PERFORMANCE OF MACHINE LEARNING AND DEEP LEARNING MODELS FOR PREDICTING RAINFALL IN A LARGE WATERSHED: CASE STUDY OF BENGAWAN SOLO RIVER BASIN, INDONESIA

Jumadi JUMADI [1*,2], ***Kuswaji Dwi PRIYONO*** [3], ***Ali Hasan ABDULLAH*** [3], ***Supari SUPARI*** [4], ***Hamza AIT ZAMZAMI*** [5], ***Farha SATTAR*** [6], ***Muhammad NAWAZ*** [7], ***Hamzah HASYIM*** [8] ***& Steve CARVER*** [9]

DOI: 10.21163/GT_2026.212.05

ABSTRACT

Predicting rainfall in large, heterogeneous watersheds remains among the most important hydrological challenges. This research investigates the effectiveness of both ML (Machine Learning) and DL (Deep Learning) for predicting spatiotemporal rainfall in the Bengawan Solo watershed, Indonesia. Satellite rainfall data from CHIRPS (spatial resolution: 0.05°) were prepared and sampled for the period 1981–2024. The data set contained 523 grid points. We employed nine ML and DL algorithms: Random Forest (RF), Extreme Gradient Boosting (XGB), Support Vector Regression (SVR), Multilayer Perceptron (MLP), Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), Temporal Convolutional Network (TCN), Convolutional Neural Network (CNN), and Transformer. Models were trained on the samples from 1981 to 2019 and tested on 2020–2024. Performance was judged from mean absolute error (MAE), root mean square error (RMSE), mean absolute percentage error (MAPE), and coefficient of determination (R^2). XGB showed the best overall performance (MAE ≈59 mm; R^2 ≈0.73). GRU became the most competitive DL model at 60 mm (MAE ≈60 mm; R^2 ≈0.72). Temporal model analysis shows that XGB and GRU stay among the top three models with the minimum monthly errors. TCN, CNN, and Transformer exhibited higher errors and more monthly variability. XGB and GRU have average MAE values of ~59–60 mm and R^2 values of ~0.71–0.72 across most grids. MAE values for TCN, CNN, and Transformer are greater than 76 mm, while R^2 values are lower. The data we obtained indicate that using ensemble decision tree models and recurrent neural networks across large tropical areas yields greater stability and more reliable spatiotemporal rainfall predictions than more sophisticated DL architectures.

Keywords: *Rainfall prediction; Machine learning; Deep learning; CHIRPS; Bengawan Solo; Spatial analysis; Temporal analysis.*

1. INTRODUCTION

Rainfall prediction is essential for effectively planning water resources, managing disasters, and managing agricultural areas in the tropics (Hong et al., 2018; Praveen et al., 2020). The Bengawan Solo watershed in Indonesia is an area critical to food security, water supply, and flood control (Musiyam et al., 2025). The watershed is a large, socio-economically critical watershed where rainfall

[1*]*Faculty of Geography, Muhammadiyah University Surakarta, Indonesia.*
Corresponding author: jumadi@ums.ac.id (JJ)
[2]*INTI International University, Malaysia*
[3]*Faculty of Geography, Muhammadiyah University Surakarta, Indonesia; kuswaji.priyono@ums.ac.id (KDP), e100210132@student.ums.ac.id (AHA)*
[4]*Agency for Meteorology, Climatology and Geophysics (BMKG), Indonesia; supari@bmkg.go.id (SS)*
[5]*University of Hassan II Casablanca, Casablanca, Morocco; hamza.aitzamzami-etu@etu.univh2c.ma (HAZ)*
[6]*Charles Darwin University, Australia; Farha.Sattar@cdu.edu.au (FS)*
[7]*National University of Singapore, Singapore; geomn@nus.edu.sg (MH)*
[8]*Universitas Sriwijaya, Ogan Ilir, South Sumatera, Indonesia; hamzah@fkm.unsri.ac.id (HH)*
[9]*School of Geography, University of Leeds, United Kingdom; S.J.Carver@leeds.ac.uk (SC)*

variability directly affects water availability for irrigation and is closely linked to recurrent hydrometeorological impacts on agriculture and a large population. The basin also exhibits pronounced physical heterogeneity, ranging from flat plains to hilly and volcanic terrain, with diverse soils and land uses, producing strong spatial contrasts in hydrological response and making rainfall prediction more challenging at scale (Santhyami et al., 2025). Thus, long-term rainfall data and rainfall predictors are crucial. Nonetheless, the effects of high climatic variability, topographical complexity, and limited field observation data are major obstacles to the long-term accuracy of rainfall prediction (Funk et al., 2015; Kundu et al., 2022).

Data-driven methods have revolutionized hydrometeorological modeling over time. Machine Learning (ML) methods such as Random Forest (RF), Extreme Gradient Boosting (XGB), Support Vector Regression (SVR), and Multilayer Perceptron (MLP) to examine in depth the associations between predictive variables and rainfall predictors (Sachindra et al., 2018; Miao et al., 2021) are used more frequently to discover nonlinear relationships between predictors and rainfall. Concurrently, deep learning (DL) approaches like Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), Convolutional Neural Network (CNN), Temporal Convolutional Network (TCN), and Transformer architectures depict complex spatio-temporal patterns in climate time series (Shi et al., 2015; Chattopadhyay et al., 2020; Espeholt et al., 2022).

Many studies have shown that DL techniques have great potential for accurately modeling hydrometeorological phenomena (Aswin & Geetha, 2020; Hernández et al., 2021), including the prediction of severe and rare rainfall events (Laptev et al., 2017; Lim & Zohren, 2021). But most of that performance depends on the vastness of the data in terms of temporal depth, with a fairly uniform spatial distribution. In data-limited regions, DL models are often weak under time-series conditions due to overfitting, poor training, and generalization issues (Willard et al., 2021; Das et al., 2022). Although ML and DL techniques have made significant progress in predicting rainfall, quite a few research gaps still exist. Comparative studies assessing the performance of ML and DL in tropical regions, such as the Bengawan Solo River Basin, with complex spatial and even topographical features, are limited, especially when using high-resolution satellite data (such as CHIRPS) to inform annual rainfall forecasting.

Furthermore, it is rare for model training to rely on long time series, spanning 40 years or more, despite the fact that long time spans are necessary to account for long-term climate dynamics and interannual variability that impact prediction accuracy (Gu et al., 2020). Furthermore, spatially and temporally differentiated error analysis is underused, even though it is important to identify specific locations and time intervals of model inaccuracy to better prepare for the next model (Cioffi et al., 2023). To fill in the gap, this study conducted a comprehensive evaluation of nine ML and DL models using monthly CHIRPS data for 1981-2024, aggregated to annual rainfall at 98 sample points in the Bengawan Solo watershed. Models were trained on historical data and tested on 2020–2024 conditions, with metrics including MAE, RMSE, MAPE, R^2, spatial errors, and temporal errors. The best-performing model was then adopted as the basis for predicting annual rainfall for 2025–2030. In conclusion, this study assesses the predictive performance of ML and DL in the Bengawan Solo watershed region, using the long-term CHIRPS (1981–2024) and 2025–2030 periods to estimate annual rainfall and to determine which method is best suited to tropical regions with scarce data. The literature shows that when remote sensing series are discontinuous, statistical interpolation approaches can reconstruct long-term trajectories useful for analysis (Haidu et al., 2024).

2. METHOD

2.1. Study Area

The Bengawan Solo River Basin (**Fig. 1**) is the largest river system on Java Island, covering an area of approximately 16,100 km² and serving as the primary source of water for domestic needs and agricultural irrigation. Hydrologically, the watershed is divided into three sub-basins: the Upper Bengawan Solo, the Madiun River, and the Lower Bengawan Solo.

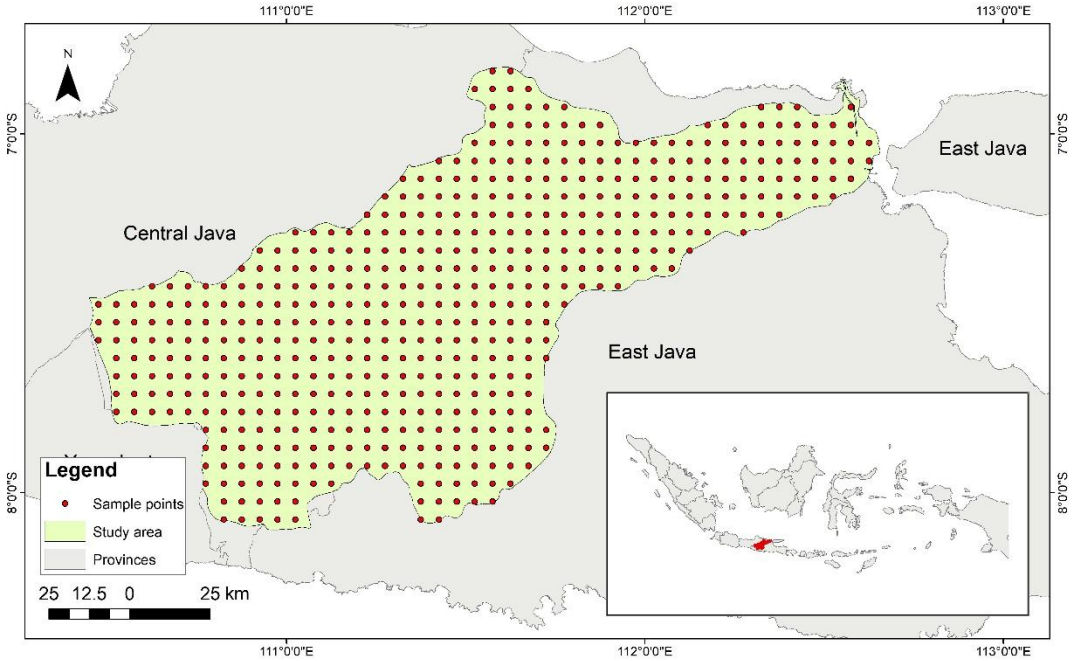

Fig. 1. Study Area and Sample Points.

Its upstream area (approximately 6,000 km²) lies between 110º13'7.16"–110º26'57.10" East Longitude and 7º26'33.15"–8º6'13.81" South Latitude, with predominantly flat topography but undulating in the northeast–northwest near the mountains; flow supply comes from the Merapi–Merbabu volcanic complex in the west and Mount Lawu in the east. The contrasting topographic variations and land use, as well as the occurrence of seasonal floods and droughts, make the Bengawan Solo River Basin a priority location for hydrometeorological studies and water resources management planning (Jumadi et al, 2024; Jumadi et al, 2025).

2.2. Research Framework

This research was conducted in several stages, including data collection, preprocessing, annual rainfall prediction using ML and DL, model evaluation (using data from 2020-2024), selection of the best model, prediction for 2025-2030, data aggregation, interpolation, and spatiotemporal analysis (**Fig. 2**).

2.3. Data and Data Sources

Data aggregation, interpolation, and spatiotemporal analysis were performed in several stages. The study took advantage of satellite rainfall data collected from the Climate Hazards Group InfraRed Precipitation with Stations (CHIRPS) product, issued by the Climate Hazards Center at the University of California, Santa Barbara (Funk et al., 2015). CHIRPS is a global precipitation dataset for rainfall estimation with high spatial resolution (~0.05°, or ±5 km) and long time span (1981 to present). This is a global dataset of infrared satellite observation data, reanalysis data, and rainfall data from land stations. This hybrid allows CHIRPS to generate consistent rainfall estimations, even in parts of the world where ground observation networks are weak such as Indonesia (Ceccherini et al., 2015; Gu et al., 2020). In hydrology, the comparative evaluation of monthly gap imputation methods has proven essential for the stability of analyses (cf. monthly flow study), which supports our focus on the continuity of series (Magyari-Sáska et al., 2025).

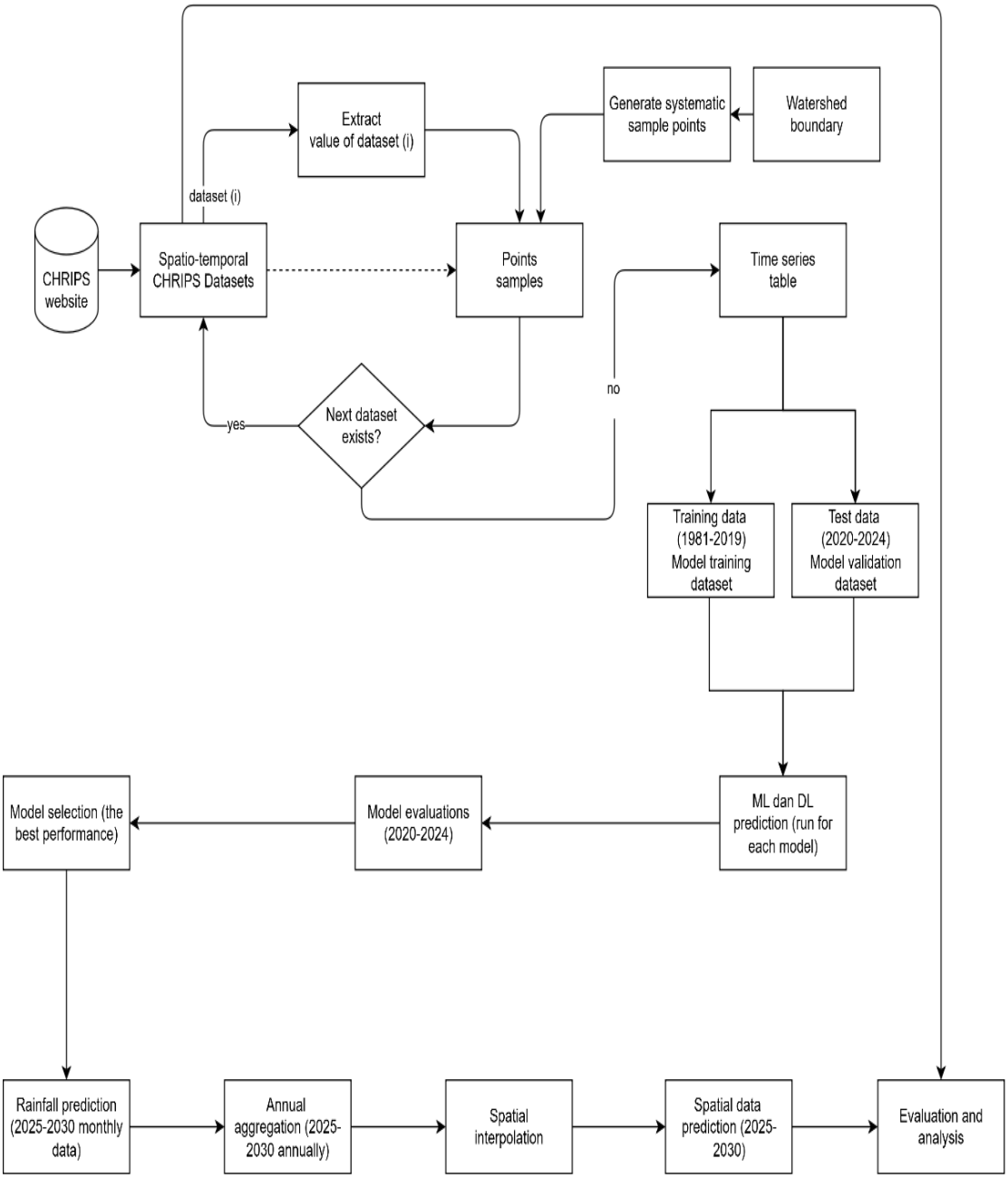

Fig. 2. Research Framework.

The selection of CHIRPS for this study involved three elements. First, its high spatial resolution allows for a better representation of rainfall variability in areas with complex topography such as the Bengawan Solo River Basin. Second, its time series spanning more than four decades (1981-2024) allows a better understanding of long-term trends and interannual climate variability. Third, the availability of open-access data that is routinely updated enables replication of models as well as future updates, which is the key principle of data openness in scientific research (Huffman et al., 2020).

2.4. Determination of Prediction Points

All pixels from CHIRPS were extracted in order to allow for the model prediction to capture spatial variety of rainfall throughout the Bengawan Solo watershed (**Fig. 1**). The choice of this method was motivated by its ability to provide an even distribution of points across the study area, eliminating spatial bias when observation points cluster around specific positions (Hijmans et al., 2005). The result was 523 points, evenly distributed over the basin. Each point's geographic coordinates were recorded, and a monthly rainfall time series spanning 44 years (1981–2024) was used. This technique is also helpful for error analysis in spatial contexts and is central to this study. An analysis of accuracy disparities, including from edge vs. center of the watershed (or lowlands vs. mountains) can be made under equal spreading of points.

2.5. Development and Configuration of Machine Learning and Deep Learning Models

The development of an annual rainfall prediction model in this study involves two main approaches, namely Machine Learning (ML) and Deep Learning (DL). The selection of these two groups of methods is based on the consideration that ML has advantages in handling medium-sized datasets with relatively limited feature space, while DL is designed to extract complex patterns from extensive, temporal, or spatial data (Goodfellow et al., 2016; Zhang et al., 2022). The hyperparameterization of the model is presented in **Table 1**.

2.5.1. Machine Learning Model

Four ML algorithms, such as Random Forest (RF), Extreme Gradient Boosting (XGB), Support Vector Regression (SVR), and Multi-Layer Perceptron (MLP), are applied. RF was selected because it can accommodate non-linear relationships and reduce the overfitting of medium-sized datasets (Breiman, 2001). Some important parameters, such as the number of trees (n_estimators) and the maximum tree depth (max_depth), were optimized using grid search. A gradient boosting method called XGB was selected for its ease of computation and ability to handle noisy data (Chen & Guestrin, 2016).

The learning_rate, max_depth, and number of estimators were fine-tuned for the validation data to minimize the error. SVR was employed for testing the performance of kernel-based approaches, which map data to higher-dimensional spaces to model nonlinear relationships (Smola & Schölkopf, 2004). The radial basis function (RBF) kernel was used, and the parameters C and gamma were determined via parameter search. For the ML group, we utilized MLP as a shallow neural network-based baseline. The MLP architecture had multiple hidden layers, combined with ReLU activation functions and a linear output layer for regression.

2.5.2. Deep Learning Model

Time and space structures in the data are processed over five different DL models: Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), Temporal Convolutional Network (TCN), Convolutional Neural Network (CNN), and Transformer. To make use of LSTM, this model features two sequential LSTM layers with dropout to reduce overfitting. GRU is a simpler yet computationally efficient variant of LSTM (Cho et al., 2014).

The number of hidden units and dropout rate parameters were optimized based on validation performance. TCN was proposed as a non-recurrent alternative to time-series modeling with dilated causal convolutions, facilitating long-term dependency modeling (Bai et al., 2018). A CNN was applied to extract local temporal features from monthly rainfall data, using a 1D convolutional layer, pooling, and dense layers. Transformer is applicable to investigate the capability in fitting temporal dependencies without sequential restriction induced by the self-attention mechanism (Vaswani et al., 2017). The design involved a sequence of encoder layers combining multi-head attention and normalization.

Table 1.

ML and DL Hyperparameters.

Model	Type	Input window	Key hyperparameters / architecture
RF	ML	24 months (24 rainfall lags)	RandomForestRegressor(n_estimators=100, random_state=42); other parameters use scikit-learn defaults (e.g., max_depth=None, min_samples_split=2, min_samples_leaf=1, max_features="sqrt", bootstrap=True).
XGB	ML	24 months	XGBRegressor(n_estimators=100, random_state=42); other parameters follow XGBoost defaults (e.g., typical defaults such as max_depth=6, learning_rate=0.3, subsample=1.0, colsample_bytree=1.0, objective="reg:squarederror" depending on library version).
SVR	ML	24 months	SVR() with scikit-learn defaults: kernel="rbf", C=1.0, epsilon=0.1, gamma="scale", degree=3.
MLP	ML	24 months	MLPRegressor(hidden_layer_sizes=(64, 32), max_iter=500, random_state=42); other parameters follow defaults: activation="relu", solver="adam", learning_rate_init=0.001, alpha=0.0001, etc.
LSTM	DL	24 months (24 × 1)	Input(shape=(24, 1)) → LSTM(64, activation="relu") → Dense(1, activation="linear"); compiled with optimizer="adam", loss="mse"; trained for epochs=60, batch_size=32.
GRU	DL	24 months	Input(shape=(24, 1)) → GRU(64, activation="relu") → Dense(1, activation="relu"); compiled with optimizer="adam", loss="mse"; trained for epochs=60, batch_size=32.
TCN	DL	24 months	Input(shape=(24, 1)) → TCN(nb_filters=32, kernel_size=3, dilations=[1, 2, 4, 8]) → Dense(1, activation="linear"); compiled with optimizer="adam", loss="mse"; trained for epochs=60, batch_size=32.
CNN	DL	24 months	Input(shape=(24, 1)) → Conv1D(64, kernel_size=3, activation="relu", padding="causal") → BatchNormalization() → Conv1D(32, kernel_size=3, activation="relu", padding="causal") → BatchNormalization() → GlobalAveragePooling1D() → Dense(1, activation="linear"); compiled with optimizer="adam", loss="mse"; trained for epochs=60, batch_size=32; predictions clipped to non-negative values during recursive forecasting.
Transformer	DL	24 months	Input(shape=(24, 1)) → transformer_encoder(head_size=8, num_heads=2, ff_dim=32, dropout=0.1) → GlobalAveragePooling1D() → Dropout(0.1) → Dense(1, activation="linear"); compiled with optimizer="adam", loss="mse"; trained for epochs=60, batch_size=32.

2.5.3. Implementation

All models were developed using Python with scikit-learn, TensorFlow, pandas, numpy, and several other libraries. Model training and prediction were performed on Google Colaboratory Pro using high-RAM cloud devices to speed up computation (https://tinyurl.com/y23kfsyh). Each model generates monthly rainfall predictions at 98 sample points, which are then aggregated into annual totals for performance evaluation and spatio-temporal analysis of prediction errors.

2.6. Model Training, Validation, and Evaluation Scheme

The model's training and validation methods were developed specifically to ensure that the predicted findings were statistically sound and robust to annual climate variability in the Bengawan Solo River Basin. The dataset was split into two time subsets: the training period (1981–2019) and the validation period (2020–2024). This separation was done chronologically (time-based split) to avoid data leakage, as future data would otherwise be used for model training. This is frequently used to make time series predictions, as seasonal patterns and climate trends may evolve (Hyndman & Athanasopoulos, 2018). For training purposes, all models were trained on 523 normalized points from monthly rainfall data.

We configured the data in a temporal order (24 months) for deep learning-based models that could capture long-term dependencies. The performance of our model is evaluated with four overall criteria: Mean Absolute Error (MAE) (Equation 1), Root Mean Squared Error (RMSE) (Equation 2), Mean Absolute Percentage Error (MAPE) (Equation 3), and coefficient of determination (R^2) (Equation 4). Such composite metrics are chosen to provide a complete picture of the model: an absolute accuracy rate, the reliability of these measures, and the proportionality of all errors. In general, these metrics measure the difference between the predicted value ($\hat{y}_i$) and the observed value (y_i).

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad \text{.....(1)}$$

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

$$\mathrm{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \quad \text{.....(3)}$$

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

The evaluation results were interpreted by applying the average metric values across all sample points to identify the model with the best overall performance. The model with the best performance during the validation period was selected and used to produce annual rainfall projections for 2025–2030.

2.7. Creation of Spatial Maps of Rainfall Projections with IDW

The final stage of this research methodology is to translate rainfall forecasting results for the year into spatial representations, such as maps. This spatial visualization is critical for understanding the distribution of rainfall across geographical intervals, which is not available from prediction value

tables alone. To accomplish this, the Inverse Distance Weighting (IDW) interpolation technique was selected, as it is simple, flexible for handling irregular data, and works well for environmental variables with pronounced spatial autocorrelation, such as rainfall (Lu & Wong, 2008; Phoophathong et al., 2025).

The interpolation process was carried out separately for each projection year (2025–2030) and for each prediction model. For each projection year, annual rainfall at each sample point was first obtained from the monthly predictions. Aggregation is performed by summing the monthly rainfall predictions (in millimeters) at each sample point for the period from January to December of each year, specifically for the projection years 2025 to 2030. The IDW is implemented in ArcGIS 10.8, using the geographic coordinates of 523 sample points and the corresponding yearly prediction values as inputs. To address this, we use the WGS 84 projection system; it helps preserve compatibility with CHIRPS data while preventing distortion due to distance. The output grid resolution was set to 0.01° (~1 km) to achieve sufficient spatial detail without overloading the computation. The IDW map outputs were then spatially analyzed to identify areas with significant rainfall increase and decrease trends during the projection period.

3. RESULTS

3.1. Model Performance Comparison

A set of nine annual rainfall prediction models was evaluated for their performance for the 2020–2024 validation period based on four main metrics, namely Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE) and coefficient of determination (R^2) (**Fig. 3**). These results indicate that XGBoost (XGB) performed the best at all metrics. The MAE and RMSE of 59 mm and 80 smm, respectively, indicate low absolute and quadratic errors, and the model has an R^2 of 0.73, indicating excellent ability to explain the variability in annual rainfall.

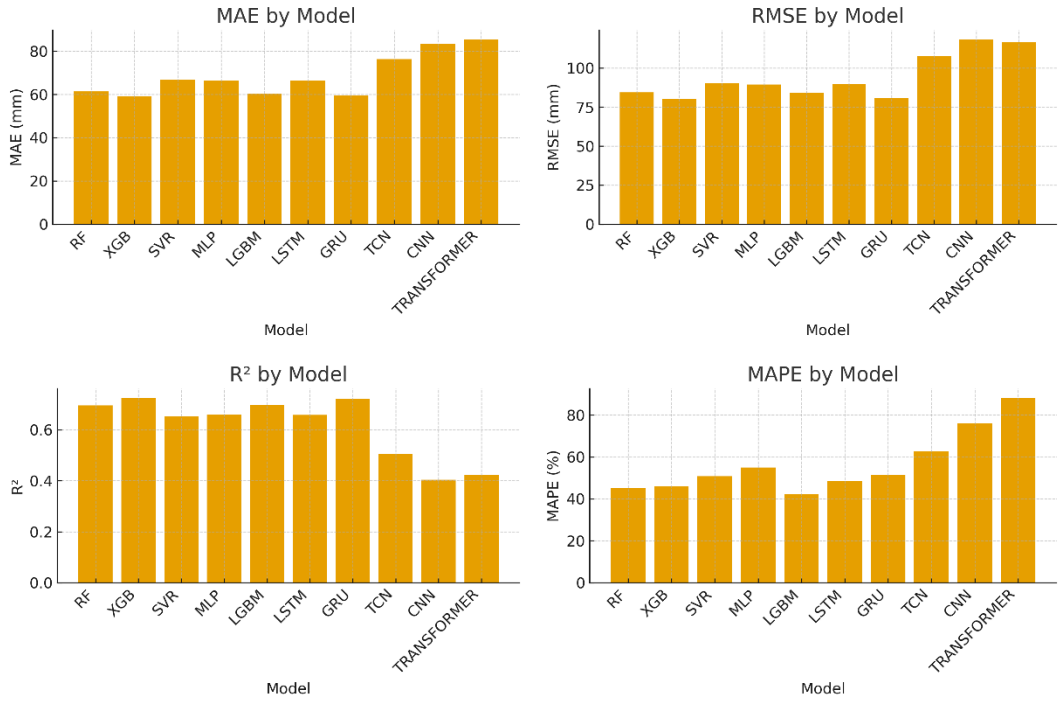

Fig. 3. Measurement error results for all models.

Of the two models, Extreme Gradient Boosting (XGB) ranked 2nd, with an MAE of 55.140 mm and an R^2 of 0.767. XGB handled data non-linearity well but performed slightly worse than RF. XGB is sensitive to the learning rate and the number of trees, which is likely to cause instability in predictions for long-term datasets where year-to-year data exhibits high variation. We have seen that the Multi-Layer Perceptron (MLP), with an MAE of 46.762 mm and an R^2 of 0.817, performed relatively well, although lower than RF. The lower strength of MLPs compared to ensemble-based models like RF is attributed to the poor ability of feedforward ANNs to capture long-term temporal patterns without a direct memory mechanism (Wu et al., 2010). Although performance was surprising, with sequence-based models (e.g., LSTM, GRU, TCN, CNN, and Transformer) achieving very high MAE (>88 mm, up to 202 mm for LSTM, TCN, CNN, and Transformer), some architectures also yielded negative R^2 values.

Such excessive overfitting is likely an effect of the time series being too short, even though it spans 44 years, for detecting stable annual climate patterns in complex DL architectures with a large number of parameters (Lim & Zohren, 2021). This condition aligns with the results of Laptev et al. (2017), who note that DL forecasting models for meteorological time series require hundreds of thousands of observations or hundreds of years of simulation data to generalize consistently. These findings support the idea that ensemble-based machine learning models, such as RF, remain more stable in tropical environments with limited data. These results serve not only as a technical framework for hydrology researchers and practitioners but also contribute to the literature by showing the practical limitations of DL in low-data areas and by pointing to opportunities for future study to develop lighter DL architectures and more realistic regularization.

3.2. Spatial Error Analysis

Spatial error analysis was performed to understand the distribution of prediction errors at 523 sample points in the Bengawan Solo watershed during the 2020–2024 validation period (**Fig. 4–7**).

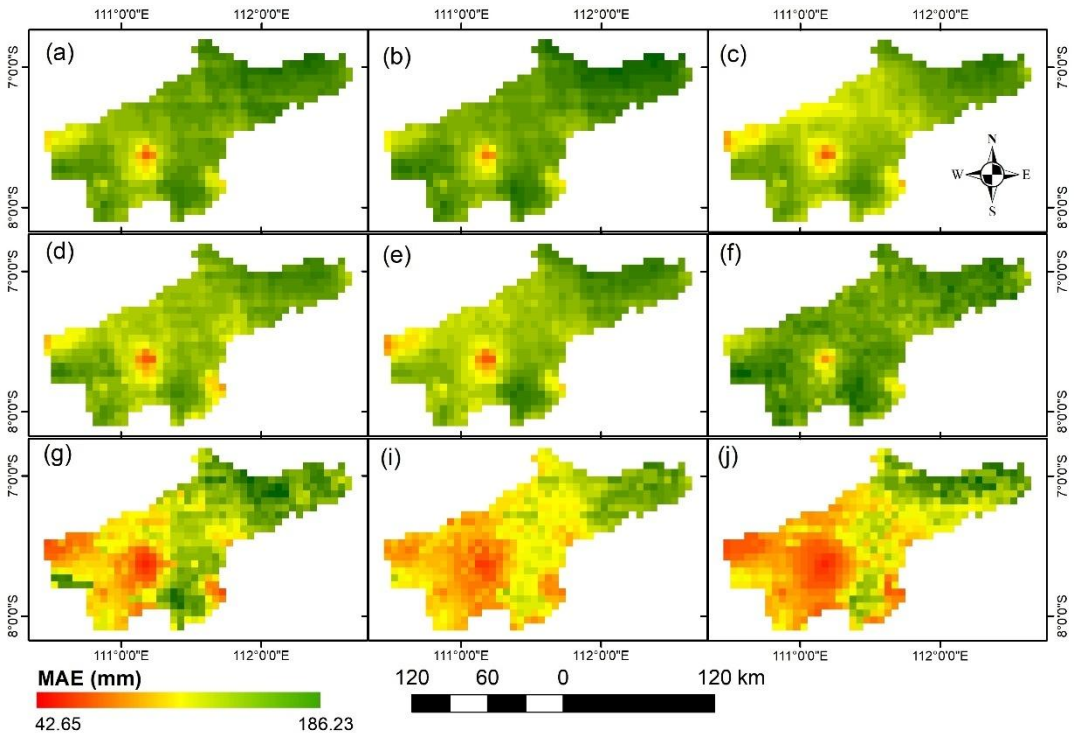

Fig. 4. Spatial MAE Distribution (mm). (a) RF, (b) XGB, (c) SVR, (d) MLP, I LSTM, (f) GRU, (g) TCN, (h) CNN, (i) TRANSFORMER.

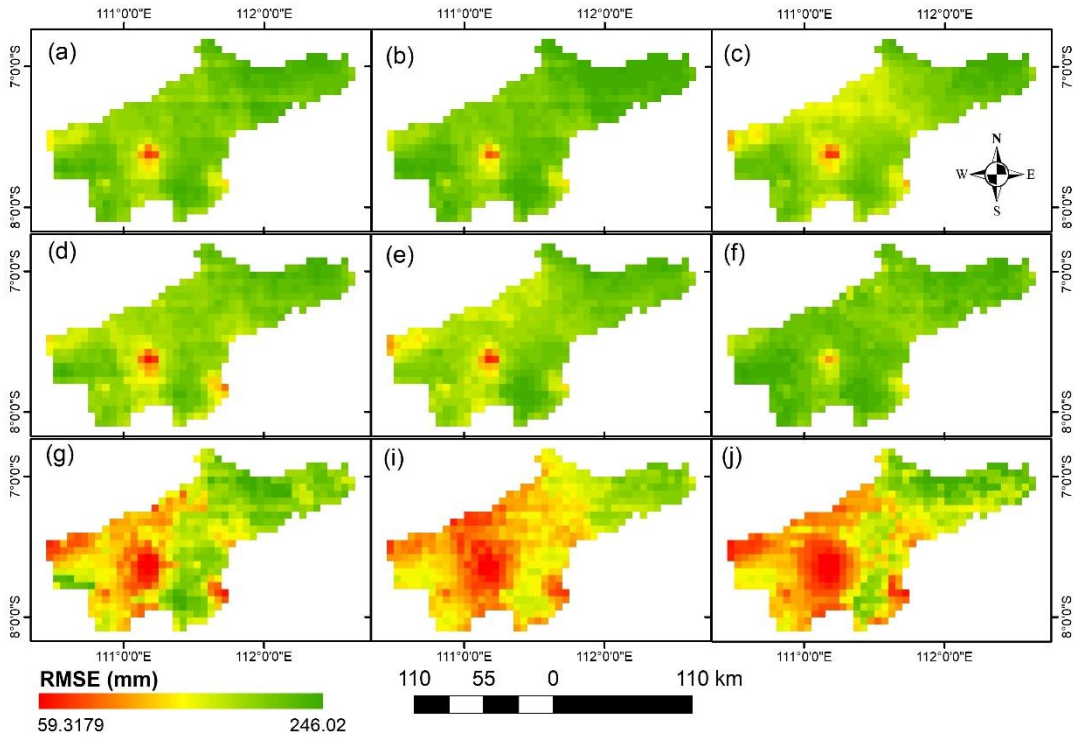

Fig. 5. Spatial RMSE Distribution (mm). (a) RF, (b) XGB, (c) SVR, (d) MLP, I LSTM, (f) GRU, (g) TCN, (h) CNN, (i) TRANSFORMER.

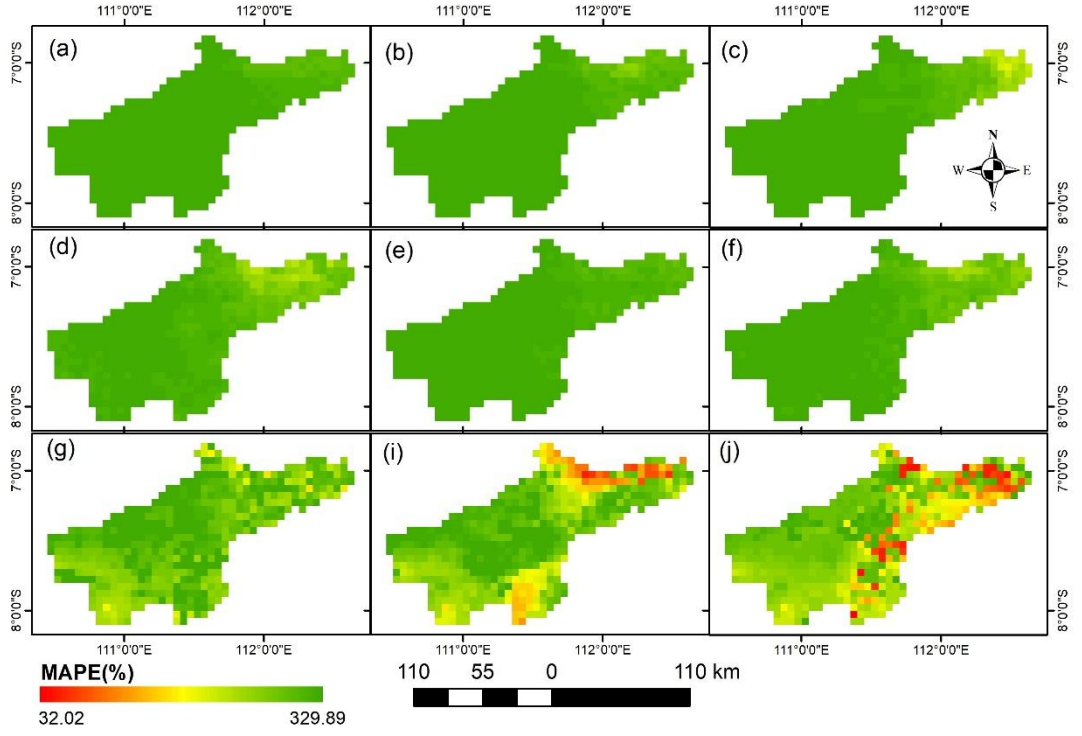

Fig. 6. Spatial MAPE Distribution (%). (a) RF, (b) XGB, (c) SVR, (d) MLP, I LSTM, (f) GRU, (g) TCN, (h) CNN, (i) TRANSFORMER.

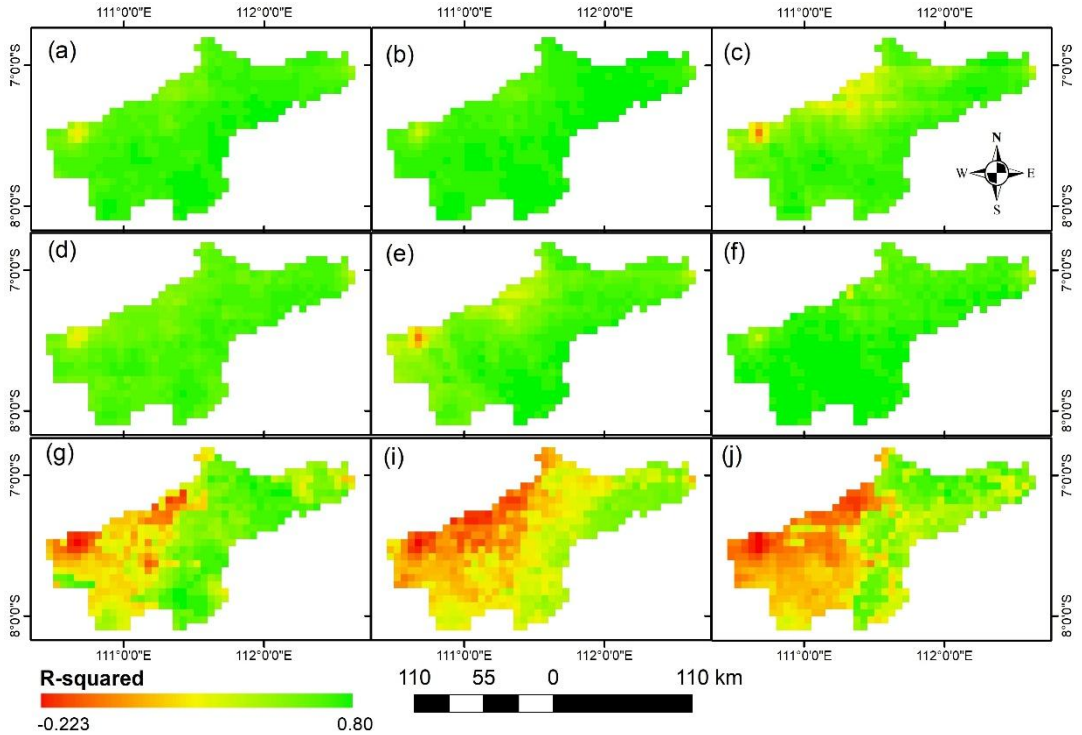

Fig. 7. Spatial R2 Distribution. (a) RF, (b) XGB, (c) SVR, (d) MLP, I LSTM, (f) GRU, (g) TCN, (h) CNN, (i) TRANSFORMERs.

This is critical, given that although aggregated metrics (e.g., MAE, RMSE) provide a baseline picture of model performance, the distribution of errors at spatial scales often reveals underperformance not evident at the aggregate level (Cioffi et al., 2023).

Values of the MAE for points indicate that the error distribution is not homogeneous. Despite the fact that the Random Forest (RF) model generally achieves the best results, it still exhibits localized spatial error distribution, especially at key points in the watershed in the east and west. The points are usually in the parts of the area that are more diverse in topography, e.g., in the foothills of Mount Lawu to the east and hills in the northwest. These topographic differences probably impact the precision of CHIRPS satellite rainfall predictions and the precision of model predictions, respectively. On the other hand, in the central plains of the watershed, where the topography is similar and near the main Bengawan Solo River, there are lower prediction errors.

This suggests that the model is more adept at learning the relationship between the historical rainfall data and prediction patterns in regions free from extreme topographic gradients. Other, e.g., those models used, such as XGB and MLP, exhibit error patterns as in the RF, but they generate more error intensity in almost all locations. With the SVR model, the error distribution is scattered and not so well matched to the topography pattern, which reflects its sensitivity to the high variation of training data. Moreover, deep learning-based models, namely LSTM, GRU, TCN, CNN, and Transformer, have large average errors and spatial instability. There exist very large errors in points in mountain regions and in transition zones between mountains and plains that exceed the watershed mean MAE by more than 3 times. This confirms the overfitting assumption in DL, which occurs when the model cannot learn spatial variability effectively.

The data also supports this observation with an increased error gradient on the MAE distribution map for IDW interpolation results towards east and west of the watershed boundaries. This fact appears to be in line with earlier works (Funk et al., 2015; Gu et al., 2020), which found that satellite-based rainfall prediction is more prone to vary when the orographic environment and the number of rainfall stations in the location are not very high.

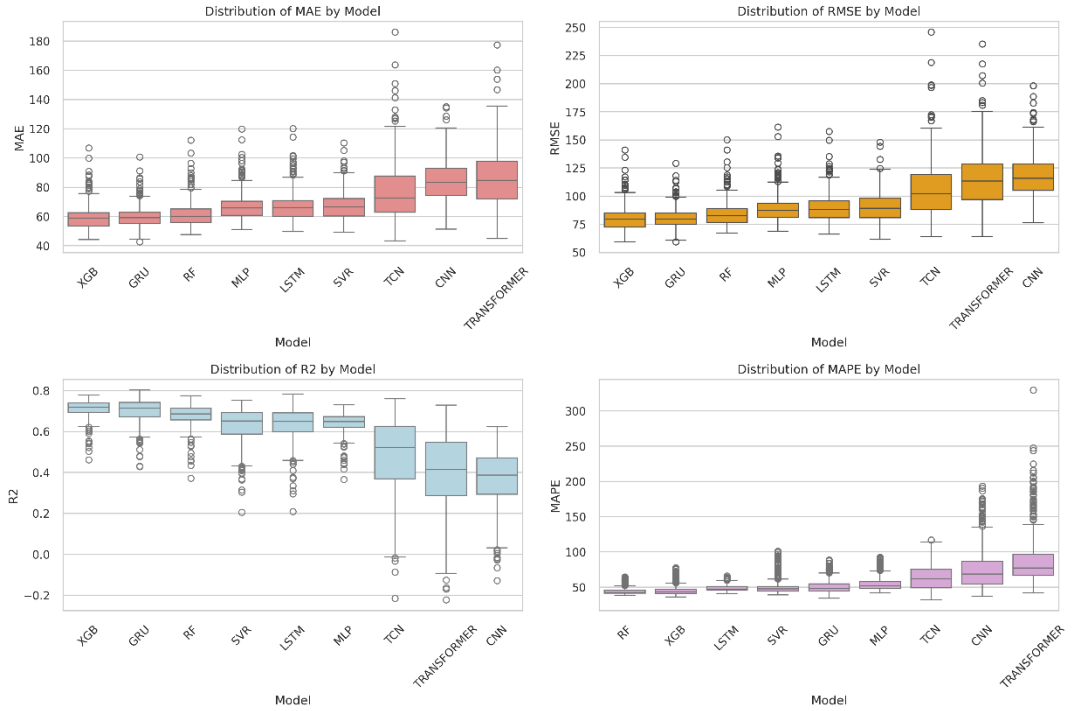

Fig. 8. Boxplot Comparison of Error Metrics.

This spatial study has a practical dimension, as it implies that, even though RF is technically well-known for its worldwide performance, it could still be improved by integrating CHIRPS data with higher-resolution data or by pre-conditioning the model to mitigate topographical bias. In addition, model evaluation can always be conducted to consider the spatial, not overall, aspect, in order to detect and target prediction failures in some regions for future study.

The spatial error comparison boxplot (**Fig. 8**) reveals the marked disparity between models in accounting for the variability in precipitation at 523 grid points in the Bengawan Solo watershed. On average, XGB and GRU have the lowest median MAE and RMSE, and their interquartile ranges (IQRs) are rather narrow, which implies good and constant (or comparable) performance in most locations. RF marginally is behind (median MAE slightly higher; upper tail slightly longer), but it also outperforms SVR and MLP by a long margin of difference because both have larger errors and wider distributions. On the other hand, convolutional and attention-based deep learning models (TCN, CNN, Transformer) seem obviously less robust: median MAE and RMSE are high, interquartile ranges are wide, and the large number of extreme outliers suggests spatial instability and no ability to follow rainfall trends for some grids. The boxplot pattern only confirms that the decision tree ensembles and GRUs are spatially superior to the more complex deep learning architectures used with the dataset.

3.3. Temporal Error Analysis

Temporal error analysis was performed to evaluate variation in model performance over time during the 2020–2024 validation period (**Fig. 9-12**). This approach aims to identify specific periods during which the model degrades in accuracy, thereby providing deeper insights into seasonal factors and extreme climate events that affect predictions (Zhou et al., 2022).

Results from temporal evaluation show that the XGBoost (XGB) model, which has good overall performance, tends to have more prediction errors during transitional months (March–April and October–November, for example).

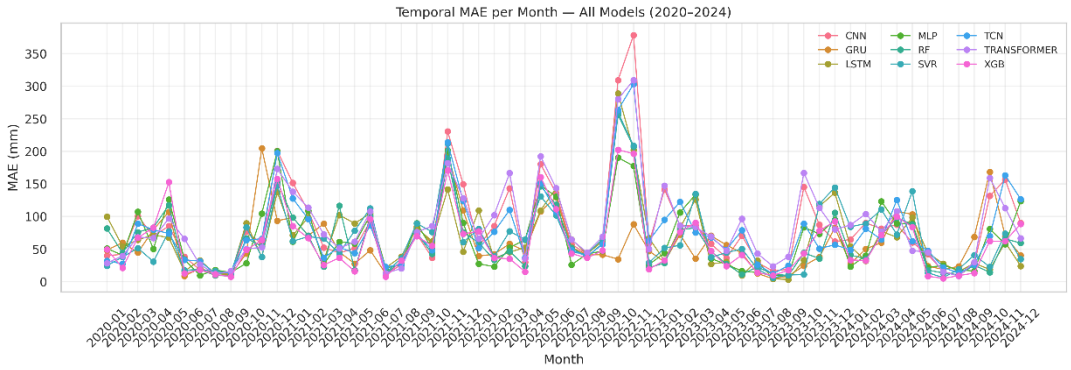

Fig. 9. Temporal Metric of MAE.

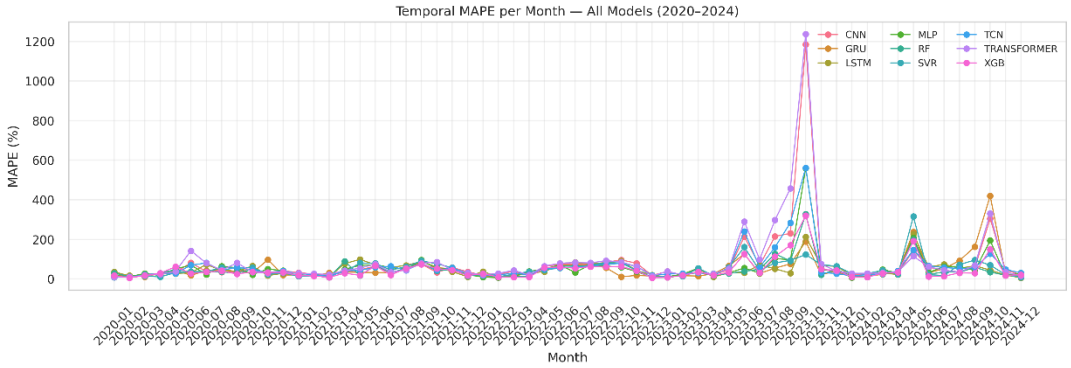

Fig. 10. Temporal Metric of MAPE.

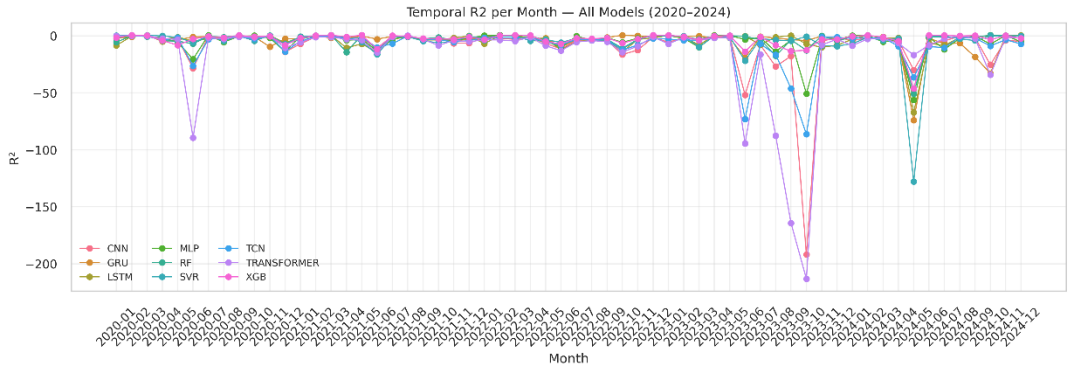

Fig. 11. Temporal Metric of R^2.

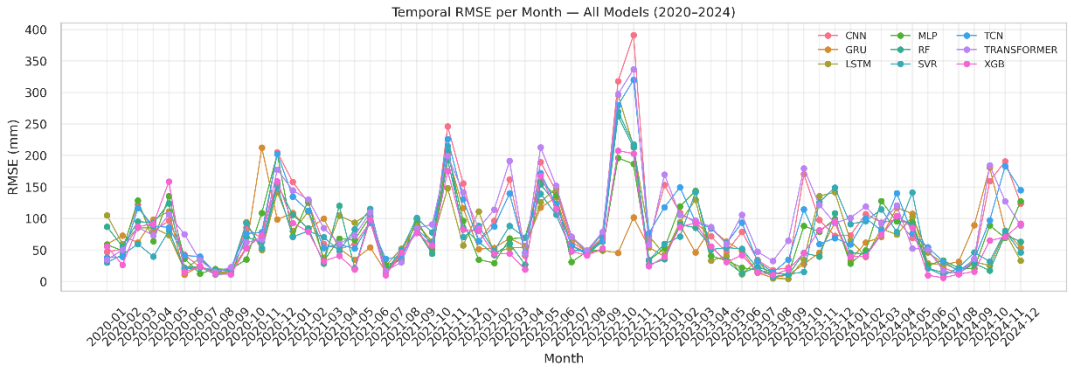

Fig. 12. Temporal Metric of RMSE.

At these junctures, monsoon variations in predominant wind directions and erratic weather patterns also occur in areas of tropical Indonesia such as the Bengawan Solo watershed. Because these patterns are not completely periodic, the abrupt changes in rainfall intensity and location can be hard to capture in models based on historical data due to the fact that these patterns are highly influenced by global climate variability such as El Niño–Southern Oscillation (ENSO) and Indian Ocean Dipole (IOD) (McBride et al., 2003; Lestari et al., 2021).

Major mistakes are also found in years characterized by extreme climate anomalies. That is, for example, from October to December 2020 (strong La Niña), rainfall was well above normal, leading to large discrepancies between predictions and observed rainfall. On the opposite hand, in 2023 (a drier year as we have seen in the rain shadow of El Niño), there were errors early in the rainy season with the model predicting heavy rainfall, yet the rainfall did not match the average. RF and GRU models have temporal trends similar to that of XGB, except they oscillate with error a little further. In contrast, convolutional and attention-based deep learning models (TCN, CNN, Transformer) show greater, less stable errors over time, because some models have very high sensitivity to monthly outliers, and one or two "bad" months can significantly increase the overall error.

Overall, this pattern confirms that all models really struggle with rainfall prediction during transition seasons or extreme climate processes. In terms of temporal and operational challenges, these relate primarily to (1) the inability to better capture long-term climate variability, despite employing a 44-year time series, and (2) monthly temporal resolution, which fails to fully reflect the intra-monthly dynamics affecting rainfall accumulation.

3.3. Prediction Results

Spatial predictions of annual rainfall with different models (RF, XGB, SVR, MLP, LSTM, GRU, CNN, TCN, Transformer) exhibit a fairly consistent distribution pattern in the upper Bengawan Solo region, but with varying intensities between models (**Fig. 13**). Overall, the spatial dynamics of traditional ML models (RF in particular) display relatively smoother and more stable spatial patterns, whereas DL models like LSTM, GRU, and CNN exhibit sharper spatial fluctuations, especially in mountainous areas. The spatial distribution indicates that the highest annual rainfall is anticipated to occur in areas adjacent to the Merapi, Merbabu, and Lawu mountains. It mirrors the orographic conditions in the area, in which moist air rising produces higher orographic rainfall. In contrast, the central part of the watershed tends to have lower and more homogeneous rainfall. These are consistent with the RF, XGB, and MLP models, and in fact, DL models like CNN or Transformer often result in extreme predictions at certain points, which can be attributed to overfitting issues.

The differences between the two models are apparent in the sharpness of the spatial gradients. DL-based models typically can be "sensitive" to local variations, so that predictions are usually quite heterogeneous. This might prove their capacity for finding local anomalies, but it also poses possibilities of instability in predicting. When compared to this, RF models give smoother distributions, which can be practically put to use (like water planning, flood reduction, etc). This phenomenon is also in line with the results of Chen et al. (2022) and Zandi et al. (2022): model integration must balance the trade-off between generalization and spatial sensitivity.

In sum, the spatial results of models show that traditional machine learning models such as Random Forest (RF), Extreme Gradient Boosting (XGB), and Multi-Layer Perceptron (MLP) are effective in keeping spatial rainfall predictions stable and provide consistent and relatively homogeneous distributions in nearly every region of the Bengawan Solo watershed.

On the other hand, in the case of deep learning models, Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), Convolutional Neural Network (CNN), and Transformer, the prediction variability is higher and is likely to give large values at multiple places, particularly where orographic complexity is high. Another interesting discovery was that the largest spatial variations are found in mountainous zones. This emphasizes the necessary relevance of topographical parameters and/or orographic processes for the calibration process and the construction of spatiotemporal rainfall prediction models in tropical regions.

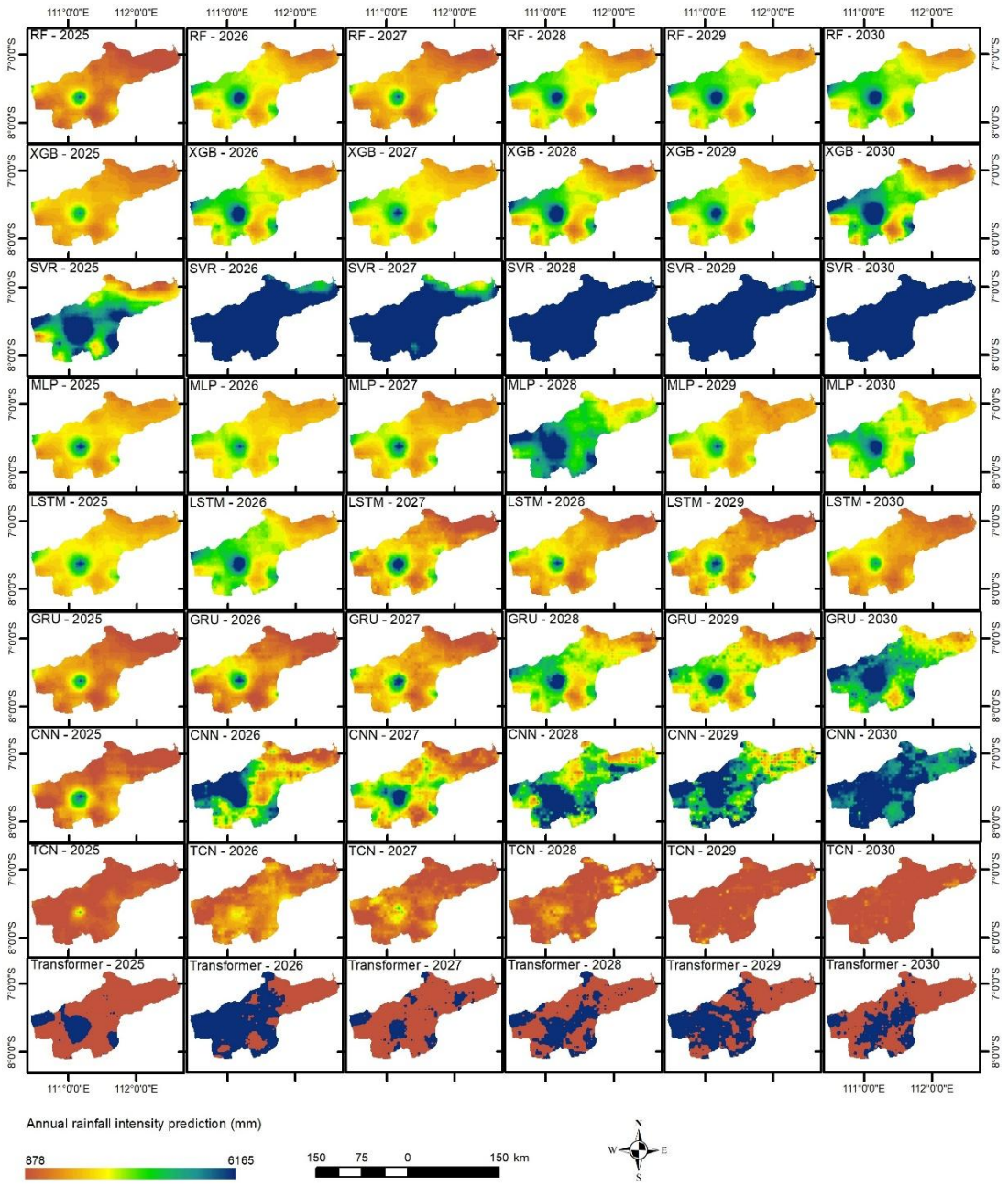

Fig. 13. Spatial Prediction Results for Annual Rainfall 2025–2030 for all models.

4. DISCUSSIONS

This work demonstrates that the decision tree-based gradient boosting models, with XGBoost (XGB), are most trustworthy for predicting spatiotemporal rainfall for the Bengawan Solo River Basin. XGB is the most consistent in providing overall accuracy (MAE ≈ 59 mm; RMSE ≈ 80 mm; $R^2 \approx 0.73$), followed by RF and GRU for both temporally (month-to-month) and spatially (location-to-location) data. The supremacy of the XGB-RF relationship agrees with that of CHIRPS satellite

rainfall data in tropical regions, where rainfall strongly, but not linearly, correlates with its controlling features and is subject to noise and observational bias. In this light, ensemble decision trees are very powerful since they can capture nonlinear interaction in a complicated way without strict distributional assumptions, and they provide less sensitivity to hyperparameter selection than large-capacity deep learning models (Breiman, 2001; Hong et al., 2021; Sachindra et al., 2018). Bagging and boosting mechanisms help the model to smooth out local uncertainties between trees, leading to stable predictions over 523 grids with a vast variation in topography and microclimates.

Comparison with the literature shows that these findings are consistent with previous studies reporting the superiority of ensemble decision tree models for highly variable hydrometeorological data. Hong et al. (2021) and Sachindra et al. (2018) demonstrated that RF and gradient boosting models tend to outperform other nonlinear methods when data are limited, spatially heterogeneous, and contain difficult-to-model noise. In this study, despite the long temporal sample size (44 years), the spatial sample size was only 523 grid points within a single watershed, so the sample-to-parameter ratio was much more favorable for decision tree models than for very deep deep learning architectures. These results reinforce that, for data configurations like this, a "moderate but robust capacity" strategy is more effective than "very large capacity but highly dependent on data volume and diversity."

On the other hand, deep learning models exhibit more diverse behavior. GRU emerged as the most competitive deep learning model, with performance close to XGB ($R^2 \approx 0.72$) and fairly good temporal stability. This indicates that lighter recurrent architectures with fewer parameters, such as GRUs, are more suitable for moderate-length monthly time series because they require fewer parameters to learn and are easier to calibrate. Conversely, LSTM showed performance not far from MLP and SVR. At the same time, convolutional and attention-based architectures (TCN, CNN, Transformer) tended to have higher MAE and RMSE and lower R^2, with significant variability in errors across months and locations. This pattern aligns with findings by Chattopadhyay et al. (2020), Aswin & Geetha (2021), and reviews by Lim & Zohren (2021), which state that advanced deep learning, especially CNN and Transformer, only show clear advantages when: (i) the temporal resolution is very high (daily or sub-daily), (ii) the sample size is very large, and (iii) additional rich spatial/climate information is available. For monthly resolution and limited domains such as a single watershed, this large capacity can easily lead to overfitting and instability.

Spatial analysis of error distribution confirms that XGB and GRU not only outperform on average but are also more robust across most of the watershed area. Both models maintain an average MAE of 59–60 mm and an R^2 of 0.71–0.72 across most grids, while RF lags slightly behind with marginally worse MAE and R^2. Conversely, TCN, CNN, and Transformer show much higher spatial MAE (≈76–85 mm) and lower R^2 (≈0.38–0.50), with many grids performing near or even below the climatological baseline. The spatial error patterns also reveal that areas with steep topographic gradients, such as the slopes of Merbabu, Merapi, and Lawu, as well as the upper reaches of the watershed, tend to have higher errors. This is consistent with studies by Gu et al. (2020) and Funk et al., which indicate that satellite-based rainfall estimates in tropical mountainous regions are often biased due to sensor limitations in capturing small-scale convective clouds and complex orographic effects. Therefore, some of the errors observed across all models, including XGB and GRU, also reflect the limitations of the "ground truth" rainfall representation itself.

From a temporal perspective, the analysis of monthly MAE and RMSE shows that all models exhibit higher errors during the monsoon transition months (March–April and October–November) and in years with strong climate anomalies (e.g., La Niña 2020–2022 and El Niño 2023). In these conditions, rainfall intensity and distribution change rapidly and often do not follow periodic patterns that can be captured by models based on historical time series. Nevertheless, XGB and GRU are relatively consistent in maintaining the highest performance rankings across most months, while TCN, CNN, and Transformer exhibit sharp, unstable error spikes. This indicates that the capacity to model short-term non-linearity alone is not sufficient; models must also be able to generalize to rare but impactful climate regime changes. These findings align with McBride et al. (2003) and Lestari et al. (2021), which emphasize that rainfall variability in Indonesia is heavily influenced by ENSO and

IOD, making purely time-series-based local models potentially struggle when faced with rare extreme events.

Compared to other regions, model performance patterns in the Bengawan Solo River Basin show a strong correlation with the hydrometeorological context and data structure. In the North-Western Himalayas, dominated by steep mountainous topography with sharp elevation gradients, multivariable daily station data, and a time span of 1980–2021, DL models (specifically Bi-LSTM and LSTM) provide the lowest RMSE/MAE and outperform ML, while among ML models, ANN and KNN outperform RF and SVR, and accuracy is highly sensitive to elevation (Wani et al., 2024). In five large UK cities with a temperate maritime climate and hourly rainfall data, Stacked-LSTM and Bidirectional-LSTM architectures also reportedly outperform XGBoost and ML ensembles, although they still struggle to capture very abrupt rainfall surges (Barrera-Animas et al., 2022).

Conversely, a study in the Aligarh District of India showed that for daily and monthly rainfall with standard meteorological predictors, CatBoost and RF provide strong correlation and outperform SVR, confirming the reliability of tree ensembles for more aggregated time scales (Abdullah & Said, 2025). In African cities with humid tropical to dry Mediterranean climates, DL (especially single RNN) performs best for highly non-linear daily rainfall, with relative humidity and antecedent rainfall as key predictors (Samson & Aweda, 2025). In this context, the dominance of XGB, RF, and GRU in monthly CHIRPS-based Bengawan Solo annual rainfall prediction can be understood as a consequence of a combination of: a tropical monsoon climate with a strong monsoon but averaged daily pattern, watershed-scale orographic heterogeneity, and limitations of atmospheric predictors, thus significantly reducing the advantages of very deep DL architectures compared to other regions and time scales.

In terms of data validity, CHIRPS is supported by a broad body of validation studies across very different hydroclimatic settings. At the global scale, a recent synthesis shows that CHIRPS generally attains correlations above 0.7 at monthly and seasonal scales and is widely judged suitable for hydroclimatic analysis, with reduced skill mainly in very arid and high-mountain regions (Du et al., 2023).

Regionally, CHIRPS has been shown to outperform or rival reanalysis and other satellite products in complex tropical and subtropical terrains: for example across Ethiopia, where it consistently outperforms ERA5 and reproduces the seasonal cycle and interannual variability (Ahmed et al., 2024; Geleta & Deressa, 2020), in the tropical Andes and Antioquia where it captures orographic gradients and ENSO-related variability (López-Bermeo et al., 2022; Rivera et al., 2018), in the Amazon basin where annual totals are reproduced with R^2 ~0.98 (da Motta Paca et al., 2020), over the Qinghai–Tibet Plateau where CHIRPS compares favorably with MSWEP at monthly scales (Liu et al., 2019), and in semi-arid and drought-prone regions such as Northeast Brazil and eastern Africa (Paredes-Trejo et al., 2017; Dinku et al., 2018). These studies consistently report that CHIRPS performs best for monthly–annual aggregates and large-basin applications, while underestimating some local extremes, a trade-off that aligns with our use of 0.05° monthly CHIRPS data aggregated to annual rainfall over a large tropical watershed.

Several limitations of this study should be noted to contextualize the results. First, the temporal resolution of the input data is limited to monthly aggregation, so intra-monthly dynamics, such as daily extreme rainfall intensity or consecutive rainfall events, are not explicitly represented, even though these phenomena significantly contribute to monthly accumulation. Second, although 523 grid points provide much better spatial coverage than 98 points, the representation of microclimates in mountainous areas and narrow valleys remains far from perfect. Third, the model in this study has not yet integrated global and regional climate predictors, such as the Niño 3.4 index, the Dipole Mode Index (DMI), or the MJO, which are important for explaining seasonal rainfall variability in Indonesia. Fourth, the approach used is entirely based on statistical data without explicit linkage to physical processes within the hydrological system, so the mechanistic interpretation of some error patterns remains limited.

The practical implications and directions for further research that emerge from these findings are quite clear. For operational applications such as early flood warning and water resource planning in the Bengawan Solo watershed, ensemble decision tree models (XGB, RF) and relatively lightweight recurrent networks (GRU) are currently the most realistic choices, offering the best trade-off between accuracy, spatio-temporal stability, and computational complexity. On the other hand, the potential of deep learning (DL) is not fully exhausted; improvements can be achieved by (i) enriching features with ENSO, IOD, and MJO indices, (ii) increasing the temporal resolution of training to daily or decadal scales, (iii) applying more aggressive regularization and early stopping, and (iv) exploring transfer learning from larger climate domains.

Additionally, developing hybrid approaches that combine statistical/machine learning models with physical hydrological models can help reduce structural ambiguities and enhance generalization capabilities, especially under climate change scenarios. Thus, this study not only demonstrates that machine learning, particularly XGB and RF, still outperforms most advanced deep learning architectures in limited-data configurations but also provides a concrete roadmap for improving spatio-temporal rainfall prediction models in tropical regions with complex topography. Furthermore, a useful future step would be to integrate a self-imputation and self-quality assessment framework to stabilize the lunar series before training and applying various types of optimizers (Haidu et al., 2025; Sriwahyuni et al., 2025).

5. CONCLUSION

This study compares nine ML and DL algorithms for predicting rainfall in the Bengawan Solo River Basin using CHIRPS data from 1981 to 2024. Results show decision tree-based gradient boosting models, especially XGBoost (XGB), perform best during 2020–2024 (MAE ≈ 59 mm; RMSE ≈ 80 mm; $R^2 \approx 0.73$), followed by Random Forest (RF) and Gated Recurrent Unit (GRU), with minor differences. GRU was the top DL model ($R^2 \approx 0.72$), while CNN, TCN, and Transformer architectures had larger errors and variability. Findings indicate ensemble decision trees and lightweight recurrent neural networks are more reliable than complex DL models in large tropical river basins with spatial heterogeneity and moderate time series length.

Spatially, accuracy is higher in the basin's center and lower in peripheral or mountainous areas, affected by microclimate, orographic effects, and satellite limitations. Temporally, errors peak during monsoon transitions (March–April, October–November) and during climate anomalies, challenging models during ENSO and IOD events. Practically, XGB, RF, and GRU are suitable for operational rainfall forecasts, aiding water management and flood mitigation in tropical regions. Limitations include monthly data resolution, satellite biases, and a lack of external climate predictors like ENSO and IOD indices.

Future research should integrate high-resolution data, include climate indicators, use regularization and transfer learning to enhance DL, and explore hybrid models. Overall, this study highlights trade-offs and opportunities in ML and DL for tropical rainfall prediction, serving as a guide for developing accurate, adaptive hydrometeorological systems in Indonesia and similar areas.

ACKNOWLEDGEMENT

This study is funded by the HIT Program of 2025 of Universitas Muhammadiyah Surakarta.

REFERENCES

Abdullah, M., & Said, S. (2025). Performance evaluation of machine learning regression models for rainfall prediction. *Iranian Journal of Science and Technology, Transactions of Civil Engineering*, *49*(4), 4231-4250.

Ahmed, J. S., Buizza, R., Dell'Acqua, M., Demissie, T., & Pè, M. E. (2024). Evaluation of ERA5 and CHIRPS rainfall estimates against observations across Ethiopia. *Meteorology and Atmospheric Physics*, *136*(3), 17.

Alikhanov, B., Pulatov, B., Samiev, L., (2024). Impact of Climate Change on the Cryosphere of the Ugam Chatkal National Park, Bostonliq District, Uzbekistan, During the Post-Soviet Period, Based on Remote Sensing and Statistical Analysis. *Forum Geografi*, 38, 302–316. https://doi.org/10.23917/forgeo.v38i3.4405

Al-Samrraie, L.A., Abdalla, A.M., Alrawashdeh, K.A.-B., Bsoul, A.A., Awad, M.A., Alzboon, K., Al-Taani, A.A., (2025). Deep Learning Models Based on CNN, RNN, and LSTM for Rainfall Forecasting: Jordan as a Case Study. *Mathematical Modeling of Engineering Problems,* 12.

Amini, E., Zolfaghari, A., Kaboli, H., Rahimi, M., (2022). Estimation of rainfall erosivity map in areas with limited number of rainfall stations (case study: Semnan Province). *Iranian Journal of Soil and Water Research,* 53, 2027–2044.

Avand, M., Moradi, H.R., Ramazanzadeh Lasboyee, M., (2021). Spatial prediction of future flood risk: an approach to the effects of climate change. *Geosciences*, 11, 25.

Baig, F., Ali, L., Faiz, M.A., Chen, H., Sherif, M., (2024). How accurate are the machine learning models in improving monthly rainfall prediction in hyper arid environment? *Journal of Hydrology,* 633, 131040.

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. *Machine Learning with Applications*, *7*, 100204.

Chen, C., Zhang, Q., Kashani, M.H., Jun, C., Bateni, S.M., Band, S.S., Dash, S.S., Chau, K.-W., (2022). Forecast of rainfall distribution based on fixed sliding window long short-term memory. Engineering *Applications of Computational Fluid Mechanics* 16, 248–261. https://doi.org/10.1080/19942060.2021.2009374

CHIRPS: Rainfall Estimates from Rain Gauge and Satellite Observations | Climate Hazards Center - UC Santa Barbara [WWW Document], n.d. URL https://www.chc.ucsb.edu/data/chirps (accessed 8.30.2025).

da Motta Paca, V. H., Espinoza-Davalos, G. E., Moreira, D. M., & Comair, G. (2020). Variability of trends in precipitation across the Amazon River basin determined from the CHIRPS precipitation product and from station records. *Water*, *12*(5), 1244.

Das, P., Posch, A., Barber, N., Hicks, M., Duffy, K., Vandal, T., Singh, D., Werkhoven, K. van, Ganguly, A.R., (2024). Hybrid physics-AI outperforms numerical weather prediction for extreme precipitation nowcasting. *Climate and Atmospheric Science,* 7, 282.

Dinku, T., Funk, C., Peterson, P., Maidment, R., Tadesse, T., Gadain, H., & Ceccato, P. (2018). Validation of the CHIRPS satellite rainfall estimates over eastern Africa. *Quarterly Journal of the Royal Meteorological Society*, *144*, 292-312.

Du, H., Tan, M. L., Zhang, F., Chun, K. P., Li, L., & Kabir, M. H. (2024). Evaluating the effectiveness of CHIRPS data for hydroclimatic studies. *Theoretical and Applied Climatology*, *155*(3), 1519-1539.

El Hafyani, M., El Himdi, K., El Adlouni, S.-E., (2024). Improving monthly precipitation prediction accuracy using machine learning models: a multi-view stacking learning technique. *Frontiers in Water,* 6, 1378598.

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

Fang, L., Shao, D., (2022). Application of long short-term memory (LSTM) on the prediction of rainfall-runoff in karst area. *Frontiers in Physics*, 9, 790687.

Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., Husak, G., Rowland, J., Harrison, L., Hoell, A., (2015). The Climate Hazards Infrared Precipitation with Stations—A New Environmental Record for Monitoring Extremes. *Scientific Data*, 2, 1–21.

Geleta, C. D., & Deressa, T. A. (2021). Evaluation of climate hazards group infrared precipitation station (CHIRPS) satellite-based rainfall estimates over Finchaa and Neshe Watersheds, Ethiopia. *Engineering Reports*, *3*(6), e12338.

Gu, J., Liu, S., Zhou, Z., Chalov, S.R., Zhuang, Q., (2022). A stacking ensemble learning model for monthly rainfall prediction in the Taihu Basin, China. *Water,* 14, 492.

Haidu, I., El Orfi, T., Magyari-Sáska, Z., Lebaut, S., & El Gachi, M. (2024). Modeling the Long-Term Variability in the Surfaces of Three Lakes in Morocco with Limited Remote Sensing Image Sources. *Remote Sensing*, *16*(17), 3133. https://doi.org/10.3390/rs16173133

Haidu, I., Magyari-Sáska, Z., & Magyari-Sáska, A. (2025). Spatio-Temporal Gap Filling of Sentinel-2 NDI45 Data Using a Variance-Weighted Kalman Filter and LSTM Ensemble. *Sensors*, *25*(17), 5299. https://doi.org/10.3390/s25175299

Haq, D.Z., Novitasari, D.C.R., Hamid, A., Ulinnuha, N., Farida, Y., Nugraheni, R.D., Nariswari, R., Rohayani, H., Pramulya, R., Widjayanto, A., (2021). Long short-term memory algorithm for rainfall prediction based on El-Nino and IOD data. *Procedia Computer Science,* 179, 829–837.

Hu, C., Wu, Q., Li, H., Jian, S., Li, N., Lou, Z., (2018). Deep learning with a long short-term memory networks approach for rainfall-runoff simulation. *Water,* 10, 1543.

Jumadi, J., Danardono, D., Roziaty, E., Ulinuha, A., Supari, S., Choy, L. K., Sattar, F., Nawaz, M. (2025). AI-Driven Ensemble Learning for Spatio-Temporal Rainfall Prediction in the Bengawan Solo River Watershed, Indonesia. *Sustainability*, 17(20), 9281. https://doi.org/10.3390/su17209281

Jumadi, J., Danardono, D., Priyono, K. D., Roziaty, E., Masruroh, H., Rohman, A., Amin, C., Hadibasyir, H. Z., Fikriyah, V. N., Nawaz, M., Sattar, F., & Lotfata, A. (2024). Utilizing Open Access Spatial Data for Flood Risk Mapping: A Case Study in the Upper Solo Watershed. Geoplanning: *Journal of Geomatics and Planning*, 11(2), 189-204. https://doi.org/10.14710/geoplanning.11.2.189-204

Kasihairani, D., Hidayat, R., Supari, S., (2024). Assessing the Reliability of Predicted Decadal Surface Temperatures in Southeast Asia. *Forum Geografi,* 38, 413–425. https://doi.org/10.23917/forgeo.v38i3.5402

Kim, S., Shin, J.-Y., Heo, J.-H., (2025). Assessment of Future Rainfall Quantile Changes in South Korea Based on a CMIP6 Multi-Model Ensemble. *Water,* 17, 894.

Koya, S.R., Roy, T., (2024). Temporal Fusion Transformers for Streamflow Prediction: Value of Combining Attention with Recurrence. *Journal of Hydrology,* 637, 131301.

Kumar, V., Kedam, N., Kisi, O., Alsulamy, S., Khedher, K.M., Salem, M.A., (2025). A Comparative Study of Machine Learning Models for Daily and Weekly Rainfall Forecasting. *Water Resours Manage,* 39, 271–290. https://doi.org/10.1007/s11269-024-03969-8

Kundu, S., Biswas, S.K., Tripathi, D., Karmakar, R., Majumdar, S., Mandal, S., (2023). A review on rainfall forecasting using ensemble learning techniques. e-Prime-Advances in Electrical Engineering, *Electronics and Energy*, 6, 100296.

Laptev, N., Yosinski, J., Li, L.E., Smyl, S., (2017). Time-series extreme event forecasting with neural networks at Uber, in: *International Conference on Machine Learning*. SN, pp. 1–5.

Lim, B., Zohren, S., (2021). Time-series forecasting with deep learning: a survey. *Phil. Trans. R. Soc.,* A. 379, 20200209. https://doi.org/10.1098/rsta.2020.0209

Liu, J., Shangguan, D., Liu, S., Ding, Y., Wang, S., & Wang, X. (2019). Evaluation and comparison of CHIRPS and MSWEP daily-precipitation products in the Qinghai-Tibet Plateau during the period of 1981–2015. *Atmospheric Research*, *230*, 104634.

López-Bermeo, C., Montoya, R. D., Caro-Lopera, F. J., & Díaz-García, J. A. (2022). Validation of the accuracy of the CHIRPS precipitation dataset at representing climate variability in a tropical mountainous region of South America. *Physics and Chemistry of the Earth, Parts A/B/C*, *127*, 103184.

Magyari-Sáska, Z., Haidu, I., & Magyari-Sáska, A. (2025). Experimental Comparative Study on Self-Imputation Methods and Their Quality Assessment for Monthly River Flow Data with Gaps: Case Study to Mures River. *Applied Sciences*, *15*(3), 1242. https://doi.org/10.3390/app15031242

Miao, Q., Pan, B., Wang, H., Hsu, K., Sorooshian, S., (2019). Improving monsoon precipitation prediction using combined convolutional and long short-term memory neural network. *Water*, 11, 977.

Musiyam, M., Jumadi, J., Fikriyah, V. N., Masruroh, H., Septiyani, E. D., Amin, C., Hadibasyir,H.Z., Sattar, F., Nawaz, M. (2025). Flood Risk Analysis Using Spatial Synthetic Population in the Upper Bengawan Solo Watershed, Indonesia. *Forum Geografi*, 39(3), 383-397. https://doi.org/10.23917/forgeo.v39i3.13611

Naik, R., Majhi, B., (2025). Explainable AI reverse verification approach for monthly rainfall prediction in Chhattisgarh, India. *Theor Appl Climatol*, 156, 412. https://doi.org/10.1007/s00704-025-05645-2

Nelson, B.K., (1998). Time Series Analysis Using Autoregressive Integrated Moving Average (ARIMA) Models. *Academic Emergency Medicine*, 5, 739–744. https://doi.org/10.1111/j.1553-2712.1998.tb02493.x

Ni, L., Wang, D., Singh, V.P., Wu, J., Wang, Y., Tao, Y., Zhang, J., (2020). Streamflow and rainfall forecasting by two long short-term memory-based models. *Journal of Hydrology,* 583, 124296.

Pan, X., Hou, J., Gao, X., Chen, G., Li, D., Imran, M., Li, X., Yang, N., Ma, M., Zhou, X., (2025). LSTM Model-Based Rapid Prediction Method of Urban Inundation with Rainfall Time Series. *Water Resour Manage,* 39, 661–688. https://doi.org/10.1007/s11269-024-03972-z

Papacharalampous, G., Tyralis, H., Doulamis, N., Doulamis, A., (2025). Ensemble learning for uncertainty estimation with application to the correction of satellite precipitation products. *Machine Learning: Earth,* 1, 015004.

Paredes-Trejo, F. J., Barbosa, H. A., & Kumar, T. L. (2017). Validating CHIRPS-based satellite precipitation estimates in Northeast Brazil. *Journal of arid environments*, *139*, 26-40.

Phoophathong, T., Laosuwan, T., Sangpradid, S., Uttaruk, Y., & Angkahad, T. (2025). Improvement of Traditional and Hybrid Interpolation Techniques Using Support Vector Machine for Land Surface Temperature Analysis in Urban Areas. *Geographia Technica*, *20*(1). DOI: 10.21163/GT_2025.201.21

Praveen, B., Talukdar, S., Shahfahad, Mahato, S., Mondal, J., Sharma, P., Islam, A.R.M.T., Rahman, A., (2020). Analyzing trend and forecasting of rainfall changes in India using non-parametric and machine learning approaches. *Scientific reports*, 10, 10342.

Priyana, Y., Jumadi, Anna, A.N., Rudiyanto, (2023). Farmers' adaptation strategies in facing drought disasters (A case study in some areas of the Bengawan Solo watershed). *Proceedings of the international summit on education, technology, and humanity, AIP Conf. Proc.* 2727, 050026.

Rivera, J. A., Marianetti, G., & Hinrichs, S. (2018). Validation of CHIRPS precipitation dataset along the Central Andes of Argentina. *Atmospheric Research*, *213*, 437-449.

Samson, T. K., & Aweda, F. O. (2025). Comparative study of single and hybrid deep learning models for daily rainfall prediction in selected African cities. *Scientific Reports*, *15*(1), 42718.

Santhyami, S., Jumadi, J., Priyono, K. D., Rahayu, T., Sari, D. N., Othman, M., & Rudi, R. (2025). Spatio-Temporal Mo-deling for the Analysis of Hydrological Drought and its Impact on Rice

Production in the Upper Bengawan Solo Basin, Central Java, Indonesia. *Geographia Technica*, *20*(2). DOI: 10.21163/GT_2025.202.16

Schaffer, A.L., Dobbins, T.A., Pearson, S.-A., (2021). Interrupted time series analysis using autoregressive integrated moving average (ARIMA) models: a guide for evaluating large-scale health interventions. *BMC Med Res Methodol,* 21, 58. https://doi.org/10.1186/s12874-021-01235-8

Shetty, S., Dharmendra, D., Bankapur, S., Prasad, P., (2024). HydroStack: A Hybrid Meta-Ensemble Machine Learning Framework for Accurate Annual Rainfall Prediction, in: 2024 *8th International Conference on I-SMAC* (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC). IEEE, pp. 1926–1934.

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., Woo, W., (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28.

Sivadasan, E.T., Sundaram, N.M., Santhosh, R., (2025). Deep Learning for Energy Forecasting Using Gated Recurrent Units and Long Short-Term Memory. *Journal of Intelligent Systems & Internet of Things*, 14.

Slater, L.J., Arnal, L., Boucher, M.-A., Chang, A.Y.-Y., Moulds, S., Murphy, C., Nearing, G., Shalev, G., Shen, C., Speight, L., (2023). Hybrid forecasting: blending climate predictions with AI models. *Hydrology and Earth System Sciences,* 27, 1865–1889.

Sriwahyuni, L., Nurdiati, S., Nugrahani, E. H., & Najib, M. K. (2025). Performance of Machine Learning for Imputing Missing Daily Rainfall Data in East Java Under Multiple Satellite Data Models. *Geographia Technica*, 20(1). DOI: 10.21163/GT_2025.201.23

Wani, O. A., Mahdi, S. S., Yeasin, M., Kumar, S. S., Gagnon, A. S., Danish, F., ... & Mattar, M. A. (2024). Predicting rainfall using machine learning, deep learning, and time series models across an altitudinal gradient in the North-Western Himalayas. *Scientific Reports*, *14*(1), 27876.

Ward, P.J., Van Pelt, S.C., De Keizer, O., Aerts, J.C.J.H., Beersma, J.J., Van Den Hurk, B.J.J.M., Te Linde, A.H., (2014). Including climate change projections in probabilistic flood risk assessment. *J Flood Risk Management,* 7, 141–151. https://doi.org/10.1111/jfr3.12029

Willard, J., Jia, X., Xu, S., Steinbach, M., Kumar, V., (2020). Integrating physics-based modeling with machine learning: *A survey. arXiv preprint arXiv*:2003.04919 1, 1–34.

Winsemius, H.C., Aerts, J.C., Van Beek, L.P., Bierkens, M.F., Bouwman, A., Jongman, B., Kwadijk, J.C., Ligtvoet, W., Lucas, P.L., Van Vuuren, D.P., (2016). Global drivers of future river flood risk. *Nature Climate Change,* 6, 381–385.

Yang, Y., Tan, G., Shen, Z., Zhang, Y., Fei, Q., Liu, X., Dogar, M.A., (2025). Integrating Physical Dynamics into Ensemble ML for Improved Monthly Rainfall Forecasting. *Earth Syst Environ*. https://doi.org/10.1007/s41748-025-00691-2

Yin, W., Zhou, C., Tian, Y., Qiu, H., Zhang, W., Chen, H., Liu, P., Zhao, Q., Kong, J., Yao, Y., (2025). Accurate Rainfall Prediction Using GNSS PWV Based on Pre-Trained Transformer Model. *Remote Sensing,* 17, 2023.

Zandi, O., Zahraie, B., Nasseri, M., Behrangi, A., (2022). Stacking machine learning models versus a locally weighted linear model to generate high-resolution monthly precipitation over a topographically complex area. *Atmospheric Research*, 272, 106159. https://doi.org/10.1016/j.atmosres.2022.106159

Zhou, Y., Cui, Z., Lin, K., Sheng, S., Chen, H., Guo, S., Xu, C.-Y., (2022). Short-term flood probability density forecasting using a conceptual hydrological model with machine learning techniques. *Journal of Hydrology*, 604, 127255.