Improving Streamflow Forecasting Efficiency Using Signal Decomposition Approaches

Dinesh Kumar Vishwakarma, Salim Heddam, Arpit Gaur, Ravindra Kumar Tiwari, Özgür Kişi, Anurag Malik, Chetak Bishnoi, Abed Alataway, Ahmed Z. Dewidar, Mohamed A. Mattar

Water Resources Management
https://doi.org/10.1007/s11269-025-04258-8

Improving Streamflow Forecasting Efficiency Using Signal Decomposition Approaches

Dinesh Kumar Vishwakarma[1] · Salim Heddam[2] · Arpit Gaur[3] · Ravindra Kumar Tiwari[4] · Ozgur Kisi[5,6] · Anurag Malik[7] · Chetak Bishnoi[7] · Abed Alataway[8] · Ahmed Z. Dewidar[8] · Mohamed A. Mattar[8]

Received: 24 October 2024 / Accepted: 27 May 2025

Abstract
This study introduces a novel approach utilizing the Maximal Overlap Discrete Wavelet Transform (MODWT) to enhance daily streamflow forecasting at two USGS stations (14211500 and 14211550) from 1998 to 2021. The MODWT is integrated with three machine learning models: Extremely Randomized Trees (ERT), Artificial Neural Networks (ANN), and Gaussian Process Regression (GPR). Autocorrelation and partial autocorrelation functions were employed to determine relevant lags and generate multiple input variables, which were then analyzed through MODWT to derive multi-resolution analysis features. The hybrid model incorporating MODWT significantly improved prediction accuracy. Among the methods, ANN with MODWT (ANN6_MODWT) demonstrated superior performance compared to standalone ANN, ERT, and GPR models. ANN6_MODWT achieved improvements of 15.60%, 24.70%, 39.74%, and 28.34% in terms of correlation coefficient (R), Nash-Sutcliffe efficiency (NSE), root mean square error (RMSE), and mean absolute error (MAE) at USGS 14211550, and 13.50%, 23.80%, 46.47%, and 34.06% at USGS 14211500. These results underscore the potential of MODWT for enhancing streamflow prediction accuracy.

Keywords Forecasting · Streamflow · Extremely randomized trees · Artificial neural network · Gaussian process regression · Maximum overlap discrete wavelet transform

Notation

WT	Wavelet transform
MODWT	Maximal overlap discrete wavelet transform
ERT	Extremely randomized trees
ANN	Artificial neural network
GPR	Gaussian process regression
USGS	United states geological survey
R	Pearson Correlation coefficient
NSE	Nash-Sutcliffe efficiency
RMSE	Root mean square error

Extended author information available on the last page of the article

Published online: 24 June 2025

MAE	Mean absolute error
FT	Fourier transform
SVR	Support vector regression
ML	Machine learning
MLR	Multiple linear regression
GWO	Gray Wolf optimization
VMD	Variational mode decomposition
EEMD	Ensemble empirical mode decomposition
DWT	Discrete wavelet transforms
LSTM	Long short-term memory
WNN	Wavelet ANN
BNN	Bootstrap neural network
BWNN	Bootstrap wavelet neural network
WSVR	Wavelet SVR
BSVR	Bootstrap SVR
BWSVR	Bootstrap wavelet SVR
GRNN	General regression neural network
RBFNN	Radial basis function neural network
MLPNN	Multilayer perceptron neural network
PSO	Particle swarm optimization
ANFIS	Adaptive network-based fuzzy inference system
WANFIS	Wavelet ANFIS
LWL	Locally weighted learning
RS	Random-subspace
RF	Random-forest
AR	Additive regression
Bagg	Bagging
Dagg	Dagging
GEP	Gene expression programming
DT	Decision tree
TB	Tree boost
DTF	Decision tree forest
CNN	Convolutional neural network
WI	Willmott's index ()
DL	Deep learning
X_{max}	Mean value of daily discharge (Q)
X_{mean}	Maximum of daily discharge (Q)
X_{min}	Minimum of daily discharge (Q)
S_x	Standard deviation of daily discharge (Q)
C_v	Coefficient of variation of daily discharge (Q)
Q	Daily discharge
ACF	Autocorrelation function
PACF	Partial autocorrelation function

1 Introduction

1.1 Background

Reliable and timely streamflow information plays a vital role in hydrologic studies, water resource planning, flood management, and environmental monitoring (Poursaeid 2025). River discharge is a key hydrologic variable that reflects the integrated response of a watershed to meteorological and catchment conditions (Duong et al. 2025). However, maintaining a network of in-situ gauging stations that continuously monitor streamflow is challenging due to financial, logistical, and technical constraints (Tran et al. 2025). These limitations often lead to missing or insufficient discharge data, especially in remote or underdeveloped regions. To overcome these gaps, the scientific community increasingly turns to data-driven methods and machine learning (ML) techniques that can model nonlinear and complex hydrological processes with relatively low computational cost. ML approaches provide a promising alternative to traditional conceptual and physical-based models by efficiently learning patterns from historical hydro-meteorological data (Singh et al. 2022). Their integration into hydrologic forecasting frameworks has shown great potential in improving the accuracy and timeliness of streamflow prediction.

1.2 Literature Review

The limitations of distributed conceptual hydrological models and the growing need for high-precision runoff prediction have increased interest in machine learning-based approaches (Bouaziz et al. 2021; Fenicia and McDonnell 2022). These data-driven models offer a more flexible framework for handling nonlinear, nonstationary hydrologic data and are particularly suited for runoff forecasting. Numerous studies have explored and benchmarked various ML algorithms and hybrid approaches (Khosravi et al. 2022; Singh et al. 2022; Ditthakit et al. 2023; Vatanchi et al. 2023). For instance, Yu et al. (2020) evaluated the Fourier transform (FT) hybridized with support vector regression (SVR-FT) and extreme gradient boosting (XGBoost-FT) at China's Three Gorges Dam. The SVR-FT demonstrated superior performance with an RMSE of 136.93 m^3/s and NSE of 0.999, outperforming the XGBoost-FT.

Tikhamarine et al. (2020) investigated the effect of integrating metaheuristic optimization algorithms like the Grey Wolf Optimizer (GWO) with ML models. Their study on Nile River discharge forecasting showed significant improvement in prediction metrics (e.g., NSE, RMSE) when applied optimization. Likewise, Zuo et al. (2020) demonstrated that signal decomposition techniques such as variational mode decomposition (VMD) and discrete wavelet transform (DWT) significantly enhance the predictive power of deep learning models like LSTM. Similarly, Saraiva et al. (2021) found that hybrid models such as bootstrap wavelet neural networks (BWNN) offered high accuracy across varying forecast horizons, with correlation coefficients above 0.98 even for 13-day-ahead predictions.

Further comparative studies, Zounemat-Kermani et al. (2021), revealed that hybrid ML models incorporating optimization and decomposition techniques yielded higher accuracy than standalone models. Wavelet ANFIS (WANFIS) performed better than other models for predicting daily discharge on the Koyna River in India (Bajirao et al. 2021). Adnan et al. (2021) advanced the field by combining ensemble learning with local learning techniques for the Jhelum River Basin, where LWL-AR exhibited RMSE and R-values of 223.9 m^3/s and 0.867, respectively.

Wavelet-based preprocessing, particularly the Maximal Overlap Discrete Wavelet Transform (MODWT), has shown consistent improvements across various hydrologic applications. Khan et al. (2023) combined gene expression programming (GEP) with MODWT, achieving up to 70% improvement in NSE and a 50% reduction in RMSE. Similarly, Quilty and Adamowski (2021) demonstrated the effectiveness of MODWT in enhancing model accuracy for monthly runoff and rainfall forecasting. In a related context, Barzegar et al. (2021) reported that CNN-LSTM-MODWT provided significant performance gains for lake level prediction, confirming MODWT's ability to extract meaningful features from nonstationary hydrologic time series.

Despite the widespread use of models like ANN, SVR, and LSTM, the Extremely Randomized Trees (ERT) algorithm remains underutilized in hydrological forecasting. The potential of ERT, a highly randomized ensemble tree-based method known for its computational efficiency and robustness against overfitting, warrants further exploration in runoff prediction tasks, especially when combined with decomposition techniques like MODWT.

1.3 Novelty and Purpose of the Current Research

In the present study, a new model, ERT, has been introduced, which is not usually used for runoff prediction. The ERT is combined with the discrete wavelet transform with maximum overlap. To address this research gap, several objectives are to be achieved in this study, including (i) demonstrating the use of highly randomized trees based on discrete wavelet transforms with maximum overlap for modeling streamflow in the two United States Geological Survey (USGS) stations, (ii) assessing predictive accuracy of the selected standalone machine learning approaches (ANN, ERT and GPR), (iii) coupling standalone ANN, ERT and GPR algorithm with MODWT to check the predictive accuracy and quantify the effect of pre-processing input combination on model results, and (iv) comparing ERT and ERT_MODWT with those of GPR, ANN, GPR_MODWT, and ANN _MODWT. The novelty of the proposed approach should be demonstrated by the improvement gained in the model's performance and the high forecasting accuracy obtained compared to the single models without decomposition.

2 Material and Methodology

2.1 Study Area and Data

This study focuses on two United States Geological Survey (USGS) streamflow gauging stations located in Oregon, USA (Fig. 1): (i) USGS 14211500 on Johnson Creek near Sycamore in Multnomah County (latitude 45°28′38.84″, longitude 122°30′28.87″, NAD83), and (ii) USGS 14211550 on Johnson Creek near Milwaukie in Clackamas County (latitude 45°27′11″, longitude 122°38′31″, NAD27). Daily discharge data (Q) for both stations were acquired from the USGS website (http://or.water.usgs.gov) for the period from 01 January 1998 to 10 November 2021, resulting in 8715 observations. The dataset was partitioned into training (70%, 6103 records; 01 January 1998 to 16 September 2014) and validation (30%,

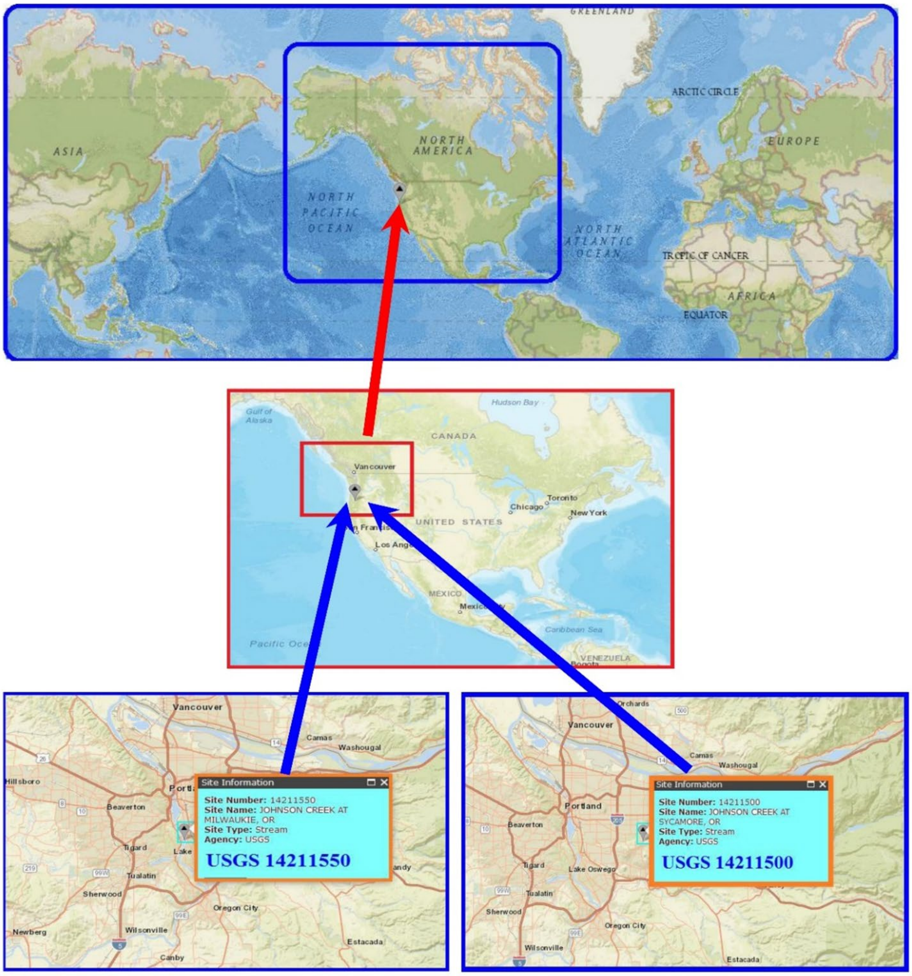


Fig. 1 A 2D map depicting the locations of the two USGS stations

2612 records; 17 September 2014 to 10 November 2021) subsets. Descriptive statistics (Table 1) reveal that USGS 14211550 generally exhibits higher mean flow and variability compared to USGS 14211500, as indicated by larger standard deviation and coefficient of variation (Cv).

Autocorrelation (ACF) and partial autocorrelation (PACF) analyses (Fig. 2) guided the selection of six time-lagged discharge inputs: Q(t–1) to Q(t–6). These lags were found to retain significant correlation with Q(t), making them suitable predictors for all machine learning models. Subsequently, each input series was decomposed using the maximal overlap discrete wavelet transform (MODWT) into ten subseries via multiresolution analysis (MRA) (Fig. 3). These decomposed signals served as enhanced input features for the hybrid ML models (ERT, GPR, and ANN), as illustrated in the modeling framework (Fig. 4).

Table 1 The streamflow statistical parameters for the two stations

Variables	Subset	Unit	X_{max}	X_{mean}	X_{min}	S_x	C_v
USGS 14211500							
Q	Training	Cu. ft/s	1580.000	50.849	0.270	90.164	1.773
	Validation	Cu. ft/s	1100.000	49.153	0.350	82.835	1.685
	Whole	Cu. ft/s	1580.000	50.340	0.270	88.033	1.749
USGS 14211550							
Q	Training	Cu. ft/s	1730.000	71.364	9.000	103.357	1.448
	Validation	Cu. ft/s	1400.000	73.480	9.140	109.397	1.489
	Whole	Cu. ft/s	1730.000	71.999	9.000	105.203	1.461

X_{mean}: mean; X_{max}: maximum; X_{min}: minimum; S_x: standard deviation; C_v: coefficient of variation

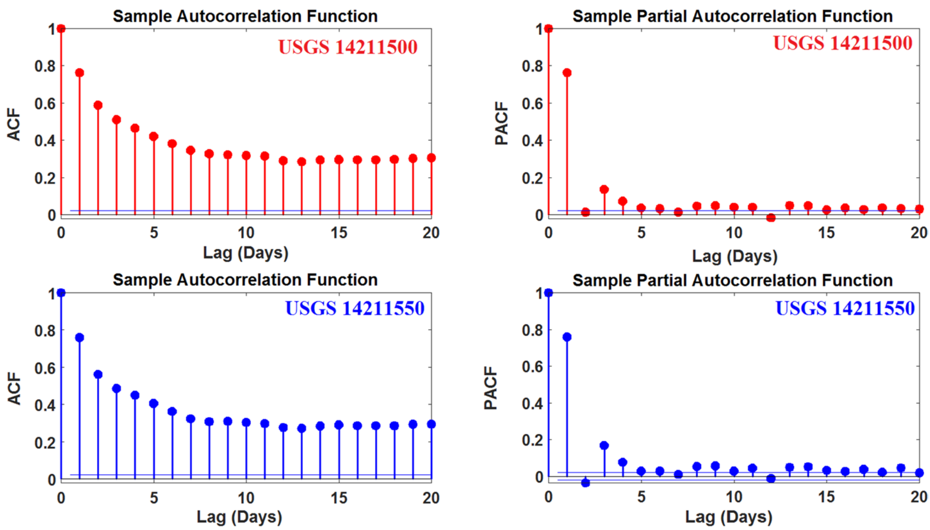


Fig. 2 Autocorrelation (ACF) and partial autocorrelation function (PACF) of the sample for daily streamflow (*Q*)

2.2 Machine Learning Models

This section provides brief descriptions of the approaches used by ML, such as Extremely Randomized Trees (ERT), Gaussian Process Regression (GPR), Artificial Neural Network (ANN), and Maximal Overlap Discrete Wavelet Transform (MODWT), as well as a description of the steps taken to develop the model.

2.2.1 Extremely Randomized Trees (ERT)

Extremely randomized trees (ERT) consist of a set of decision trees (DT) similar to the Random Forest Algorithm (RFR) (Geurts et al. 2006). However, unlike the RFR algorithm based on bootstrap sampling, ERT uses all the data to create each tree and train it (Fig. 5); the cut point is entirely random.

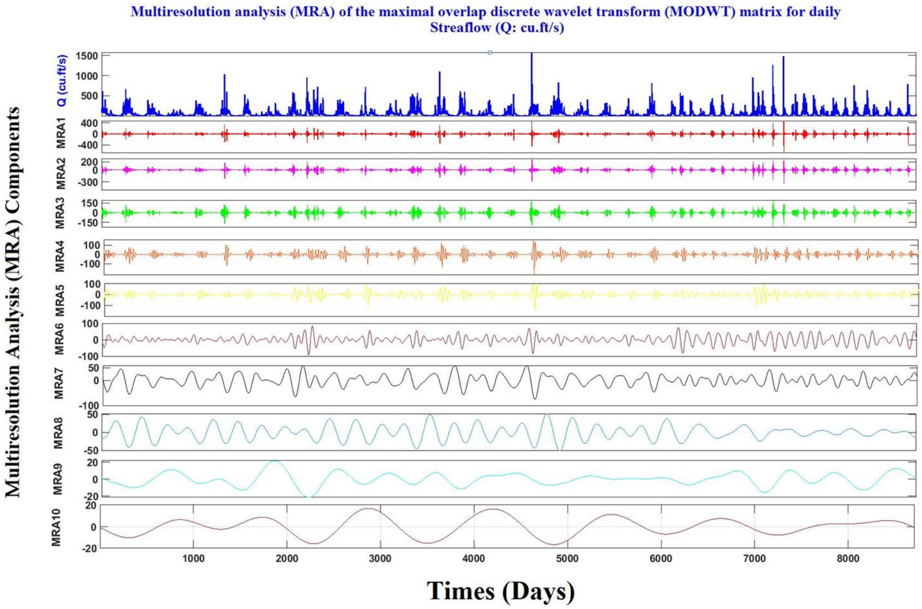


Fig. 3 The daily streamflow dataset (Q) has been decomposed into Multi-Resolution Analysis (MRA) components using the Maximal Overlap Discrete Wavelet Transform (MODWT) algorithm

Randomly selecting the optimal bifurcation attribute of the DT results in a great advantage as the calculated variance becomes more stable (Cheng et al. 2021b). Moreover, ERT selects the attributes and the intersection of the DT without prior assumption and not uniformly, but the selection is highly randomized (Yang et al. 2021). To develop an ERT model, a suite of DT, many randomly selected attributes (K) are considered for splitting a node (n min), and the ERT creates an M-independent DT (Roy et al. 2021). Numerous examples of ERT applications can be found in the literature, such as Internet big data recovery (Cheng et al. 2021a), coal and biomass co-pyrolysis prediction (Wei et al. 2022), and nitrogen dioxide estimation (Huang et al. 2022).

2.2.2 Gaussian Process Regression (GPR)

Gaussian process regression was introduced by (Rasmussen and Williams 2006). The GPR can be classified into non-parametric algorithms to construct robust regression models based on the Bayesian learning paradigm (Li et al. 2022). The GPR is mainly used to build robust Bayesian regression models that can handle nonstationary data sets accurately. Mathematically, the GPR can be formulated as follows (Fig. 6):

$$Y_n = f\left(x_n\right) + \mu_{\ n} \tag{1}$$

In the above equation, Y_n is the dependent variable to be modeled, x_n is the independent variable, and $\mu_{\ n}$ corresponds to the Gaussian distributed measurement noise. The noise is calculated as the difference between the measured and target output (the actual output). The

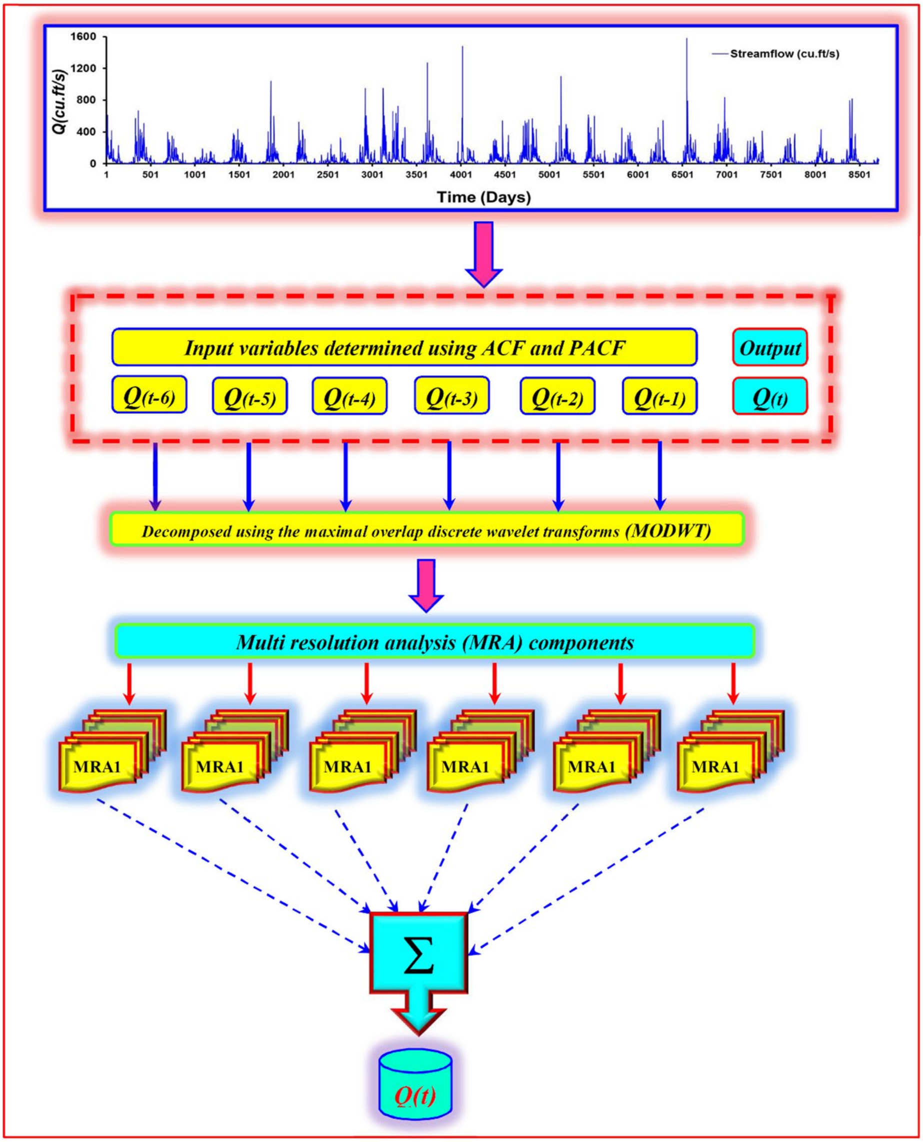


Fig. 4 The flowchart of the MODWT, ERT, ANN, and GPR models

noise results from various errors, especially the presence of outliers. The noise variance is than calculated using the kernel function (i.e., Gaussian function), which assumes that the noise is Independent and identically distributed. The GPR can use the vector (x) to output (Y) using a training dataset as follows (Li et al. 2022):

$$training\ dataet = T = (x_i, y_i)_{i=1}^{N} \tag{2}$$

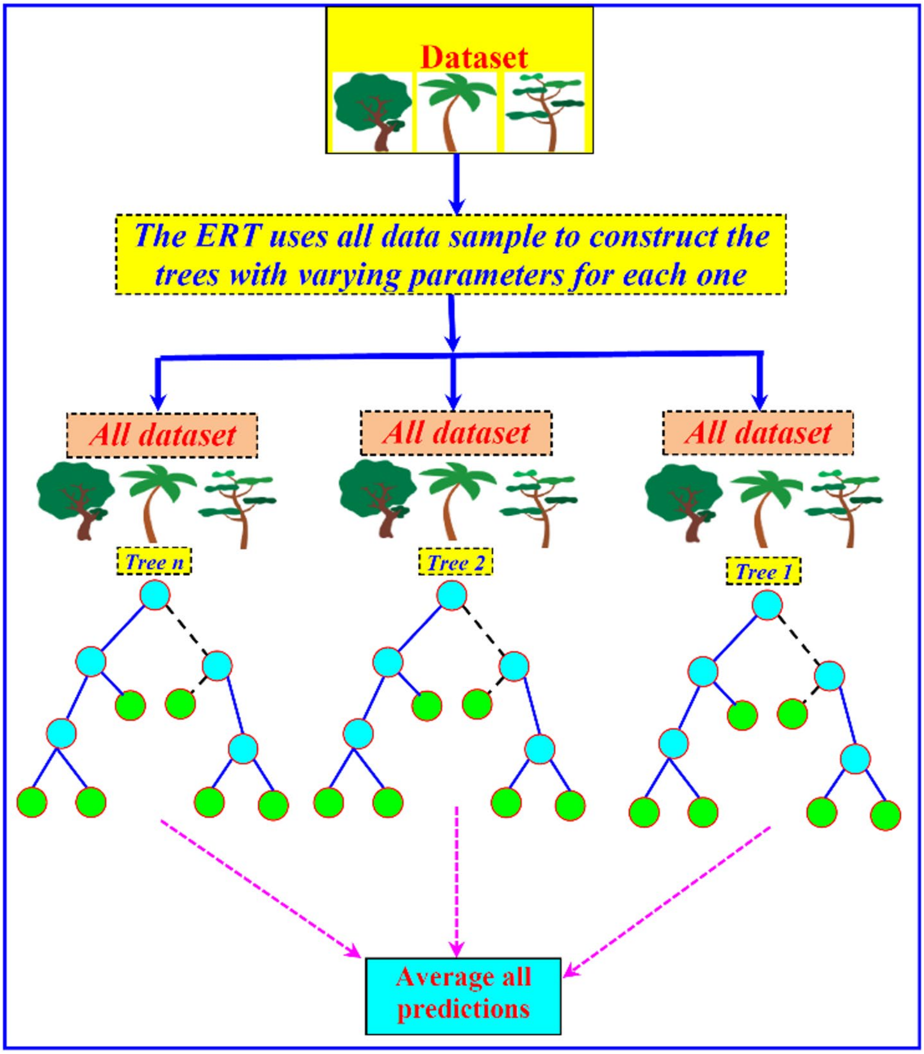


Fig. 5 Extremely randomized trees (ERT) architecture

x_i is the input vector with N dimension and y_i is the target vector. The GPR defines a probability distribution over functions as follows (Yang et al. 2018; Liu et al. 2018):

$$f\left(x\right)\sim GP\left(m\left(x\right),k\left(x,x'\right)\right) \tag{3}$$

Where $m(x)$ and $\kappa(x, x')$ are the mean and covariance functions calculated as follows (Yang et al. 2018):

$$m\left(x\right)=E\left[f\left(x\right)\right] \tag{4}$$

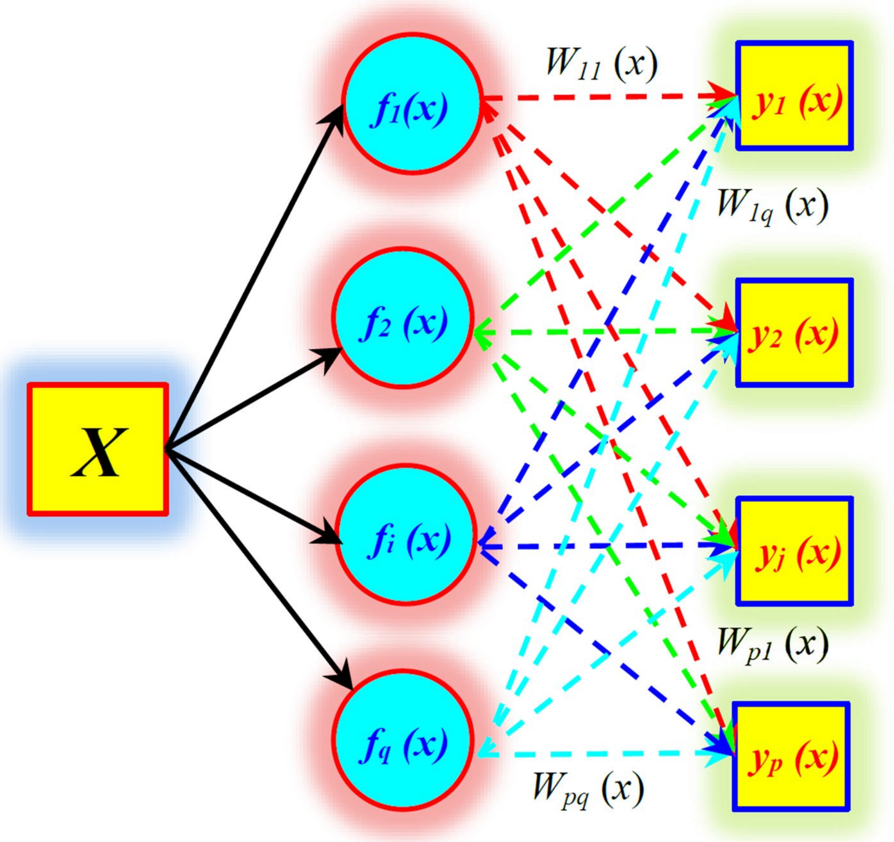


Fig. 6 Structure of the Gaussian process regression (GPR) network

$$k(x, x') = E\left[(f(x) - m(x))(f(x') - m(x'))^T\right] \tag{5}$$

where x' and x are random variables that denote the labeling of dimensions, x denotes a dimensional feature in the set, and x' denotes another dimensional feature that is not identical (Cai et al. 2023).

The task of regression is to create a new input x' to achieve the predicted distribution for the corresponding values of the observational data y based on the dataset. The GPR's most critical aspect is the covariance function, and generally speaking, the squared exponential (SE) is used and expressed as follows (Zhou et al. 2022):

$$k_{SE}(x, x') = \sigma_f^2 exp\left(-\frac{1}{2}\frac{(x - x')^T(x - x')}{\sigma_l^2}\right) \tag{6}$$

Where x and x' are selected randomly from the dataset, σ_l^2 is the characteristic length scale; and σ_f^2 is the signal standard deviation (Zhou et al. 2022). Using gradient-based algorithms, the hyperparameters of the covariance function σ_l, σ_f can be estimated. More detail about the GPR approach can be found in (Zhou et al. 2022).

2.2.3 Artificial Neural Network (ANN)

An artificial neural network (ANN) is a machine learning type modeled on how the human brain works (Shukla et al. 2021; Elbeltagi et al. 2022). The ANN uses a sigmoid activation function for the hidden layer nodes and has been used to solve various engineering problems, especially in water management (Shukla et al. 2021; Elbeltagi et al. 2022). By combining the brain's learning capabilities and a network structure with multiple layers, ANN can significantly improve the ability to map complex and difficult nonlinear tasks (Talei 2022). Figure 7 shows the general structure of an ANN model with a single output and n inputs.

According to Fig. 7, an unknown nonlinear system can be expressed as follows:

$$Y = f(x_i) \tag{7}$$

Y is the model output (i.e., the outflow at time t), x_i is the input variable (i.e., the outflow at time t-n), and f corresponds to the unknown function the model ANN determines. In the

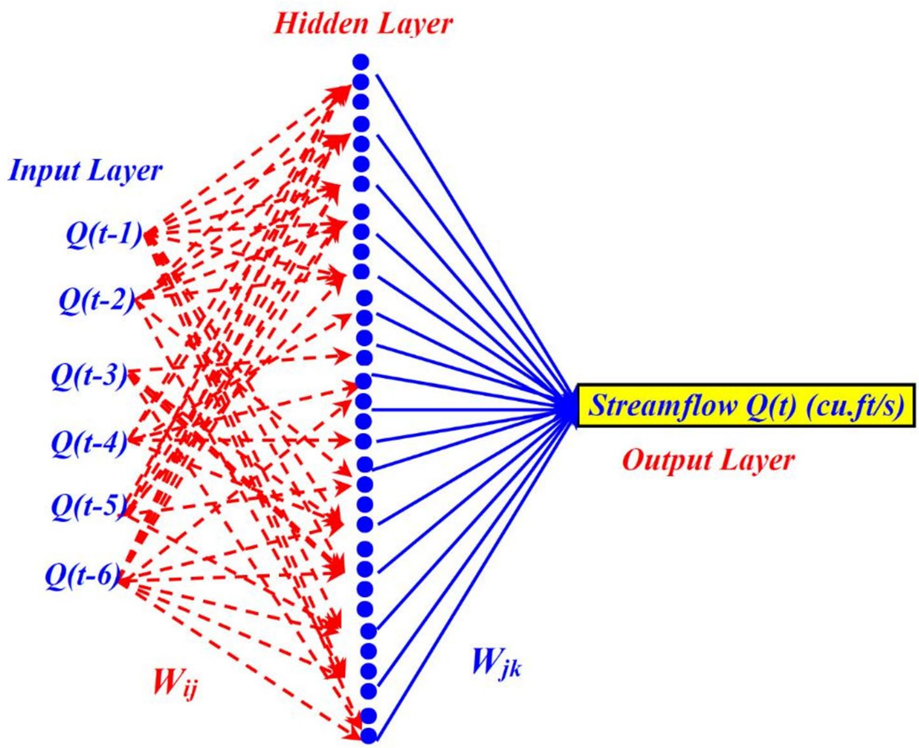


Fig. 7 Flowchart of the ANN model

ANN paradigm, the main objective is to find the appropriate parameter sets (i.e., the W_{ij} and W_{jk}) using a learning algorithm.

The backpropagation algorithm is used iteratively for training purposes to minimize the mean square error (MSE) between the measured and predicted output. Each layer is connected to the previous layer, and all relevant information must be transferred from the input to the output layer. Overall, the main operations of the ANN can be summarized as (i) the initialization step, where the weights and biases were initialized randomly, (ii) the pairs of input and output variables were presented to the model using the training subset and served as a database, (iii) the actual output is then calculated according to Eqs. 12 and 13, (iv) the adjustment of the weights and biases, (v) the learning procedure is repeated several times until the desired output was obtained. The best-obtained model parameters were then frozen and used to test the generalization capabilities of the model.

To determine the net input and output for each neuron in the hidden layer, the following calculation is used:

$$A_j = \sum {}_{j=1}^{m}(W_{ij} \times x_i) + \beta_j \tag{8}$$

$$Y_k = \sum {}_{j=1}^{m}(W_{jk} \times f(A_j)) + \alpha_1 \tag{9}$$

A_j is the net input of the hidden neuron j, β_j is the bias of the hidden neuron j, W_{ij} is the weight linking the input neuron i to the hidden neuron j, x_i corresponds to one of the inputs variables, $f(A_j)$ corresponds to the net output of the hidden neuron j, W_{jk} is the weight linking the hidden neuron j to the output neuron k ($k = 1$), and α_1 is the bias of the output layer. Further information regarding the ANN paradigm can be found elsewhere (Hornik et al. 1989; Hornik 1991; Simon 1999). The hyper-parameters of the ANN model were selected as follows: only one hidden layer with thirteen neurons with a sigmoidal activation function, and only one output neuron with linear (identity function).

2.2.4 Maximal Overlap Discrete Wavelet Transforms

The discrete wavelet transforms with maximum overlap (MODWT) is a generalization of the standalone discrete wavelet transform (DWT) and is a nonorthogonal transform algorithm that decomposes a nonlinear signal into multiple sub-signals or details and a single residual (Zhang et al. 2019; Sadek and Abdulrazak 2021). For any nonlinear time series X, by choosing the degree of decomposition equal to (0, there are two principal components of MODWT, namely wavelet coefficients, $\widetilde{W}_{j,\,l}$ and scaling coefficients, $\widetilde{Z}_{j,\,l}$ which can be defined as follows (Mouatadid et al. 2019; Ghimire et al. 2019; Ashok and Yadav 2020; Patnaik et al. 2021; Dubey et al. 2022):

$$\widetilde{W}_{j,\,l} = \sum {}_{l=0}^{L_{j-1}} \widetilde{\phi}_{j,l} X_{t-l mod N} \tag{10}$$

$$\widetilde{Z}_{j,\,l} = \sum {}_{l=0}^{L_{j-1}} \widetilde{\delta}_{j,l} X_{t-l mod N} \tag{11}$$

$$\tilde{\phi}_{j,l} = \frac{\phi_{j,l}}{2^{\frac{j}{2}}} \tag{12}$$

$$\tilde{\delta}_{j,l} = \frac{\delta_{j,l}}{2^{\frac{j}{2}}} \tag{13}$$

Where j and l are the level of decomposition and the length of the filter ($l = 1, 2, \cdots$, L) with $L_j = (2^j-1)(L-1)+1$, the $\tilde{W}_{j,\ l}$ and $\tilde{Z}_{j,\ l}$ are the wavelet coefficients and the scaling for the corresponding time series X. $\phi_{j,l}$ and $\delta_{j,l}$ correspond to the wavelet and scaling filter, respectively, of DWT, and, $\tilde{\phi}_{j,l}$ and $\tilde{\delta}_{j,l}$ are the wavelet and scaling filters of MODWT (Mouatadid et al. 2019; Ghimire et al. 2019; Patnaik et al. 2021). The wavelet coefficient should be accompanied by a suitable mother wavelet function and the decomposition ratio (Kulkarni and Sahasrabudhe 2017). There are several mother wavelet functions, including Haar wavelet ''haar'', Daubechies wavelet with extreme phase and N vanishing moments ''dbN'', where N is a positive integer from 1 to 45, and Symlets wavelet with N vanishing moments ''symN'', where N is a positive integer from 1 to 45 (Shukla et al. 2021). In the present study, we chose ''db4'' as the mother wavelet function and ten as the decomposition degree by trial and error.

2.3 Evaluating the Performance of the Models

Several measures were used to evaluate the accuracy of the daily runoff forecast, including root mean square error (RMSE), mean absolute error (MAE), correlation coefficient (R), and Nash-Sutcliffe efficiency (NSE). The expressions are given as:

$$RMSE = \sqrt{\frac{1}{N}\sum{}_{i=1}^{N}\left[(Q_{obs,i}) - (Q_{est,i})_i\right]^2}, (0 \leq\ RMSE < +\infty\ ) \tag{14}$$

$$MAE = \frac{1}{N}\sum{}_{i=1}^{N}\left|Q_{obs,i} - Q_{est,i}\right|, (0 \leq\ MAE < +\infty\ ) \tag{15}$$

$$R = \left[\frac{\frac{1}{N}\sum{}_{i=1}^{N}\left(Q_{obs,i} - \bar{Q}_{obs}\right)\left(Q_{est,i} - \bar{Q}_{est}\right)}{\sqrt{\frac{1}{N}\sum{}_{i=1}^{N}\left[(Q_{obs,i}) - \bar{Q}_{obs}\right]^2}\sqrt{\frac{1}{N}\sum{}_{i=1}^{N}\left[Q_{est,i} - \bar{Q}_{est}\right]^2}}\right], (-1 < R \leq\ +1) \tag{16}$$

$$NSE = 1 - \left[\frac{\sum{}_{i=1}^{N}\left[Q_{obs} - Q_{est}\right]^2}{\sum{}_{i=1}^{N}\left[Q_{obs,i} - \bar{Q}_{obs}\right]^2}\right], (-\infty\ < NSE \leq\ 1) \tag{17}$$

$\bar{Q}_{obs}$ and $\bar{Q}_{est}$ are the average measured and forecasted daily streamflow (*Q*), respectively, Q_{obs}, and Q_{est} specify the observed and forecasted daily streamflow (*Q*: *cu. ft/s*) for i^{th} observations. N shows the number of data points.

3 Results and Discussion

This study applies three machine learning techniques: ERT, GPR, and standalone ANN. The models are developed without signal preprocessing, but at a later stage, the MODWT algorithm is used for signal decomposition. Six input combinations were selected based on ACF and PACF, using the flux (Q) measured at earlier time points as the input variable (Table 2). The results from the two stations are presented and discussed below.

3.1 Results at the USGS 14211500

This section presents the comparative performance analysis of three machine learning models, i.e., Extremely Randomized Trees (ERT), Gaussian Process Regression (GPR), and Artificial Neural Network (ANN), applied to daily streamflow prediction at the USGS 14211500 station, with and without Maximal Overlap Discrete Wavelet Transform (MODWT) preprocessing. Six input combinations, representing increasing lag times from Q(t − 1) to Q(t − 6), were evaluated. The evaluation metrics/and performance metrics for both the training and validation periods of all developed modes are highlighted in Fig. 8, while scatter plots, discharge comparisons (time series) only for ERT, ANN, and GPR integrated with MODWT, are illustrated in Figs. 9 and 10.

Overall, the application of MODWT significantly enhanced the prediction accuracy across all models, supporting the hypothesis that decomposition improves model sensitivity to temporal patterns in hydrologic data (Fig. 8). Without MODWT preprocessing, ANN and ERT models demonstrated comparable predictive capabilities. For the ANN models, the validation phase yielded R values ranging from 0.809 to 0.822 (mean = 0.817), NSE from 0.655 to 0.675 (mean = 0.668), RMSE between 47.214 and 48.677 ft^3/s (mean = 47.945 ft^3/s), and MAE from 19.222 to 19.977 ft^3/s (mean = 19.547 ft^3/s). ERT models showed similar performance, with R values between 0.804 and 0.822 (mean = 0.817), NSE ranging from 0.646 to 0.673 (mean = 0.665), RMSE from 47.362 to 49.293 ft^3/s (mean = 47.856 ft^3/s), and MAE between 19.727 and 20.539 ft^3/s (mean = 20.104 ft^3/s). In contrast, the GPR models exhibited substantially lower accuracy, with R values from 0.612 to 0.817 (mean = 0.657), NSE ranging from 0.372 to 0.666 (mean = 0.434), RMSE between 47.900 and 65.675 ft^3/s (mean = 62.049 ft^3/s), and MAE from 19.717 to 28.254 ft^3/s (mean = 26.168 ft^3/s) (Fig. 8). These results confirm that ANN and ERT models were more robust and stable, In contrast GPR was more sensitive to increasing input dimensionality, leading to performance degradation beyond the first input combination.

ANN and ERT both benefited from shorter input sequences, particularly Q(t − 1) and Q(t − 2), However, increasing the number of input lags did not yield further improvements and, in some cases, reduced accuracy. For example, ANN1 and ERT2 delivered the best

Table 2 The input combinations of different models

ERT	GPR	MLPNN	Input combination	Output
ERT1	GPR1	ANN1	*Q(t-1)*	*Q(t)*
ERT2	GPR2	ANN2	*Q(t-2)*, *Q(t-1)*	*Q(t)*
ERT3	GPR3	ANN3	*Q(t-3)*, *Q(t-2)*, *Q(t-1)*	*Q(t)*
ERT4	GPR4	ANN4	*Q(t-4)*, *Q(t-3)*, *Q(t-2)*, *Q(t-1)*	*Q(t)*
ERT5	GPR5	ANN5	*Q(t-5)*, *Q(t-4)*, *Q(t-3)*, *Q(t-2)*, *Q(t-1)*	*Q(t)*
ERT6	GPR6	ANN6	*Q(t-6)*, *Q(t-5)*, *Q(t-4)*, *Q(t-3)*, *Q(t-2)*, *Q(t-1)*	*Q(t)*

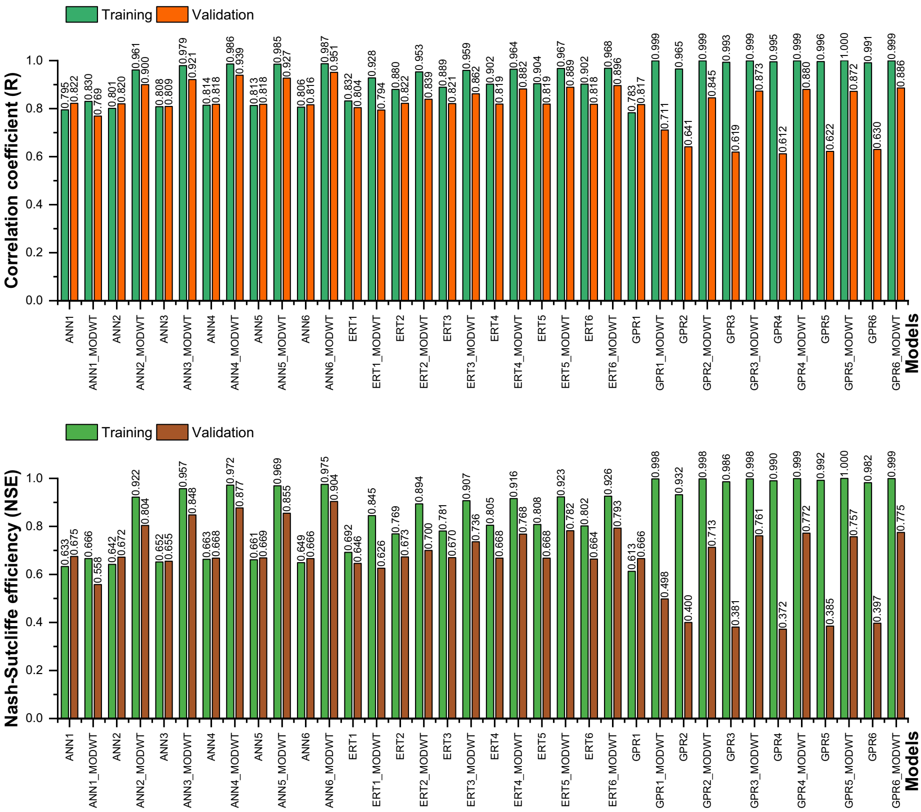


Fig. 8 Prediction accuracy of various forecasting models at the USGS 14211500 station

standalone results, while ANN3 and ERT3 showed marginal declines in NSE and increases in RMSE. The Gaussian Process Regression (GPR) models exhibited a marked decline in predictive performance with the inclusion of additional input lags. Specifically, the GPR1 model achieved a correlation coefficient (R) of 0.817 and a Nash–Sutcliffe Efficiency (NSE) of 0.666, whereas the performance of GPR6 declined significantly, with R decreasing to 0.612 and NSE to 0.372. This suggests that the GPR model is more sensitive to input dimensionality, likely due to its kernel-based architecture.

The application of MODWT in the second analysis phase improved model performance, particularly for input combinations beyond Q(t −1). When using only Q(t −1), MODWT marginally decreased performance in ANN1, ERT1, and GPR1. Specifically, ANN1_MODWT showed reductions of 5.30% in R, 11.70% in NSE, 14.31% in RMSE, and 29.15% in MAE compared to ANN1. ERT1_MODWT exhibited respective declines of 1.00%, 2.00%, 2.66%, and 10.63%, and GPR1_MODWT similarly saw reductions across all four metrics (Fig. 8).

However, starting from Q(t −2), MODWT had a consistently positive effect. ANN_MODWT models demonstrated continuous improvement from ANN2_MODWT to ANN6_MODWT. For ANN6_MODWT, validation performance reached R= 0.951, NSE =0.904, RMSE =25.617 ft³/s, and MAE =12.978 ft³/s—representing the best overall performance

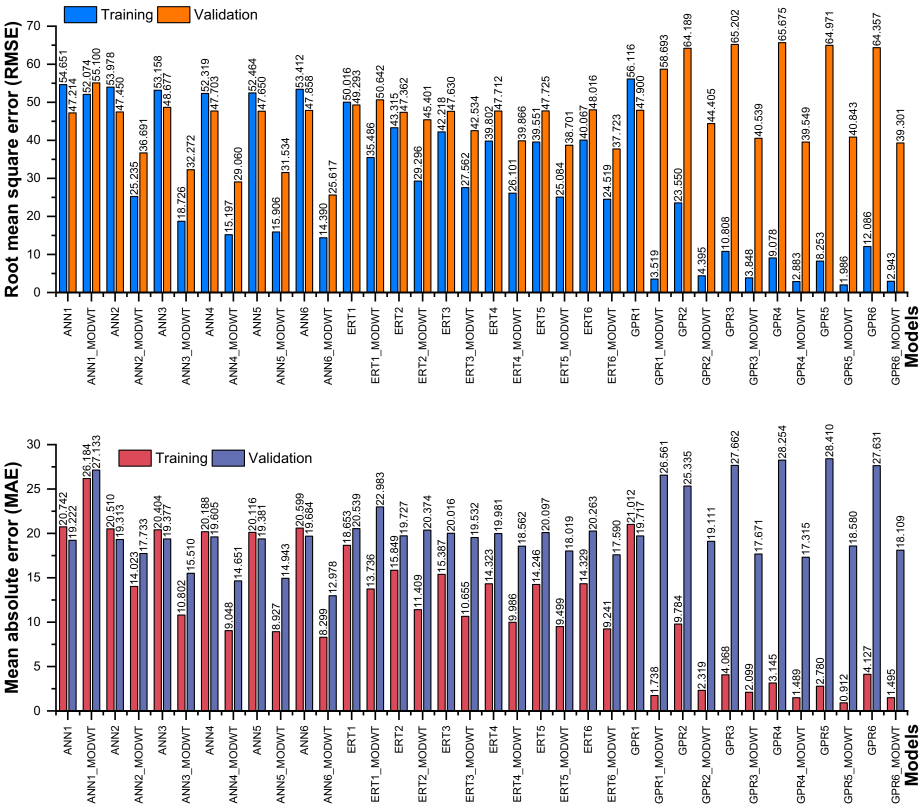


Figure 8 (continued)

(Figs. 8, 9 and 10). Compared to ERT6_MODWT, ANN6_MODWT improved by 5.50% in R, 11.10% in NSE, 32.09% in RMSE, and 26.22% in MAE. Against GPR6_MODWT, improvements were even more pronounced at 6.50%, 12.90%, 34.81%, and 28.33%, respectively (Figs. 8, 9 and 10).

On average, the MODWT enhanced ANN model performance across the validation phase was 8.4% (R), 14.02% (NSE), 26.62% (RMSE), and 12.15% (MAE). For ERT models, mean improvements were 4.14% (R), 6.40% (NSE), 11.42% (RMSE), and 2.95% (MAE). GPR models showed the most significant relative gains: 18.76% in R, 27.91% in NSE, 29.26% in RMSE, and 25.26% in MAE, indicating that GPR models benefit disproportionately from MODWT preprocessing.

Comprehensive visual comparisons using boxplots, violin plots, radar charts, and Taylor diagrams (Fig. 11a–d) further corroborate these findings. Representative plots illustrate the best-performing models' performance during the validation phase at USGS station 14211500. The boxplots and violin plot in Fig. 11a-b highlight the density and spread of predicted versus observed discharge. ANN6_MODWT closely replicates the distribution of the observed data, followed by ERT6_MODWT and GPR6_MODWT. Standalone models (ANN6, ERT6, and GPR6) depict wider variance and lower central tendency alignment.

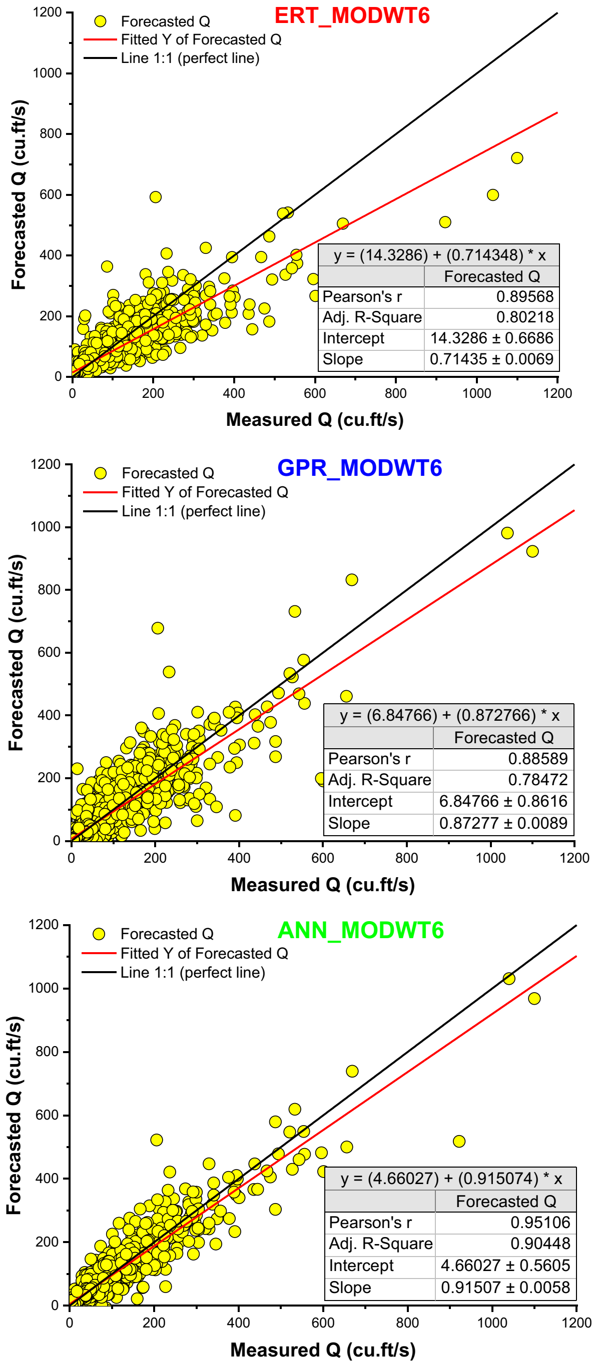


Fig. 9 Scatterplots measured against forecasted daily streamflow (Q) at the USGS 14211500 for the validation stage: ERT_MODWT6 (left), GRP_MODWT6 (middle), and ANN_MODWT6 (right)

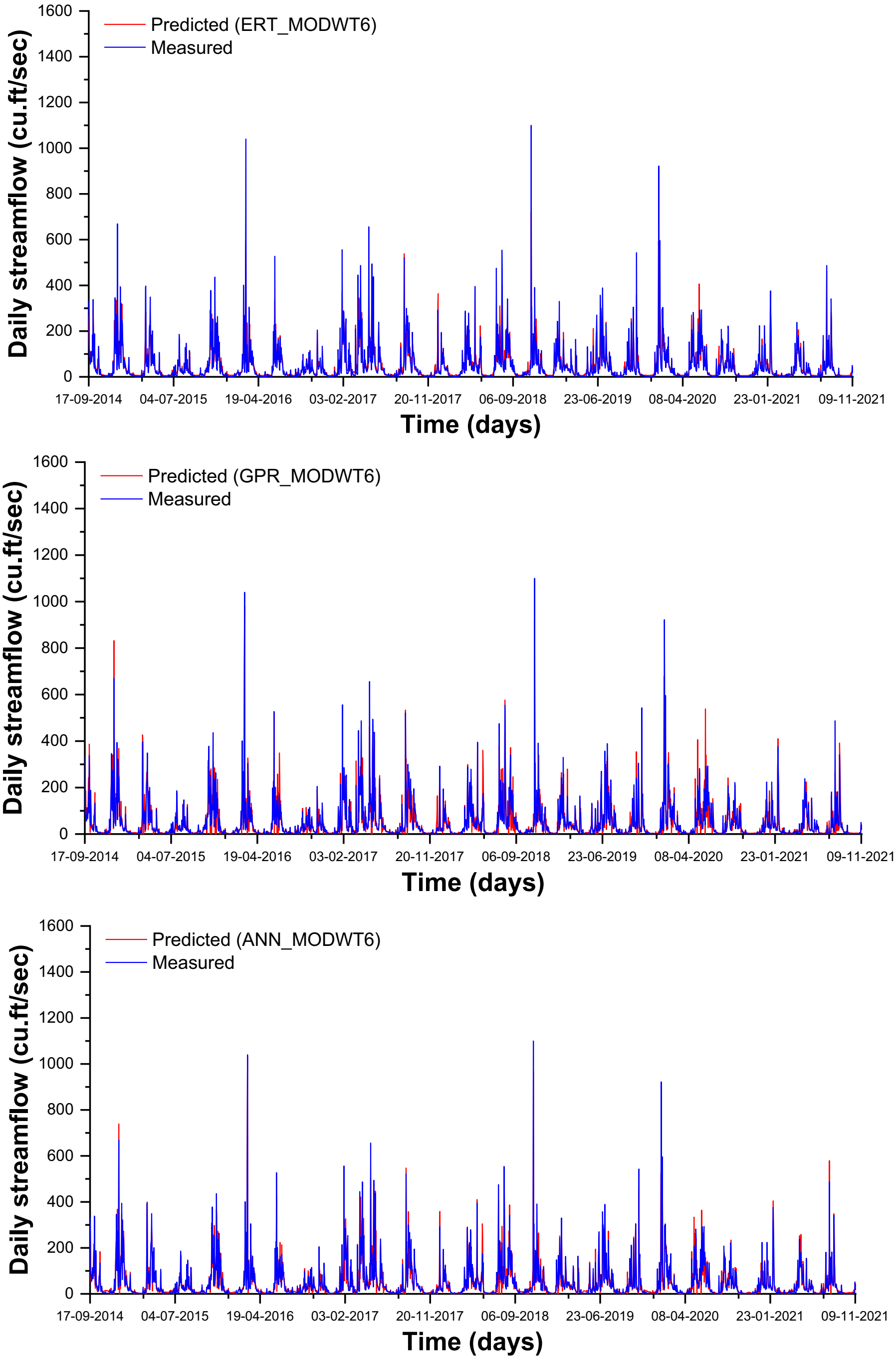


Fig. 10 Comparison between observed and predicted daily streamflow (Q) at the USGS 14211500 during the validation. In the figure, ERT_MODWT6 (upper panel), GRP_MODWT6 (middle panel), and ANN_MODWT6 (lower panel)

The radar chart (Fig. 11c) consolidates performance indicators for each model. ANN6_MODWT occupies the largest polygonal area, indicating consistent superiority across R, NSE, RMSE, and MAE metrics. ERT6_MODWT follows closely, while GPR6_MODWT, although improved, trails both.

Taylor diagrams (Fig. 11d) provide an integrated visualization of RMSE, correlation (R), and standard deviation. The proximity of ANN6_MODWT to the reference point representing observed data reflects its optimal balance of low error and high correlation. ERT6_MODWT is positioned nearby, while GPR6_MODWT, although improved over its standalone counterpart, is more distant.

These findings demonstrate that the hybrid ANN6_MODWT model is the most effective in predicting daily streamflow at the USGS 14211500 station. Its superior performance underscores the value of MODWT as a signal decomposition technique, especially when paired with deep learning architectures. Although ERT shows competitive accuracy and efficiency, and GPR benefits substantially from MODWT, ANN combined with MODWT consistently outperforms the alternatives across all statistical indicators and visual validation tools.

3.2 Results at the USGS 14211550

The performance of the ANN, ERT, and GPR models at USGS station 14211550 is summarized in Fig. 12. The results indicate that incorporating MODWT substantially enhanced model performance. Across all models, the R values ranged from 0.577 to 0.775 (mean = 0.709), NSE from 0.336 to 0.600 (mean = 0.511), RMSE from 69.170 to 88.464 ft^3/s (mean = 76.841 ft^3/s), and MAE from 26.907 to 41.232 ft^3/s (mean = 33.287 ft^3/s). Both ANN and ERT models significantly outperformed GPR models in predictive accuracy.

Increasing the number of input variables from Q(t − 1) to Q(t − 6) did not consistently improve model performance. In some cases, particularly with GPR, performance declined as more input lags were added. For instance, the best GPR result (GPR6) had lower accuracy than ANN2, which achieved R = 0.783, NSE = 0.612, RMSE = 68.12 ft^3/s, and MAE = 26.35 ft^3/s.

The scatter plots illustrating the performance of the three best MODWT-enhanced models are presented in Fig. 13, In contrast, Fig. 14 time-series plots depict the comparison between observed and forecasted daily discharge (Q) at USGS 14211550 for the validation phase. The three best MODWT-enhanced models—ANN6_MODWT, ERT6_MODWT, and GPR6_MODWT—are featured. MODWT preprocessing produced clear improvements in all models. For ANN, MODWT raised mean R by 13.67%, NSE by 20.68%, and reduced RMSE and MAE by 31.77% and 17.75%, respectively, when comparing ANN models with and without MODWT. Notably, ANN6_MODWT outperformed ANN1_MODWT by 14.70% (R), 23.10% (NSE), 38.51% (RMSE), and 39.77% (MAE), underscoring the benefit of using multiple lags in combination with wavelet decomposition (Fig. 12). For ERT, the best-performing configuration was ERT6_MODWT. While it delivered strong results, it remained less accurate than ANN6_MODWT, with relative differences of −3.5% (R), −3.5% (NSE), + 13.10% (RMSE), and + 12.64% (MAE). Interestingly, GPR6_MODWT slightly outperformed ERT6_MODWT with minor gains of approximately 0.60% (R), 0.30% (NSE), 0.47% (RMSE), and 1.27% (MAE), yet still fell short of ANN6_MODWT's performance (Fig. 12).

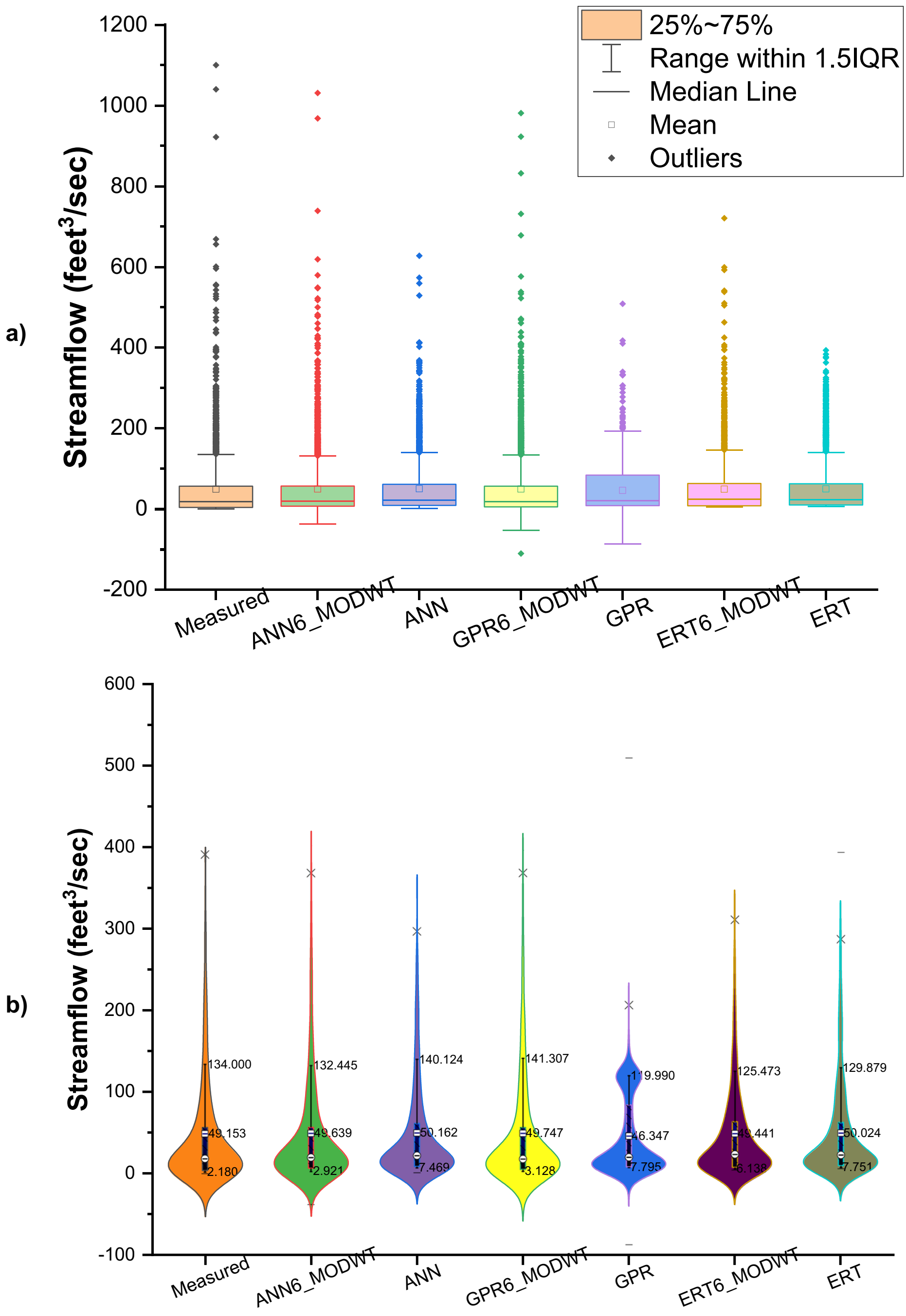


Fig. 11 Performance plots of best models during validation at USGS 14211500: **a**) boxplot, **b**) violin plot, **c**) radar chart, **d**) Taylor diagram

c)

d)

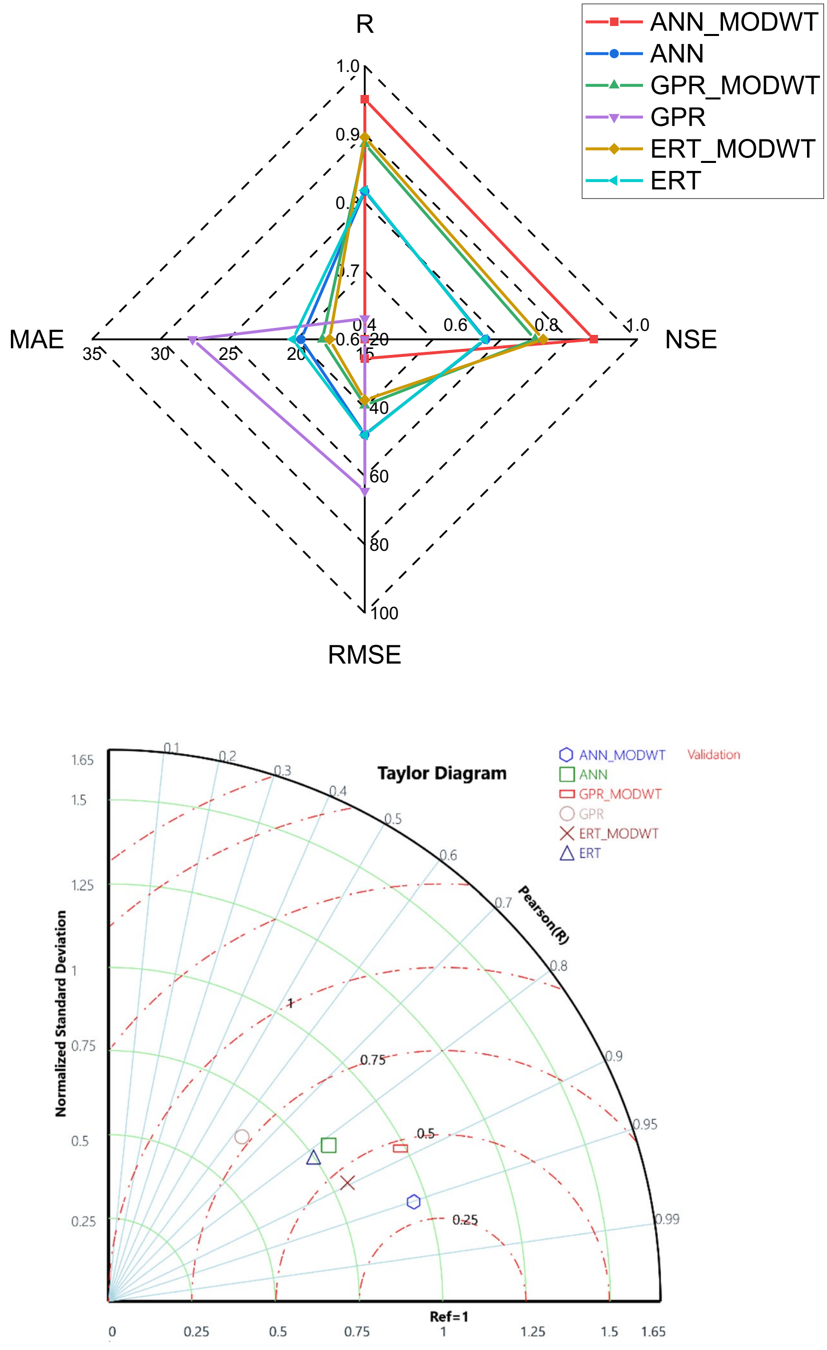


Figure 11 (continued)

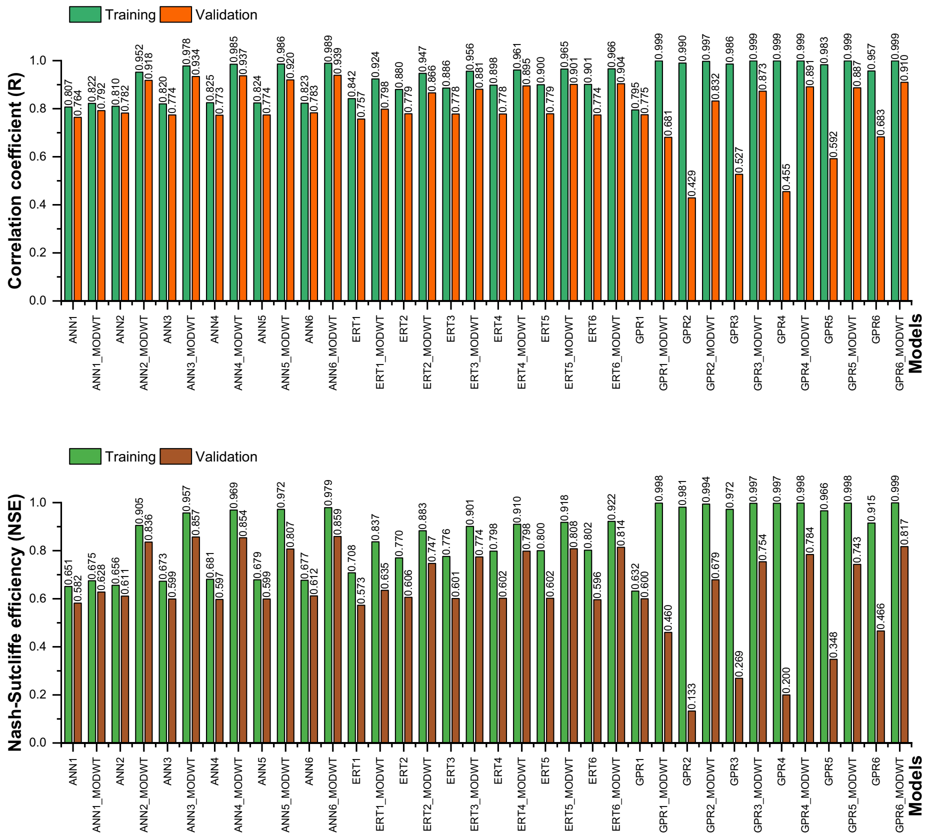


Fig. 12 Performances of different forecasting models at the USGS 14211550 station

Boxplots and violin plots in Fig. 15a and b further illustrate the agreement between predicted and observed values. ANN6_MODWT demonstrated the closest match to observed runoff distributions, followed by ERT6_MODWT and GPR6_MODWT. Standalone models (ANN6, ERT6, and GPR6) exhibited wider variance and lower central tendency alignment. Radar plots (Fig. 15c) provide a comparative visual of the performance metrics across the best models. ANN6_MODWT covered the largest polygonal area, reflecting superior R, NSE, RMSE, and MAE values. Both ERT6_MODWT and GPR6_MODWT also showed improved metrics relative to their standalone versions, but ANN6_MODWT remained dominant. Taylor diagrams (Fig. 15d) further demonstrate the ANN6_MODWT model's superiority. This plot simultaneously represents correlation, standard deviation, and RMSE in a single frame. ANN6_MODWT lies closest to the reference point (Ref = 1), indicating strong correlation and minimal error compared to ERT6_MODWT and GPR6_MODWT.

From a broader perspective, several conclusions can be drawn. First, ANN and ERT consistently outperformed GPR across all configurations, regardless of the number of input lags. Second, increasing the number of input variables beyond two lags ($Q(t-1)$ and $Q(t-2)$) often yielded diminishing or even negative returns, particularly in GPR. Only the ERT model showed performance gains from the first to the second lag, whereas

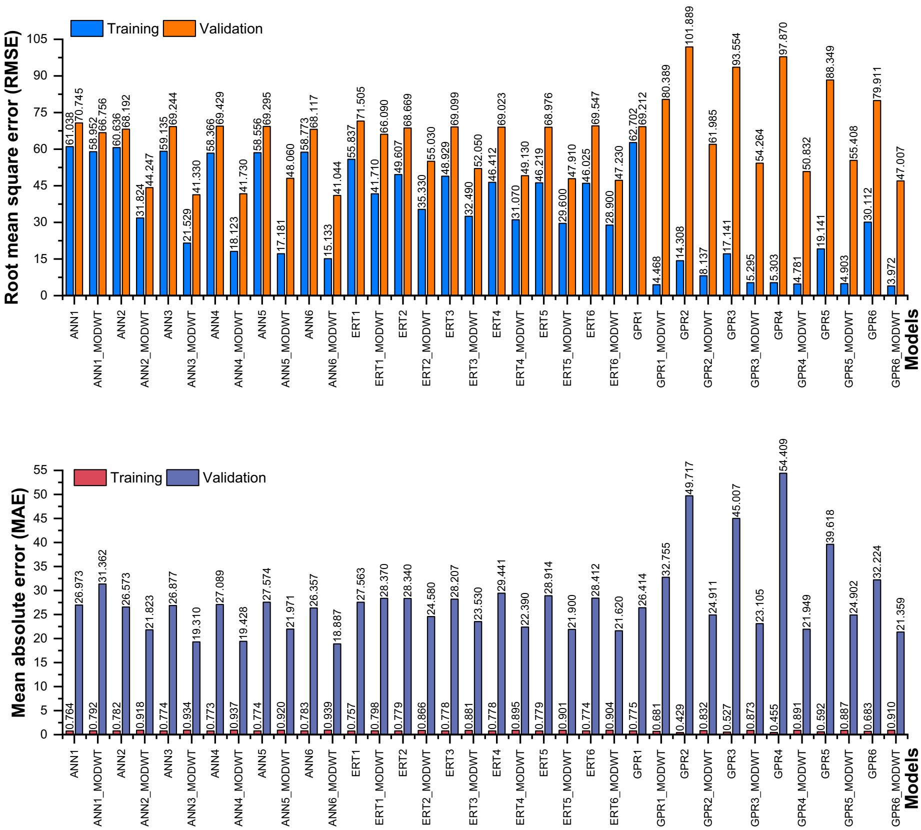


Figure 12 (continued)

ANN exhibited a slight decline, and GPR showed a more pronounced deterioration. These observations suggest that, in the absence of preprocessing, using only one or two previous discharge values is sufficient for acceptable accuracy. However, when MODWT is employed, adding additional lags becomes beneficial. The MODWT decomposed each lag into multiple resolution subcomponents, capturing both high- and low-frequency features of the streamflow data. This enhanced the model's capacity to extract meaningful patterns and temporal dependencies not captured by raw input alone.

In the final phase of the study, MODWT was applied to all six lagged discharge inputs, resulting in multi-resolution analysis (MRA) subseries for model training. As the number of decomposed inputs increased, model coherence between observed and predicted discharge improved across training and validation periods. By the sixth input combination, the ANN6_MODWT model produced the most accurate results in terms of R, NSE, RMSE, and MAE, demonstrating that MODWT effectively captures the available signal content in the streamflow series. In summary, the findings demonstrate the superiority of MODWT-enhanced models, particularly ANN6_MODWT, in streamflow forecasting at USGS 14211550. While all models benefited from wavelet decomposition, ANN showed the highest sensitivity to enhanced input features and consistently achieved the best overall

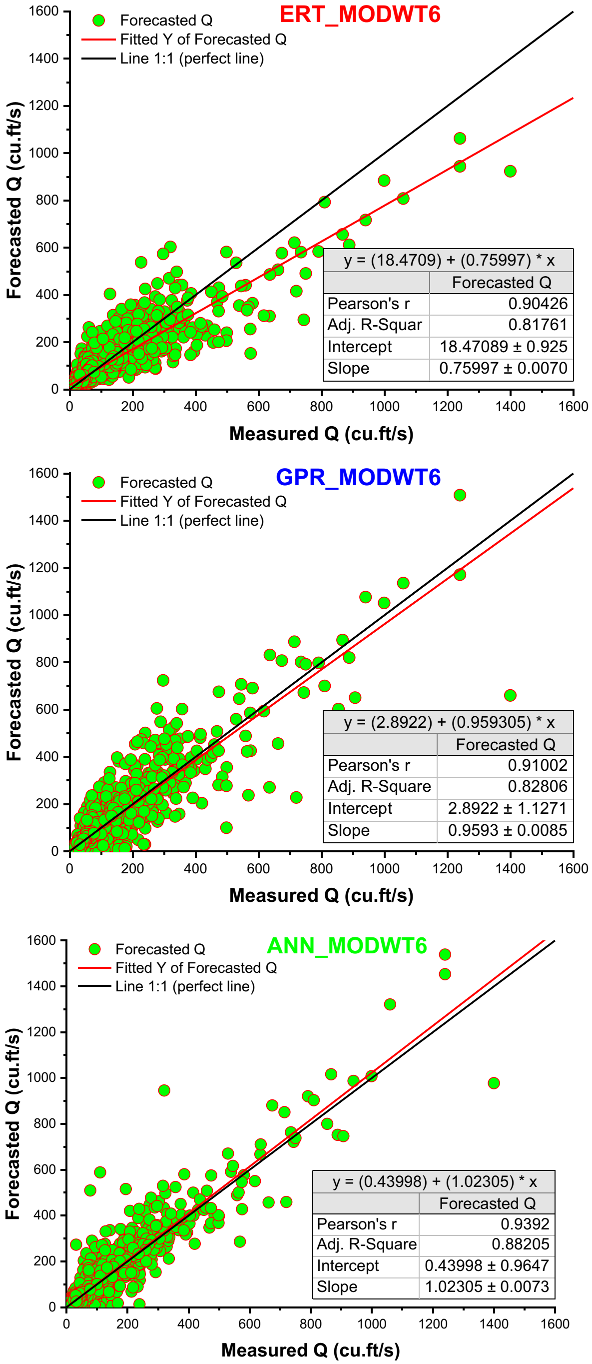


Fig. 13 Scatter plots of measured versus predicted daily discharge (Q) at USGS 14211550 for the validation phase: ERT_MODWT6 (left), GRP_MODWT6 (center), and ANN _MODWT6 (right)

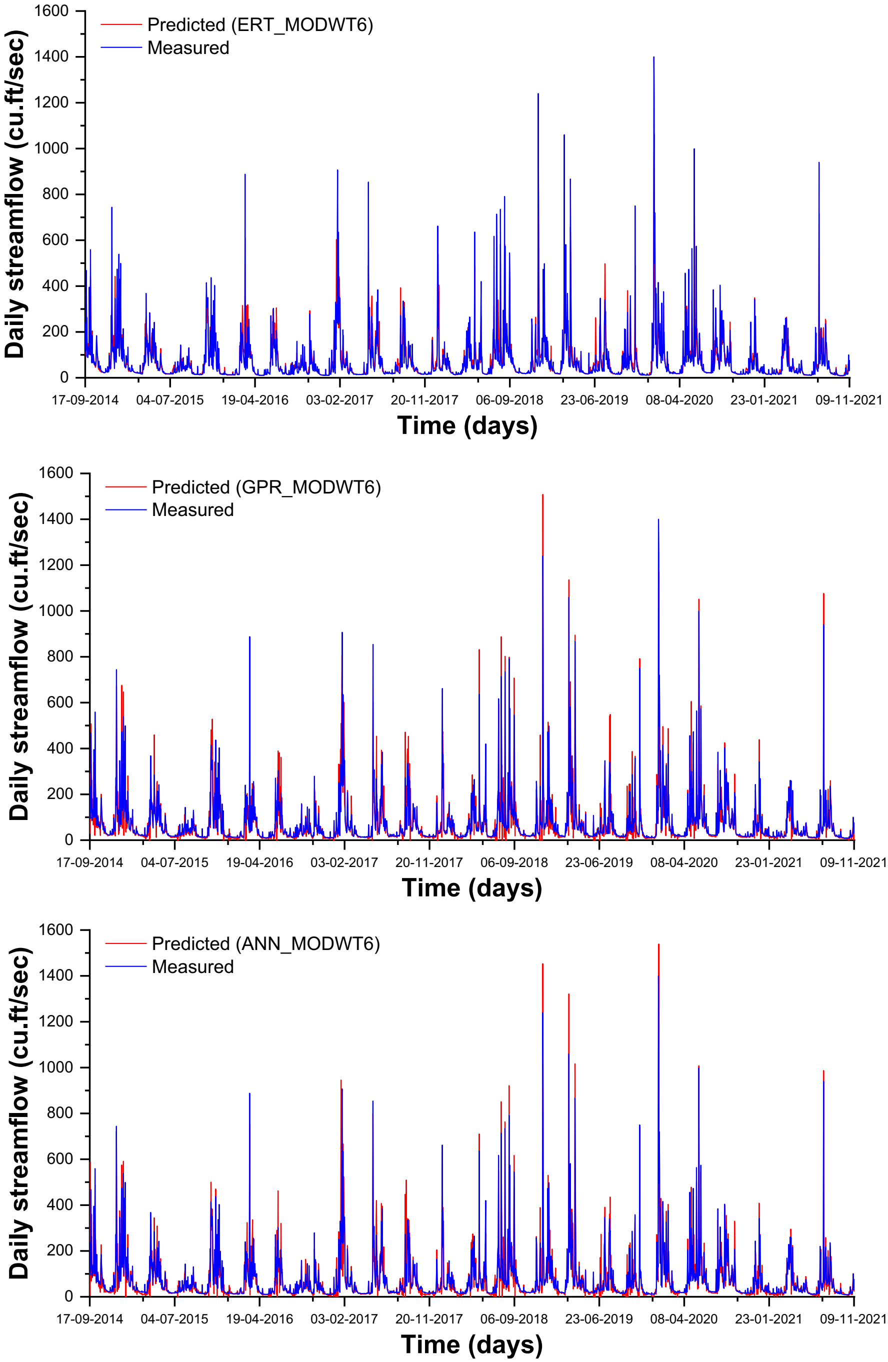


Fig. 14 Comparison between measured and predicted daily runoff (Q) at USGS 14211550 for the validation phase: ERT_MODWT6 (top panel), GRP_MODWT6 (middle panel), and ANN _MODWT6 (bottom panel)

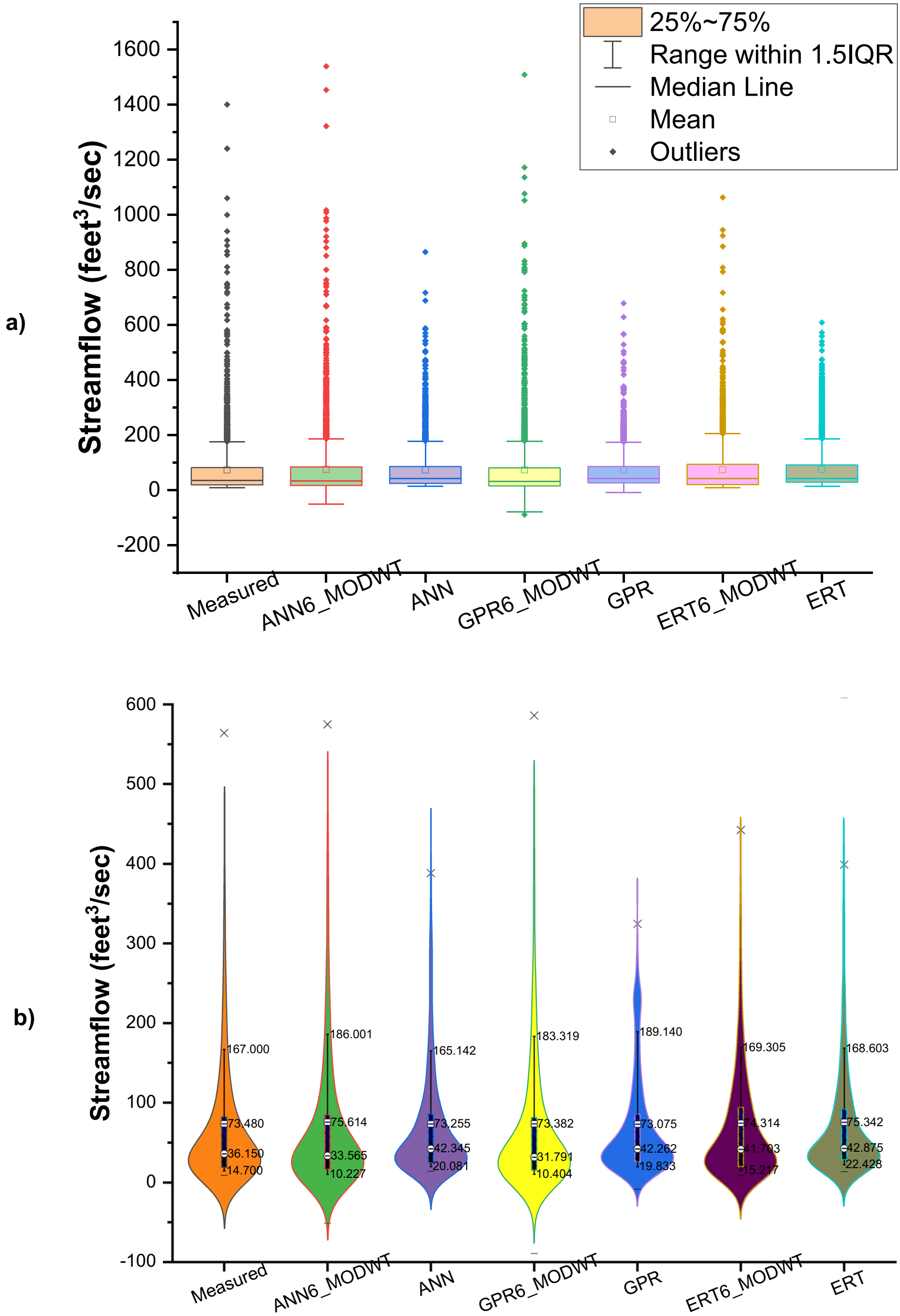
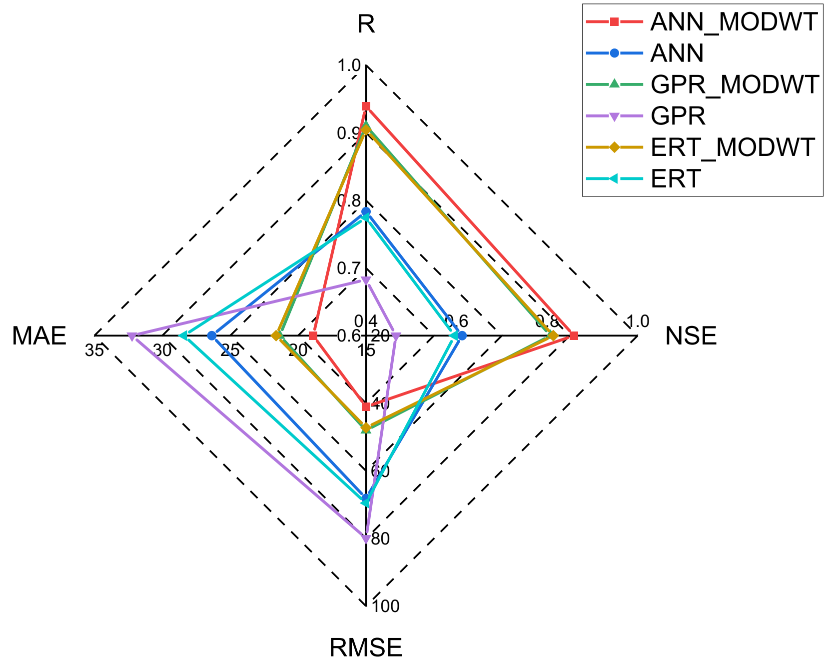


Fig. 15 Performance plots of best models during validation at USGS 14211550: **a**) boxplot, **b**) violin plot, **c**) radar chart, **d**) Taylor diagram

c)



d)

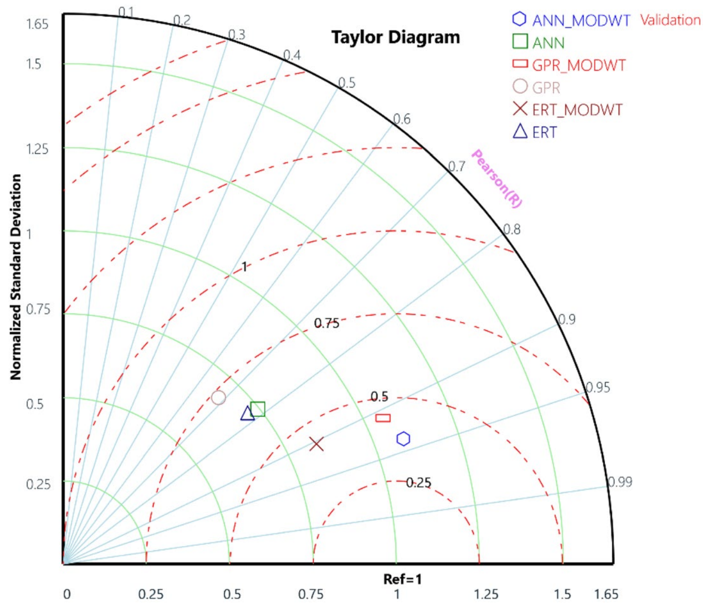


Figure 15 (continued)

performance. These results support the integration of signal preprocessing techniques like MODWT with advanced machine learning algorithms for robust hydrological forecasting.

4 Discussion

The literature review highlights the growing interest in machine learning-based artificial intelligence methods for improving runoff forecasting. Various models, such as ANN, SVM, RF, EDB, M5 Tree, and ensemble algorithms, have been explored for their effectiveness in predicting streamflow (Martinho et al. 2023). The study contributes by introducing the ERT model coupled with MODWT for signal decomposition, which, as demonstrated, significantly enhances prediction accuracy. The initial analysis without signal preprocessing revealed that all three machine learning models (ERT, GPR, and ANN) exhibited significant correlations between observed and forecasted streamflow data. The predictive ability of the ANN6 standalone model outperformed ERT6 and GPR6. However, the results indicated that the model's performance was notably improved when MODWT was applied for signal decomposition. This emphasizes the importance of preprocessing techniques in refining the input data for more accurate predictions (Mouatadid et al. 2019).

The results show that MODWT had a substantial positive impact on model performance (Alenezy et al. 2021; Jiang et al. 2021; Esmaeili-Gisavandani et al. 2022). Among the models, ANN consistently outperformed ERT and GPR, demonstrating the sensitivity of ANN to the MODWT preprocessing (Alenezy et al. 2022; Zhang et al. 2024). The study suggests that MODWT provides a crucial signal decomposition method, particularly when dealing with nonlinear and nonstationary discharges. The results indicate that ANN6_MODWT consistently outperformed ERT6_MODWT and GPR6_MODWT in terms of statistical indices R, NSE, RMSE, and MAE. Compared to previous studies on this topic applied in different modelling contexts and using different data sets, the output from this study was consistent (Mouatadid et al. 2019; Ashok and Yadav 2020; Sadek and Abdulrazak 2021; Heddam et al. 2023). The current study's use of ANN with MODWT aligns with these findings, showing that wavelet-based pre-processing (like MODWT) can significantly boost ANN's performance. Saraiva et al. (2021) used wavelet decomposition with SVR, ANN, and bootstrap neural networks for daily discharge forecasting, concluding that the incorporation of signal decomposition methods (e.g., wavelet transforms) generally enhances prediction accuracy across various models. This is consistent with the findings of the current study, where MODWT was a crucial component in improving all machine learning models' performances.

The thorough analysis includes each model's scatter plots, performance metrics, and comparative assessments. Using Taylor diagrams and radar charts enhances the visualization of model performance. The diagrams clearly illustrate the superiority of ANN6_MODWT over ERT6_MODWT and GPR6_MODWT. The radar charts provide a concise summary of statistical metrics, reinforcing the effectiveness of ANN coupled with MODWT. Geurts et al. (2006), the original developers of the ERT algorithm, have shown that ERT can handle large datasets and noisy data. The study's results align with these findings, but ERT, even when coupled with MODWT, was less effective than ANN in handling complex, nonlinear streamflow processes. Peng et al. (2017) explored the use of Empirical Wavelet Transform (EWT) with ANN and SVR, indicating that wavelet preprocessing generally benefits machine learning models, but GPR's relative performance lagged behind ANN. The current study similarly shows that while GPR_MODWT improves performance, it is still less competitive compared to ANN and ERT models.

It was observed that the ANN outperformed ERT in capturing complex, non-linear relationships (e.g., streamflow prediction), the ERT has several advantages that make it an attractive choice for certain applications. ERT is simpler to implement and interpret, as decision trees provide clear, understandable rules, whereas ANN operates as a "black box." It also requires less data preprocessing—no feature scaling or encoding of categorical variables is needed. Additionally, ERT is less prone to overfitting due to its inherent randomness, handles missing data naturally, and trains faster, especially on smaller datasets. These advantages make ERT a robust and efficient option in cases where model transparency, training speed, and ease of use are important.

4.1 Future Directions

The study paves the way for future research directions in runoff prediction. Further investigations could explore the application of the proposed models in different geographical locations, evaluate the robustness of the models under varying climatic conditions, and consider additional input variables for a more comprehensive analysis. In conclusion, this study introduces a novel approach to runoff prediction by combining the ERT model with MODWT for signal decomposition. The results demonstrate the efficacy of this approach, particularly when applied to the ANN model, leading to significant improvements in prediction accuracy. The findings contribute valuable insights to the evolving landscape of machine learning applications in hydrology and water resources management.

5 Conclusion

This study proposed an integrated approach for daily runoff prediction by coupling three machine learning models—Artificial Neural Network (ANN), Extremely Randomized Trees (ERT), and Gaussian Process Regression (GPR)—with Maximal Overlap Discrete Wavelet Transform (MODWT). Using six time-lagged discharge inputs identified through autocorrelation analysis, the models were trained and validated at two USGS stations (14211500 and 14211550) under scenarios with and without MODWT preprocessing. The results demonstrate that integrating MODWT significantly enhanced the predictive accuracy of all three models. Among them, the ANN_MODWT6 model achieved the highest accuracy at both stations, showing improvements of 15.60% (R), 24.70% (NSE), 39.74% (RMSE), and 28.34% (MAE) at USGS 14211550, and 13.50%, 23.80%, 46.47%, and 34.06%, respectively, at USGS 14211500. Similarly, ERT_MODWT6 outperformed its standalone version, improving by 13.00–21.80% at station 14211550 and 7.80–21.43% at station 14211500 across the same metrics. Despite being the least effective standalone model, even GPR benefited substantially from MODWT, closing much of the performance gap.

These findings confirm that MODWT effectively extracts meaningful signal components from time series data, enhancing model learning. The performance differences between the two stations are attributed to variability in hydrological characteristics and flow patterns. Overall, integrating MODWT with machine learning models proves to be a powerful strategy for improving runoff forecasting accuracy. Future research should explore additional decomposition techniques and datasets to further validate the robustness of this approach.

Acknowledgements The authors extend their appreciation to Ongoing Research Funding program - Research Chairs (ORF-RC-2025-5501), King Saud University, Riyadh, Saudi Arabia. The author expresses gratitude to the USGS data survey for their invaluable support, without which this study would not have been possible. Special appreciation is also extended to the dedicated staff of the USGS web server for providing the necessary data for this research.

Author Contributions Conceptualization, S.H., and D.K.V.; Methodology, S.H., and D.K.V.; Software, S.H., D.K.V.; Validation, S.H., and D.K.V.; Formal Analysis, S.H., M.A.M., and D.K.V.; Investigation, S.H.; Resources, S.H.; Data Curation, S.H., and D.K.V.; Writing– Original Draft Preparation, S.H., R.K.T., and D.K.V.; Writing– Review & Editing, A.M., A.G., O.K., D.K.V., C.B. A.A., A.Z.D. and M.A.M.; Visualization, S.H., M.A.M., and D.K.V.; Supervision, S.H., and D.K.V.; Project Administration, S.H. All authors have read and agreed to the published version of the manuscript.

Funding This research was funded by Ongoing Research Funding program - Research Chairs (ORF-RC-2025-5501), King Saud University, Riyadh, Saudi Arabia.

Data Availability The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethical Approval Not applicable.

Consent to Participate Author consent to participate in this research/publication.

Consent to Publish This research does not have any consent to publish.

Competing interests The author declares that they have no conflicts of interest.

References

Adnan RM, Jaafari A, Mohanavelu A et al (2021) Novel ensemble forecasting of streamflow using locally weighted learning algorithm. Sustainability 13:5877. https://doi.org/10.3390/su13115877

Alenezy AH, Ismail MT, Wadi S, Al et al (2021) Forecasting stock market volatility using hybrid of adaptive network of fuzzy inference system and wavelet functions. J Math 2021:1–10. https://doi.org/10.1155/2021/9954341

Alenezy AH, Ismail MT, Jaber JJ et al (2022) Hybrid fuzzy inference rules of descent method and wavelet function for volatility forecasting. PLoS ONE 17:e0278835. https://doi.org/10.1371/journal.pone.0278835

Ashok V, Yadav A (2020) A real-time fault detection and classification algorithm for transmission line faults based on MODWT during power swing. Int Trans Electr Energy Syst 30:e12164. https://doi.org/10.1002/2050-7038.12164

Bajirao TS, Kumar P, Kumar M et al (2021) Potential of hybrid wavelet-coupled data-driven-based algorithms for daily runoff prediction in complex river basins. Theor Appl Climatol 145:1207–1231. https://doi.org/10.1007/s00704-021-03681-2

Barzegar R, Aalami MT, Adamowski J (2021) Coupling a hybrid CNN-LSTM deep learning model with a boundary corrected maximal overlap discrete wavelet transform for multiscale lake water level forecasting. J Hydrol (Amst) 598:126196. https://doi.org/10.1016/j.jhydrol.2021.126196

Bouaziz M, Medhioub E, Csaplovisc E (2021) A machine learning model for drought tracking and forecasting using remote precipitation data and a standardized precipitation index from arid regions. J Arid Environ 189:104478. https://doi.org/10.1016/j.jaridenv.2021.104478

Cai L, Yin H, Lin J, Zhao D (2023) An update-strategy-based Gaussian process regression method for aeroengines fault prediction. IEEE Trans Industr Inf 1–10. https://doi.org/10.1109/TII.2023.3283103

Cheng H, Shi Y, Wu L et al (2021a) An intelligent scheme for big data recovery in internet of things based on multi-attribute assistance and extremely randomized trees. Inf Sci (N Y) 557:66–83. https://doi.org/10.1016/j.ins.2020.12.041

Cheng H, Wu L, Li R et al (2021b) Data recovery in wireless sensor networks based on attribute correlation and extremely randomized trees. J Ambient Intell Humaniz Comput 12:245–259. https://doi.org/10.1007/s12652-019-01475-z

Ditthakit P, Pinthong S, Salaeh N et al (2023) Comparative study of machine learning methods and GR2M model for monthly runoff prediction. Ain Shams Eng J 14:101941. https://doi.org/10.1016/j.asej.2022.101941

Dubey K, Jain YK, Sharma V (2022) A comparative study of deep learning based MODWT-LSTM and machine learning model to predict environmental parameters of polyhouse. In: 2022 IEEE international conference on current development in engineering and technology (CCET). IEEE, pp 1–8

Duong TD, Tran VN, Nguyen TV (2025) Evaluating rainfall-runoff generation mechanisms of deep learning models using a process-based rainfall-runoff model. Water Resources Management. https://doi.org/10.1007/s11269-025-04231-5

Elbeltagi A, Kushwaha NL, Rajput J et al (2022) Modelling daily reference evapotranspiration based on stacking hybridization of ANN with meta-heuristic algorithms under diverse agro-climatic conditions. Stochastic Environmental Research and Risk Assessment. https://doi.org/10.1007/s00477-022-02196-0

Esmaeili-Gisavandani H, Farajpanah H, Adib A et al (2022) Evaluating ability of three types of discrete wavelet transforms for improving performance of different ML models in estimation of daily-suspended sediment load. Arab J Geosci 15:29. https://doi.org/10.1007/s12517-021-09282-7

Fenicia F, McDonnell JJ (2022) Modeling streamflow variability at the regional scale: (1) perceptual model development through signature analysis. J Hydrol (Amst) 605:127287. https://doi.org/10.1016/j.jhydrol.2021.127287

Geurts P, Ernst D, Wehenkel L (2006) Extremely randomized trees. Mach Learn 63:3–42. https://doi.org/10.1007/s10994-006-6226-1

Ghimire S, Deo RC, Raj N, Mi J (2019) Wavelet-based 3-phase hybrid SVR model trained with satellite-derived predictors, particle swarm optimization and maximum overlap discrete wavelet transform for solar radiation prediction. Renew Sustain Energy Rev 113:109247. https://doi.org/10.1016/j.rser.2019.109247

Heddam S, Merabet K, Difi S et al (2023) River water temperature prediction using hybrid machine learning coupled signal decomposition: EWT versus MODWT. Ecol Inf 78:102376. https://doi.org/10.1016/j.ecoinf.2023.102376

Hornik K (1991) Approximation capabilities of multilayer feedforward networks. Neural Netw 4:251–257. https://doi.org/10.1016/0893-6080(91)90009-T

Hornik K, Stinchcombe M, White H (1989) Multilayer feedforward networks are universal approximators. Neural Netw 2:359–366. https://doi.org/10.1016/0893-6080(89)90020-8

Huang C, Sun K, Hu J et al (2022) Estimating 2013–2019 NO2 exposure with high spatiotemporal resolution in China using an ensemble model. Environ Pollut 292:118285. https://doi.org/10.1016/j.envpol.2021.118285

Jiang Z, Sharma A, Johnson F (2021) Variable transformations in the spectral domain– implications for hydrologic forecasting. J Hydrol (Amst) 603:126816. https://doi.org/10.1016/j.jhydrol.2021.126816

Khan MT, Shoaib M, Albano R et al (2023) Intercomparison and assessment of stand-alone and wavelet-coupled machine learning models for simulating rainfall-runoff process in four basins of Pothohar region, Pakistan. Atmos (Basel) 14:452. https://doi.org/10.3390/atmos14030452

Khosravi K, Golkarian A, Tiefenbacher JP (2022) Using optimized deep learning to predict daily streamflow: a comparison to common machine learning algorithms. Water Resour Manage 36:699–716. https://doi.org/10.1007/s11269-021-03051-7

Kulkarni PG, Sahasrabudhe AD (2017) Investigations on mother wavelet selection for health assessment of lathe bearings. Int J Adv Manuf Technol 90:3317–3331. https://doi.org/10.1007/s00170-016-9664-3

Li W, Fan Y, Ringbeck F et al (2022) Unlocking electrochemical model-based online power prediction for lithium-ion batteries via Gaussian process regression. Appl Energy 306:118114. https://doi.org/10.1016/j.apenergy.2021.118114

Liu H, Yang C, Huang M et al (2018) Modeling of subway indoor air quality using Gaussian process regression. J Hazard Mater 359:266–273. https://doi.org/10.1016/j.jhazmat.2018.07.034

Martinho AD, Hippert HS, Goliatt L (2023) Short-term streamflow modeling using data-intelligence evolutionary machine learning models. Sci Rep 13:13824. https://doi.org/10.1038/s41598-023-41113-5

Mouatadid S, Adamowski JF, Tiwari MK, Quilty JM (2019) Coupling the maximum overlap discrete wavelet transform and long short-term memory networks for irrigation flow forecasting. Agric Water Manag 219:72–85. https://doi.org/10.1016/j.agwat.2019.03.045

Patnaik B, Mishra M, Bansal RC, Jena RK (2021) MODWT-XGBoost based smart energy solution for fault detection and classification in a smart microgrid. Appl Energy 285:116457. https://doi.org/10.1016/j.apenergy.2021.116457

Peng T, Zhou J, Zhang C, Fu W (2017) Streamflow forecasting using empirical wavelet transform and artificial neural networks. Water (Basel) 9:406. https://doi.org/10.3390/w9060406

Poursaeid M (2025) A novel approach for water resources management to accurate river flow prediction: utilization of linear programming and weighted ensemble learning. Water Resour Manag. https://doi.org/10.1007/s11269-025-04230-6

Quilty J, Adamowski J (2021) A maximal overlap discrete wavelet packet transform integrated approach for rainfall forecasting– a case study in the Awash river basin (Ethiopia). Environ Model Softw 144:105119. https://doi.org/10.1016/j.envsoft.2021.105119

Rasmussen CE, Williams CKI (2006) Gaussian processes for machine learning. Springer

Roy P, Pal SC, Arabameri A et al (2021) Climate and land use change induced future flood susceptibility assessment in a sub-tropical region of India. Soft Comput 25:5925–5949. https://doi.org/10.1007/s00500-021-05584-w

Sadek I, Abdulrazak B (2021) A comparison of three heart rate detection algorithms over ballistocardiogram signals. Biomed Signal Process Control 70:103017. https://doi.org/10.1016/j.bspc.2021.103017

Saraiva SV, Carvalho F, de Santos O CAG, et al (2021) Daily streamflow forecasting in Sobradinho reservoir using machine learning models coupled with wavelet transform and bootstrapping. Appl Soft Comput 102:107081. https://doi.org/10.1016/j.asoc.2021.107081

Shukla R, Kumar P, Vishwakarma DK et al (2021) Modeling of stage-discharge using back propagation ANN-, ANFIS-, and WANN-based computing techniques. Theor Appl Climatol. https://doi.org/10.1007/s00704-021-03863-y

Simon H (1999) Neural networks: a comprehensive foundation, 2nd edn. Prentice hall, Upper Saddle River; New Jersey, p 07458

Singh AK, Kumar P, Ali R et al (2022) An integrated Statistical-Machine learning approach for runoff prediction. Sustainability 14:8209. https://doi.org/10.3390/su14138209

Talei A (2022) Artificial intelligence in real-time rainfall-runoff modelling and flood forecasting. In: Ong HL, Doong R, Naguib R et al (eds) Artificial intelligence and environmental sustainability. Algorithms for intelligent systems. Springer Nature Singapore, Singapore, pp 87–103

Tikhamarine Y, Souag-Gamane D, Najah Ahmed A et al (2020) Improving artificial intelligence models accuracy for monthly streamflow forecasting using grey Wolf optimization (GWO) algorithm. J Hydrol (Amst) 582:124435. https://doi.org/10.1016/j.jhydrol.2019.124435

Tran VN, Nguyen HD, Van Khuong H et al (2025) Reconstructing long-term daily streamflow data at the discontinuous monitoring station in the ungauged transboundary basin using machine learning. Water Resour Manag. https://doi.org/10.1007/s11269-025-04109-6

Vatanchi SM, Etemadfard H, Maghrebi MF, Shad R (2023) A comparative study on forecasting of long-term daily streamflow using ANN, ANFIS, BiLSTM and CNN-GRU-LSTM. Water Resour Manage 37:4769–4785. https://doi.org/10.1007/s11269-023-03579-w

Wei H, Luo K, Xing J, Fan J (2022) Predicting co-pyrolysis of coal and biomass using machine learning approaches. Fuel 310:122248. https://doi.org/10.1016/j.fuel.2021.122248

Yang Y, Li S, Li W, Qu M (2018) Power load probability density forecasting using Gaussian process quantile regression. Appl Energy 213:499–509. https://doi.org/10.1016/j.apenergy.2017.11.035

Yang B, Ding Y, Lu Z et al (2021) Intelligent computing of positive switching impulse breakdown voltage of rod-plane air gap based on extremely randomized trees algorithm. Electr Eng 103:3177–3187. https://doi.org/10.1007/s00202-021-01307-4

Yu X, Wang Y, Wu L et al (2020) Comparison of support vector regression and extreme gradient boosting for decomposition-based data-driven 10-day streamflow forecasting. J Hydrol (Amst) 582:124293. https://doi.org/10.1016/j.jhydrol.2019.124293

Zhang T, Chen W, Li M (2019) Classification of inter-ictal and ictal EEGs using multi-basis MODWPT, dimensionality reduction algorithms and LS-SVM: A comparative study. Biomed Signal Process Control 47:240–251. https://doi.org/10.1016/j.bspc.2018.08.038

Zhang J, Liu H, Bai W, Li X (2024) A hybrid approach of wavelet transform, ARIMA and LSTM model for the share price index futures forecasting. North Am J Econ Finance 69:102022. https://doi.org/10.1016/j.najef.2023.102022

Zhou Q, Li Y, Zhao D et al (2022) Transferable representation modelling for real-time energy management of the plug-in hybrid vehicle based on k-fold fuzzy learning and Gaussian process regression. Appl Energy 305:117853. https://doi.org/10.1016/j.apenergy.2021.117853

Zounemat-Kermani M, Mahdavi-Meymand A, Hinkelmann R (2021) A comprehensive survey on conventional and modern neural networks: application to river flow forecasting. Earth Sci Inf 14:893–911. https://doi.org/10.1007/s12145-021-00599-1

Zuo G, Luo J, Wang N et al (2020) Decomposition ensemble model based on variational mode decomposition and long short-term memory for streamflow forecasting. J Hydrol (Amst) 585:124776. https://doi.org/10.1016/j.jhydrol.2020.124776

Authors and Affiliations

Dinesh Kumar Vishwakarma[1] · Salim Heddam[2] · Arpit Gaur[3] · Ravindra Kumar Tiwari[4] · Ozgur Kisi[5,6] · Anurag Malik[7] · Chetak Bishnoi[7] · Abed Alataway[8] · Ahmed Z. Dewidar[8] · Mohamed A. Mattar[8]

Dinesh Kumar Vishwakarma
dinesh.vishwakarma4820@gmail.com

Mohamed A. Mattar
mmattar@ksu.edu.sa

Salim Heddam
heddamsalim@yahoo.fr

Arpit Gaur
arpit.gaur@montana.edu

Ravindra Kumar Tiwari
ravindratiwari15@gmail.com

Ozgur Kisi
ozgur.kisi@th-luebeck.de

Anurag Malik
amalik19@pau.edu

Chetak Bishnoi
chetakbishnoi@pau.edu

Abed Alataway
aalataway@ksu.edu.sa

Ahmed Z. Dewidar
adewidar@ksu.edu.sa

1 Department of Irrigation and Drainage Engineering, Govind Ballabh Pant University of Agriculture and Technology, Pantnagar, Uttarakhand 263145, India

2 Faculty of Science, Agronomy Department, Hydraulics Division, University 20 Août 1955, Route El Hadaik, BP 26, Skikda, Algeria

3 Montana State University, Bozeman, MT, USA

4 Department of Post Harvest Technology, College of Horticulture and Forestry, Rani Lakshmi Bai Central Agricultural University, Jhansi 284003, Uttar Pradesh, India

5 Department of Civil Engineering, Lübeck University of Applied Sciences, 23562 Lübeck, Germany

6 Department of Civil Engineering, Ilia State University, Tbilisi 0162, Georgia

7 Punjab Agricultural University, Regional Research Station, Bathinda-151001, Punjab, India

8 Prince Sultan Bin Abdulaziz International Prize for Water Chair, Prince Sultan Institute for Environmental, Water and Desert Research, King Saud University, P.O. Box 2454 Riyadh 11451, Saudi Arabia