Forest Model A (R^2 = 0.91, MAE = 1.17 mm, RMSE = 1.38 mm, MSE = 5.25 mm, EVS = 0.92).

Appendix 17(a) depicts a scatter plot with an R2 value of 0.87 of the association between precipitations measured and predicted by the SVR Model A for the Mankunia Station. Appendix 17(b) depicts the predicted precipitation (in mm) vs. the measured precipitation (in mm) for the SVR Model A at the Mankunia Station. The figure depicts a perfect match between predicted and measured precipitation. Appendix 17(c) depicts the predicted precipitation (in mm) vs. the measured precipitation (in mm) using the Decision Tree Model A for the Mankunia Station. The graphic also shows the coefficient of determination (R2 = 0.67), which indicates how well the trend line fits the data.

5. CONCLUSION

The study showed the potential of ML methods for creating accurate precipitation prediction models for hydrological systems. SVR, RF, DT, and MRL methods were investigated, and the performance of the models was assessed using five evaluation metrics. The findings demonstrated that the RF and SVR models outperformed the other models in offering high accuracy precipitation forecast for all four stations in the Ambica River basin. This study emphasizes how crucial it is to take into account local meteorological variables like temperature, humidity, wind speed, and precipitation when creating precise precipitation prediction models. The study's findings also highlight the need for more research to confirm these models' efficacy in additional domains with distinct characteristics. Overall, effective precipitation prediction models are critical for efficient water resource management, particularly in places where irrigation and agriculture are strongly reliant on rainfall. ML approaches offer a feasible solution for addressing this issue and improving precipitation forecasting accuracy. Our future goals include creating a rainfall prediction model that takes into account wind direction, sea surface temperature, sunshine data, cloud cover, and climate indices. We also want to look into how climate change is affecting rainfall, and further research should be done in regions that have both tropical monsoon and other distinct climates, such as semi-arid climates. It is advised that more study and development be done to improve the models' suitability for predicting urban precipitation.

DECLARATIONS

All authors have read, understood, and have complied as applicable with the statement on 'Ethical responsibilities of Authors' as found in the Instructions for Authors.

FUNDING

This research received no external funding.

DATA AVAILABILITY STATEMENT

All relevant data are included in the paper or its Supplementary Information.

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

Akkem, Y., Biswas, S. K. & Varanasi, A. 2023 Smart farming using artificial intelligence: A review. *Eng. Appl. Artif. Intell.* **120**, 105899. doi:10.1016/j.engappai.2023.105899.

Ali, S., Liu, Y., Ishaq, M., Shah, T., Abdullah, Ilyas, A. & Din, I. U. 2017 Climate change and its impact on the yield of major food crops: Evidence from Pakistan. *Foods* **6**(6), 39.

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

Basha, C. Z., Bhavana, N., Bhavya, P. & Sowmya, V. 2020 Rainfall Prediction using Machine Learning & Deep Learning Techniques. In: *2020 Int. Conf. Electron. Sustain. Commun. Syst.* IEEE. pp. 92–97. doi:10.1109/ICESC48915.2020.9155896.

Benyahya, L., Caissie, D., St-Hilaire, A., Ouarda, T. B. M. & Bobée, B. 2007 A review of statistical water temperature models. *Can. Water Resour. J.* **32**(3), 179–192. doi:10.4296/cwrj3203179.

Cho, D., Yoo, C., Im, J. & Cha, D. 2020 Comparative assessment of various machine learning-based bias correction methods for numerical weather prediction model forecasts of extreme air temperatures in urban areas. *Earth Sp. Sci.* **7**(4). doi:10.1029/2019EA000740.

Choudhury, P., Allen, R. T. & Endres, M. G. 2021 Machine learning for pattern discovery in management research. *Strateg. Manage. J.* **42**(1), 30–57. doi:10.1002/smj.3215.

Cojbasic, S., Dmitrasinovic, S., Kostic, M., Sekulic, M. T., Radonic, J., Dodig, A. & Stojkovic, M. 2023 Application of machine learning in river water quality management: A review. *Water Sci. Technol.* **88**(9), 2297–2308. https://doi.org/10.2166/wst.2023.331.

Cramer, S., Kampouridis, M., Freitas, A. A. & Alexandridis, A. K. 2017 An extensive evaluation of seven machine learning methods for rainfall prediction in weather derivatives. *Expert Syst. Appl.* **85**, 169–181. doi:10.1016/j.eswa.2017.05.029.

Dastorani, M., Mirzavand, M., Dastorani, M. T. & Sadatinejad, S. J. 2016 Comparative study among different time series models applied to monthly rainfall forecasting in semi-arid climate condition. *Nat. Hazards* **81**(3), 1811–1827. doi:10.1007/s11069-016-2163-x.

Dhal, P. & Azad, C. 2022 A comprehensive survey on feature selection in the various fields of machine learning. *Appl. Intell.* **52**(4), 4543–4581. doi:10.1007/s10489-021-02550-9.

Espeholt, L., Agrawal, S., Sønderby, C., Kumar, M., Heek, J., Bromberg, C., Gazen, C., Andrychowicz, M., Hickey, J., Bell, A. & Kalchbrenner, N. 2022 Deep learning for twelve hour precipitation forecasts. *Nature communications* **13**(1), 1–10.

Gadze, J. D., Bamfo-Asante, A. A., Agyemang, J. O., Nunoo-Mensah, H. & Opare, K. A.-B. 2021 An investigation into the application of deep learning in the detection and mitigation of DDOS attack on SDN controllers. *Technologies* **9**(1), 14. doi:10.3390/technologies9010014.

Gong, Y., Zhang, Y., Lan, S. & Wang, H. 2016 A comparative study of artificial neural networks, support vector machines and adaptive neuro fuzzy inference system for forecasting groundwater levels near Lake Okeechobee, Florida. *Water Resour. Manage.* **30**(1), 375–391. doi:10.1007/s11269-015-1167-8.

Jobson, J. D. 1991 *Multiple Linear Regression*, pp. 219–398. doi:10.1007/978-1-4612-0955-3_4.

Kumar, V., Yadav, S. M., 2021 Real-time flood analysis using artificial neural network. In: *Recent Trends in Civil Engineering.* (Pathak, K. K., Bandara, J. M. S. J. & Agrawal, R., eds). Lecture Notes in Civil Engineering, Vol. 77. Springer, Singapore, pp. 973–986 https://doi.org/10.1007/978-981-15-5195-6_71, doi:10.1007/978-981-15-5195-6_71.

Kumar, V. & Yadav, S. M. 2022 A state-of-the-Art review of heuristic and metaheuristic optimization techniques for the management of water resources. *Water Supply* **22**(4), 3702–3728. doi:10.2166/ws.2022.010.

Kusiak, A., Wei, X., Verma, A. P. & Roz, E. 2013 Modeling and prediction of rainfall using radar reflectivity data: A data-mining approach. *IEEE Trans. Geosci. Remote Sens.* **51**(4), 2337–2342. doi:10.1109/TGRS.2012.2210429.

Ludwig, F., Slobbe, E. v. & Cofino, W. 2014 Climate change adaptation and integrated water resource management in the water sector. *J. Hydrol.* **518**, 235–242. doi:10.1016/j.jhydrol.2013.08.010.

Markovics, D. & Mayer, M. J. 2022 Comparison of machine learning methods for photovoltaic power forecasting based on numerical weather prediction. *Renewable Sustainable Energy Rev.* **161**, 112364. doi:10.1016/j.rser.2022.112364.

Mehta, D. J. & Kumar, V. Y. 2021 Water productivity enhancement through controlling the flood inundation of the surrounding region of Navsari Purna River, India. *Water Productivity J.* **1**(2), 11–20.

Mehta, D. J. & Kumar, Y. V. 2022 Flood Modelling Using HEC-RAS for Purna River, Navsari District, Gujarat, India. In: Kumar, P., Nigam, G.K., Sinha, M.K., Singh, A. (eds) *Water Resources Management and Sustainability. Advances in Geographical and Environmental Sciences*. Springer, Singapore. https://doi.org/10.1007/978-981-16-6573-8_11.

Mehta, D. & Yadav, S. M. 2021 Analysis of Long-Term Rainfall Trends in Rajasthan, India. In: Jha, R., Singh, V. P., Singh, V., Roy, L. B., Thendiyath, R. (eds) *Climate Change Impacts on Water Resources. Water Science and Technology Library*, vol 98. Springer, Cham. https://doi.org/10.1007/978-3-030-64202-0_26.

Mehta, D., Waikhom, S., Yadav, V., Lukhi, Z., Eslamian, S., Furze, J. N., 2022a Trend analysis of rainfall: A case study of Surat City in Gujarat, Western India. In: *Earth Systems Protection and Sustainability* (Furze, J. N., Eslamian, S., Raafat, S. M. & Swing, K., eds). Springer, Cham. https://doi.org/10.1007/978-3-030-98584-4_8.

Mehta, D. J., Eslamian, S. & Prajapati, K. 2022b Flood modelling for a data-scare semi-arid region using 1-D hydrodynamic model: A case study of Navsari Region. *Model. Earth Syst. Environ* **8**(2), 2675–2685.

Méndez, M., Merayo, M. G. & Núñez, M. 2023 Machine learning algorithms to forecast air quality: a survey. *Artif Intell Rev* **56**, 10031–10066 (2023). https://doi.org/10.1007/s10462-023-10424-4.

Moosavi, A., Rao, V. & Sandu, A. 2021 Machine learning based algorithms for uncertainty quantification in numerical weather prediction models. *J. Comput. Sci.* **50**, 101295. doi:10.1016/j.jocs.2020.101295.

Nunno, F., Granata, F., Pham, Q. B. & Marinis, G. d. 2022 Precipitation forecasting in Northern Bangladesh using a hybrid machine learning model. *Sustainability* **14**(5), 2663. doi:10.3390/su14052663.

Pan, Q., Harrou, F. & Sun, Y. 2023 A comparison of machine learning methods for ozone pollution prediction. *J. Big Data* **10**, 63. https://doi.org/10.1186/s40537-023-00748-x.

Patel, A., Keriwala, N., Mehta, D., Shaikh, M. & Eslamian, S. 2023 Flood Resilient Plan for Urban Area: A Case Study. In: Eslamian, S., Eslamian, F. (eds) *Disaster Risk Reduction for Resilience*. Springer, Cham. https://doi.org/10.1007/978-3-031-22112-5_8.

Pérez-Alarcón, A., Garcia-Cortes, D., Fernández-Alvarez, J. C. & Martínez-González, Y. 2022 Improving monthly rainfall forecast in a watershed by combining neural networks and autoregressive models. *Environ. Process.* **9**(3), 53. doi:10.1007/s40710-022-00602-x.

Quoc Bao Pham, Q. B. P., Abba, S. I., Usman, A. G., Nguyen Thi Thuy Linh, N. T. T. L., Vivek Gupta, V. G., Anurag Malik, A. M., Costache, R. & Doan Quang Tri, D. Q. T. 2019 Potential of hybrid data-intelligence algorithms for multi-station modelling of rainfall. *Water Resour Manage* **33**(15), 5067–5087. https://doi.org/10.1007/s11269-019-02408-3.

Rahman, A., Abbas, S., Gollapalli, M., Ahmed, R., Aftab, S., Ahmad, M., Khan, M. A. & Mosavi, A. 2022 Rainfall Prediction System Using Machine Learning Fusion for Smart Cities. *Sensors* **22**(9), 3504. https://doi.org/10.3390/s22093504.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N. & Prabhat, F. 2019 Deep learning and process understanding for data-driven Earth system science. *Nature* **566**(7743), 195–204.

Salman, A. G., Kanigoro, B. & Heryadi, Y. 2015 Weather forecasting using deep learning techniques. In: *2015 Int. Conf. Adv. Comput. Sci. Inf. Syst.* IEEE. pp. 281–285. doi:10.1109/ICACSIS.2015.7415154.

Schumann, G., Kirschbaum, D., Anderson, E. & Rashid, K. 2016 Role of earth observation data in disaster response and recovery: From science to capacity building. pp. 119–146. doi:10.1007/978-3-319-33438-7_5.

Shah, U., Garg, S., Sisodiya, N., Dube, N. & Sharma, S. 2018 Rainfall prediction: Accuracy enhancement using machine learning and forecasting techniques. In: *2018 Fifth Int. Conf. Parallel, Distrib. Grid Comput*. IEEE. pp. 776–782. doi:10.1109/PDGC.2018.8745763.

Shams, M. Y., Elshewey, A. M., El-kenawy, E. S. M., Ibrahim, A., Talaat, F. M. & Tarek, Z. 2023 Water quality prediction using machine learning models based on grid search method. *Multimedia Tools and Applications* 1–28. https://doi.org/10.1007/s11042-023-16737-4.

Sharma, K. V., Kumar, V., Singh, K. & Mehta, D. J. 2023 LANDSAT 8 LST pan sharpening using novel principal component based downscaling model. *Remote Sens. Appl. Soc. Environ.* 100963. Elsevier B.V.. doi:10.1016/j.rsase.2023.100963.

Singh, K., Singh, B., Sihag, P., Kumar, V. & Sharma, K. V. 2023 Development and application of modeling techniques to estimate the unsaturated hydraulic conductivity. *Model. Earth Syst. Environ.*. doi:10.1007/s40808-023-01744-z.

Sun, H., Lui, S., Huang, X., Sweeney, J. & Gong, Q. 2023 Effects of randomness in the development of machine learning models in neuroimaging studies of schizophrenia. *Schizophr. Res.* **252**, 253–261. doi:10.1016/j.schres.2023.01.014.

Tang, T., Jiao, D., Chen, T. & Gui, G. 2022 Medium- and long-term precipitation forecasting method based on data augmentation and machine learning algorithms. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **15**, 1000–1011. doi:10.1109/JSTARS.2022.3140442.

Themeßl, M. J., Gobiet, A. & Leuprecht, A. 2011 Empirical-statistical downscaling and error correction of daily precipitation from regional climate models. *Int. J. Climatol.* **31**(10), 1530–1544. doi:10.1002/joc.2168.

Wang, X. & Xie, H. 2018 A review on applications of remote sensing and Geographic Information Systems (GIS) in water resources and flood risk management. *Water* **10**(5), 608. doi:10.3390/w10050608.

Wang, X., Liang, G., Zhang, Y., Blanton, H., Bessinger, Z. & Jacobs, N. 2020 Inconsistent performance of deep learning models on mammogram classification. *J. Am. Coll. Radiol.* **17**(6), 796–803. doi:10.1016/j.jacr.2020.01.006.

Weyn, J. A., Durran, D. R. & Caruana, R. 2020 Improving data-driven global weather prediction using deep convolutional neural networks on a cubed sphere. *J. Adv. Model. Earth Syst.* **12**(9). doi:10.1029/2020MS002109.

Yoon, S.-S. 2019 Adaptive blending method of radar-based and numerical weather prediction QPFs for urban flood forecasting. *Remote Sens.* **11**(6), 642. doi:10.3390/rs11060642.

Zhang, Y., Liu, J. & Shen, W. A. 2022 Review of ensemble learning algorithms used in remote sensing applications. *Appl. Sci.* **12**, 8654. https://doi.org/10.3390/app12178654.

Zhu, M., Wang, J., Yang, X., Zhang, Y., Zhang, L., Ren, H., Wu, B. & Ye, L. 2023 A review of the application of machine learning in water quality evaluation. *Eco-Environ. Health* **1**(2), 107–116. https://doi.org/10.1016/j.eehl.2022.06.001.

First received 1 November 2023; accepted in revised form 20 March 2024. Available online 30 March 2024