STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **16**, August 2026, pp 1682–1696.
Published online in International Academic Press (www.IAPress.org)

Flood Forecasting Using Non-linear Autoregressive Exogenous Neural Networks with Radial Basis Function

Sumia Hussin Mohamed Abdualkarim [1], Muhammad Fadhil Marsani [1], Mohsen Ayyash [1,*], and Caiyan Long [1,2]

[1] *School of Mathematical Sciences, Universiti Sains Malaysia, 11800 Penang, Malaysia*
[2] *School of Integrated Circuits and New Energy, Guangzhou College of Technology and Business, Foshan City, 528138, China*

Abstract Accurate river water-level (WL) forecasting is essential for flood preparedness and early-warning decision-making. However, WL fluctuations are highly nonlinear and time-dependent, making accurate prediction challenging when using conventional time-series approaches. This study aims to improve short-term water-level forecasting in the Selangor River basin using an Intelligent Committee Machine Learning (ICML) framework that integrates Radial Basis Function Support Vector Machine Regression (RBF-SVR) and Nonlinear Autoregressive Exogenous (NARX) models. The performance of the proposed framework was evaluated by comparing the individual NARX and RBF models with two ensemble approaches: simple averaging (ENS-AVG) and genetic algorithm-based weighted averaging (ENS-GA). Forecasts were generated at 3-, 6-, and 12-hour lead times using rainfall measurements and historical water-level observations as input variables. Model performance was assessed on the training and testing datasets using standard statistical criteria: root mean square error (RMSE), mean absolute error (MAE), and the coefficient of determination R^2. The results showed that forecasting accuracy declined as the lead time increased, which depicts the greater uncertainty associated with longer-horizon hydrological prediction. Using the training dataset, ENS-GA achieved the best overall performance across the three lead times, which indicates the effectiveness of GA-based weight optimisation. Using the testing dataset, however, NARX consistently outperformed the other models, producing the lowest RMSE and MAE and the highest R^2 values at the 3-, 6-, and 12-hour horizons. At the 6-hour testing horizon, NARX achieved an RMSE of 0.26922, MAE of 0.16238, and R^2 of 0.88934. Although the ensemble architectures demonstrated superior precision to the baseline RBF-SVM model, they failed to outperform the individual NARX model in generalisation capability. Hence, the study emphasises that temporal NARX modelling provides highly reliable short- and medium-term WL forecasts, which reinforces its usefulness for flood forecasting in the study area.

Keywords Intelligent Committee Machine Learning (ICML) models, Non-linear Autoregressive Exogenous (NARX) models; Radial Basis Function (RBF); Flood forecasting

AMS 2010 subject classifications 62M10, 62M20, 62P12, 68T05.

DOI: 10.19139/soic-2310-5070-3471

1. Introduction

Floods are common and destructive phenomena in many regions of the world and remain a major global challenge, causing significant destruction, income losses, and population displacement. During the period 2000–2018, the world experienced a total flood-inundation area of approximately 2.23 million km^2, affecting about 290 million people [1]. Intense precipitation in upstream watersheds can cause rapid and substantial rises in river water levels, increasing the risk of severe flooding in downstream catchments [2]. Therefore, accurate monitoring of river water

*Correspondence to: Mohsen Ayyash. School of Mathematical Sciences, Universiti Sains Malaysia, 11800 Penang, Malaysia. Email: [mhayyash@usm.my; ayash.mohsen@gmail.com]

ISSN 2310-5070 (online) ISSN 2311-004X (print)

level (WL) and flow rate is essential for effective flood forecasting. Reliable WL and discharge information can strengthen early-warning systems and support timely structural and non-structural mitigation measures to reduce risks to human life and property [3].

Recent years have shown increasing interest in the use of modern technologies, including satellite imagery and hydrological models, to improve flood forecasting accuracy. At the same time, climate change and more unstable weather patterns have increased the demand for reliable forecasting systems that can support timely flood preparedness and response [4]. Machine learning (ML) has also contributed substantially to improving short- and long-term forecasting performance, with recent developments emphasising hybrid models, data decomposition techniques, ensemble learning, and optimised neural-network architectures [5].

Although classical linear time-series techniques have achieved good forecasting performance in previous studies [6, 7, 8, 9], their effectiveness may be limited when flood processes involve nonlinear responses, rapid temporal changes, and complex hydrological interactions. Modern statistics and machine learning have introduced a wide range of nonlinear predictive approaches, including artificial neural networks (ANNs), deep learning, decision trees (DTs), random forests (RFs), and support vector machines (SVMs), which enable researchers to analyse large and complex datasets and extract meaningful predictive patterns [10]. Across various countries, including Cameroon, Bangladesh, Southeast Europe, Malaysia, Japan, and Colombia, ANN-based models have demonstrated promising performance in flood forecasting applications [11, 12, 13, 14, 15, 16]. Convolutional neural networks (CNNs) have also been employed for rapid urban flood prediction by integrating rainstorm–inundation databases with conventional modelling frameworks [17]. Similarly, a hybrid model combining AdaBoostRegressor with Lion Swarm Optimisation was proposed to improve forecasting accuracy [18]. Frisk and Johansson compared the forecasting performance of the Stormwater Management Model (SWMM), a hydrodynamic model, with machine-learning approaches based on the NARX framework [19]. Kurt examined the meteorological drivers of water-level variations in the Great Lakes basin using Multiple Linear Regression (MLR), Nonlinear Autoregressive with Exogenous inputs (NARX), Facebook Prophet (FB-Prophet), and Long Short-Term Memory (LSTM) networks [20]. Duangkhwan et al. [21] applied Gated Recurrent Units (GRUs), LSTM models, and CNNs, reporting that GRUs outperformed LSTM models in their forecasting setting. Jiken et al. [22] investigated the use of sediment grain size as an indicator for assessing flood-related damage. Therefore, artificial intelligence-based models have demonstrated superior capabilities for simulating complex physical flood processes [23], while ANNs continue to offer substantial potential for environmental modelling, hydrology, and weather prediction [24]. Ahmad et al. [25] compared RF, DT, and Feed Forward Neural Network (FFNN) models with the NAM rainfall–runoff model for the Vesubie River in France and found that FFNN was comparatively less efficient. CNNs have also estimated historical flood depth and duration from RF records and showed value for flood-impact modelling [26].

Nonlinear Autoregressive with Exogenous Inputs (NARX) neural networks are effective for modelling complex nonlinear time-series relationships and generating accurate short-term flood forecasts by incorporating historical observations and exogenous inputs [27, 28]. Renteria Mena et al. [16] proposed a multivariate NARX model that outperformed a simplified ARX model for water-level prediction in the Atrato River, Colombia [16]. They also compared linear and nonlinear ARX/NARX models with LSTM models for water-level forecasting at stations along the Atrato River. Similarly, Ruslan et al. [28] indicated that NARX provided superior flood-water-level predictions compared with the Extended Kalman Filter (EKF), while Zehra [29] showed that NARX outperformed Support Vector Machine (SVM)-based flood-estimation methods. Faruq et al. [3] further indicated that the Nonlinear Autoregressive Neural Network with Exogenous Inputs (NARXNN) produced lower forecasting errors than Radial Basis Function (RBF) models for 12-hour-ahead predictions. In addition to NARX, RBF models are effective for handling nonlinear and dynamic systems, offering a relatively simple structure for approximating continuous functions in real-time forecasting applications [30]. For example, a study conducted by Moradian et al. [31] proposed a two-step hydrodynamic–machine learning framework for flood modelling in a coastal–fluvial regime, comparing SVR, SVM, RBF, LR, GPR, DT, and ANN models at the pixel level, and found that the RBF model achieved the best performance for rainfall–runoff modelling in the Cork City, Ireland. Another study conducted by Tiwari et al. [32] showed that the wavelet-enhanced WMTM5NN model improved hydrological forecasting accuracy, achieving R^2 values close to 0.97 and outperforming conventional RBFNN and MTM5NN models.

Although NARX and RBF models have demonstrated strong individual forecasting performance in hydrological and flood-related applications, their optimal integration for ensemble-based flood-water-level prediction remains underexplored. Previous studies have proposed Intelligent Committee Machine Learning Flood Forecasting (ICML-FF) models that combine methods such as RBF, ANFIS, SVM, and LSTM using averaging, K-NN, and genetic algorithm (GA)-based strategies, with applications to the Kelantan River in Malaysia [2, 3]. The literature suggests that individual models, such as ANFIS and RBF, can achieve strong forecasting performance in specific applications; however, ensemble-based committee models often provide greater robustness by combining the strengths of multiple learners. Earlier comparative studies involving NARX and RBF mainly focused on the performance of the individual models and did not develop an optimised ensemble framework that explicitly integrates both approaches. In addition, existing ensemble studies have commonly combined ANFIS, SVM, LSTM, and RBF models but have rarely incorporated NARX, despite its established ability to capture nonlinear temporal dependencies in hydrological time series data. This highlights a clear methodological gap in the development of a GA-optimized NARX-RBF ensemble model for short-term flood water-level forecasting.

Thus, it is vital to address this gap for the Selangor River basin, where monsoon-driven floods and rising water levels continue to threaten downstream communities, infrastructure, and flood management systems. Accurate short-term water-level forecasting is therefore essential for strengthening early-warning capacity, supporting timely decision-making, and reducing flood-related damage. Although optimisation algorithms have been applied to enhance individual forecasting models, such as particle-filter-enhanced SVM [33], Grey Wolf Optimiser-based ANFIS [34], Harris Hawks Optimisation-based ANFIS [35], and GA-based predictive models [36, 37, 38], few studies have focused on optimising the weights of NARX–RBF ensemble forecasts for flood-prone Malaysian river basins. Simple averaging treats all models as equally important, which may not represent their actual predictive contributions. In contrast, GA-based weighted averaging allows model-specific weights to be assigned according to forecasting performance, thereby improving the joint predictive capability of the ensemble [39]. Previous studies have also shown that weighted voting and optimised weighting schemes can enhance ensemble accuracy and stability compared with equal-weighting methods [40, 41, 42, 43]. Moreover, appropriate lag structures are important in time-series forecasting because lagged variables help models capture temporal dependencies and improve prediction accuracy in hydrological and other dynamic systems [42, 44].

This study aims to improve short-term river water-level forecasting for the Selangor River by integrating Nonlinear Autoregressive with Exogenous Inputs (NARX) and Radial Basis Function Support Vector Machine Regression (RBF-SVR) models within an Intelligent Committee Machine Learning framework. In particular, the current study compares the forecasting performance of individual NARX and RBF-SVR models with two ensemble strategies: simple averaging (ENS-AVG) and genetic algorithm-based weighted averaging (ENS-GA). Forecasts are generated at 3-, 6-, and 12-hour lead times, and model accuracy is evaluated using root mean square error (RMSE), mean absolute error (MAE), and the coefficient of determination (R^2). By developing a GA-optimized NARX–RBF ensemble model with data-driven lag inputs, this study addresses the lack of an optimised ensemble forecasting framework for the Selangor River and contributes to more reliable short-term flood early-warning systems.

2. Data and Methodology

2.1. Data and study variables

The study focuses on the Selangor River in Malaysia, which originates near Kuala Kubu Bharu in the east and flows westward before discharging into the Strait of Malacca at Kuala Selangor. The river is the most important to the ecology and economy of the area, as it supplies all households, agriculture and industry, especially in the highly populated areas of Selangor, with the required water. The river corridor also exhibits high biodiversity, with diverse flora and fauna supporting activities such as agriculture and aquaculture. Additionally, the Selangor River is an important component of local flood control, with adequate river flow management contributing to the minimisation of flood risks during heavy rainfall and to the enhancement of drainage system performance [45]. Figure 1 shows the location of the study area. ‘

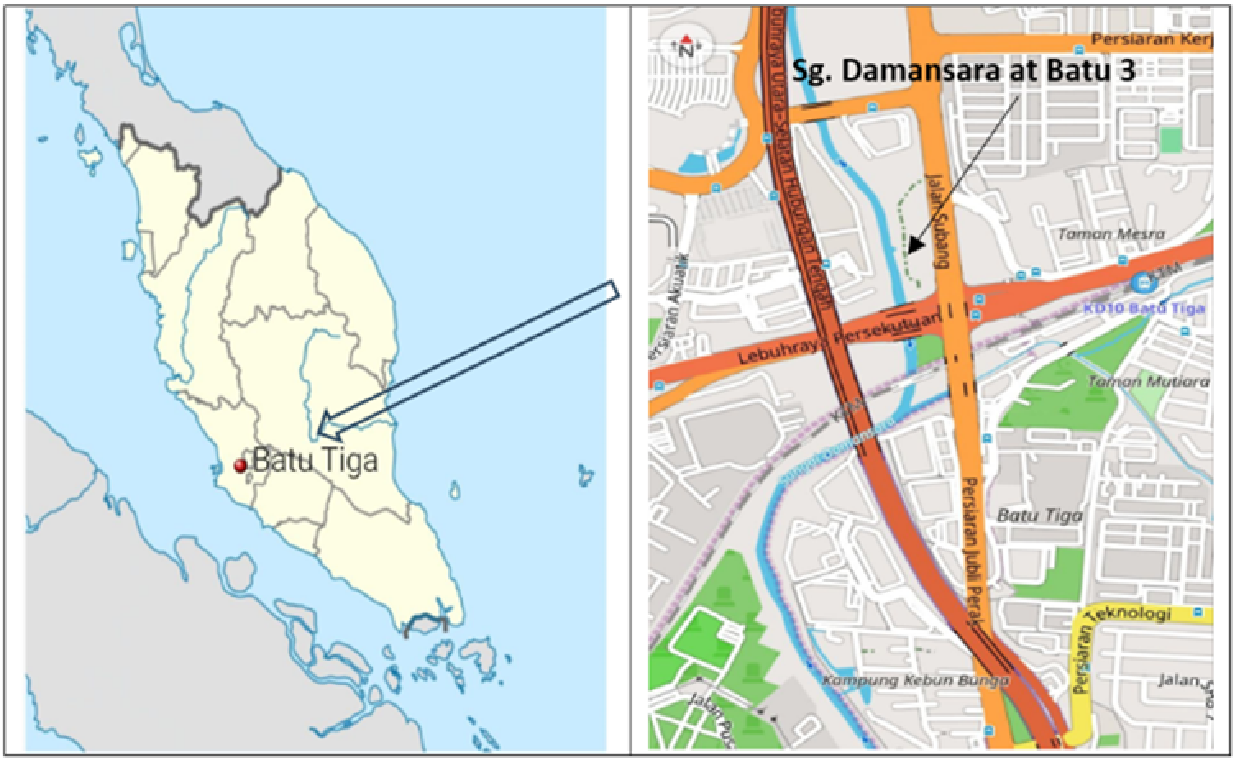

Figure 1. Location of the study area

WL measurements were collected over a specified period. RF data were obtained from station LDG at Segi Tinggi, Bahagian Mary, Selangor (station ID: 0210071RF), which recorded total rainfall in millimetres. Water-level measurements were collected at Rantau Panjang, Selangor (station ID: 0210241WL), in metres. The data covered the period from January 1, 2010, at 00:00 hours to December 31, 2023, at 23:00 hours, which forms a total of 122,712 hours of continuous measurements. The lagged values of the WL were also incorporated into the model to enhance its predictive accuracy. As illustrated in Table 1, the final input parameters comprise concurrent rainfall and lagged water-level data, which represent key hydrological drivers associated with flood dynamics.

Table 1. Input variables

Outcome variable	Predictor variables		
Water level (WL)	Rainfall (RF)	WL-lag1	WL-lag2
Value at (t)	Value at (t)	Value at $(t-1)$	Value at $(t-2)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
Value at (t_n)	Value at (t_n)	Value at (t_{n-1})	Value at (t_{n-2})

2.2. Data Preprocessing

2.2.1. Data imputation and normalisation

To ensure a complete and continuous dataset, the initial phase of data preprocessing addressed missing values using linear interpolation. Linear interpolation for estimating missing values can be formulated as shown in Equation (1):

$$y = y_1 + \frac{y_2 - y_1}{x_2 - x_1}(x - x_1) \tag{1}$$

where y is the estimated value at time step x, x_1 and x_2 denote the time steps of the nearest valid observations immediately preceding and succeeding the missing observation, respectively, and y_1 and y_2 represent their corresponding recorded values.

The study also normalises the data to the standard range of 0 to 1 to simplify model computation and improve accuracy by mitigating the impact of outliers, minimising noise, and stabilising data variance [51]. The min–max normalisation procedure used in this study is expressed as follows:

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{2}$$

where X is the original value, X' is the normalized value, $X_{\min}$ is the minimum observed value of the variable, and $X_{\max}$ is the maximum observed value of the variable.

2.2.2. *Data partitioning*

To measure the robustness and generalisability of the model, the dataset was divided into two segments: 80% for training and 20% for testing. Regarding the 3-hour time period, the total dataset comprised 40,903 samples, of which 32,722 were assigned to the training phase and 8,181 to the testing phase. For 6-hour forecasting periods, the dataset comprised 20,452 samples, of which 16,362 were assigned to training and 4,090 to testing. Finally, for the 12-hour time period, the total number of samples was 10,226, with 8,181 allocated to training and 2,045 to testing.

2.3. *Methodology*

The forecasted water level, M_{t+i}, was modelled as a function of the current rainfall measurement, RF_t, and the two previous water level observations, WL_{t-1} and WL_{t-2}. The forecasting structure is represented as follows:

$$M_{t+i} = f\left(RF_t, WL_{t-1}, WL_{t-2}\right), \quad i \in \{3, 6, 12\} \tag{3}$$

where t denotes the current time step, i represents the forecasting horizon, which was set to 3, 6, and 12 steps ahead, WL refers to river water level measured in metres, and RF denotes rainfall measured in mm/h. The function $f(\cdot)$ represents the intelligent forecasting model, which was estimated using either the RBF, NARX, or ICML algorithm.

2.3.1. ***Nonlinear Autoregressive Network with Exogenous Inputs***

NARX is widely used in machine learning for modelling nonlinear dynamic systems. NARX models are discrete-time nonlinear autoregressive models that incorporate both past values of the target variable and past values of external input variables [47]. Its structure enables the network to capture temporal dependencies by incorporating lagged outputs, lagged inputs, and exogenous predictors in the forecasting process [48]. The general mathematical form of the NARX model is expressed as follows:

$$y(t) = f\left[y(t-1), y(t-2), \ldots, y(t-n_y), u(t-1), u(t-2), \ldots, u(t-n_u)\right] \tag{4}$$

where $y(t)$ is the output variable at time t, $y(t-1), y(t-2), \ldots, y(t-n_y)$ are the lagged output values, and n_y is the maximum output delay. The terms $u(t-1), u(t-2), \ldots, u(t-n_u)$ represent the lagged exogenous input variables, where n_u is the maximum input delay. In this study, the exogenous variables included hydrological and meteorological predictors, mainly RF and WL measurements. The lagged output values represent the model memory and allow the network to capture the temporal dependence in flood-level behaviour. The use of input and output delays enables the model to account for time-lag effects and nonlinear relationships in the hydrological time series.

A recurrent NARX neural network was employed in this study and implemented in a closed-loop mode to capture temporal dependencies. The model was used for multi-step-ahead water-level forecasting at lead times of 3, 6, and 12 hours. The network architecture consisted of one hidden layer with five neurons, selected based on preliminary trial-and-error experiments to balance model complexity and predictive performance. A sigmoid activation function was used in the hidden layer, whereas a linear activation function was applied in the output layer, which is suitable for regression-based forecasting tasks. Grid search with cross-validation was used to tune hyperparameters and confirmed that 5 hidden neurons provided an appropriate trade-off between forecasting accuracy and computational efficiency [52].

2.3.2. *Radial Basis Function Support Vector Regression*

SVR is a kernel-based learning method designed to estimate nonlinear relationships between a set of input predictors and a continuous response variable. In this study, SVR with an RBF (RBF-SVR) kernel was employed to forecast river water level using rainfall and lagged water-level observations as predictors. The input vector was defined as $\mathbf{x}_t = (RF_t, WL_{t-1}, WL_{t-2})$, where RF_t denotes rainfall at time t, and WL_{t-1} and WL_{t-2} denote the first and second lagged water-level observations, respectively. The response variable was the forecasted water level WL_{t+h}, where h represents the forecasting horizon of 3, 6, or 12 hours.

The RBF kernel was selected because of its ability to capture nonlinear relationships by mapping the input variables into a higher-dimensional feature space [53, 55]. The RBF kernel between two input vectors, $\mathbf{x}_i$ and $\mathbf{x}_j$, is defined as

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right), \tag{5}$$

where $\|\mathbf{x}_i - \mathbf{x}_j\|^2$ is the squared Euclidean distance between the two input vectors and σ is the kernel-width parameter that controls the smoothness of the fitted function.

The SVR prediction function for a new input vector $\mathbf{x}$ is expressed as

$$\hat{y}(\mathbf{x}) = \sum_{i \in SV} (\alpha_i - \alpha_i^*) K(\mathbf{x}_i, \mathbf{x}) + b, \tag{6}$$

where $\hat{y}(\mathbf{x})$ is the predicted water level, SV denotes the set of support vectors, α_i and α_i^* are the dual coefficients estimated during model training, and b is the bias term.

The ϵ-SVR model was estimated by minimising the following objective function:

$$\min_{\mathbf{w}, b, \xi_i, \xi_i^*} \frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^*), \tag{7}$$

subject to

$$\begin{aligned} y_i - \left(\mathbf{w}^T \phi(\mathbf{x}_i) + b\right) &\leq \epsilon + \xi_i, \\ \left(\mathbf{w}^T \phi(\mathbf{x}_i) + b\right) - y_i &\leq \epsilon + \xi_i^*, \\ \xi_i, \xi_i^* &\geq 0, \end{aligned} \tag{8}$$

where C is the regularisation parameter controlling the trade-off between model complexity and prediction error, ϵ defines the width of the insensitive loss zone, and ξ_i and ξ_i^* are slack variables that allow deviations beyond the ϵ-tube.

In the empirical implementation, the RBF-SVR model was trained separately for the 3-, 6-, and 12-hour forecasting horizons. The main hyperparameters, including the regularisation parameter C, the kernel-width parameter σ, and the ϵ-insensitive loss parameter, were selected based on model-tuning procedures to improve predictive accuracy and avoid overfitting. After training, the fitted SVR models were used to generate multi-step-ahead forecasts of river water level [54, 55, 56].

2.3.3. *Intelligent Committee Machine Learning*

The Intelligent Committee Machine Learning (ICML) framework combines the strengths of the RBF-SVR and NARX models to improve overall prediction accuracy. By integrating forecasts from both models, the ensemble approach reduces model-specific errors and produces more robust, accurate water-level predictions. The proposed combination of these models represents the main contribution of the present study. The proposed ICML framework was implemented in three stages. First, water-level forecasts were generated independently using the RBF-SVR and NARX models. Second, the individual forecasts were combined using a simple averaging ensemble method. This method calculates the arithmetic mean of the raw outputs of the two models, assuming that both models contribute equally to the final prediction. It does not require transformation or weighting and therefore provides a baseline ensemble forecast. Third, an optimised weighted averaging approach was applied. In this stage, the weights assigned to the RBF-SVR and NARX forecasts were optimised using a Genetic Algorithm (GA) to determine the optimal contribution of each model. The weighted averaging was applied directly to the raw forecast outputs without additional transformation. The final ensemble forecast was generated for the 3-, 6-, and 12-hour forecasting horizons.

Simple Averaging Ensemble The simple averaging ensemble (ENS-AVG) computes the ensemble prediction as the arithmetic mean of the individual model outputs. In this study, the forecasts obtained from the RBF-SVR and NARX models were combined as follows:

$$\hat{y}_{\text{ensemble}} = \frac{1}{2}\left(\hat{y}_{\text{RBF}} + \hat{y}_{\text{NARX}}\right) \tag{9}$$

where $\hat{y}_{\text{ensemble}}$ is the ensemble forecast, $\hat{y}_{\text{RBF}}$ is the forecast produced by the RBF-SVR model, and $\hat{y}_{\text{NARX}}$ is the forecast produced by the NARX model. This approach is straightforward to implement and can be effective when the errors of the individual models are weakly correlated. However, simple averaging assigns equal weight to all models regardless of their predictive performance, which may reduce the overall accuracy of the ensemble [3]. To overcome this limitation, the present study also employed a GA-based weighted averaging method.

Genetic-Algorithm Weighted Averaging The Genetic-Algorithm weighted averaging ensemble (ENS-GA) assigns model-specific weights to the forecasts based on the predictive performance of each model. The weights are optimised to minimise the overall prediction error [50]. In this study, the GA was used to determine the optimal weights by simulating the natural process of evolution. The algorithm begins with an initial population, where each individual represents a candidate set of weights for the base models. The fitness of each individual is then evaluated using an objective function based on prediction error and model constraints. Individuals are ranked according to their fitness, and new generations are produced through selection, crossover, and mutation. This process continues until the stopping criterion is reached, after which the best set of weights is selected for ensemble prediction [36].

In the present implementation, the GA optimised the weights of the RBF-SVR and NARX forecasts by minimising a composite error measure based on the Huber loss. The weights were encoded as real-valued chromosomes, with each chromosome representing the weight assigned to a specific model. The GA was configured with a population size of 50 and a maximum of 100 iterations, with the initial weight vectors generated randomly within the interval $[0, 1]$.

The GA monitored the best fitness value across generations. If no improvement in the best fitness value was observed over 50 consecutive generations, the algorithm was assumed to have converged and was stopped early. The objective function was designed to identify the optimal weight vector $\mathbf{W} = (w_1, w_2)$ that minimises the total Huber loss across the validation dataset, as follows:

$$\min_{\mathbf{W}} J(\mathbf{W}) = \sum_{i=1}^{N} L_\delta\left(y_i - \hat{y}_i(\mathbf{W})\right) \tag{10}$$

where y_i is the observed value at time step i, and $\hat{y}_i(\mathbf{W})$ is the weighted ensemble forecast, defined as

$$\hat{y}_i(\mathbf{W}) = w_1 \hat{y}_{\text{RBF},i} + w_2 \hat{y}_{\text{NARX},i}. \tag{11}$$

The Huber loss function $L_\delta(r)$ is defined as

$$L_\delta(r) = \begin{cases} \frac{1}{2}r^2, & \text{if } |r| \leq \delta, \\ \delta\left(|r| - \frac{1}{2}\delta\right), & \text{if } |r| > \delta, \end{cases} \tag{12}$$

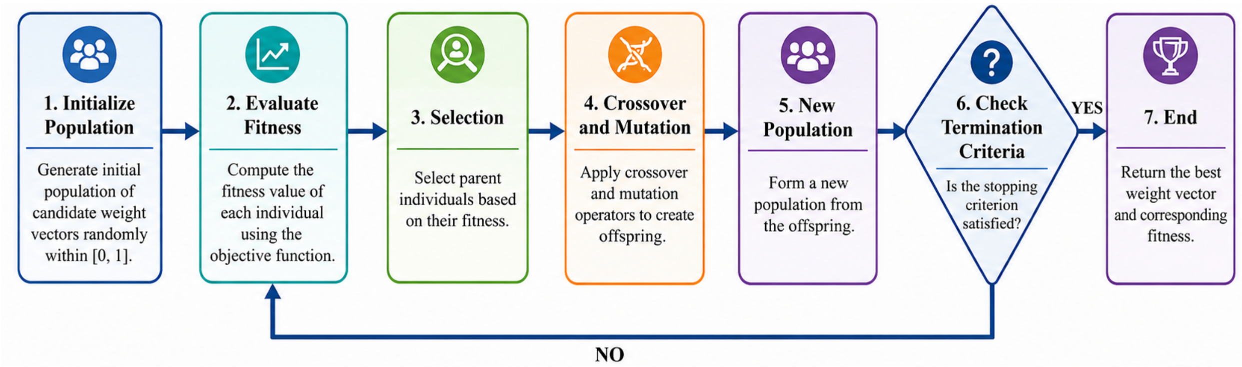

Figure 2. Flowchart of the proposed genetic algorithm

where $r = y_i - \hat{y}_i(\mathbf{W})$ is the residual, and δ is the threshold parameter. In this study, $\delta = 1$ was used.

Finally, the optimised ensemble forecast was obtained by combining the base model predictions using the optimal weights:

$$\hat{y}^*_{\text{ensemble}} = w_1 \hat{y}_{\text{RBF}} + w_2 \hat{y}_{\text{NARX}}, \tag{13}$$

subject to the following constraints:

$$w_1, w_2 \in [0, 1], \qquad w_1 + w_2 = 1. \tag{14}$$

Figure 2 illustrates the proposed GA framework. Candidate weight vectors are first initialised within the interval $[0, 1]$, followed by iterative fitness evaluation, selection, crossover, and mutation until the stopping criterion is met and the optimal weight vector is identified.

2.3.4. *Performance Evaluation Metrics*

The forecasting performance of the proposed models was evaluated using three commonly used regression accuracy metrics: root mean square error (RMSE), mean absolute error (MAE), and the coefficient of determination (R^2) [57, 58, 59]. RMSE quantifies the overall size of forecasting errors while assigning greater influence to larger deviations, whereas MAE expresses the mean absolute difference between observed and predicted values in the original measurement scale. The R^2 statistic evaluates the proportion of variation in the observed values explained by the model. Lower values of RMSE and MAE, together with higher values of R^2, indicate better forecasting performance. The evaluation metrics are defined as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{15}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{16}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{17}$$

where n is the number of observations, y_i is the observed value, $\hat{y}_i$ is the predicted value, and $\bar{y}$ is the mean of the observed values. Figure 3 illustrates the overall methodological framework.

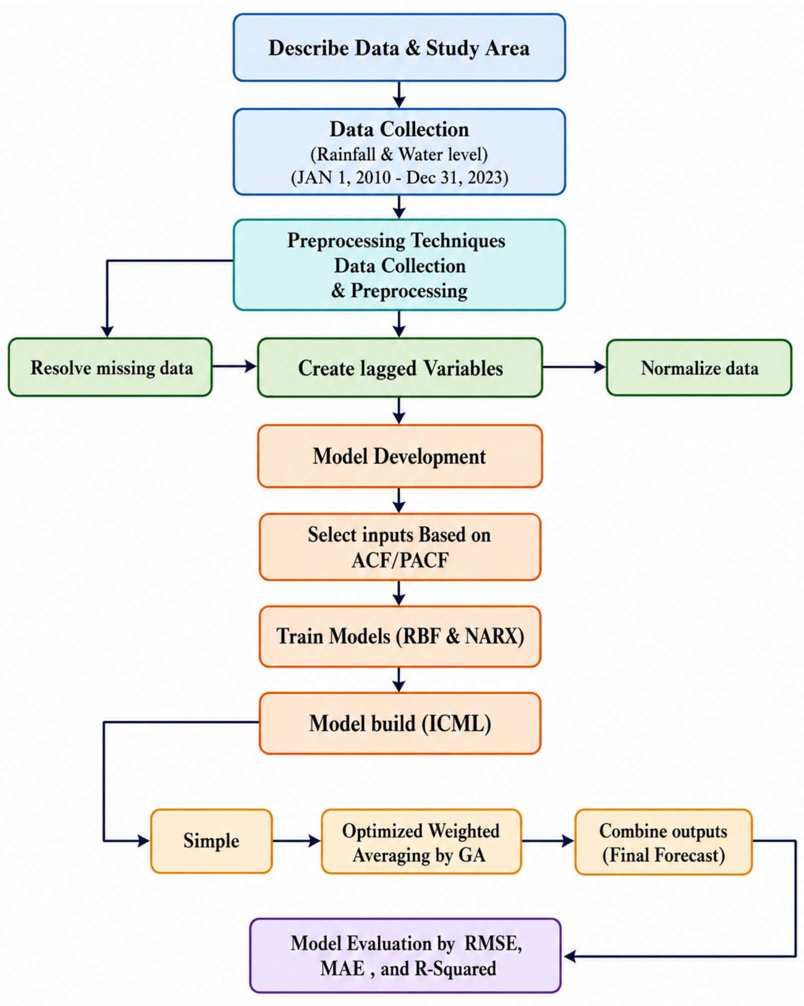

Figure 3. Flowchart of the methodology

3. Results and Discussions

3.1. *Descriptive statistics*

Table 2 shows the descriptive statistics of the WL and RF variables across the three different sampling periods. The descriptive statistics show that WL remained relatively stable, with mean values ranging from 4.9803 to 4.9817 m and coefficients of variation (CV) close to 0.155. In contrast, RF exhibited substantially greater variability, as indicated by high CVs ranging from 8.1721 to 13.5556. RF was also highly right-skewed and leptokurtic, particularly at the 12-hour interval, indicating an intermittent pattern dominated by many zero or low-rainfall observations and a limited number of extreme rainfall events.

3.2. *Training and Testing Performance of the Forecasting Models*

Table 3 displays the predictive performance of the RBF-NN, NARX-NN, ENS-AVG, and ENS-GA models across 3-, 6-, and 12-hour forecasting horizons during both the training and testing phases. During training, all four models demonstrated high predictive accuracy at the 3-hour horizon, with RMSE values ranging from 0.08969 to 0.09142

Table 2. Descriptive statistics of water level and rainfall across different sampling intervals

Sampling interval	Variable	Mean	Median	SD	Minimum	Maximum	Skewness	Kurtosis	CV
3 hours	Water level (m)	4.9807	4.76	0.7743	2.78	9.57	1.2464	4.5313	0.1554
	Rainfall (mm/h)	0.2722	0.00	2.2246	0.00	156.00	20.3124	794.1443	8.1721
6 hours	Water level (m)	4.9803	4.76	0.7741	2.78	9.55	1.2469	4.5321	0.1554
	Rainfall (mm/h)	0.2749	0.00	2.3660	0.00	156.00	24.1155	1086.175	8.6040
12 hours	Water level (m)	4.9817	4.77	0.7728	2.78	9.52	1.2298	4.4980	0.1551
	Rainfall (mm/h)	0.1446	0.00	1.9602	0.00	156.00	54.4227	4000.494	13.5556

and R^2 values exceeding 0.985. The ENS-GA model consistently achieved the best training performance across all forecasting horizons, producing the lowest RMSE values of 0.08969, 0.21304 and 0.34942 for the 3-, 6-and 12-hour horizons, respectively and the highest R^2 values. Nevertheless, the performance differences among the models were relatively small during training, particularly at the shortest forecasting horizon. This indicates that all models successfully captured the underlying patterns within the training data.

When the prediction horizon was expanded from 3 to 12 hours, forecasting accuracy declined across all architectures. This decay pattern aligns perfectly with recent empirical studies in hydrological time-series modelling (2023–2025), which demonstrate that expanding the forecast lead time exponentially accentuates error propagation due to cumulative lag uncertainties in rainfall-runoff mechanisms [60, 61].

The examination of the testing dataset revealed more pronounced discrepancies in generalisation capacity, highlighting the structural advantages of certain architectures. The NARX-NN model achieved the strongest generalisation capability across all forecasting horizons. At the 3-hour horizon, it produced the lowest RMSE of 0.12542, the lowest MAE of 0.06441, and the highest R^2 of 0.97598. The outperformance is sustained at the 6- and 12-hour horizons, with the NARX-NN maintaining robust R^2 values of 0.88934 and 0.76462, respectively.

The findings of the current study are in line with recent literature that reported recurrent architectures outperform static feedforward networks over medium-term horizons [62, 63]. The explicit benefit of the NARX-NN lies in its internal memory feedback loops. Recent benchmarks by Liu et al. [62] confirm that while static models evaluate data as isolated snapshots, recurrent feedback configurations effectively internalise the dynamic catchment response time and physical inertia of river basins, making them much more resilient to extended lead times.

Conversely, while the static RBF-NN model proved highly competent using the training dataset, its testing performance deteriorated drastically over extended lead times, with its R^2 dropping sharply from 0.85921 at 3 hours to 0.55424 at the 12-hour horizon. The substantial drop in performance indicates a classic susceptibility to overfitting and an inability to adapt to unseen hydrological variability. It reflects a limitation of static neural networks in handling non-linear time-series extrapolation, a limitation widely critiqued in recent literature, which highlights that architectures lacking temporal data-routing mechanisms inherently fail when extreme flow variations occur downstream [64]. Nonetheless, the ENS-GA framework failed to outperform the standalone NARX-NN model in the testing dataset. It can be explained by a critical constraint identified in recent ensemble learning literature (2024–2026) [65, 66]. The generalisation limits of a weighted ensemble are fundamentally constrained by the performance floor of its weakest base learner. Because the RBF-NN suffered severe performance degradation during testing, it acted as a ”predictive drag” on the ensemble. Even though the GA-driven weighting mechanism successfully minimised the impact of the weaker model by shifting numerical focus onto the dominant predictor during training, the structural superiority of the independent, temporally informed NARX-NN remained unmatched when confronting entirely unseen testing data [65, 66]. Recent literature and our empirical results confirm that selecting a fundamentally robust, recurrent network topology remains the most critical factor for reliable, operational flood-warning frameworks [63, 66].

Thus, the results demonstrate the expected decline in forecasting accuracy as the prediction horizon increases, reflecting the greater uncertainty associated with longer-lead hydrological forecasting. Although ENS-GA achieved the best performance during training, NARX-NN delivered the most reliable test performance and therefore demonstrated the strongest generalisation ability. The findings support the use of temporally informed neural network models for water-level forecasting and confirm that ensemble methods can enhance robustness by integrating complementary predictive models [57, 55].

Table 3. Training and testing performance of the forecasting models

Algorithm	T ahead hour	Training			Testing		
		RMSE	MAE	R^2	RMSE	MAE	R^2
RBF-SVR	3	0.09142	0.05657	0.98567	0.30365	0.18729	0.85921
	6	0.21559	0.12685	0.92027	0.39906	0.26516	0.75687
	12	0.35500	0.22032	0.78303	0.53906	0.36560	0.55424
NARX-NN	3	0.09031	0.05158	0.98602	0.12542	0.06441	0.97598
	6	0.21548	0.13451	0.92035	0.26922	0.16238	0.88934
	12	0.35171	0.23459	0.78703	0.39171	0.26557	0.76462
ENS-AVG	3	0.08977	0.05307	0.98618	0.18431	0.11813	0.94812
	6	0.21363	0.12895	0.92171	0.30417	0.20246	0.85875
	12	0.35069	0.22498	0.78827	0.43412	0.30003	0.71089
ENS-GA	3	0.08969	0.05221	0.98621	0.13367	0.07386	0.97271
	6	0.21304	0.13211	0.92215	0.27746	0.18402	0.88246
	12	0.34942	0.22989	0.78980	0.39819	0.27679	0.75677

3.3. Comparison of Observed and Predicted Water Levels Across Lead Times

Figure 4 presents the comparison between the observed training water-level series and the fitted values generated by the NARX, RBF-SVR, ENS-AVG, and ENS-GA models at three forecasting horizons: 3 hours (A), 6 hours (B), and 12 hours (C). The subplots show that the model outputs closely follow the temporal fluctuations in the training data, particularly at shorter forecasting horizons. The dense variations in the series reflect the dynamic and highly variable behaviour of water levels over time. As the forecasting horizon increases, greater divergence between observed and predicted values becomes apparent, indicating increasing difficulty in longer-lead forecasting. Overall, the figure provides visual evidence that the models learned the main water-level patterns during training before being evaluated on the test dataset.

Figure 5 compares the observed and predicted water levels for the 3-, 6-, and 12-hour forecasting horizons. Panels A, B, and C show the test results for the 3-, 6-, and 12-hour-ahead forecasts, respectively. The observed water-level series is represented by the test data, while the predicted values are generated by the NARX, RBF-SVR, ENS-AVG, and ENS-GA models. Overall, the predicted series closely follows the observed water-level pattern, particularly at the 3-hour forecasting horizon. However, the deviation between observed and predicted values becomes more noticeable as the forecasting horizon increases, especially for the 12-hour-ahead forecast. This indicates that longer lead-time prediction is more challenging due to increasing uncertainty in the hydrological process. The plots also show that the models generally capture the main fluctuations and peak patterns in water level, which is important for practical flood forecasting applications.

4. Conclusion

This study evaluated the performance of individual and ensemble machine-learning models for forecasting river water levels at 3-, 6-, and 12-hour lead times using rainfall and antecedent water-level data as input variables. The findings confirmed the expected inverse relationship between forecast lead time and model accuracy, with predictive performance decreasing at longer horizons due to the increasing uncertainty and variability of catchment dynamics.

In the training dataset, the ENS-GA model outperformed other models across all horizons, recording the lowest RMSE values and the highest R^2 values, which indicates that the GA was effective in optimising the ensemble weighting structure during model training. However, the testing results showed that the standalone NARX-NN model achieved the strongest generalisation performance across all lead times. By producing the lowest RMSE

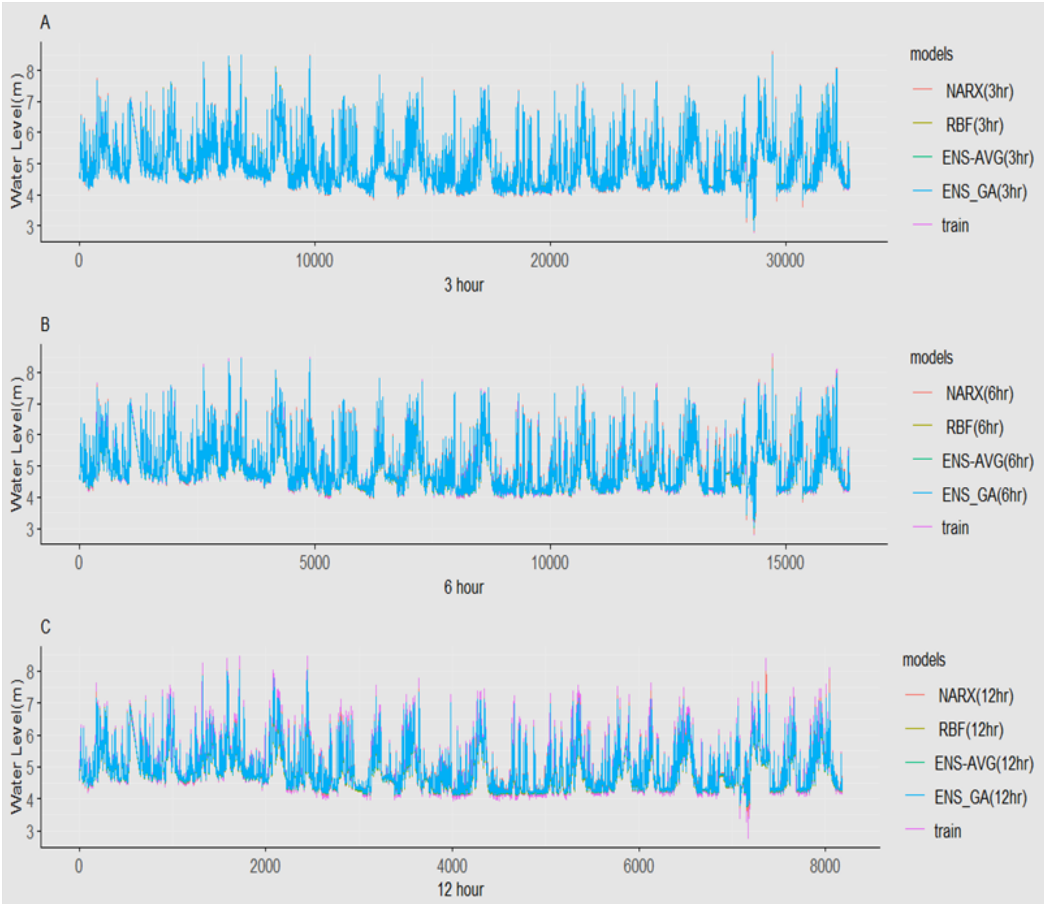

Figure 4. Comparison of observed training data and model-fitted values across the 3-, 6-, and 12-hour forecasting periods

and MAE values, together with the highest R^2 values at the 3-, 6-, and 12-hour horizons, the NARX-NN model demonstrated a superior ability to capture the temporal dependencies between RF, previous WL, and future river-level fluctuations.

The forecasted water-level series closely matched the observed patterns, with the strongest agreement observed at the 3-hour forecasting horizon. As the lead time increased to 6 and 12 hours, the discrepancy between observed and predicted values became more apparent, consistent with the decline in predictive accuracy shown by the evaluation metrics. Among the evaluated models, NARX-NN demonstrated the most stable agreement with the observed data, whereas the ensemble approaches, ENS-AVG and ENS-GA, outperformed the individual RBF-SVR model.

To conclude, the findings highlight the effectiveness of temporally informed neural-network models, particularly NARX-NN, for short- to medium-term river water-level forecasting. Although the proposed ENS-GA framework improved ensemble performance compared with simple averaging and the standalone RBF-SVR model, it did not surpass the generalisation performance of the NARX-NN model on the test data. Therefore, within the scope of this study, NARX-NN can be considered the most reliable model for operational flood forecasting. Future research should consider incorporating additional hydrometeorological predictors, validating the models across different river basins, and exploring advanced hybrid or deep-learning architectures to improve forecasting stability at longer lead times.

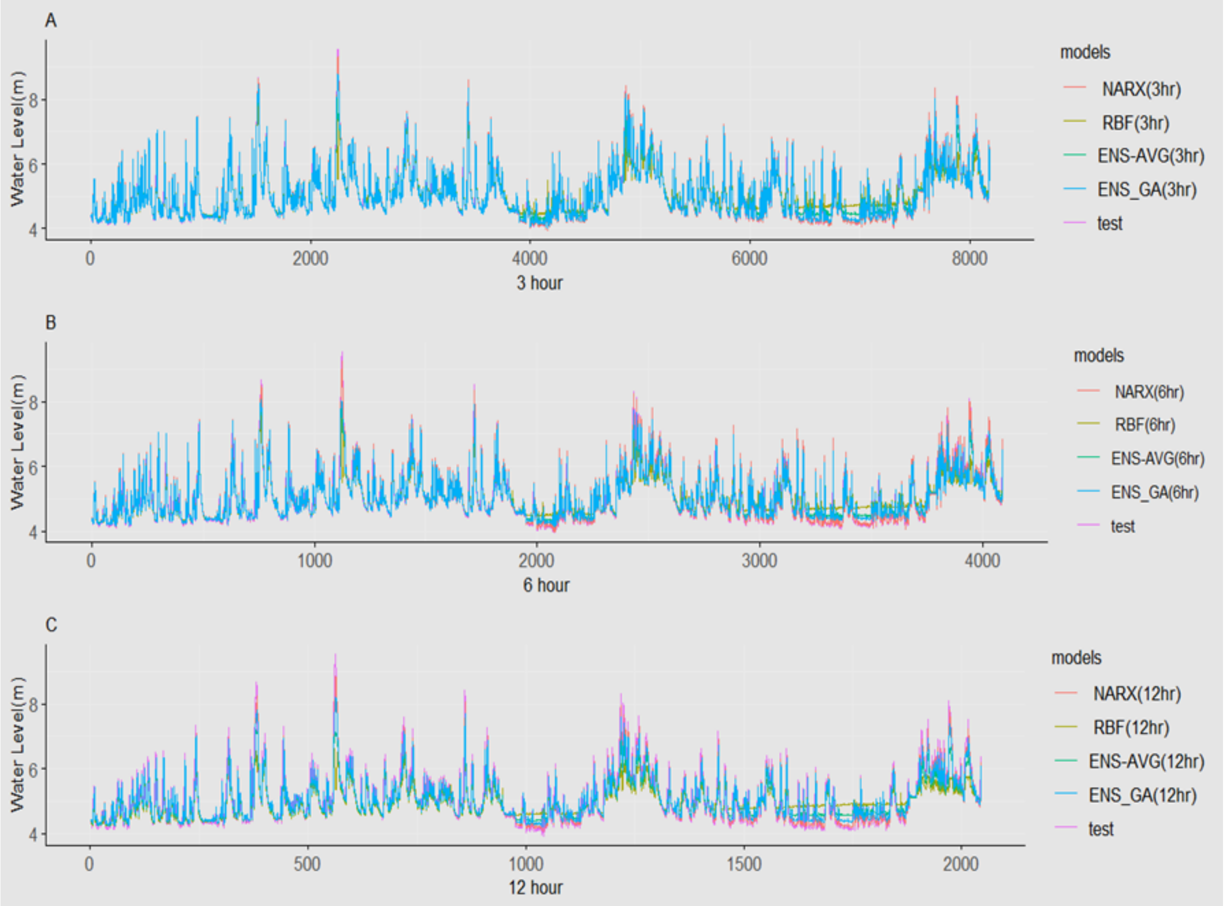

Figure 5. Comparison of observed testing data and model-fitted values across the 3-, 6-, and 12-hour forecasting periods

Acknowledgement

The authors would like to thank the editor and reviewers for their valuable comments and suggestions, which helped improve the quality of the study.

REFERENCES

1. I. Sahdar, D. Rohmat, W. A. Pranoto, and Solehudin, *Hydraulic modeling for flood control scenarios in Akelaka Watershed, North Maluku, Indonesia*, International Journal of GEOMATE, vol. 27, no. 120, pp. 49–59, 2024.
2. A. Faruq, S. F. M. Hussein, A. Marto, and S. S. Abdullah, *Flood river water level forecasting using ensemble machine learning for early warning systems*, IOP Conference Series: Earth and Environmental Science, vol. 1091, no. 1, p. 012041, 2022.
3. A. Faruq, S. S. Abdullah, A. Marto, M. A. A. Bakar, S. F. M. Hussein, and C. M. C. Razali, *The use of radial basis function and non-linear autoregressive exogenous neural networks to forecast multi-step ahead of time flood water level*, International Journal of Advanced Intelligent Informatics, vol. 5, no. 1, pp. 1–10, 2019.
4. T. Sayama, M. Yamada, Y. Sugawara, and D. Yamazaki, *Ensemble flash flood predictions using a high-resolution nationwide distributed rainfall-runoff model: Case study of the heavy rain event of July 2018 and Typhoon Hagibis in 2019*, Progress in Earth and Planetary Science, vol. 7, pp. 1–18, 2019..
5. N. Byaruhanga, D. Kibirige, S. Gokool, and G. Mkhonta, *Evolution of flood prediction and forecasting models for flood early warning systems: A scoping review*, Water, vol. 16, no. 13, p. 1763, 2024.
6. A. Saikhu, A. Z. Arifin, and C. Fatichah, *Rainfall forecasting by using autoregressive integrated moving average, single input and multi input transfer function*, in Proceedings of the 11th International Conference on Information and Communication Technology Systems (ICTS), pp. 85–90, 2017.

7. W. M. Wong, M. Y. Lee, A. S. Azman, L. A. F. Rose, and F. P. T. dan Teknousahawan, *Development of short-term flood forecast using ARIMA model*, International Journal of Mathematics, vol. 15, pp. 68–75, 2021.
8. X. Zhang, P. Liu, L. Cheng, Z. Liu, and Y. Zhao, *A back-fitting algorithm to improve real-time flood forecasting*, Journal of Hydrology, vol. 562, pp. 140–150, 2018.
9. W. M. Wong, S. K. Subramaniam, F. S. Feroz, I. D. Subramaniam, and L. A. F. Rose, *Flood prediction using ARIMA model in Sungai Melaka, Malaysia*, International Journal of Advanced Trends in Computer Science and Engineering, vol. 9, no. 4, pp. 5287–5295, 2020.
10. P. Ghorpade, A. Gadge, A. Lende, H. Chordiya, G. Gosavi, A. Mishra, and N. Shaikh, *Flood forecasting using machine learning: A review*, in Proceedings of the 8th International Conference on Smart Computing and Communication (ICSCC), pp. 32–36, 2021.
11. F. Y. Dtissibe, A. A. A. Ari, H. Abboubakar, A. N. Njoya, A. Mohamadou, and O. Thiare, *A comparative study of machine learning and deep learning methods for flood forecasting in the Far-North Region, Cameroon*, Scientific African, vol. 23, pp. 1–18, 2024.
12. J. F. Ruma, M. S. G. Adnan, A. Dewan, and R. M. Rahman, *Particle swarm optimization-based LSTM networks for water level forecasting: A case study on Bangladesh river network*, Results in Engineering, vol. 17, pp. 1–14, 2023.
13. M. Castangia et al., *Transformer neural networks for interpretable flood forecasting*, Environmental Modelling & Software, vol. 160, pp. 1–9, 2023.
14. A. Dehghani et al., *Comparative evaluation of LSTM, CNN, and ConvLSTM for hourly short-term streamflow forecasting using deep learning approaches*, Ecological Informatics, vol. 75, pp. 1–12, 2023.
15. K. Vidyalashmi et al., *Analysing the performance of the NARX model for forecasting the water level in the Chikugo River estuary, Japan*, Environmental Research, vol. 251, no. 1, p. 118531, 2024.
16. J. B. Renteria-Mena, D. Plaza, and E. Giraldo, *Comparative analysis of nonlinear methods for multivariable water level prediction: The case study of the Atrato River*, Journal of Electrical and Computer Engineering, vol. 2024, p. 2894031, 2024.
17. Y. Liao, Z. Wang, X. Chen, and C. Lai, *Fast simulation and prediction of urban pluvial floods using a deep convolutional neural network model*, Journal of Hydrology, vol. 624, p. 129945, 2023.
18. S. E. Priestly, K. Raimond, Y. Cohen, J. Brema, and D. J. Hemanth, *Evaluation of a novel hybrid lion swarm optimization–AdaBoost regressor model for forecasting monthly precipitation*, Sustainable Computing: Informatics and Systems, vol. 39, p. 100884, 2023.
19. F. Frisk and O. Johansson, *Comparative evaluation of water level forecasting using IoT sensor data: Hydrodynamic model SWMM vs. machine learning models based on NARX framework*, Water, vol. 16, no. 19, p. 2776, 2024.
20. O. Kurt, *Model-based prediction of water levels for the Great Lakes: A comparative analysis*, Earth Science Informatics, vol. 17, no. 4, pp. 3333–3349, 2024.
21. W. Duangkhwan, C. Ekkawatpanit, D. Kositgittiwong, W. Kompor, and C. Petpongpan, *Deep learning-based flood inundation prediction in the Pattani River Basin*, GEOMATE Journal, vol. 28, no. 125, pp. 133–140, 2025.
22. Y. Jiken, K. Watanabe, and N. Saito, *Study on sediment grain-size measurement and calculations at multiple points on the sandbar in class B river*, GEOMATE Journal, vol. 24, no. 104, pp. 93–100, 2023.
23. R. Tabbussum and A. Q. Dar, *Performance evaluation of artificial intelligence paradigms—artificial neural networks, fuzzy logic, and adaptive neuro-fuzzy inference system for flood prediction*, Environmental Science and Pollution Research, vol. 28, no. 20, pp. 25265–25282, 2021.
24. M. Uaisova, B. Zharlykassov, D. Aldasheva, A. Artykbayeva, and P. Radchenko, *The use of ANN and machine learning algorithms to predict road surface deterioration*, GEOMATE Journal, vol. 27, no. 121, pp. 136–143, 2024.
25. M. Ahmad, M. A. Al Mehedi, M. M. S. Yazdan, and R. Kumar, *Development of machine learning flood model using artificial neural network (ANN) at Var River*, Liquids, vol. 2, no. 3, pp. 147–160, 2022.
26. H. W. Wang, G. F. Lin, C. T. Hsu, S. J. Wu, and S. S. Tfwala, *Long-term temporal flood predictions made using convolutional neural networks*, Water, vol. 14, no. 24, p. 4134, 2022.
27. J. B. Renteria-Mena, D. Plaza, and E. Giraldo, *Multivariable NARX based neural networks models for short-term water level forecasting*, Engineering Proceedings, vol. 39, no. 1, pp. 60–79, 2023.
28. F. A. Ruslan, A. M. Samad, Z. M. Zain, and R. Adnan, *Flood prediction using NARX neural network and EKF prediction technique: A comparative study*, in Proceedings of the IEEE 3rd International Conference on System Engineering and Technology, Shah Alam, Malaysia, pp. 203–208, 2013.
29. N. Zehra, *Prediction analysis of floods using machine learning algorithms (NARX & SVM)*, International Journal of Sciences: Basic and Applied Research, vol. 49, no. 2, pp. 24–34, 2020.
30. V. Kumar, K. V. Sharma, T. Caloiero, D. J. Mehta, and K. Singh, *Comprehensive overview of flood modeling approaches: A review of recent advances*, Hydrology, vol. 10, no. 7, p. 141, 2023.
31. S. Moradian, A. AghaKouchak, S. Gharbia, C. Broderick, and A. I. Olbert, *Forecasting of compound ocean-fluvial floods using machine learning*, Journal of Environmental Management, vol. 364, p. 121295, 2024.
32. D. K. Tiwari, V. Kumar, A. Goyal, K. M. Khedher, and M. A. Salem, *Comparative analysis of data driven rainfall-runoff models in the Kolar River Basin*, Results in Engineering, vol. 23, p. 102682, 2024.
33. P. Insom et al., *A support vector machine-based particle filter method for improved flooding classification*, IEEE Geoscience and Remote Sensing Letters, vol. 12, no. 9, pp. 1943–1947, 2015.
34. M. Dehghani, A. Seifi, and H. Riahi-Madvar, *Novel forecasting models for immediate-short-term to long-term influent flow prediction by combining ANFIS and grey wolf optimization*, Journal of Hydrology, vol. 576, pp. 698–725, 2019.
35. N. Arya Azar, S. Ghordoyee Milan, and Z. Kayhomayoon, *Predicting monthly evaporation from dam reservoirs using LS-SVR and ANFIS optimized by Harris hawk's optimization algorithm*, Environmental Monitoring and Assessment, vol. 193, no. 11, pp. 1–14, 2021.
36. I. Elkhrachy et al., *Landslide susceptibility mapping and management in Western Serbia: An analysis of ANFIS- and SVM-based hybrid models*, Frontiers in Environmental Science, vol. 11, p. 1218954, 2023.
37. E. Dodangeh et al., *Novel hybrid intelligence models for flood-susceptibility prediction: Meta optimization of the GMDH and SVR models with the genetic algorithm and harmony search*, Journal of Hydrology, vol. 590, p. 125423, 2020.

38. T. C. Yang, P. S. Yu, K. H. Lin, C. M. Kuo, and H. W. Tseng, *Predictor selection method for the construction of support vector machine (SVM)-based typhoon rainfall forecasting models using a non-dominated sorting genetic algorithm*, Meteorological Applications, vol. 25, no. 4, pp. 510–522, 2018.
39. S. K. Bozchaloei and M. Vafakhah, *Regional analysis of flow duration curves using adaptive neuro-fuzzy inference system*, Journal of Hydrologic Engineering, vol. 20, no. 12, p. 06015008, 2015.
40. A. Ekbal and S. Saha, *Weighted vote-based classifier ensemble for named entity recognition: A genetic algorithm-based approach*, ACM Transactions on Asian Language Information Processing, vol. 10, no. 2, 2011.
41. A. Ekbal and S. Saha, *A multiobjective simulated annealing approach for classifier ensemble: Named entity recognition in Indian languages as case studies*, Expert Systems with Applications, vol. 38, no. 12, pp. 14760–14772, 2011.
42. H. Kim, H. Kim, H. Moon, and H. Ahn, *A weight-adjusted voting algorithm for ensembles of classifiers*, Journal of the Korean Statistical Society, vol. 40, no. 4, pp. 437–449, 2011.
43. X. Zhang, J. W. Bao, B. Chen, and W. Huang, *Evaluation and comparison of two deep convection parameterization schemes at convection-permitting resolution*, Monthly Weather Review, vol. 149, no. 10, pp. 3419–3432, 2021.
44. X. Zhou, B. Xue, Y. Wang, and G. Wang, *Evaluating land surface temperature variation and its responses to climate changes and human activities during 1980–2015: A case study of the Yarlung Zangbo River, southeastern Tibetan Plateau*, SSRN preprint, 2015.
45. R. Cheah et al., *Geospatial modelling of watershed peak flood discharge in Selangor, Malaysia*, Water, vol. 11, no. 12, p. 2490, 2019.
46. K. Lukoseviciute and M. Ragulskis, *Evolutionary algorithms for the selection of time lags for time series forecasting by fuzzy inference systems*, Neurocomputing, vol. 73, nos. 10–12, pp. 2077–2088, 2010.
47. H. Xie, H. Tang, and Y.-H. Liao, *Time series prediction based on NARX neural networks: An advanced approach*, in Proceedings of the International Conference on Machine Learning and Cybernetics, vol. 3, 2009.
48. Y. Essam et al., *Predicting streamflow in Peninsular Malaysia using support vector machine and deep learning algorithms*, Scientific Reports, vol. 12, no. 1, pp. 1–46, 2022.
49. I. M. Yassin et al., *Multi-layer perceptron (MLP)-based nonlinear auto-regressive with exogenous inputs (NARX) stock forecasting model*, International Journal of Advanced Science, Engineering and Information Technology, vol. 7, no. 3, pp. 1098–1103, 2017.
50. B. Li et al., *Optimized neural network combined model based on the induced ordered weighted averaging operator for vegetable price forecasting*, Expert Systems with Applications, vol. 168, p. 114232, 2021.
51. J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Morgan Kaufmann, 2012.
52. W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S*, 4th ed. New York: Springer, 2002.
53. V. N. Vapnik, *The Nature of Statistical Learning Theory*, New York: Springer, 1995.
54. C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.
55. A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, no. 3, pp. 199–222, 2004.
56. C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 1–27, 2011.
57. R. J. Hyndman and A. B. Koehler, "Another look at measures of forecast accuracy," *International Journal of Forecasting*, vol. 22, no. 4, pp. 679–688, 2006.
58. C. J. Willmott and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance," *Climate Research*, vol. 30, no. 1, pp. 79–82, 2005.
59. D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R^2 is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Computer Science*, vol. 7, p. e623, 2021.
60. M. S. Rahman, M. S. H. Inan, and M. Al-Asad, "Multi-hour ahead river water level forecasting using machine learning: Analysis of error propagation across extended lead times," *Journal of Hydrology*, vol. 620, p. 129412, 2023.
61. H. Zhang, K. Li, and P. Multi, "Deep learning and recurrent feedback frameworks in flood forecasting: Scaling lead times from hours to days," *Hydrology and Earth System Sciences*, vol. 28, no. 4, pp. 891–910, 2024.
62. Y. Liu, J. Wang, and X. ChenBox, "Recurrent versus static neural networks for multi-step ahead streamflow forecasting: The role of basin physical inertia memory," *Water Resources Research*, vol. 60, no. 2, p. e2023WR035124, 2024.
63. T. H. Nguyen, S. Park, and D. H. Kim, "Temporally aware machine learning architectures for operational flood early-warning systems: A comprehensive benchmark," *Journal of Hydro-environment Research*, vol. 52, pp. 15–29, 2025.
64. S. L. Zubaidi, M. Al-Bahrani, and S. Ortega-Martorell, "Limitations of static feedforward neural networks in long-term hydro-meteorological forecasting under climate variability," *Environmental Modelling & Software*, vol. 165, p. 105718, 2023.
65. L. Chen, X. Gu, and J. Smith, "Evaluating the predictive drag of weak base learners in evolutionary-optimized ensemble frameworks for hydrological time-series," *Information Sciences*, vol. 658, p. 119982, 2024.
66. Z. Wang and P. Liu, "Why evolutionary weighting algorithms fail to salvage weak learners in multi-step hydrological ensemble forecasting," *Advances in Water Resources*, vol. 183, p. 104590, 2025.