KSÜ Mühendislik Bilimleri Dergisi, 29(3), 2026
Araştırma Makalesi

KSU J Eng Sci, 29(3), 2026
Research Article

Kahramanmaras Sutcu Imam University
Journal of Engineering Sciences

Geliş Tarihi : 20.02.2026
Kabul Tarihi : 25.06.2026

Received Date : 20.02.2026
Accepted Date : 25.06.2026

COMPARATIVE EVALUATION OF MACHINE LEARNING-BASED MODELS FOR RAINFALL–RUNOFF PREDICTION

YAĞIŞ-AKIŞ TAHMİNİ İÇİN MAKİNE ÖĞRENİMİNE DAYALI MODELLERİN KARŞILAŞTIRMALI DEĞERLENDİRMESİ

Bestami TAŞAR[1] (ORCID: 0000-0003-4842-3937)

[1] Iskenderun Technical University, Department of Civil Engineering, Hatay, Türkiye

*Sorumlu Yazar / Corresponding Author: *Bestami TAŞAR*, bestami.tasar@iste.edu.tr

ABSTRACT

Accurate rainfall–runoff prediction is important in hydrological modeling. However, watershed processes are nonlinear and dynamic. This study examined the effect of different input configurations on daily streamflow prediction using data-driven methods. A total of 1,826 daily observations covering the period 01/01/2020–09/13/2025 were used. The dataset was split into 1,461 for training and 365 for testing. Three model sets were developed and evaluated using Multiple Linear Regression (MLR), Multilayer Perceptron (MLP), M5 Model Tree (M5 TREE), and Random Tree (RT). Model performance was assessed using RMSE, MAE, MAPE, NSE, R, and R^2. The results showed that prediction accuracy improved as the input structure became more comprehensive. Model Set 3 produced the best overall results. MLP3 achieved the lowest RMSE (3.582 m^3/s) and the highest NSE (0.974) and R^2 (0.974). M5 TREE3 produced the lowest MAE (2.118 m^3/s) and MAPE (4.601%). In contrast, RT models showed weaker and less stable performance. Overall, the results showed that input configuration is as important as algorithm selection in rainfall–runoff modeling. The findings also indicated that nonlinear machine learning models can provide more reliable predictions when suitable lagged variables are included.

Keywords: Rainfall–runoff modeling, streamflow forecasting, machine learning, multilayer perceptron, M5 model tree.

ÖZET

Yağış-akış tahmini, hidrolojik modellemede önemli bir konudur. Ancak havza süreçleri doğrusal olmayan ve dinamik bir yapı göstermektedir. Bu çalışmada, farklı girdi konfigürasyonlarının günlük akım tahmini üzerindeki etkisi veri odaklı yöntemlerle incelenmiştir. Çalışmada 01/01/2020–13/09/2025 dönemini kapsayan toplam 1826 günlük gözlem kullanılmıştır. Veri setinin 1461 adedi eğitim, 365 adedi ise test amacıyla ayrılmıştır. Üç farklı model seti oluşturulmuş ve bunlar Çoklu Doğrusal Regresyon (MLR), Çok Katmanlı Algılayıcı (MLP), M5 Model Ağacı (M5 TREE) ve Rastgele Ağaç (RT) yöntemleri ile değerlendirilmiştir. Model performansı RMSE, MAE, MAPE, NSE, R ve R^2 ölçütleri ile belirlenmiştir. Sonuçlar, girdi yapısı genişledikçe tahmin başarısının arttığını göstermiştir. En iyi genel sonuçlar Model Set 3 ile elde edilmiştir. MLP3 en düşük RMSE (3,582 m^3/s) ile en yüksek NSE (0,974) ve R^2 (0,974) değerlerini vermiştir. M5 TREE3 ise en düşük MAE (2,118 m^3/s) ve MAPE (%4,601) değerlerini üretmiştir. Buna karşılık, RT modelleri daha zayıf ve daha kararsız sonuçlar vermiştir. Genel olarak, bulgular girdi konfigürasyonunun yağış-akış modellemesinde algoritma seçimi kadar önemli olduğunu göstermiştir. Ayrıca, uygun gecikmeli değişkenler kullanıldığında doğrusal olmayan makine öğrenmesi modellerinin daha güvenilir tahminler üretebildiği belirlenmiştir.

Anahtar Kelimeler: Yağış-akış modellemesi, akarsu debisi tahmini, makine öğrenimi, çok katmanlı algılayıcı, M5 model ağacı.

This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0).

INTRODUCTION

Streamflow prediction plays a central role in water resources management and supports a wide range of hydrological applications, from flood mitigation to drought assessment and hydropower planning. As climate variability increases, the need for reliable discharge forecasts has become even more pronounced. Yet, the rainfall–runoff process is inherently nonlinear and influenced by multiple interacting factors, including precipitation variability, temperature fluctuations, basin characteristics, land use patterns, and subsurface conditions. Capturing these interactions with conventional deterministic or purely linear models remains challenging, particularly during extreme flow conditions.

Recent developments in data-driven modeling have introduced alternative tools capable of addressing this complexity. Machine learning approaches, such as Multilayer Perceptron (MLP), Support Vector Machines (SVM), Random Forest (RF), and tree-based regression methods, including M5 Model Trees, have been increasingly adopted in hydrological forecasting studies. These methods offer the advantage of learning the relationship between inputs and outputs directly from observational data, without requiring explicit formulation of the underlying physical processes. As a result, they provide flexible modeling frameworks for representing nonlinear watershed behavior.
Although a substantial body of research has compared the predictive accuracy of various machine learning algorithms for streamflow forecasting, far fewer studies have explicitly examined the role of model configuration. In practice, forecasting performance is influenced not only by the choice of algorithm but also by input variable selection, lag structure, and parameter settings. Evaluating alternative model structures under identical data conditions, therefore, becomes essential to isolate the true contribution of each methodological component. Traditional physics-based and statistical models have long provided the foundation for rainfall–runoff analysis. However, the multivariate and nonlinear nature of hydrological systems often challenges the assumptions embedded in linear or conceptual formulations (Beven, 2001; Moradkhani et al., 2005). Linear regression models, for example, offer interpretability but may struggle to reproduce peak flows or abrupt discharge transitions (Hipel & McLeod, 1994). This limitation has driven the growing adoption of machine learning approaches that can represent nonlinear dependencies without explicitly defining physical relationships. Among these approaches, artificial neural networks—particularly MLP architectures—have frequently demonstrated strong predictive capability when sufficient training data are available (ASCE Task Committee, 2000; Maier & Dandy, 2000). Yet, their performance is closely linked to architectural design choices and the selection of relevant inputs (Dawson & Wilby, 2001). Tree-based methods offer an alternative modeling philosophy. The M5 Model Tree, for instance, partitions the predictor space and applies localized linear models, enabling simultaneous representation of nonlinear segmentation and local linear behavior (Quinlan, 1992; Wang & Witten, 1997). Randomized tree approaches can detect complex data patterns, though their stability may decrease in the presence of noisy or highly variable flow conditions (Breiman, 2001). More recent developments in deep learning have introduced recurrent neural network architectures, such as Long Short-Term Memory (LSTM), which are specifically designed to capture temporal dependencies in hydrological time series (Kratzert et al., 2018, 2019). Large-sample hydrology studies have further emphasized the scalability and transferability of data-driven methods across diverse basins (Shen, 2018). In parallel, ensemble strategies and hybrid frameworks combining feature selection or optimization techniques have reported additional improvements in forecasting skill (Tyralis et al., 2019; Mosavi et al., 2018; Hao et al., 2020; Tikhamarine et al., 2020). Despite these advances, comparatively fewer investigations have isolated the effect of input configuration while maintaining a consistent dataset and evaluation framework. Moreover, reliance on a single performance indicator may obscure important differences in model behavior. A balanced assessment using both error-based (RMSE, MAE, MAPE) and efficiency-based (NSE, R, R^2) metrics provides a more comprehensive understanding of predictive reliability (Legates & McCabe, 1999). Within this perspective, examining how alternative input structures interact with different algorithms remains a relevant and timely research question. In recent years, planning for hydraulic and water infrastructure has extensively used artificial intelligence techniques, commonly known as black-box models (Tasar et al., 2017; Demirci, 2019; Unes et al., 2020; Guzel, 2025). Recent studies have highlighted the growing role of hybrid and optimized machine learning approaches in rainfall–runoff and streamflow modeling (Katipoğlu and Sarıgöl, 2023; Dalkilic et al., 2023; Katipoğlu, 2023). Hybrid frameworks combining physically based and data-driven models (Kumar et al., 2024; Guo et al.) and optimization-enhanced techniques such as SVM–SSA (Samantaray et al., 2022) have demonstrated notable improvements in predictive accuracy. Decomposition-based methods such as EEMD–ANN (Liao et al., 2023) and conventional ANN applications (Verma, 2022) further confirm the ability of machine learning models to capture nonlinear runoff dynamics. Overall, these findings indicate that model structure and input configuration critically influence forecasting performance.

Accordingly, the main objective of this study was to evaluate how alternative input configurations affect model performance under identical data conditions and to identify the most suitable structure for daily streamflow prediction. To examine this issue, three alternative model configurations were constructed for daily streamflow prediction. Each configuration was tested using four different approaches: Multiple Linear Regression (MLR), Multilayer Perceptron (MLP), M5 Model Tree (M5 TREE), and Random Tree (RT). Model skill was assessed on independent test data through a combination of error-based (RMSE, MAE, and MAPE) and efficiency-based (NSE, R, and R^2) performance metrics. Rather than focusing solely on algorithm comparison, this study emphasizes how input configuration affects forecasting reliability. Previous studies have generally focused on comparing the predictive performance of different modeling techniques, while the influence of input structure has received less direct attention under identical data conditions. Therefore, there remains a need for studies that evaluate how alternative input configurations affect model performance within a consistent rainfall–runoff prediction framework. By evaluating alternative setups under the same data conditions, the analysis seeks to clarify whether improvements in prediction accuracy arise primarily from algorithm selection or from the structure of the input variables themselves. The results are intended to provide methodological insight for data-driven rainfall–runoff modeling and to inform the design of more robust operational forecasting frameworks.

MATERIALS AND METHODS

Study Area Data Description

The streamflow data analyzed in this study were obtained from the United States Geological Survey (USGS) gauging station 04121970, located on the Muskegon River near Croton, Michigan, USA. The station is located at approximately 43.41°N, 85.66°W, with an elevation of about 238 m above mean sea level. The contributing drainage area is roughly 5,990 km^2 (2,313 square miles), representing a substantial portion of the Lower Peninsula of Michigan. The basin lies within a humid continental climate zone characterized by cold winters and moderately warm summers. Precipitation occurs throughout the year; however, seasonal snow accumulation and subsequent spring snowmelt strongly influence annual peak discharges. As a result, the watershed exhibits a mixed hydrological regime in which both rainfall-driven and snowmelt-driven processes contribute to streamflow generation. Lower discharge values are typically observed during late summer, when evapotranspiration rates increase and precipitation inputs decline. The long-term mean annual discharge at the Croton station ranges between 70 and 75 m^3/s (approximately 2,400–2,600 ft^3/s), though considerable seasonal and interannual variability is evident. Spring months generally correspond to the highest flows, while late summer and early autumn are associated with reduced discharge levels. The watershed is dominated by forested areas, wetlands, and agricultural lands shaped by glacial geomorphology typical of the Great Lakes region. The basin is characterized by a mixed hydrological regime and a heterogeneous physiographic structure. Forested areas, wetlands, and agricultural lands constitute the dominant land cover, while glacial geomorphological features typical of the Great Lakes region influence infiltration, runoff generation, and baseflow response. These catchment characteristics contribute to the nonlinear rainfall–runoff behavior observed in the study area. Variations in soil permeability, land cover distribution, and basin topography influence infiltration, surface runoff, and baseflow contributions, leading to nonlinear flow responses. The availability of continuous daily discharge records from USGS provides a consistent and reliable basis for model development and independent validation. Owing to its climatic variability and physiographic diversity, the basin offers a suitable framework for assessing how alternative machine learning configurations perform under changing hydrological conditions. The study area is shown in Figure 1.

Data Description and Model Configurations

Daily hydrometeorological records, including observed streamflow and selected predictor variables relevant to discharge estimation, were used in the analysis. The dataset covered the period from 01/01/2020 to 09/13/2025 and consisted of 1,826 daily observations. Streamflow data were obtained from the United States Geological Survey (USGS), while meteorological variables were derived from the PRISM dataset, both at a daily temporal resolution. The available time series was split into two segments: 1,461 observations for model training and 365 for independent testing. The training portion was employed to calibrate model structures and optimize parameters, whereas the testing period was retained exclusively to evaluate predictive skill under unseen conditions. A chronological train–test split was preferred over cross-validation to preserve the dataset's temporal structure and avoid data leakage. The use of an independent test period also helped reduce the risk of overfitting by evaluating model performance on unseen data. Prior to model development, the dataset was screened for missing, inconsistent, and potentially erroneous records. As part of the data-cleaning procedure, observations with missing values were removed from the study. This

procedure improved data consistency and reduced the potential influence of incomplete or erroneous observations on model performance. Before model development, the input and output variables were normalized to the range 0–1 to eliminate scale differences among variables and improve model performance. Extreme values were examined during the data screening stage; however, they were not automatically excluded from the dataset, as they may reflect actual hydrological conditions rather than measurement errors. The temporal distribution of the observed discharge data and the applied train–test partition is illustrated in Figure 2.

Figure 1. Location of the Study Area and the USGS Gauging Station 04121970 in Michigan, USA

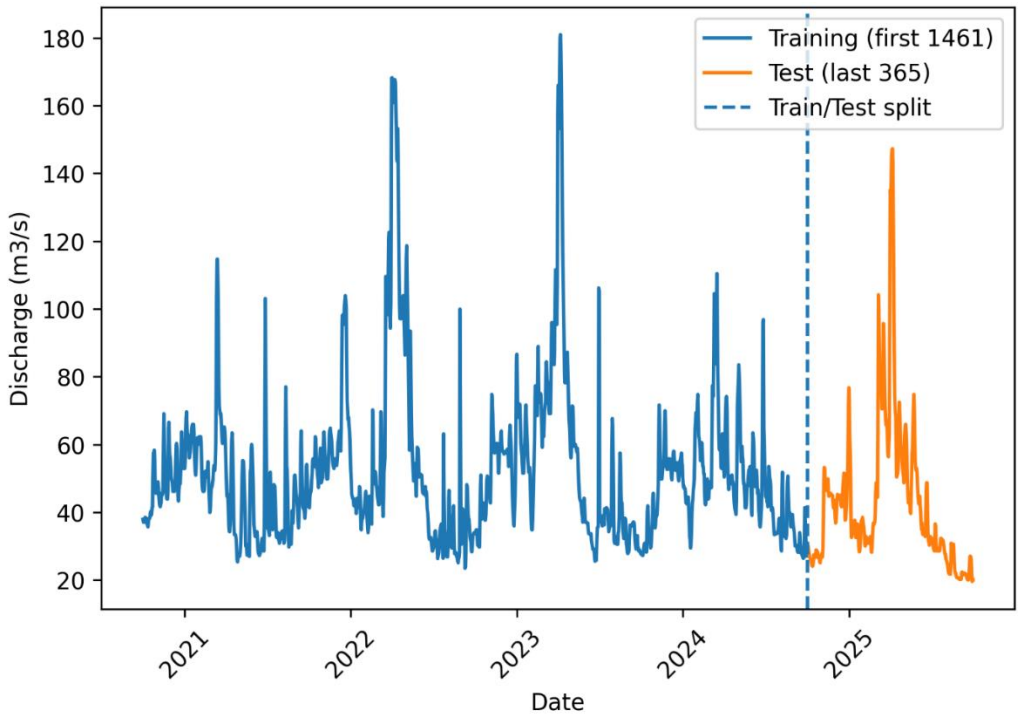

Figure 2. Observed Daily Discharge Time Series Showing the Training and Testing Periods Used in This Study

To explore how input structure affects forecasting accuracy, three alternative model configurations were constructed. Each configuration differed in its selection of input variables and lag structure, allowing temporal dependencies within the streamflow series to be represented in different ways. Model Set 1 included air temperature, current precipitation, and one-day lagged streamflow (Q_{t-1}). Model Set 2 was developed by adding one-day lagged precipitation (P_{t-1}) to the first configuration (Model Set 1). Model Set 3 was developed by extending Model Set 2 by including an additional two-day lagged streamflow variable (Q_{t-2}). In this way, the effect of different combinations of meteorological inputs and hydrological memory components on predictive performance was systematically examined.

Hydrological systems typically exhibit memory effects, meaning that current discharge conditions are influenced by antecedent conditions. For this reason, lagged discharge values and relevant meteorological inputs were incorporated at varying combinations across the configurations. By structuring the inputs differently, it becomes possible to assess whether changes in prediction skill originate from algorithm choice or from the way temporal information is organized. Each configuration was evaluated using four predictive approaches, namely Multiple Linear Regression (MLR), Multilayer Perceptron (MLP), M5 Model Tree (M5 TREE), and Random Tree (RT).

Applying the same algorithms to each configuration ensures a consistent basis for comparison between linear, neural network, and tree-based modeling strategies under identical data conditions.

It should be noted that the use of same-day meteorological predictors in this study represents a retrospective simulation setting rather than a strict real-time operational forecasting framework. Same day temperature values may not be fully available at the time of prediction in practical forecasting applications unless forecasted meteorological inputs are used. Therefore, the developed models should be interpreted as offline rainfall–runoff simulations based on historical observations, not as directly deployable real-time forecasting models.

Multiple Linear Regression (MLR)

Multiple Linear Regression was used as a baseline model to establish a reference performance level. The MLR model expresses streamflow as a linear combination of selected input variables:

$$y_0 = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \quad (1)$$

In this equation "y_0" represents dependent variable, "x_1, x_2, x_3…in" represents independent variables, "β_1, β_2, β_3 ...β_n" represents independent regression coefficient and "β_0" represents intercept. Although MLR assumes linear relationships between predictors and the output, it remains valuable for its interpretability and computational efficiency. In the present study, daily streamflow (Q_t) was treated as the dependent variable, and the independent variables were defined according to the model configuration. Air temperature was represented by minimum (T_{min}), mean (T_{mean}), and maximum (T_{max}) temperature values. Accordingly, three different multiple linear regression formulations were developed for Model Sets 1, 2, and 3, and their corresponding equations are presented in Eqs. (2)–(4).

$$Q_t = 1.76 + 0.302 * T_{min} - 0.6 * T_{mean} + 0.24 * T_{max} + 0.58 * P_t + 0.95 * Q_{t-1} \quad (2)$$

$$Q_t = 1.36 + 0.48 * T_{min} - 1.06 * T_{mean} + 0.50 * T_{max} + 0.57 * P_t + 0.22 * P_{t-1} + 0.95 * Q_{t-1} \quad (3)$$

$$Q_t = 2.03 + 0.486 * T_{min} - 1.068 * T_{mean} + 0.51 * T_{max} + 0.54 * P_t + 0.072 * P_{t-1} + 1.213 * Q_{t-1} - 0.27 * Q_{t-2} \quad (4)$$

As shown in Eqs. (2)–(4), Model Set 1 was constructed using temperature variables, current precipitation (P_t), and one-day lagged streamflow (Q_{t-1}). Model Set 2 extended this structure by including one-day lagged precipitation (P_{t-1}), whereas Model Set 3 further extended Model Set 2 by incorporating an additional two-day lagged streamflow variable (Q_{t-2}).

Since T_{min}, T_{mean}, and T_{max} were simultaneously used as temperature-related predictors in the regression formulations, an additional multicollinearity check was performed using Pearson correlation and Variance Inflation Factor (VIF) analyses. The results indicated strong correlations among the temperature variables, particularly between T_{min}–T_{mean} ($R = 0.973$), T_{mean}–T_{max} ($R = 0.982$), and T_{min}–T_{max} ($R = 0.911$). The VIF results also confirmed severe multicollinearity among the temperature-related predictors, whereas the VIF values of precipitation and lagged discharge variables remained low. These findings confirm that the three temperature variables are not fully independent, as they represent different daily statistics of the same thermal regime. Therefore, the regression coefficients from the MLR models should be interpreted with caution and not regarded as isolated physical effects

of individual temperature variables. Instead, the MLR models were used only as baseline predictive models within the comparative modeling framework of this study.

Multilayer Perceptron (MLP)

The Multilayer Perceptron is a feedforward artificial neural network capable of modeling nonlinear relationships. The network architecture consists of an input layer, one or more hidden layers, and an output layer. Neurons in each layer are connected through weighted links and nonlinear activation functions.

The general mathematical structure of an MLP with one hidden layer can be expressed as:

$$Q_t = f\left(\sum_{j=1}^{m} w_j \cdot g\left(\sum_{i=1}^{N} v_{ij} \cdot X_{i,t} + b_j\right) + c\right) \quad (5)$$

where $g(\cdot)$ and $f(\cdot)$ are activation functions, v_{ij} and w_j are connection weights, and b_j and c are bias terms. As shown in Figure 3, the MLP architecture comprises lagged discharge values, precipitation inputs (t and t−1), and temperature statistics (T_{min}, T_{max}, T_{mean}), which are processed through a single hidden layer to produce the predicted discharge.

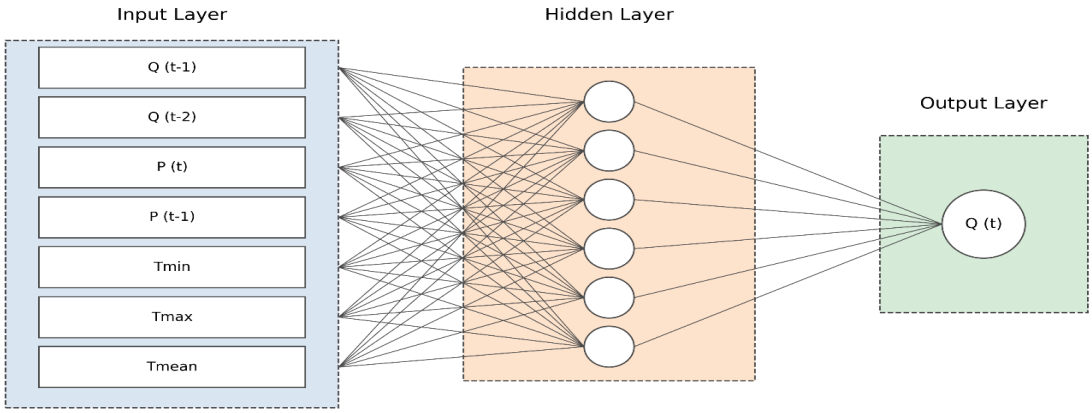

Figure 3. Architecture of the Multilayer Perceptron (MLP) Model used in Model Set 3

The network parameters were optimized using a supervised learning algorithm to minimize prediction error on the training dataset.

M5 Model Tree

The M5 Model Tree algorithm is a tree-based regression technique that partitions the input space into subregions and fits linear regression models at the terminal nodes. This hybrid structure allows the model to capture both nonlinear partitioning behavior and local linear relationships. The model construction process involves recursive splitting based on variance reduction, followed by pruning and smoothing to improve generalization.

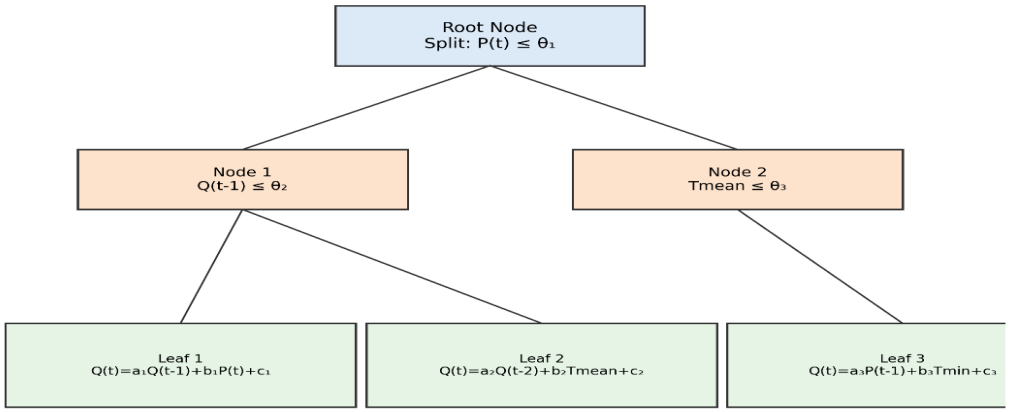

Figure 4. Structure of the M5 Model Tree

The conceptual structure of the M5 Model Tree applied in this study is illustrated in Figure 4. The model recursively partitions the input space based on splitting criteria that minimize intra-node variance. Unlike conventional regression

trees, linear regression equations are constructed at each terminal node, enabling the model to represent local linear relationships within different subregions of the dataset. This hybrid structure allows the M5 approach to capture both nonlinear decision boundaries and locally linear discharge dynamics. The M5 decision tree algorithm was originally developed by Quinlan (1992). Witten and Frank (2005) provided a comprehensive description of this technique. The M5 algorithm operates by generating a regression sequence through iterative division of the sample space. Readers who want to learn more about the M5 model tree can examine Quinlan (1992).

Random Tree (RT)

The Random Tree algorithm is a stochastic decision tree method that introduces randomness in feature selection during tree construction. By randomly selecting input variables at each split, the method aims to reduce overfitting and improve generalization. However, as a single-tree method, its performance may be more sensitive to noise and data variability than that of ensemble-based extensions.

The MLP, M5 TREE, and RT models were implemented in the WEKA environment. The main model parameters were determined according to each method and finalized through trial-and-error during model development. Prediction performance and consistency on the independent test dataset were considered when selecting the final model settings. The final MLP model was configured with a hidden layer size of 1, a learning rate of 0.3, a momentum coefficient of 0.2, and 500 training iterations. The M5 TREE model was developed with a minimum of 4 instances per leaf and a batch size of 100. The RT model was configured with a batch size of 100, a minimum of 1 instance per leaf, a minimum variance proportion of 0.001, and a seed value of 1.

Performance Evaluation Criteria

Model performance was assessed using an independent test dataset and quantified through several complementary statistical indicators, including Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Nash–Sutcliffe Efficiency (NSE), the correlation coefficient (R), and the coefficient of determination (R^2). The RMSE and MAE measure absolute prediction errors, while the MAPE provides relative percentage error. NSE evaluates predictive skill relative to the mean of observations. R and R^2 quantify the strength of the linear association between observed and predicted values. Using multiple performance indicators ensures a comprehensive and unbiased evaluation of model behavior under different hydrological conditions.

RESULTS AND DISCUSSION

The predictive performances of the three model configurations were evaluated using independent test data. The comparison was conducted based on error-based metrics (RMSE, MAE, and MAPE) and efficiency-based indicators (NSE, R, and R^2). The results reveal clear differences between the model structures and input configurations. The mathematical formulations of these performance metrics are presented in Equations (6)–(10).

$$RMSE = \sqrt{\frac{1}{N} * \left(\sum_{i=1}^{N} \mathrm{Q}i_{observation} - \mathrm{Q}i_{prediction}\right)^2} \quad (6)$$

$$MAE = \frac{1}{N} * \sum_{i=1}^{N} \left| Qi_{observation} - Qi_{prediction} \right| \quad (7)$$

$$MAPE = \frac{100}{N} * \frac{\sum_{i=1}^{N} Qi_{observation} - Qi_{prediction}}{\sum_{i=1}^{N} Qi_{observation}} \quad (8)$$

$$NSE = 1 - \frac{\sum_{i=1}^{N} Qi_{observation} - Qi_{prediction}}{\sum_{i=1}^{N} Qi_{observation} - \overline{Qi_{observed}}} \quad (9)$$

$$R = \frac{\sum_{i=1}^{N} (Qi_{observation} - \overline{Qi_{observed}})(Qi_{prediction} - \overline{Qi_{prediction}})}{\sqrt{\sum_{i=1}^{N} (Qi_{observation} - \overline{Qi_{observed}})^2} \sqrt{\sum_{i=1}^{N} (Qi_{prediction} - \overline{Qi_{prediction}})^2}} \quad (10)$$

where N denotes the number of data sets, and Qi denotes the runoff data. Table 1 displays the model results. Overall, Model Set 3 demonstrated the highest predictive capability among the three configurations. In particular, the MLP3 model achieved the lowest RMSE (3.582 m^3/s) and the highest NSE (0.974), R (0.989), and R^2 (0.974) values. These results indicate that the MLP architecture's nonlinear mapping capability effectively captured the complex hydrological relationships embedded in the dataset. The high correlation and efficiency values suggest strong agreement between observed and predicted discharge values during the test period. Model Set 2 also produced robust results, particularly with the MLP2 model (RMSE = 3.861 m^3/s; NSE = 0.970). While slightly inferior to Model Set 3, this configuration still demonstrated strong generalization performance. In contrast, Model Set 1 yielded higher

error values and lower efficiency scores, suggesting that the input configuration in this set may not fully capture the temporal dependencies governing streamflow variability. These findings confirm that input structure plays a decisive role in model performance, sometimes even more than algorithm selection alone. The comparative performance of all model configurations is presented in Table 1.

The predictive performance of the models is quantitatively summarized in Table 1 and visually illustrated in Figures 5–7 for the test period. As reported in Table 1, Model Set 1 exhibits moderate accuracy, with M5 TREE1 achieving the lowest RMSE (4.392 m³/s) and highest NSE (0.961), while RT1 shows substantially weaker performance (NSE = 0.892). These findings are consistent with the dispersion patterns shown in Figure 5, where predictions from M5 TREE1 and MLP1 are more closely aligned with the 1:1 reference line, whereas RT1 shows larger deviations and greater scatter. For Model Set 2, Table 1 indicates an overall improvement in predictive accuracy, particularly for MLP2 (RMSE = 3.861 m³/s; NSE = 0.970) and M5 TREE2 (MAE = 2.278 m³/s), which is also reflected in the tighter clustering of points around the identity line in Figure 6. The strongest overall performance is observed in Model Set 3 (Table 1), where MLP3 attains the highest NSE (0.974) and lowest RMSE (3.582 m³/s), while M5 TREE3 minimizes absolute errors (MAPE = 4.601%). As illustrated in Figure 7, the predicted and observed discharges exhibit a high degree of agreement, with reduced dispersion and improved alignment compared to the previous configurations. Overall, the graphical comparisons in Figures 5–7 support the quantitative findings presented in Table 1, confirming the progressive improvement in prediction accuracy from Model Set 1 to Model Set 3.

Table 1. Statistical Performance Metrics of Model Sets

Model	RMSE (m^3/s)	MAE (m^3/s)	MAPE (%)	NSE	R	R^2
MLR1	4.65	2.511	5.53	0.956	0.978	0.956
MLP1	4.536	2.498	5.383	0.958	0.979	0.958
M5 TREE1	**4.392**	**2.362**	**5.069**	**0.961**	**0.980**	**0.961**
RT1	7.277	5.27	14.858	0.892	0.950	0.892
MLR2	4.288	2.498	5.686	0.963	0.981	0.963
MLP2	**3.861**	2.592	6.082	**0.970**	**0.986**	**0.970**
M5 TREE2	3.881	**2.278**	**5.043**	0.969	0.985	0.969
RT2	5.887	3.987	9.899	0.929	0.966	0.929
MLR3	3.92	2.215	4.961	0.969	0.984	0.969
MLP3	**3.582**	2.36	5.917	**0.974**	**0.989**	**0.974**
M5 TREE3	3.825	**2.118**	**4.601**	0.970	0.985	0.970
RT3	7.646	4.711	11.35	0.881	0.940	0.881

Discharge distribution comparisons between observed and simulated streamflow (left panels) and corresponding scatter plots (right panels) for (a) MLR1, (b) MLP1, (c) M5 TREE1, and (d) RT1. Statistical performance metrics (NSE, R^2, and R) are provided within each scatter plot. Across the three model configurations, a progressive improvement in predictive performance was observed from Model Set 1 to Model Set 3, highlighting the influence of input structure on forecasting accuracy. Within Model Set 1, M5 TREE1 yielded the best overall results (RMSE = 4.392 m³/s, NSE = 0.961), slightly outperforming MLP1 (RMSE = 4.536 m³/s, NSE = 0.958) and MLR1 (RMSE = 4.65 m³/s, NSE = 0.956), whereas RT1 produced substantially weaker performance (RMSE = 7.277 m³/s, NSE = 0.892). In Model Set 2, accuracy further improved, with MLP2 achieving the lowest RMSE (3.861 m³/s) and highest NSE (0.970), while M5 TREE2 provided the lowest absolute errors (MAE = 2.278 m³/s; MAPE = 5.043%). Model Set 3 demonstrated the strongest overall results, where MLP3 attained the best goodness-of-fit (RMSE = 3.582 m³/s; NSE = 0.974; R^2 = 0.974), and M5 TREE3 minimized absolute deviations (MAE = 2.118 m³/s; MAPE = 4.601%). When models are compared across configurations, MLPs consistently produce the highest NSE and R^2 values, confirming neural networks' ability to capture nonlinear rainfall–runoff dynamics. M5 Model Tree showed stable and competitive performance, particularly in reducing absolute error metrics, reflecting the advantage of localized linear modeling. Although MLR achieved relatively strong results (e.g., MLR3 NSE = 0.969), it remained slightly inferior to nonlinear approaches. In contrast, Random Tree models consistently exhibited the lowest predictive accuracy (NSE range: 0.881–0.929), suggesting that single-tree structures may be insufficient to represent complex hydrological variability.

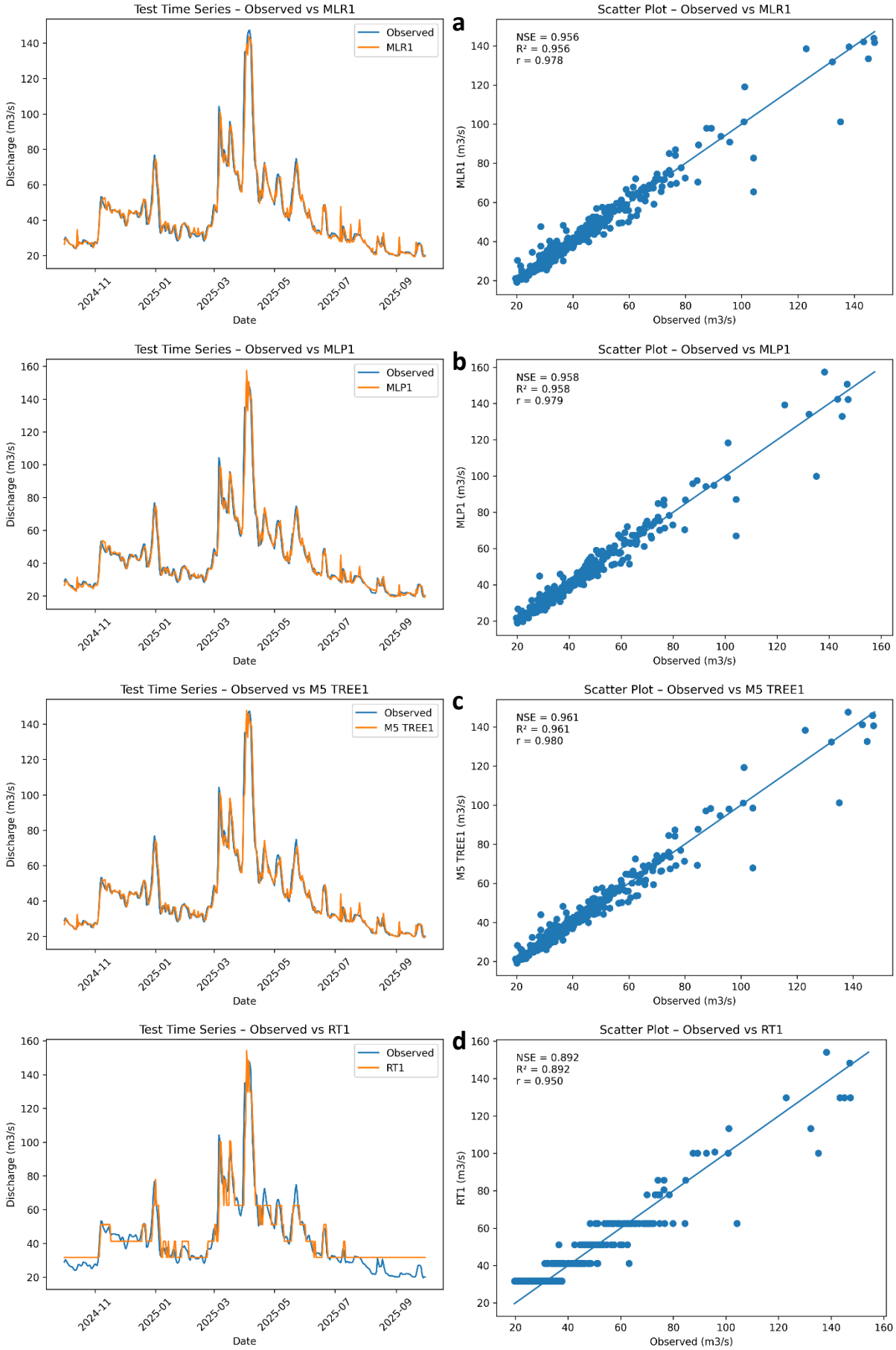

Figure 5. Observed and Predicted Discharge Comparisons for Model Set 1 During the Test Period

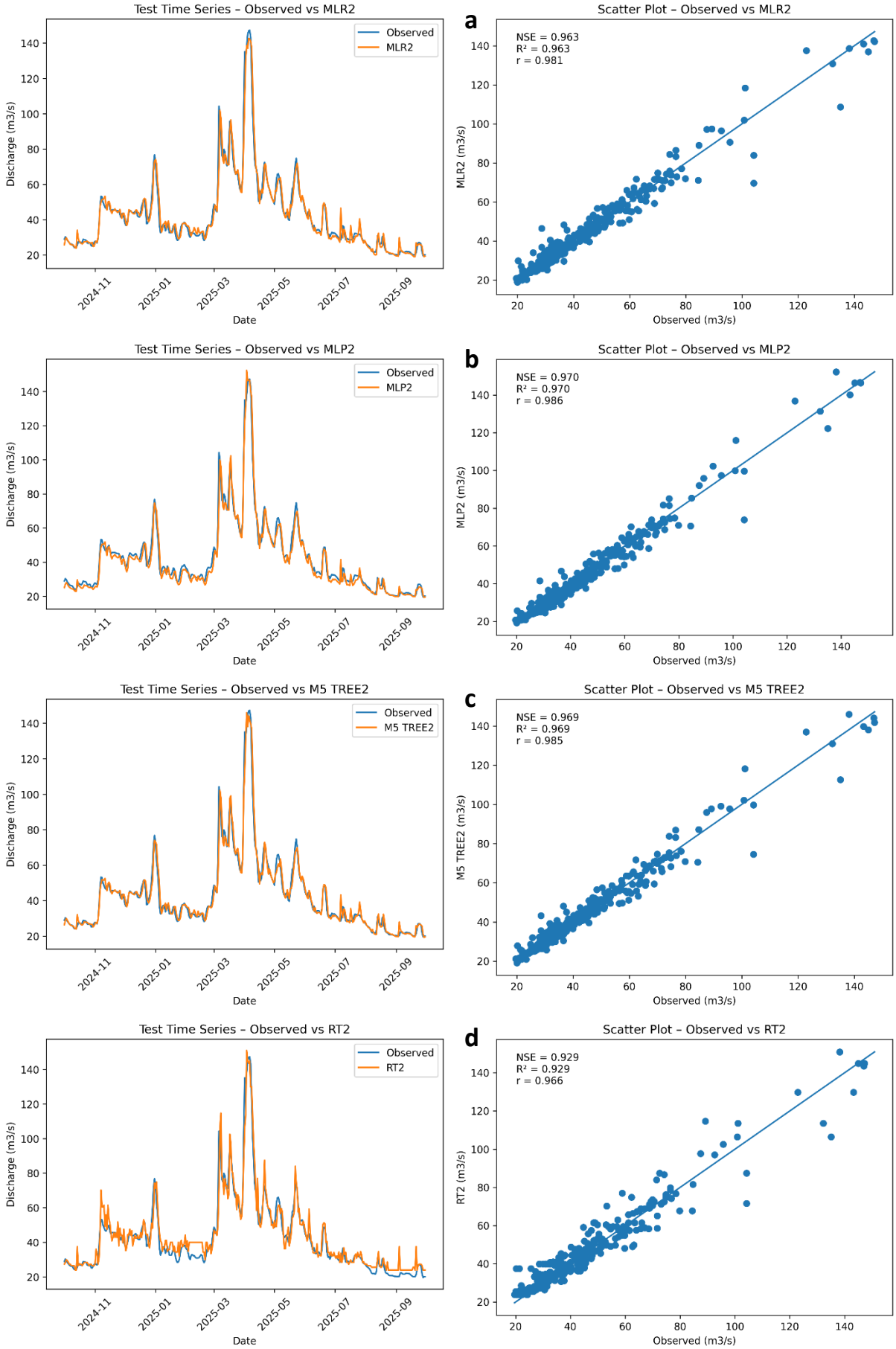

Figure 6. Observed and Predicted Discharge Comparisons for Model Set 2 during the Test Period

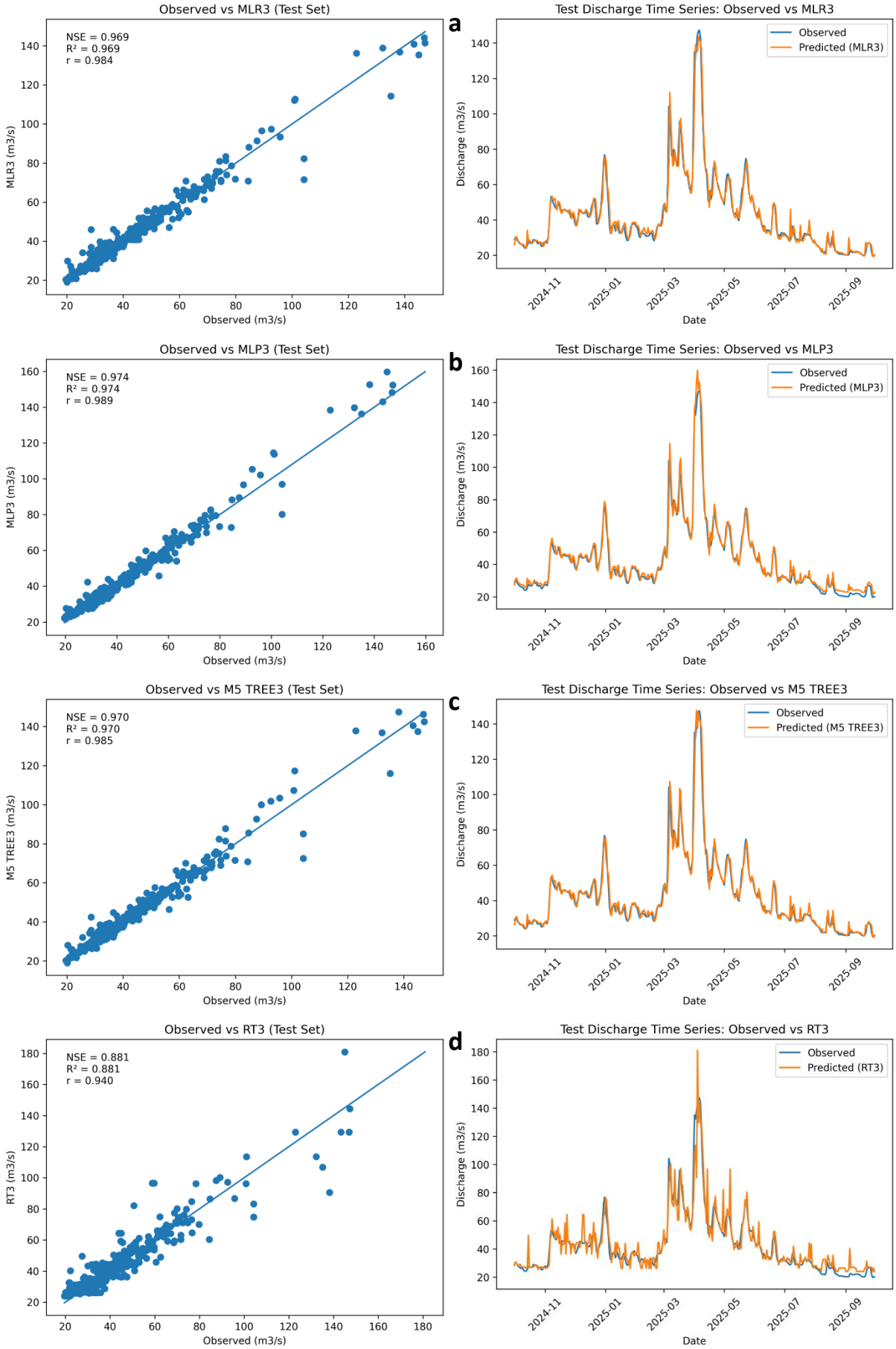

Figure 7. Observed and Predicted Discharge Comparisons for Model Set 3 during the Test Period

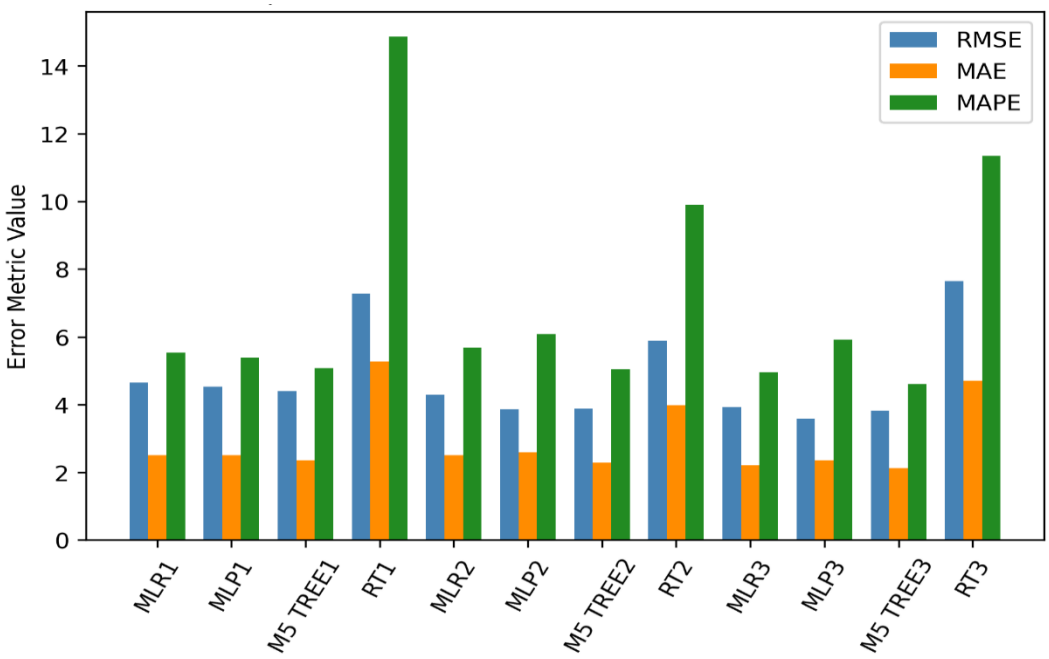

Figure 8. Comparative Error Performance of Model Sets 1–3 based on RMSE, MAE, and MAPE Metrics

The comparative error performance of all models is illustrated in Figure 8. It is evident that Model Set 3 consistently produced lower RMSE and MAE values compared to Model Sets 1 and 2. Among the evaluated algorithms, MLP3 achieved the minimum RMSE, indicating superior predictive accuracy in terms of overall error magnitude. Similarly, M5 TREE3 yielded the lowest MAE and MAPE values, suggesting improved robustness in minimizing absolute and percentage deviations. In contrast, the Random Tree models exhibited notably higher errors across all configurations, particularly in MAPE, reflecting reduced stability during extreme flow conditions. Overall, Figure 2 confirms that nonlinear machine learning approaches outperform the linear baseline model in terms of predictive reliability.

To provide a more comprehensive comparison of model performance, a rank-based evaluation was also considered across the selected performance metrics. This comparison supported the overall superiority of the best-performing configurations by considering the collective behavior of RMSE, MAE, MAPE, NSE, R, and R^2 rather than relying on a single metric. To provide a more comprehensive comparison of model performance, a rank-based analysis was conducted across all evaluation metrics. For RMSE, MAE, and MAPE, lower values were considered better, whereas for NSE, R, and R^2, higher values were considered better. The models were ranked separately for each metric, and the average rank was then calculated for each model, as presented in Table 2. According to the Rank analysis, M5 TREE3 achieved the best overall performance with the lowest average rank (2.08), followed by MLP3 (2.67), M5 TREE2 (3.75), and MLR3 (3.83). Although MLP3 produced the best RMSE, NSE, R, and R^2 values, M5 TREE3 ranked first overall due to its superior MAE and MAPE performance. In contrast, the RT models consistently ranked lowest overall.

The findings should be interpreted with some caution. Since the analysis was conducted on a single watershed, the observed model behavior may partly reflect the specific hydrological characteristics of the study basin. Although the dataset was sufficient for model development and testing, broader datasets and real-time forecasting applications were not considered in the present study. For real-time operational forecasting, future studies should consider previous-day temperature variables, other lagged meteorological predictors, or forecasted meteorological inputs. Moreover, dimensionality reduction techniques such as principal component analysis or factor analysis, or the selection of a single representative temperature variable based on correlation analysis, may be useful to reduce redundancy among temperature-related predictors and improve the interpretability of regression-based models.

In addition, future studies may employ dimensionality reduction techniques such as principal component analysis or factor analysis, or select a single representative temperature variable based on correlation analysis, to reduce redundancy among temperature-related predictors and improve the interpretability of regression-based models.

Table 2. Rank-Based Comparison of Model Performance across All Evaluation Metrics

Model	RMSE Rank	MAE Rank	MAPE Rank	NSE Rank	R Rank	R^2 Rank	Average Rank	Overall Rank
M5 TREE3	2.0	1.0	1.0	2.5	3.5	2.5	2.08	1
MLP3	1.0	4.0	8.0	1.0	1.0	1.0	2.67	2
M5 TREE2	4.0	3.0	3.0	4.5	3.5	4.5	3.75	3
MLR3	5.0	2.0	2.0	4.5	5.0	4.5	3.83	4
MLP2	3.0	9.0	9.0	2.5	2.0	2.5	4.67	5
M5 TREE1	7.0	5.0	4.0	7.0	7.0	7.0	6.17	6
MLR2	6.0	6.5	7.0	6.0	6.0	6.0	6.25	7
MLP1	8.0	6.5	5.0	8.0	8.0	8.0	7.25	8
MLR1	9.0	8.0	6.0	9.0	9.0	9.0	8.33	9
RT2	10.0	10.0	10.0	10.0	10.0	10.0	10.00	10
RT1	11.0	12.0	12.0	11.0	11.0	11.0	11.33	11
RT3	12.0	11.0	11.0	12.0	12.0	12.0	11.67	12

CONCLUSIONS

This study evaluated the effects of different model configurations and machine learning algorithms on daily streamflow prediction performance. Three input configurations were tested using four approaches: MLR, MLP, M5 Model Tree, and Random Tree. Model performance was assessed using RMSE, MAE, MAPE, NSE, R, and R^2 on independent test data.

The findings indicate that both algorithm selection and input configuration significantly influence predictive accuracy. Among the evaluated setups, Model Set 3 consistently produced superior results, with MLP3 achieving the lowest RMSE and highest NSE and R^2 values. M5 TREE3 demonstrated strong performance in minimizing absolute and percentage errors. In contrast, Random Tree models exhibited comparatively weaker results across all configurations.

Overall, nonlinear machine learning models outperformed the linear baseline approach, highlighting the importance of capturing complex hydrological relationships. The results emphasize that optimized input structure and lag selection are critical for improving forecasting reliability. These findings support the applicability of machine learning techniques in retrospective rainfall–runoff simulation and provide methodological guidance for the development of future operational forecasting frameworks.

This study is based on a single watershed; therefore, the obtained findings reflect the hydrological characteristics of the selected basin. In addition, the models were evaluated using historical data in an offline setting, and real-time forecasting performance was not examined. Although model performance was assessed using multiple statistical indicators, a formal statistical significance test was not included in the present study. Future studies should evaluate the proposed modeling framework using broader datasets, different catchments, and real-time forecasting conditions to test its wider applicability and strengthen comparative model assessment.

ACKNOWLEDGEMENTS

The author gratefully acknowledges the U.S. Geological Survey (USGS) for making the hydrological data used in this study publicly available.

Artificial Intelligence Contribution Statement

Artificial intelligence tools were used only as supportive tools in this manuscript. Grammarly was used for grammar and spelling corrections, while ChatGPT was used to improve text clarity and assist with the visual presentation of figures. All analyses, scientific interpretations, and conclusions were performed by the author.

REFERENCES

ASCE Task Committee on Application of Artificial Neural Networks in Hydrology. (2000). Artificial neural networks in hydrology. I: Preliminary concepts. *Journal of Hydrologic Engineering, 5*(2), 115–123. https://doi.org/10.1061/(ASCE)1084-0699(2000)5:2(115)

Beven, K. (2001). *Rainfall–runoff modelling: The primer*. John Wiley & Sons.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Dalkilic, H. Y., Kumar, D., Samui, P., Dixon, B., Yesilyurt, S. N., & Katipoğlu, O. M. (2023). Application of deep learning approaches to predict monthly stream flows. *Environmental Monitoring and Assessment, 195*, 705. https://doi.org/10.1007/s10661-023-11331-5

Dawson, C. W., & Wilby, R. L. (2001). Hydrological modelling using artificial neural networks. *Progress in Physical Geography, 25*(1), 80–108. https://doi.org/10.1177/030913330102500104

Demirci, M. (2019). Destek vektör makineleri ve M5 karar ağacı yöntemleri kullanılarak yağış akış ilişkisinin tahmini. *Dicle Üniversitesi Mühendislik Fakültesi Mühendislik Dergisi*, *10*(3), 1113-1124. https://doi.org/10.24012/dumf.525658

Guo, J., Liu, Y., Zou, Q., Ye, L., Zhu, S., & Zhang, H. (2023). Study on optimization and combination strategy of multiple daily runoff prediction models coupled with physical mechanism and LSTM. *Journal of Hydrology*, *624*, 129969. https://doi.org/10.1016/j.jhydrol.2023.129969

Guzel, H. (2025). Estimation of Rainfall-Runoff Relationship Using Soft Computing Techniques. *Applied Ecology & Environmental Research*, *23*(2), 1923-1936. https://doi.org/10.15666/aeer/2302_19231936

Hao, Z., Singh, V. P., & Xia, Y. (2020). Seasonal drought prediction: Advances, challenges, and future prospects. *Reviews of Geophysics, 56*(1). https://doi.org/10.1002/2016RG000549

Hipel, K. W., & McLeod, A. I. (1994). *Time series modelling of water resources and environmental systems*. Elsevier.

Katipoğlu, O. M. (2023). Modeling the effect of meteorological variables on streamflow estimation: Application of data mining techniques in mixed rainfall–snowmelt regime Munzur River, Türkiye. *Environmental Science and Pollution Research, 30*, 96312–96328. https://doi.org/10.1007/s11356-023-29220-2

Katipoğlu, O. M., & Sarıgöl, M. (2023). Improving the accuracy of rainfall-runoff relationship estimation using signal processing techniques, bio-inspired swarm intelligence and artificial intelligence algorithms. *Earth Science Informatics, 16*, 3125–3141. https://doi.org/10.1007/s12145-023-01081-w

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A., Hochreiter, S., & Nearing, G. (2019). Toward improved predictions in ungauged basins: Exploiting the power of machine learning. *Water Resources Research, 55*(12), 11344–11354. https://doi.org/10.1029/2019WR026065

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., & Nearing, G. (2018). Rainfall–runoff modelling using long short-term memory (LSTM) networks. *Hydrology and Earth System Sciences, 22*, 6005–6022. https://doi.org/10.5194/hess-22-6005-2018

Kumar, S., Choudhary, M. K., & Thomas, T. (2024). A hybrid technique to enhance the rainfall-runoff prediction of physical and data-driven model: a case study of Upper Narmada River Sub-basin, India. *Scientific Reports*, *14*(1), 26263. https://doi.org/10.1038/s41598-024-77655-5

Legates, D. R., & McCabe, G. J. (1999). Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. *Water Resources Research, 35*(1), 233–241. https://doi.org/10.1029/1998WR900018

Liao, S., Wang, H., Liu, B., Ma, X., Zhou, B., & Su, H. (2023). Runoff forecast model based on an EEMD-ANN and meteorological factors using a multicore parallel algorithm. *Water Resources Management*, *37*(4), 1539-1555. https://doi.org/10.1007/s11269-023-03442-y

Maier, H. R., & Dandy, G. C. (2000). Neural networks for the prediction and forecasting of water resources variables: A review. *Environmental Modelling & Software, 15*(1), 101–124. https://doi.org/10.1016/S1364-8152(99)00007-9

Moradkhani, H., Hsu, K. L., Gupta, H., & Sorooshian, S. (2005). Uncertainty assessment of hydrologic model states and parameters. *Water Resources Research, 41*(5). https://doi.org/10.1029/2004WR003604

Mosavi, A., Ozturk, P., & Chau, K. W. (2018). Flood prediction using machine learning models: Literature review. *Water, 10*(11), 1536. https://doi.org/10.3390/w10111536

Quinlan, J. R. (1992). Learning with continuous classes. In *Proceedings of the 5th Australian Joint Conference on Artificial Intelligence* (pp. 343–348). https://doi.org/10.1142/1897

Samantaray, S., Das, S. S., Sahoo, A., & Satapathy, D. P. (2022). Monthly runoff prediction at Baitarani river basin by support vector machine based on Salp swarm algorithm. *Ain Shams Engineering Journal*, *13*(5), 101732. https://doi.org/10.1016/j.asej.2022.101732

Shen, C. (2018). A transdisciplinary review of deep learning research and its relevance for water resources scientists. *Water Resources Research, 54*(11), 8558–8593. https://doi.org/10.1029/2018WR022643

Taşar, B., Kaya, Y. Z., Varçin, H., Üneş, F., & Demirci, M. (2017). Forecasting of suspended sediment in rivers using artificial neural networks approach. *International Journal of Advanced Engineering Research and Science*, *4*(12), 79-84. https://doi.org/10.22161/ijaers.4.12.14

Tikhamarine, Y., Souag-Gamane, D., Ahmed, A. N., Kisi, O., & El-Shafie, A. (2020). Improving artificial intelligence models accuracy for monthly streamflow forecasting using data preprocessing techniques. *Journal of Hydrology, 582*, 124435. https://doi.org/10.1016/j.jhydrol.2019.124435

Tyralis, H., Papacharalampous, G., & Langousis, A. (2019). A brief review of random forests for water scientists and practitioners and their recent history in water resources. *Water, 11*(5), 910. https://doi.org/10.3390/w11050910

U.S. Geological Survey. (n.d.). *National Water Information System: Web interface. https://waterdata.usgs.gov/nwis

Üneş, F., Demirci, M., Zelenakova, M., Çalışıcı, M., Taşar, B., Vranay, F., & Kaya, Y. Z. (2020). River flow estimation using artificial intelligence and fuzzy techniques. *Water*, *12*(9), 2427. https://doi.org/10.3390/w12092427

Verma, R. (2022). ANN-based rainfall-runoff model and its performance evaluation of Sabarmati River Basin, Gujarat, India. *Water Conservation Science and Engineering*, *7*(4), 525-532. https://doi.org/10.1007/s41101-022-00160-1

Wang, Y., & Witten, I. H. (1997). Induction of model trees for predicting continuous classes. In *Proceedings of the European Conference on Machine Learning* (pp. 128–137).

Witten, I.H., Frank, E. (2005). Data mining: practical machine learning tools and techniques (2nd edn.). Morgan Kaufmann, San Francisco.