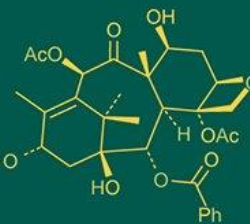
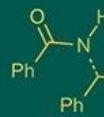
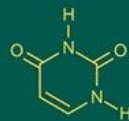
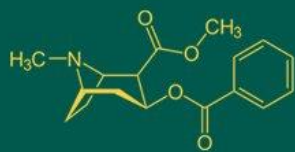


International Journal of Advanced Biochemistry Research



ISSN Print: 2617-4693
ISSN Online: 2617-4707
NAAS Rating (2026): 5.29
IJABR 2026; SP-10(7): 602-613
www.biochemjournal.com
Received: 13-05-2026
Accepted: 15-06-2026

Roshini Priya CH
Department of Statistics and Mathematics, College of Agriculture, Professor Jayashankar Telangana State Agricultural University (PJ TSAU), Rajendranagar, Hyderabad, Telangana, India

A Meena
Department of Statistics and Mathematics, College of Agriculture, Professor Jayashankar Telangana State Agricultural University (PJ TSAU), Rajendranagar, Hyderabad, Telangana, India

Santosh A Rathod
ICAR-National Institute of Abiotic Stress Management (ICAR-NIASM), Malegaon Khurd, Baramati, Pune District, Maharashtra, India

B Balaji Naik
Agro Climate Research Centre (ACRC), Agricultural Research Institute (ARI), Professor Jayashankar Telangana State Agricultural University (PJ TSAU), Rajendranagar, Hyderabad, Telangana, India

Corresponding Author:
Roshini Priya CH
Department of Statistics and Mathematics, College of Agriculture, Professor Jayashankar Telangana State Agricultural University (PJ TSAU), Rajendranagar, Hyderabad, Telangana, India

Comparative evaluation of statistical, machine learning and deep learning models for multivariate daily rainfall forecasting

Roshini Priya CH, A Meena, Santosha Rathod and B Balaji Naik

DOI: <https://www.doi.org/10.33545/26174693.2026.v10.i7Sh.9043>

Abstract

Aims: To compare statistical, machine learning and deep learning models for multivariate daily rainfall prediction in Dandepalle Mandal, Telangana, India, and assess their usefulness for mandal-level agricultural advisory support.

Study Design: Comparative time-series forecasting study using secondary daily weather observations.

Place and Duration of Study: Dandepalle Mandal, Mancherial district, Telangana, India, using daily rainfall, maximum temperature and minimum temperature data for 1995-2024.

Methodology: Rainfall distribution was examined for zero-rainfall frequency, skewness and kurtosis. Log1p transformation was used to reduce skewness. Machine learning models used lagged rainfall, rolling statistics, calendar indicators and rain/no-rain history features, while neural architectures used sequential rainfall-temperature inputs. ARIMAX, SARIMAX and GARCH variants were used as statistical baselines. Models were evaluated using RMSE, R^2 and residual diagnostics on the original rainfall scale.

Results: The shallow Decision Tree regressor gave the lowest test RMSE (8.7019 mm) with $R^2 = 0.1075$, followed by XGBoost (RMSE = 8.8180 mm; $R^2 = 0.0836$) and SVR (RMSE = 8.8497 mm; $R^2 = 0.0770$). Deep learning models did not outperform the best machine learning models, with RNN and LSTM giving RMSE values of 8.8634 and 8.8868 mm, respectively. Significant Ljung-Box statistics indicated remaining autocorrelation.

Conclusion: Parsimonious machine learning models, particularly Decision Tree, XGBoost and SVR, provided the most useful local rainfall prediction results. However, low R^2 values and residual autocorrelation show that rainfall-temperature inputs alone do not fully capture daily monsoon dynamics.

Keywords: Daily rainfall forecasting, Dandepalle, Telangana, decision tree, SVR, XGBoost, agrometeorological advisory

1. Introduction

Rainfall is a key weather factor for farming in semi-arid tropical areas. Decisions about planting, crop growth, irrigation, nutrient use and pest management rely mostly on when and how much it rains. At the farm level, uncertainty in rainfall is not only a water management problem, but it also affects crop production. Unexpected rainfall can lead to nutrient loss, poor field accessibility, waterlogging or pesticide runoff. These risks are particularly important in rainfed and partially irrigated areas of peninsular India, where the southwest monsoon is a key water source but is often unpredictable, with irregular rainfall, dry spells, high variability and occasional extreme weather events.

India has made significant investments in weather services, rainfall monitoring and crop advisories based on weather conditions. However, the effectiveness of these forecasts for farmers depends on their accuracy at the local level. While state or district forecasts are useful for overall planning, farmers at the mandal level need more precise rainfall information because of differences in soil type, land use and rainfall patterns within districts. This mismatch in scale creates continuing challenges for agricultural advisories. For example, a rainfall forecast that is accurate at the district level may still miss a specific mandal, while sudden heavy rain may not be captured by broader forecasts. Thus, there is a clear need for localized rainfall prediction models that can be regularly updated with current weather data and used to inform farmers' decisions.

Telangana provides a suitable study area for developing forecasting models because its agriculture is heavily affected by changes in monsoon rainfall, dry spells, and extreme rainfall events. The state has northern, central, and southern regions with different farming conditions and crops. Farmers grow major crops like cotton, rice, maize, pulses, and oilseeds in these areas. The success of these crops depends on rainfall, irrigation, and soil quality. Dandepalle Mandal, in the northern part of Telangana, faces changes in seasonal rainfall that impact farming practices, irrigation choices, and how farmers manage crop risks. Therefore, having daily rainfall forecasts at this local level is important for farmers and those providing agricultural support. Predicting daily rainfall is challenging due to its statistical characteristics. Unlike temperature, rainfall is unpredictable and can fluctuate widely, often showing many days with no rain at all, while a few days have high values. This causes problems for traditional statistical models, which often assume a normal distribution. Some models might work well on dry days but struggle during heavy rainfall, or they might predict rain on too many dry days, leading to higher average errors. Therefore, evaluating these models requires considering not just overall accuracy but also how well they handle dry days, the presence of many zeros, biases, and their ability to identify rare but significant rainfall events important for agriculture.

Previous rainfall trend analysis for Telangana also showed that rainfall behaviour varies across months and seasons, with annual and seasonal trends requiring localized interpretation for agricultural planning (Jyostna *et al.*, 2022)^[17].

Traditional forecasting methods like ARIMA, SARIMA, and GARCH models remain valuable because they are easy to understand and allow for assessment of prediction quality. However, these methods may not be effective for predicting daily monsoon rainfall, which is influenced by complex interactions among previous rainfall, temperature, humidity, weather patterns, and local conditions. Machine learning techniques can be a better option, as they can identify non-linear relationships using features such as past data and seasonal patterns without requiring a specific mathematical formula. Deep learning techniques, such as ANNs, CNNs, RNNs, and LSTMs, are commonly used to predict rainfall because they can identify patterns in sequential data. Ray *et al.* (2020)^[32] integrated wavelet transform with ANN and WNN for Indian monsoon rainfall forecasting and reported that the hybrid approach improved forecasting performance compared with individual ANN and WNN models. Rathod *et al.* (2021)^[28] also developed a two-stage spatiotemporal time-series modelling approach for rice yield prediction, showing the importance of combining statistical modelling, nonlinear learning and decision-support relevance in agricultural forecasting. LSTM models are particularly useful for rainfall forecasting, as they can retain important information during both wet and dry periods. Recently, Jyostna *et al.* (2025)^[16] proposed a multiscale EEMD-LSTM rainfall forecasting approach for Telangana, where rainfall series were decomposed into intrinsic mode functions before LSTM modelling. This supports the potential of hybrid decomposition-based deep learning models for improving rainfall prediction. However, these models need enough data, proper regularization, and suitable input variables. When the dataset includes only local rainfall and temperature, deep learning may not yield better predictions, especially when

daily rainfall is influenced by random events and other atmospheric factors not included in the data inputs.

This study focuses on improving daily rainfall prediction at the mandal level in Dandepalle by comparing various statistical, machine learning, and deep learning models. The key goals include: (i) analyzing the rainfall data distribution in Dandepalle; (ii) creating features for machine learning models and structured inputs for deep learning; (iii) assessing different models such as Decision Tree, SVR, XGBoost, Random Forest, various neural networks, and others; (iv) evaluating the models based on metrics like RMSE, R^2 , and residual analysis; and (v) applying the results to enhance crop planning and agricultural advice at the mandal level.

2. Review of Literature

Rainfall forecasting has gradually moved from classical statistical models to machine learning and deep learning approaches. The rainfall data show seasonal patterns, nonlinear behaviour, many days with zero rainfall, and occasional heavy rainfall events. Because of these characteristics, predicting rainfall at the local and daily scales is still difficult. Previous rainfall forecast studies in India have also shown that rainfall varies strongly across regions, making local-level forecasting important for agricultural planning and water management (Pai *et al.*, 2014; Durai and Bhavani Bhate, 2020; Chandu *et al.*, 2025)^[5, 9, 25].

Classical statistical time-series models used for rainfall forecasting. The ARIMA model became popular because it can capture temporal patterns using autoregressive, differencing, and moving-average components (Box & Jenkins, 1970; Box *et al.*, 2015)^[2, 3]. ARIMA works well when the data has a clear linear pattern and can be made stationary. Later, SARIMA was used to capture seasonal changes in time-series data, especially when there is a regular seasonal cycle (Hyndman & Khandakar, 2008)^[15]. In agricultural weather studies, SARIMA has also been used to model potential evapotranspiration, which is important for irrigation planning (Patil *et al.*, 2022)^[26]. But daily rainfall often has sudden changes, dry spells, and extreme events. These characteristics of rainfall are challenging for linear models to accurately capture.

GARCH-type models have also been used in time-series studies to show how the variance changes over time. These models are helpful when the amount of variation in the data changes from one period to another. But ARIMA, SARIMA, and GARCH models still rely on statistical assumptions and may not fully capture the nonlinear nature of rainfall. So, these models are good as baseline methods, but they may not perform well when rainfall is highly irregular and contains many zeros.

As computational methods improved, machine learning models started to be used for rainfall forecasting. Decision Tree models split the data into groups based on rules and capture threshold-type relationships between rainfall and other variables. This helps because rainfall might depend on factors such as previous rainfall, season, temperature, and other weather conditions. Random Forest was developed to improve on single Decision Trees. It trains many trees and averages their results, which helps stabilise predictions and reduce overfitting (Breiman, 2001)^[4]. In rainfall forecasting, Random Forest can capture the nonlinear interactions between rainfall, temperature, and lagged weather variables. Pham *et al.* (2021)^[27] also showed that tree-based and support vector methods can help with rainfall downscaling and

extreme rainfall prediction. As Random Forest averages many trees, it can sometimes smooth out rare heavy rainfall events and miss sudden extremes. Support Vector Regression is another important machine learning method used for rainfall prediction. SVR uses kernel functions to capture nonlinear relationships between input variables and rainfall, introduced by Cortes and Vapnik (1995)^[7], and later studies showed that SVR can be useful for rainfall forecasting when the relationship between predictors and rainfall is not linear (Hong and Pai, 2007)^[14, 25]. In agricultural and weather forecasting, SVR has also shown good performance compared with classical models. Gorlapalli *et al.* (2022)^[12], in a water-stress forecasting study showed that artificial intelligence models such as SVR can be useful in agrometeorological forecasting. Similarly, Saha *et al.* (2021)^[33] used a tuned SVR for cotton production forecasting and reported that SVR-based models can effectively handle nonlinear agricultural time-series patterns.

Boosting-based models such as XGBoost have also been applied to weather and rainfall data. XGBoost builds trees sequentially, where each new tree tries to correct the errors made by the earlier trees. It combines nonlinear modelling, regularization and learning-rate control, which makes it useful for noisy datasets (Chen and Guestrin, 2016)^[6]. In rainfall forecasting, XGBoost can capture complex relationships between rainfall, temperature, lag variables and calendar-based features. Mishra *et al.* (2024)^[22] showed the usefulness of XGBoost for modelling rainfall patterns in India. El Hafyani *et al.* (2024)^[10] also reported that machine learning approaches, including tree-based boosting methods, can improve precipitation prediction when suitable meteorological inputs are available.

Artificial Neural Networks have also been widely used in rainfall and hydrological forecasting. ANN models can learn nonlinear relationships without requiring strict statistical assumptions. Earlier studies showed that neural networks can be useful for rainfall and hydrological prediction tasks (Luk *et al.*, 2001; Dawson and Wilby, 2001; Nanda *et al.*, 2013)^[8, 21, 24]. However, ordinary ANN models do not naturally preserve the order of time unless lagged inputs are carefully included. Therefore, their performance in daily rainfall forecasting depends strongly on the choice of input variables, lag structure and training procedure.

Deep learning models were later introduced to better handle sequential and complex rainfall patterns. Convolutional Neural Networks can detect short-term local patterns within input sequences and have been used in precipitation nowcasting and time-series applications (Shi *et al.*, 2015). However, CNN models may not perform well when rainfall does not follow stable short-term patterns. Recurrent Neural Networks are designed for sequential data because they use information from previous time steps. This makes them useful for time-series forecasting, but simple RNNs may struggle to learn long-term dependencies.

Long Short-Term Memory models were developed to overcome the limitations of simple RNNs by using memory cells that can retain useful information for longer periods (Hochreiter and Schmidhuber, 1997)^[13]. LSTM models have been successfully applied in rainfall, flood and hydrological forecasting studies (Kratzert *et al.*, 2018; Le *et al.*, 2019; Zhang *et al.*, 2020)^[18, 20, 37]. However, the performance of LSTM depends on the strength of the temporal signal, the quality of input variables and the amount of training data. When rainfall data contain many zero-rainfall days and only

a few extreme events, even LSTM models may find it difficult to predict heavy rainfall accurately.

Recent studies have also shown that advanced deep learning and AI-based weather models can improve precipitation forecasting when rich spatial and atmospheric information is available. For example, radar-based and large-scale weather forecasting systems have shown strong performance in short-term and medium-range precipitation forecasting (Ravuri *et al.*, 2021; Espeholt *et al.*, 2022; Lam *et al.*, 2023)^[11, 19, 31]. However, such models usually require large datasets, spatial weather fields, satellite/radar information or numerical weather prediction inputs. At the local mandal scale, where only rainfall and temperature data are available, simpler machine learning models may still be more practical and interpretable.

Several recent studies have compared statistical, machine learning and deep learning methods for rainfall forecasting. Wani *et al.* (2024)^[36] compared machine learning, deep learning and time-series models for rainfall prediction across different elevation zones and showed that model performance depends strongly on location, data quality and input variables. Mishra *et al.* (2025)^[23] also compared statistical and deep learning models for rainfall and temperature prediction in South Asia and reported that no single model performs best in all situations. These findings support the need to compare different model families under a common validation framework rather than assuming that complex models will always perform better.

Agricultural forecasting studies by Rathod and co-workers also highlight the importance of weather-based modelling, model diagnostics and practical decision support. Rathod *et al.* (2022)^[29] used climate variables, count time-series models, and machine learning approaches to predict rice gall midge populations. Although the target variable was pest incidence rather than rainfall, the study is relevant because it demonstrates how weather-driven agricultural forecasting can combine exploratory analysis, model comparison, and diagnostic checks. Saha *et al.* (2020)^[33] also demonstrated the importance of spatio-temporal modelling in rainfall forecasting, showing that rainfall behaviour is influenced by both time and location.

Overall, the literature shows a clear movement from classical linear models to nonlinear machine learning and deep learning methods. ARIMA, SARIMA and GARCH-type models are useful as statistical baselines, but they may not fully handle nonlinear and zero-inflated rainfall behaviour. Machine learning models such as Decision Tree, Random Forest, SVR and XGBoost provide more flexibility for rainfall prediction. Deep learning models such as ANN, CNN, RNN and LSTM can capture more complex temporal patterns, especially when enough sequential and atmospheric information is available. However, their success depends on data quality, input variables and proper tuning.

Despite these advances, there is still a need for more local-scale studies that compare statistical, machine learning and deep learning models for daily rainfall forecasting under highly zero-inflated conditions. Many earlier studies focused on district, basin, regional or national scales, while mandal-level daily rainfall prediction has received less attention. The present study addresses this gap by evaluating ARIMAX, SARIMAX, GARCH, Decision Tree, Random Forest, SVR, XGBoost, ANN, CNN, RNN and LSTM models under a common validation framework. By using rainfall and temperature data along with lag, rolling and calendar-based

features, this study provides a localized assessment of daily rainfall prediction for Dandepalle Mandal, Telangana.

3. Materials and Methods

3.1 Study area and data description

In the present study, secondary daily data on maximum temperature, minimum temperature and rainfall for the period 1995-2024 were obtained from Rajendranagar for Dandepalle Mandal in Mancherial district, Telangana, India. The mandal falls under the Northern Telangana Zone and is located at 18.99° N latitude and 79.15° E longitude, with an altitude of 173 m above mean sea level. The agroclimatic region is characterized by monsoon-dependent rainfall, warm summers, black and red soil tracts, and cropping systems that include rice, cotton, maize and pulses depending on irrigation availability. The study region is important because of its

susceptibility to dry spells and intense rainfall. These weather conditions directly affect planting schedules, irrigation timing and other agricultural practices. In total, the dataset includes 10,958 daily observations for the period January 1995 to December 2024. Training
The collected secondary data on daily maximum temperature, minimum temperature and rainfall were systematically organized, evaluated and statistically analyzed to identify long-term trends and variability patterns. The analysis focused on understanding seasonal and annual variations in rainfall over the study period from 1995 to 2024, with emphasis on their possible influence on agricultural activities in Dandepalle Mandal of Mancherial district, Telangana. The location map of Dandepalle Mandal in Mancherial district is shown in Fig. 1.

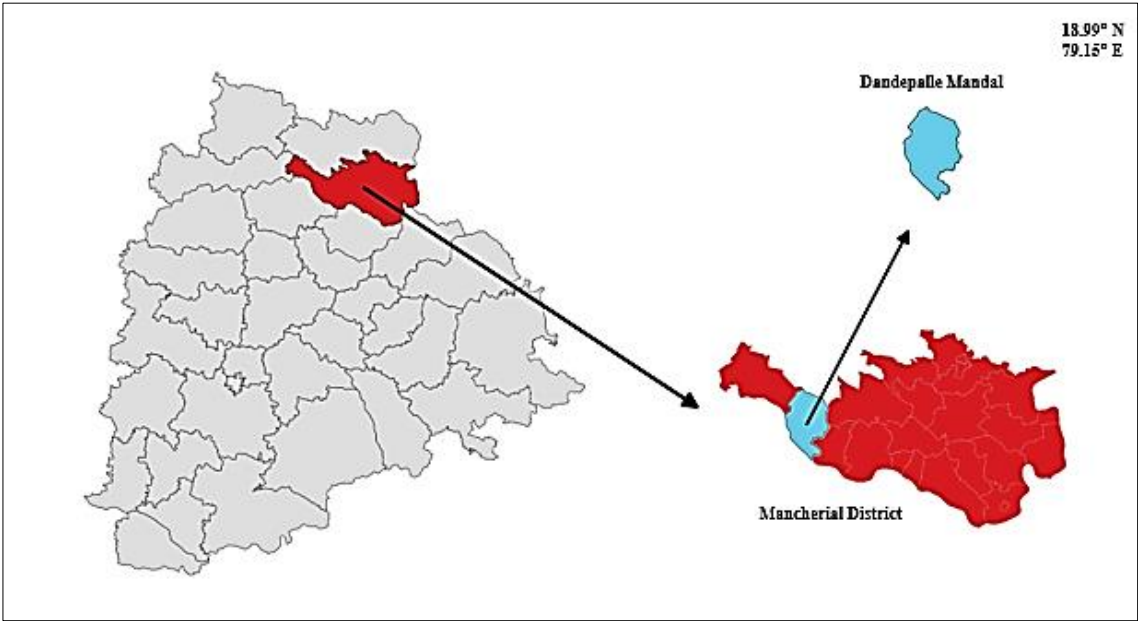


Fig 1: Location map of Dandepalle Mandal, Mancherial district, Telangana

3.2 Methodological framework

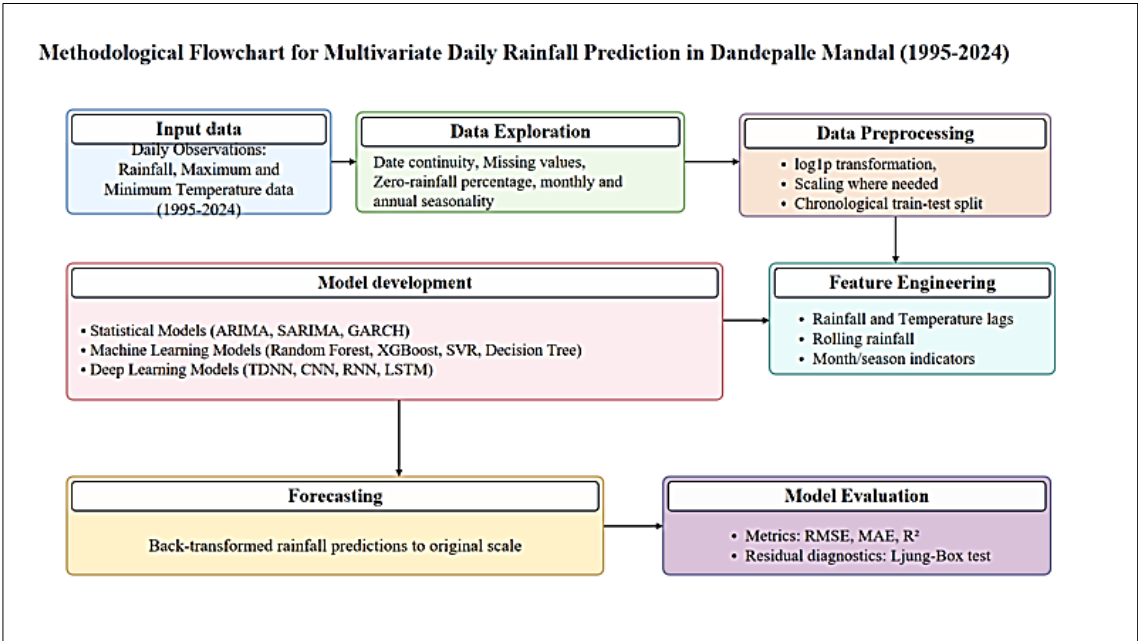


Fig 2: Methodological workflow for multivariate daily rainfall prediction in Dandepalle Mandal

3.3 Data preprocessing and feature engineering

For rainfall modelling, log1p transformation was applied to reduce skewness and the influence of extreme rainfall values.

$$y_t^* = \log(1 + y_t) \quad (1)$$

After prediction, values were converted back to the original scale using:

$$\hat{y}_t = \exp(\hat{y}_t^*) - 1 \quad (2)$$

Where y_t is the original rainfall value, y_t^* is the transformed rainfall value, and $\hat{y}_t$ is the back-transformed predicted rainfall.

In the machine learning models, lagged rainfall and temperature values, rolling rainfall and temperature statistics, and rain/no-rain history information were used as input features. For deep learning models, sequential input data were created using rainfall, maximum temperature, and minimum temperature, with the sequence length depending on the model. Continuous variables were standardized wherever needed, especially for Support Vector Regression (SVR), and scaled for neural network training. All model evaluation metrics, including RMSE, MAE, and R^2 , were calculated using rainfall values on the original scale after back-transforming the predictions.

3.4 Model development

3.4.1 Linear trend analysis

Linear trend analysis was used to examine the direction and rate of change in the weather variables over time.

$$y_t = \beta_0 + \beta_1 t + \varepsilon_t \quad (3)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (4)$$

where y_t is the observed value at time t , β_0 is the intercept, β_1 is the trend coefficient, ε_t is the random error, and R^2 represents the goodness of fit.

3.4.2 Statistical, machine learning and deep learning models:

The weather variables, namely maximum temperature, minimum temperature and rainfall, were modelled using statistical, machine learning and deep learning approaches. The models used in the study included linear trend analysis, ARIMAX, SARIMAX, GARCH, SVR, Random Forest, XGBoost, ANN, CNN, TDNN, RNN and LSTM.

In the multivariate models, lagged rainfall values were used along with maximum temperature and minimum temperature as exogenous predictors.

3.4.2.1 ARIMAX Model

The Autoregressive Integrated Moving Average with Exogenous Variables (ARIMAX) model is an extension of the ARIMA model in which external explanatory variables are included along with the autoregressive, differencing and moving average components.

The general form of the ARIMAX model is expressed as

$$\phi_p(B)(1-B)^d(Y_t - \beta_0 - \sum_{j=1}^k \beta_j X_{jt}) = \theta_q(B)\varepsilon_t \quad (5)$$

where Y_t is rainfall at time t , X_{jt} is the j^{th} exogenous variable at time t , β_0 is the intercept, β_j are regression coefficients, B is the backshift operator, p is the autoregressive order, d is the differencing order, q is the moving-average order, $\phi_p(B)$ is the autoregressive polynomial, $\theta_q(B)$ is the moving-average polynomial and ε_t is the random error term.

3.4.2.2 SARIMAX Model

The Seasonal Autoregressive Integrated Moving Average with Exogenous Variables (SARIMAX) model is an extension of the SARIMA model in which seasonal components and external explanatory variables are included in the same forecasting framework. The general form of the SARIMAX model is expressed as:

$$\phi_p(B)\phi_p(B^s)(1-B)^d(1-B^s)^D(Y_t - \beta_0 - \sum_{j=1}^k \beta_j X_{jt}) = \theta_q(B)\theta_q(B^s)\varepsilon_t \quad (6)$$

where Y_t is rainfall at time t , X_{jt} is the j^{th} exogenous variable at time t , β_0 is the intercept, β_j are regression coefficients, B is the backshift operator, p is the autoregressive order, d is the differencing order, q is the moving-average order, $\phi_p(B)$ is the autoregressive polynomial, $\theta_q(B)$ is the moving-average polynomial and ε_t is the random error term.

3.4.2.3 Generalized Autoregressive Conditional Heteroskedasticity (GARCH)

The Generalized Autoregressive Conditional Heteroskedasticity (GARCH) model, proposed by Bollerslev (1986) [1], is employed to model time series with time-varying conditional variance. The GARCH (p, q) model is defined as:

$$h_t = a_0 + \sum_{i=1}^q a_i \varepsilon_{t-i}^2 + \sum_{j=1}^p b_j h_{t-j} \quad (7)$$

The most popular GARCH model in applications has been the GARCH (1, 1) model.

3.4.2.4 Support Vector Regression (SVR)

SVR was used for nonlinear regression through kernel-based transformation.

$$f(x_i) = w^T \phi(x_i) + b \quad (8)$$

$$\min_{w, b, \xi, \xi^*} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n (\xi_i + \xi_i^*) \quad (9)$$

Subject to

$$\begin{aligned} y_i - f(x_i) &\leq \varepsilon + \xi_i \\ f(x_i) - y_i &\leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* &\geq 0 \end{aligned} \quad (10)$$

Where C is the regularization parameter, ε is the insensitive loss margin, and ξ_i, ξ_i^* are slack variables.

3.4.2.5 Random Forest (RF): Random Forest was applied as an ensemble tree-based model, where final prediction was obtained by averaging the outputs of individual trees.

$$\hat{y} = \frac{1}{M} \sum_{m=1}^M T_m(x) \quad (11)$$

where M is the number of trees and $T_m(x)$ is the prediction from the m^{th} tree.

3.4.2.6 Extreme Gradient Boosting (XGBoost)

XGBoost was used as a regularized gradient boosting model.

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i), \quad f_k \in \mathcal{F} \dots \quad (12)$$

$$\text{Obj} = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k) \dots \quad (13)$$

$$\Omega(f_k) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2 \quad (14)$$

where l is the loss function and $\Omega(f_k)$ is the regularization term.

3.4.2.7 Time Delay Neural Network (TDNN)

TDNN was used to model nonlinear relationships using lagged weather inputs.

$$\hat{y}_{t+h} = F(y_t, y_{t-1}, \dots, y_{t-p}) \dots \quad (15)$$

$$z = \sum_{i=1}^n w_i x_i + b \dots \quad (16)$$

$$a = \sigma(z) \dots \quad (17)$$

Where x_i, w_i, b, z and σ represent input, weight, bias, weighted sum and activation function, respectively.

3.4.2.8 Recurrent Neural Network (RNN)

RNN was used for sequential time series modelling by retaining information from previous time steps.

$$h_t = \sigma(W_{xh}x_t + W_{hh}h_{t-1} + b_h) \dots \quad (18)$$

$$\hat{y}_t = W_{hy}h_t + b_y \dots \quad (19)$$

where h_t is the hidden state at time t , x_t is the input and $\hat{y}_t$ is the predicted output.

3.4.2.9 Long Short-Term Memory (LSTM)

LSTM was used to capture short-term and long-term temporal dependencies using gated memory cells.

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \dots \quad (20)$$

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \dots \quad (21)$$

$$\tilde{C}_t = \tanh(W_C[h_{t-1}, x_t] + b_C) \dots \quad (22)$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \dots \quad (23)$$

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \dots \quad (24)$$

$$h_t = o_t \odot \tanh(C_t) \dots \quad (33)$$

where f_t, i_t, o_t, C_t , and h_t represent forget gate, input gate, output gate, cell state and hidden state, respectively.

3.5 Evaluation metrics and residual diagnostics

Remaining autocorrelation was examined using the Ljung-Box statistic, and first-order residual correlation was assessed using the Durbin-Watson statistic.

Model performance was evaluated using Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) and Coefficient of Determination (R^2).

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \dots \quad (25)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \dots \quad (26)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \dots \quad (27)$$

where y_i and $\hat{y}_i$ are the observed and predicted values, respectively, and n is the total number of observations.

4. Results and Discussion

4.1 Descriptive statistics of original and log1p-transformed rainfall data

The descriptive statistics of rainfall showed that the daily rainfall data were highly skewed with substantial variability. Among the 10,958 observations, 7,646 days (69.78%) recorded zero rainfall, while 3,312 days (30.22%) were classified as wet days. The original rainfall series had a mean of 3.95 mm and a median of 0.00 mm, indicating that dry days dominated the study period. The maximum daily rainfall was 448.10 mm, while the 99th percentile was 63.53 mm, highlighting the occurrence of a few extreme rainfall events. The high skewness (9.27) and kurtosis (160.22) values further confirmed the presence of a highly asymmetric distribution with extreme observations. After applying the log1p transformation, the mean changed to 0.56 and the standard deviation decreased from 14.03 to 1.09. The skewness was reduced from 9.27 to 1.95, while the kurtosis decreased from 160.22 to 2.74. These changes indicate that the transformation effectively reduced the influence of extreme rainfall values and improved the data's distributional properties while preserving zero-rainfall observations. Therefore, the transformed rainfall series was more suitable for subsequent statistical and machine learning analyses.

Table 1. Summary statistics of original and log1p-transformed rainfall series.

Statistic	Original Rainfall (mm)	Log1p Rainfall
No. of Observations	10,958	10,958
Mean	3.95	0.56
Median	0.00	0.00
Standard Deviation	14.03	1.09

Variance	196.95	1.19
Minimum	0.00	0.00
99 th Percentile	63.53	4.17
Maximum	448.10	6.11
Skewness	9.27	1.95
Kurtosis	160.22	2.74
Zero Rainfall Days	7,646	7,646
Zero Rainfall Percentage	69.78%	69.78%
Wet Rainfall Days	3,312	3,312
Wet Rainfall Percentage	30.22%	30.22%

Fig. 3 presents the log1p-transformed daily rainfall series for Dandepalle during 1995-2024. The rainfall data exhibit significant day-to-day variability, highlighting the

irregularity of rainfall events. The 365-day rolling mean displays moderate fluctuations, while the fitted linear trend suggests a slight increase in rainfall over time.

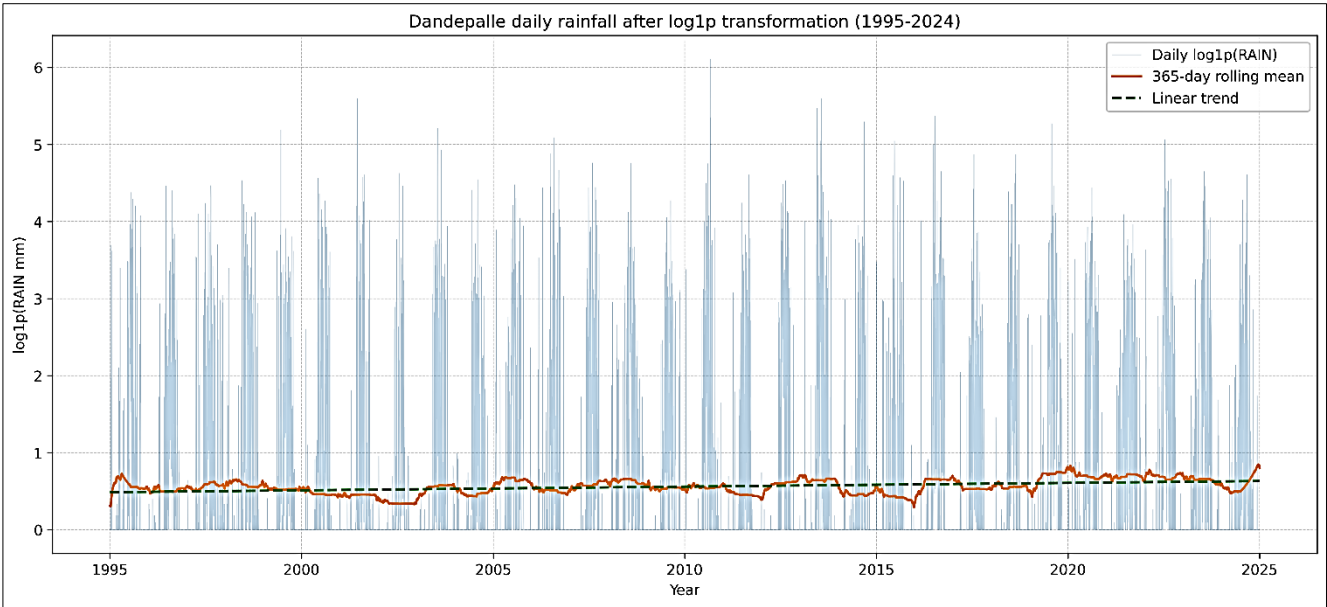


Fig 3: Long-term variability and trend of log-transformed daily rainfall in Dandepalle Mandal (1995-2024).

Fig. 4 further illustrates the distributional nature of the rainfall series before and after log1p transformation. The original rainfall distribution had a strong concentration near

zero with a long right tail, while the transformed series reduced the influence of extreme rainfall values and retained the zero-rainfall observations.

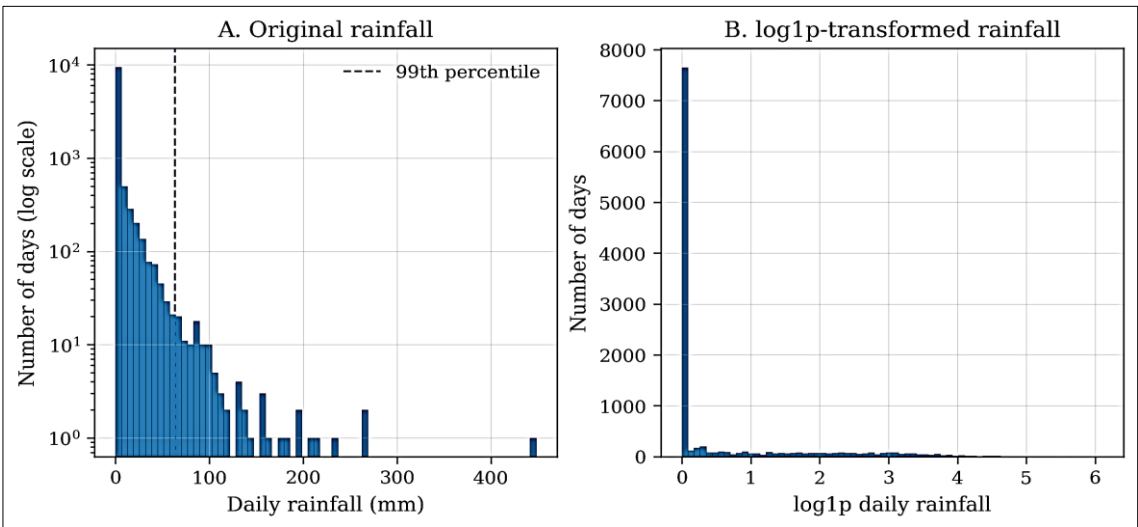


Fig 4: Distribution of original and log1p-transformed daily rainfall in Dandepalle Mandal during 1995-2024

Fig. 5 shows the monthly rainfall distribution during the study period. Rainfall was concentrated mainly during the

southwest monsoon months, while the non-monsoon months recorded low rainfall totals and larger dry-period dominance.

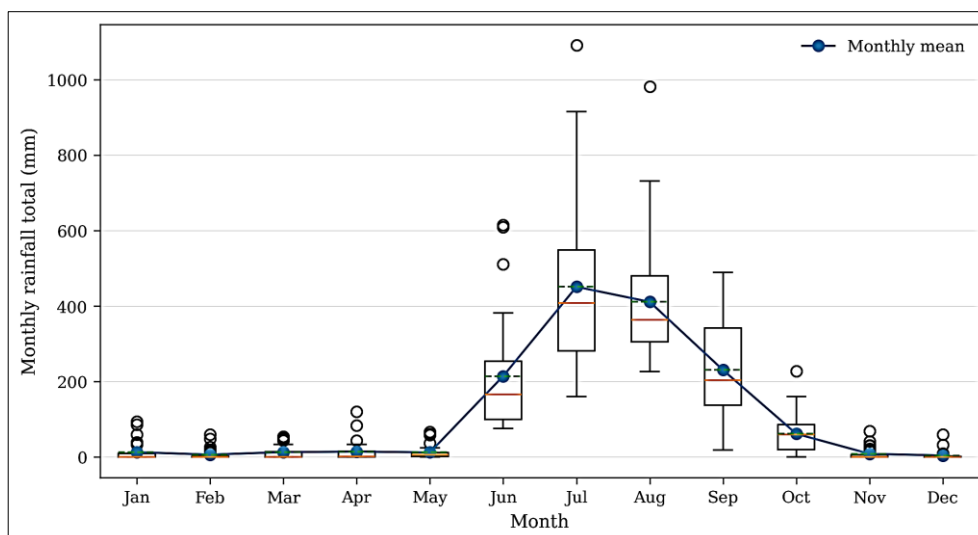


Fig 5: Monthly rainfall distribution in Dandepalle Mandal during 1995-2024

Fig. 6 presents the annual rainfall total and fitted linear trend for Dandepalle Mandal. The annual totals show clear inter-

annual variability, while the fitted trend indicates a slight increasing tendency over the 1995-2024 period.

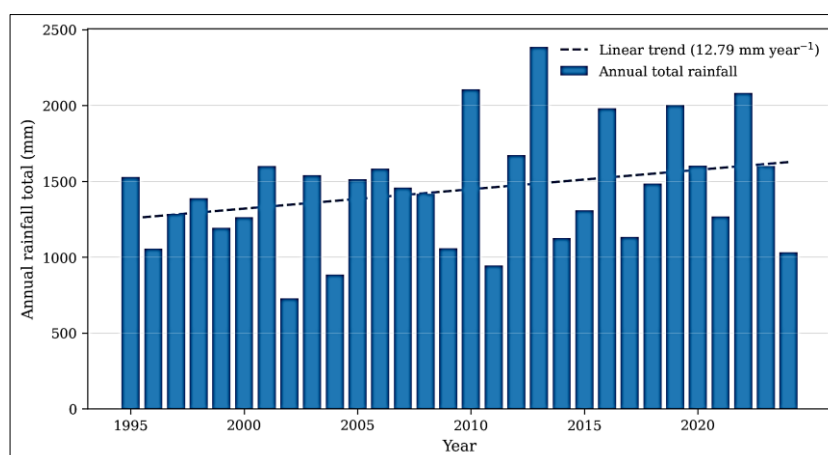


Fig 6: Annual rainfall trend in Dandepalle Mandal during 1995-2024.

4.2 Comparative model performance and diagnostic interpretation

In the present study, various statistical, machine learning and deep learning approaches were used for the prediction of rainfall. To evaluate their performance, several metrics such as RMSE, R^2 , residual mean, Ljung-Box statistics and Durbin-Watson statistics were used.

The comparative performance of the statistical, machine learning and deep learning models is presented in Table 2. Among the various models tested, the Decision Tree model achieved the lowest test RMSE of 8.7019 mm, with an R^2 of 0.1075. XGBoost ranked second, with an RMSE of 8.8180 mm and an R^2 value of 0.0836. The SVR model also performed at a similar level, with an RMSE of 8.8497 mm and an R^2 of 0.0770. The difference among the Decision Tree, XGBoost and SVR models was small, indicating that these models performed nearly similarly for daily rainfall prediction in Dandepalle Mandal. This result supports the use of nonlinear machine learning approaches for agricultural and weather-based time-series forecasting. Gorlapalli *et al.* (2022)^[12], in a water-stress forecasting study, compared time-series and artificial intelligence models and demonstrated the value of AI-based approaches, such as SVR, in agro-meteorological forecasting. Similarly, Saha *et al.* (2021)^[33]

used a tuned SVR to forecast cotton production and reported that SVR-based methods can outperform traditional time-series models when nonlinear patterns are present. In the present study, SVR outperformed ARIMAX, GARCH, and SARIMAX, suggesting that kernel-based regression was better suited to the nonlinear nature of daily rainfall.

XGBoost ranked second and outperformed SVR, most deep learning and traditional time-series models. Similar findings were reported by Mishra *et al.* (2024)^[22], who identified that XGBoost was useful for modelling rainfall patterns in India. El Hafyani *et al.* (2024)^[10] also reported the usefulness of machine learning models, including XGBoost-type approaches, for precipitation prediction using meteorological inputs. In the present study, XGBoost was useful but did not outperform the Decision Tree. This may be because the rainfall series had many zero-rainfall days and only a few extreme rainfall events, making it difficult for boosting models to learn a strong and stable rainfall signal.

The deep learning models showed moderate performance. Among them, RNN performed better than the other neural network models, with an RMSE of 8.8634 mm and an R^2 value of 0.0741. It was followed by LSTM, CNN and ANN. Although LSTM is generally more advanced than a simple RNN, it performed slightly worse in this study. This suggests

that long-term memory did not provide much extra advantage for the present dataset. One possible reason is that the available input variables were limited to rainfall, maximum temperature and minimum temperature. Daily rainfall is also influenced by other atmospheric factors such as humidity, pressure, wind movement, cloud cover, monsoon circulation and local convection. Since these variables were not included, the LSTM model may not have had sufficient information to effectively learn rare and extreme rainfall events. Wani *et al.* (2024) [36] also showed that the performance of machine learning, deep learning and time-series models in rainfall prediction depends strongly on data quality, location and available meteorological inputs. Mishra *et al.* (2025) [25] further reported that no single model performs uniformly well for rainfall and temperature forecasting across different South Asian regions, which agrees with the present findings. CNN and ANN showed weaker performance than RNN and LSTM. ANN can capture nonlinear relationships, but it does not naturally preserve time order unless lag variables are included. Therefore, these models may not have captured the temporal behaviour of rainfall as effectively as the recurrent models. These results show that deep learning models do not always outperform simpler models simply because they are more complex. For highly variable rainfall data, model performance depends not only on model architecture but also on the strength of the rainfall signal and the quality of input variables.

The traditional time-series models, namely ARIMAX, GARCH and SARIMAX, showed the weakest prediction performance. ARIMAX recorded an RMSE of 9.4367 mm and a R² value of 0.0495. GARCH and SARIMAX performed even weaker, with RMSE values of 9.5817 mm and 9.6324 mm, respectively. Their low R² values show that these models performed worse than a simple mean-based prediction. This suggests that the daily rainfall series in Dandepalle Mandal was too irregular and intermittent for these models to capture effectively. ARIMAX mainly captures linear autocorrelation, GARCH focuses on changing variance, and SARIMAX captures regular seasonal patterns. However, rainfall in the study area appears to be influenced by irregular wet and dry spells and sudden extreme rainfall events, which are difficult to represent using these models alone. Though ARIMAX and SARIMAX did not perform well in the present study, they remain useful statistical baselines in agricultural weather forecasting. Patil *et al.* (2022) [26] used seasonal ARIMA for

forecasting potential evapotranspiration for irrigation planning. This shows that SARIMA-type models can be useful when the weather variable has a more regular seasonal structure.

The Ljung-Box test was used to check whether the model errors still had time-based patterns. In this study, the significant Ljung-Box results showed that some temporal structure still remained in the model errors. This means that even the better-performing models did not fully capture the rainfall process. These remaining patterns may be linked to wet spells, dry spells, seasonal rainfall persistence or other atmospheric variables that were not included in the model. Rathod *et al.* (2022) [29], in their weather-based rice gall midge prediction study, also emphasized the importance of model diagnostics and comparison when developing agricultural forecasting models. Decision Tree, XGBoost and SVR were more effective because they could capture nonlinear and threshold-based rainfall behaviour. The deep learning models gave moderate results but did not outperform the best machine learning models. This suggests that complex neural networks may need richer input data, such as humidity, pressure, wind speed, cloud cover, soil moisture or spatial rainfall information, to improve daily rainfall prediction. Overall, the Decision Tree model was the most effective model in terms of RMSE, followed by XGBoost and then SVR. However, the low R² values across all models confirm that daily rainfall forecasting remains a difficult task. The models reduced prediction error to some extent, but they could not explain most of the day-to-day variation in rainfall. This is mainly because rainfall is controlled by complex atmospheric processes, seasonal changes, local conditions and sudden extreme events. From an agricultural perspective, the results suggest that machine learning models, especially Decision Tree, XGBoost and SVR, can support rainfall-based decision-making in Dandepalle Mandal. These models may be useful for identifying general rainfall trends to inform irrigation scheduling, field operations, pesticide application, and farm planning. However, their outputs should not be used as exact rainfall values. The predictions should be combined with field knowledge, official weather forecasts and real-time observations. Future improvement may be possible by adding more weather variables, atmospheric predictors, soil moisture information, spatial rainfall data and separate rain/no-rain evaluation metrics.

Table 2: Comparative evaluation of forecasting models based on test-period accuracy and residual diagnostics.

Sl. No.	Model	Test RMSE (mm)	Test R ²	Ljung-Box χ^2 test
1	Decision Tree	8.7019	0.1075	$\chi^2 = 61.8566, p < 0.0001$
2	XGBoost	8.8180	0.0836	$\chi^2 = 55.0278, p < 0.0001$
3	SVR	8.8497	0.0770	$\chi^2 = 33.2530, p < 0.001$
4	RNN	8.8634	0.0741	$\chi^2 = 36.1103, p < 0.0001$
5	LSTM	8.8868	0.0692	$\chi^2 = 39.1495, p < 0.0001$
6	ANN	8.9663	0.0525	$\chi^2 = 53.4851, p < 0.0001$
7	Random Forest	9.0417	0.0365	$\chi^2 = 43.1790, p < 0.0001$
9	ARIMAX	9.4367	0.0495	$\chi^2 = 205.2420, p < 0.0001$
10	GARCH	9.5817	0.0820	$\chi^2 = 212.5929, p < 0.0001$
11	SARIMAX	9.6324	0.0935	$\chi^2 = 212.802, p < 0.0001$

Figs. 7A-C present separate observed and predicted daily rainfall values during the 2024 test period for the three best-performing machine learning models. The separate model-wise plots show that the models followed the general wet-

season pattern, but several high rainfall peaks were underestimated. This supports the interpretation that RMSE values alone should be interpreted with caution for highly intermittent rainfall data.

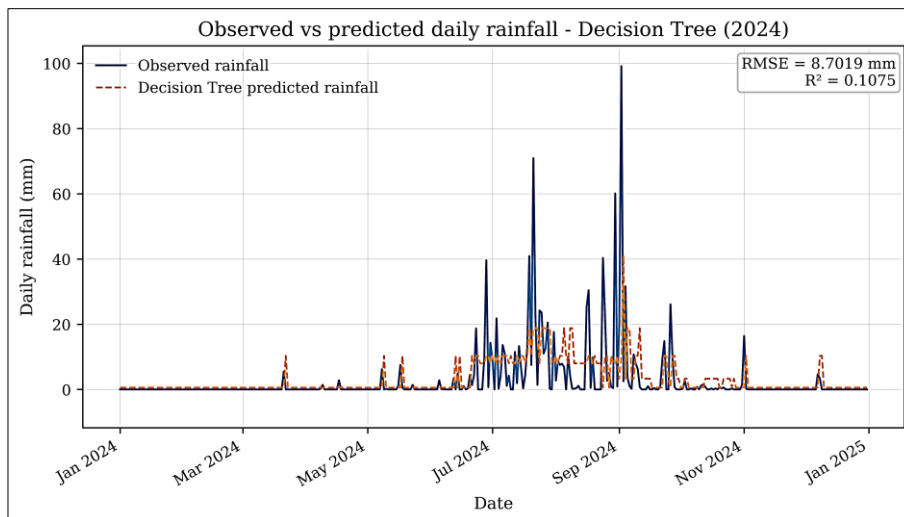


Fig 7A: Observed and predicted daily rainfall during the test period using the Decision Tree model in Dandepalle Mandal.

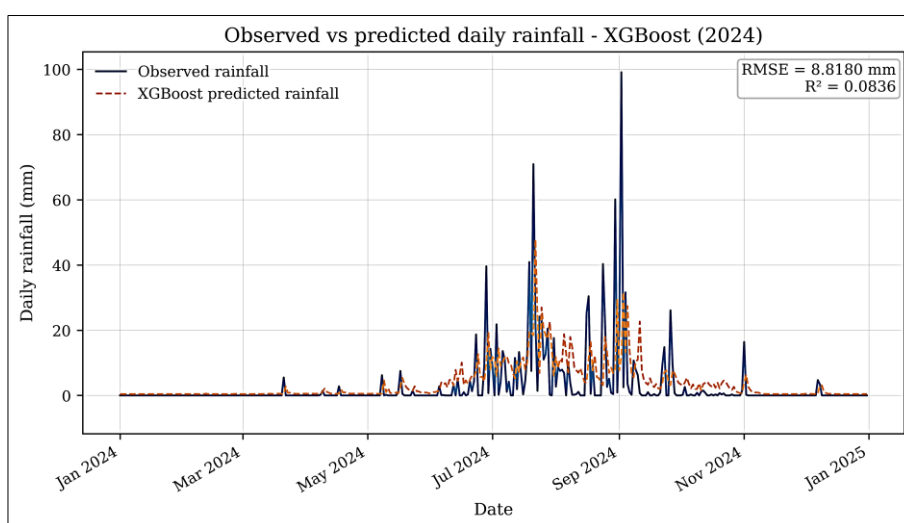


Fig 7B: Observed and predicted daily rainfall during the test period using the XGBoost model in Dandepalle Mandal.

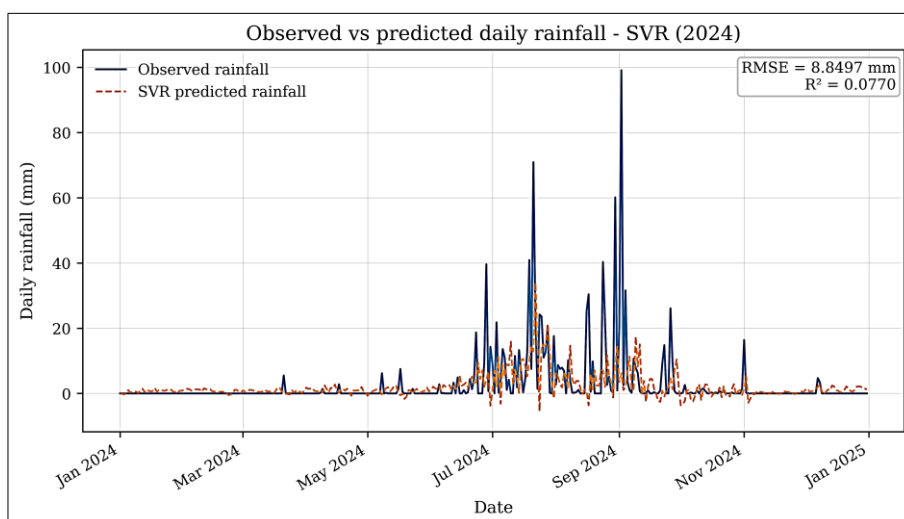


Fig 7C: Observed and predicted daily rainfall during the test period using the SVR model in Dandepalle Mandal.

5. Conclusion

In this study, statistical, machine learning and deep learning models were used to forecast daily rainfall in Dandepalle Mandal, Telangana. The input data included rainfall, maximum temperature and minimum temperature. The rainfall data were very irregular, with many days having no

rain and only a few days having high or heavy rainfall. As a result, all the models had difficulty accurately predicting daily rainfall.

Out of all the models, the Decision Tree achieved the lowest RMSE. XGBoost and SVR also performed well, almost similar to the Decision Tree. This shows that nonlinear

machine learning models performed better on this rainfall data than traditional time-series models.

The deep learning models gave moderate results, but they did not do better than the best machine learning models. This suggests that using complex neural networks does not always give better results, especially when there are only a few input variables. In this case, rainfall, maximum temperature, and minimum temperature were not enough to fully explain the changes in daily rainfall. The traditional time-series models like ARIMAX, GARCH, and SARIMAX did not perform well and has very low R^2 values. This means they could not capture the irregular and sudden changes in daily rainfall in the study area. One reason for this could be the large number of dry days, short wet periods, and sudden heavy rainfall events. Even though the Decision Tree model had the lowest prediction error, the R^2 values for all models remained low. The Ljung-Box test also showed that time-dependent patterns remained in the model errors. From an agricultural point of view, these results show that machine learning models like Decision Tree, XGBoost, and SVR can help with rainfall-based decisions in Dandepalle Mandal. They may be useful for irrigation scheduling, crop planning, pesticide use, and planning field operations. Still, the predictions are not exact rainfall values and should be used together with weather forecasts, field knowledge, and real-time observations. Future studies may also explore hybrid wavelet-ANN, wavelet-WNN and EEMD-LSTM approaches to capture multiscale and nonlinear rainfall behaviour more effectively by adding more weather variables like relative humidity, wind speed, solar radiation, atmospheric pressure, cloud cover, and soil moisture. Using spatial rainfall data, satellite weather information, and separate models for rain vs no rain could also improve predictions. Overall, these results show that simple, easy-to-understand machine learning models can support local rainfall advisory systems, but more work is needed to achieve reliable forecasting.

Acknowledgements

The authors gratefully acknowledge the Department of Agrometeorology, Professor Jayashankar Telangana State Agricultural University (PJTSAU), and the Indian Council of Agricultural Research (ICAR) for providing the meteorological data used in this study.

Authors' Contributions

- Roshini Priya C H: Methodology, Software, Formal analysis, Data curation, Writing - original draft.
- Meena A: Supervision, Methodology, Validation, review and editing.
- Santosha Rathod: Conceptualization, Resources, review and editing.
- B. Balaji Naik: Resources, Validation, Supervision, review and editing.
- Competing Interests
- Authors have declared that no competing interests exist.
- Data Availability
- The meteorological data used in this study were obtained from Agro Climate Research Centre (ACRC), Agricultural Research Institute (ARI), PJTSAU, Rajendranagar.
- Consent
- Not applicable.
- Ethical Approval

- This study used meteorological data only and did not involve human participants, animals or private personal information.

References

1. Bollerslev T. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*. 1986;31(3):307-327.
2. Box GEP, Jenkins GM. Time series analysis: forecasting and control. San Francisco: Holden-Day; 1970.
3. Box GEP, Jenkins GM, Reinsel GC, Ljung GM. Time series analysis: forecasting and control. 5th ed. Hoboken (NJ): Wiley; 2015.
4. Breiman L. Random forests. *Machine Learning*. 2001;45(1):5-32.
5. Chandu K, Dharma Raju A, Kumar SVJ, Satyanarayana GCh, *et al*. Rainfall in Telangana, India: a decadal trend analysis. *Proceedings of the Indian National Science Academy*. 2025.
6. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; 2016. p. 785-794.
7. Cortes C, Vapnik V. Support-vector networks. *Machine Learning*. 1995;20(3):273-297.
8. Dawson CW, Wilby RL. Hydrological modelling using artificial neural networks. *Progress in Physical Geography*. 2001;25(1):80-108.
9. Durai VR, Bhavani Bhate P. Evaluation of district-level rainfall forecasts from India Meteorological Department. *Meteorological Applications*. 2020;27:e1961.
10. El Hafyani MEM, El Himdi KEK, El Adlouni SE. Improving monthly precipitation prediction accuracy using machine learning models: a multi-view stacking learning technique. *Frontiers in Water*. 2024.
11. Espeholt L, Agrawal S, Sønderby C, Kumar M, Heek J, Bromberg C, *et al*. Deep learning for twelve hour precipitation forecasts. *Nature Communications*. 2022;13:5145.
12. Gorlapalli A, Kallakuri S, Sreekanth PD, Patil R, Bandumula N, Ondrasek G, *et al*. Characterization and prediction of water stress using time series and artificial intelligence models. *Sustainability*. 2022;14(11):6690.
13. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Computation*. 1997;9(8):1735-1780.
14. Hong WC, Pai PF. Potential assessment of the support vector regression technique in rainfall forecasting. *Water Resources Management*. 2007;21(3):495-513.
15. Hyndman RJ, Khandakar Y. Automatic time series forecasting: the forecast package for R. *Journal of Statistical Software*. 2008;27(3):1-22.
16. Jyostna B, Meena A, Rathod S, Tuti MD, Choudhary K, Lama A, *et al*. Multiscale rainfall forecasting using a hybrid ensemble empirical mode decomposition and LSTM model. *Modeling Earth Systems and Environment*. 2025;11:119. DOI:10.1007/s40808-024-02220-y.
17. Jyostna B, Meena A, Rathod S, Tuti MD, Kallakuri S, Bhanusree D, Reddy BNK. Trend analysis of rainfall in Telangana State (India) using advanced statistical approaches. *International Journal of Environment and Climate Change*. 2022;12(11):3404-3412. DOI:10.9734/IJECC/2022/v12i111391.

18. Kratzert F, Klotz D, Brenner C, Schulz K, Herrnegger M. Rainfall-runoff modelling using Long Short-Term Memory networks. *Hydrology and Earth System Sciences*. 2018;22(11):6005-6022.
19. Lam R, Sanchez-Gonzalez A, Willson M, Wirsberger P, Fortunato M, Alet F, *et al.* Learning skillful medium-range global weather forecasting. *Science*. 2023.
20. Le XH, Ho HV, Lee G, Jung S. Application of Long Short-Term Memory neural network for flood forecasting. *Water*. 2019;11(7):1387.
21. Luk KC, Ball JE, Sharma A. An application of artificial neural networks for rainfall forecasting. *Mathematical and Computer Modelling*. 2001;33(6-7):683-693.
22. Mishra P, Al Khatib AMG, Yadav S, Ray S, Lama A, Kumari B, *et al.* Modeling and forecasting rainfall patterns in India: a time series analysis with XGBoost algorithm. *Environmental Earth Sciences*. 2024;83:163.
23. Mishra P, Ray S, Lal P, Nair SB, Matuka A, Tashkandy Y, Emam W. Climate modeling for South Asia: statistical and deep learning for rainfall and temperature prediction. *Scientific Reports*. 2025;15:38235.
24. Nanda SK, Tripathy DP, Nayak SK, Mohapatra S. Prediction of rainfall in India using artificial neural network models. *International Journal of Intelligent Systems and Applications*. 2013;5(12):1-22.
25. Pai DS, Rajeevan M, Sreejith OP, Mukhopadhyay B, Satbha NS. Development of a new high spatial resolution daily gridded rainfall dataset over India. *MAUSAM*. 2014;65(1):1-18.
26. Patil R, Nagaraj DM, Polisgowdar BS, Rathod S. Forecasting potential evapotranspiration for Raichur district using seasonal ARIMA model. *MAUSAM*. 2022;73(2):433-440.
27. Pham QB, Yang TC, Kuo CM, Tseng HW, Yu PS. Combining random forest and least square support vector regression for improving extreme rainfall downscaling. *Water*. 2021;13(6):725.
28. Rathod S, Saha A, Patil R, Ondrasek G, Gireesh C, Anantha MS, , *et al.* Two-stage spatiotemporal time series modelling approach for rice yield prediction and advanced agroecosystem management. *Agronomy*. 2021;11(12):2502. doi:10.3390/agronomy11122502.
29. Rathod S, Yerram S, Arya P, Katti G, Rani J, Padmakumari AP, *et al.* Climate-based modeling and prediction of rice gall midge populations using count time series and machine learning approaches. *Agronomy*. 2022;12(1):22.
30. Raval M, Sivashanmugam P, Pham V, Gohel H, Kaushik A, Wan Y. Automated predictive analytics tool for rainfall forecasting. *Scientific Reports*. 2021;11:17704.
31. Ravuri S, Lenc K, Willson M, Kangin D, Lam R, Mirowski P, *et al.* Skillful precipitation nowcasting using deep generative models of radar. *Nature*. 2021;597(7878):672-677.
32. Ray M, Singh KN, Ramasubramanian V, Paul RK, Mukherjee A, Rathod S. Integration of wavelet transform with ANN and WNN for time series forecasting: an application to Indian monsoon rainfall. *National Academy Science Letters*. 2020;43(6):509-513. doi:10.1007/s40009-020-00887-2.
33. Saha A, Singh KN, Ray M, Rathod S, Choudhury S. Modelling and forecasting cotton production using tuned-support vector regression. *Current Science*. 2021;121(8):1090-1098.
34. Senjaliya J, Vagharia V, Shah SM. A comprehensive review of machine learning and deep learning approaches for rainfall forecasting: current progress, challenges, and future directions. *Theoretical and Applied Climatology*. 2026.
35. Shi X, Chen Z, Wang H, Yeung DY, Wong WK, Woo WC. Convolutional LSTM network: a machine learning approach for precipitation nowcasting. In: *Advances in Neural Information Processing Systems*. 2015.
36. Wani OA, Mahdi SS, Yeasin M, Kumar SS, Gagnon AS, Danish F, Al-Ansari N, El-Hendawy S, Mattar MA. Predicting rainfall using machine learning, deep learning, and time series models across an altitudinal gradient in the North-Western Himalayas. *Scientific Reports*. 2024.
37. Zhang Z, Dong J, Georgakis C. Short-term forecasting of rainfall using LSTM. *Geoscientific Model Development*. 2020;13:1-15.