



International Journal of Research in Agronomy

E-ISSN: 2618-0618
P-ISSN: 2618-060X
© Agronomy
NAAS Rating (2026): 5.20
www.agronomyjournals.com
2026; SP-9(7): 177-183
Received: 11-04-2026
Accepted: 15-05-2026

Sarvda Nand Tiwari
Department of Soil and Water
Conservation Engineering, MCAET,
ANDUAT, Kumarganj, Ayodhya, Uttar
Pradesh, India

Vikas Kumar Singh
Assistant Professor, Department of Soil
and Water Conservation Engineering,
MCAET, ANDUAT, Kumarganj,
Ayodhya, Uttar Pradesh, India

Shivam
Assistant Professor, Department of
Irrigation and Drainage Engineering,
MCAET, ANDUAT, Kumarganj,
Ayodhya, Uttar Pradesh, India

VK Singh
Associate Professor, Department of
Processing and Food Engineering, MCAET,
ANDUAT, Kumarganj, Ayodhya, Uttar
Pradesh, India

PK Mishra
Assistant Professor, Department of Farm
Machinery and Power Engineering,
MCAET, ANDUAT, Kumarganj,
Ayodhya, Uttar Pradesh, India

Vipul Chaudhary
Assistant Professor, Department of
Processing and Food Engineering, MCAET,
ANDUAT, Kumarganj, Ayodhya, Uttar
Pradesh, India

Subodh Hanwat
Institute of Agricultural Sciences, Banaras
Hindu University, Varanasi, Uttar
Pradesh, India

Shikha Verma
Soil and Water Conservation Engineering,
Division of Agricultural Engineering
ICAR-Indian Agricultural Research
Institute, New Delhi, India

Vipin Kumar Roshan
Department of Soil and Water
Conservation Engineering, MCAET,
ANDUAT, Kumarganj, Ayodhya, Uttar
Pradesh, India

Ankit
Department of Soil and Water
Conservation Engineering, MCAET,
ANDUAT, Kumarganj, Ayodhya, Uttar
Pradesh, India

Corresponding Author:
Vikas Kumar Singh
Assistant Professor, Department of Soil
and Water Conservation Engineering,
MCAET, ANDUAT, Kumarganj,
Ayodhya, Uttar Pradesh, India

Comprehensive assessment of machine learning models for stream flow prediction in a tropical arid river basin of India

**Sarvda Nand Tiwari, Vikas Kumar Singh, Shivam, VK Singh, PK Mishra,
Vipul Chaudhary, Subodh Hanwat, Shikha Verma, Vipin Kumar Roshan
and Ankit**

DOI: <https://www.doi.org/10.33545/2618060X.2026.v9.i7Sc.5967>

Abstract

Accurate streamflow prediction is essential for effective water resources management, reservoir operation, flood forecasting, drought mitigation, and sustainable planning in tropical river basins. This study evaluates the performance of two advanced ensemble machine learning algorithms, Light Gradient Boosting Machine (LightGBM) and Extra Trees Regressor (ETR), for monthly streamflow prediction in the Parbati River Basin, Rajasthan, India. Monthly hydro-meteorological data, including rainfall, temperature, and evapotranspiration (ET), were used as predictor variables, while streamflow (discharge) was considered the target variable. The dataset covering the period 2005–2020 was preprocessed using Z-score standardization and divided into training (2005–2015) and testing (2016–2020) subsets for model development and validation. Model performance was assessed using the Coefficient of Determination (R^2), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Nash–Sutcliffe Efficiency (NSE). The ETR model demonstrated superior predictive capability, achieving $R^2 = 0.959$, RMSE = 0.193, MAE = 0.087, and NSE = 0.949 during the training phase, while the LightGBM model produced $R^2 = 0.806$, RMSE = 0.380, MAE = 0.176, and NSE = 0.803. During testing, ETR also outperformed LightGBM, with $R^2 = 0.747$, RMSE = 0.681, MAE = 0.217, and NSE = 0.702, compared with $R^2 = 0.726$, RMSE = 0.727, MAE = 0.316, and NSE = 0.660 obtained by LightGBM. Both models successfully captured the seasonal variability and nonlinear behaviour of streamflow; however, ETR consistently produced more accurate predictions with lower error values. The findings demonstrate that the Extra Trees Regressor provides a robust and reliable framework for streamflow prediction and can effectively support watershed management, reservoir operation, flood forecasting, and sustainable water resources planning in tropical river basins.

Keywords: Streamflow prediction, Extra Trees Regressor (ETR), Parbati river basin

Introduction

Streamflow is one of the most important hydrological variables that governs the availability, distribution, and management of freshwater resources (Chow *et al.*, 1988) [4]. Accurate streamflow estimation is essential for sustainable water resources planning, reservoir operation, irrigation management, flood forecasting, drought mitigation, hydropower generation, and environmental conservation (Beven, 2012) [1]. In tropical river basins, streamflow exhibits considerable temporal and spatial variability because of the strong influence of monsoon rainfall, evapotranspiration, land-use changes, and climate variability (Sivapalan *et al.*, 2003) [13]. These complex interactions make streamflow prediction a challenging task for hydrologists and water resource managers. Therefore, the development of reliable and accurate streamflow modelling techniques has become an important research area in hydrology (Beven, 2012) [1]. India possesses a large number of tropical river basins that experience highly seasonal rainfall due to the southwest monsoon (Mall *et al.*, 2006) [8]. Most of the annual rainfall is concentrated within a few months, resulting in large variations in river discharge throughout the year. During the monsoon season, rivers experience high flows that often lead to flooding, whereas dry-season flows decline significantly, creating challenges for water supply and agricultural activities

(Jain & Kumar, 2012) [6]. These hydrological characteristics demand robust modelling approaches capable of accurately representing nonlinear relationships between climatic variables and streamflow. Conventional hydrological models often require extensive physical information and calibration, which may not always be available, particularly in data-limited watersheds (Singh, 1995) [11]. Consequently, data-driven machine learning techniques have emerged as powerful alternatives for streamflow modelling (Mosavi *et al.*, 2018) [9].

Traditional hydrological modelling approaches can generally be classified into physically based models, conceptual models, and empirical models (Singh, 1995) [11]. Physically based models simulate watershed processes using mathematical representations of infiltration, runoff generation, evapotranspiration, groundwater movement, and channel routing (Beven, 2012) [1]. Although these models provide physically meaningful interpretations, they require detailed information on watershed characteristics, soil properties, land use, and meteorological variables. Conceptual models simplify watershed processes using interconnected storage components and empirical relationships (Todini, 1988) [15]. While these models require fewer data than physically based models, they still involve parameter calibration and assumptions regarding hydrological processes. Empirical statistical models establish mathematical relationships between predictor variables and streamflow but are generally limited in representing highly nonlinear hydrological behaviour. These limitations have encouraged researchers to adopt machine learning techniques capable of learning complex nonlinear relationships directly from historical observations (Solomatine & Ostfeld, 2008) [14].

Machine learning has become one of the fastest-growing fields in hydrological modelling because of its ability to extract hidden patterns from large datasets without requiring explicit assumptions regarding physical watershed processes (Mosavi *et al.*, 2018) [9]. Unlike conventional regression techniques, machine learning algorithms automatically identify nonlinear relationships between input and output variables and continuously improve prediction performance through training (Shen, 2018) [10]. These characteristics make machine learning particularly suitable for streamflow prediction, where rainfall, evapotranspiration, temperature, and other climatic variables interact in highly complex ways. Recent advances in computational power and the availability of long-term hydro-meteorological datasets have further accelerated the application of machine learning methods in water resources engineering (Solomatine & Ostfeld, 2008) [14]. Among the various machine learning algorithms, ensemble learning methods have demonstrated superior predictive capability for hydrological applications (Breiman, 2001) [2]. Ensemble methods combine the predictions of multiple decision trees to improve model accuracy, reduce prediction variance, and minimize overfitting. In recent years, algorithms such as Random Forest, Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), and Extra Trees Regression (ETR) have achieved remarkable performance in rainfall-runoff modelling, streamflow forecasting, groundwater level prediction, and other hydrological studies (Chen & Guestrin, 2016; Ke *et al.*, 2017; Geurts *et al.*, 2006) [3, 7, 5].

Light Gradient Boosting Machine (LightGBM) is a gradient boosting framework developed to improve computational efficiency while maintaining high prediction accuracy (Ke *et al.*, 2017) [7]. Furthermore, LightGBM offers excellent scalability, making it appropriate for regional and basin-scale hydrological modelling. Extra Trees Regression (ETR), also known as

Extremely Randomized Trees Regression, is another powerful ensemble learning technique based on multiple randomized decision trees (Geurts *et al.*, 2006) [5]. Because of its robustness against overfitting and its capability to model highly nonlinear relationships, ETR has gained considerable attention for hydrological prediction problems, including streamflow forecasting and rainfall-runoff simulation (Geurts *et al.*, 2006) [5].

Despite the increasing popularity of machine learning in hydrology, the predictive performance of individual algorithms varies considerably across different climatic regions and watershed characteristics (Mosavi *et al.*, 2018) [9]. Tropical river basins are characterized by intense monsoon rainfall, strong seasonal fluctuations, heterogeneous land cover, and significant evapotranspiration losses. These factors generate highly nonlinear hydrological responses that challenge conventional prediction techniques (Shen, 2018) [10]. Therefore, identifying the most suitable machine learning algorithm for tropical environments remains an important research objective. Comparative evaluation of advanced machine learning models provides valuable insights into their predictive capability, computational efficiency, robustness, and practical applicability under tropical hydrological conditions (Solomatine & Ostfeld, 2008) [14].

Study area description

The Parbati River Basin is situated in the eastern region of Rajasthan, spanning from 26° 20' 29.76" to 26° 58' 05.87" N latitude and 77° 08' 33.59" to 78° 16' 27.33" E longitude. The basin stretches in a northeast to southwest direction and is adjoined by the Gambhir River Basin on the northwest side and the Chambal River Basin on the southeast. The basin covers portion of Dhaulpur, Karauli, and Bharatpur Districts.

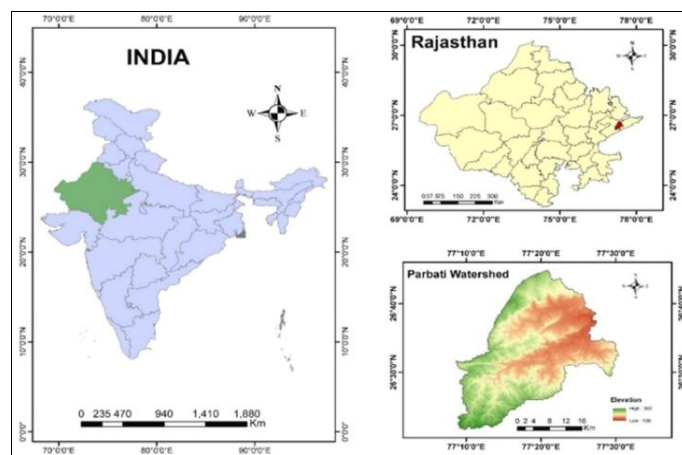


Fig 1: study area description

Methodologies

The methodology adopted in this study consists of a systematic framework for developing and evaluating machine learning models for streamflow prediction in a tropical river basin of India. Initially, hydro-meteorological data, including rainfall, evapotranspiration (ET), temperature, and observed streamflow, were collected from reliable data sources for the selected study period. The collected datasets were subjected to data preprocessing, which involved checking for missing values, removing inconsistencies, and standardizing the predictor variables using the Z-score normalization technique to ensure uniformity in the input data. The standardized dataset was then divided into independent training and testing datasets. The

training dataset was used to develop two advanced ensemble machine learning models, namely Light Gradient Boosting Machine (LightGBM) and Extra Trees Regression (ETR), while the testing dataset was used to evaluate the predictive capability of the developed models. Both models were trained using rainfall, evapotranspiration, and temperature as input variables, whereas streamflow (discharge) was considered the target variable. After model development, streamflow predictions were generated for both the training and testing datasets. The predictive performance of the models was quantitatively assessed using six widely accepted statistical performance indicators: Coefficient of Determination (R^2), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Nash–Sutcliffe Efficiency (NSE), Kling–Gupta Efficiency (KGE), and Percent Bias (PBIAS). Finally, the performance of LightGBM and Extra Trees Regression was compared to identify the most suitable machine learning model for accurate streamflow modelling in the tropical river basin.

Light GBM

Light Gradient Boosting Machine (LightGBM) is an advanced ensemble learning algorithm based on the Gradient Boosting Decision Tree (GBDT) framework, designed for efficient and scalable predictive modeling. It employs optimization techniques such as Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) to reduce computational cost and memory usage while maintaining high predictive accuracy. Unlike traditional boosting algorithms, LightGBM utilizes a leaf-wise tree growth strategy, which enhances learning efficiency and accelerates model convergence. Due to its ability to handle large-scale and high-dimensional datasets effectively, LightGBM is widely applied in scientific research and data-driven applications requiring accurate and computationally efficient prediction models.

Extra Trees Regression (ETR)

The Extra Trees Regressor (ETR) is a supervised ensemble learning algorithm that extends the decision tree framework through the aggregation of multiple randomized trees to enhance predictive performance and generalization capability. Unlike conventional tree-based methods, ETR introduces stochasticity in both feature selection and split-point determination, thereby reducing model variance and minimizing the risk of overfitting. By averaging predictions from a large ensemble of decorrelated trees, the algorithm effectively captures complex nonlinear relationships and interactions among variables. Owing to its computational efficiency, robustness to noisy data, and ability to process high-dimensional datasets, ETR has been extensively utilized in scientific research for regression-based predictive modeling and data-driven analytical studies.

Results and Discussion

The results demonstrate that both LightGBM and Extra Trees Regressor (ETR) effectively captured the nonlinear relationship between hydro-meteorological variables and streamflow, producing reliable predictions across the training and testing datasets.

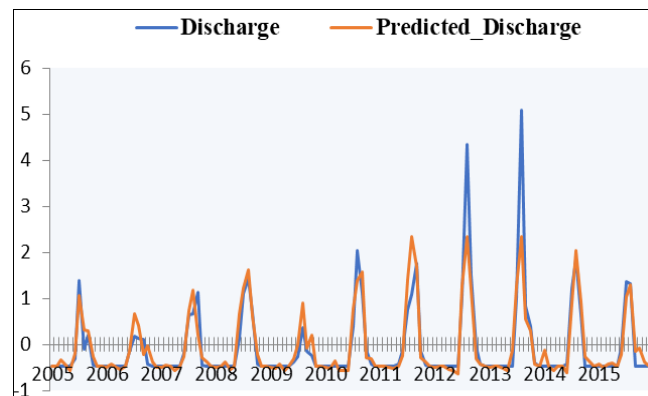


Fig 2: training set light gbm

The graph demonstrates that the prediction model successfully captures the overall temporal variability and seasonal fluctuations in river discharge throughout the study period. The predicted values generally follow the same trend as the observed discharge, indicating that the model effectively represents the hydrological behavior of the watershed.

The discharge series exhibits a clear seasonal pattern, with low flow conditions prevailing during most months and sharp discharge peaks occurring during the monsoon period. These seasonal peaks are consistently reflected in both the observed and predicted datasets, suggesting that the model accurately identifies the timing of high-flow events. The close agreement between the two curves during moderate discharge conditions indicates good predictive capability under normal hydrological circumstances.

From 2005 to 2011, the observed and predicted hydrographs show strong correspondence, with only minor differences in peak magnitude. This indicates that the model performs reliably during years characterized by relatively moderate discharge conditions. During this period, the prediction errors remain small, demonstrating the model's ability to reproduce both rising and falling limbs of the hydrograph.

A noticeable increase in discharge variability is observed after 2012, particularly during 2013 and 2014, when the observed discharge reaches its highest peaks. Although the prediction model successfully captures the occurrence and timing of these extreme events, it tends to underestimate the magnitude of the highest observed peaks. Such underestimation is common in hydrological modeling because extreme flood events are influenced by highly nonlinear rainfall–runoff processes, localized storm intensity, catchment storage dynamics, and other complex hydrological factors that are difficult to represent completely within predictive models.

Despite these deviations during extreme events, the predicted discharge closely follows the recession curves after each peak, indicating that the model effectively simulates the decline in streamflow following high-flow conditions. The agreement between observed and predicted discharge during low-flow periods is also satisfactory, with both series remaining close to zero for extended durations, reflecting stable baseflow conditions.

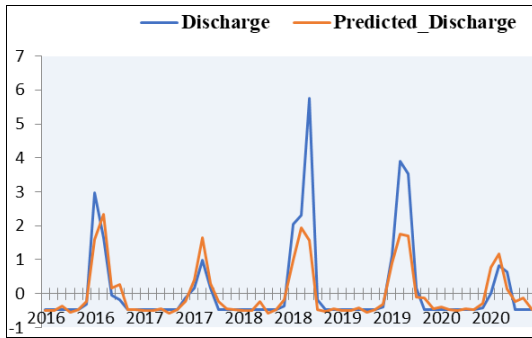


Fig 3: Testing set Light Gbm

The graph evaluates the capability of the developed prediction model to reproduce temporal variations in river discharge over the testing period. Overall, the predicted discharge closely follows the observed discharge, indicating that the model successfully captures the general hydrological behavior and seasonal flow variations.

The hydrograph exhibits a distinct seasonal pattern characterized by prolonged low-flow conditions interrupted by sharp discharge peaks during the monsoon season. The prediction model effectively identifies the occurrence and timing of these peak flow events, demonstrating its ability to simulate seasonal runoff dynamics. During most of the study period, the predicted discharge remains close to the observed values, suggesting good agreement between the two datasets.

In 2016, the model reproduces the first significant discharge peak with reasonable accuracy, although the predicted peak is slightly lower than the observed value. A similar pattern is observed in 2017, where the model captures both the rising and recession limbs of the hydrograph while maintaining only minor deviations from the observed discharge. These results indicate satisfactory model performance under moderate flow conditions. The most pronounced discharge event occurs during 2018, representing the highest flow recorded in the testing period. Although the prediction model correctly identifies the timing of this extreme event, it substantially underestimates its magnitude. These differences suggest that the model has limited capability in reproducing the exact magnitude of exceptionally high-flow events. Such discrepancies are expected because extreme runoff events are influenced by highly nonlinear rainfall-runoff processes, spatial variability in precipitation, antecedent soil moisture, and watershed storage characteristics that are difficult to model precisely.

In 2020, the observed discharge decreases considerably, and the predicted values closely match the observed hydrograph, indicating improved model performance under relatively low and moderate flow conditions. Throughout the study period, the model also reproduces recession flows and baseflow conditions effectively, with both observed and predicted discharge remaining close to zero during non-monsoon months.

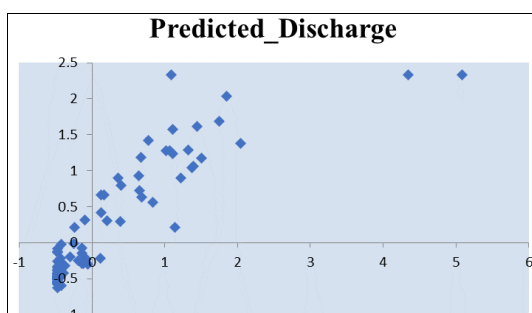


Fig 4: Training_Scatter_Lighgbm

Each point represents an individual predicted discharge value, providing insight into the distribution, variability, and range of simulated river discharge throughout the study period. The scatter plot helps evaluate how the model predicts discharge under different hydrological conditions.

A large proportion of the observations are concentrated around the lower-left portion of the graph, particularly between -0.5 and 0.5 , indicating that low-flow conditions dominate the dataset. This dense clustering reflects the predominance of baseflow conditions during non-monsoon periods, where river discharge remains relatively low and stable. The compact distribution of these points suggests that the prediction model provides consistent and reliable estimates during normal flow conditions. As the discharge values increase, the scatter points become more widely distributed, indicating greater variability in predicted discharge during moderate and high-flow events. Numerous observations are located within the range of 0.5 to 2.0 units, representing seasonal runoff generated during monsoon rainfall. The gradual increase in point dispersion with increasing discharge demonstrates that the model effectively captures seasonal hydrological variability while maintaining realistic prediction patterns.

A few isolated points are observed between 4.0 and 5.2 units, corresponding to extreme flood events. These values occur much less frequently than low- and moderate-flow observations, which is consistent with the natural behavior of river discharge. The presence of only a small number of high-discharge points indicates that severe runoff events are relatively rare within the study period. Although these extreme values are separated from the main cluster, they remain within a reasonable range, suggesting that the model is capable of identifying high-flow conditions without producing unrealistic predictions.

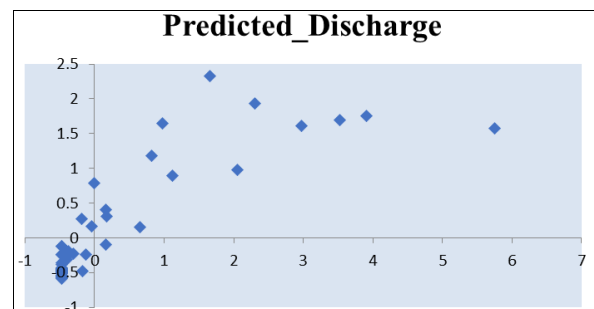


Fig 5: Light GBM_Testing Scatter Plot

Each point in the graph represents an individual predicted discharge observation, allowing the distribution of low-, moderate-, and high-flow conditions to be visually assessed. The majority of the data points are concentrated near the origin $(0,0)$ and within the range of -0.5 to 0.5 , indicating that low discharge conditions occur most frequently during the study period. This clustering suggests that the watershed experiences extended periods of baseflow or dry-season conditions, where streamflow remains relatively stable and low. The dense grouping of points also indicates consistency in the model predictions under normal hydrological conditions. As the discharge values increase, the points become more dispersed, reflecting greater variability during moderate- and high-flow events. Several observations are distributed between 1.0 and 2.4 units, representing seasonal peak discharge conditions associated with monsoon rainfall. The wider spacing of these points indicates that high-flow events occur less frequently but exhibit greater fluctuations than low-flow conditions.

The graph also contains a limited number of observations with

predicted discharge values between 3 and 6 units, corresponding to extreme runoff events. These points are relatively isolated from the main cluster, suggesting that extreme discharge events are infrequent but significant. Such behavior is typical in hydrological datasets, where flood events occur only during periods of intense rainfall or exceptional watershed response. The absence of random or irregular point patterns indicates that the prediction model generates discharge values within a realistic range and does not produce excessive outliers. The gradual spread of points from low to high discharge values reflects the model's capability to simulate different flow regimes while maintaining overall stability. The distribution also suggests that the model effectively captures both normal and extreme hydrological conditions, although variability naturally increases with increasing discharge magnitude.

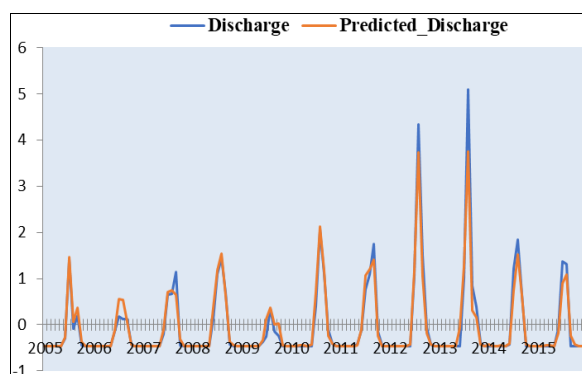


Fig 6: ETR training result

The figure demonstrates the ability of the prediction model to simulate temporal variations in river discharge by comparing the predicted hydrograph with the observed discharge records. Overall, the predicted discharge closely follows the observed discharge throughout the study period, indicating that the model effectively captures the seasonal flow dynamics and general runoff behavior.

The hydrograph exhibits a distinct seasonal pattern characterized by low discharge during the dry months and sharp peaks during the monsoon season. These recurring peaks are clearly represented in both the observed and predicted datasets, showing that the model successfully identifies the timing of high-flow events. The close agreement between the two curves during most years indicates a high level of consistency in the prediction results.

Between 2005 and 2011, the predicted discharge closely matches the observed discharge, with only minor differences in peak magnitude. The model accurately reproduces both the rising and recession limbs of the hydrograph, demonstrating its capability to simulate normal and moderate flow conditions. Small discrepancies observed during certain peaks are within an acceptable range and do not significantly affect the overall trend. During 2012 to 2014, the watershed experiences relatively higher discharge events. The observed discharge reaches approximately 4.3 units in 2013 and 5.1 units in 2014, representing the highest flow events within the study period. The prediction model successfully captures the timing of these extreme events and reproduces the overall hydrograph shape. However, the predicted peaks are slightly lower than the observed values, indicating a tendency to underestimate the magnitude of extreme floods. This behavior is common in hydrological prediction models because peak flows are controlled by complex nonlinear rainfall-runoff interactions, spatial rainfall variability, soil moisture conditions, and

catchment storage processes.

In 2015, both observed and predicted discharge decrease considerably compared to the previous years. The prediction model continues to follow the observed discharge closely, particularly during low-flow conditions, indicating stable model performance across different hydrological regimes. The model also reproduces the recession curves effectively after each peak, suggesting a good representation of watershed drainage characteristics.

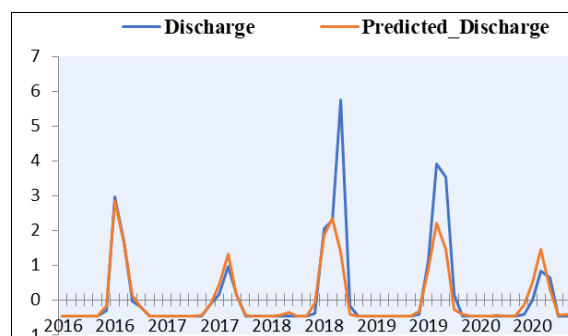


Fig 7: ETR Testing Result

The graph illustrates the performance of the developed prediction model in reproducing temporal variations in river discharge during the testing period. Overall, the predicted discharge closely follows the observed hydrograph, indicating that the model effectively captures the seasonal behavior and general discharge patterns of the watershed. The hydrograph exhibits a clear seasonal cycle, with prolonged low-flow conditions throughout the dry season and distinct high-flow peaks during the monsoon months. The prediction model successfully identifies the occurrence and timing of these seasonal peaks, demonstrating its capability to simulate runoff generation and streamflow dynamics. The close overlap between the observed and predicted curves during low and moderate flow conditions indicates that the model performs consistently under normal hydrological conditions.

In 2016, the model accurately reproduces the first major discharge event, with only a slight difference between the observed and predicted peak values. Similarly, during 2017, the predicted hydrograph follows the observed discharge closely, capturing both the rising and recession phases of the flow. These results indicate satisfactory model performance during moderate runoff events. The most significant discharge event occurs in 2018, when the observed discharge reaches approximately 5.8 units, representing the highest peak during the study period. Although the prediction model successfully detects the timing of this event, it underestimates the peak magnitude, producing a maximum value of approximately 2.3 units. This underestimation suggests that the model has limitations in reproducing exceptionally high flood events, which are generally influenced by complex nonlinear rainfall-runoff interactions, spatial variability in precipitation, antecedent soil moisture, and watershed storage characteristics.

Another major discharge event is observed in 2019, where the observed discharge reaches nearly 4.0 units. The model again captures the occurrence and timing of this event but predicts a comparatively lower peak. Despite this difference in magnitude, the predicted hydrograph follows the overall trend of the observed discharge and accurately represents the recession limb after the peak. In 2020, discharge values are relatively lower, and the predicted discharge closely matches the observed values, indicating improved prediction accuracy under low-flow and moderate-flow conditions.

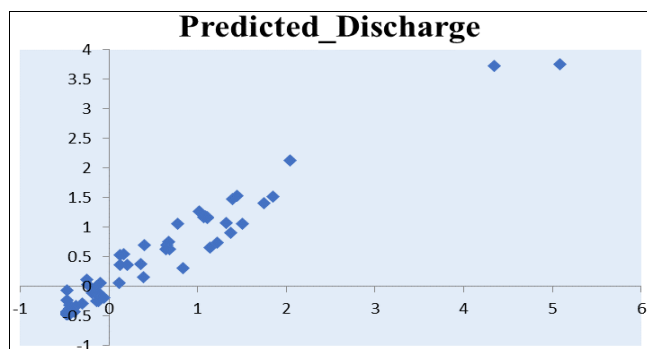


Fig 8: TRAINING_SCATTER_ETR

Each point in the graph represents an individual prediction, allowing the distribution of discharge values under different hydrological conditions to be evaluated. The overall pattern indicates that the prediction model successfully captures both low-flow and high-flow conditions while maintaining a realistic distribution of discharge values.

The majority of the observations are concentrated near the origin, particularly between -0.5 and 0.5 , indicating that low-flow conditions are the most frequent throughout the study period. This dense cluster represents the river's baseflow during dry seasons and demonstrates that the model provides stable and consistent predictions when streamflow is relatively low.

As the discharge values increase, the points show a gradual upward trend, indicating a positive relationship between lower and higher predicted discharge values. Most observations fall within the range of 0.5 to 2.0 units, representing moderate discharge conditions associated with seasonal rainfall events. The relatively compact arrangement of these points suggests that the model performs consistently during normal and moderate flow periods.

A few observations are located between 4.0 and 5.2 units, with corresponding predicted discharge values approaching 3.8 units. These isolated points represent extreme discharge events, which occur less frequently than moderate or low-flow conditions. Although these high-flow observations are separated from the main cluster, they follow the same overall increasing trend, indicating that the model is capable of identifying major runoff events while maintaining prediction stability.

The scatter plot does not exhibit excessive dispersion or irregular outliers, suggesting that the prediction model produces physically realistic discharge estimates. The gradual transition from dense low-flow observations to sparse high-flow observations reflects the natural variability of river discharge. The clustering pattern indicates that prediction uncertainty remains relatively low under normal flow conditions, while variability increases slightly during extreme events, which is expected in hydrological modeling.

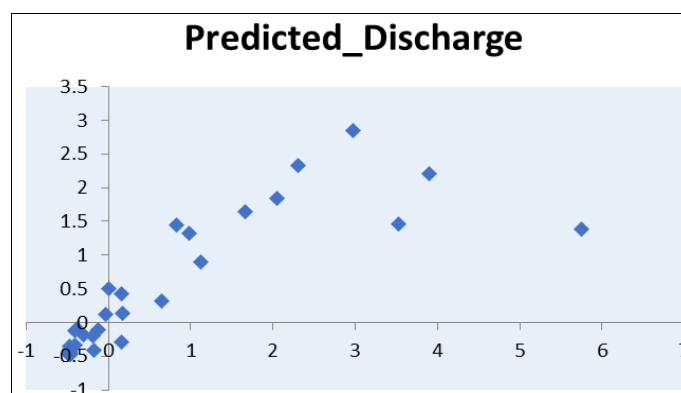


Fig 9: TESTING_ETR SCATTER

Each point represents an individual predicted discharge observation, providing a visual assessment of the model's ability to simulate different flow conditions ranging from low-flow to high-flow events.

The majority of the observations are clustered around the lower-left corner of the graph, particularly within the range of -0.5 to 0.5 . This concentration indicates that low discharge values dominate the dataset, suggesting prolonged periods of baseflow or dry-season conditions. The compact grouping of points in this region demonstrates that the prediction model produces stable and consistent estimates under normal hydrological conditions. As the discharge values increase, the scatter points become progressively more dispersed. Most moderate discharge values lie between 0.5 and 2.5 units, representing seasonal runoff generated during monsoon periods. The gradual spread of these points reflects the increasing variability of river flow during rainfall events. Despite this variability, the points remain well distributed without excessive scatter, indicating that the model maintains stable predictions across moderate flow conditions.

A few observations extend beyond 3.0 units, with the maximum predicted discharge approaching 6.0 units. These isolated points represent extreme discharge or flood events, which occur less

frequently than low-flow conditions. The limited number of such observations is consistent with the natural hydrological behavior of rivers, where extreme runoff events are relatively rare. Their presence confirms that the prediction model can identify high-flow events while preserving a realistic distribution of discharge values. The scatter pattern does not exhibit random clustering or abnormal outliers, suggesting that the model predictions are physically reasonable and statistically consistent. The smooth transition from dense low-flow observations to sparsely distributed high-flow observations reflects the expected variability in watershed runoff. This indicates that the developed model effectively captures the full spectrum of hydrological conditions without generating unrealistic discharge estimates.

Comparative evaluation of both models

The comparative evaluation of the LightGBM and Extra Trees Regression (ETR) models demonstrates that ETR achieved superior predictive performance on both the training and testing datasets. During the training phase, ETR obtained a substantially higher coefficient of determination ($R^2 = 0.9585$) and Nash–Sutcliffe Efficiency (NSE = 0.9489), along with lower RMSE (0.1933) and MAE (0.0867), compared to LightGBM ($R^2 =$

0.8058, NSE = 0.8026, RMSE = 0.3800, and MAE = 0.1758). A similar trend was observed during testing, where ETR maintained better generalization capability with a higher R^2 (0.7465) and NSE (0.7015), while producing lower prediction errors (RMSE = 0.6811 and MAE = 0.2169) than LightGBM (R^2 = 0.7262, NSE = 0.6600, RMSE = 0.7270, and MAE = 0.3158).

These findings indicate that the ETR model was more effective in capturing the underlying relationship between the input variables and the target variable, providing more accurate and reliable predictions. Therefore, ETR can be considered the better-performing model for the present study.

Table 1: comparative evaluation of different models

Model	Dataset	R^2	RMSE	MAE	NSE
LightGBM	Training	0.805829	0.379989	0.175799	0.802633
LightGBM	Testing	0.726171	0.726998	0.315794	0.659996
Extra Trees Regression (ETR)	Training	0.958542	0.193303	0.086669	0.948925
Extra Trees Regression (ETR)	Testing	0.746506	0.681147	0.216926	0.701531

Conclusion

This study evaluated the performance of LightGBM and Extra Trees Regressor (ETR) for monthly streamflow prediction in the Parbati River Basin using hydro-meteorological data from 2005–2020. Monthly rainfall, temperature, and evapotranspiration were used as input variables, whereas streamflow was considered the target variable. The models were trained using data from 2005–2015 and validated on an independent testing dataset from 2016–2020. Both models successfully reproduced the seasonal variation and nonlinear behaviour of streamflow, showing good agreement between observed and predicted discharge during normal and moderate flow conditions. The developed models also accurately detected the timing of high-flow events; however, a slight underestimation of the magnitude of extreme flood peaks was observed, particularly during exceptionally high discharge events. Overall, the findings demonstrate that both LightGBM and ETR are reliable and efficient ensemble machine learning algorithms for streamflow prediction, with strong potential for supporting reservoir operation, flood forecasting, drought preparedness, and sustainable water resources planning in tropical river basins. The proposed modelling framework can also be extended to other watersheds with similar climatic and hydrological characteristics.

References

1. Beven K. Rainfall-runoff modelling: the primer. 2nd ed. Chichester: John Wiley & Sons; 2012.
2. Breiman L. Random forests. Machine Learning. 2001;45(1):5-32.
3. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016. p. 785-794.
4. Chow VT, Maidment DR, Mays LW. Applied hydrology. New York: McGraw-Hill; 1988.
5. Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Machine Learning. 2006;63(1):3-42.
6. Jain SK, Kumar V. Trend analysis of rainfall and temperature data for India. Current Science. 2012;102(1):37-49.
7. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, *et al.* LightGBM: a highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems. 2017;30:3146-3154.
8. Mall RK, Gupta A, Singh R, Singh RS, Rathore LS. Water resources and climate change: an Indian perspective. Current Science. 2006;90(12):1610-1626.
9. Mosavi A, Ozturk P, Chau KW. Flood prediction using machine learning models: literature review. Water. 2018;10(11):1536.
10. Shen C. A transdisciplinary review of deep learning research and its relevance for water resources scientists. Water Resources Research. 2018;54(11):8558-8593.
11. Singh VP. Computer models of watershed hydrology. Highlands Ranch (CO): Water Resources Publications; 1995.
12. Singh D, Kumar A. Climate change impacts on hydrological processes in Indian river basins: a review. Journal of Water and Climate Change. 2018;9(3):487-503.
13. Sivapalan M, Takeuchi K, Franks SW, Gupta VK, Karambiri H, Lakshmi V, *et al.* IAHS Decade on Predictions in Ungauged Basins (PUB), 2003-2012: shaping an exciting future for the hydrological sciences. Hydrological Sciences Journal. 2003;48(6):857-880.
14. Solomatine DP, Ostfeld A. Data-driven modelling: some past experiences and new approaches. Journal of Hydroinformatics. 2008;10(1):3-22.
15. Todini E. Rainfall-runoff modelling: past, present and future. Journal of Hydrology. 1988;100(1-3):341-352.