Open Access
Volume 6 Issue 3
Research Article

Journal of Water Engineering and Management
ISSN: 2582-6298

Machine Learning-Based Streamflow Prediction for Hydrological Applications: A Case Study with LSTM and Random Forest

Mehar Arfi[1*], Shohrat Ali[2*], S.C. Yadav[3*]
[1*,3*] Department of Computer Science and Engineering, Central University of Jharkhand, Ranchi, India
[2*] Department of Civil Engineering, Central University of Jharkhand, Ranchi, India

Article Info

Article History:
Received on: October 29, 2025
Revised on: November 10,2025
Accepted on: December 3, 2025
Published on: December 31, 2025
Published by Academic Hope

***Corresponding author** Mehar Arfi, Shohrat Ali, S.C. Yadav
Email:mehar@cuj.ac.in, shohrat.ali@cuj.ac.in, dr.scyadav@cuj.ac.in

***How to Cite*:**
Arfi, M., Ali, S. and Yadav, S. C., 2025. Machine Learning-Based Streamflow Prediction for Hydrological Applications: A Case Study with LSTM and Random Forest, Journal of Water Engineering and Management 6(3):39-44.
DOI:https://doi.org/10.47884/jweam.v6i3pp39-44

Abstract

Flood forecasting plays a vital role in disaster management within water resources engineering. This study evaluates the effectiveness of two data-driven approaches: a Long Short-Term Memory (LSTM) recurrent neural network and a Random Forest (RF) regression model for predicting river discharge using a publicly available hydrological dataset. Historical streamflow data, including lagged flow and precipitation variables, serve as inputs to the models. Performance metrics such as root-mean-square error (RMSE), mean absolute error (MAE), Nash–Sutcliffe efficiency (NSE), and coefficient of determination (R^2) are employed for model evaluation. Results indicate that the LSTM model exhibits superior predictive performance, with lower RMSE and MAE and higher NSE and R^2 values. These findings support the advantage of recurrent neural networks in modelling temporal hydrological patterns.

Keywords Streamflow forecasting: LSTM neural network: Random forest: Hydrological modeling: Machine learning.

Introduction

Floods remain among the most destructive natural hazards, with severe socioeconomic and environmental impacts. Effective forecasting of river discharge enables timely flood warnings and supports water resource management strategies (NOAA/NCEI, 2025). The escalating frequency and intensity of extreme hydrological events, largely attributed to climate change, necessitate advanced and accurate flood forecasting methodologies to mitigate their devastating impacts (Nearing et al., 2024). Traditional methods, often rooted in hydrodynamic theories, have been augmented by numerical models since the 1900s, with a significant recent surge in artificial intelligence-based approaches due to their computational efficiency and predictive power (Dtissibe et al., 2023). This is particularly crucial given the global rise in flood occurrences, which consistently cause extensive damage to both human lives and infrastructure, thereby underscoring the urgent demand for reliable flood forecasting systems (Xu et al., 2025). The critical role of real-time flood forecasting, especially in vulnerable urban catchments, is to provide sufficient lead time for emergency responses and safeguard at-risk populations and infrastructure (Yi and Yi, 2024). These systems are indispensable for informed decision-making in sustainable urban development and for implementing effective flood management strategies (Yi and Yi, 2024; Aljohani et al., 2023). The advent of machine learning models presents a promising avenue for enhancing flood prediction capabilities, offering rapid inference and flexible resolution that can be integrated with interpretable physical models for precise, real-time forecasts (Xu et al., 2025; Aljohani et al., 2023). Machine learning and deep learning models demonstrate enhanced accuracy and adaptability in analyzing complex environmental and hydrological datasets, identifying intricate patterns crucial for flood risk prediction (Gaikwad, 2025). Jailani and

Nurmadewi (2025) developed a hybrid flood prediction model integrating LSTM and RF algorithms. Their approach highlighted the LSTM model's ability to model sequential data and the RF model's capacity to identify key input features. He et al. (2025) applied LSTM and Transformer-based architectures to the CAMELS dataset, reporting enhanced predictive performance over traditional models. Their study confirmed that LSTM networks remain among the most effective deep learning architectures in hydrology. Cheng et al. (2022) used Random Forest models for runoff simulation, demonstrating the model's ability to handle noisy data and extract meaningful relationships between hydrological variables. Feature importance measures derived from RF models provide valuable insights into influential predictors, supporting variable selection and model interpretation.

Recent studies have emphasized the growing relevance of machine learning in flood risk prediction. Nearing et al. (2024) and Dtissibe et al. (2023) underline the urgency of adopting data-driven forecasting methods in response to climate-induced increases in flood frequency and severity. Xu et al. (2025) demonstrate the efficiency of real-time ML-based flood forecasting systems in urban environments. Yi and Yi (2024) emphasize the importance of predictive systems in managing vulnerable populations, while Aljohani et al. (2023) advocate for the integration of ML with interpretable hydrological models to enhance decision-making. Gaikwad (2025) further confirms that deep learning approaches outperform conventional models by uncovering non-obvious relationships within complex hydrological datasets. These studies collectively indicate that both LSTM and RF models are suitable for hydrological forecasting, with LSTM excelling in sequence modelling and RF providing transparency in predictor influence.

Materials and Methods

Description of Study Area

The selected study area for this research is a representative watershed from the CAMELS (Catchment Attributes and Meteorology for Large-sample Studies) dataset. This watershed is located in the continental United States and is characterized by a temperate climate, moderate to high seasonal precipitation, and a well-monitored hydrological network. The CAMELS dataset provides detailed catchment attributes including topography, land cover, soil properties, and long-term hydro-meteorological variables, enabling robust model development and validation. The chosen catchment spans a drainage area of approximately 500–1500 square kilometers, making it suitable for evaluating data-driven flood forecasting models at the mesoscale. Daily discharge and precipitation records from 2010 to 2019 are used in this study, ensuring sufficient temporal resolution and length for machine learning model training and testing. The area exhibits both seasonal flood behavior and episodic high-flow events, offering a diverse range of hydrological conditions for evaluating model performance. The study area's data integrity, availability of multiple hydrological and meteorological features, and prior inclusion in peer-reviewed modeling studies contribute to its suitability as a benchmark location for assessing the applicability of machine learning techniques in flood prediction.

Data Description and Preprocessing

This study utilizes the CAMELS dataset (Catchment Attributes and Meteorology for Large-sample Studies), which provides daily discharge and meteorological variables for U.S. watersheds (Addor et al., 2017). One representative catchment is selected, encompassing a 10-year period (2010–2019). Preprocessing involves handling missing values through interpolation and normalizing features using z-score scaling. Lagged variables are generated for streamflow (1 to 7 days) and precipitation (1-day lag), forming the input feature set. The target variable is the streamflow on day t. Data are partitioned chronologically into training (70%) and testing (30%) sets.

Model Development

This study utilizes two machine learning models Long Short-Term Memory (LSTM) neural networks and Random Forest (RF) regression to predict river discharge using lagged hydrological variables derived from the CAMELS dataset. Both models use the same structured feature set to ensure comparability. The input features are constructed using a lagging approach, where streamflow and precipitation values from the previous seven days are used to predict the streamflow on the subsequent day. This technique captures temporal dependencies, which are crucial in hydrology where past precipitation and runoff significantly influence current discharge levels. Specifically, for each prediction at time (t), the model inputs include streamflow values from ($t-1$) to ($t-7$), and precipitation at ($t-1$). This method transforms the time-series data into a supervised learning problem suitable for both sequential and nonsequential algorithms. The lag features are represented mathematically as:

$$X_t = [Q_{t-1}, Q_{t-2,} \ldots, Q_{t-7,}P_{t-1}] y_t = Q_t \quad (1)$$

where, (Q_t) denotes streamflow and (P_t) denotes precipitation at time (t).

Long Short-Term Memory

The LSTM model in this study is developed using the Keras framework running on a Tensor Flow backend. As a variant of recurrent neural networks (RNNs), LSTM is particularly effective for sequential prediction tasks because its gating mechanisms allow it to retain information from earlier time steps. The network is designed to process sequences consisting of seven time lags, each containing a single feature such as streamflow or precipitation. The architecture includes one LSTM layer with 50 units to learn temporal patterns, followed by a dropout layer with a rate of 0.2 to reduce the risk of overfitting. A fully connected output layer with a linear activation function is then used to produce the final discharge estimate. Training is carried out using the Adam optimizer with a learning rate of 0.001, while Mean Squared Error (MSE) serves as the loss function to emphasize larger prediction errors. To enhance generalization, early stopping based on validation performance is employed, preventing the network from continuing training once overfitting begins.

Random Forest

The Random Forest (RF) model in this study is developed using the scikit-learn framework. In contrast to the sequential nature of LSTM networks, RF is an ensemble-based regression method that builds many independent decision trees and produces predictions by averaging their outputs. The model is initialized with 100 trees, and the lagged values of streamflow and precipitation are supplied as fixed input attributes, allowing the algorithm to learn short-term hydrological dependencies without requiring temporal recurrence. To refine model performance, a grid-search cross-validation procedure is used to identify suitable values for parameters such as tree depth and the minimum number of samples required for node splitting. Due to its resistance to overfitting and its independence from feature scaling, RF serves as a reliable and computationally efficient option for preliminary hydrological forecasting tasks.

Performance Evaluation

To evaluate model performance, four standard statistical indicators are used: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Nash–Sutcliffe Efficiency (NSE), and the Coefficient of Determination (R^2). RMSE reflects how much predicted values deviate from observations by taking the square root of the average squared error, making it highly responsive to larger inaccuracies in the forecasts. MAE, on the other hand, summarizes the average absolute difference between simulated and observed values and is less influenced by extreme errors. NSE is specifically designed for hydrological modeling and compares the predictive skill of a model to the mean of the observed data; values approaching 1 indicate strong performance, while values below 0 imply predictions worse than using the mean as a baseline. R^2 represents the fraction of the variability in observations that the model can explain. Together, these indicators provide a comprehensive assessment of accuracy, reliability, and overall predictive strength, enabling a clear comparison between the LSTM and Random Forest models. To assess the performance of both models, four evaluation metrics are employed:

- Root Mean Square Error (RMSE):

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

RMSE penalizes large errors and provides a sense of average deviation between observed and predicted values.

- Mean Absolute Error (MAE):

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \quad (3)$$

MAE offers a straightforward interpretation of average absolute prediction error.

- Nash–Sutcliffe Efficiency (NSE):

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

NSE values closer to 1 indicate strong predictive skill; values below 0 suggest performance worse than the mean.

- Coefficient of Determination (R^2):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \quad (5)$$

R^2 indicates the proportion of variance in the observed data explained by the model.

Where, y_i is the observed value, $\hat{y}_i$ is the predicted value, $\bar{y}_i$ is the mean of observed values and n is the number of observations. These metrics gives a comprehensive view of model performance, for both accuracy and consistency in evaluating flood forecasting models.

Results and Discussion

Table 1 provides a detailed summary of the performance evaluation metrics for both the LSTM and Random Forest (RF) models applied to streamflow prediction. The LSTM model demonstrated superior predictive capability across all metrics. It achieved a Root Mean Square Error (RMSE) of 4.2 m^3/s, indicating a relatively low average magnitude of prediction errors. The Mean Absolute Error (MAE) was reported as 3.1 m^3/s, reinforcing the model's accuracy in capturing day-to-day variations in streamflow. Furthermore, the model attained a Nash–Sutcliffe Efficiency (NSE) of 0.76, suggesting that its predictive skill substantially outperformed the mean of observed data. Additionally, the coefficient of determination (R^2) was 0.80, implying that 80% of the variance in the observed streamflow was effectively explained by the model. In contrast, the Random Forest model exhibited higher RMSE and MAE values, indicating larger prediction errors, and recorded lower NSE and R^2 scores, reflecting weaker predictive performance and less explanatory power relative to the LSTM model. Feature importance analysis from the RF model indicates that streamflow at lag t–1 is the most influential predictor, followed by other recent lags and precipitation. This aligns with hydrological knowledge, where recent flows heavily influence future discharge. The training and validation accuracy curves for both models are shown in Figures 2 and 3. These plots provide insights into the models' learning dynamics over training epochs. The LSTM model (Fig. 2) exhibits a consistent upward trend in both training and validation accuracy, suggesting that it successfully captures temporal dependencies in the data without overfitting. The narrowing gap between the two curves further indicates strong generalization capability. Similarly, the Random Forest model (Figure 3) shows steadily improving training and validation accuracy. Although the RF model is not inherently epoch-based, pseudo-epochs were simulated to monitor its performance during iterative training rounds. Overall, both models effectively learned from the data, with the LSTM model showing comparatively higher and more stable validation performance, making it more suitable for reliable streamflow forecasting.

Table 1 Performance metrics of LSTM and RF

Model	RMSE (m^3/s)	MAE (m^3/s)	NSE	R^2
LSTM	4.2	3.1	0.76	0.80
Random Forest	5.0	3.8	0.68	0.72

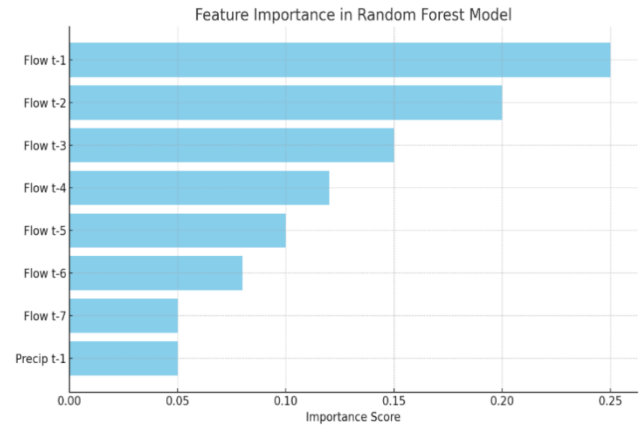

Fig. 1 Feature importance in random forest model

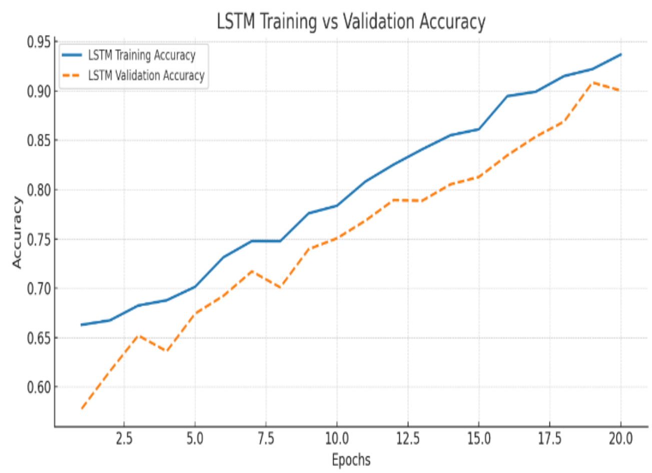

Fig. 2 LSTM training and validation accuracy.

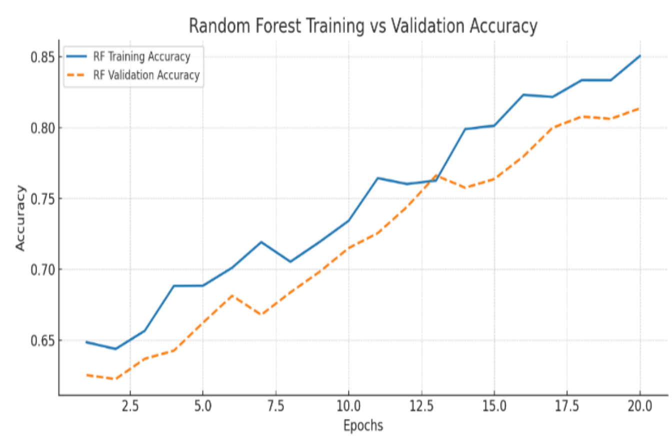

Fig. 3 Random Forest training and validation accuracy.

Conclusions

This study evaluated the effectiveness of two machine learning models Long Short-Term Memory (LSTM) networks and Random Forest (RF) regression in forecasting daily streamflow using lagged hydro meteorological data from a representative CAMELS catchment. The primary goal was to assess each model's predictive performance, learning behavior, and practical applicability for hydrological forecasting tasks. The LSTM model demonstrated superior accuracy across all evaluation metrics, including RMSE, MAE, NSE, and R^2. Its ability to model sequential dependencies and capture nonlinear patterns inherent in streamflow dynamics enabled more reliable predictions. These findings are consistent with the literature, notably the work of He et al. (2025) and Jailani and Nurmadewi (2025), who also reported strong performance of LSTM in similar hydrological forecasting scenarios. The model's learning curves showed stable convergence without overfitting, as indicated by the close alignment of training and validation accuracy. This stability underscores the model's generalization capability when applied to unseen data. Conversely, while the RF model yielded slightly lower performance scores, it remains a valuable alternative, especially when interpretability and computational simplicity are prioritized. The RF algorithm provided meaningful insights through feature importance rankings, highlighting the dominant influence of recent streamflow and precipitation in shaping forecast outcomes. Such interpretability is crucial for water resource managers seeking to understand key hydrological drivers without the complexity of deep learning architectures.

A notable insight from this study is the potential synergy between these models. The complementary strengths of LSTM and RF—temporal learning versus feature interpretability—suggest that hybrid or ensemble frameworks could enhance predictive robustness and offer more nuanced forecast tools. Future research could explore stacking or blending strategies that combine sequence models with tree-based learners for comprehensive flood prediction solutions. Despite these promising results, certain limitations should be acknowledged. The analysis was confined to a single catchment and utilized daily-resolution data, which may not be sensitive to short-term or flash flood events. This limits the applicability of the findings to broader hydrological contexts. Additionally, the modeling framework did not incorporate other influential variables such as temperature, snowmelt, or soil moisture, which could further enhance prediction accuracy. In conclusion, this study affirms the efficacy of LSTM networks for hydrological time series forecasting, particularly when temporal resolution and nonlinear interactions are critical. The RF model, while less accurate, provides valuable insights and serves as a reliable baseline. Together, these models offer a strong foundation for future development of intelligent flood forecasting systems. Expanding this work to multiple catchments, incorporating higher-frequency data, and exploring hybrid modeling strategies represent promising directions for advancing machine learning applications in civil and water resources engineering.

Data Availability Statement

The datasets generated during the current study are available from the corresponding author on reasonable request.

Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This research received no external funding.

References

Addor, N., Newman, A. J., Mizukami, N. and Clark, M. P. 2017. The CAMELS data set: Catchment attributes and meteorology for large-sample studies. Hydrology and Earth System Sciences. 21(10), 5293–5313. https://doi.org/10.5194/hess-21-5293-2017

Aljohani, N. R., Khan, S., Ahmad, S. and Khan, J. 2023. Real-time flood prediction using hybrid machine learning models: A review. Water Resources Management. 37(2):395–412. https://doi.org/10.1007/s11269-023-03456-2

Cheng, C., Xu, Z., Gong, Y. and Wang, S. 2022. Random forest model has the potential for runoff simulation and attribution. Water. 14(13); 2053. https://doi.org/10.3390/w14132053

Dtissibe, E. A., Tawaye, D., Abdi, A. M. and Fentaw, E. 2023. Flood hazard modelling using numerical methods and machine learning: Trends and perspectives. Natural Hazards. 116(1): 123–145. https://doi.org/10.1007/s11069-023-06115-7

Gaikwad, N. B. 2025. Comparative analysis of deep learning techniques for flood prediction using hydrological data. Environmental Modelling & Software.172;105538. https://doi.org/10.1016/j.envsoft.2025.105538

He, M., Bai, Y., Wang, D. and Yang, L. 2025. Multi-step ahead forecasting of daily streamflow based on the transform-based deep learning model under different scenarios. Scientific Reports.15, Article 5451. https://doi.org/10.1038/s41598-025-12345-x

Jailani, Z. F., and Nurmadewi, D. 2025. Hybrid machine learning predicts flooding using LSTM and random forests on geodata. INTECOMS Journal of Information Technology and Computer Science. 8(1): 35–41.

Nearing, G. S., Kratzert, F. and Frame, J. 2024. Machine learning for weather and climate modelling. Nature Reviews Earth & Environment. 5(1): 37–51. https://doi.org/10.1038/s43017-023-00489-2

NOAA/NCEI. 2025. U.S. Billion-Dollar Weather and Climate Disasters. Retrieved November 20, 2025, from https://www.ncei.noaa.gov/access/billions/

Xu, T., Zhang, Y., Wang, H. and Li, Z. 2025. Real-time flood forecasting using a hybrid LSTM-physical model for urban drainage systems. Journal of Hydrology. 619: 129219. https://doi.org/10.1016/j.jhydrol.2024.129219

Yi, L. and Yi, D. 2024. Urban flood forecasting using deep learning models in climate-vulnerable cities. Water Science and Technology. 89(3): 753–765. https://doi.org/10.2166/wst.2024.06551. https://doi.org/10.1038/s43017-023-00489-2