Journal of Hydrology 680 (2026) 136387

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Research papers

Enhancing river flow prediction accuracy in data scarce catchments using a time varying ensemble deep learning approach

Angel I. Tumushabe[a], Seith N. Mugume[a,*], Johanna Sörensen[b,c]

[a] Department of Civil and Environmental Engineering, Makerere University, P.O. Box 7062, Kampala, Uganda
[b] Department of Water Resources Engineering, Lund University, P.O. Box 118, Lund, Sweden
[c] United Nations University Hub on Water in a Changing Environment (WICE) at Lund University, United Nations University Institute for Water, Environment and Health (UNU-INWEH), Lund, Sweden

ARTICLE INFO

This manuscript was handled by Ashok Mishra, Editor-in-Chief, with the assistance of Jiang-jiang Zhang, Associate Editor

Keywords:
Bayesian model averaging
Data scarcity
Deep learning
Ensemble modelling
River flow prediction
Time-varying dynamic model averaging

ABSTRACT

Ensemble deep learning (EDL) methods offer an effective way to reduce model structural uncertainty and improve the reliability of river flow predictions, especially in catchments with limited data. This study evaluated the performance of a Time-Varying Dynamic Model Averaging (TV-DMA) ensemble for integrating heterogeneous machine learning and deep learning architectures, including FFNN, 1D-CNN, LSTM, CNN-LSTM and Diff-FFNN-LSTM models, for river flow prediction in two hydrologically distinct, data-scarce tropical catchments in the Upper White Nile Basin. Results show that the heterogeneous, first-order differential processing Diff-FFNN-LSTM model consistently achieved the highest individual-model river flow predictive performance (NSE $\geq$ 0.980), due to its ability to stabilise non-stationary river flow time series by mitigating boundary effects associated with decomposition methods, thereby improving representation of rapid flow transitions. Additionally, the TV-DMA resulted in more superior performance in both the Semliki (NSE = 0.910; PBIAS = −0.42%) and Tochi (NSE = 0.831; PBIAS = −0.34%) catchments when compared to the Bayesian Model Averaging (BMA) ensemble (Semliki: NSE = 0.848; PBIAS = −0.20% and Tochi: NSE = 0.710, PBIAS = −18.86%). BMA and TV-DMA also improved computational efficiency by 98–98.4% and 77–81%, respectively, relative to the process-based HEC-HMS model. The findings highlight the principal value of TV-DMA's dynamic update of individual model weights based on predictive error distributions, which enhances river flow predictive performance, particularly in the rainfall-runoff-dominated Tochi catchment. This EDL approach can improve the accuracy and efficiency of river flow modelling and support the development of flood forecasting and early warning systems in data scarce regions.

1. Introduction

Accurate and reliable prediction of river flows is essential for designing resilient water infrastructure in regions experiencing climate change-induced extremes such as flooding, drought and water scarcity (Ghimire et al., 2021; Liu et al., 2022; Mugume et al., 2024). Conventional physically based or conceptual hydrological models simulate runoff using mathematical representations of processes such as evapotranspiration, infiltration, effective rainfall and flow routing (Liu et al., 2022; Mugume et al., 2024; Young et al., 2017). Although widely applied, these models require extensive observed hydro-meteorological data, substantial calibration effort, considerable computational resources, higher modeller skill and may contain inherent uncertainties in model structure which limit their application in data-scarce catchments (Mount et al., 2016; Pareta, 2023). Consequently, the development and application of machine learning (ML) and deep learning (DL) models that are more computationally efficient and leverage empirical data to capture complex nonlinear relationships between hydro-meteorological inputs and outputs without requiring explicit representation of underlying physical processes is a subject of current research (Kratzert et al., 2019a; Mao et al., 2021; Mugume et al., 2024).

Despite these advances, individual ML and DL models are constrained by differences in model structure, predictive skill across hydrological conditions and requirement of extensive observed data for efficient training (Kratzert et al., 2018; Mugume et al., 2024; Staudinger et al., 2025). Feedforward neural networks have limited capability to

* Corresponding author.
E-mail address: smugume@gmail.com (S.N. Mugume).

https://doi.org/10.1016/j.jhydrol.2026.136387
Received 24 May 2026; Received in revised form 14 August 2026; Accepted 7 September 2026
Available online 9 September 2026

represent long-term temporal dependencies, while recurrent architectures such as LSTM are constrained by challenges of model interpretability and uncertainties in predicting extreme events (Baste et al., 2025; Kratzert et al., 2019a, 2019b, 2018; Mao et al., 2021).

Furthermore, hybrid architectures such as CNN-LSTM can improve the representation of temporal dynamics and peak flow predictions. However, their performance may decline in catchments characterised by high flows and extensive irrigation, suggesting that no single architecture performs consistently across contrasting river flow regimes (Kratzert et al., 2018; Liu et al., 2022; Pokharel and Roy, 2024; Staudinger et al., 2025).

To mitigate the challenges of individual and conventional hybrid ML and DL models architectures, recent studies have investigated the effectiveness of heterogeneous hybrid ML and DL models, bootstrapping and physics-based data augmentation in enhancing river flow predictive performance (Murungi et al., 2026; Saraiva et al., 2021; Tiwari and Chatterjee, 2011). Lin et al. (2021) applied a hybrid first-order differential Feed-Forward Neural Network–Long Short-Term Memory (Diff-FFNN-LSTM) model for hourly river flow prediction in the Andun River Basin, China. The Diff-FFNN-LSTM model outperformed the statistical, FFNN, LSTM, Diff-FFNN and Diff-LSTM models. The superior performance was attributed to the first-order differential pre-processing, which stabilised the non-stationary river flow series, avoided the backward induction and boundary effects associated with decomposition methods, and provided a simpler and more efficient framework for river flow prediction (Lin et al., 2021).

Recent studies have increasingly explored multi-source heterogeneous data fusion, use of larger sets of model input features, hybrid data-driven model architectures and ensemble learning to improve river flow prediction and minimise uncertainty (Ding et al., 2026). For example, recent work has demonstrated that integrating multi-source heterogeneous data that included hydro-meteorological, remote sensing, catchment attributes and vegetation data can improve representation of the spatial–temporal features and enhance the river flow predictive performance of deep learning models (Ding et al., 2026; Modi et al., 2025; Solanki et al., 2025). Similarly, multi-model frameworks that combine process-based hydrological models with data-driven ML and DL have improved streamflow predictions compared with individual models, although improvements under high-flow conditions remain limited (Solanki et al., 2025).

However, most recent studies have largely focused on either enriching the input data or combining model outputs rather than dynamically integrating heterogeneous deep learning architectures according to their time-varying predictive skill. In addition, the application of bootstrapping and physics-based data augmentation to improve individual ML and DL model river flow predictive performance in data-scarce catchments in Sub-Saharan Africa has shown potential, particularly when training data were limited, but its effectiveness may vary with data availability and catchment hydro climatic characteristics (Murungi et al., 2026).

The aforementioned studies demonstrate the value of enriching input and training datasets, while highlighting the need for dynamic ensemble modelling approaches (EMAs) that exploit complementary strengths among heterogeneous architectures under contrasting hydrological conditions. Consequently, there is a need to evaluate time-varying ensemble strategies that adapt the contributions of heterogeneous deep learning models to changing hydrological conditions.

Recent studies have increasingly applied conventional EMAs, including Weighted Averaging (WA), Simple Model Averaging (SMA), Linear Stacking (LS), Bayesian Model Averaging (BMA) and Extra Tree Regression (ETR), to combine predictions from multiple ML, DL and hybrid models and to reduce their individual performance limitations and uncertainty (Wegayehu and Muluneh, 2024). These approaches typically assign fixed model weights during calibration and retain them throughout the prediction period which allows the ensemble to combine complementary strengths of heterogeneous base learners and to improve river flow prediction accuracy and reliability (Ruan et al., 2022; Tumushabe and Mugume, 2025; Zhu and Willems, 2025).

Wegayehu & Muluneh (2024) applied WA, BMA and ETR to aggregate the Gated Recurrent Unit (GRU), LSTM, multi-layer perceptron (MLP), CNN-GRU, support vector regression (SVR), Lasso and Extreme Gradient Boosting (XGBoost) and linear regression models for riverflow prediction in the Baro Akobo catchment in Ethiopia. Their results showed that the Extra Tree Regression Super Ensemble outperformed the WA and BMA approaches, demonstrating the advantages of nonlinear ensemble aggregation for capturing complex hydrological relationships in river catchments.

Tu et al. (2024) utilised AutoGluon, a stacking ensemble machine-learning model, to reconstruct long-term natural river flows for the Pearl River Delta and evaluated the performance of these reconstructed inflows based on observed flow records during undisturbed periods. The reconstruction suggested superior performance of the AutoGluon model, when compared to individual Multiple Linear Regression (MLR), Support Vector Machine (SVM), and Random Forest (RF) models.

Qian et al. (2025) also applied a stacking ensemble that combined individual MLP, LSTM, and Transformer models for forecasting of river flows into the Hongze Lake in China. The study results suggested that the stacking ensemble model outperformed the base models, reducing forecast RMSE by 17–25% and increasing NSE by 0.12–0.20. The above EMAs assign fixed or globally optimised time-invariant weights to competing models, typically estimated during the training phase and applied unchanged during validation (Zounemat-Kermani et al., 2021).

While effective under stationary conditions, globally optimised fixed-weight ensembles cannot adapt to temporal changes in model skill associated with seasonal transitions, rapidly varying rainfall-runoff processes and non-stationary hydro-climatic conditions, particularly in tropical catchments with bimodal rainfall patterns (Ajami et al., 2007; Duan et al., 2007; Raftery et al., 2005; Ruan et al., 2022; Taye and Dyer, 2024). Recent studies have further expanded ensemble hydrological forecasting through the use of dynamic weighting, feature fusion and uncertainty-aware frameworks. Dynamic weighting approaches, such as Time-Varying Dynamic Model Averaging (TV-DMA), overcome the limitations of fixed-weight EMAs through updating model weights at each time step throughout the simulation period based on the prediction error distributions. This approach allows the ensemble to continuously adapt to changing hydro-meteorological conditions.

This dynamic weighting mechanism improves the representation of non-stationary rainfall-runoff processes and enhances simulation stability by leveraging the strengths of diverse base learner architectures (Ruan et al., 2022). The TV-DMA ensemble modelling approach is particularly relevant for minimizing prediction biases in tropical catchments, where hydrological responses are strongly influenced not only by base flows but also seasonal bimodal rainfall patterns and rapid runoff generation (Taye and Dyer, 2024).

Sheikh & Coulibaly (2025) developed a dynamic weighting approach based on time-series features (TSF-Ws) and demonstrated improved deterministic prediction of high and low flows when compared with conventional streamflow-based weighting schemes. Similarly, Xu et al. (2026) developed a Fourier-based Time-Varying Parameter Bayesian LSTM (F-TV-BLSTM) framework that dynamically updates model parameters while integrating the FourCastNet precipitation forecasts for probabilistic flood forecasting in the Yalong River Basin, China. Their results showed that dynamic parameter adaptation and multi-source data fusion improved NSE by up to 6.3% and reduced RMSE and MAE by more than 30%.

In another recent study, He et al. (2026) integrated multi-source hydroclimatic data with a hybrid Temporal Convolutional Network (NVWT) framework combining variational mode decomposition and wavelet threshold denoising for short- to long-term river flow forecasting. The NVWT model outperformed GRU and Transformer architectures and improved uncertainty quantification, highlighting the growing importance of feature fusion and signal processing in reliable prediction of river flows.

In a recent study, Ruan et al. (2022) applied the Time-Varying Dynamic Model Averaging (TV-DMA) approach to integrate Back Propagation Feedforward Neural Network (FFNN), Gated Recurrent Unit (GRU) and Long Short-Term Memory (LSTM) models for streamflow prediction in the Huaihe River Basin, China. The TV-DMA ensemble achieved superior predictive performance compared with the individual base learners. However, to our knowledge, the application of TV-DMA to a more heterogeneous ensemble comprising of individual and hybrid deep learning architectures, including differential processing, has not been investigated for river flow prediction in tropical catchments characterised by contrasting base flow- and rainfall-runoff-dominated regimes. This study therefore focuses on investigating the performance of TV-DMA in integrating the FFNN, 1D-CNN, LSTM, CNN-LSTM and Diff-FFNN-LSTM models, with the latter incorporating first-order differential processing to stabilise non-stationary flow sequences. This provides a distinct framework for evaluating whether dynamic weighting can exploit complementary model strengths across contrasting hydrological regimes.

In view of this, the specific objectives of this study were to:

a) Evaluate the performance of the individual FFNN, 1D-CNN and LSTM, as well as the hybrid CNN-LSTM and Diff-FFNN-LSTM models in simulating river flows in two hydrologically distinct catchments drained by the Semliki and Tochi Rivers, in the Upper White Nile Basin, East Africa;
b) Assess the predictive accuracy and reliability of TV-DMA ensemble modelling approach in integrating these models through benchmarking its performance against the conventional BMA approach and
c) Examine the model generalisation and transferability across two hydrologically distinct flow regimes through case studies of the Semliki and Tochi Rivers.

This paper introduces a novel approach that utilises the Time-Varying Dynamic Model Averaging (TV-DMA) ensemble framework to integrate heterogeneous individual and hybrid machine-learning and deep-learning architectures, including the heterogeneous Diff-FFNN-LSTM model, for predicting river flow in data-scarce tropical catchments. The Diff-FFNN-LSTM model enhances the ensemble by stabilising non-stationary river flow series through first-order differential processing, thereby reducing backward induction and boundary effects common in decomposition methods and improving prediction reliability and accuracy. The study also demonstrates the framework's applicability through tests on two hydrologically distinct tropical catchments in the Upper Nile Basin characterised by the base-flow-dominated and the rainfall-runoff-driven flow regimes. Evaluating the model's generalisation and transferability across these contrasting regimes provides practical evidence of its potential for reliable and accurate river flow prediction in other data-scarce regions with limited hydrometeorological data. These results provide a strong basis for developing future forecasting and flood warning systems in regions with limited data.

2. Methodology

The study developed and evaluated two ensemble deep learning (EDL) approaches, namely Bayesian Model Averaging (BMA) and Time-Varying Dynamic Model Averaging (TV-DMA), to enhance the reliability and accuracy of river flow predictions. The two ensemble modelling approaches were applied to integrate predictions from FFNN, LSTM, 1D-CNN, hybrid CNN-LSTM and Diff-FFNN-LSTM models developed for river flow prediction in the Semliki and Tochi catchments of the Upper White Nile Basin. The performance of individual ML and DL models was also benchmarked against physically based HEC-HMS simulations for both catchments which provided a consistent reference for model evaluation (Fig. 1).

The primary steps in the adopted study approach include collecting and preprocessing of catchment and hydrometeorological data, feature engineering and data normalisation, training, validation and testing of models with five different machine learning and deep learning architectures, creating two ensemble schemes, namely Bayesian Model Averaging (BMA) and Time-Varying Dynamic Model Averaging (TV-DMA), and evaluation of model performance. These methods were independently applied to the Semliki and Tochi River catchments in the Upper White Nile Basin, to assess the ensemble's generalisation capability.

2.1. Study area and data sets

Two tropical river catchments were selected for this study: Semliki and Tochi. The selection of these two catchments provides an opportunity to evaluate model performance across contrasting hydrological regimes: the baseflow-dominated, high-discharge Semliki River catchment versus the three-times-smaller, rainfall run-off driven, low-discharge Tochi River catchment. This approach enables assessment of ensemble generalisation capability across diverse tropical hydrological conditions.

2.1.1. Semliki river catchment

The Semliki catchment (Fig. 2a) is located along the Uganda-Democratic Republic of Congo border, covering an area of approximately 6,273 km^2. The catchment is drained by the 140 km transboundary Semliki River, which discharges into Lake Albert at coordinates 1.1881°N, 33.0210°E. The average annual rainfall ranges from about 900 mm in the low lying Semliki Delta to 2,600 mm in the Rwenzori Mountains. The River Semliki originates from Lake Edward (913 m above sea level, mASL) and flows through the Rift Valley to Lake Albert (619 mASL), draining the northern and western parts of the Rwenzori Mountain ranges, traversing a low-lying flat area whose geology mainly consists of rift valley sediments, gneisses and argillites and with gleysols as the predominant soil type (NELSAP-CU, 2020). The observed river flow data recorded between 1950 and 2011 indicate interannual variability, ranging from 50 to 150 m^3/s during low flow conditions to 200 to 500 m^3/s during flood conditions (NELSAP-CU, 2020).

Semliki exhibits a mixed hydrological regime: predominantly baseflow dominated under normal conditions but shifting to rainfall-driven during extreme storm events, a characteristic behaviour of tropical catchments with significant subsurface storage. The initial flow conditions in the Semliki River are influenced by the upstream Lake Edward through storage of excess inflows during rainy periods and maintenance of baseflows during dry periods (Russell and Johnson, 2006). In recent years, the lower Semliki River catchment has experienced an increasing frequency and magnitude of catastrophic fluvial flooding events (Mugume et al., 2026) which underscores the critical need for reliable flood early-warning systems to minimise anticipated flooding impacts, losses and damages.

Daily river flow data for the Semliki River were recorded at Bweramule gauging station (Station ID: 85205; 0.958°N, 30.187°E) from 1 March 1950 to 17 June 1978, when hydrological monitoring ceased due to political instability (MWE, 2013). Measurements resumed from 10 January 2006 to 31 December 2012 (Refer to Supplementary Fig. S1). Daily rainfall data were obtained from Kyembogo meteorological station (Station ID: 89300180; 0.683°N, 30.333°E) for the period 1990 to 2021 (Refer to Supplementary Fig. S2). For this study, a five-year overlapping period (2007 to 2011) with concurrent river flow and rainfall records was used for further hydrological modelling.

2.1.2. Tochi catchment

The Tochi catchment covers an area of approximately 2,270 km^2 and is drained by the Tochi River, with its outlet at 2.226°N, 32.342°E. The

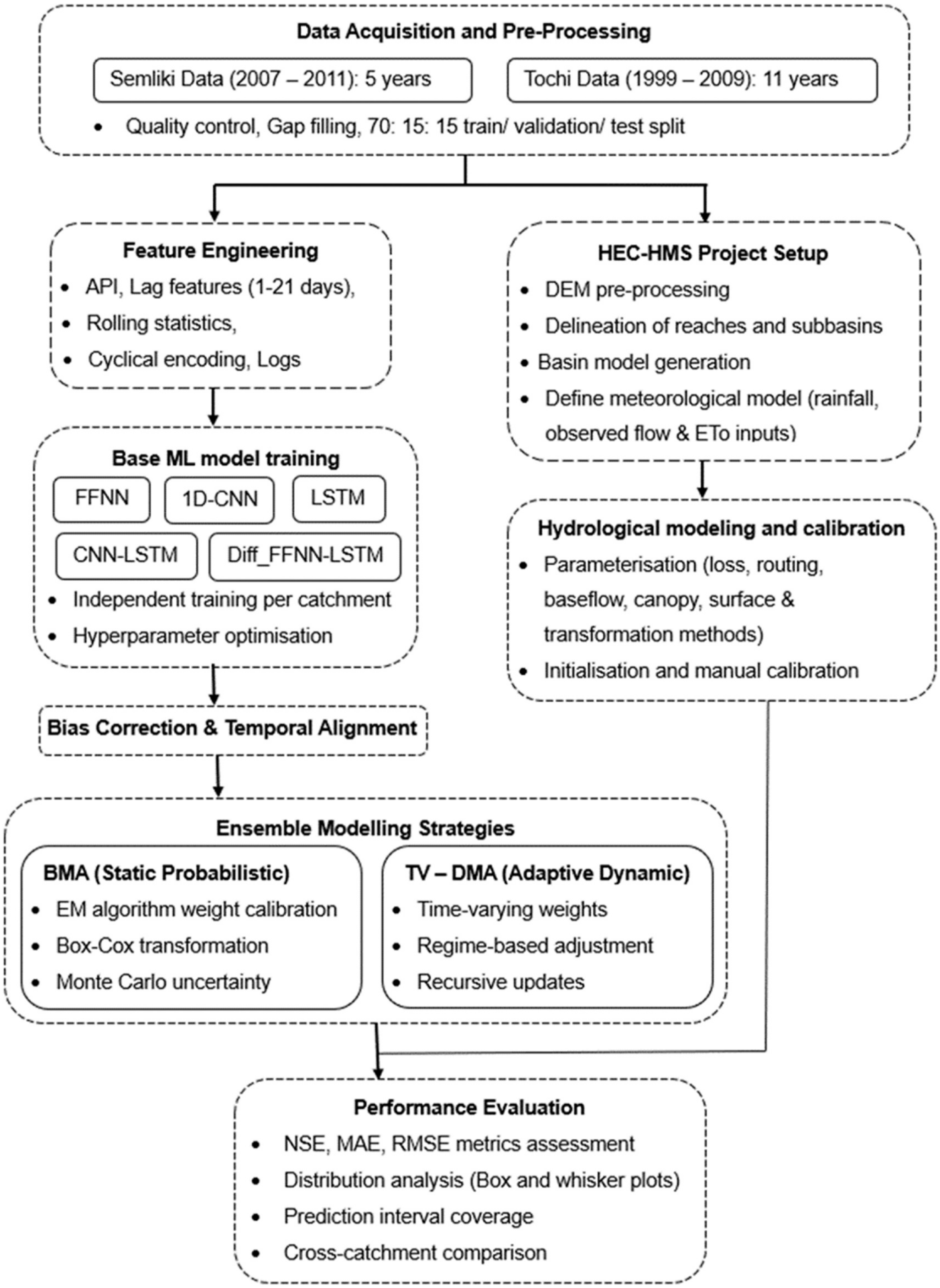

Fig. 1. Adopted study approach.

catchment is relatively flat with elevations ranging from 1,000 to 1,300 mASL, and it receives substantially higher precipitation than the Semliki catchment, with mean annual rainfall of 1,274 mm (range: 99–1,546 mm) and mean annual streamflow of 9.7 m^3/s. The catchment (Fig. 2b) exhibits a strongly rainfall-driven hydrological regime characterised by rapid and pronounced discharge responses following rainfall events. Recent trends indicate rainfall is more erratic, intensifying both flood and drought risks (JICA, 2014). Baseflow levels are minimal, often nearing zero in the dry season, except for some minor groundwater seepage in wetlands. The soils in the Tochi basin are largely ferrallitic (lateritic/ferrisols) sandy loams, which have low water storage capacity (rapid interflow) and erode easily under cultivation (MWE, 2018). This flashy response pattern reflects the relatively low soil permeability and limited storage capacity within the catchment, which contrasts sharply with the Semliki's mixed-regime behaviour. Observed daily rainfall and river flow measurements spanning nearly 11 years (1999–2009) were obtained from SW-R Tochi station (Station ID: 83212, 2.226°N, 32.34°E) and Gulu Meteorological Station (Station ID: 87320000, 2.78°N,

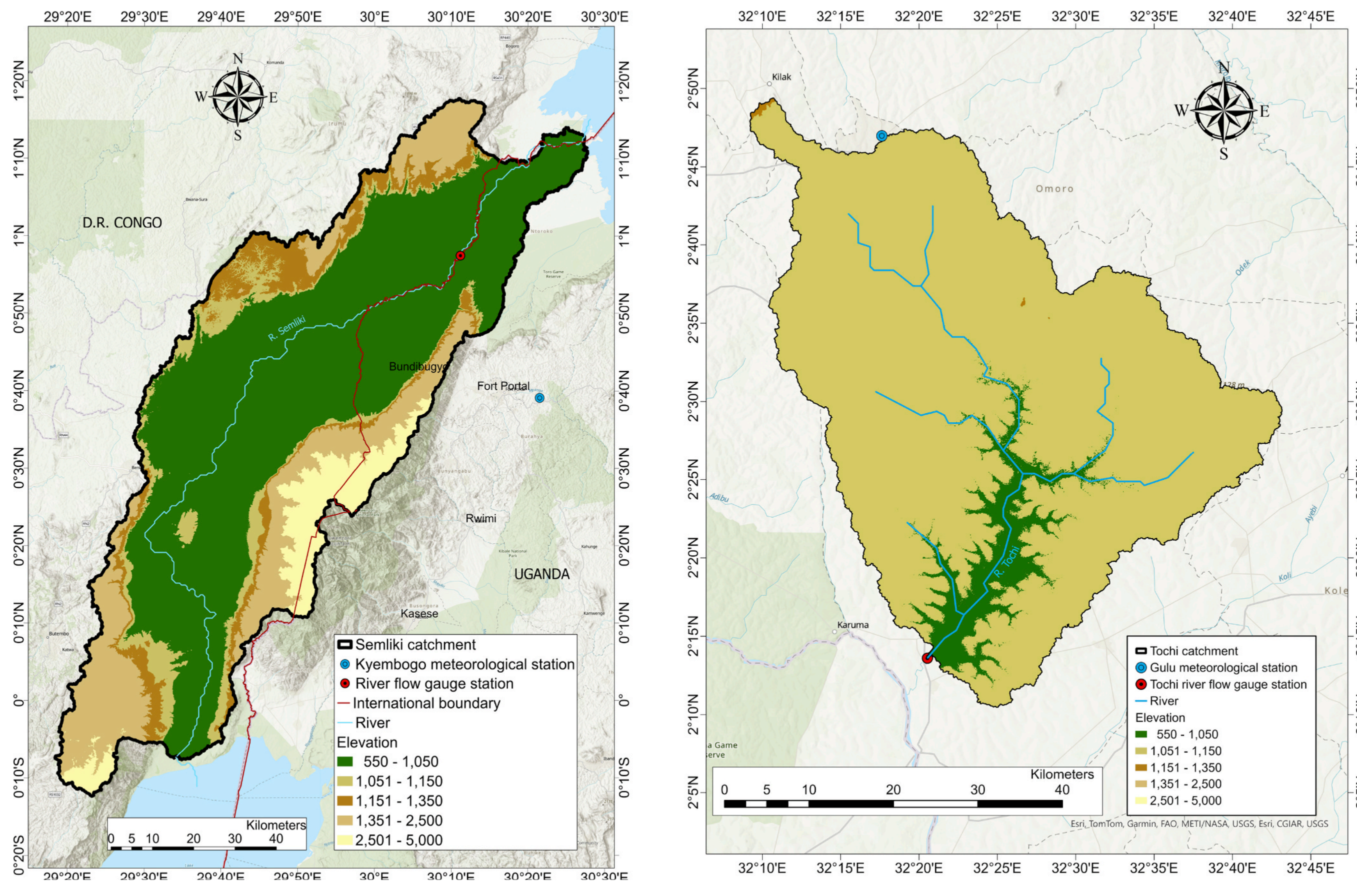

Fig. 2. Delineated (a) Semliki and (b) Tochi catchments. Note the difference in scale.

32.28°E).

2.1.3. *Data pre-processing and normalisation*

Hydro-meteorological time series were subjected to standard quality control procedures to ensure consistency and reliability prior to model development. Short data gaps of fewer than five consecutive days were infilled using linear interpolation to preserve temporal structure and autocorrelation in the series, a widely adopted approach in hydrological applications (Teegavarapu and Chandramouli, 2005). Predictor variables were subsequently normalised to improve numerical stability and learning efficiency of deep learning models. Min–max scaling was applied to variables with bounded, approximately uniform distributions, while robust scaling based on the interquartile range was used for skewed variables and those containing outliers, following established deep learning practice in hydrology (Kratzert et al., 2018). The dataset was split into training, validation, and test subsets using a 70:15:15 chronological split ratio, with all normalisation parameters derived exclusively from the training data and consistently applied to the remaining subsets to avoid information leakage.

2.1.4. *Physically-based hydrological modelling*

The Hydrologic Engineering Center – Hydrologic Modeling System (HEC-HMS) model was developed and applied to simulate rainfall-runoff processes in the Tochi and Semliki catchments. HEC-HMS is a comprehensive semi-distributed hydrologic simulation software developed by the US Army Corps of Engineers (USACE) for modelling precipitation-runoff processes in tree-like watershed systems (Bartles et al., 2023). The model is one of the most widely used in global hydrological modelling and was chosen for its flexible 'toolbox' approach, which allows for choice from multiple mathematical methods for infiltration losses, surface runoff transformation, base flow representation, and channel routing across diverse geographic regions and hydrologic conditions (Bartles et al., 2023).

The individual catchments were delineated into sub catchments from 30 m DEMs using the GIS capabilities of the HEC-HMS modelling environment to create the basin models. The meteorological models were created using pre-processed precipitation, discharge, and evapotranspiration data. Various parameterisation methods were employed to analyse the impact of canopy and land surface interference on hydrological processes. These methods included simple canopy, simple surface, linear reservoir analysis, Soil Moisture Accounting (SMA), Clarke unit hydrograph, and Muskingum routing, respectively, for assessing canopy interference, surface loss, baseflow, runoff loss, hydrograph transformation, and river routing. Initial values for model calibration were estimated based on soil characteristics extracted from the FAO digital soil map database, land use data generated using GIS based spatial analysis of Landsat imagery and the physical characteristics of each catchment.

2.1.5. *Feature engineering*

Feature engineering was performed to represent catchment memory, seasonality and delayed hydrological responses, using variables derived from available meteorological inputs. Seasonal variability was captured using cyclic temporal features based on the day of year, encoded with sine and cosine transformations to avoid artificial discontinuities at annual boundaries (Kratzert et al., 2019a; Schutte et al., 2024). Antecedent catchment wetness conditions were represented using the Antecedent Precipitation Index, which accounts for rainfall persistence through an exponential decay formulation (Descroix et al., 2002). Lagged rainfall variables and rolling precipitation statistics over short and extended windows were included to characterise both rapid runoff generation and longer-term storage effects (Jain and Srinivasulu, 2006).

For the rainfall-runoff dominated Tochi catchment, rainfall inputs were log-transformed by a log(1 + x) formulation to reduce skewness associated with extreme events and improve model training stability (e.g., Chadalawada et al. 2020).

2.2. Base machine learning and deep learning models

In this study, the FFNN, 1D-CNN, LSTM, hybrid CNN-LSTM, and Diff-LSTM-CNN models were developed and independently trained as base learners for the Semliki and Tochi catchments. During training, hyper parameter optimisation (HPO) was conducted using the random search method, which involves independent draws from a defined search space until the specified values giving optimal performance are achieved (Bergstra and Bengio, 2012; Kannengießer et al., 2025). The random search method has demonstrated superior efficiency compared to grid search because it explores the hyper parameter space more thoroughly, resulting in highly accurate model performance both empirically and theoretically with fewer computational resources (Bergstra and Bengio, 2012). All random searches were conducted in a deterministic Google Colab environment using Numpy's legacy global random number generator (numpy.random.seed) alongside Python's built-in random module, both of which implement the Mersenne Twister (MT19937) algorithm. A fixed seed value of 42 was used to initialise Numpy's random state (np.random.seed(42)) (Bergstra and Bengio, 2012; Hosseini et al., 2025).

The pre-defined set of architectural configuration values that dictated the HPO search space in each trial for testing, presented in Supplementary Information Table S1, was based on commonly tested literature-based ranges for machine learning. The sampling was done for a number of trials (40–50 trials) per model architecture, with validation loss/performance as the primary selection criterion, as it directly reflects model generalisation capacity and helps avoid overfitting (Apicella et al., 2026). For each trial, a configuration was drawn from the search space using the seeded RNG, the corresponding model architecture was constructed, trained and evaluated on the validation period. The optimal configuration for the final model performance evaluation on the unbiased test set was selected based on the minimum validation loss, achieved when the early stopping process halts training after the validation loss fails to improve for a specified number of epochs (patience). The underlying theory and model architecture of the investigated ML and DL models are described in detail in subsections 2.2.1, 2.2.2, 2.2.3, 2.2.4 and 2.2.5.

2.2.1. Feed-forward neural network (FFNN)

The FFNN algorithm has a non-recurrent model architecture and results in superior performance in capturing non-linear hydrological processes, including river flows (Mugume et al., 2024; Rahman et al., 2022; Tibangayuka et al., 2022). The architecture of the FFNN model used comprises an input layer, a hidden layer, and an output layer (Tibangayuka et al., 2022). The process of determining the ANN model architecture includes specifying the number of network layers, neurons in the layers, iterations, the learning rate, the activation function, the momentum of the coefficient, and the determination of the normalisation method (Koycegiz and Buyukyildiz, 2022). The mathematical representation of the ANN model based on FFNN is presented in Eq. (1) and was based on earlier studies on the standard artificial neuron (Rumelhart et al., 1986). A detailed description of the conventional ANN model architecture is provided in recent studies (Rahman et al., 2022; Tibangayuka et al., 2022; Turhan, 2021). The detailed description of the catchment-specific FFNN model configuration is provided in Supplementary Information Table S2.

$$y = f\left(\sum_{i=1}^{N} w_{ij}x_i + b_j\right) \tag{1}$$

where

y is the output of the j_{th} neuron;
w_{ij} represents the connection weight from the i_{th} neuron in the preceding layer to the j_{th} neuron of the current layer;
x_i represents the sequence of inputs, b_j is the bias of the j_{th} neuron, and;
f represents the activation function.

2.2.2. Long short-term memory (LSTM) model

LSTM network is a special Recurrent Neural Network (RNN) variant developed to solve the vanishing gradient and explosion problems during training of long sequences (Fu et al., 2022; Liu et al., 2022). In LSTM models, multiple repeated memory blocks are connected through a special chain structure. In each memory block, three special gate structures, i.e. forget, input and output gates, are designed to protect and control the cell state. These gating mechanisms enable LSTMs to maintain long-term dependencies while selectively forgetting irrelevant information, which is crucial for capturing both immediate rainfall-runoff responses and longer-term base-flow contributions. The underlying LTSM equations are specified in Eqs. (2), (3), (4), (5) and (6) (Liu et al., 2022). Detailed descriptions of LSTM models are provided in earlier studies (Kratzert et al., 2019b, 2018; Staudinger et al., 2025)

$$C_t = f_t \times C_{t-1} + i_t \times \widetilde{C}_t \tag{2}$$

$$f_t = \sigma\left(w_{xf} \times x_t + w_{hf} \times h_{t-1} + b_f\right) \tag{3}$$

$$i_t = \sigma(w_{xi} \times x_t + w_{hi} \times h_{t-1} + b_i) \tag{4}$$

$$C_t = \sigma(w_{xc} \times x_t + w_{hc} \times h_{t-1} + b_c)h_t = o_t \times \tanh(C_t) \tag{5}$$

$$o_t = \sigma(w_{xo} \times x_t + w_{ho} \times h_{t-1} + b_o)h_t \tag{6}$$

Where: f_t is the forget gate in the current period; i_t the input gate in the current period; O_t the output gate in the current period; $\widetilde{C}_t$ the abstract cell state in the current period; C_t and C_{t-1} the cell state in the current and previous periods; h_t and h_{t-1} the output of the hidden layer in the current and previous periods; x_t the input of the input layer in the current period; w_x the weight matrices and b the bias vectors and *tanh* the hyperbolic tangent function and σ the sigmoid function.

The detailed description of the catchment-specific LSTM model configuration is provided in Supplementary Information Table S3.

2.2.3. One-Dimensional Convolutional Neural Network (1D-CNN) model

Convolutional Neural Network (CNN) models were initially developed to process 2D signals, which are widely used in image and video processing (Fu et al., 2022; Kabir et al., 2020; Lecun et al., 2015). Kiranyaz et al. (2015) proposed the first 1D-CNN to handle sequential data (1D signals). 1D-CNNs efficiently extract spatial features from time-series data and can identify key patterns that influence river flow dynamics. CNNs consist of an input layer, a convolution with an activation function layer, a pooling layer, a fully dense connected layer, and a final output layer (Gao et al., 2024). Convolutional layers apply filters to local regions of the input data to extract features, analogous to how receptive fields in the visual cortex of the human brain respond to localised visual stimuli (Tsangaratos et al., 2023). The convolution and pooling layers extract and map features from the input data to the hidden layer. The fully dense connected layer integrates extracted features and maps them to the final output layer for prediction. The output feature map of a 1D convolution layer can be written using Eq. (7) (Ishida et al., 2024).

$$v_i^p = \sum_{n=1}^{N}\sum_{m=0}^{M-1} k_{mn}^p u_n(i+m) + b_p \tag{7}$$

Where:v_i^p is the output feature value at time step i for output channel p; i the time index of the output sequence ;k_{mn}^p the convolution kernel weight for position m, input channel n, and output channel p; $u_n(i+m)$

the input value of channel n at time $(i+m)$; N the number of input channels/observations ; M the kernel size (length of convolution filter) and b_p the bias term for output channel p. The detailed description of the catchment-specific 1D-CNN model configuration is provided in Supplementary Information Table S4.

2.2.4. CNN-LSTM model

The hybrid CNN-LSTM model uses CNN layers to extract features from streamflow time series, while LSTM networks utilise these features for temporal prediction (Ghimire et al., 2021; Pokharel and Roy, 2024). Hybrid models integrating CNNs and LSTMs have demonstrated superior accuracy compared to standalone models across various forecast horizons, as they combine CNN's feature extraction capabilities with LSTMs' temporal modelling strengths. The CNN-LSTM model utilises Equation (8) to predict river flows (Hu et al., 2024).

$$Q_t = LSTM(CNN(X_t)) \tag{8}$$

Where Q_t is the predicted river flow at time t;X_t is the input meteorological variable at time t; $CNN(\bullet)$ denotes the convolutional feature extraction process and $LSTM(\bullet)$ is the sequential learning process capturing temporal dependencies.

The detailed description of the catchment-specific CNN-LSTM model configuration is provided in Supplementary Information Table S5.

2.2.5. Diff-FFNN-LSTM model

The Diff-FFNN-LSTM is a novel architecture designed to address streamflow prediction through differencing, targeting the change in flow (ΔQ) rather than absolute values. In this research, the first-order difference (Eq. (9)) was applied to stabilise the river flow time series (Lin et al., 2021).

$$\Delta Q_{t+1} = Q_{t+1} - Q_t \tag{9}$$

Where: n is the length of data series; Q_t and Q_{t+1} are the recorded data at time t and $t+1$ respectively; ΔQ_{t+1} the first order difference based on Q_t and Q_{t+1} that are used as inputs for the prediction model.

The differencing strategy offers several theoretical and empirical

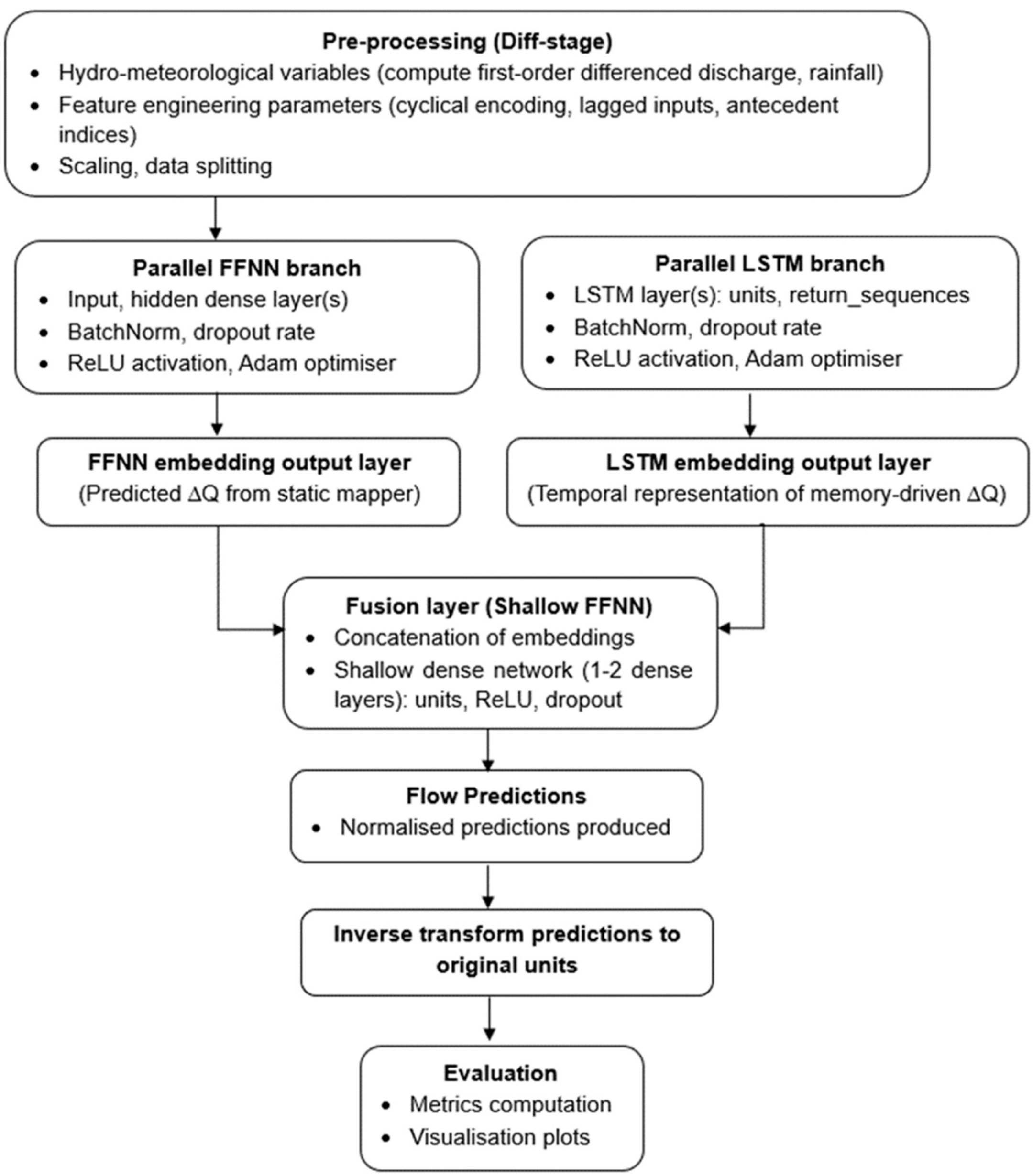

Fig. 3. Diff-FFNN-LSTM model workflow.

advantages for river flow prediction (Fig. 3). These include prediction of changes rather than absolute values which places emphasis on dynamic flow transitions such as rising limbs during storm events and recession curves following peak flows that are critical for flood forecasting applications. The strategy also shifts the model's focus toward rapid changes, which are characteristic of flood responses, and improves model generalisation to flow regimes not well represented in the training data. The detailed description of the catchment-specific Diff-FFNN-LSTM model configuration is provided in Supplementary Information Table S6.

2.3. Ensemble modelling strategies

The study implemented two averaging ensemble strategies: a regularised Bayesian Model Averaging (BMA) scheme, and the Time-Varying Dynamic Model Averaging (TV-DMA) method, following established ensemble modelling theory (Dion et al., 2021; Gelete et al., 2023; Raftery et al., 2010; Ruan et al., 2022).

2.3.1. Bayesian Model Averaging (BMA)

The BMA strategy blends predictions from multiple models by assigning weights based on their posterior probabilities, effectively integrating model performance with uncertainty (Raftery et al., 2005; Ajami et al., 2007). The predictive distribution for a hydrological variable of interest, y_t, relies on these posterior weights and is expressed by Eq. (10).

$$p(y_t \backslash D) = \sum_{k=1}^{K} w_k \times p(y_t \backslash M_k, D) \tag{10}$$

where w_k is the posterior probability (weight) of model k; $p(y_t \backslash M_k, D)$, the prediction from model, k; M_k, D, the kth candidate model; k the number of models and D the training data, with the constraint $\sum_{k=1}^{K} w_k \times p = 1$.

Predictions from the base models are first temporarily aligned to a common period by trimming initial time steps to accommodate the maximum lookback which ensures consistent training data across architectures with different sequence lengths.

Prior to BMA calibration, a two-stage pre-processing was applied. First, slope-intercept bias correction ($f_{corr,k} = a_k f_k + b_k$) was applied in the original discharge space only when it improved the NSE relative to the uncorrected model, k and $|a_k| > 0.05$. Here, f_k denotes raw predictions from the base model, k, while a_k and b_k are the slope and intercept of the corresponding linear regression, respectively. This criterion limited unnecessary bias correction for skilful models while supporting the Gaussian error assumption of normal likelihood used in EM-BMA. Secondly, the Box-Cox transformation, with the transformation parameter, λ optimised using observed river flows was applied to further satisfy the Gaussian errors required by the expectation–maximisation (EM) algorithm.

Model weights and competent variances were then estimated using the EM algorithm applied in the original prediction space, with log-sum-exp stabilisation and a variance floor (10^{-6} to 5×10^{-4}) to prevent numerical collapse during M-step updates. Post-EM, predictive variances were scaled by an empirically calibrated factor (1.0–2.0, targeting 90% coverage on calibration data) to achieve nominal reliability. In the E-step, posterior responsibilities z_{ik} were computed for each observation–model pair, and in the M-step, weights were updated using the Dirichlet-regularised formula specified in Eq. (11) (Vrugt; and Robinson, 2007).

$$w_k^{(new)} = \frac{\sum_{i=1}^{n} z_{ik} + \alpha}{n + \alpha K} \tag{11}$$

where z_{ik} represents the posterior probability that observation i comes from model k; α the Dirichlet regularisation parameter ; n the total number of observations/training data ; w_k the weight assigned to model k and k the number of models.

Effective model diversity was then quantified as $K_{eff} = \frac{1}{\sum w_k^2}$. The final operational weights were obtained using a hybrid scheme that combines pure BMA weights with validation-performance-based weights through a convex combination (Eq. (12)).

$$w_{final,\,k} = (1 - \beta) \times w_{BMA,k} + \beta \times w_{val,k} \tag{12}$$

where $\beta = 0.3$ is the adaptation strength; $w_{BMA,k}$ the pure BMA weights, and; $w_{val,k}$ the validation performance-based weights

Predictive uncertainty was characterised using Monte Carlo sampling from the fitted BMA mixture distribution (Huang and Merwade, 2023). For each lead time, component models were selected based on BMA weights, $w_{final,k}$, and realisations were drawn from their Gaussian predictive distributions, yielding an ensemble of river flow trajectories from which point forecasts, prediction intervals, and coverage statistics (e.g., 90% credible intervals) were derived. Ensemble predictions were calculated as weighted means of the bias-corrected and transformed forecasts. Predictive uncertainty was estimated using 10,000–20,000 Monte Carlo samples drawn from Gaussian mixture distribution, followed by inverse transformation and truncation at zero to ensure non-negative river flow values.

2.3.2. Time-Varying Dynamic Model Averaging (TV-DMA)

Time-Varying Dynamic Model Averaging (TV-DMA) extends Bayesian Model Averaging by allowing model weights to evolve in response to changes in predictive performance and hydrological regime. Unlike static averaging, which assumes time-invariant relationships between predictors and streamflow, TV-DMA updates the contribution of each base model as new observations become available, making it well-suited for non-stationary systems with seasonal shifts, structural changes, or evolving catchment conditions (Ruan et al., 2022). The ensemble prediction at time, t is given by Eq. (13).

$$\hat{y}_t = \sum_{k=1}^{K} w_{k,t} \times \left(f_{k,t} + \mu_k\right) \tag{13}$$

Where $w_{k,t}$ is the time-varying weight for model k; $f_{k,t}$ the raw prediction, and μ_k the bias correction term and k the number of models

The first stage of the TV-DMA workflow involves characterising residuals from each model using summary statistics and distributional fitting. Mean error, standard deviation, Kolmogorov–Smirnov statistics, and sum of squared errors are computed, and several candidate distributions (Gaussian, Student-t, and Cauchy) are fitted to the error series. This step forms the basis for future adjustments to reduce bias and quantify uncertainty.

In the second stage, model-specific additive bias corrections are applied, and initial static weights are derived from bias-corrected performance. Bias terms, μ_k, are estimated as the mean residual over the calibration period. (Gelete et al., 2023). Initial ensemble weights, $w_{k,0}$, are then set proportional to the inverse of the bias-corrected sum of squared errors (SSE), that is, $w_{k,0} \propto \left(\frac{SSE_{min}}{SSE_k}\right)^{\gamma}$ which ensures that higher-skill models received larger starting weights, while other models still retain non-zero influence.

The TV-DMA, with its dynamic weight evolution algorithm that adjusts $w_{k,t}$ according to recent absolute prediction errors, follows the outlined steps below while applying regime-adaptive factors (Duan et al., 2007; Raftery et al., 2010; Ruan et al., 2022). At each time step, t, the ensemble:

a) Applied flow-regime adjustment: low flows ($Q <$ 33rd percentile): best model $\times$ 1.2; high flows ($Q >$ 67th percentile): other models $\times$ 0.6, best model $\times$ 1.8; then normal flows: uniform) (e.g. Arsenault et al. 2015).

b) Computed a weighted prediction: $\hat{y}_t = \sum w_{k,t} \times f_{k,t}$
c) Calculated individual model errors: $\varepsilon_{k,t} = y_t - f_{k,t}$
d) Using a fast-learning rate, α, tracked model performance, $perf_{k,t}$, via exponential smoothing: $perf_{k,t} = \alpha|\varepsilon_{k,t}| + (1-\alpha)perf_{k,t-1}$,
e) Updated the weight, $w_{k,t+1} \propto \left(\frac{1}{perf_k + \varepsilon}\right)$
f) Enforces constraints: best-model floor and variance floor, then
g) Renormalises $\sum w = 1$

The percentile-based flow-regime adjustment is a novel approach that enhances the TV-DMA framework in this study and is conceptually consistent with the broad regime-aware ensemble, following work by Arsenault et al. (2015), and Oudin et al. (2006), which showed that different averaging schemes can improve performance differently across different flow states and highlighted the necessity of time-varying indices to maintain accuracy across the hydrograph's wet and dry periods.

The value of 1.2 was selected to reflect a moderate increase of approximately 20% in the weight of the best-performing model during low-flow conditions in order to strengthen its contribution without allowing it to dominate the ensemble. In addition, the values of 1.8 and 0.6 were selected to reflect controlled asymmetric adjustment during high-flow periods, comprising an 80% increase (1.8) for the best-performing model and a 40% reduction (0.6) for the remaining models. These were applied to minimise the tendency for under-prediction of peak flows in multi-model simulations and to limit unreliable individual model contributions to the ensemble (e.g. Wan et al., 2021).

The fast-learning rate provides gradual adaptation, preventing excessive volatility, while the regime factors and floors ensure stability across hydrological conditions, both of which are crucial characteristics for operational forecasting applications.

2.4. Model performance evaluation

Standard hydrologic performance metrics were employed to comprehensively assess the predictive performance and reliability of the individual ML and DL, hybrid and ensemble models (Moriasi et al., 2007 and Singh et al., 2007). The Nash Sutcliffe Efficiency (NSE) Index, Mean Absolute Error (MAE), Coefficient of Determination (R^2), Root Mean Square Error (RMSE) and Percent Bias (PBIAS) (Eqs. (14), (15), (16), (17) and (18), respectively) were computed and used to evaluate the performance of the developed models (Gupta and Kling, 2011; Legates and Mccabe Jr, 1999; Ritter and Muñoz-Carpena, 2013). The performance of the models, as assessed by the NSE criteria, was classified according to the following thresholds: "Very Good" (NSE $>$ 0.75), "Good" (0.65 $<$ NSE $\leq$ 0.75), or "Satisfactory" (0.50 $<$ NSE $\leq$ 0.65), with values $\leq$ 0.50 considered "Unsatisfactory". Following the guidelines by Singh et al. (2007) for the RSR logic (RMSE-observations standard deviation ratio), the error indices (RMSE and MAE) were judged relative to the observed data's standard deviation (σ_{obs}), where an error $\leq 0.5\ \sigma_{obs}$ was considered "Very Good". The rating was expanded to include a "Good" classification for errors between 0.5 σ_{obs} and 0.6 σ_{obs}, and a "Satisfactory" rating for errors between 0.6 σ_{obs} and 0.7 σ_{obs}. The errors exceeding 0.7 σ_{obs} were classified as "unsatisfactory".

$$NSE = 1 - \frac{\sum_{t=1}^{N}\left(Q_{obs,t} - Q_{sim,t}\right)^2}{\sum_{t=1}^{N}\left(Q_{obs,t} - \overline{Q}_{obs}\right)^2} \tag{14}$$

$$MAE = \frac{1}{N}\sum_{t=1}^{N}\left|Q_{obs,t} - Q_{sim,t}\right| \tag{15}$$

$$R^2 = \left[\frac{\sum_{i=1}^{N}\left(Q_{obs,t} - \overline{Q}_{obs}\right)\times\left(Q_{sim,t} - \overline{Q}_{sim}\right)}{\sqrt{\sum_{i=1}^{N}\left(Q_{obs,t} - \overline{Q}_{obs,t}\right)^2}\sqrt{\sum_{i=1}^{N}\left(Q_{sim,t} - \overline{Q}_{sim}\right)^2}}\right]^2 \tag{16}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{t=1}^{N}\left(Q_{obs,t} - Q_{sim,t}\right)} \tag{17}$$

$$PBIAS = \left[\frac{\sum_{i=1}^{n}\left(Q_{obs,t} - Q_{sim,t}\right)\times 100}{\sum_{i=1}^{n}\left(Q_{obs,t}\right)}\right] \tag{18}$$

where $Q_{obs,t}$ *and* $Q_{sim,t}$ are the observed and simulated river flows at time t; $\overline{Q}_{obs}$ is the mean observed river flows; $\overline{Q}_{sim}$ is the mean of the simulated river flows and N is the number of time steps.

The quantitative statistical performance measures above were supplemented with graphical analysis using box plots. Graphical techniques are essential in hydrological model evaluation as they provide a visual comparison of simulated and measured data, offering a first overview of model performance. They go beyond the aggregated indices (e.g., NSE), which can mask specific errors in variability and enable diagnostic evaluation of flow variability (Gupta et al., 2009). Specifically, box and whisker plots are used to examine data distributions and seasonal variations which allows for a robust assessment of how well the model reproduces the statistical characteristics of observed streamflow. This graphical approach provides a visual summary that includes the five conventional values: the extremes (maximum and minimum), the upper and lower hinges (quartiles), and the median (Mcgil et al., 1978). The box plots use a central mark within each box to represent the median (50% exceedance) forecast; the upper and lower boundaries of the box indicate the 25% and 75% exceedance forecasts, respectively; the whiskers (horizontal lines) above and below the box denote the 10% and 90% exceedance forecasts, respectively; and the circular markings illustrate the outliers. Flow duration curve (FDC) analysis was performed to complement the metric-and-hydrograph-based evaluation of the ensemble model forecasts by assessing the full river flow distribution and the ability of the ensemble to reproduce extreme flow conditions. A FDC expresses the percentage of time that a flow value is equalled or exceeded by ranking flows in descending order, and assigning each ranked value an exceedance probability using the Weibull plotting-position formula (Eq. (19)) (Ridolfi et al., 2020)

$$P_i = \frac{i}{n+1}\times 100 \tag{19}$$

where i is the rank after sorting flows in descending order, and n is the total number of observations.

The FDCs were developed for the unbiased test period ensemble predictions, and comparison was done against the corresponding observed flow, and contributing individual models (Westerberg et al., 2011).

To evaluate model predictive performance under extreme flow conditions and assess tail-fitting accuracy on FDC, two regime-specific metrics proposed by Yilmaz et al. (2008) were utilised: the percentage bias in the low-flow volume, %BiasFLV (Eq. (20)) and percentage bias in the high-flow volume, %BiasFHV (Eq. (21)).

$$\%\text{BiasFLV} = \frac{\sum_{l=1}^{L}\left[\log(Q_{Sl}) - \log(Q_{SL})\right] - \sum_{l=1}^{L}\left[\log(Q_{Ol}) - \log(Q_{OL})\right]}{\sum_{l=1}^{L}\left[\log(Q_{Ol}) - \log(Q_{OL})\right]} \times 100 \tag{20}$$

$$\%\text{BiasFHV} = \frac{\sum_{h=1}^{H} Q_{sh} - Q_{Oh}}{\sum_{h=1}^{H} Q_{Oh}} \times 100 \tag{21}$$

The %BiasFHV quantifies the model's ability to predict peak flow conditions, defined as flows exceeding the 2% exceedance probability of the FDC. In contrast, the %BiasFLV evaluates model performance during low flow conditions corresponding to the 70%–100% exceedance probabilities of the FDC (Yilmaz et al., 2008).

3. Results

3.1. Physically-based HEC-HMS model performance evaluation

The HEC-HMS model calibration and validation results for the Semliki catchment are adopted from Mugume et al. (2026) while in the present study, the HEC-HMS model was newly developed, calibrated and validated for the Tochi River catchment. For the Semliki HEC-HMS model, the Nash Sutcliffe Efficiency (NSE) and Percentage Bias (PBIAS) values during calibration were computed as 0.67 and −5.7%, respectively. Furthermore, the computed NSE and PBIAS values during validation were computed as 0.62 and + 10.6% respectively (Mugume et al., 2026), which suggested good performance of the developed HEC-HMS model during calibration and validation. The results also showed that the HEC-HMS model tended to underestimate the peak flows during the validation period (Fig. 4).

For the Tochi River catchment, NSE and PBIAS values during calibration and validation were computed as 0.60 and +3.3%, respectively. In addition, the NSE and PBIAS values during validation were computed as 0.60 and −20.5%, respectively. These NSE results suggest satisfactory performance of the developed Tochi HEC-HMS model during calibration. The NSE results also suggest satisfactory model performance during validation (Moriasi et al., 2007). Furthermore, the PBIAS results of +3.2% and −20.5% suggest satisfactory performance of the developed HEC-HMS model. The results also show that the HEC-HMS model tended to overestimate low flows and underestimated peak flows during both calibration and validation periods (Fig. 5).

3.2. Machine learning and deep learning model performance evaluation

3.2.1. Semliki river catchment

The performance of individual ML and DL models in predicting the Semliki River flows was investigated. Among individual models, LSTM achieved the best performance (NSE = 0.891). The hybrid Diff-FFNN-LSTM model produced the highest overall performance (NSE = 0.980) (Table 3). Models generally achieved higher NSE values in the baseflow-dominated Semliki catchment (0.842–0.980), likely due to inclusion of 3-day flow lags representing base flow persistence (Table 1). The study results also suggest that FFNN captured general temporal trends and low flows but underestimated peak flows. The 1D-CNN improved peak timing and magnitude, while LSTM demonstrated stable generalisation without overfitting. The hybrid CNN-LSTM improved the representation of peak flows and recession behaviour and generally outperformed individual CNN and LSTM models, although slight peak flow underestimation remained. The heterogenous hybrid differential Diff-FFNN-LSTM model showed strong agreement between observed and simulated flows across flow conditions (Fig. 6).

The box plot analysis shows that all models reproduced observed median flows across calibration, validation and testing, with errors below 5%, although FFNN showed minor systematic bias during the testing period (Fig. 7). In addition, the interquartile range ratios were close to 1.0 for all models, indicating good representation of flow variability. High flow representation was generally adequate but was underestimated by FFNN and LSTM models. Furthermore, the low flow predictions were most accurate during model testing. Representation of extreme outliers was strongest during model calibration and declined during the testing phase. The largest reduction in extreme outliers was observed for LSTM and 1D-CNN models (Fig. 7).

3.2.2. Tochi river catchment

The performance of five individual ML and DL models in simulating the Tochi river flows was investigated. The results suggest that the FFNN achieved the best performance (NSE = 0.698). The hybrid Diff-FFNN-LSTM model produced the highest overall performance (NSE = 0.984) (Table 2). Model performance in the rainfall-driven Tochi catchment ranged from NSE values of 0.600 to 0.984 (see Fig. 8).

The results further suggest that FFNN captured overall temporal patterns and mean flow conditions but underestimated peak flows and produced timing errors during transitional phases (Fig. 8). The 1D-CNN improved peak flow predictive performance and recession curve representation, reducing overall error relative to FFNN. LSTM further improved flow regime and peak representation, outperforming both FFNN and 1D-CNN. The CNN-LSTM hybrid combined local pattern recognition with sequential memory, thereby improving peak timing and the simulation of recession flow. The Diff-FFNN-LSTM model showed the best performance, accurately reproducing peak magnitudes, stable baseflow conditions and recession behaviour in the Tochi catchment.

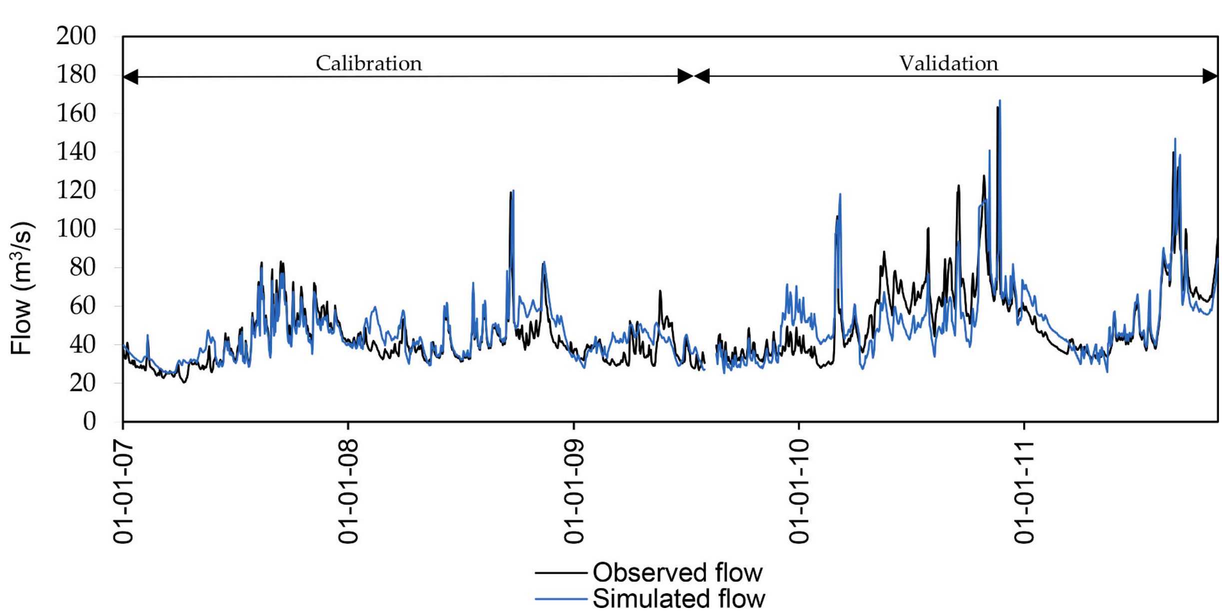

Fig. 4. Observed and simulated Semliki river flows during the calibration (2007–2009) and validation (2009–2011) periods (Mugume et al., 2026).

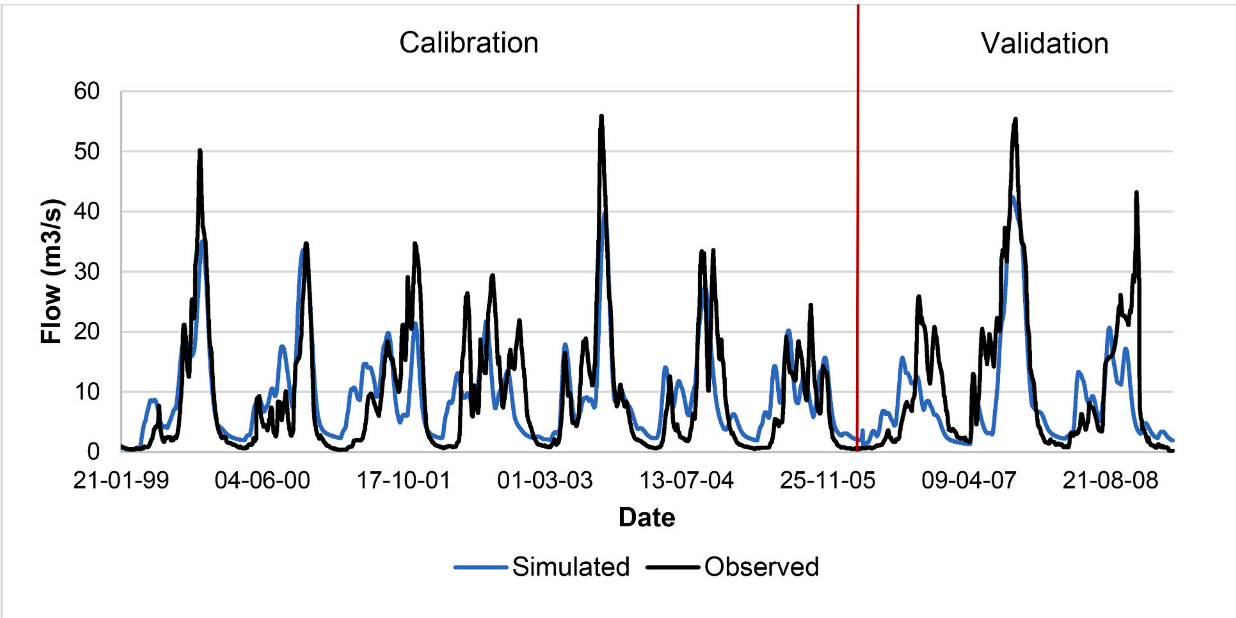

Fig. 5. Observed and simulated Tochi river flows during the calibration (1999–2005) and validation (2005–2008) periods.

Table 1
Performance of the developed FFNN, 1D-CNN, LSTM, CNN-LSTM and Diff-FFNN-LSTM models for the Semliki River catchment during model calibration, validation and testing.

Model	Performance Indicator	Calibration	Validation	Test
FFNN	NSE	0.750	0.448	0.842
	MAE	3.65	8.14	4.71
	RMSE	6.10	13.63	8.19
1D-CNN	NSE	0.875	0.693	0.887
	MAE	2.81	6.47	4.79
	RMSE	4.20	9.77	6.96
LSTM	NSE	0.886	0.550	0.891
	MAE	2.62	7.88	4.04
	RMSE	4.06	10.64	6.52
CNN-LSTM	NSE	0.859	0.641	0.900
	MAE	2.81	6.74	3.89
	RMSE	4.42	10.37	6.47
Diff-FFNN-LSTM	NSE	0.985	0.958	0.980
	MAE	1.00	2.40	1.15
	RMSE	1.42	3.76	2.07

Rating	Very Good	Good	Satisfactory	Unsatisfactory
Color				

The box plot analysis shows that the Diff-FFNN-LSTM model reproduced median flows with errors below 5% across all periods, indicating stable generalisation, while other models showed systematic overestimation exceeding 20% (Fig. 9). The river flow variability was generally preserved across models, with interquartile range ratios close to 1.0. However, the results also showed that the CNN-LSTM and LSTM models under-predicted the river flows. The Diff-FFNN-LSTM provided the most consistent variability representation.

The whisker analysis showed strong high-flow representation during calibration, but increasing underestimation during validation and testing. LSTM underestimated high flows across all periods, whereas Diff-FFNN-LSTM maintained near-observed high river flow values. Most models, except the Diff-FFNN-LSTM model, underestimated low flows. Extreme high-flow outliers during testing were poorly captured by all models except Diff-FFNN-LSTM, indicating that other models had limited ability to predict the variability of extreme flood flows (Fig. 9).

3.3. BMA ensemble modelling strategies

3.3.1. Semliki river catchment

The BMA ensemble model implementation involved pre-processing and calibration steps to ensure robust probabilistic forecasts (Ajami et al., 2007; Raftery et al., 2005). To satisfy the Gaussian error assumption required by the EM algorithm, the Box-Cox transformation applied optimal power transformation parameter, $\lambda = -0.703$ for the Semliki catchment. This transformation normalised the error distributions of all base model predictions. Bias correction was applied in the original space using linear regression between the base model predictions and the observed set to remove systematic errors that could distort likelihood-based weighting (Table 3).

The EM algorithm converged after 414 iterations with a log-likelihood of 9.95×10^{-8} and an effective model count, K_eff, of 2.53 (out of 5 models). The convergence showed successful probabilistic model combination, while the low K_eff indicated more weight concentration toward superior-performing models (Fig. 10).

The BMA weight distribution results suggest that the FFNN model (variance = 3.7×10^{-5}) was the most dominant model (BMA weight =

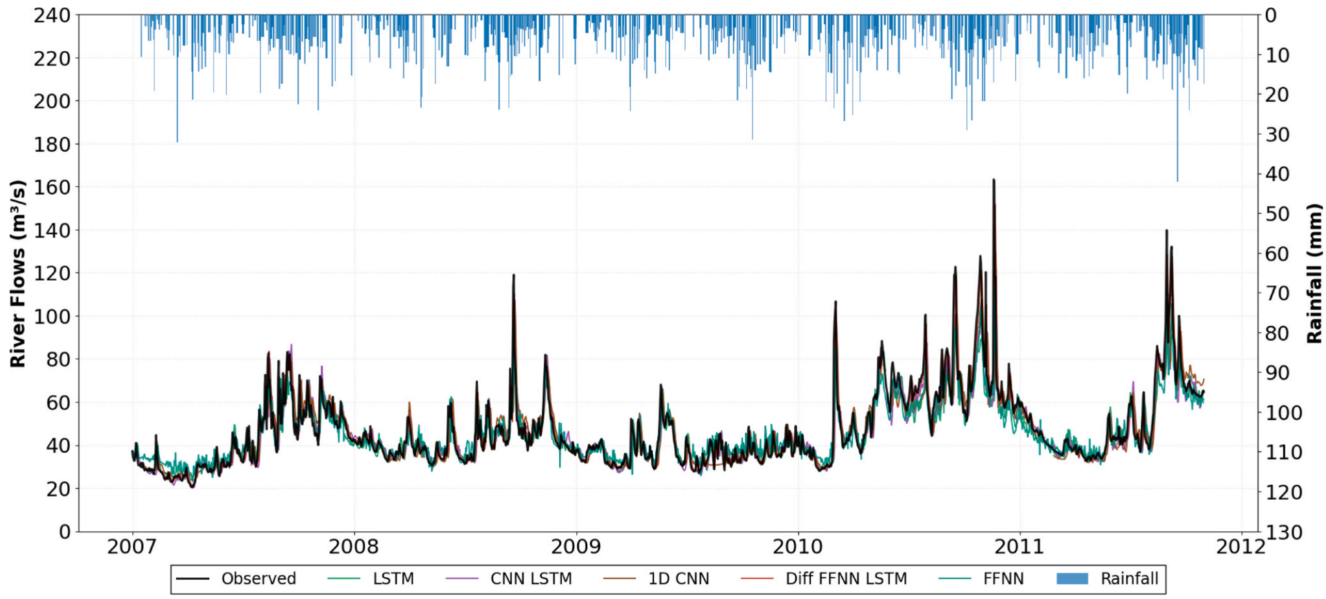

Fig. 6. Daily rainfall, observed and simulated Semliki river flows during model calibration, training and testing periods for LSTM, CNN-LSTM, 1D-CNN, Diff-FFNN-LSTM, and FFNN models.

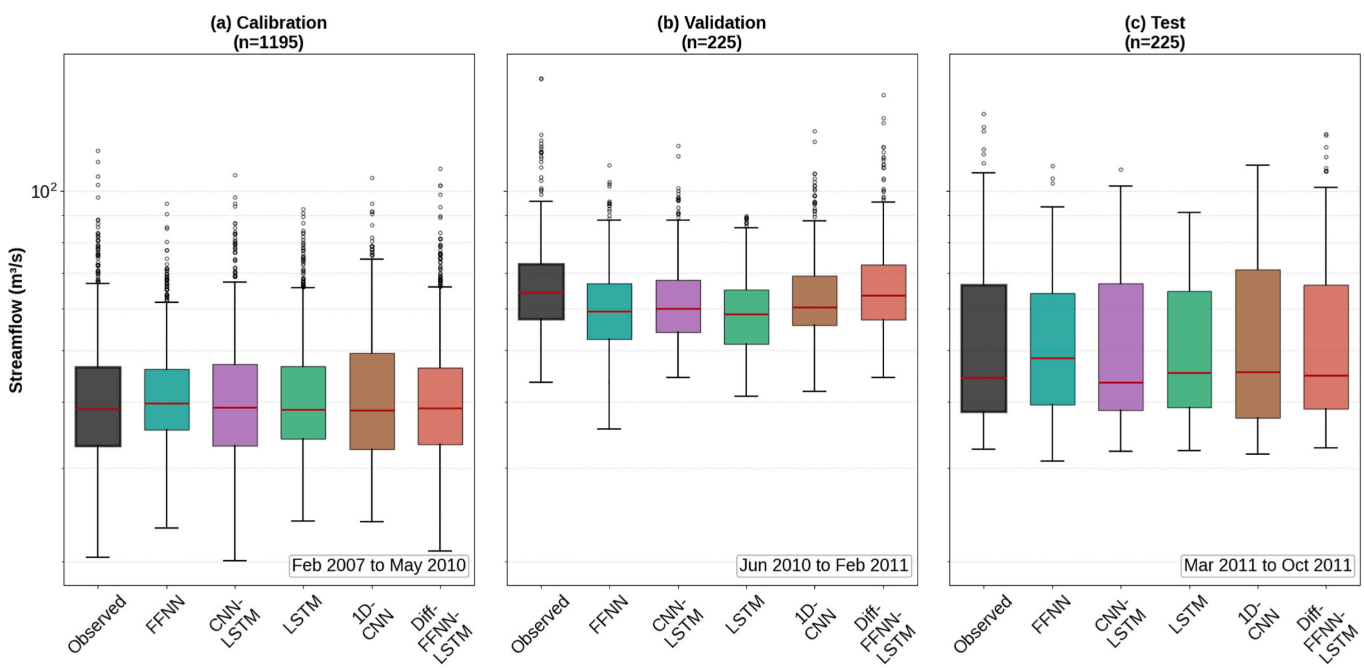

Fig. 7. Box plot showing the observed simulated Semliki River flow distribution during ML, DL and hybrid model calibration, validation, and testing. The box plots show the simulated flow distribution characteristics for each model. The box range represents the interquartile range ($Q_1 - Q_3$), and the middle line indicates the median.

85.6%), followed by the 1D-CNN model (6.4%). On the contrary, the Diff-FFNN-LSTM model received minimal weight (1.4%). The BMA's adaptive weighting demonstrates the BMA's ability to automatically reduce the influence of poorly performing models.

The study results also suggest that the BMA ensemble improved performance relative to most individual models (Table 4), achieving NSE = 0.848 and the lowest bias (−0.2%) among all models, demonstrating strong simulation skill for the Semliki catchment (Fig. 11). The simulated peak discharge is close to the observed, but with a slight time lag for all peaks. The model also showed strong uncertainty quantification, with 90% and 95% prediction interval coverages of 94.2% and 97.8% (Fig. 11). The time series analysis indicates that BMA combines complementary model strengths and widens the prediction interval during peak flow events which suggests higher forecast uncertainty during extreme conditions.

The flow duration curve, presented in (Fig. S3), showed that the ensemble model closely reproduced the catchment's flow regime, with good agreement between the observed and simulated river flows across both the upper and lower flow tails of the curve. Furthermore, the curve generated by the ensemble model more effectively replicated the observed trend across the entire flow range when compared with the base models which demonstrated its superior performance.

Table 2
Performance of the developed HEC-HMS, FFNN, 1D-CNN, LSTM, CNN-LSTM and Diff-FFNN-LSTM models for the Tochi River catchment during model calibration, validation and testing.

Model	Performance Indicator	Calibration	Validation	Test
HEC-HMS	NSE	0.603	0.592	
	MAE	4.65	5.36	
	RMSE	6.31	7.73	
FFNN	NSE	0.582	0.422	0.698
	MAE	4.26	6.85	3.93
	RMSE	6.48	9.01	6.76
1D-CNN	NSE	0.674	0.520	0.650
	MAE	3.74	6.43	3.11
	RMSE	5.73	8.33	5.80
LSTM	NSE	0.744	0.536	0.600
	MAE	3.36	6.11	3.65
	RMSE	5.06	8.09	6.57
CNN-LSTM	NSE	0.754	0.662	0.694
	MAE	3.71	5.61	4.18
	RMSE	4.98	6.89	6.80
Diff-FFNN-LSTM	NSE	0.981	0.980	0.984
	MAE	0.41	0.36	0.38
	RMSE	0.63	0.61	0.79

3.3.2. Tochi river catchment

The implementation of the BMA ensemble involved systematic pre-processing to ensure robust probabilistic forecasts in the hydrological environment. A Box-Cox transformation with optimal $\lambda = 0.099$ was applied to normalise error distributions. The EM algorithm converged after 268 iterations which indicated greater difficulty in achieving stable weight estimates due to the variable flow regime. The final weight distribution showed a strong concentration toward the Diff-FFNN-LSTM (84.9%) (Fig. 12).

The remaining weight was distributed among CNN-LSTM (9.6%), LSTM (5.4%), FFNN (0.1%), and 1D-CNN (0%). The ensemble recorded an effective model count K_eff, of 1.43 (out of 5), indicating low model diversity. However, the non-zero weights of CNN-LSTM and LSTM still showed that BMA's likelihood formulation did not fully converge on the best-performing base model but retained contributions from multiple models.

The ensemble slightly outperformed the best individual model across several metrics (Table 5), with NSE improvement and RMSE reduction of 0.42% and MAE reduction of 7%, indicating that modest improvement resulted from the ensemble model combination. The uncertainty analysis showed well-calibrated prediction intervals, with 90% and 95% coverage rates of 93.1% and 95.0%, which closely match nominal targets for reliable probabilistic operational hydrological forecasting. The study results also suggest that the prediction intervals widened during high flow events and narrowed during base flow periods (Fig. 13), which reflected the expected uncertainty patterns under extreme conditions. In Fig. 13, the time series comparison shows strong ensemble model performance in capturing rapid flow dynamics, including rising limbs, peak timing and magnitude, and recession behaviour. Despite reproducing the observed flow distribution reasonably well, with good agreement in normal and low flows, as seen in the FDC (Fig. S4), the ensemble tended to under predict the extreme peaks in the left tail of the curve. The individual model curves exhibited a similar pattern, remaining clustered around the ensemble curve, which shows that the ensemble combines the complementary strengths of the base model while improving overall consistency with the observed flow regime.

3.4. TV-DMA ensemble modelling

3.4.1. Semliki river catchment

The TV-DMA ensemble was implemented to enable adaptive model weighting that responds to changing hydrological conditions. Based on an initial performance screening (NSE $>$ 0.7), all five models were selected for the combination so as to balance model diversity with computational efficiency. The static initial weights computed based on inverse bias-corrected sum of squared errors (SSE) were as follows: FFNN: 0.2578 (26%); LSTM: 0.1496 (15%); 1D-CNN: 0.1440 (14%); CNN-LSTM: 0.1585 (16%); and Diff-FFNN-LSTM: 0.2902 (29%).

The dynamic weight update mechanism incorporated two key features: a regime-adaptive learning rate ($\alpha\_fast = 0.15$) to enable rapid weight adjustment in response to changing model performance, and a minimum weight constraint (0.6) to ensure that the superior model maintained dominant influence. The weight evolution analysis represents the adaptive nature of the ensemble (Fig. S1), with the FFNN model generally dominating, especially during high flow periods (August–October), the CNN-LSTM given slightly more weight during rising flows (June–July), the 1D-CNN and Diff-FFNN-LSTM with higher weights during low flows (April–May), and stable weighting during the recession period (October).

The results of the ensemble modelling suggest that the TV-DMA exhibited strong adaptive performance, with near-zero bias and accuracy close to the results obtained using the hybrid Diff-FFNN-LSTM model (Table 6). The ensemble's lower bias compared to the

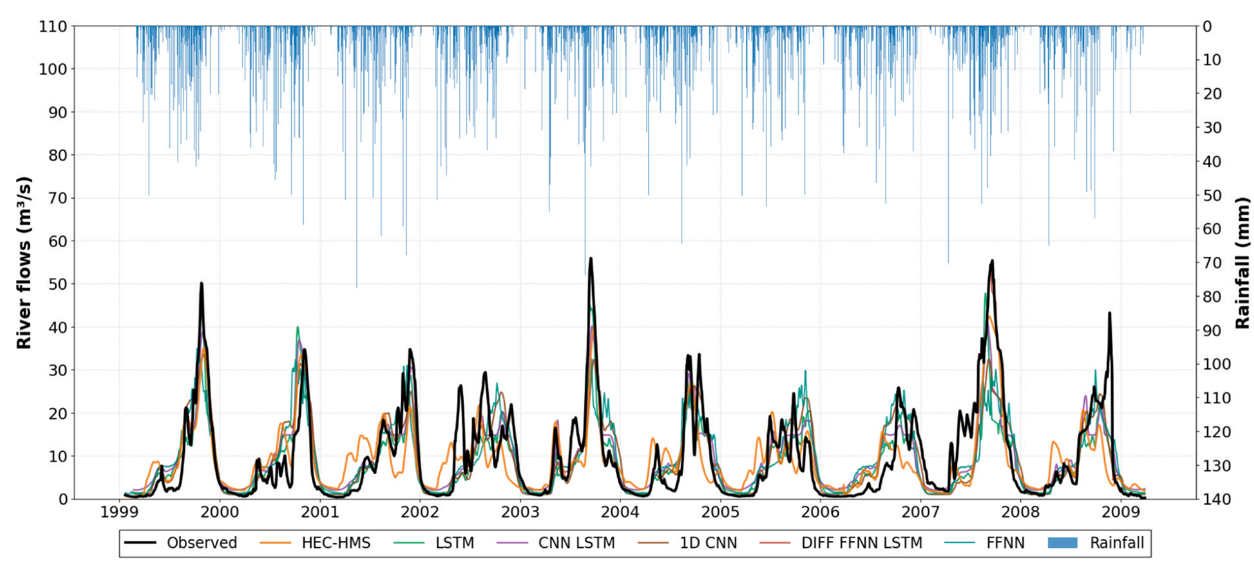

Fig. 8. Daily rainfall, observed and simulated Tochi River flows during model calibration, training and testing periods for HEC-HMS, LSTM, CNN-LSTM, 1D-CNN, Diff-FFNN-LSTM, and FFNN models.

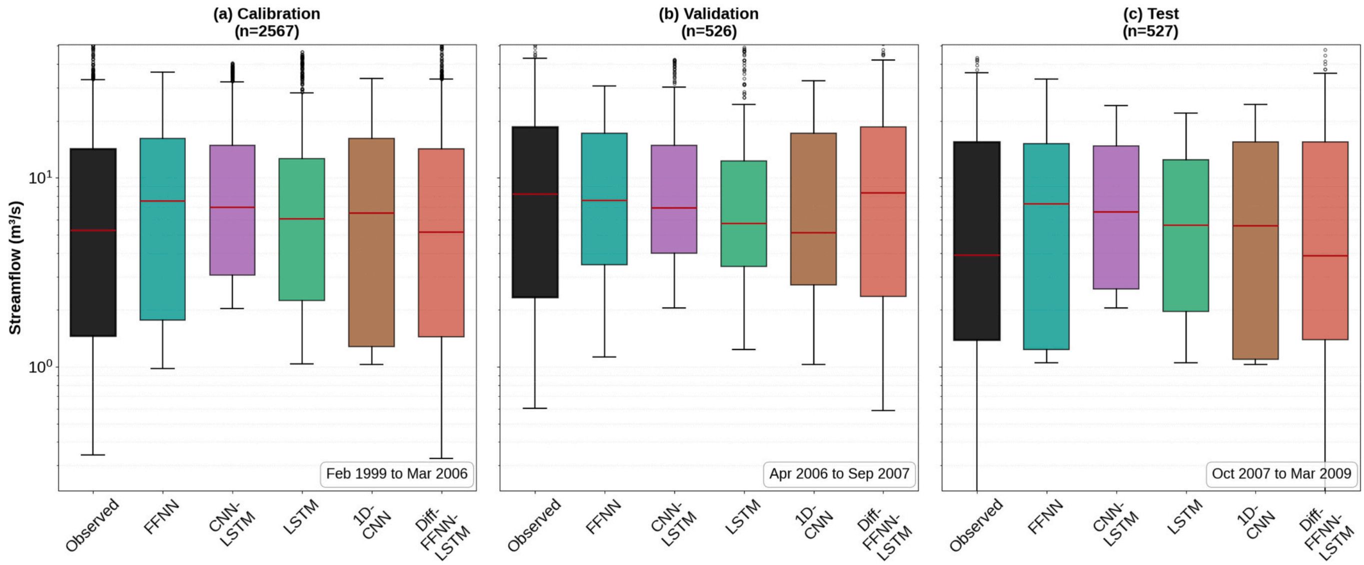

Fig. 9. Box plot showing the observed simulated Tochi River flow distribution during ML, DL and hybrid model calibration, validation, and testing. The box plots show the simulated flow distribution characteristics for each model. The box range represents the interquartile range ($Q_1 - Q_3$), and the middle line indicates the median.

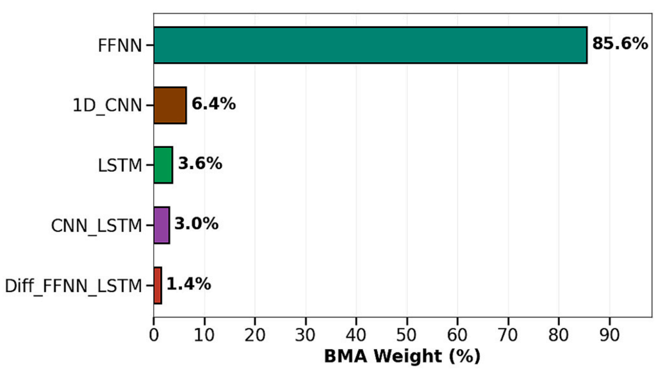

Fig. 10. Final BMA weights and model contributions for the Semliki River catchment.

Table 3
Bias correction results – Semliki BMA.

Model	Slope	Intercept	NSE before BC	NSE after BC	ΔNSE
FFNN	1.15	−4.49	0.745	0.769	+0.0244
LSTM	1.08	−2.03	0.673	0.683	+0.0095
1D-CNN	0.92	3.95	0.669	0.675	+0.0054
CNN-LSTM	0.95	2.92	0.657	0.660	+0.0036
Diff-FFNN-LSTM	0.91	4.25	0.745	0.752	+0.0075

Table 4
BMA ensemble performance comparison for the Semliki River catchment.

Model	PBIAS (%)	NSE	RMSE	MAE (m^3/s)
FFNN	−1.85	0.847	8.37	5.11
LSTM	−0.94	0.771	10.25	6.32
1D-CNN	−3.05	0.730	11.12	6.73
CNN-LSTM	−4.03	0.749	10.74	6.42
Diff-FFNN-LSTM	−1.99	0.841	8.55	4.88
BMA ensemble	−0.20	0.848	8.34	4.94

individual models demonstrates the effectiveness of the continuous adaptive weighting utilised by the TV-DMA method.

The hydrograph analysis showed good performance in capturing both the timing and magnitude of flow variations of the catchment (Fig. 14). This predictive accuracy was further supported by the FDC analysis (Fig. S5), which showed that the ensemble curve reproduced the general trend of the flow regime and reduced divergence from the observed by the base models, with only slight underestimation of extreme peaks.

3.4.2. Tochi river catchment

The TV-DMA ensemble was implemented using targeted model selection to balance performance and computational efficiency. Based on preliminary screening (NSE > 0.5), the CNN-LSTM and Diff-FFNN-LSTM models were retained while weaker models were excluded. The results of the weight evolution (Supplementary Information Fig. S11) showed dynamic redistribution, with Diff-FFNN-LSTM model dominating major peak events (January to February 2008), the CNN-LSTM model gaining influence during transition periods, and stable dominance of the best model during moderate flows (March to May 2008).

The study results also suggest that the TV-DMA ensemble outperformed all individual models (Table 5), achieving an NSE value of 0.831, which was higher than the NSE values for the BMA (ΔNSE = +0.047) and the Diff-FFNN-LSTM models (ΔNSE = +0.012) (Table 7). The ensemble maintained near-zero bias and improved log-space NSE by 1.6%, indicating improved low-flow prediction. The comparison of river flow hydrographs during the test period showed improved representation of multi-peak events and rapid recession behaviour typical of flashy river catchment regimes (Fig. 15). The ensemble model's ability to improve prediction accuracy was also demonstrated by the flow-duration-curve analysis (Fig. S6), in which near-perfect agreement with the observed curve was obtained.

3.5. Extreme river flow analysis

Tables 8 and 9 present the FHV and FLV calculations for the Semliki and Tochi River flows. The computed FDC bias indicates clear differences in the ability of the evaluated models to reproduce the observed streamflow regime. For Semliki, the FFNN and Diff-FFNN-LSTM individual models resulted in slightly low (acceptable) underestimation of peak and low flows. The TV-DMA model also resulted in the lowest overall bias and the closest agreement with the observed flow-duration

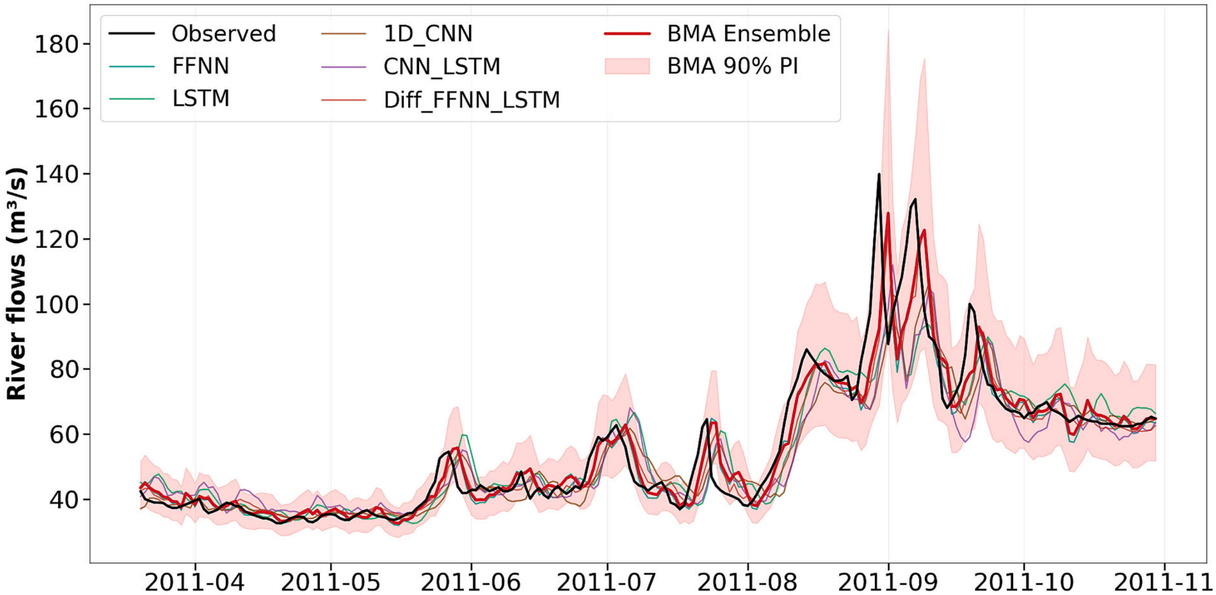

Fig. 11. BMA ensemble predictions compared to observed river flows for the Semliki test period.

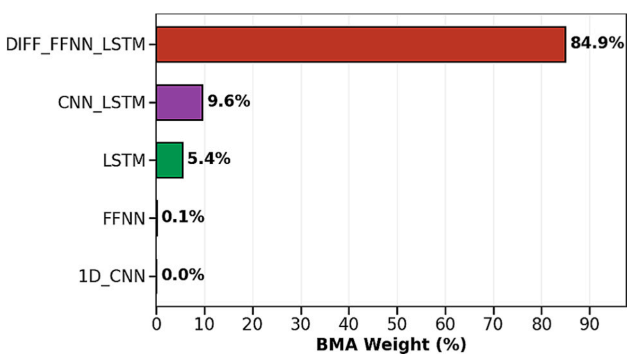

Fig. 12. Final BMA weights for the Tochi River catchment.

Table 5
BMA ensemble performance comparison for the Tochi River catchment.

Model	PBIAS (%)	NSE	RMSE	MAE
FFNN	−22.41	0.640	5.93	3.48
LSTM	−17.43	0.605	6.21	3.10
CNN-LSTM	−4.78	0.641	5.92	3.76
Diff-FFNN-LSTM	−6.75	0.707	5.34	2.47
BMA ensemble	−18.86	0.710	5.32	2.30

curve. Furthermore, the performances of the TV-DMA and BMA ensemble models were comparable. For Tochi, only the Diff-FFNN-LSTM individual model resulted in excellent estimation of peak and low flows. Furthermore, the TV-DMA model also resulted in very low (acceptable) underestimation of peak flows and moderate overestimation of low flows and its performance was superior when compared to corresponding results obtained using the BMA ensemble model. These results demonstrate that dynamic ensemble weighting improves representation of both flood and low-flow conditions while reducing individual model structural prediction errors.

4. Results discussion

The presented study combined individual ML, DL and hybrid model architectures (FFNN, LSTM, 1D-CNN, CNN-LSTM, and Diff-FFNN-LSTM) using the Time-Varying Dynamic Model Averaging (TV-DMA) approach, which was evaluated through case studies of the Semliki and Tochi catchments in the Upper White Nile basin. In addition, the performance of the TV-DMA was benchmarked against corresponding results obtained using the BMA approach. This discussion interprets the main findings on the reliability of physically-based HEC-HMS, the individual ML, DL and hybrid ML models, and the comparative performance evaluation of the developed EDL methods.

4.1. Performance evaluation of individual base learner models

The HEC-HMS model for the Semliki River catchment, which was developed in an earlier study (Mugume et al., 2026), demonstrated reliable performance in predicting the prevailing river flow conditions in the base flow-dominated Semliki River catchment. The study findings for the rainfall-runoff dominated Tochi River catchment showed less satisfactory performance of the developed HEC-HMS model, particularly in simulating peak flows. Overall, the HEC-HMS model performance for both catchments was less superior when compared to the tested data-driven models. The current study's findings agree with Murungi et al. (2026), who demonstrated that the length of training data tended to influence the performance of individual ML and DL models: i.e., ML and DL models can outperform physically-based models as data availability increases. However, in instances when the training data is limited, process-based models or hybrids of process-based and data-driven models become more reliable. In addition, integration of ML models with data augmentation approaches has also been shown to improve the reliability of river flow predictions (Kratzert et al., 2018).

The study results also suggest that in the Semliki River catchment, the LSTM model (NSE = 0.891) outperformed other individual models. The superior performance of the LSTM model in the Semliki catchment may be attributed to its ability to learn long-term temporal dependencies and catchment storage dynamics (e.g., groundwater and snowmelt) which could have enabled it to gain hydrological memory. Base flow-dominated catchments such as those with snowmelt contributions exhibit stronger hydrological memory and flow persistence, characteristics that are well suited to the memory architecture of LSTM networks (Kratzert et al., 2019b; Sabzipour et al., 2023). In contrast, rainfall-runoff-dominated catchments are characterised by rapid hydrological responses and flashy flow regimes that are generally more difficult to represent using recurrent architectures alone (Kratzert et al., 2019b, 2018).

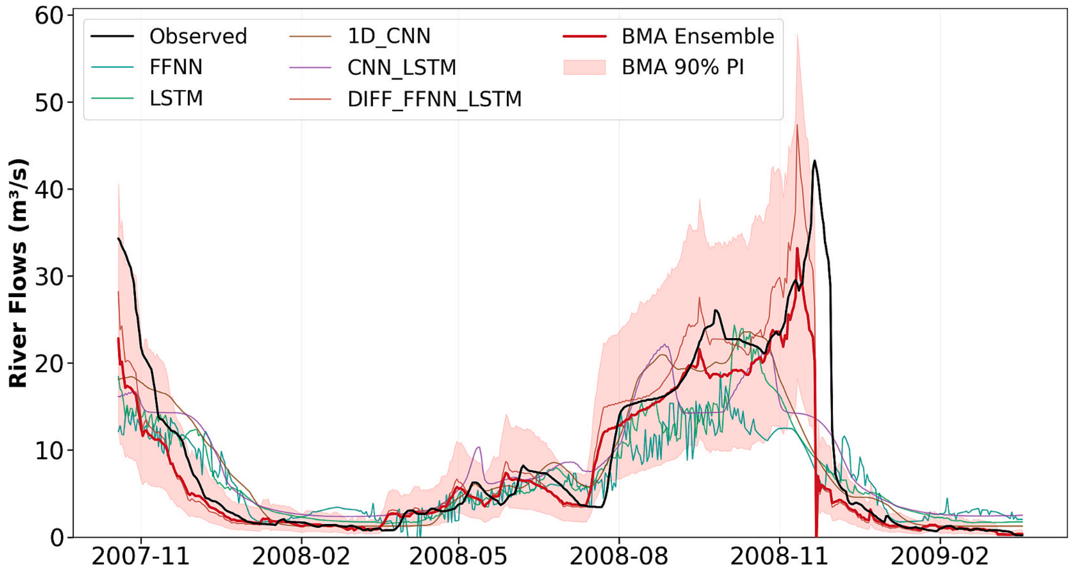

Fig. 13. BMA comparison during the test period for the Tochi River catchment.

Table 6
TV-DMA ensemble performance comparison for the Semliki River catchment.

Model	PBIAS (%)	NSE	MAE (m^3/s)	RMSE (m^3/s)
FFNN	−5.14	0.911	4.20	6.37
LSTM	−4.33	0.808	5.82	9.39
1D-CNN	−2.33	0.797	6.35	9.45
CNN-LSTM	−4.56	0.823	5.83	9.01
Diff-FFNN-LSTM	−0.54	0.925	3.31	5.85
TV-DMA Ensemble	−0.42	0.910	3.81	6.42

In contrast, the study findings showed that the FFNN model resulted in superior performance (NSE = 0.698) when compared to other base learner models in the rainfall-runoff dominated Tochi catchment. These findings reflect the rapid hydrological response of the Tochi catchment, where rainfall-runoff generation is driven primarily by short-term meteorological inputs rather than long-term flow persistence. These findings agree with previous studies which have shown that FFNN model performance is influenced by catchment characteristics and that these models can effectively represent nonlinear rainfall-runoff relationships in rapidly responding catchments (Shamseldin et al., 2002; Toth and Brath, 2007). The present findings are also consistent with Staudinger et al. (2025), who demonstrated that ANN models can effectively exploit the strong autocorrelation inherent in hydrological time series through lagged input windows, despite lacking an explicit memory mechanism.

Furthermore, the superior performance of the FFNN in the Tochi catchment could also be attributed to the longer observational record, which comprised approximately 11 years of concurrent rainfall and river flow data compared with 5 years for the Semliki catchment. A larger training dataset provides a more comprehensive representation of hydrological variability, thereby improving model calibration and predictive performance (Toth and Brath, 2007).

The study findings also demonstrated superior performance of the

Table 7
TV-DMA ensemble performance comparison for the Tochi River catchment.

Model	PBIAS (%)	NSE	NSE_log	MAE	RMSE
CNN-LSTM	−3.20	0.651	0.729	3.30	5.69
Diff-FFNN-LSTM	−0.24	0.819	0.891	1.61	4.10
TV-DMA Ensemble	−0.34	0.831	0.905	1.63	3.96

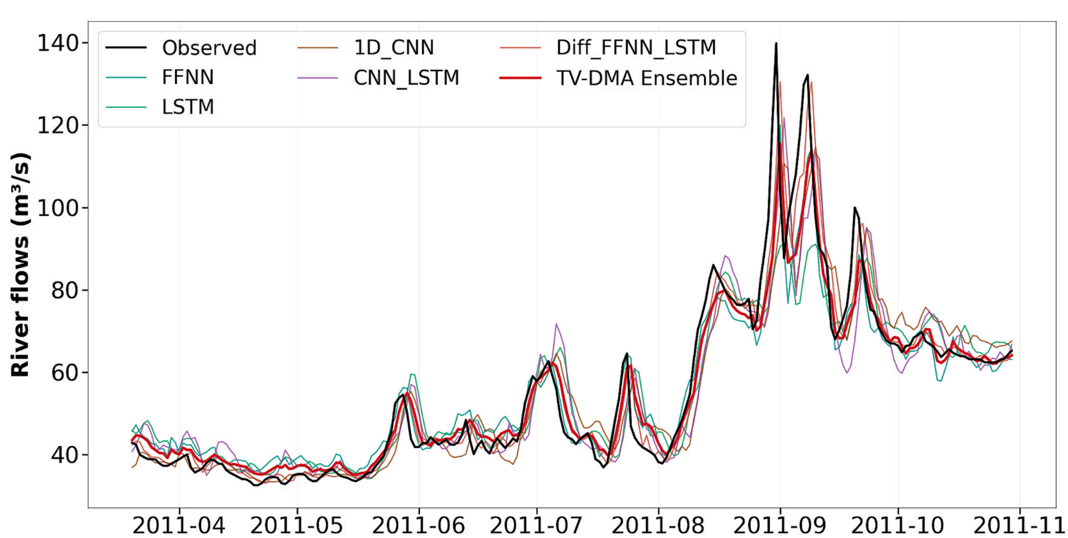

Fig. 14. TV-DMA test period comparison for the Semliki River catchment.

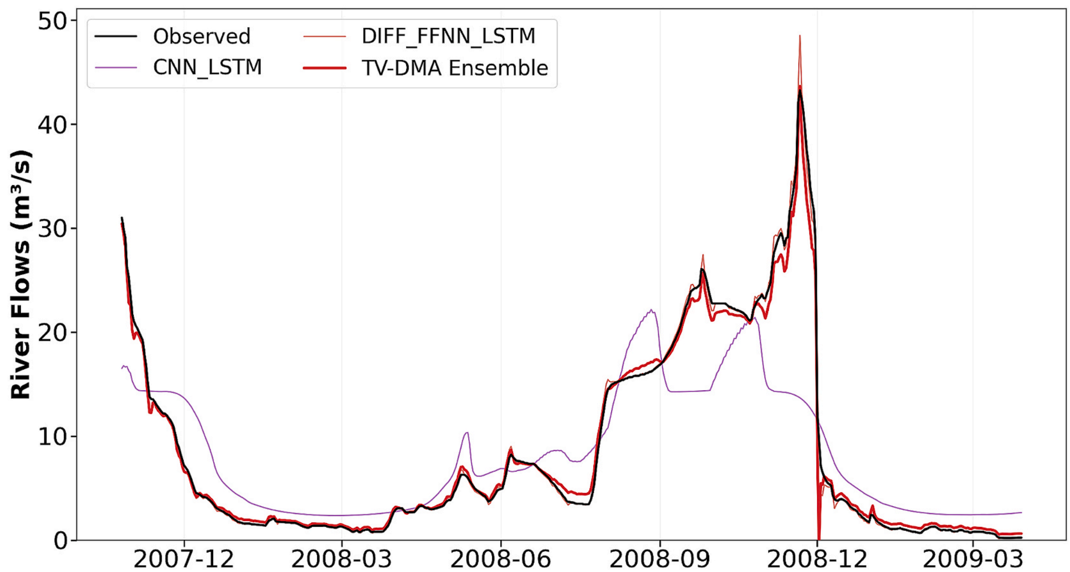

Fig. 15. TV-DMA test period hydrograph for the Tochi River catchment.

Table 8
Percentage Bias (%Bias) in FHV and FLV for Semliki River Flows.

Model	% BiasFHV	% BiasFLV	Model Performance
FFNN	−15.833	−8.805	low (acceptable) underestimation of peak and low flows
LSTM	−32.016	−18.601	large underestimation of peak flows, very low (acceptable) underestimation of low flows
1D_CNN	−15.064	18.104	low (acceptable) underestimation of peak and low flows
CNN_LSTM	−13.354	−6.172	low (acceptable) underestimation of peak and low flows
DIFF_FFNN_LSTM	−4.374	0.08	Excellent reproduction of high and low flows
TV_DMA_Ensemble	−15.251	16.571	low (acceptable) underestimation of peak flows and low (acceptable) overestimation of low flows
BMA_Ensemble	−7.612	−19.699	low (acceptable) underestimation of peak and low flows

Table 9
Percentage Bias (%Bias) in FHV and FLV for Tochi River Catchment.

Model	% BiasFHV	%BiasFLV	Model Performance
FFNN	−58.353	23.541	large underestimation of peak flows, moderate overestimation of low flows
LSTM	−52.828	34.208	large underestimation of peak flows, large overestimation of low flows
1D_CNN	−39.331	98.299	large underestimation of peak flows and large overestimation of peak flows
CNN_LSTM	−43.870	94.741	large underestimation of peak flows, large overestimation of low flows
DIFF_FFNN_LSTM	0.730	1.455	Excellent reproduction of peak and low flows
TV_DMA_Ensemble	−7.997	44.329	very low (acceptable) underestimation of peak flows, moderate over estimation of low flows
BMA_Ensemble	−26.945	−380.304	moderate underestimation of peak flows, large underestimation of low flows

heterogeneous Diff-FFNN-LSTM hybrid model in both catchments (NSE $\geq$ 0.980). Our study findings agree with Yin et al. (2022), Liu et al. (2024), and Feng et al. (2026), which demonstrated the potential of recent hybrid and attention-based deep learning architectures in improving river flow predictions by combining temporal modelling with enhanced feature extraction. These results also agree with the findings of Lin et al. (2021), Mugume et al. (2024) and Pokharel & Roy (2024), who attributed the superior performance of hybrid models to their ability to combine the strengths of individual models, resulting in superior river flow predictive performance in data-scarce regions. Furthermore, the Diff-FFNN-LSTM model also provided more consistent reproduction of observed river flow distributions (central tendency, flow variability, peak and low flows, and extreme events), as indicated by the flow duration curve analysis. This improved performance can be attributed to the first-order differential processing algorithm, which stabilises input time series by reducing non-stationarity and enhancing the representation of dynamic flow transitions, as well as the independent training of the FFNN and LSTM sub-models (Kratzert et al., 2019a; Lin et al., 2021). These Diff-FFNN-LSTM model features likely improved simulation of rapid flow transitions, including rising limbs, peak flows and recession behaviour, which are critical for reliable river flow prediction in data-scarce catchments.

4.2. Performance and reliability of ensemble modelling strategies

The results demonstrate the value of combining diverse ML, DL and hybrid architectures using dynamic ensemble weighting. TV-DMA achieved high deterministic accuracy and near-zero systematic bias in both Semliki (NSE = 0.910; PBIAS = −0.42%) and Tochi (NSE = 0.831; PBIAS = −0.34%). By updating model weights based on time-varying prediction errors, TV-DMA adapted the ensemble contribution to changing flow conditions, improving model reliability under non-stationary conditions. This finding is consistent with Ruan et al. (2022) who reported superior rainfall runoff predictive performance and lower prediction uncertainty for TV-DMA compared to individual FFNN, LSTM, and GRU models. However, their study used three individual neural architectures. In our current study, we have demonstrated through the use of TV-DMA to integrate diverse and heterogeneous model architectures that included the hybrid first order differential processing Diff-FFNN-LSTM model, which incorporates first-order differential processing, greatly enhances river flow predictive performance.

In addition, the dynamic weight evolution also showed hydrologically plausible behaviour, with greater contributions from Diff-FFNN-LSTM during low flows and FFNN during peak flows, indicating that TV-DMA can exploit complementary model strengths across flow regimes. Ruan et al. (2022) similarly reported that TV-DMA maintained lower error rates than individual models in both flood and non-flood seasons, although performance declined during high-flow periods.

In contrast, BMA produced lower deterministic performance, with NSE values of 0.848 and 0.710 for Semliki and Tochi, respectively. Nevertheless, BMA reduced prediction errors relative to most individual models and resulted in PBIAS values of −0.20% for Semliki and −18.86% for Tochi River catchments, respectively. These results are consistent with Darbandsari and Coulibaly, (2020) who showed that the entropy-based BMA approach can improve both the deterministic and probabilistic peak river flow predictions while accounting for uncertainty arising from differences among model structures. The study findings also agree with Huang et al. (2019) who showed that the BMA outperformed individual Random Forest (RF), Artificial Neural Network (ANN) and Support Vector Machine (SVM) models for daily river flow forecasting.

However, the BMA uncertainty bands widened during peak flows and narrowed during sustained low flows, indicating greater predictive uncertainty under extreme conditions. This behaviour agrees with Moknatian and Mukundan (2023), who showed that despite the high uncertainty, BMA substantially improved the proportion of observations contained within prediction intervals compared with individual objective-function-based simulations during high-flow periods when model divergence is greater (Darbandsari and Coulibaly, 2020; Moknatian and Mukundan, 2023). These findings indicate that BMA may provide useful probabilistic information even when its deterministic performance is lower than that of TV-DMA.

The differences between the ensemble approaches also highlight the importance of model diversity and relative predictive skill. Ensemble gains were modest where Diff-FFNN-LSTM already provided the strongest deterministic predictions (Sharma et al., 2019). The stronger TV-DMA improvement in Tochi may therefore reflect greater variation in model skill under its rapidly changing rainfall-runoff regime. TV-DMA reproduced the observed flow-duration curve more accurately and reduced predictive bias, suggesting that adaptive weighting may be particularly useful where rainfall-runoff relationships vary rapidly over time.

The contrasting results between Semliki and Tochi further indicate that ensemble performance is influenced by catchment hydrological characteristics. TV-DMA resulted in greater improvement in Tochi, whereas BMA achieved stronger bias reduction and prediction-interval performance in Semliki. The substantially lower BMA bias in Semliki may reflect its stronger flow persistence and base flow contribution, which favour probabilistic averaging of individual models with similar temporal behaviour.

However, the current study was limited to two tropical, data-scarce catchments with contrasting flow regimes, and the relative advantages of the two ensemble strategies may depend on rainfall variability, catchment storage, data availability, diversity and predictive skill of the constituent ML and DL models. Future research work, considering, a larger set of hydro-climatic regions is therefore required to establish the transferability of these findings.

It should also be noted that the Tochi observational record was approximately twice as long as that of the Semliki catchment, which likely improved model training, parameter-tuning stability and ensemble calibration, resulting in more robust predictive performance. In contrast, the shorter Semliki record may have constrained model exposure to flow variability and reduced the stability of uncertainty estimates.

With respect to computational efficiency, the BMA and TV-DMA ensemble model runs utilised approximately 0.07 and 0.80 min. Relative to the process-based HEC-HMS model, the BMA ensemble reduced computational time by 98.0–98.4%, while the TV-DMA ensemble achieved a reduction of 77–81%. Although the TV-DMA ensemble incurred a longer runtime because model weights were updated at each time step, it remained substantially more computationally efficient than HEC-HMS, supporting its application in near real-time operational river flow forecasting where both prediction accuracy and computational speed are critical (Refer to Supplementary Information Table S8).

4.3. Model transferability and generalisability

The study results suggest that transferability of the individual machine learning (ML) and deep learning (DL) models from the Semliki to the Tochi catchment was constrained by differences in hydrological regimes and data availability. The best-performing individual LSTM model trained on Semliki's hydrological conditions underperformed, on the unbiased test set, when applied to Tochi's rainfall-driven flow conditions. This aligns with findings by Ma et al. (2024), who reported limited spatial transferability of ML and DL models in river flow prediction without local data integration.

However, the individual Diff-FFNN-LSTM model and the TV-DMA ensemble model resulted in comparable performance in Semliki and Tochi catchments, respectively. These findings further suggest that use of the differential processing approach facilitates reduction of catchment-specific non-stationarity in data-scarce environments. In addition, the TV-DMA demonstrates enhanced adaptability by adjusting model weights in response to local hydrological conditions, thereby improving transferability across catchments. However, to further enhance model generalisability and transferability, future research is required to improve dynamic weighting ensemble model training using diverse data sets that encompass a wide range of hydrological conditions that include arid, snow-dominated and regulated rivers (e.g., Ueda et al., 2025).

4.4. Ensemble modelling constraints

Sequence-based deep learning models such as LSTM and CNN-LSTM utilised different optimised lookback window lengths during pre-training, necessitating temporal alignment of base model predictions before ensemble generation. This required the trimming of early time steps based on the longest lookback length to obtain a common, valid window consistent across all base learner models.

In this study, ensemble calibration was undertaken using prediction windows aligned with the maximum lookback lengths of 30 days for Tochi and 28 days for Semliki to establish a common-day evaluation period for consistent comparison with ensemble predictions (e.g. Kratzert et al. (2019) and Ruan et al. (2022). This adjustment reduced the effective ensemble training dataset which decreased the recomputed test performance metrics by approximately 0.3–13% for most models, with larger reductions of 20–30% for the FFNN model. These findings highlight the sensitivity of performance metrics to evaluation window harmonisation which is an important consideration in ensemble hydrological modelling.

4.5. Study limitations and future research directions

The key study limitations included limited observed hydrometeorological data, potential uncertainty in input data quality, and evaluation using two tropical catchments, which may constrain generalisability and transferability of the developed dynamic weighting EDL approach to other catchments. Similar hydrological deep learning studies have shown that model reliability and transferability improve when trained on larger multi-basin datasets with more diverse hydroclimatic conditions and expanded predictor sets (Kratzert et al., 2019b; Nearing et al., 2021). In addition, the ensemble evaluation was limited to a specific set of ML, DL and hybrid base learners (i.e. FFNN, 1D-CNN, LSTM, CNN-LSTM and Diff-FFNN-LSTM). However, other advanced deep learning architectures may yield different ensemble behaviour. Additionally, to the authors' knowledge, no previous studies have reported specific multiplier ranges for adjusting river flow regime responses in ensemble deep learning frameworks, which necessitated use of a conservative range so as to maintain numerical stability while accounting for model structural and parameter uncertainty (e.g. Shu et al., 2023). In view of the identified study limitations, future research should evaluate TV-DMA transferability across broader hydro climatic regions and incorporate predictors derived from bias-corrected reanalysis datasets. Furthermore, future research should investigate the effect of dynamic weight evolution and model prediction uncertainty in TV-DMA approaches, and develop state-dependent functions to systematically calibrate regime-dependent multipliers through hyper parameterisation of different multiplier combinations. Future work could also extend model predictive capabilities to multi-day lead times through encoder–decoder model architectures and investigate the effectiveness of hybrid physics-informed ensemble deep learning approaches in enhancing river flow predictive accuracy and reliability in data-scarce environments.

5. Conclusions and recommendations for future research

Accurate river flow prediction is essential for resilient and sustainable water resources management including the design, operation or rehabilitation of hydraulic infrastructure for water supply, irrigation, flood control and hydropower dams, particularly in regions that are increasingly exposed to climate change induced extreme events. This study evaluated an ensemble deep learning approach that integrates FFNN, 1D-CNN, LSTM, CNN-LSTM and the heterogenous Diff-FFNN-LSTM models using the Time-Varying Dynamic Model Averaging (TV-DMA) ensemble in two hydrologically contrasting, data-scarce tropical catchments in the Upper White Nile Basin. The results demonstrate that model diversity can improve prediction reliability when individual architectures exhibit complementary strengths across flow regimes.

The Diff-FFNN-LSTM model consistently provided the strongest individual-model performance, with NSE values of 0.980 and 0.984 for Semliki and Tochi, respectively. Its performance indicates that first-order differential processing can provide a useful strategy for stabilising non-stationary flow sequences and improving the representation of changing flow conditions. This suggests that differential processing is particularly relevant where rapid transitions in river flows constrain the performance of conventional deep learning model architectures.

Furthermore, the TV-DMA provided the strongest overall ensemble performance, achieving NSE values of up to 0.910 and PBIAS below 0.5%. Its principal practical value is its ability to update model contributions as predictive skill changes over time, rather than relying on fixed model weights. This adaptive formulation provides a basis for operational forecasting systems in which different models can contribute according to prevailing hydrological conditions. Previous ensemble forecasting research similarly identifies adaptive model combination and uncertainty treatment as important directions for improving the reliability and operational usefulness of river flow forecasts (Ruan et al., 2022; Troin et al., 2021).

The findings show that the benefits of dynamic ensemble modelling depend on catchment characteristics and the predictive diversity of the constituent models. TV-DMA performed particularly well in the rainfall-runoff-dominated Tochi catchment, whereas BMA provided competitive probabilistic performance in Semliki. Therefore, the results do not imply universal superiority of TV-DMA or BMA, but indicate that adaptive weighting can be advantageous where model skill varies with hydrological conditions. The proposed framework is consequently most relevant as a flexible forecasting strategy for catchments where non-stationary flow behaviour and limited observations make reliance on a single model or fixed ensemble weights less desirable.

However, this research faced several limitations. First, the analysis relied on a relatively short observed hydro-meteorological data from only two catchments, making the results sensitive to input data quality, sampling period and the representation of extreme hydrological conditions. Deep learning models can be particularly sensitive to observational errors and out-of-sample extremes, while limited observations can constrain model transferability across catchments (Murungi et al., 2026). Second, the ensemble was restricted to five base learner architectures, namely FFNN, 1D-CNN, LSTM, CNN-LSTM and Diff-FFNN-LSTM, and therefore did not assess newer architectures such as Transformers or Graph Neural Networks. Third, although TV-DMA dynamically adjusted model contributions, the weighting mechanism tended to concentrate weights on the strongest-performing models, potentially reducing the diversity of the ensemble and limiting the benefits of model combination when one architecture dominates predictive skill. Finally, the findings were evaluated in only two contrasting tropical catchments and therefore should not be assumed to be transferable to other hydroclimatic regimes without further validation.

Furthermore, future research should: (a) Evaluate the transferability of TV-DMA across a broader range of tropical and non-tropical hydroclimatic regimes using predictors derived from bias-corrected reanalysis datasets; (b) Investigate the relationship between time-varying model weights and prediction uncertainty and assess its performance using of real-time observations and meteorological forecasts to support operational river flow forecasting; (c) Extend model predictive capabilities to multi-day lead times using encoder-decoder architectures and (d) Investigate the effectiveness of hybrid physics-informed ensemble deep learning approaches in enhancing river flow predictive accuracy and reliability in data-scarce environments.

Overall, the study demonstrates that reliable and accurate river flow prediction in data-scarce tropical catchments benefits from combining catchment-sensitive pre-processing, heterogeneous hybrid deep learning model architectures with differential processing to capture rapid flow transitions, non-stationarity-aware network design, and dynamic ensemble modelling strategies. The developed approach can enable reliable, computationally efficient simulation of river flows and inform future development of river flow forecasting and flood early-warning systems in data-scarce regions.

CRediT authorship contribution statement

Angel I. Tumushabe: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Seith N. Mugume:** Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Conceptualization. **Johanna Sörensen:** Writing – review & editing, Writing – original draft, Supervision, Investigation.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used OpenAI's ChatGPT to improve readability and language. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of competing 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.

Acknowledgements

Acknowledgement is given to the Directorate of Water Resources Management (DWRM), Ministry of Water and Environment, for providing the data sets used in the study. The insights of the Editors and the two anonymous reviewers are also greatly acknowledged.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jhydrol.2026.136387.

Data availability

Data will be made available on request.

References

Ajami, N.K., Duan, Q., Sorooshian, S., 2007. An integrated hydrologic Bayesian multimodel combination framework : confronting input, parameter, and model structural uncertainty in hydrologic prediction. Water Resour. Res. 43, 1–19. https://doi.org/10.1029/2005WR004745.

Apicella, A., Isgr, F., Pollastro, A., Prevete1, R., 2026. Don't stop me now: rethinking validation criteria for model parameter selection. Fisciano (Salerno), Italy. https://doi.org/10.48550/arXiv.2602.22107.

Arsenault, R., Gatien, P., Renaud, B., Brissette, F., Martel, J., 2015. A comparative analysis of 9 multi-model averaging approaches in hydrological continuous streamflow simulation. J. Hydrol. 529, 754–767. https://doi.org/10.1016/j.jhydrol.2015.09.001.

Bartles, M., Brauer, T., Ho, D., Fleming, M., Karlovits, G., Van Jay Pak, N., Willis, J., 2023. HEC-HMS User's Manual. Davis, California.

Baste, S., Klotz, D., Espinoza, E.A., Bardossy, A., Loritz, R., 2025. Unveiling the limits of deep learning models in hydrological extrapolation tasks. Hydrol. Earth Syst. Sci. 29, 5871–5891. https://doi.org/10.5194/hess-29-5871-2025.

Bergstra, J., Bengio, Y., 2012. Random search for hyper-parameter optimization. J. Mach. Learn. Res. 13, 281–305.

Chadalawada, J., Herath, H.M.V.V., Babovic, V., 2020. Hydrologically informed machine learning for rainfall-runoff modeling: a genetic programming-based toolkit for automatic model induction. Water Resour. Res. 56, 1–23. https://doi.org/10.1029/2019WR026933.

Darbandsari, P., Coulibaly, P., 2020. Introducing entropy-based Bayesian model averaging for streamflow forecast. J. Hydrol. 591, 1–17. https://doi.org/10.1016/j.jhydrol.2020.125577.

Descroix, L., Nouvelot, J.-F., Vauclin, M., 2002. Evaluation of an antecedent precipitation index to model runoff yield in the western Sierra Madre (North-west Mexico). J. Hydrol. 263, 114–130. https://doi.org/10.1016/S0022-1694(02)00047-1.

Ding, C., Wang, Z., Tan, Z., Zhang, X., 2026. Multi-source heterogeneous data-driven interpretable model based on transformer and Kolmogorov-Arnold networks for streamflow prediction. Expert Syst. Appl. 328, 1–25. https://doi.org/10.1016/j.eswa.2026.132862.

Dion, P., Martel, J., Arsenault, R., 2021. Hydrological ensemble forecasting using a multi-model framework. J. Hydrol. 600, 1–15. https://doi.org/10.1016/j.jhydrol.2021.126537.

Duan, Q., Ajami, N.K., Gao, X., Sorooshian, S., 2007. Multi-model ensemble hydrologic prediction using Bayesian model averaging. Adv. Water Resour. 30, 1371–1386. https://doi.org/10.1016/j.advwatres.2006.11.014.

Feng, X., Lin, K., Zhuang, L., Xu, Y., 2026. A rapid flood inundation integrated surrogate model based on a CNN-LSTM-ATT deep learning framework. J. Hydrol. Reg. Stud. 64, 1–23. https://doi.org/10.1016/j.ejrh.2026.103166.

Fu, G., Jin, Y., Sun, S., Yuan, Z., Butler, D., 2022. The role of deep learning in urban water management: a critical review. Water Res. 223, 1–16. https://doi.org/10.1016/j.watres.2022.118973.

Gao, W., Liao, Y., Chen, Y., Lai, C., He, S., Wang, Z., 2024. Enhancing transparency in data-driven urban pluvial flood prediction using an explainable CNN model. J. Hydrol. 645, 1–16. https://doi.org/10.1016/j.jhydrol.2024.132228.

Gelete, G., Nourania, V., Gokcekus, H., Gichamo, T., 2023. Ensemble physically based semi-distributed models for the rainfall-runoff process modeling in the data-scarce Katar catchment, Ethiopia. J. Hydroinformatics 25, 567–592. https://doi.org/10.2166/hydro.2023.197.

Ghimire, S., Yaseen, Z.M., Farooque, A.A., Deo, R.C., 2021. Streamflow prediction using an integrated methodology based on convolutional neural network and long short-term memory networks. Sci. Rep. 11, 1–26. https://doi.org/10.1038/s41598-021-96751-4.

Gupta, H.V., Kling, H., 2011. On typical range, sensitivity, and normalization of mean squared error and nash-sutcliffe efficiency type metrics. Water Resour. Res. 47, 2–4. https://doi.org/10.1029/2011WR010962.

Gupta, H.V., Kling, H., Yilmaz, K.K., Martinez, G.F., 2009. Decomposition of the mean squared error and NSE performance criteria : implications for improving hydrological modelling. J. Hydrol. 377, 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003.

He, Y., Wang, Z., Cheng, H., Ding, W., 2026. Multi-scale feature fusion and uncertainty quantification in streamflow prediction: a temporal convolutional network approach with hybrid denoising. Environ Model Softw. 198, 1–18. https://doi.org/10.1016/j.envsoft.2026.106879.

Hosseini, F., Prieto, C., Alvarez, C., 2025. Ensemble learning of catchment-wise optimized LSTMs enhances regional rainfall-runoff modelling – case study: Basque Country. Spain. J. Hydrol. 646, 1–17. https://doi.org/10.1016/j.jhydrol.2024.132269.

Hu, F., Yang, Q., Yang, J., Luo, Z., Shao, J., 2024. Incorporating multiple grid-based data in CNN-LSTM hybrid model for daily runoff prediction in the source region of the Yellow River Basin. J. Hydrol.: Reg. Stud. 51, 1–14. https://doi.org/10.1016/j.ejrh.2023.101652.

Huang, H., Liang, Z., Li, B., Wang, D., Hu, Y., 2019. Combination of multiple data-driven models for long-term monthly runoff predictions based on Bayesian model averaging. Water Resour. Manag. 33, 3321–3338. https://doi.org/10.1007/s11269-019-02305-9.

Huang, T., Merwade, V., 2023. Improving Bayesian model averaging for ensemble flood modeling using multiple Markov chains Monte Carlo sampling. Water Resour. Res. 59, 1–20. https://doi.org/10.1029/2023WR034947.

Ishida, K., Ercan, A., Nagasato, T., Kiyama, M., Amagasaki, M., 2024. Use of one-dimensional CNN for input data size reduction in LSTM for improved computational efficiency and accuracy in hourly rainfall-runoff modeling. J. Environ. Manage. 359, 1–13. https://doi.org/10.1016/j.jenvman.2024.120931.

Jain, A., Srinivasulu, S., 2006. Integrated approach to model decomposed flow hydrograph using artificial neural network and conceptual techniques. J. Hydrol. 317, 291–306. https://doi.org/10.1016/j.jhydrol.2005.05.022.

JICA, 2014. Rural Road Master Planning in Amuru and Nwoya Districts.

Kabir, S., Patidar, S., Xia, X., Liang, Q., Neal, J., Pender, G., 2020. A deep convolutional neural network model for rapid prediction of fluvial flood inundation. J. Hydrol. 590, 1–16. https://doi.org/10.1016/j.jhydrol.2020.125481.

Kannengießer, N., Hasebrook, N., Morsbach, F., Zöller, M., Franke, J.K.H., Lindauer, M., Hutter, F., Sunyaev, A., 2025. Practitioner motives to use different hyperparameter optimization methods. ACM Trans. Comput. Interact. 32, 1–33. https://doi.org/10.1145/3745771.

Kiranyaz, S., Ince, T., Hamila, R., Gabbouj, M., 2015. Convolutional Neural Networks for patient-specific ECG classification. In: 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 2608–2611. https://doi.org/10.1109/EMBC.2015.7318926.

Koycegiz, C., Buyukyildiz, M., 2022. Estimation of streamflow using different artificial neural network models. OKU J. Inst. Sci. Technol. 5, 1141–1154. https://doi.org/10.47495/okufbed.1037242.

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., Herrnegger, M., 2018. Rainfall-runoff modelling using Long Short-Term Memory (LSTM) networks. Hydrol. Earth Syst. Sci. 22, 6005–6022. https://doi.org/10.5194/hess-22-6005-2018.

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A.K., 2019a. Toward improved predictions in ungauged basins: exploiting the power of machine learning. Water Resour. Res. 55, 11–344-11–354. https://doi.org/10.1029/2019WR026065.

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., Nearing, G., 2019b. Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. Hydrol. Earth Syst. Sci. 23, 5089–5110.

Lecun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444. https://doi.org/10.1038/nature14539.

Legates, D.R., Mccabe Jr, G.J., 1999. Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. Water Resour. Res. 35, 233–241. https://doi.org/10.1029/1998WR900018.

Lin, Y., Wang, D., Wang, G., Qiu, J., Long, K., Du, Y., Xie, H., Wei, Z., Shangguan, W., Dai, Y., 2021. A hybrid deep learning algorithm and its application to streamflow prediction. J. Hydrol. 601, 1–10. https://doi.org/10.1016/j.jhydrol.2021.126636.

Liu, J., Bian, Y., Lawson, K., Shen, C., 2024. Probing the limit of hydrologic predictability with the Transformer network. J. Hydrol. 637, 1–8. https://doi.org/10.1016/j.jhydrol.2024.131389.

Liu, G., Tang, Z., Qin, H., Liu, S., Shen, Q., Qu, Y., 2022. Short-term runoff prediction using deep learning multi-dimensional ensemble method. J. Hydrol. 609, 1–12. https://doi.org/10.1016/j.jhydrol.2022.127762.

Ma, K., Shen, C., Xu, Z., He, D., 2024. Transfer learning framework for streamflow prediction in large-scale transboundary catchments: sensitivity analysis and applicability in data-scarce basins. J. Geog. Sci. 34, 963–964. https://doi.org/10.1007/s11442-024-2235-x.

Mao, G., Wang, M., Liu, J., Wang, Z., Wang, K., Meng, Y., Zhong, R., Wang, H., Li, Y., 2021. Comprehensive comparison of artificial neural networks and long short-term memory networks for rainfall-runoff simulation. Phys. Chem. Earth 123, 1–12. https://doi.org/10.1016/j.pce.2021.103026.

Mcgil, R., John, T.W., Wayne, L.A., 1978. Variations of box plots. Am. Stat. 32, 12–16. https://doi.org/10.1080/00031305.1978.10479236.

Modi, P., Jennings, K., Wobus, C., Livneh, B., Kasprzyk, J., Small, E., 2025. Using deep learning in ensemble streamflow forecasting: exploring the predictive value of explicit snowpack information. J. Adv. Model. Earth Syst. 17, 1–22. https://doi.org/10.1029/2024MS004582.

Moknatian, M., Mukundan, R., 2023. Uncertainty analysis of streamflow simulations using multiple objective functions and Bayesian Model Averaging. J. Hydrol. 617, 1–16. https://doi.org/10.1016/j.jhydrol.2022.128961.

Moriasi, D.N., Arnold, J., Liew, M.W.V., 2007. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. Trans. ASABE 50, 885–900. https://doi.org/10.13031/2013.23153.

Mount, N., Maier, H., Toth, E., Elshorbagy, A., Solomatine, D., Chang, F.-J., Brahart, R., 2016. Data-driven modelling approaches for socio-hydrology: opportunities and challenges within the Panta Rhei Science Plan. Hydrol. Sci. J. 61, 1192–1208. https://doi.org/10.1080/02626667.2016.1159683.

Mugume, S.N., Murungi, J., Kato, R., Tumwesigye, E., 2026. Modelling the influence of future climate and land-use change on fluvial flood inundation in data-scarce regions using CMIP6 ensembles, FABDEM and a validated 2D hydrodynamic model. Hydrol. Res. 57, 473–500. https://doi.org/10.2166/nh.2026.103.

Mugume, S.N., Murungi, J., Nyenje, P.M., Sempewo, J.I., Okedi, J., Sörensen, J., 2024. Development and application of a hybrid artificial neural network model for simulating future stream flows in catchments with limited in situ observed data. J. Hydroinformatics 26, 1944–1969. https://doi.org/10.2166/hydro.2024.066.

Murungi, J., Mugume, S.N., Loots, I., 2026. Enhancing the performance of machine learning models for simulation of river flows in data-scarce catchments using bootstrapping and physics-based data augmentation. Eng. Appl. Artif. Intel. 181, 1–18. https://doi.org/10.1016/j.engappai.2026.115603.

MWE, 2018. Catchment Management Plan for Tochi Irrigation Scheme – Draft. Nairobi, Kenya https://doi.org/1766503544.

MWE, 2013. Uganda National Water Resources Assessment. Kampala, Uganda.

Nearing, G.S., Kratzert, F., Sampson, A.K., Pelissier, C.S., Klotz, D., Frame, J.M., Prieto, C., Gupta, H.V., 2021. What role does hydrological science play in the age of machine learning ? Water Resour. Res. 57, 1–15. https://doi.org/10.1029/2020WR028091.

Nelsap-CU, 2020. Semliki Delta Transboundary Wetland Management Plan 2020–2030. Kigali, Rwanda.

Oudin, L., Andre, V., Mathevet, T., Perrin, C., Michel, C., 2006. Dynamic averaging of rainfall-runoff model simulations from complementary model parameterizations. Water Resour. Res. 42, 1–10. https://doi.org/10.1029/2005WR004636.

Pareta, K., 2023. Hydrological modelling of largest braided river of India using MIKE Hydro River software with rainfall runoff, hydrodynamic and snowmelt modules. Water Clim. Chang. 14, 1314–1338. https://doi.org/10.2166/wcc.2023.484.

Pokharel, S., Roy, T., 2024. A parsimonious setup for stream flow forecasting using CNN-LSTM. Hydroinformatics 26, 2751–2761. https://doi.org/10.2166/hydro.2024.114.

Qian, X., Wang, B., Chen, J., Fan, Y., Mo, R., Xu, C., Liu, W., 2025. An explainable ensemble deep learning model for long-term streamflow forecasting under multiple uncertainties. J. Hydrol. 662, 1–17. https://doi.org/10.1016/j.jhydrol.2025.133968.

Raftery, A.E., Gneiting, T., Balabdaoui, F., Polakowski, M., 2005. Using bayesian model averaging to calibrate forecast ensembles. Mon. Weather Rev. 133, 1155–1174. https://doi.org/10.1175/MWR2906.1.

Raftery, A.E., Kárný, M., Ettler, P., 2010. Online prediction under model uncertainty via dynamic model averaging : application to a cold rolling mill online prediction under model uncertainty via dynamic model averaging: application to a cold rolling mill. Technometrics 1706, 52–56. https://doi.org/10.1198/TECH.2009.08104.

Rahman, K.U., Pham, Q.B., Jadoon, K.Z., Shahid, M., Kushwaha, D.P., Khedher, Z.D., Mohammadi, B., Mohamed, K., Anh, D.T., 2022. Comparison of machine learning and process-based SWAT model in simulating streamflow in the Upper Indus Basin. Appl. Water Sci. 12. https://doi.org/10.1007/s13201-022-01692-6.

Ridolfi, E., Kumar, H., Bárdossy, A., 2020. A methodology to estimate flow duration curves at partially ungauged basins. Hydrol. Earth Syst. Sci. 24, 2043–2060. https://doi.org/10.5194/hess-24-2043-2020.

Ritter, A., Muñoz-Carpena, R., 2013. Performance evaluation of hydrological models: statistical significance for reducing subjectivity in goodness-of-fit assessments. J. Hydrol. 480, 33–45. https://doi.org/10.1016/j.jhydrol.2012.12.004.

Ruan, J., Cui, Y., Xiang, K., Song, Y., 2022. Improved runoff forecasting based on time-varying model averaging method and deep learning. PLoS One 17, 1–20. https://doi.org/10.1371/journal.pone.0274004.

Rumelhart, D.E., Hintont, G.E., Williams, R.J., 1986. Learning representations by back-propagating errors. Nature 323, 533–536. https://doi.org/10.1038/323533a0.

Russell, J.M., Johnson, T.C., 2006. The water balance and stable isotope hydrology of Lake Edward, Uganda-Congo. J. Great Lakes Res. 32, 77–90. https://doi.org/10.3394/0380-1330(2006)32[77:TWBASI]2.0.CO;2.

Sabzipour, B., Arsenault, R., Troin, M., Martel, J., Mai, J., 2023. Comparing a long short-term memory (LSTM) neural network with a physically-based hydrological model for streamflow forecasting over a Canadian catchment. J. Hydrol. 627, 1–14. https://doi.org/10.1016/j.jhydrol.2023.130380.

Saraiva, S.V., Carlvalho, F. de O., Augusto, C., Santos, C.A.G., Barreto, L.C., Freire, P.K. de M.M.F., 2021. Daily streamflow forecasting in Sobradinho Reservoir using machine learning models coupled with wavelet transform and bootstrapping. Appl. Soft Comput. J. 102, 1–11. https://doi.org/10.1016/j.asoc.2021.107081.

Schutte, C., van der Laan, M., van der Merwe, B., 2024. Leveraging historic stream flow and weather data with deep learning for enhanced stream flow predictions. J. Hydroinformatics 26, 835–852. https://doi.org/10.2166/hydro.2024.268.

Shamseldin, A.Y., Nasr, A.E., Connor, K.M.O., 2002. Comparison of different forms of the multi-layer feed-forward neural network method used for river flow forecasting. Hydrol. Earth Syst. Sci. 6, 671–684.

Sharma, S., Siddique, R., Reed, S., Ahnert, P., Mejia, A., 2019. Hydrological model diversity enhances stream flow forecast skill at short- to medium-range timescales. Water Resour. Res. 55, 1510–1530. https://doi.org/10.1029/2018WR023197.

Sheikh, M.R., Coulibaly, P., 2025. Introducing time series features based dynamic weights estimation framework for hydrologic forecast merging. J. Hydrol. 654, 1–18. https://doi.org/10.1016/j.jhydrol.2025.132872.

Shu, Z., Zhang, J., Wang, L., Jin, J., Cui, N., Wang, G., Sun, Z., Liu, Y., Bao, Z., Liu, C., 2023. Evaluation of the impact of multi-source uncertainties on meteorological and hydrological ensemble forecasting. Engineering 24, 212–228. https://doi.org/10.1016/j.eng.2022.06.007.

Singh, J., Knapp, H.V., Demissie, M., 2007. Hydrologic modeling of the Iroquois river watershed using HSPF and SWAT. J. Am. Water Resour. Assoc. 41, 343–360. https://doi.org/10.1111/j.1752-1688.2005.tb03740.x.

Solanki, H., Vegad, U., Kushwaha, A., Mishra, V., 2025. Improving streamflow prediction using multiple hydrological models and machine learning methods. Water Resour. Res. 61, 1–22. https://doi.org/10.1029/2024WR038192.

Staudinger, M., Herzog, A., Loritz, R., Houska, T., Pool, S., Spieler, D., Wagner, P.D., 2025. How well do process-based and data-driven hydrological models learn from limited discharge data? Hydrol. Earth Syst. Sci. 29, 5005–5029. https://doi.org/10.5194/hess-29-5005-2025.

Taye, M.T., Dyer, E., 2024. Hydrologic extremes in a changing climate: a review of extremes in East Africa. Curr. Clim. Chang. Reports 10, 1–11. https://doi.org/10.1007/s40641-024-00193-9.

Teegavarapu, R.S.V., Chandramouli, V., 2005. Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records. J. Hydrol. 312, 191–206. https://doi.org/10.1016/j.jhydrol.2005.02.015.

Tibangayuka, N., Mulungu, D.M.M., Izdori, F., 2022. Evaluating the performance of HBV, HEC-HMS and ANN models in simulating streamflow for a data scarce high-humid tropical catchment in Tanzania. Hydrol. Sci. J. 67, 2191–2204. https://doi.org/10.1080/02626667.2022.2137417.

Tiwari, M.K., Chatterjee, C., 2011. A new wavelet-bootstrap-ANN hybrid model for daily discharge forecasting. J. Hydroinformatics 13, 500–519. https://doi.org/10.2166/hydro.2010.142.

Toth, E., Brath, A., 2007. Multistep ahead streamflow forecasting: role of calibration data in conceptual and neural network modeling. Water Resour. Res. 43, 1–11. https://doi.org/10.1029/2006WR005383.

Troin, M., Arsenault, R., Wood, A.W., Brissette, F., Martel, J.-L., 2021. Generating ensemble streamflow forecasts: a review of methods and approaches over the past 40 years. Water Resour. Res. 57, 1–48. https://doi.org/10.1029/2020WR028392.

Tsangaratos, P., Ilia, I., Chrysafi, A., Matiatos, I., Chen, W., Hong, H., 2023. Applying a 1D convolutional neural network in flood susceptibility assessments – the case of the Island of Euboea. Greece. Remote Sens. 15, 1–30. https://doi.org/10.3390/rs15143471.

Tu, T., Li, Y., Duan, K., Zhao, T., 2024. Enhancing physically-based hydrological modeling with an ensemble of machine-learning reservoir operation modules under heavy human regulation using easily accessible data. J. Environ. Manage. 359, 121044. https://doi.org/10.1016/j.jenvman.2024.121044.

Tumushabe, A.I., Mugume, S.N., 2025. Ensemble deep learning approaches for enhanced river flow prediction in the transboundary lower semliki river catchment. In: Nones, M., Yang, Y., Varrani, A., Adriasola, J., Carrillo, J.M. (Eds.), 6th IAHR Young Professionals Congress, 3-5 December 2025 (online). International Association for Hydro-Environment Engineering and Research (IAHR), Madrid, Spain and Beijing China, pp. 136–138.

Turhan, E., 2021. A comparative evaluation of the use of artificial neural networks for modeling the rainfall–runoff relationship in water resources management. J. Ecol. Eng. 22, 166–178. https://doi.org/10.12911/22998993/135775.

Ueda, F., Tanouchi, H., Egusa, N., Yoshihiro, T., 2025. Building a generalized pre-training model to predict river water-level from radar rainfall. Water 17, 1–19. https://doi.org/10.3390/w17243449.

Vrugt;, J.A., Robinson, B.A., 2007. Treatment of uncertainty using ensemble methods: comparison of sequential data assimilation and Bayesian model averaging. Water Resour. Res. 43, 1–15. https://doi.org/10.1029/2005WR004838.

Wan, Y., Chen, J., Xu, C.Y., Xie, P., Qi, W., Li, D., Zhang, S., 2021. Performance dependence of multi-model combination methods on hydrological model calibration strategy and ensemble size. J. Hydrol. 603. https://doi.org/10.1016/j.jhydrol.2021.127065.

Wegayehu, E.B., Muluneh, F.B., 2024. Comparing conceptual and super ensemble deep learning models for streamflow simulation in data-scarce catchments. J. Hydrol.: Reg. Stud. 52, 1–24. https://doi.org/10.1016/j.ejrh.2024.101694.

Westerberg, I.K., Guerrero, J., Younger, P.M., Beven, K.J., Seibert, J., Halldin, S., Freer, J. E., 2011. Calibration of hydrological models using flow-duration curves. Hydrol. Earth Syst. Sci. 2205–2227. https://doi.org/10.5194/hess-15-2205-2011.

Xu, H., Zhou, Y., Xia, T., Chen, H., Chang, F., Xu, C., 2026. Elevating predictive reliability: time-varying parameter Bayesian deep learning techniques for flood probability forecasting. J. Hydrol. 664, 1–15. https://doi.org/10.1016/j.jhydrol.2025.134597.

Yilmaz, K.K., Gupta, H.V., Wagener, T., 2008. A process-based diagnostic approach to model evaluation: application to the NWS distributed hydrologic model. Water Resour. Res. 44, 1–18. https://doi.org/10.1029/2007WR006716.

Yin, H., Guo, Z., Zhang, X., Chen, J., Zhang, Y., 2022. RR-Former: Rainfall-runoff modeling based on Transformer. J. Hydrol. 609, 1–18. https://doi.org/10.1016/j.jhydrol.2022.127781.

Young, C.-C., Liu, W.-C., Wu, M.-C., 2017. A physically based and machine learning hybrid approach for accurate rainfall-runoff modeling during extreme typhoon events. Appl. Soft Comput. 53, 205–216. https://doi.org/10.1016/j.asoc.2016.12.052.

Zhu, B., Willems, P., 2025. Ensembles of machine learning and hydrodynamic numerical modeling for salinity simulations in a tidal estuary. J. Hydrodyn. 1–17. https://doi.org/10.2166/hydro.2025.090.

Zounemat-Kermani, M., Batelaan, O., Fadaee, M., Hinkelmann, R., 2021. Ensemble machine learning paradigms in hydrology: a review. J. Hydrol. https://doi.org/10.1016/j.jhydrol.2021.126266.