Water and Soil Management and Modeling

Online ISSN: 2783 - 2546

University of Mohaghegh Ardabili

Integration of conceptual hydrological and machine learning models via output augmentation for enhanced streamflow prediction

Bewketu Assefa Mulu [1]*, Fasikaw Atanaw Zimale [2], Mulugeta Genanu Kebede[3]

[1] PhD Candidate in Geoinformation and Earth Observation for Hydrology, Faculty of Meteorology and Hydrology, Arba Minch Water Technology Institute, Arba Minch University, Arba Minch, Ethiopia

[2] Associate Professor, Faculty of Civil and Water Resources Engineering, Bahir Dar Institute of Technology, Bahir Dar University, Bahir Dar, Ethiopia

[3] Assistant Professor, Institute of Geophysics, Space Science, and Astronomy, Atmospheric and Oceanic Sciences Unit, Addis Ababa University, Addis Ababa, Ethiopia

Abstract

Quantifying water resources is essential for developing evidence-based management strategies. Hydrological models play a great role in estimating streamflow, particularly in regions with limited flow measurement infrastructure. This study evaluates the integration of the GR4J conceptual hydrological model with Machine Learning (ML) techniques, Random Forest (RF), Extreme Learning Machine (ELM), eXtreme Gradient Boosting (XGB), and Long Short-Term Memory (LSTM) networks to improve daily streamflow prediction in the Bilate River watershed. Though GR4J captures general hydrological trends, its limitations in modeling nonlinear dynamics and extreme flows necessitate advanced approaches by augmenting GR4J's simulated outputs with climate input features to train the ML models. The integrated models GR4J-RF, GR4J-ELM, GR4J-XGB, and GR4J-LSTM combine GR4J's physical interpretability with ML's capability to capture complex and nonlinear relationships, addressing the shortcomings of both the conceptual and ML methods. Findings of the study demonstrate significant improvements over standalone GR4J, with GR4J-LSTM and GR4J-XGB achieving the highest test performance (NSE of 0.77, KGE of up to 0.86), GR4J-RF excelling in training fit (train NSE of 0.87) with gaps in generalization, and GR4J-ELM offering computational efficiency with comparable performance (test NSE of 0.74). These findings highlight the potential of integrated modeling to improve streamflow prediction in data-limited regions, supporting applications such as flood prediction and drought monitoring.

Keywords: GR4J, Conceptual hydrological model, Machine Learning, Integrated model, Lake Abaya-Chamo Basin

Article Type: Research Article

*Corresponding Author, E-mail: bewketu_assefa@dmu.edu.et

Citation: Mulu, B. A., Zimale, F. A., & Kebede, M. G. (2025). Integration of conceptual hydrological and machine learning models via output augmentation for enhanced streamflow prediction, Water and Soil Management and Modeling, 5(4), 95-115. doi: 10.22098/mmws.2025.18128.1648

Received: 20 August 2025, Received in revised form: 16 September 2025, Accepted: 19 October 2025, Published online: 07 November 2025

Water and Soil Management and Modeling, Year 2025, Vol. 5, No. 4, pp. 95-115.

Publisher: University of Mohaghegh Ardabili

1. Introduction

The design, operation, and management of hydraulic structures such as dams and appurtenant structures, diversion headworks, and bridges heavily rely on accurately measured or correctly estimated streamflow data. Therefore, quantifying streamflow is critical, particularly in data-scarce regions, where hydrological models serve as essential tools for streamflow estimation (Ma et al., 2018; Sezen et al., 2019; Adane et al., 2021; Bargam et al., 2024). These models are broadly categorized into process-based hydrological models (PBHMs) and data-driven approaches, each with distinct strengths and limitations. PBHMs, which simulate hydrological processes through mathematical representations (Dessie et al., 2014; Al-Mukhtar & Al-Yaseen, 2019), face challenges related to structural uncertainties, parameter calibration, and computational demands (Clark et al., 2017; S. Liu et al., 2022). Although lumped models like GR4J simplify spatial variability (Shi et al., 2011), distributed models such as SWAT require extensive datasets (Zamani et al., 2021; Janjić & Tadić, 2023), limiting their applicability in data-poor regions. On the other hand, machine learning (ML) and deep learning (DL) models excel at capturing complex, nonlinear relationships (Li et al., 2022) but often lack interpretability and are prone to overfitting when training data is limited (Shen et al., 2018; Kapoor et al., 2023).

To overcome these limitations, recent research has explored hybrid modeling, integrating PBHMs with ML techniques to leverage their complementary strengths (Humphrey et al., 2016; Kumanlioglu & Fistikoglu, 2019). Various integration strategies have been proposed, including model output fusion (Zhang et al., 2020), physics-guided ML where physical constraints are embedded into loss functions (Khandelwal et al., 2020), and feature augmentation using PBHM-simulated variables as additional ML inputs (He et al., 2021). For instance, Humphrey et al. (2016) demonstrated improved monthly flow predictions by coupling a conceptual model with ANN, while He et al. (2021) showed enhanced daily streamflow simulation through GR4J-LSTM integration. However, most existing work has focused on LSTM-based integrations (Kwak et al., 2022; Mei et al., 2024), leaving other ML methods such as random forests (RF), extreme learning machines (ELM), and XGBoost (XGB) relatively unexplored. Additionally, few studies rigorously compare hybrid models against standalone ML approaches (Hao & Bai, 2023) to determine whether performance gains stem from integration or simply from the use of ML.

In the Abaya-Chamo Lake basin of Ethiopia, where the studied watershed is found, previous hydrological studies have predominantly relied on PBHMs such as SWAT (Ayalew et al., 2023; Mada & Nannawo, 2023; Beza et al., 2024; Darota et al., 2024), HEC-HMS (Ibrahim et al., 2024), and MIKE11-NAM (Nannawo et al., 2022), often at coarse temporal resolutions due to concerns about input data variability. This study seeks to address these gaps by evaluating multiple integrated modeling approaches (GR4J combined with RF, ELM, XGB, and LSTM) through feature augmentation of GR4J simulated outputs (simulated flow and state variables) as input variables to the ML techniques. By systematically assessing different integration strategies in a data-scarce watershed, Bilate, this research contributes to the broader hydrological literature by: 1) identifying optimal integration configurations other than the commonly used LSTM approach; 2) clarifying whether GR4J simulated output-augmented integrations enhance predictive skill than the standalone base model; and 3) providing a transferable framework for improving streamflow predictions in understudied watersheds of the basin.

2. Materials and Methods

2.1. Study Area Description

This study focused on the Abaya-Chamo Lake basin, specifically examining the Bilate River watershed, one of the major tributaries of Lake Abaya. The study area encompasses the Bilate River watershed upstream of the hydrological station near Alaba Kulito town, covering approximately 2,008 km². Geographically, the watershed extends between 7°16'6"N to 8°6'45"N latitude and 37°46'32"E to 38°14'55"E longitude (see Figure 1).

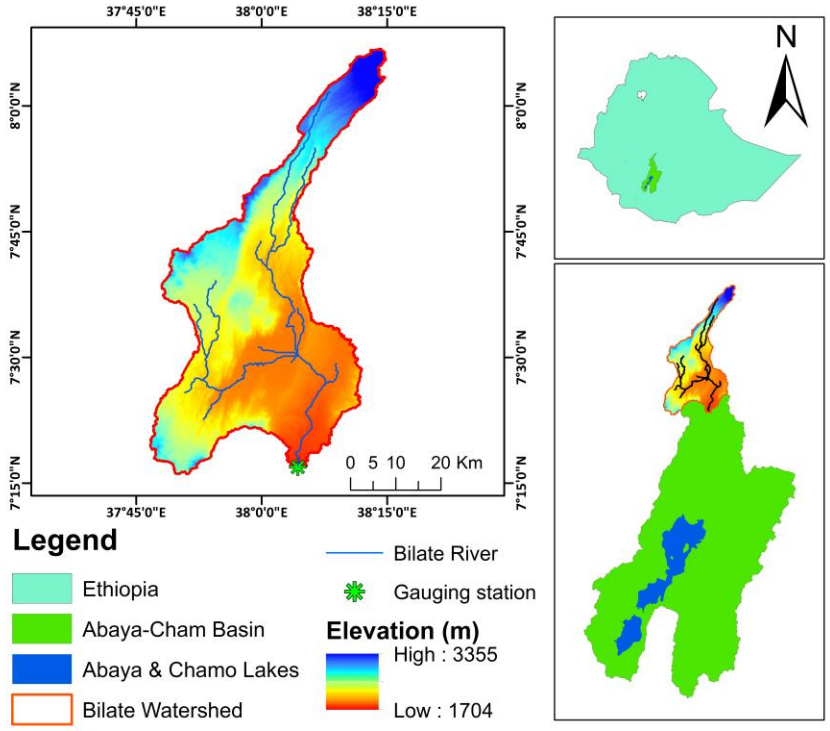

Figure 1. Geographical location of the study area

2.2. Data and data sources

This study utilized multiple datasets to characterize the hydro-climatic settings of the Bilate River watershed from 1985 to 2015. Primary meteorological data, including daily precipitation and minimum/maximum temperature records, were obtained from the Ethiopian Meteorological Institute (EMI). Corresponding daily streamflow measurements were collected from the Ministry of Water and Energy (MoWE) of Ethiopia for the station near Alaba Kulito town. To address missing values in precipitation and temperature, we used the Enhancing National Climate Services (ENACTS) gridded dataset. This dataset has employed an advanced blending algorithm to combine ground measurements with satellite estimates and has demonstrated strong agreement with observed station data (Dinku et al., 2018). Recently, Woldemariam et al. (2025) further validated ENACTS's applicability for hydrological studies in the Rift Valley basin of Ethiopia. Table 1 presents detailed specifications of all datasets used in this research.

Table 1. Hydro-meteorological datasets used in this study

Dataset	Description	Spatial Resolution	Temporal Resolution	Source	Time span
Precipitation	Station based	--	daily	EMA	1985 - 2015
	ENATS da	4 km			1985 - 2015
Temperature	Station based	--			1985 - 2015
	ENATS data	4 km			1985 - 2015
Streamflow	Station	--		MoWE	1985 - 2015

2.3. Methodology

This study is aimed at integrating the GR4J conceptual hydrological model with RF, ELM, XGB, and LSTM ML networks to simulate daily streamflow in the Bilate River watershed, with particular focus on a hydrological station near

Alaba Kulito town. The methodology utilized GR4J simulated outputs, including simulated flow and state variables (production store and routing store) as hydrologically significant features, augmented with climate datasets to train the ML models. These integrated approaches leverage both hydrological processes understanding of the conceptual model and data-driven learning capabilities of ML models. The complete methodological workflow implemented in this study is demonstrated in Figure 2.

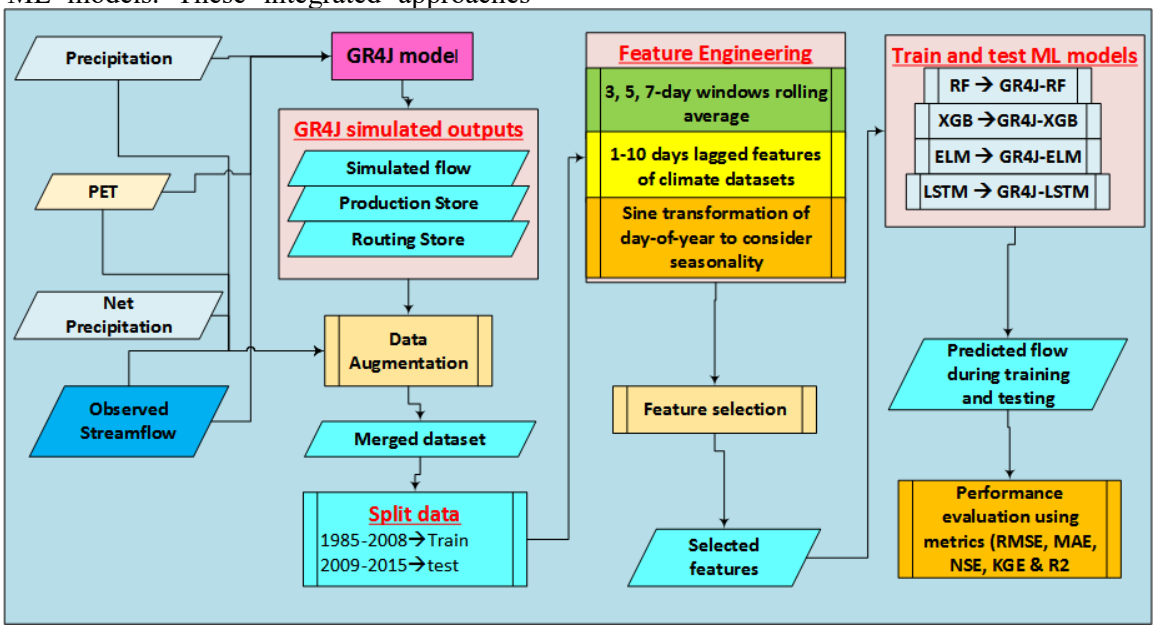

Figure 2. Conceptual framework for integration of the GR4J conceptual model with ML model

2.3.1. Data preparation

The station-based precipitation and temperature records had missing values, which were filled using the 4 km resolution gridded ENACTS weather dataset. Similarly, missing timestamps in the streamflow data were imputed using Random Forest regression, a robust machine learning approach for hydrological data imputation and rectification (Hamzah et al., 2021).

Then, as PET is a critical input for the GR4J model, it was computed using *Enku's temperature method* (Enku & Melesse, 2013), an empirical approach validated for Ethiopia. This method provides reliable PET estimates while requiring only temperature data, making it suitable for data-scarce regions.

$$PET = \frac{(T_{max})^n}{k} \quad (1)$$

Where PET is the potential evapotranspiration (mm/day); T_{max} is the daily maximum temperature in °C; $n = 2.5$; $k = 48*T_{mm} - 330$ for combined wet and dry conditions, where T_{mm} is the long-term daily mean maximum temperature in oC.

Moreover, the streamflow records at the target station also exhibited statistical inconsistencies. Adjustments were done using empirical cumulative distribution function (CDF) matching and quantile mapping techniques (Nguyen et al., 2024) to secure temporal consistency, an imperative step for reliable hydrological modeling. Additionally, the streamflow data had high-frequency noise that distorted hydrological patterns and compromised model performance. The application of low-level wavelet transformation reduced this noise while maintaining critical hydrological signals in the time series. Then, datasets were converted to NumPy arrays to meet the input requirements for both the GR4J hydrological and the ML models.

2.3.2. Conceptual hydrological model (GR4J)

The GR4J (Génie Rural à 4 parameters Journalize) was selected for this study as a parsimonious yet robust conceptual hydrological

model (Anshuman et al., 2021). As demonstrated by Perrin et al. (2003), this four-parameter model effectively simulates daily streamflow using only precipitation and potential evapotranspiration inputs, making it particularly suitable for data-scarce regions. Its structure captures key hydrological processes through four calibrated parameters: X_1 represents the maximum capacity of the production store (mm), X_2 governs groundwater exchange, X_3 determines the routing store capacity (mm), and X_4 controls the unit hydrograph time (days) (Anshuman et al., 2021; Asgari et al., 2025). To further improve model performance, an additional calibration parameter, scale factor (S_f), was incorporated, intending to mitigate measurement errors and model structural limitations.

This model was implemented in Python utilizing the Digital Earth Africa Sandbox and its computational resources (dual-core processor, 16GB RAM) (Digital Earth Africa, 2025). The model architecture was designed to process daily climatic inputs, specifically precipitation (P) and potential evapotranspiration (PET) time series.

The simulation function was developed to process daily inputs of precipitation (P) and potential evapotranspiration (PET) over *n* time steps. For each time step *t*, a stepwise and simplified algorithm was implemented to successfully run the GR4J model. The stepwise algorithm employed is presented as follows.

i. Production store update

Production store update with precipitation:

$$S_t = \begin{cases} S_{t-1} + P_t & \text{if } P_t > 0 \\ S_{t-1} & \text{if } P_t \le 0' \end{cases} \qquad S_t = \min(S_t, x_1) \qquad (2)$$

Effect of Evapotranspiration:

$$S_t = S_t - E_t, \quad S_t = \max(S_t, 0) \qquad (3)$$

ii. Calculation of percolation

Percolation is the water infiltrated from the production store to the routing store (Perrin et al., 2003).

Water transfer from S to R was calculated using a nonlinear function:

$$\mathrm{Perc}_t = \left[\sqrt{S_t} * \left(1 - \max\left(0, 1 - \frac{S_t}{x_1 + \epsilon}\right)^{\frac{1}{1+\max(x_2, 10^{-8})}}\right)\right]^{1.2} \qquad (4)$$

Where $S_t = S_t - \mathrm{Perc}_t$; $R_t = R_{t-1} + \mathrm{Perc}_t$ and ϵ = very small number for numerical stability

iii. **Runoff Generation:**

Direct runoff (Q_{dt}) is computed from the routing store using a nonlinear runoff-generating equation.

$$Qd_t = \left[R_t * \left(1 - \exp\left(-\frac{1}{x_3}\right)\right)\right]^{1.2} \qquad (5)$$

To ensure non-negative routing storage:

$$R_t = \max(R_t - Qd_t, 0) \qquad (6)$$

Total simulated flow was scaled by applying a scale factor (S_f)

$$Q_t = Qd_t \cdot S_f \qquad (7)$$

Then the above algorithm returns time series GR4J simulated flow and corresponding values of the two state variables, production store and routing store (Dambré et al., 2024).

The optimal parameter values for the GR4J model were determined through automated calibration utilizing the Differential Evolution (DE) algorithm. The optimization process employed KGE as the objective function, which evaluates model performance by simultaneously considering correlation, variability, and biases.

2.3.3. Long Short-Term Memory (LSTM)

In this study, LSTM with residual connection was used for its performance in capturing sequential features characteristics to predict streamflow (Le et al., 2021). LSTM architecture is a special type of Recurrent Neural Network (RNN) which is capable of learning long-term dependencies in sequence prediction problems (Wegayehu & Muluneh, 2022). The Architecture of an LSTM network built around memory cells whose information is manipulated by three gates: an input gate, a forget gate, and an output gate (Gers & Cummins, 2000).

The input layer accepts sequences of shape [L, F], where L is the sequence length and F is the number of features at each time step t. This sliding window approach is common in hydrological time series forecasting (Xiang et al., 2020; Le et al., 2021). For a given time step t, the input vector is:

$$x_t = \left[F_{1,t}, F_{2,t}, F_{3t}, \ldots, F_{F,t}\right] \in \mathbb{R}^F, \quad t = 1,2,\ldots,L \qquad (8)$$

The input sequence for a sample at time t can be expressed as:

$$X_t = [x_{t-L+1}, x_{t-L+2}, \dots, x_t] \in \mathbb{R}^{L \times F} \quad (9)$$

Each LSTM layer processed the input sequence using a recurrent structure with memory cells, governed by forget, input, and output gates.

Forget gate: information that is no longer useful in the cell state is rejected by the forget gate (Gers & Cummins, 2000). It takes X_t (input at time t) and h_{t-1} (previous cell output), and multiplies weight followed by the addition of bias (Siami-Namini et al., 2019). The result passes through an activation function, which gives a binary output, 0 and 1. If the cell state output is 0, the information is forgotten, and if it is 1, the information is retained for future use (Greff et al., 2017).

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \quad (10)$$

Input gate: The addition of information to the cell state is done by the input gate. The information is first gated by a *sigmoid* function, which uses the inputs h_{t-1} and X_t to filter and decide which values to remember. The *tanh* function is then applied to generate a vector with values ranging from -1 to +1 that contains all of the possible values from h_{t-1} and x_t. Finally, the vector and regulated values are multiplied to obtain useful information. This two-step process creates a candidate value that is added to the state (Graves et al., 2013).

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \quad (11)$$

$$\tilde{\tilde{C}}_t = \tanh(W_c.[h_{t-1}, x_t] + b_c) \quad (12)$$

$$C_t = (f_t.C_{t-1} + i_t.\tilde{C}_t) \quad (13)$$

Output gate: The output gate is in charge of extracting useful information from the current cell state and presenting it as output (Kratzert et al., 2018). To begin, a vector is created by applying the *tanh* function to the cell. The information is then regulated using the sigmoid function and filtered by the values to be remembered via h_{t-1} and X_t inputs. Finally, the vector and regulated values are multiplied and sent as output and input to the next cell (Greff et al., 2017).

$$O_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \quad (14)$$

$$h_t = O_t * \tanh(C_t) \quad (15)$$

Where W_f, W_i, W_c, W_o: weight metrics for forget, input, candidate, and output gate; bf, bi, bc, bo: bias vectors; *tanh:* activation function, *σ: sigmoid activation function;* X_t*:* input at time t, h_{t-1}*:* previous cell output, f_t*:* information retained at forget gate, i_t*:* regulated information at input gate, $\tilde{C}_t$*:* vector ranging -1 to 1, C_t*:* useful information at input gate, O_t*:* output, h_t*:* input to the next cell.

2.3.4. Random Forest (RF)

RF is an ensemble machine learning algorithm that builds multiple decision trees during training and combines their predictions, typically averaging them for regression tasks (Breiman, 2001). Its ability to capture nonlinear relationships and manage multi-dimensional datasets makes it a prevalent choice for hydrological modeling (Tyralis & Papacharalampous, 2019). By introducing randomness through bootstrapped training samples and random feature selection at each tree split, RF improves model generalization and robustness (Cutler et al., 2012). Compared to deep learning models, RF requires minimal hyperparameter tuning, offering computational efficiency that is particularly valuable for hydrological applications.

Recent studies highlight RF's effectiveness in streamflow forecasting. For instance, Li et al. (2019) demonstrated its strong performance in predicting daily streamflow, while Zhang & Thorburn (2022) noted its resilience to missing data and noise, common challenges in hydrological datasets. These qualities make RF well-suited for modeling complex hydrological systems.

In basins with limited data, such as the Bilate watershed, RF offers a powerful, data-driven solution for streamflow prediction. Its ability to integrate diverse environmental data sources enhances its utility in water resource management, making it an essential tool for researchers.

2.3.5. Extreme Learning Machine (ELM)

ELM is a single-hidden-layer feedforward neural network popular for its fast training and robust generalization (Huang et al., 2006). Unlike conventional neural networks that rely on iterative backpropagation, ELM randomly initializes the weights and biases of its input layer and analytically computes the output layer weights using least-squares optimization (Huang et al., 2011). This approach significantly reduces computational demands while preserving high

predictive accuracy, making ELM effective for hydrological modeling, where nonlinear relationships are dominant (Yaseen et al., 2016). Its ability to model nonlinear functions without extensive parameter tuning makes it well-suited for real-time forecasting in regions with limited data (Adnan et al., 2019). ELM has comparable performance with deep learning models like LSTM networks, in situations with limited training data (Tao et al., 2022).

2.3.6. Extreme Gradient Boosting (XGB)

XGB, an ensemble machine learning method, is excellent in regression and classification by an optimized gradient boosting framework. It uses regularization to mitigate overfitting and captures nonlinear relationships, making it suitable for hydrological modeling. By iteratively constructing decision trees that correct errors from prior iterations, XGB achieves fast training and scalability for large datasets (Chen & Guestrin, 2016).

Recent studies demonstrate that XGB often surpasses conventional models like Random Forest and Artificial Neural Networks in both accuracy and computational efficiency, especially for high temporal resolution predictions (Hao & Bai, 2023).

2.3.7. Integration of GR4J with LSTM, RF, ELM, and XGB

This study proposed an integrated modeling framework that augments simulation outputs of the GR4J model to train ML models, LSTM, RF, ELM, and XGB. The primary objective is to enhance streamflow prediction by combining GR4J simulated outputs, including streamflow and state variables (production store, *PS,* and routing store, *RS*), with climate data (precipitation, *P*, potential evapotranspiration, *PET*, and net precipitation) and their derived features. The input features used in this study can be categorized into three. 1) Hydrological features: representing watershed response dynamics (GR4J-simulated streamflow, *PS*, *RS*). 2) Climate features: External drivers of hydrological processes (*P*, *PET, P_net*). 3) Temporal features: Capturing seasonal patterns (sine-transformed day of the year).

To ensure the selection of the most relevant input features for the models, a series of preprocessing steps was employed. Initially, temporal lags ranging from 1 to 10 days were incorporated to capture the influence of antecedent conditions on streamflow, accounting for the delayed effects of meteorological inputs. Next, seasonality was encoded using a sine transformation of the day-of-year to effectively represent annual recurring patterns in the dataset and enhance the models' ability to capture seasonal patterns. Then, rolling statistics were computed with moving average windows of 3, 5, and 7 days to smooth short-term variability while preserving meaningful signals in the time series. This step reduced noise in the data, making the models focus on significant signals. Finally, feature selection was performed using Pearson's correlation to identify the most important and non-redundant features. Features with an absolute correlation coefficient (|r|) greater than 0.4 with the target were retained. To mitigate multicollinearity, features exhibiting an absolute feature-to-feature correlation greater than 0.8 were assessed, and the feature with the lower target correlation was removed, thus keeping a balanced and effective features for the models.

2.3.7.1. Data Splitting and Normalization

A two-way data split was applied in this study to segment the data into training and testing sets to train and evaluate the ML models. A temporal split method was applied to maintain chronological order, ensuring the training period preceded the testing period. Training from 1985 to 2008, used for model training, and testing from 2009 to 2015 for performance assessment. Splitting of the data was conducted before applying any feature engineering tasks, such as rolling means, to prevent data leakage between training and testing datasets.

To ensure ML's training stability and better results, all variables, both in training and testing datasets, were normalized. The Min-Max scaler was applied to scale the features and the target to

the range 0 to 1 (Liu et al., 2020). MinMaxScaler uses this formula.

$$X_{scaled} = \frac{X_t - X_{min}}{X_{max} - X_{min}} \quad (16)$$

Where X_{scaled}: normalized value at time t; X_t: Value before normalization at time t; X_{max}: the maximum value within the time series; X_{min}: The minimum value within the time series.

2.3.8. Hyperparameter tuning

Hyperparameters are predefined parameters that cannot be learned during model training, yet they significantly influence an ML model's performance. Properly optimized hyperparameters boost model accuracy and optimize training efficiency.

For LSTM models, essential hyperparameters include the number of LSTM layers, units per layer, activation function, dropout rates, regularization parameters, learning rate, batch size, and the number of epochs (Bartz-beielstein & Zaefferer, 2023).

Hyperparameter tuning for the LSTM model was performed using Keras Tuner, an open-source optimization tool integrated with Python 3.8 and TensorFlow/Keras. The model was compiled using the Adam optimizer and mean squared error (MSE) loss function (Wegayehu & Muluneh, 2021, 2022). The tuning process encompassed 75 trials, with each trial running for a maximum of 50 epochs and incorporating EarlyStopping with a patience of 10 epochs to mitigate overfitting.

Following the initial tuning, additional trial-and-error adjustments were made to refine the hyperparameters further. Key LSTM hyperparameters, such as sequence length and the number of epochs, were determined through iterative experimentation to identify optimal values for the model. The optimization process was done on the Digital Earth Africa Sandbox Computing platform, equipped with a dual-core processor and 16 GB of RAM, providing computational resources for the tuning task.

Hyperparameters for RF, ELM, and XGB were tuned using the Grid Search method, which exhaustively evaluates predefined hyperparameter combinations through cross-validation. For detailed information on the hyperparameters and optimal values, see Table 2.

Table 2. Optimal hyperparameters for FR, ELM, XGB, and LSTM

Hyperparameters for RF		
Hyperparameter	Optimal Values	Description
n_estimators	200	Number of trees
max_depth	10	Maximum depth of trees
min_samples_split	2	Minimum samples to split a node
min_samples_leaf	10	minimum samples at a leaf node
Hyperparameters for ELM		
n_hidden	20	Number of hidden neurons
activation	sigmoid	Activation function
Hyperparameters for XGB		
n_estimators	100	Number of boosting rounds
max_depth	3	Maximum depth of trees
learning_rate	0.05	Step size shrinkage
subsample	0.8	Fraction of sample per tree
colsample_bytree	0.6	Fraction of features per tree
Hyperparameters for LSTM		
num_layers	3	Number of LSTM layers
units_1	128	Number of units in the first layer.
units_2	64	Number of units in the second layer
units_3	32	Number of units in the third layer
dropout_rate_1	0.5	Dropout rate for the first layer
dropout_rate_2	0.4	Dropout rate for the second layer
dropout_rate_3	0.4	Dropout rate for the third layer
l1_reg	8.00E-06	L1 regularization to penalize large weights.
l2_reg	8.30E-06	L2 regularization to penalize large weights.
learning_rate	0.0001	Learning rate for the Adam optimizer

batch_size	16	Number of samples per gradient update
Sequence_length	15	Number of time steps in the LSTM input sequence
Epochs	100	Number of training iterations

2.3.9. Model Performance Evaluation Metrics

Hydrological model performance is usually assessed using multiple metrics, each offering unique understandings into different aspects of model accuracy (Moriasi et al., 2007). In this study, we employed five widely recognized evaluation metrics in hydrology, namely, NSE, R^2, RMSE, MAE, and KGE. These metrics are calculated as follows (Nash & Sutcliffe, 1970; Duc & Sawada, 2023):

$$NSE = 1 - \frac{\sum(Q_{0bs} - Q_{sim})^2}{\sum(Q_{0bs} - \bar{Q}_{0bs})^2} \quad (17)$$

$$R^2 = \frac{(\sum(Q_{obs} - \bar{Q}_{obs})(Q_{sim} - \bar{Q}_{sim}))^2}{\Sigma(Q_0 - \bar{Q}_0)^2(Q_{sim} - \bar{Q}_{sim})^2} \quad (18)$$

$$KGE = 1 - \sqrt{(r-1)^2 + (a-1)^2 + (\beta - 1)^2} \quad (19)$$

$$RMSE = \sqrt{\frac{\Sigma(Q\_sim - Q\_obs)^2}{n}} \quad (20)$$

$$MAE = \frac{\sum|Q_{sim} - Q_{0bs}|}{n} \quad (21)$$

Where Q_{obs} and Q_{sim} represent observed and simulated streamflow at time step i, and $\bar{O}_{obs}$ and $\bar{O}_{sim}$ denote their respective means, n is the number of observations, and r, α, and β correspond to the correlation, variability ratio, and bias components of KGE

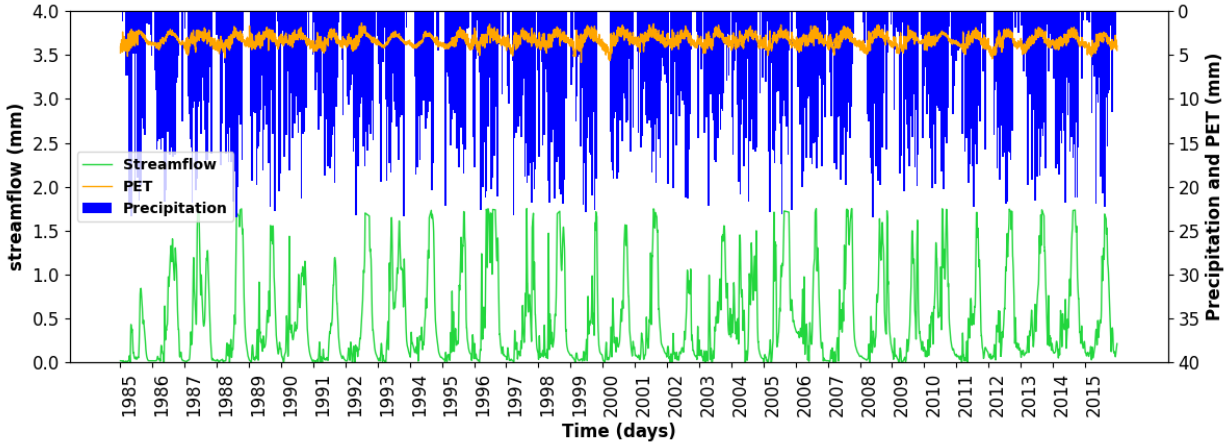

Figure 3. Time series GR4J model input datasets

3. Results and Discussion

3.1. GR4J Conceptual Model

After applying appropriate data processing steps, including handling missing values, adjusting data consistency, and correcting outliers, the time series input datasets for the GR4J model (precipitation *P*, potential evapotranspiration *PET*, and observed streamflow Q_{obs}) are plotted and shown in Figure 3. The statistical characteristics of the data are summarized in Table 3.

As shown in Table 3, Precipitation exhibits significant variability, as indicated by its large standard deviation, and is positively skewed, suggesting that while most days experience minimum rainfall values, there are occasional instances of heavy rainfall. In contrast, Potential Evapotranspiration (PET) remains relatively stable, showing small fluctuations and a nearly symmetric distribution. Streamflow, although more variable than PET, is less so compared to precipitation. It also follows a positive skew, indicating that streamflow values are generally low, with occasional high flow events.

Table 3. Statistical characteristics of the GR4J's input data

Variables/Stat.	min	max	mean	Std.	skewness	kurtosis	sample size
Precipitation (mm)	0.00	23.53	2.99	4.34	1.80	3.16	11322
PET (mm)	1.38	5.85	3.34	0.63	0.08	-0.32	11322
Streamflow (mm)	0.00	1.75	0.48	0.51	1.14	0.07	11322

3.2. GR4J Model Calibration

The GR4J model parameters were calibrated using DE, applying the approach of Napiorkowski et al. (2023), with KGE serving as the objective function. This study optimized the four standard GR4J parameters (X_1, X_2, X_3, *and* X_4) along with an additional scale factor (S_f). The resulting calibrated parameter values (see Table 4) were 939.42 mm for X1, -0.70 mm for X2, 1.00 mm for X3, 2.73 days for X4, and 0.60 for Sf.

Table 4. GR4J model calibration parameters and their corresponding values

Parameter	Description	Optimal value
X_1	Maximum capacity of the production store (mm)	939.42
X_2	Groundwater exchange coefficient (mm)	-0.70
X_3	A day ahead, maximum capacity of the routing stores (mm)	1.00
X_4	The time base of the unit hydrograph (days)	2.73
S_f	Scale factor	0.60

The GR4J model produced three outputs: daily simulated flow and two state variables, the production store and the routing store. To visualize the time series GR4J simulated outputs, see Figure 4.

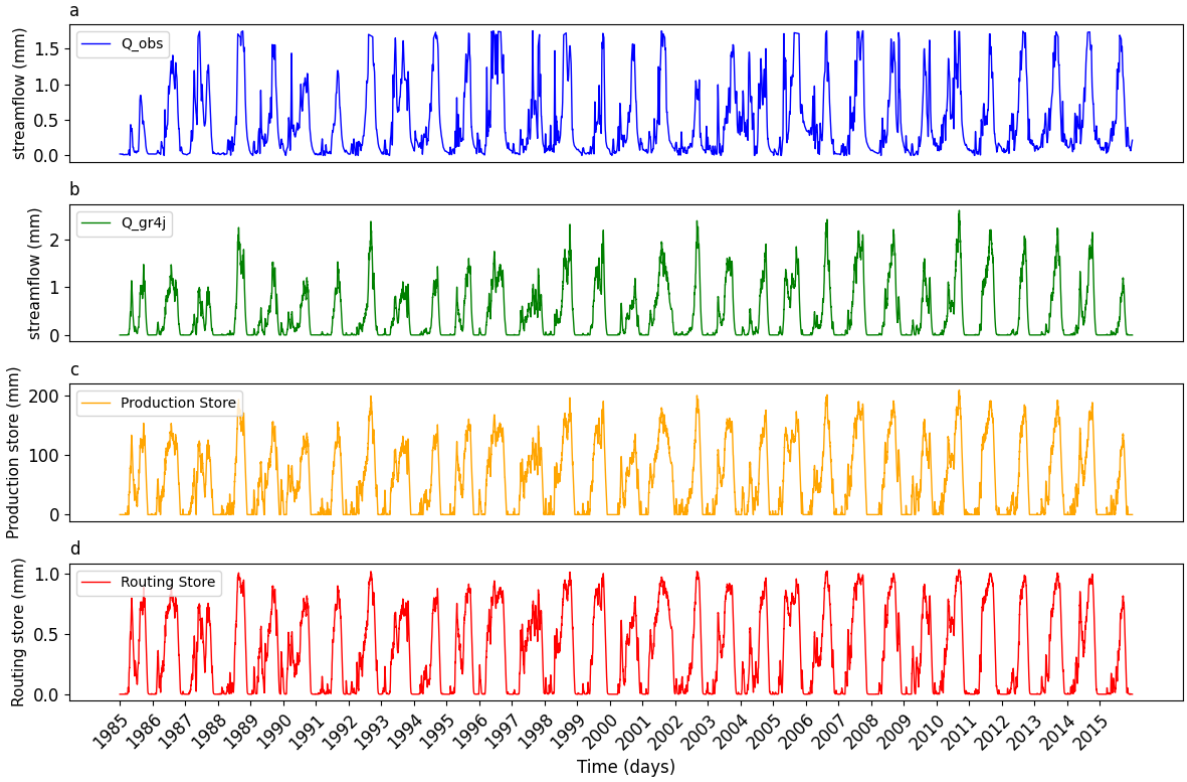

Figure 4. a: Observed streamflow time series; b: GR4J simulated streamflow time series; c: Production store time series; d: Routing store time series

3.3. Feature Selection

The dataset initially comprised 55 features, encompassing original climate variables, GR4J-simulated outputs, and derived features from lagged values and rolling averages. However, not all features contribute equally to model training, and including irrelevant features can negatively affect training speed and predictive performance. Therefore, a feature selection process was executed based on Pearson's correlation with the target and with features themselves. This approach resulted in the selection of eleven key features for training and testing the proposed models.

Figure 5 presents the correlation values of the selected features with the target and among themselves, offering insights into their importance and inter-relationships.

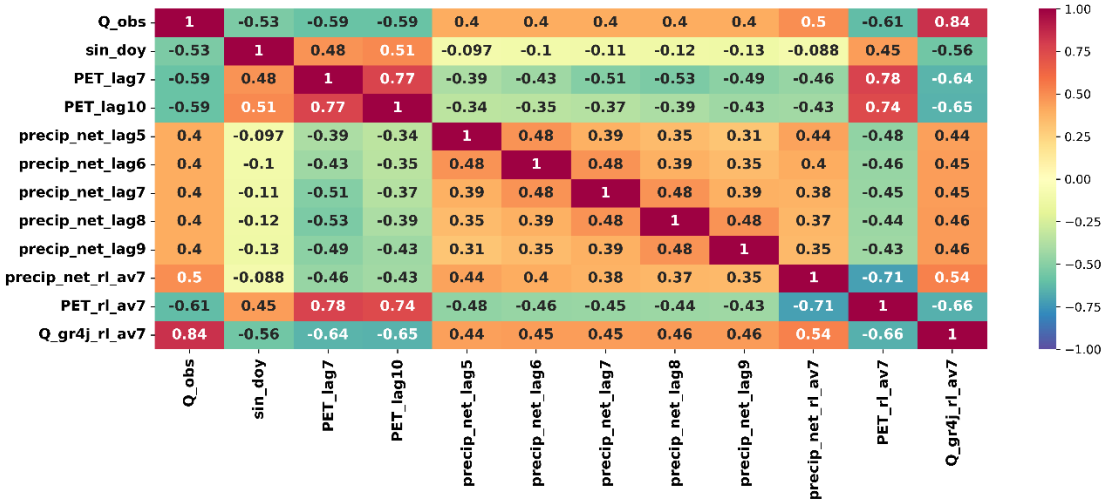

Figure 5. Heat-map showing correlation of selected features to the target, and feature to feature; it shows relatively strong correlation with the target and low correlation with each other

The heat-map visualization confirms that the selected features exhibit strong relationships with the target variable while maintaining a low correlation with each other, effectively mitigating the issue of multicollinearity and offering insights into their importance and inter-relationships.

3.4. Model's performance evaluation

The GR4J model was calibrated using data from 1985 to 2015. To facilitate a comparative analysis with the integrated models, the dataset was divided into training (1985–2008) and testing (2009–2015) periods. This consistent partitioning ensures fair performance evaluation across all five models. Although the terms 'training' and 'testing' are not conventionally used for GR4J, they are adopted here to keep alignment with the integrated models.

To achieve enhanced streamflow prediction capability, we integrated GR4J with advanced ML techniques. The results in Table 5 demonstrate significant improvements in predictive performance across all integrated models compared to the standalone GR4J, with nuanced variations among the ML integrations.

Table 5. Performance metrics and their corresponding values for all models during training and testing

Model	GR4J		GR4J-RF		GR4J-ELM		GR4J-XGB		GR4J-LSTM	
Datasets	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**
RMSE	0.302	0.347	0.107	0.153	0.151	0.148	0.140	0.141	0.149	0.138
MAE	0.204	0.255	0.073	0.107	0.104	0.102	0.099	0.098	0.102	0.096
NSE	0.65	0.53	0.87	0.72	0.73	0.74	0.77	0.76	0.74	0.77

KGE	0.80	0.69	0.87	0.82	0.79	0.83	0.80	0.83	0.79	0.86
R^2	0.70	0.70	0.87	0.73	0.73	0.75	0.77	0.77	0.74	0.79

Below, the findings are analyzed in detail, discussing models' strengths and weaknesses based on the metrics and hydrological implications.

3.4.1. Performance based on RMSE and MAE

As illustrated in Table 5, RMSE and MAE provide complementary insights into model accuracy, with RMSE emphasizing larger deviations and MAE offering a balanced measure of average prediction consistency without overweighting outliers. The standalone GR4J model exhibited a training RMSE of 0.302 mm, increasing to 0.347 mm during testing, while MAE rose from 0.204 mm to 0.255 mm. In contrast, in all integrated models (GR4J-RF, GR4J-ELM, GR4J-XGB, and GR4J-LSTM), substantial improvements were achieved.

During training, GR4J-RF achieved the lowest errors (RMSE = 0.107 mm; MAE = 0.073 mm), leveraging its ensemble tree structure to effectively model nonlinear hydro-meteorological relationships. It was followed by GR4J-XGB (with RMSE of 0.140 mm, and MAE of 0.099 mm), GR4J-LSTM (with RMSE of 0.149 mm, and MAE of 0.102 mm), and GR4J-ELM (with RMSE = 0.151 mm, and MAE = 0.104 mm). These results indicate the hybrid models' superior ability to fit observed data.

During testing, GR4J-LSTM outperformed other models (RMSE = 0.138 mm; MAE = 0.096 mm), likely due to its sequential learning capability, which effectively captures temporal dependencies in hydrological time series. GR4J-XGB closely followed (RMSE = 0.141 mm; MAE = 0.098 mm), reflecting its gradient-boosting precision, while GR4J-ELM (RMSE = 0.148 mm; MAE = 0.102 mm) and GR4J-RF (RMSE = 0.153 mm; MAE = 0.107 mm) exhibited slightly higher errors. Although GR4J-RF excelled in training, its minor generalization gap suggests potential overfitting compared to the GR4J-LSTM and GR4J-XGB. Collectively, the integrated models' reduced RMSE and MAE values reflect their enhanced predictive accuracy over the standalone GR4J, with GR4J-LSTM emerging as the most effective for operational streamflow prediction.

3.4.2. Performance based on NSE, KGE, and R^2

Based on the results shown in Table 5, further evaluation of model performance was done using the NSE, KGE, and R^2, which assess variance explanation, bias-variability-correlation balance, and trend alignment, respectively. The standalone GR4J exhibited moderate performance, whereas all integrated models demonstrated significant improvements.

GR4J achieved a training NSE of 0.65, explaining 65% of streamflow variance relative to a mean-flow baseline, but this declined to 0.53 during testing. Its KGE (integrating correlation, variability, and bias) decreased from 0.80 (training) to 0.69 (testing), while R^2 remained stable at 0.70, indicating reliable but limited trend-tracking capability. These results suggest that GR4J captures basic hydrological patterns but struggles with complex, nonlinear dynamics.

The hybrid models consistently outperformed GR4J. GR4J-RF achieved the highest training NSE (0.87), KGE (0.87), and R^2 (0.87), benefiting from its ensemble tree-based nonlinear modeling. However, its testing performance declined (NSE = 0.72; KGE = 0.82; R^2 = 0.73), indicating a slight generalization gap.

GR4J-LSTM dominated in testing, achieving the highest NSE (0.77), KGE (0.86), and R^2 (0.79), reflecting its superior generalization via sequential learning. During training, it achieved NSE of 0.74, KGE of 0.80, and R^2 of 0.74. Similarly, GR4J-XGB matched GR4J-LSTM's testing NSE (0.77) while achieving a KGE of 0.83 and R^2 of 0.77. GR4J-ELM also showed consistent improvements (training: NSE = 0.73, KGE = 0.79, R^2 = 0.73; testing: NSE = 0.74, KGE = 0.83, R^2 = 0.75), balancing efficiency and predictive capability.

In general, the integrated models outperformed the GR4J model, with GR4J-LSTM and GR4J-XGB leading in generalization (testing NSE = 0.77, KGE up to 0.86, R^2 up to 0.79), GR4J-RF excelling in training fit, and GR4J-ELM offering a computationally efficient alternative. These improvements accentuate the value of integrating ML with GR4J by augmenting simulated outputs

to capture complex hydrological dynamics, with LSTM and XGB being particularly suited for high-accuracy applications.

In addition to the performance metrics presented in Table 5, Figure 6 illustrates the time series plots of the observed and simulated streamflow of the standalone GR4J model, and its integration with the ML models.

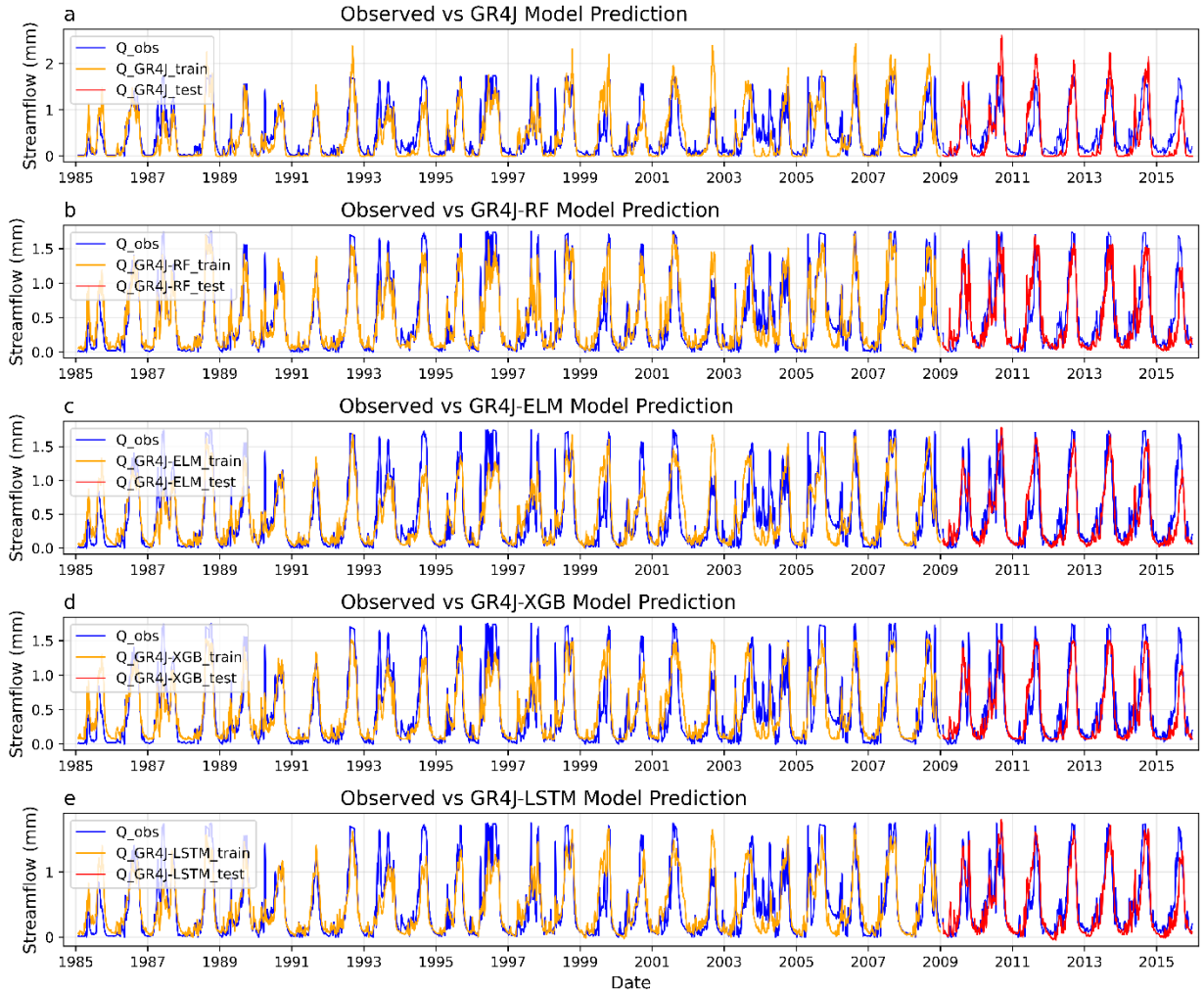

Figure 6. Time series observed and predicted streamflow during training and testing; a: for GR4J standalone model, b: for GR4J-RF integrated model, c: GR4J-ELM integrated model, d: for GR4J-XGB integrated model, and e: for GR4J-LSTM integrated model.

3.5. Discussion

The standalone GR4J model demonstrated moderate performance in streamflow prediction, effectively capturing general hydrological trends (Dambré et al., 2024). However, its limitations are apparent when addressing daily streamflow variability, particularly in highly dynamic watersheds like Bilate. With only four parameters and a simplified process representation, GR4J struggles to capture complex hydro-meteorological dynamics, a constraint well-stated in previous studies (Perrin et al., 2003; Kunnath-Poovakka & Eldho, 2019; Anshuman et al., 2021; Kodja et al., 2023). Figure 6a highlights these constraints, revealing systematic biases, overestimation of peak flows, and underestimation of low flows, consistent with findings by Kapoor et al. (2023) and Yang et al. (2024). These limitations emphasize the need for structural enhancements to improve predictive

performance, especially under extreme flow conditions.

3.5.1. Advantages of GR4J-ML Models

The integration of GR4J with ML methods, RF, ELM, XGB, and LSTM networks remarkably improved predictive accuracy across both training and testing phases. This enhancement arises from synergizing GR4J's physical interpretability with ML's ability to model nonlinear relationships and temporal dependencies (Tian et al., 2018; Sezen & Partal, 2022; Mei et al., 2024).

During training, all integrated models outperformed standalone GR4J, with GR4J-RF achieving the highest NSE (0.87) and KGE (0.87). However, while GR4J's testing performance declined (NSE = 0.53; KGE = 0.69), the integrating models maintained high accuracy, particularly GR4J-LSTM (NSE = 0.77; KGE = 0.86), demonstrating superior generalization across flow regimes. These results align with Konapala et al. (2020), who indicated ML's capacity to balance correlation, variability, and bias, making integrated models ideal for data-scarce regions like the Bilate watershed, where accurate water management is critical.

Time series analyses (Figures 6b to 6e) and performance metrics (Table 5) further confirm the integrated models' excellent alignment with observed streamflow data, even during extreme events. Although GR4J-LSTM slightly outperformed others, all integrated models addressed a key limitation of purely data-driven approaches: the lack of hydrological interpretability (Mohammadi, Safari, & Vazifehkhah, 2022). By training ML models with GR4J's physically-based features, such as simulated runoff, the integrated models provide a hydrologically informed framework, particularly valuable in regions with limited observational data (Armstrong et al., 2025). This integration effectively bridges the gap between conceptual and data-driven modeling, enhancing both accuracy and operational applicability. Comparative Performance and Hydrological Implications

The success of integrated modeling aligns with recent studies advocating such approaches (Mohammadi et al., 2022; Kapoor et al., 2023; Yang et al., 2024). These models retain GR4J's physical interpretability while leveraging ML's strengths in capturing complex patterns and nonlinear relationships (Hah et al., 2022; Liu et al., 2022). For instance, GR4J-LSTM's ability to capture temporal dependencies (test NSE = 0.77, KGE = 0.86) verifies Mei et al. (2024), who demonstrated that coupling conceptual models with deep learning improves predictive performance by modeling temporal variability in streamflow. Similarly, Yang et al. (2024) emphasized the role of integrated frameworks in balancing physical practicality with data-driven adaptability to complex rainfall-runoff processes, achieving high accuracy in runoff predictions, particularly during extreme events, a finding reflected in the superior predictive performance of GR4J-LSTM and GR4J-XGB in this study.

Further support comes from Mohammadi et al. (2022), who found that combining conceptual models with ML reduces errors and enhances generalization while mitigating the "black box" nature of pure ML techniques. Results of this study reinforce this, as GR4J's simulated features provided a meaningful basis for ML algorithms, improving both accuracy and interpretability.

3.5.2. Long-term mean daily streamflow evaluation

The evaluation of long-term mean daily streamflow predictions (Figure 7) provides critical understanding into model performance across the complete hydrological spectrum. Our analysis reveals that the integrated models (GR4J-RF, GR4J-ELM, GR4J-XGB, and GR4J-LSTM) demonstrate superior accuracy in capturing the full range of flow conditions compared to the standalone GR4J model. This enhanced performance is particularly evident in their ability to reproduce both seasonal flow patterns and extreme events, signifying these integrated approaches effectively combine physical process understanding with pattern recognition capabilities.

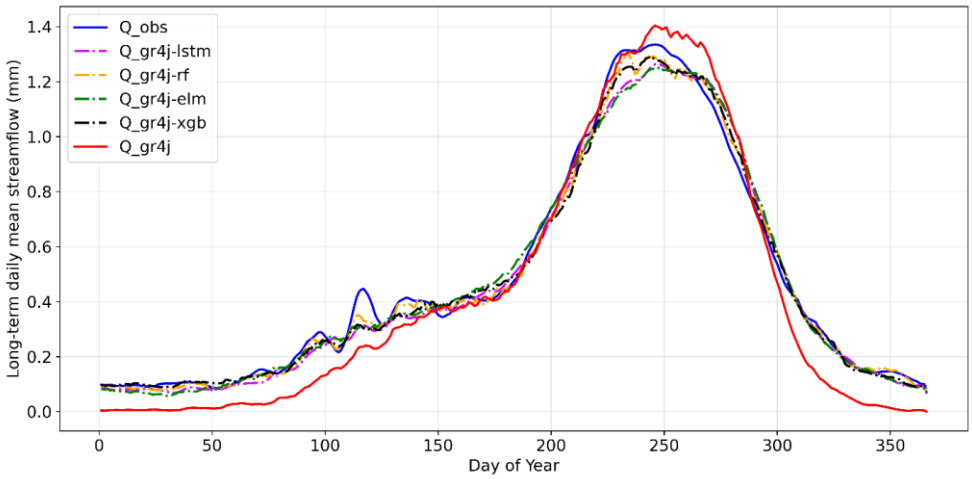

Figure 7. Long-term mean daily streamflow over days of the year

The standalone GR4J model exhibits systematic biases that show significant structural limitations. Consistent overestimation of high-flow events suggests deficiencies in its representation of quickflow generation processes, while underestimation of low flows indicates inadequate parameterization of baseflow dynamics (Clark et al., 2017). These findings align with previous studies showing conceptual models often struggle to capture nonlinear rainfall-runoff relationships, particularly in watersheds with strong seasonality (Fowler et al., 2020). The integrated models' ability to overcome these limitations supports recent arguments for combining physical and ML techniques in hydrological modeling (Shen et al., 2018).

From a practical perspective, these performance differences have significant implications for water resources management. The integrated models' accurate prediction of peak flows could substantially enhance flood early warning, reducing risks for the downstream ecosystem (Emerton et al., 2018). Similarly, their better representation of low-flow conditions enhances drought monitoring capabilities, supporting water allocation during dry seasons (Loon et al., 2016). Our findings contribute to ongoing discussions about optimal modeling approaches in data-scarce regions. Though some studies promote pure data-driven techniques, our results suggest integrated models provide significant advantages by maintaining physical interpretability while improving predictive accuracy (Nearing et al., 2020). This balance is valuable for operational water management, where both performance and interpretability are needed.

4. Conclusion

This study demonstrates that integration of the GR4J conceptual model with ML techniques, RF, ELM, XGB, and LSTM significantly enhances streamflow prediction accuracy in the Bilate River watershed. The integrated models consistently outperformed the standalone GR4J across multiple performance metrics, effectively combining GR4J's physical interpretability with ML's ability to capture nonlinear relationships and temporal dependencies. This integration successfully addresses key limitations of GR4J, particularly in modeling high daily streamflow variability and extreme flow events (high and low flows).

GR4J-LSTM and GR4J-XGB emerged as the most robust operational prediction, demonstrating superior generalization with test NSE values of up to 0.77 and KGE values reaching 0.86. These models effectively leveraged sequential learning (LSTM) and gradient-boosting precision (XGB) to improve temporal flow predictions. Meanwhile, GR4J-RF excelled in training performance (NSE = 0.87), though it exhibited a slight generalization gap during testing. GR4J-ELM provided a computationally efficient alternative, maintaining

comparative accuracy (test NSE = 0.74), making it a viable option for applications with limited computational resources.

These improvements align with prior studies promoting integrated modeling to balance physical process interpretability with ML adaptability. The integrated models' ability to preferably predict both low and peak flows supports critical water resource management applications in the Bilate River basin, such as flood and drought monitoring. Based on the findings, for achieving the highest predictive accuracy and robust generalization, particularly for capturing complex temporal patterns, GR4J-LSTM is the recommended choice, assuming sufficient computational resources and data are available. For a powerful and often more computationally efficient alternative to LSTM that also provides excellent accuracy and feature importance insights, GR4J-XGB is a recommended option. For situations where computational resources are very limited and rapid model implementation is a priority, GR4J-ELM offers a strong balance of acceptable accuracy and high computational efficiency.

Though this study confirms the value of integrated models, several challenges and opportunities remain for future research. Future work should focus on optimizing the computational demands of integrated models like GR4J-LSTM. A critical future direction is the assimilation of additional data sources, such as satellite-derived soil moisture and precipitation products, as well as climate reanalysis data, to enhance model input and potentially improve predictions, especially in data-scarce regions.

Acknowledgment

The authors extend sincere gratitude to the Ethiopian Meteorology Institute (EMI) and the Ministry of Water and Energy of Ethiopia (MoWE) for supplying critical weather and streamflow data. We also gratefully acknowledge Digital Earth Africa Sandbox for providing the computational platform.

Authors' Contribution

B.A.M.: Conceptualization, Data curation, methodology, software, Analysis, Investigation, visualization, writing original draft, writing review and editing; **F.A.Z.:** Conceptualization, methodology, Analysis, supervision, editing; **M.G.K.:** Conceptualization, methodology, Analysis, supervision, editing.

Funding

We did not receive any financial support.

Data availability statement

The data used for this research can be obtained from the corresponding author upon request.

Declarations

Ethics approval and consent to participate: not applicable

Consent for publication: not applicable

Competing interests: We declare that the authors have no competing interests.

References

Adane, G. B., Hirpa, B. A., Lim, C. H., & Lee, W. K. (2021). Evaluation and comparison of satellite-derived estimates of rainfall in the diverse climate and terrain of central and northeastern ethiopia. *Remote Sensing*, *13*(7). doi: 10.3390/rs13071275

Adnan, R. M., Liang, Z., Trajkovic, S., Zounemat-Kermani, M., Li, B., & Kisi, O. (2019). Daily streamflow prediction using optimally pruned extreme learning machine. *Journal of Hydrology*, *577*, 123981. doi: 10.1016/j.jhydrol.2019.123981

Al-Mukhtar, M., & Al-Yaseen, F. (2019). Modeling water quality parameters using data-driven models, a case study Abu-Ziriq marsh in south of Iraq. *Hydrology*, *6*(1). doi: 10.3390/hydrology6010021

Anshuman, A., Kunnath-Poovakka, A., & Eldho, T. I. (2021). Performance evaluation of conceptual rainfall-runoff models GR4J and AWBM. *ISH Journal of Hydraulic Engineering*, *27*(4), 365–374. doi: 10.1080/09715010.2018.1556124

Armstrong, W., Arsenault, R., Martel, J., Troin, M., Sabzipour, B., Brissette, F., & Mai, J. (2025). Improving multi-model ensemble streamflow forecasts by combining lumped , distributed and deep learning hydrological models. *Hydrological Sciences Journal*, *0*(0). doi: 10.1080/02626667.2025.2471430

Asgari, E., Mostafazadeh, R., & Talebi Khiavi, H. (2025). Projecting the Climate Change Impact on Water Yield in a Cold Mountainous Watershed, Ardabil. *Journal of the Earth and*

Space Physics, *50*(4), 165–177. doi: 10.22059/JESPHYS.2025.375570.1007601

Ayalew, A. D., Wagner, P. D., Tigabu, T. B., Sahlu, D., & Fohrer, N. (2023). Hydrological responses to land use and land cover change and climate dynamics in the Rift Valley Lakes Basin, Ethiopia. *Journal of Water and Climate Change*, *14*(8), 2788–2807. doi: 10.2166/wcc.2023.138

Bargam, B., Boudhar, A., Kinnard, C., Bouamri, H., Nifa, K., & Chehbouni, A. (2024). Evaluation of the support vector regression (SVR) and the random forest (RF) models accuracy for streamflow prediction under a data-scarce basin in Morocco. *Discover Applied Sciences*, *6*(6). doi: 10.1007/s42452-024-05994-z

Bartz-beielstein, E. B. T., & Zaefferer, M. (2023). Hyperparameter Tuning for Machine and Deep Learning with R. In *Hyperparameter Tuning for Machine and Deep Learning with R*. doi: 10.1007/978-981-19-5170-1

Beza, M., Tatek, E., Chala, M., & Moshe, A. (2024). Watershed hydrological responses to land use land cover changes at Bilata watershed, Rift Valley Basin, southern Ethiopia. *Water Practice and Technology*, *19*(4), 1455–1472. doi: 10.2166/wpt.2024.066

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

Chen, T., & Guestrin, C. (2016). *XGBoost : A Scalable Tree Boosting System*. 785–794. doi: 10.1145/2939672.2939785

Clark, M. P., Bierkens, M. F. P., Samaniego, L., Woods, R. A., Uijlenhoet, R., Bennett, K. E., … Peters-lidard, C. D. (2017). *The evolution of process-based hydrologic models : historical challenges and the collective quest for physical realism*. (1969), 3427–3440.

Cutler, A., Cutler, D. R., & Stevens, J. R. (2012). Random Forests. In C. Zhang & Y. Ma (Eds.), *Ensemble Machine Learning Methods and Applications* (pp. 157–175). Springer.

Dambré, K., Domiho, K. J., Faya, L., Vodounon, T., Sourou, H., & Ernest, A. (2024). *Rain-Flow Modelling Using the GR4J Model for Flood Risk Management in the Oti Watershed (Togo)*. 213–230. doi: 10.4236/ojmh.2024.144012

Darota, F. D., Borko, H. B., Adinew, C. D., & Edamo, M. L. (2024). Predicting sediment yield and locating hotspot areas in the Hamesa watershed of Ethiopia for effective watershed management. *Journal of Water and Climate Change*, *15*(4), 1855–1868. doi: 10.2166/wcc.2024.648

Dessie, M., Verhoest, N. E. C., Pauwels, V. R. N., Admasu, T., Poesen, J., Adgo, E., … Nyssen, J. (2014). Analyzing runoff processes through conceptual hydrological modeling in the Upper Blue Nile Basin, Ethiopia. *Hydrology and Earth System Sciences*, *18*(12), 5149–5167. doi: 10.5194/hess-18-5149-2014

Digital Earth Africa. (2025). *Digital Earth Africa Sandbox*.

Dinku, T., Funk, C., Peterson, P., Maidment, R., Tadesse, T., Gadain, H., & Ceccato, P. (2018). Validation of the CHIRPS satellite rainfall estimates over eastern Africa. *Quarterly Journal of the Royal Meteorological Society*, *144*(November 2017), 292–312. doi: 10.1002/qj.3244

Duc, L., & Sawada, Y. (2023). A signal-processing-based interpretation of the Nash-Sutcliffe efficiency. *Hydrology and Earth System Sciences*, *27*(9), 1827–1839. doi: 10.5194/hess-27-1827-2023

Emerton, R., Zsoter, E., Arnal, L., Cloke, H. L., Muraro, D., Prudhomme, C., … Pappenberger, F. (2018). *Developing a global operational seasonal hydro-meteorological forecasting system : GloFAS-Seasonal v1 . 0*. 3327–3346.

Enku, T., & Melesse, A. M. (2013). A simple temperature method for the estimation of evapotranspiration. *HYDROLOGICAL PROCESSES*, *2274*(November 2008), 2267–2274. doi: doi: 10.1002/hyp.9844

Fowler, K., Knoben, W., Peel, M., & Peterson, T. (2020). *Many Commonly Used Rainfall - Runoff Models Lack Long , Slow Dynamics : Implications for Runoff Projections Water Resources Research*. 1–27. doi: 10.1029/2019WR025286

Gers, F. A., & Cummins, F. (2000). Learning to Forget: Continual Prediction with LSTM.

Choice Reviews Online, *27*(09), 27-5238-27–5238. doi: 10.5860/choice.27-5238

Graves, A., Jaitly, N., & Mohamed, A. R. (2013). Hybrid speech recognition with Deep Bidirectional LSTM. *2013 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2013 - Proceedings*, 273–278. doi: 10.1109/ASRU.2013.6707742

Greff, K., Srivastava, R. K., Koutnik, J., Steunebrink, B. R., & Schmidhuber, J. (2017). LSTM: A Search Space Odyssey. *IEEE Transactions on Neural Networks and Learning Systems*, *28*(10), 2222–2232. doi: 10.1109/TNNLS.2016.2582924

Hah, D., Quilty, J. M., & Sikorska-senoner, A. E. (2022). Ensemble and stochastic conceptual data-driven approaches for improving streamflow simulations : Exploring different hydrological and data-driven models and a diagnostic tool. *Environmental Modelling and Software*, *157*(July), 105474. doi: 10.1016/j.envsoft.2022.105474

Hamzah, F. B., Hamzah, F. M., Razali, S. F. M., & Samad, H. (2021). A comparison of multiple imputation methods for recovering missing data in hydrological studies. *Civil Engineering Journal (Iran)*, *7*(9), 1608–1619. doi: 10.28991/cej-2021-03091747

Hao, R., & Bai, Z. (2023). Comparative Study for Daily Streamflow Simulation with Different Machine Learning Methods. *Water (Switzerland)*, *15*(6). doi: 10.3390/w15061179

He, S., Gu, L., Tian, J., Deng, L., Yin, J., Liao, Z., … Hui, Y. (2021). Machine learning improvement of streamflow simulation by utilizing remote sensing data and potential application in guiding reservoir operation. *Sustainability (Switzerland)*, *13*(7). doi: 10.3390/su13073645

Huang, G. Bin, Wang, D. H., & Lan, Y. (2011). Extreme learning machines: A survey. *International Journal of Machine Learning and Cybernetics*, *2*(2), 107–122. doi: 10.1007/s13042-011-0019-y

Huang, G. Bin, Zhu, Q. Y., & Siew, C. K. (2006). Extreme learning machine: Theory and applications. *Neurocomputing*, *70*(1–3), 489–501. doi: 10.1016/j.neucom.2005.12.126

Humphrey, G. B., Gibbs, M. S., Dandy, G. C., & Maier, H. R. (2016). A hybrid approach to monthly streamflow forecasting: Integrating hydrological model outputs into a Bayesian artificial neural network. *Journal of Hydrology*, *540*, 623–640. doi: 10.1016/j.jhydrol.2016.06.026

Ibrahim, A. H., Molla, D. D., & Lohani, T. K. (2024). Performance evaluation of satellite-based rainfall estimates for hydrological modeling over Bilate river basin, Ethiopia. *World Journal of Engineering*, *21*(1), 1–15. doi: 10.1108/WJE-03-2022-0106

Janjić, J., & Tadić, L. (2023). Fields of Application of SWAT Hydrological Model—A Review. *Earth (Switzerland)*, *4*(2), 331–344. doi: 10.3390/earth4020018

Kapoor, A., Pathiraja, S., Marshall, L., & Chandra, R. (2023). DeepGR4J: A deep learning hybridization approach for conceptual rainfall-runoff modelling. *Environmental Modelling and Software*, *169*(June), 105831. doi: 10.1016/j.envsoft.2023.105831

Khandelwal, A., Xu, S., Li, X., Jia, X., Stienbach, M., Duffy, C., … Kumar, V. (2020). *Physics Guided Machine Learning Methods for Hydrology*. Retrieved from http://arxiv.org/abs/2012.02854

Kodja, D. J., Mahé, G., Amoussou, E., Boko, M., Kodja, D. J., Mahé, G., … Paturel, J. E. (2023). *Assessment of the Performance of Rainfall-Runoff Model GR4J to Simulate Streamflow in Ouémé Watershed at Bonou ' s outlet (West Africa To cite this version : HAL Id : hal-04133007 Assessment of the Performance of Rainfall-Runoff Model GR4J to Simulate St*. 0–18. doi: 10.20944/preprints201803.0090.v1

Konapala, G., Kao, S., Painter, S. L., & Lu, D. (2020). *Machine learning assisted hybrid models can improve streamflow simulation in diverse catchments across the conterminous US OPEN ACCESS Machine learning assisted hybrid models can improve streamflow simulation in diverse catchments across the conterminous* .

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., & Herrnegger, M. (2018). Rainfall-runoff modelling using Long Short-Term Memory (LSTM) networks. *Hydrology and Earth System Sciences*, *22*(11), 6005–6022. doi: 10.5194/hess-22-6005-2018

Kumanlioglu, A. A., & Fistikoglu, O. (2019). Performance Enhancement of a Conceptual Hydrological Model by Integrating Artificial Intelligence. *Journal of Hydrologic Engineering*, *24*(11), 04019047. doi: 10.1061/(asce)he.1943-5584.0001850

Kunnath-Poovakka, A., & Eldho, T. I. (2019). A comparative study of conceptual rainfall-runoff models GR4J, AWBM and Sacramento at catchments in the upper Godavari river basin, India. *Journal of Earth System Science*, *128*(2), 1–15. doi: 10.1007/s12040-018-1055-8

Kwak, J., Han, H., Kim, S., & Kim, H. S. (2022). Is the deep-learning technique a completely alternative for the hydrological model?: A case study on Hyeongsan River Basin, Korea. *Stochastic Environmental Research and Risk Assessment*, *36*(6), 1615–1629. doi: 10.1007/s00477-021-02094-x

Le, X. H., Nguyen, D. H., Jung, S., Yeon, M., & Lee, G. (2021). Comparison of Deep Learning Techniques for River Streamflow Forecasting. *IEEE Access*, *9*, 71805–71820. doi: 10.1109/ACCESS.2021.3077703

Li, Xia, Xu, W., Ren, M., Jiang, Y., & Fu, G. (2022). Hybrid CNN-LSTM models for river flow prediction. *Water Supply*, *22*(5), 4902–4920. doi: 10.2166/ws.2022.170

Li, Xue, Sha, J., & Wang, Z. L. (2019). Comparison of daily streamflow forecasts using extreme learning machines and the random forest method. *Hydrological Sciences Journal*, *64*(15), 1857–1866. doi: 10.1080/02626667.2019.1680846

Liu, D., Jiang, W., Mu, L., & Wang, S. (2020). Streamflow Prediction Using Deep Learning Neural Network: Case Study of Yangtze River. *IEEE Access*, *8*, 90069–90086. doi: 10.1109/ACCESS.2020.2993874

Liu, J., Yuan, X., Zeng, J., Jiao, Y., Li, Y., Zhong, L., & Yao, L. (2022). *Ensemble streamflow forecasting over a cascade reservoir catchment with integrated hydrometeorological modeling and machine learning*. 265–278.

Liu, S., Wang, J., Wang, H., & Wu, Y. (2022). Post-processing of hydrological model simulations using the convolutional neural network and support vector regression. *Hydrology Research*, *53*(4), 605–621. doi: 10.2166/nh.2022.004

Loon, A. F. Van, Stahl, K., Baldassarre, G. Di, Clark, J., Rangecroft, S., Wanders, N., … Uijlenhoet, R. (2016). *Drought in a human-modified world : reframing drought definitions , understanding , and analysis approaches*. (1), 3631–3650. doi: 10.5194/hess-20-3631-2016

Ma, J., Sun, W., Yang, G., & Zhang, D. (2018). Hydrological Analysis Using Satellite Remote Sensing Big Data and CREST Model. *IEEE Access*, *6*, 9006–9016. doi: 10.1109/ACCESS.2018.2810252

Mada, Z. M., & Nannawo, A. S. (2023). Enhancing Understanding of Hydrologic Processes in the Shafe Watershed, Ethiopia. *Advances in Civil Engineering*, *2023*. doi: 10.1155/2023/5577851

Mei, Z., Peng, T., Chen, L., Singh, V. P., Yi, B., Leng, Z., … Xie, T. (2024). Coupling SWAT and LSTM for Improving Daily Streamflow Simulation in a Humid and Semi-humid River Basin. *Water Resources Management*, 397–418. doi: 10.1007/s11269-024-03975-w

Mohammadi, B., Safari, M. J. S., & Vazifehkhah, S. (2022). IHACRES, GR4J and MISD-based multi conceptual-machine learning approach for rainfall-runoff modeling. *Scientific Reports*, *12*(1), 1–21. doi: 10.1038/s41598-022-16215-1

Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., & Veith, T. L. (2007). Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the ASABE*, *50*(3), 885–900.

Nannawo, A. S., Lohani, T. K., Eshete, A. A., & Ayana, M. T. (2022). Evaluating the dynamics of hydroclimate and streamflow for data-scarce areas using MIKE11-NAM model in Bilate river basin, Ethiopia. *Modeling Earth Systems and Environment*, *8*(4), 4563–4578. doi: 10.1007/s40808-022-01455-x

Napiorkowski, J. J., Piotrowski, A. P., Karamuz, E., & Senbeta, T. B. (2023). Calibration of conceptual rainfall-runoff models by selected

differential evolution and particle swarm optimization variants. *Acta Geophysica*, *71*(5), 2325–2338. doi: 10.1007/s11600-022-00988-0

Nash, J. E., & Sutcliffe, J. V. (1970). River Flow Forecasting Through Conceptual Models - Part I - A Discussion of Principles. *Journal of Hydrology*, *10*(1970), 282–290.

Nearing, G. S., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J. M., … Gupta, H. V. (2020). *What Role Does Hydrological Science Play in the Age of Machine Learning ? Water Resources Research*. (2019). doi: 10.1029/2020WR028091

Nguyen, N. Y., Anh, T. N., Nguyen, H. D., & Dang, D. K. (2024). Quantile mapping technique for enhancing satellite-derived precipitation data in hydrological modelling: a case study of the Lam River Basin, Vietnam. *Journal of Hydroinformatics*, *26*(8), 2026–2044. doi: 10.2166/hydro.2024.225

Perrin, C., Michel, C., & Andréassian, V. (2003). Improvement of a parsimonious model for streamflow simulation. *Journal of Hydrology*, *279*(1–4), 275–289. doi: 10.1016/S0022-1694(03)00225-7

Sezen, C., Bezak, N., Bai, Y., & Šraj, M. (2019). Hydrological modelling of karst catchment using lumped conceptual and data mining models. *Journal of Hydrology*, *576*, 98–110. doi: 10.1016/j.jhydrol.2019.06.036

Sezen, C., & Partal, T. (2022). New hybrid GR6J-wavelet-based genetic algorithm-artificial neural network (GR6J-WGANN) conceptual-data-driven model approaches for daily rainfall–runoff modelling. *Neural Computing and Applications*, *34*(20), 17231–17255. doi: 10.1007/s00521-022-07372-5

Shen, C., Laloy, E., Albert, A., Chang, F.-J., Elshorbagy, A., Ganguly, S., … Tsai, W.-P. (2018). HESS Opinions: Deep learning as a promising avenue toward knowledge discovery in water sciences. *Hydrology and Earth System Sciences Discussions*, (April), 1–21. doi: 10.5194/hess-2018-168

Shen, C., Laloy, E., Elshorbagy, A., Albert, A., Bales, J., Chang, F. J., … Tsai, W. P. (2018). HESS Opinions: Incubating deep-learning-powered hydrologic science advances as a community. *Hydrology and Earth System Sciences*, *22*(11), 5639–5656. doi: 10.5194/hess-22-5639-2018

Shi, P., Chen, C., Srinivasan, R., Zhang, X., Cai, T., Fang, X., … Li, Q. (2011). Evaluating the SWAT Model for Hydrological Modeling in the Xixian Watershed and a Comparison with the XAJ Model. *Water Resources Management*, *25*(10), 2595–2612. doi: 10.1007/s11269-011-9828-8

Siami-Namini, S., Tavakoli, N., & Namin, A. S. (2019). The Performance of LSTM and BiLSTM in Forecasting Time Series. *Proceedings - 2019 IEEE International Conference on Big Data, Big Data 2019*, 3285–3292. doi: 10.1109/BigData47090.2019.9005997

Tao, H., Majeed, M., Abdulameer, H., Zounemat-, M., Heddam, S., Kim, S., … Falah, M. W. (2022). Neurocomputing Groundwater level prediction using machine learning models : A comprehensive review. *Neurocomputing*, *489*, 271–308. doi: 10.1016/j.neucom.2022.03.014

Tian, Y., Xu, Y. P., Yang, Z., Wang, G., & Zhu, Q. (2018). Integration of a parsimonious hydrological model with recurrent neural networks for improved streamflow forecasting. *Water (Switzerland)*, *10*(11). doi: 10.3390/w10111655

Tyralis, H., & Papacharalampous, G. (2019). A Brief Review of Random Forests for Water Scientists and Practitioners and Their Recent History in Water Resources. *Water*.

Wegayehu, E. B., & Muluneh, F. B. (2021). *Multivariate Streamflow Simulation Using Hybrid Deep*. *2021*(1).

Wegayehu, E. B., & Muluneh, F. B. (2022). Short-Term Daily Univariate Streamflow Forecasting Using Deep Learning Models. *Advances in Meteorology*, *2022*. doi: 10.1155/2022/1860460

Woldemariam, Y. A., Woldesenbet, T. A., & Alamirew, T. (2025). Evaluation and projection of CMIP6 simulations of climate variables for the Rift Valley Lakes Basin, Ethiopia. *Theoretical and Applied Climatology*, *156*(2). doi: 10.1007/s00704-025-05356-8

Xiang, Z., Yan, J., & Demir, I. (2020). A Rainfall-Runoff Model With LSTM-Based Sequence-to-Sequence Learning. *Water*

Resources Research, *56*(1), 1–17. doi: 10.1029/2019WR025326

Yang, J., Chen, F., Long, A., Sun, H., He, C., & Liu, B. (2024). Runoff simulation of the Kaidu River Basin based on the GR4J-6 and GR4J-6-LSTM models. *Journal of Hydrology: Regional Studies*, *56*, 102034. doi: 10.1016/j.ejrh.2024.102034

Yaseen, Z. M., Jaafar, O., Deo, R. C., Kisi, O., Adamowski, J., Quilty, J., & El-Shafie, A. (2016). Stream-flow forecasting using extreme learning machines: A case study in a semi-arid region in Iraq. *Journal of Hydrology*, *542*, 603–614. doi: 10.1016/j.jhydrol.2016.09.035

Zamani, M., Shrestha, N. K., Akhtar, T., Boston, T., & Daggupati, P. (2021). Advancing model calibration and uncertainty analysis of SWAT models using cloud computing infrastructure: LCC-SWAT. *Journal of Hydroinformatics*, *23*(1), 1–15. doi: 10.2166/hydro.2020.066

Zhang, R., Zen, R., Xing, J., Arsa, D. M. S., Saha, A., & Bressan, S. (2020). Hydrological Process Surrogate Modelling and Simulation with Neural Networks. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12085 LNAI*, 449–461. doi: 10.1007/978-3-030-47436-2_34

Zhang, Y., & Thorburn, P. J. (2022). Handling missing data in near real-time environmental monitoring : A system and a review of selected methods. *Future Generation Computer Systems*, *128*, 63–72. doi: 10.1016/j.future.2021.09.033