



Hybrid river stage forecasting based on machine learning with empirical mode decomposition

Salim Heddami¹ · Dinesh Kumar Vishwakarma² · Salwan Ali Abed³ · Pankaj Sharma⁴ · Nadhir Al-Ansari⁵ · Abed Alataway⁶ · Ahmed Z. Dewidar^{6,7} · Mohamed A. Mattar^{6,7,8}

Received: 9 September 2023 / Accepted: 10 January 2024 / Published online: 15 February 2024
© The Author(s) 2024

Abstract

The river stage is certainly an important indicator of how the water level fluctuates overtime. Continuous control of the water stage can help build an early warning indicator of floods along rivers and streams. Hence, forecasting river stages up to several days in advance is very important and constitutes a challenging task. Over the past few decades, the use of machine learning paradigm to investigate complex hydrological systems has gained significant importance, and forecasting river stage is one of the promising areas of investigations. Traditional in situ measurements, which are sometime restricted by the existing of several handicaps especially in terms of regular access to any points alongside the streams and rivers, can be overpassed by the use of modeling approaches. For more accurate forecasting of river stages, we suggest a new modeling framework based on machine learning. A hybrid forecasting approach was developed by combining machine learning techniques, namely random forest regression (RFR), bootstrap aggregating (Bagging), adaptive boosting (AdaBoost), and artificial neural network (ANN), with empirical mode decomposition (EMD) to provide a robust forecasting model. The singles models were first applied using only the river stage data without preprocessing, and in the following step, the data were decomposed into several intrinsic mode functions (IMF), which were then used as new input variables. According to the obtained results, the proposed models showed improved results compared to the standard RFR without EMD for which, the error performances metrics were drastically reduced, and the correlation index was increased remarkably and great changes in models' performances have taken place. The RFR_EMD, Bagging_EMD, and AdaBoost_EMD were less accurate than the ANN_EMD model, which had higher $R \approx 0.974$, $NSE \approx 0.949$, $RMSE \approx 0.330$ and $MAE \approx 0.175$ values. While the RFR_EMD and the Bagging_EMD were relatively equal and exhibited the same accuracies higher than the AdaBoost_EMD, the superiority of the ANN_EMD was obvious. The proposed model shows the potential for combining signal decomposition with machine learning, which can serve as a basis for new insights into river stage forecasting.

Keywords Forecasting · River · Stage · Lag time · RFR · Bagging · Boosting · ANN

Introduction

Flood risk management has gained significant attention in recent years, and significant effort has been made to improve flood forecasting as well as the estimation methods over time and space (Hsu et al. 2010). The variation of river stage (RS) over time is mainly governed by several factors, i.e., land cover, precipitation, topography, and vegetation, and the interaction between surface water and groundwater has been well documented and highlighted (Marques et al. 2020; Shukla et al. 2021; Vishwakarma et al. 2023c). Accurate estimation of RS based on river models, i.e., channel and

hydrological models, helps significantly in achieving and providing estimation of river flood, which can make rapid and quickly estimating records of wetlands zones (Liu et al. 2021), and knowledge of RS is also an important factor in planning and management of watersheds, the development of water resources, and the watershed/aquifer management plans (Strupczewski et al. 2001; Khatibi et al. 2012; Shukla et al. 2021; Vishwakarma et al. 2023c). While the estimation of RS can easily be obtained using hydrodynamic models, their high number of parameters makes them harder to use, and the need for data-driven models presents itself as a possible alternative (Panda et al. 2010). Chau (2006) compared two machine learning models, *multilayer perceptron neural network* (MLPNN) and the MLPNN optimized using

Extended author information available on the last page of the article

particle swarm optimization training algorithm (MLPNN-PSO), for forecasting *RS* at one, two and seven days in advance using data collected at the Shing Mun River, Hong Kong. The obtained results showed that the MLPNN-PSO was more accurate than the MLPNN at the three forecasting horizons, with correlation coefficients (*R*) between 0.92 and 0.99. Chau (2007) applied a split-step particle swarm optimization algorithm (SPS) for forecasting *RS* at one and two days ahead using data collected at the Shing Mun River, Hong Kong. The SPS was more accurate than the MLPNN, with *R* values of 0.986 and 0.979 for one and two days ahead, respectively. In order to predict daily *RS* in the Yangtze River, China, Wu et al. (2008) compared four machine learning models namely: support vector regression (SVR), nearest-neighbor regression (NNR), MLPNN, and linear regression (LR). The best accuracy was obtained using the SVR model (RMSE=0.211m), followed by the MLPNN optimized using genetic algorithm (RMSE=0.237m), the LR and the NNR with (RMSE=0.237m) and (RMSE=0.242m), respectively.

Adaptive neuro-fuzzy inference system (ANFIS), MLPNN, wavelet packet decomposition combined ANFIS (ANFIS-WPD), and wavelet packet decomposition combined MLPNN (MLPNN-WPD) were the models that Seo and Kim (2016) compared between singles and hybrid data-driven models for forecasting *RS*. In order to predict the *RS* at the time (*t*), the authors used the *RS* measured at several previous lag times, i.e., from (*t* - 1) to (*t* - 6). The hybrid models, ANFIS-WPD ($R^2=0.999$) and MLPNN-WPD ($R^2=0.988$), performed better than the single models, ANFIS ($R^2=0.963$) and MLPNN ($R^2=0.963$). Using grey neural network (GNN), Alvisi and Franchini (2012) proposed a new modeling approach for more accurate *RS* prediction. The GNN was developed using data collected at the Reno River, Italy, and the GNN was compared to the Bayesian neural network (BNN), showing its superiority. In different investigation, Fu et al. (2019) utilized the multiple additive regression trees (MART) for hourly *RS* forecasting at the Bazhang River basin in southern Taiwan. Three models were compared namely, the MART, real-time MART, and naïve MART models. The real-time MART was more accurate for all three-time steps, *t* + 1, *t* + 2, and *t* + 3, showing low RMSE and High *R* values.

For forecasting *RS* at the Schuylkill River at Berne, Philadelphia, USA, Kisi (2011) developed a hybrid model (LR-DWT) using the discrete wavelet transform (DWT) and linear regression (LR). The authors demonstrated that the LR-DWT model outperformed the MLPNN model by highlighting the significant role that the DWT played in enhancing the linear model's accuracy. Seo et al. (2016a) compared between hybrid wavelet packet decomposition combined

SVR (SVR-WPD), ANFIS-WPD, MLPNN-WPD, SVR, MLPNN and ANFIS models in forecasting *RS* in the Gam Stream watershed, South Korea. Obtained results revealed the superiority of the SVR-WPD which gives R^2 , RMSE, and MAE of approximately 0.996, 0.0256, and 0.0127, respectively. Seo et al. (2016b) used a variety of methods to optimize the SVR parameters, including genetic algorithm (GA), grid search (GS), artificial bee colony (ABC), and particle swarm optimization (PSO). The SVR-PSO and SVR-ABC were more accurate than the others models showing higher R^2 values and lower RMSE and MAE values, respectively, according to the results.

In order to maximize the usefulness of river stage studies, it is necessary to know the already published works in this subject. It is clear from the literature discussed above that machine learning was designed to meet the necessary quality of forecasting. Multiple algorithms have been developed and successfully applied. However, the major findings of the previous studies can be summarized as follow: (i) in the major cases studies discussed above, the proposed machine learning involves the combination of various features for improving the forecasting accuracy, and the comparison was done between standalone machine learning belonging to various categories, i.e., ANN, SVR, Neuroplus, among other, (ii) one major finding is that, the hybridization of machine learning with metaheuristics algorithms was found to be necessary for improving the performances of the developed models, making the comparability of the obtained results difficult, (iii) few investigations have highlighted the importance and the contribution of the preprocessing signal decomposition in improving the estimation of river stage, as signal decomposition help in capturing the high nonlinearity in the dataset. Furthermore, an important finding is that the ensemble methods are rarely reported for river stage forecasting weather with or without signal decomposition.

Therefore, for extensive research on water resources and flood management, accurate estimation of *RS* is essential for large-scale studies. However, estimation of *RS* is difficult due to the large number of variables influencing its fluctuation. The above-discussed literature review makes it clear that *RS* prediction using machine learning was widely discussed in the literature and that many models had been proposed and successfully used. Although hybrid models based on signal decomposition have been suggested for *RS* forecasting, empirical mode decomposition (EMD) performance studies have not been conducted to date. This paper therefore presents a new modeling strategy for better prediction of *RS* using machine learning models [i.e., random forest regression (RFR), bootstrap aggregating (Bagging), adaptive boosting (AdaBoost), and artificial neural network (ANN)] with the EMD.

Materials and methods

Study site

River stage data were collected from two USGS stations (Fig. 1): (i) the USGS 14210000 Clackamas River at Estacada, Clackamas County, Oregon (Latitude $45^{\circ}18'00''$, Longitude $122^{\circ}21'10''$ NAD27), and (ii) the USGS 14211499 Kelley Creek at Se 159th drive at Portland, Multnomah County, Oregon (Latitude $45^{\circ}28'37''$, Longitude $122^{\circ}29'50''$ NAD27). Data were collected for the USGS 14210000 station between January 1, 2002, and December 31, 2019, and were recorded on a daily time scale (6574 data). These data were divided into training (70%) and validation (30%), with 4598 data being used for training and 1970 for validation. In a similar manner, data were collected for the USGS 14211499 station between April 9, 2000, and December 31, 2020 on a daily time

scale (7572 data), which were split into training (70%) and validation (30%). As a result, 5297 data were used for training and 2269 for validation. For daily river stage (RS), we present the mean, maximum, minimum, standard deviation, and coefficient of variation values in Table 1 as follows: X_{mean} , X_{max} , X_{min} , S_x , and C_v . For better forecasting of daily rivers stage (RS) using only the RS measured at previous lag time. Consequently, the autocorrelation function (ACF) and partial autocorrelation function (PACF) were used to choose the best relevant time lags (Fig. 2). River stage measurements at times $(t - 1)$, $(t - 2)$, $(t - 3)$, $(t - 4)$, $(t - 5)$, and $(t - 6)$ were chosen and used as input variables according to Fig. 2, while the output variable started at the time (t) . The original chosen input variables, i.e., river stage measured at various previous lags, were divided into several intrinsic mode functions (IMF) in the second stage of the investigation using the empirical mode decomposition (EMD) approach (Fig. 3). The IMF obtained using the EMD were used as input variables. We

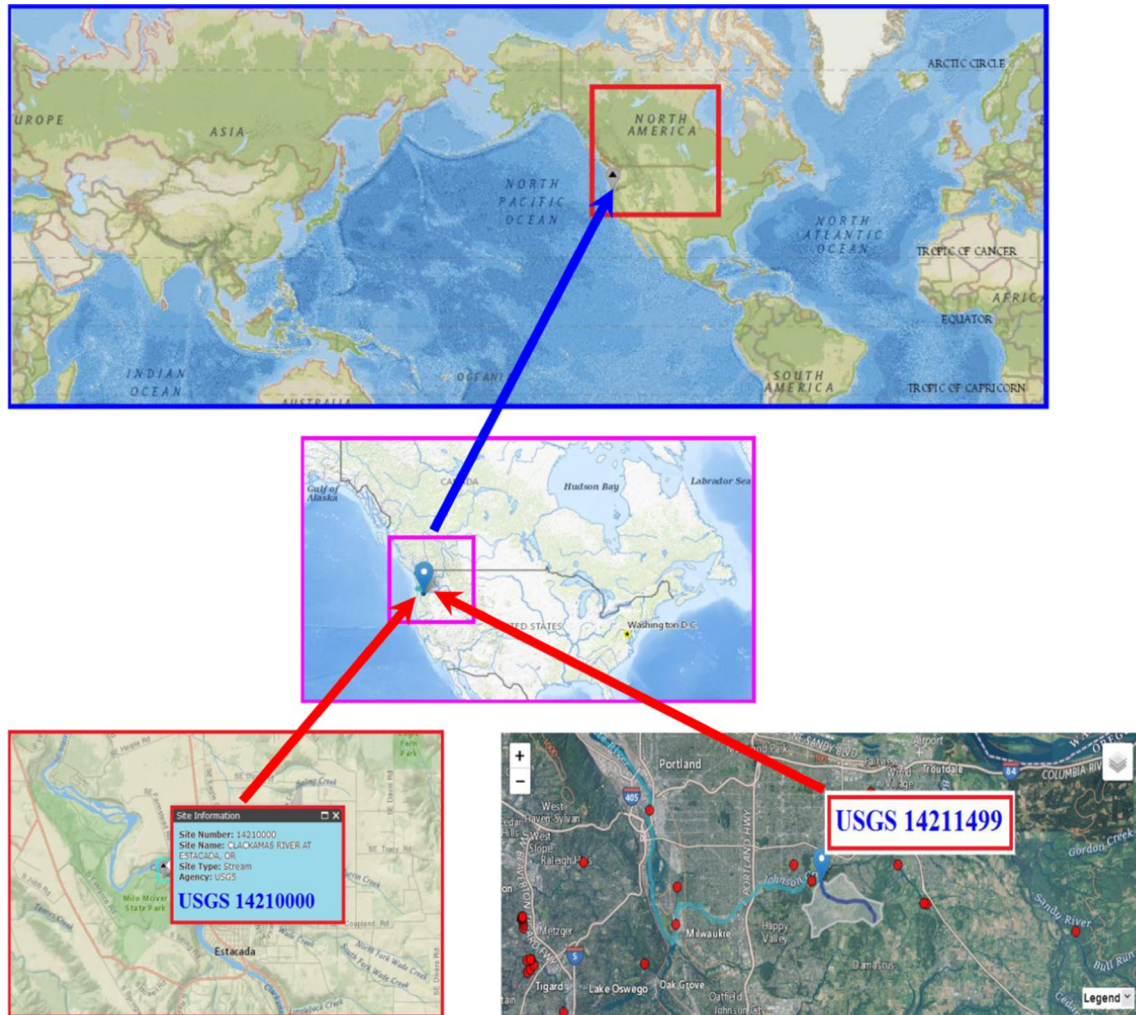


Fig. 1 Map showing the location the two USGS stations

Table 1 Statistical parameters of the river stage for the two stations

Variables	Subset	Unit	Xmax	Xmean	Xmin	Sx	Cv
USGS 14210000							
RS	Training	Feet	21.030	12.168	10.350	1.411	0.116
	Validation	Feet	20.580	12.180	10.350	1.424	0.117
	All data	Feet	21.030	12.172	10.350	1.425	0.117
USGS 14211499							
RS	Training	Feet	7.480	3.098	2.280	0.503	0.163
	Validation	Feet	5.720	3.077	2.300	0.504	0.164
	All data	Feet	7.480	3.092	2.280	0.504	0.163

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

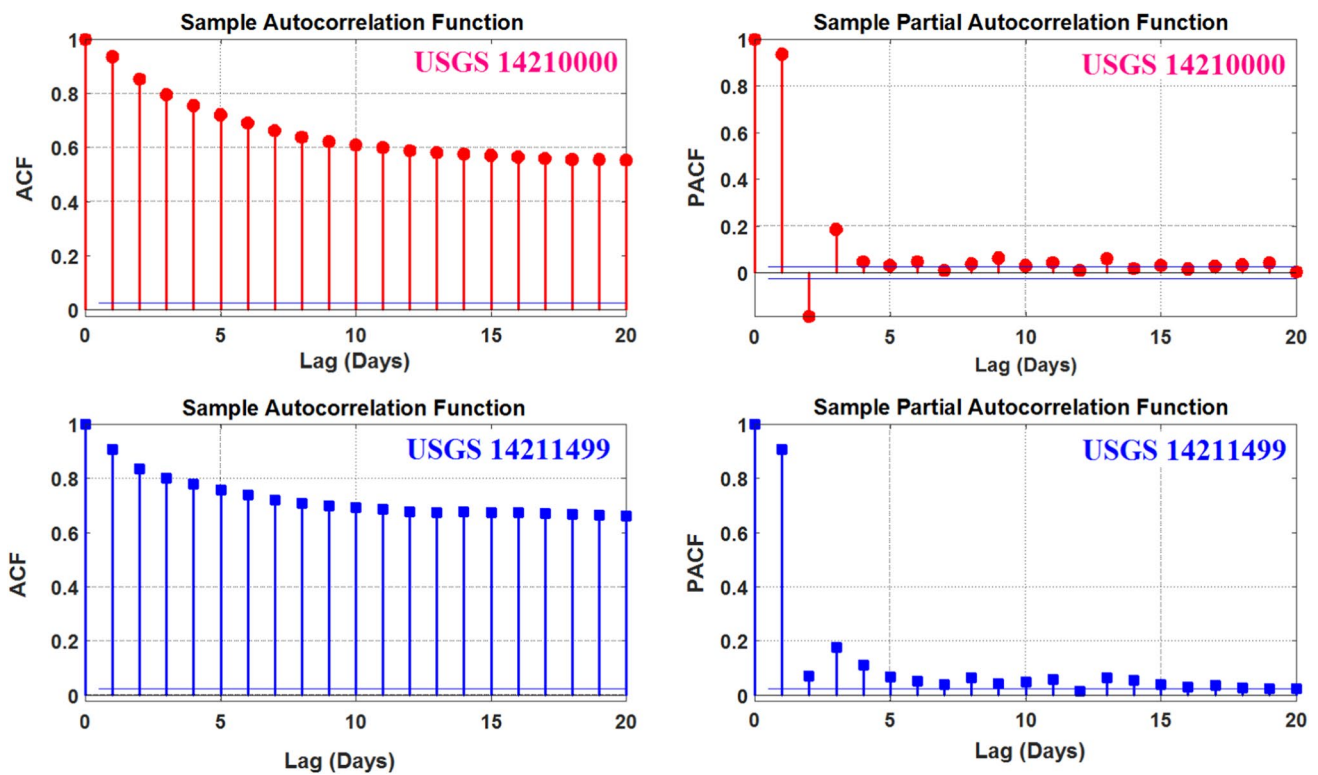


Fig. 2 Sample autocorrelation (ACF) and partial autocorrelation function (PACF) for daily river stage (RS)

used an eight-level decomposition in the current study, and the RFR, Bagging, AdaBoost and ANN all had forty-eight input variables. Figure 4 shows a flowchart of the suggested modeling approaches used in the current.

Artificial neural network (ANN)

The structure of the artificial neural network (ANN) used in the present study can be seen in Fig. 5. The network consists of three distinct parts that fit together. First, the input layer that collects the independent variables (i.e., the predictors) for which one neuron was attributed to each input variable. Second, the hidden layers with several processing neurons,

and third, the last layer, or the output layer, with only one neuron (Shukla et al. 2021; Elbeltagi et al. 2022). This kind of ANN model is called a multilayer perceptron (MLP). The mathematical operations for each layer can be described briefly as follow: Each neuron in the hidden layer computes a weighted sum of the independent variables available in the input layer (Saroughi et al. 2023); hence, for each one, we have:

$$\delta_j = \sum_{i=1}^m (W_{ij} \times x_i) + \theta_j \tag{1}$$

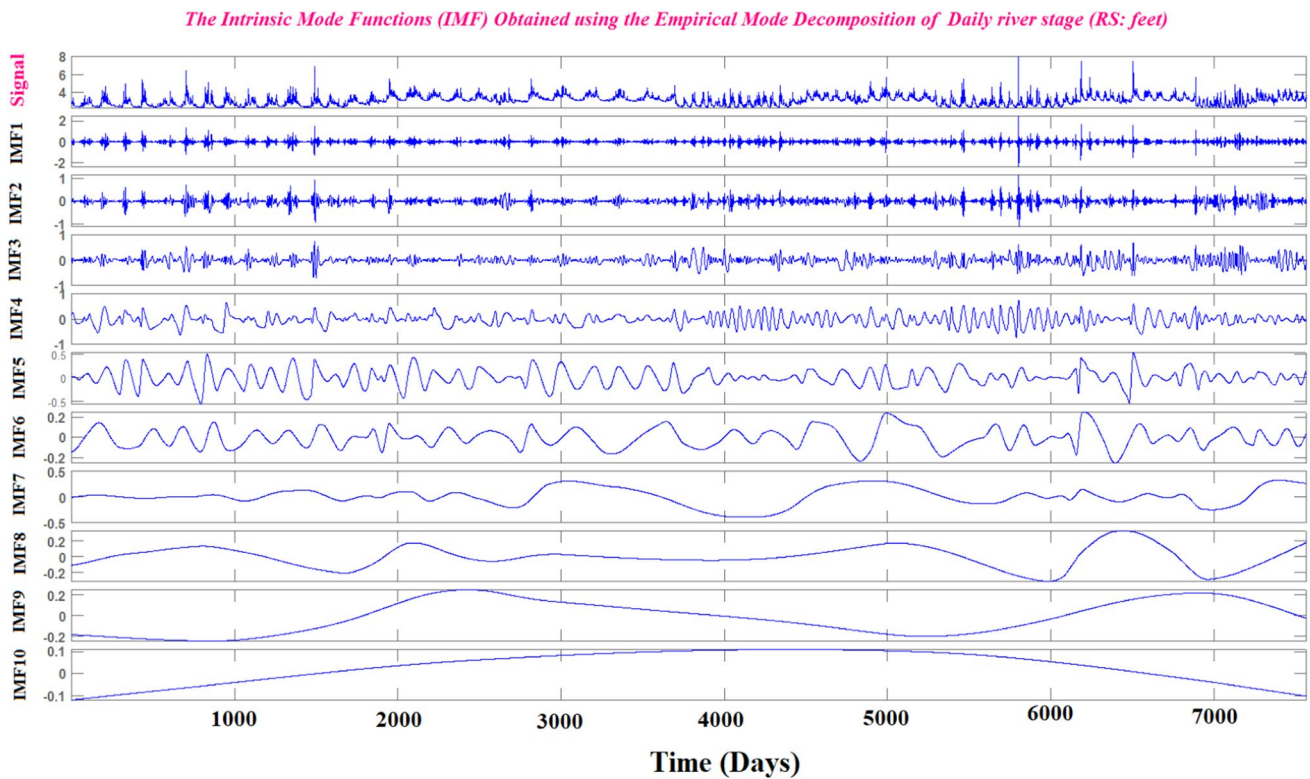


Fig. 3 Intrinsic mode functions (IMF) components of daily river stage (RS) dataset decomposed by the empirical mode decomposition (EMD) algorithm

δ_j is the weighted sum of the hidden neuron j , θ_j is the bias of the hidden neuron j , W_{ij} is the weight linking the input neuron i to the hidden neuron j , and x_i corresponds to one of the inputs variables. The computed δ_j should be moved through an activation function; in general, the transfer non-linear activation function is the sigmoid:

$$y_j = \frac{1}{1 + e^{-\delta}} \tag{2}$$

The obtained y_j value for each hidden neuron is then transferred to the output neuron as follow:

$$\gamma_k = \sum_{j=1}^m (W_{jk} \times y_j) + \vartheta_1 \tag{3}$$

The term γ corresponds to the activation value of the single output neuron; W_{jk} is the weight linking the hidden neuron j to the output neuron k ($k = 1$), and ϑ is the bias of the output layer. Similar to the hidden neurons, the output neuron uses an activation function for providing the final response, which is the linear activation function. Hornik et al. (1989); Hornik (1991); Simon (1999) all provide additional information on the ANN paradigm.

Random forest regression (RFR)

Random forest regression (RFR), developed by Breiman (2001), is an improved version of the original classification and regression tree (CART), and it is a combination of a set of decision trees (DT) models (Wang et al. 2021; Achite et al. 2023; Kumar et al. 2023), for which each one, i.e., each DT uses only a part of the overall dataset (Sun et al. 2021; Kumar et al. 2023); hence, it takes on board only their subset (Fig. 6). It is important to note that the dependent variables were included in the subset with respect to their equal probability, and the weak tree should be repartee on a different sample subset, while the final response of the RFR model is obtained by majority voting between all single trees (Bhadoria et al. 2021). The overall algorithm of the RFR can be summarized as follows: (i) the overall training dataset is randomly divided into K subsets with replacement using the bootstrap sample method (Xue et al. 2021); (ii) one CART is used for each subset; (iii) approximately two-thirds of the dataset is used for growing each tree and the remaining one-third for calculating the out-of-bag (OBB) error, thus each tree grows as

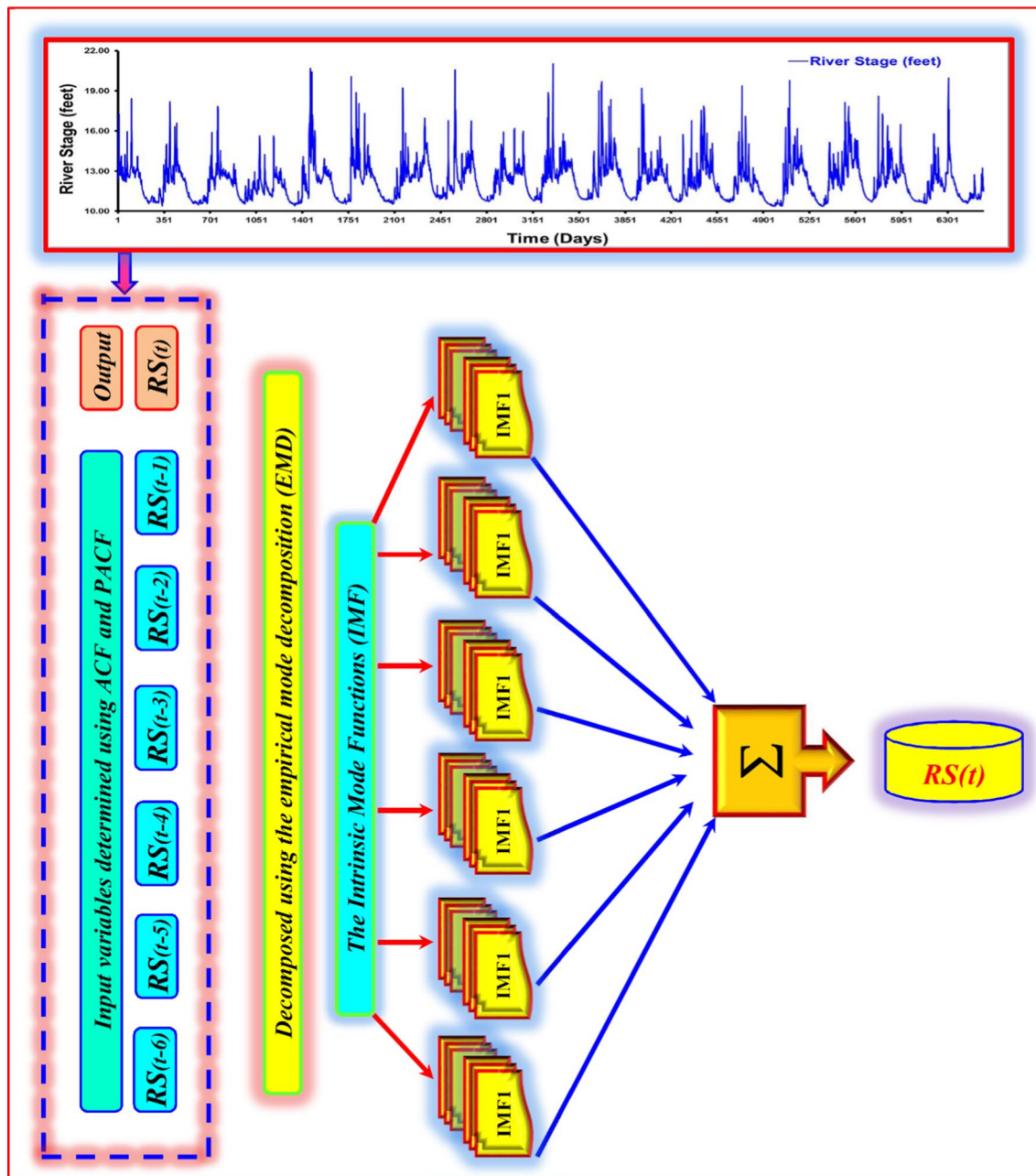


Fig. 4 Flowchart of the EMD & RFR & ANN & AdaBoost & Bag models

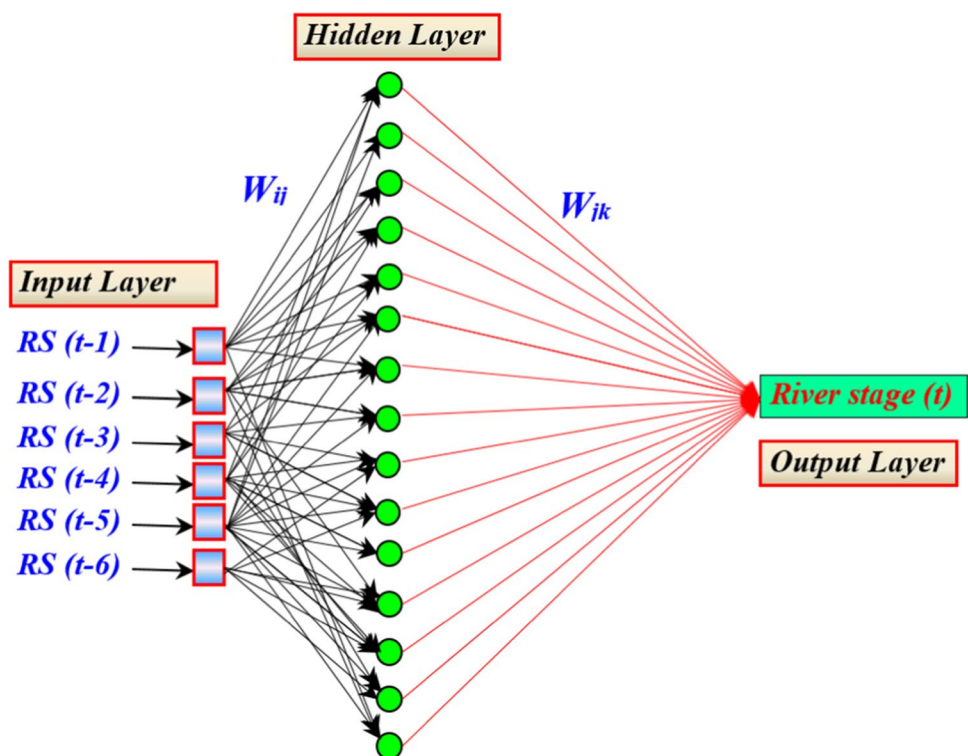
far as it becomes unable to continue the pruning process; (iv) whether used for classification or regression tasks, the final calculated output was provided by aggregation (i.e., regression) or majority voting (i.e., classification) (Xue et al. 2021; Lin et al. 2021).

Bootstrap aggregating (Bagging)

The Bagging algorithm, which is an abbreviation of Bootstrap aggregating, was developed by Breiman (1996).

Bagging is an ensemble algorithm based on the idea of majority voting, and it was proposed for improving the performances of weak classifiers through a bootstrap aggregating mechanism (Pham et al. 2017). The Bagging algorithm uses a series of parallel subsets, also called instances, from the original dataset (Fig. 7a), and it allocates a single training algorithm for each subset (Dou et al. 2020). While it was found that the size of each subset is nearly equal to the size of the original overall dataset, the random sampling with replacement helps in avoiding *falling into the trap of*

Fig. 5 Architecture of the ANN model



“duplicates and/or omissions” compared to the original dataset (Hsiao et al. 2020; Gu et al. 2022). Using the same learning algorithm, the final response of the Bagging model should be obtained by aggregating the responses of all subsets with majority voting (Tien Bui et al. 2016). During the last few years, several applications of the Bagging algorithm can be found in the literature, for example, flood probability mapping (Yariyan et al. 2020), prediction of PM_{2.5} concentration (Qiao et al. 2020), and landslide susceptibility mapping (Hu et al. 2021).

Adaptive boosting algorithm (Adaboost)

Similar to the Bagging algorithm, the Boosting algorithm belongs to the category of ensemble algorithms. However, the most significant difference between the two is that (Fig. 7b), Boosting generates the weak models sequentially, which are dependent on previous prediction results, while Bagging can generate them in parallel (Zounemat-Kermani et al. 2021). Boosting was first proposed by Bartlett et al. (1998), and later, it was described more in-depth in Schapire (2003). From a computational point of view, the Boosting used the idea of weighting the weak learner’s tacking into account for its contribution to the final prediction during the training phase and also indirectly and proportionally to its calculated error (González et al. 2020). Thus, the boosting algorithm uses a variety of weight for each learner and process by updating and optimizing the poorest calculated

errors, hence, the training dataset should be reweighted to gives the poorest learner a new larger weight for the updated training dataset (Kotsiantis 2011).

One of the most well-known ensemble algorithms is the Adaboost (Freund and Schapire 1997), which consists of two distinct components: the forward “step-by-step” algorithm and the “addition” model (Tang et al. 2020). The Adaboost aggregates the outputs of a series of weak learners as follows (i.e., the “addition” step):

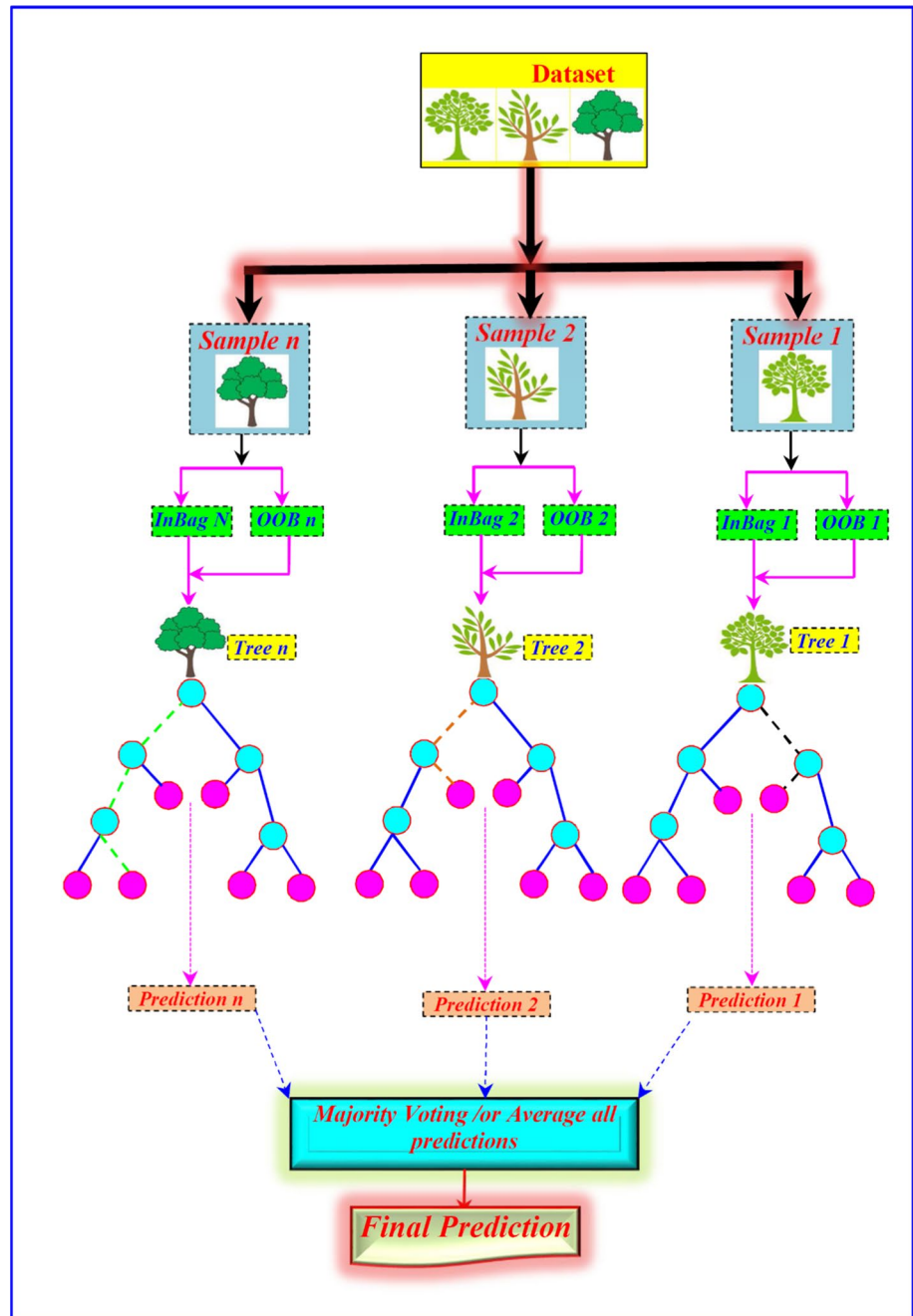
$$Z(x) = \sum_{i=1}^M W_i h_i(x) \tag{4}$$

$Z(x)$ denotes a linear combination of the weak learners; $h_i(x)$ corresponds to one of weak learners; and W_i denotes the weight attributed to the corresponding weak learner. As it is an iterative algorithm, the weight values are updated iteratively at each step, and during the forward “step-by-step”, the weak learner obtained during the previous iteration is used for training the learner for the next step and expressed as follows (Kawakita et al. 2005; Tang et al. 2020):

$$Z(x)_m = Z(x)_{m-1} + W_i h_i(x) \tag{5}$$

The $Z(x)_{m-1}$ represents the linear combination of all weak learners in the previous iteration (Kawakita et al. 2005; Tang et al. 2020). The Adaboost was successfully applied for solving several tasks, among them the prediction of fecal

Fig. 6 Structure of the Random Forest regression (RFR) model. The OOB stands for Out-Of-Bag



coliforms in rivers (EL Bilali et al. 2021), crude oil price prediction (Busari and Lim 2021), and spatial modeling of snow avalanche susceptibility Spatial modeling of snow avalanche susceptibility using hybrid and ensemble machine learning techniques (Akay 2021).

Assessment of the models' performance

Correlation coefficient (R), Nash-Sutcliffe efficiency (NSE), root mean square error (RMSE), and mean absolute error (MAE) were used to assess how well the machines learning

models for daily River stage (RS) forecasting performed. These expressions are provided (Samantaray et al. 2022; Markuna et al. 2023; Saroughi et al. 2023; Vishwakarma et al.2023a, 2023b, 2023c):

$$R = \left[\frac{\frac{1}{N} \sum_{i=1}^N (RS_{obs,i} - \overline{RS_{obs}}) (RS_{est,i} - \overline{RS_{est}})}{\sqrt{\frac{1}{N} \sum_{i=1}^n (RS_{obs,i} - \overline{RS_{obs}})^2} \sqrt{\frac{1}{N} \sum_{i=1}^n (RS_{est,i} - \overline{RS_{est}})^2}} \right], (-1 < R \leq +1) \quad (6)$$

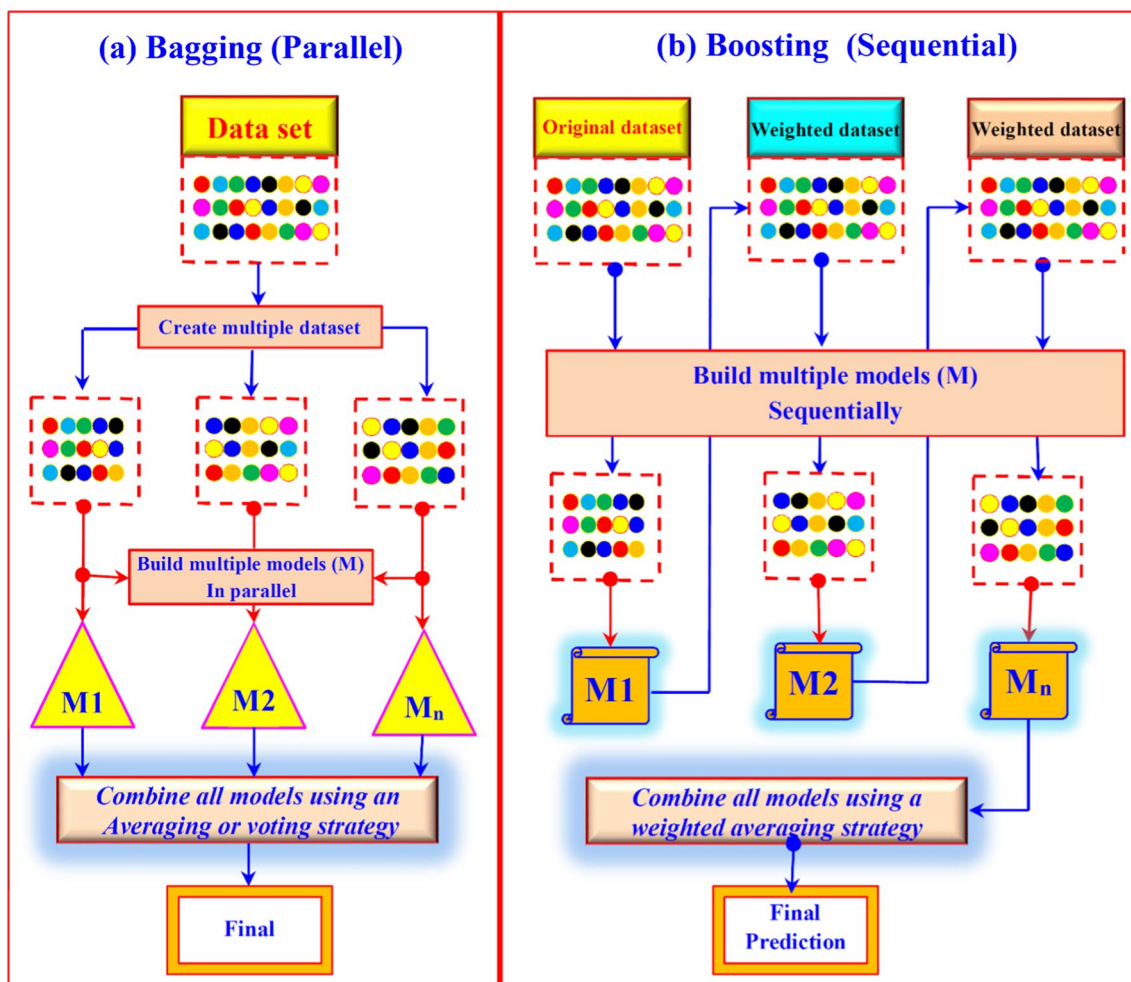


Fig. 7 Bagging and boosting architectures: **a** the Bagging create multiple datasets through random sampling with replacement, and **b** the Boosting create multiple datasets through random sampling with replacement over weighted data (adopted from Yang et al. (2019))

$$NSE = 1 - \frac{\sum_{i=1}^N [RS_{obs} - RS_{est}]^2}{\sum_{i=1}^N [RS_{obs,i} - \overline{RS_{obs}}]^2}, \quad (-\infty < NSE \leq 1) \quad (7)$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N [(RS_{obs,i}) - (RS_{est,i})]^2}, \quad (0 \leq RMSE < +\infty) \quad (8)$$

$$MAE = \frac{1}{N} \sum_{i=1}^N |RS_{obs,i} - RS_{est,i}|, \quad (0 \leq MAE < +\infty) \quad (9)$$

$\overline{RS_{obs}}$ and $\overline{RS_{est}}$ are the mean measured and mean forecasted daily river stage (*RS*), respectively, RS_{obs} and RS_{est} specifies the observed and forecasted daily river stage (*RS*: feet) for *i*th observations, and *N* shows the number of data points.

Results and discussion

Four machines learning, namely the ANN, RFR, AdaBoost, and Bagging, were applied for river stage forecasting. The models were developed without signal preprocessing, i.e., without the EMD algorithm, and in a second stage, the EMD was used for signal decomposition. Based on the ACF and PACF, six input combinations were selected (Table 2), for which the river stage (*RS*) measured at the previous lag was used as an input variable. Only the results from the validation stage are highlighted and in-depth discussed; the obtained results from the two stations are presented and discussed below.

Results at the USGS 14210000

Results obtained at the USGS 14210000 station are reported in Table 3. Scatterplots of forecasted and

Table 2 Input combinations for different models

RFR	ANN	Bagging	AdaBoost	Input combination	Output
RFR1	ANN1	Bagg1	AdaBoost1	RS ($t - 1$)	RS (t)
RFR2	ANN2	Bagg2	AdaBoost2	RS ($t - 2$), RS ($t - 1$)	RS (t)
RFR3	ANN3	Bagg3	AdaBoost3	RS ($t - 3$), RS ($t - 2$), RS ($t - 1$)	RS (t)
RFR4	ANN4	Bagg4	AdaBoost4	RS ($t - 4$), RS ($t - 3$), RS ($t - 2$), RS ($t - 1$)	RS (t)
RFR5	ANN5	Bagg5	AdaBoost5	RS ($t - 5$), RS ($t - 4$), RS ($t - 3$), RS ($t - 2$), RS ($t - 1$)	RS (t)
RFR6	ANN6	Bagg6	AdaBoost6	RS ($t - 6$), RS ($t - 5$), RS ($t - 4$), RS ($t - 3$), RS ($t - 2$), RS ($t - 1$)	RS (t)

measured daily river stage for the best models are shown in Fig. 8. In addition, comparison between forecasted and measured daily river stage is depicted in Fig. 9. Table 3 shows that the ANN was slightly more accurate than the RFR, Bagging and AdaBoost models when using standalone models without EMD. This finding is supported by the mean values of the four numerical performances. The mean R, NSE, RMSE and MAE values obtained using the ANN models were ≈ 0.938 , ≈ 0.879 , ≈ 0.508 and ≈ 0.219 , respectively. These values were marginally better than those obtained using the RFR, which showed improvement rates of approximately $\approx 0.38\%$, $\approx 0.70\%$, $\approx 2.72\%$, and $\approx 2.52\%$. The RFR models and bagging models performed similarly, with the mean R, NSE, RMSE and MAE values being 0.934, 0.872, 0.522, and 0.224, respectively. In contrast, the AdaBoost model performed the worst, with mean R, NSE, RMSE, and MAE values of ≈ 0.931 , ≈ 0.866 , ≈ 0.533 , and ≈ 0.238 , respectively. Comparing the models while taking the input combination into consideration, it is evident that increasing the number of input variables, or the number of lag times, significantly improves the model's performance. From ANN1 to ANN5, the R, NSE, RMSE, and MAE values were improved by 1.20%, 2.20%, 8.48%, and 10.63%, respectively, and the best performances were achieved using ANN5, which is more accurate than all other models. Interesting improvement rates between RFR1 and RFR4 showed values of $\approx 1.60\%$ and $\approx 3.00\%$ for R and NSE values, and the RMSE and MAE values were noticeably reduced by values of $\approx 10.85\%$ and $\approx 15.20\%$, respectively. In terms of R, NSE, RMSE, and MAE, the Bagg4 outperforms the Bagg1 by 1.50%, 2.90%, 10.179%, and 14.45%, respectively. Finally, the improvement of the AdaBoost models was less than all other models for which the improvement rates of the R, NSE, RMSE, and MAE did not exceed the values of $\approx 0.50\%$, $\approx 1.00\%$, $\approx 4.008\%$, and $\approx 8.974\%$, respectively. In any case, the superiority of one model over another was less sensitive, and none

of the proposed models could perform better by adding more input variables.

The empirical mode decomposition (EMD) algorithm was used in the second stage of the investigation to decompose signals. Each input variable, or each river stage measured at a previous lag time, was then divided into several intrinsic mode functions (IMFs) and provided to the models as new input variables. Table 3 shows that all models have improved their performances when using the EMD, and it is evident that all models' mean R, NSE, RMSE, and MAE have improved significantly. The best forecasting accuracies were achieved using the ANN5_EMD model, with improvement rates of approximately $\approx 2.78\%$, $\approx 5.35\%$, $\approx 26.40\%$, and $\approx 12.43\%$. The R, NSE, RMSE, and MAE were clearly improved when compared to the ANN5 without EMD, with improvements of $\approx 3.30\%$, $\approx 6.50\%$, $\approx 33.47\%$, and $\approx 16.66\%$, respectively. This clearly illustrates the significant role that the EMD played in capturing the high linearity in the river stage dataset, particularly due to its capabilities in reducing the errors metrics, i.e., the RMSE and MAE values. Beyond the ANN models, the improvement in forecasting accuracies of the other models was less sensitive. For example, the mean R, NSE, RMSE, and MAE of the RFR were slightly improved by $\approx 0.71\%$, $\approx 0.98\%$, $\approx 3.79\%$, and $\approx 1.26\%$, using the RFR_EMD, respectively. In addition, using the Bagg_EMD, negligible improvement was obtained compared to the Bagg without EMD. The most significant concluding remark was about the AdaBoost_EMD compared to the AdaBoost, for which the improvement rates in terms of R and NSE did not exceed $\approx 0.2\%$ and $\approx 0.33\%$, and the improvement rates of the RMSE and MAE were below the values of $\approx 1.4\%$ and $\approx 1.12\%$, respectively. Overall, the best forecasting accuracy was obtained using the ANN5_EMD, followed by the RFR5_EMD equally with the Bagg5_EMD, and the AdaBoost5_EMD was the less accurate model. Finally, in

Table 3 Effectiveness of various forecasting models at the USGS 14210000 station

Models	Training				Validation			
	R	NSE	RMSE	MAE	R	NSE	RMSE	MAE
ANN1	0.944	0.892	0.464	0.213	0.929	0.862	0.542	0.235
ANN1_EMD	0.953	0.908	0.428	0.208	0.933	0.871	0.523	0.241
ANN2	0.948	0.900	0.447	0.194	0.938	0.880	0.506	0.215
ANN2_EMD	0.976	0.952	0.309	0.161	0.974	0.948	0.331	0.184
ANN3	0.950	0.902	0.442	0.193	0.940	0.883	0.498	0.216
ANN3_EMD	0.978	0.957	0.292	0.156	0.969	0.938	0.363	0.190
ANN4	0.951	0.904	0.437	0.192	0.938	0.880	0.505	0.218
ANN4_EMD	0.981	0.961	0.277	0.148	0.973	0.947	0.336	0.177
ANN5	0.951	0.905	0.436	0.192	0.941	0.884	0.496	0.210
ANN5_EMD	0.979	0.959	0.287	0.150	0.974	0.949	0.330	0.175
ANN6	0.950	0.903	0.439	0.192	0.939	0.882	0.502	0.217
ANN6_EMD	0.980	0.961	0.279	0.151	0.969	0.939	0.361	0.181
RFR1	0.954	0.910	0.423	0.201	0.923	0.852	0.562	0.250
RFR1_EMD	0.979	0.955	0.300	0.139	0.931	0.863	0.541	0.237
RFR2	0.969	0.938	0.352	0.151	0.936	0.875	0.515	0.218
RFR2_EMD	0.983	0.963	0.270	0.122	0.937	0.875	0.517	0.226
RFR3	0.970	0.939	0.347	0.149	0.935	0.873	0.519	0.222
RFR3_EMD	0.985	0.967	0.256	0.116	0.942	0.884	0.497	0.219
RFR4	0.975	0.950	0.316	0.131	0.939	0.882	0.501	0.212
RFR4_EMD	0.986	0.969	0.249	0.112	0.944	0.887	0.490	0.217
RFR5	0.975	0.949	0.318	0.132	0.936	0.876	0.513	0.218
RFR5_EMD	0.987	0.970	0.243	0.110	0.946	0.890	0.484	0.214
RFR6	0.974	0.947	0.324	0.135	0.933	0.871	0.524	0.225
RFR6_EMD	0.987	0.971	0.240	0.109	0.945	0.889	0.486	0.215
Bagg1	0.954	0.910	0.423	0.201	0.924	0.852	0.560	0.249
Bagg1_EMD	0.979	0.956	0.297	0.138	0.931	0.863	0.539	0.237
Bagg2	0.968	0.938	0.353	0.151	0.936	0.875	0.515	0.218
Bagg2_EMD	0.983	0.963	0.270	0.123	0.939	0.877	0.512	0.220
Bagg3	0.970	0.940	0.347	0.149	0.934	0.873	0.520	0.222
Bagg3_EMD	0.985	0.967	0.257	0.117	0.943	0.884	0.497	0.217
Bagg4	0.975	0.950	0.317	0.131	0.939	0.881	0.503	0.213
Bagg4_EMD	0.986	0.969	0.249	0.113	0.944	0.887	0.490	0.216
Bagg5	0.974	0.949	0.320	0.133	0.936	0.876	0.515	0.219
Bagg5_EMD	0.987	0.970	0.242	0.110	0.946	0.890	0.484	0.214
Bagg6	0.974	0.947	0.324	0.135	0.933	0.870	0.525	0.225
Bagg6_EMD	0.987	0.971	0.240	0.108	0.945	0.890	0.485	0.215
AdaBoost1	0.950	0.903	0.439	0.210	0.925	0.854	0.557	0.245
AdaBoost1_EMD	0.970	0.940	0.345	0.204	0.923	0.850	0.564	0.289
AdaBoost2	0.955	0.912	0.418	0.198	0.932	0.868	0.531	0.234
AdaBoost2_EMD	0.980	0.960	0.281	0.179	0.939	0.882	0.501	0.220
AdaBoost3	0.956	0.915	0.412	0.195	0.934	0.871	0.524	0.234
AdaBoost3_EMD	0.981	0.963	0.271	0.175	0.931	0.864	0.537	0.229
AdaBoost4	0.957	0.916	0.409	0.194	0.933	0.869	0.527	0.235
AdaBoost4_EMD	0.983	0.966	0.261	0.172	0.935	0.872	0.521	0.226
AdaBoost5	0.958	0.917	0.407	0.195	0.932	0.868	0.529	0.240
AdaBoost5_EMD	0.984	0.968	0.252	0.164	0.940	0.883	0.498	0.219
AdaBoost6	0.958	0.918	0.405	0.194	0.932	0.868	0.529	0.239
AdaBoost6_EMD	0.984	0.968	0.251	0.163	0.932	0.867	0.532	0.218

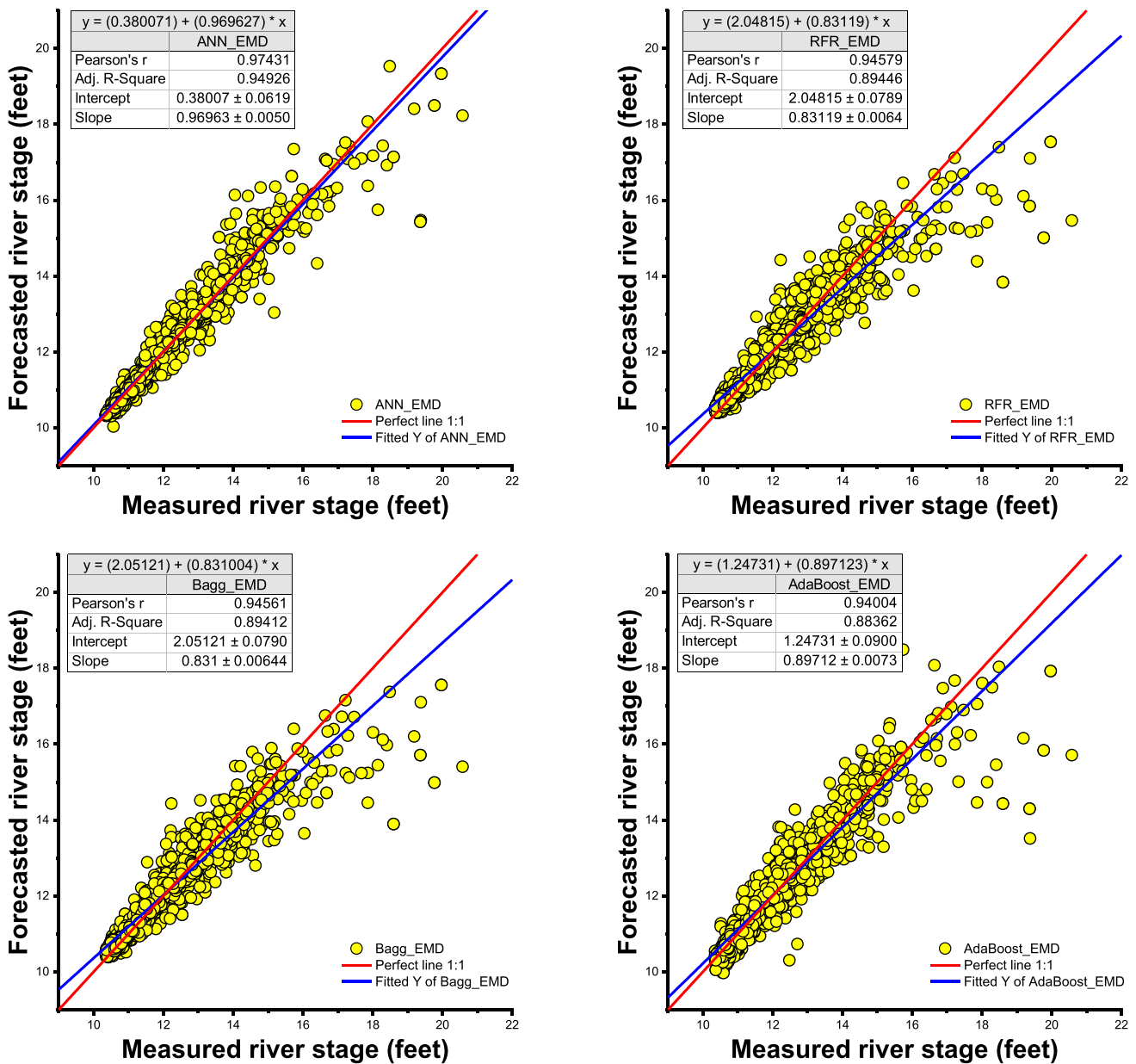


Fig. 8 Scatterplots of measured against forecasted daily river stage (RS) for the validation stage at the USGS 14210000

Fig. 10, we summarized the best-obtained results in terms of Boxplot, Violin plot, Radar plot, and Taylor diagram.

Results at the USGS 14211499

Table 4 presents the findings from the USGS 14211499. Scatterplots of forecasted and measured daily river stage for the best models are shown in Fig. 11. In addition,

comparison between forecasted and measured daily river stage is depicted in Fig. 12. Table 4 shows that the performance of the ANN models was marginally inferior to that of the RFR and Bagg models, with mean R, NSE, RMSE, and MAE values of approximately ≈ 0.932 , ≈ 0.869 , ≈ 0.182 , and ≈ 0.084 , respectively. With mean R, NSE, RMSE, and MAE values of approximately ≈ 0.935 , ≈ 0.874 , ≈ 0.179 , and ≈ 0.084 , respectively, the RFR and

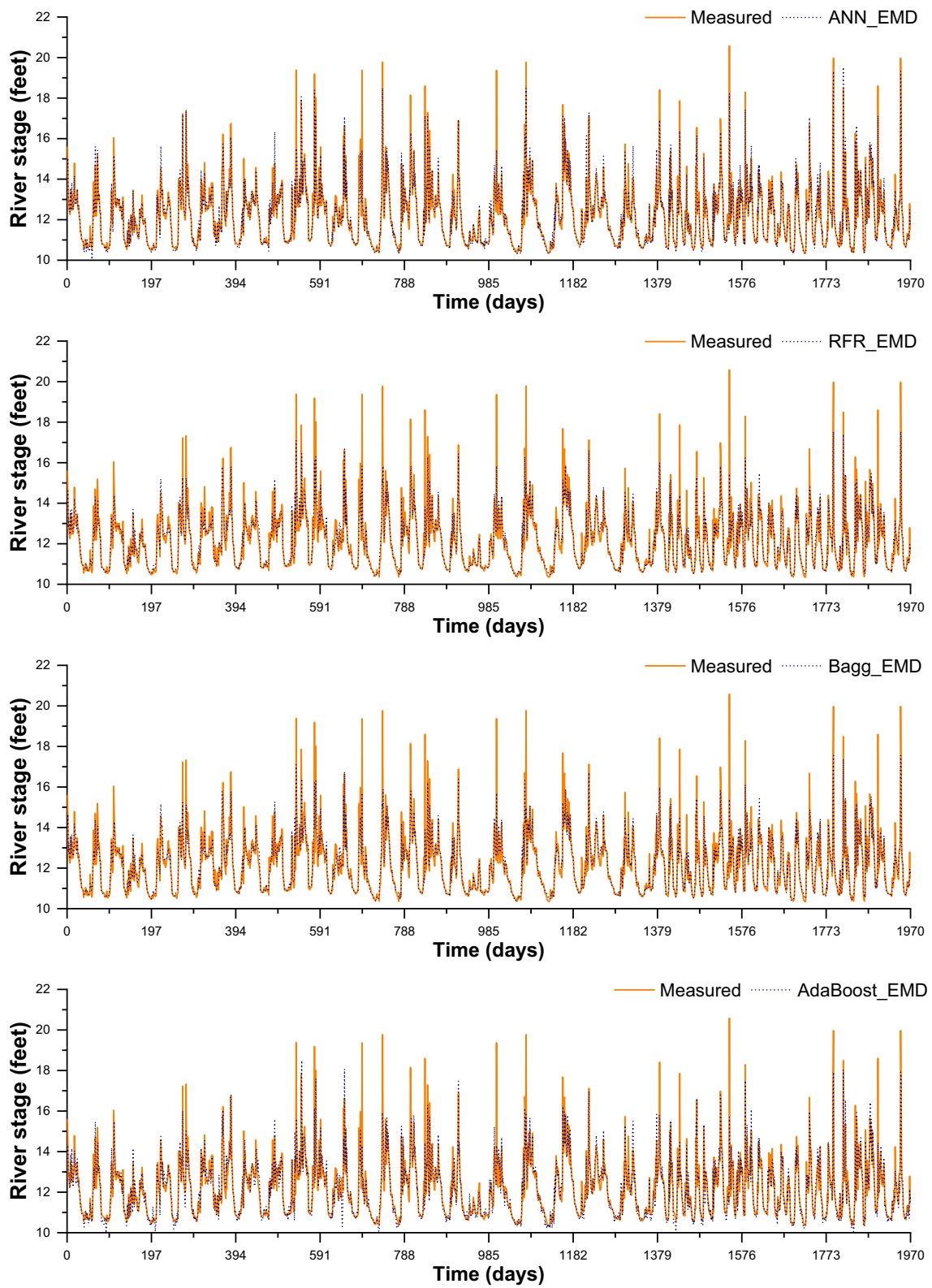


Fig. 9 A comparison of the measured and forecasted daily river stage (*RS*) for the validation stage at the USGS 14210000

Fig. 10 Examples of graphs showing models performance for the best developed algorithms during the validation stage at the USGS 14210000: **a** Boxplot, **b** Violin plot, **c** Radar plot, and **d** Taylor diagram

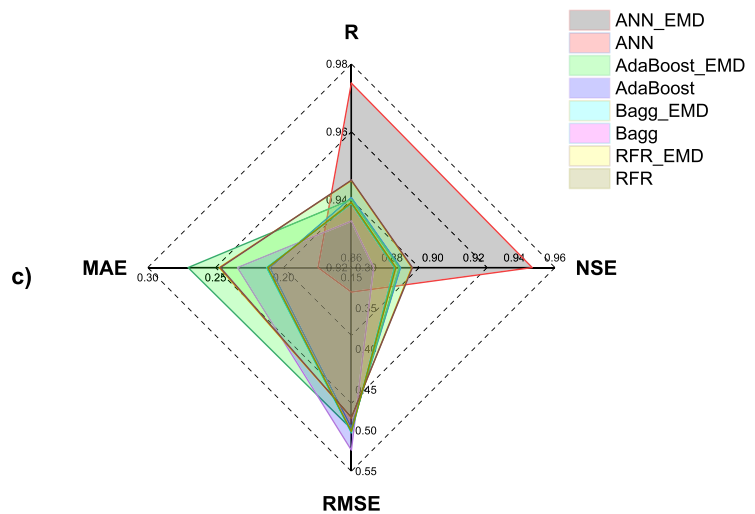
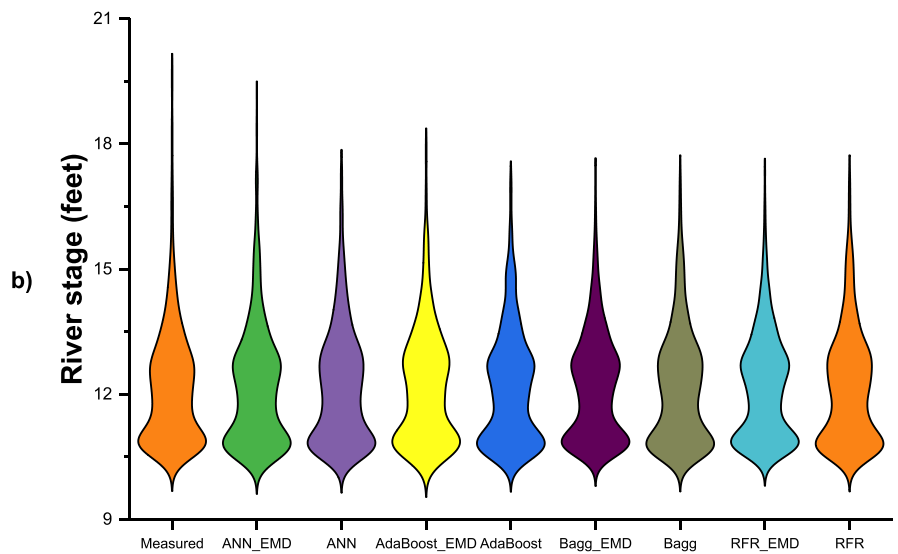
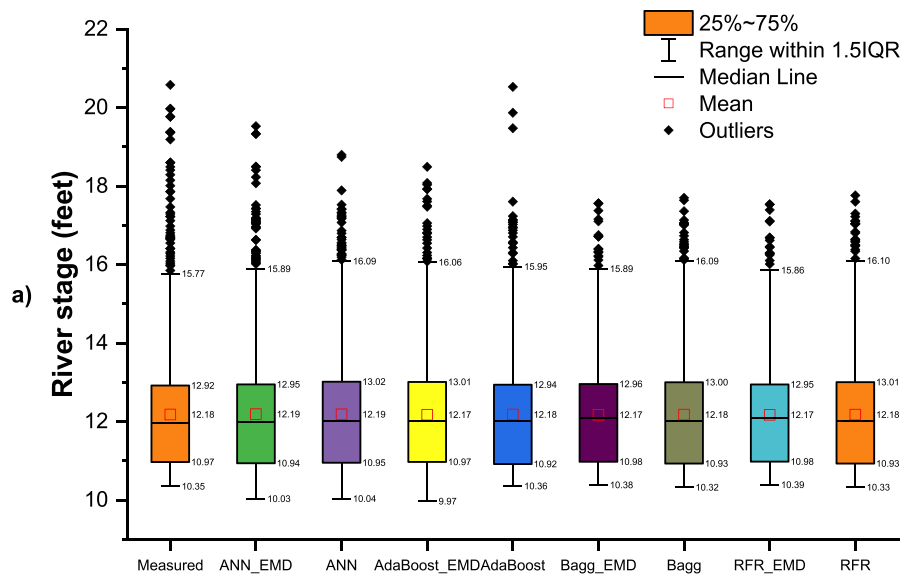
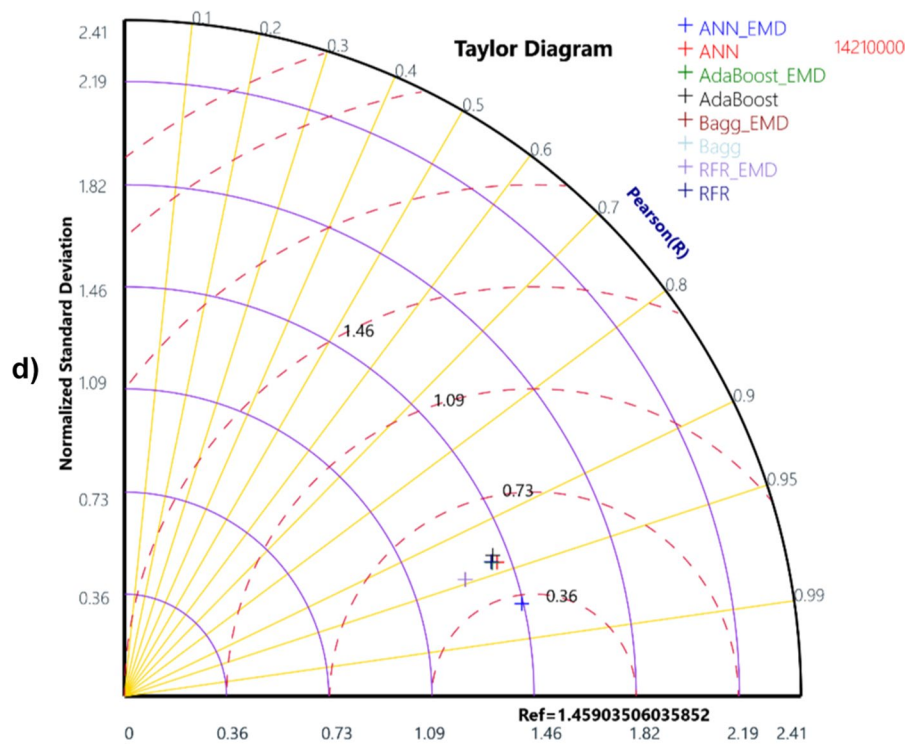


Fig. 10 (continued)



Bagg models performed equally well. The AdaBoost models were the less accurate models with mean R, NSE, RMSE, and MAE values of approximately ≈ 0.914 , ≈ 0.832 , ≈ 0.207 , and ≈ 0.093 , respectively. Only the AdaBoost models had numerical performances that were nearly identical, with barely detectable differences. Adding more input variables from one to six do not always result in a decline in the error metrics, i.e., RMSE and MAE, or an increase in the R and NSE values. It is evident that, for the RFR and Bagg models, the sixth input combination, for which the previous six lag times of the river stage were included, produced the best accuracy results (see Table 2). Finally, when using the ANN models, the ANN4 model had the highest R (≈ 0.935) and NSE (≈ 0.874) values, the lowest RMSE (≈ 0.179) and MAE (≈ 0.083) values, and it was clear that the model's performances declined after the fourth input combination. Using the EMD for improving forecasting accuracies, the ANN4_EMD was the best model, showing high R and NSE values of ≈ 0.935 and ≈ 0.913 and the lowest RMSE (≈ 0.149) and MAE (≈ 0.066). The ANN4_EMD improves the performances of the RFR6_EMD by $\approx 1.40\%$, $\approx 3.00\%$, $\approx 13.372\%$, and $\approx 31.25\%$, respectively, in terms of R, NSE, RMSE, and MAE. The Bagg_EMD models were relatively equal to the RFR_EMD with negligible differences, and the AdaBoost_EMD models were the poorest models, whatever the input combination. In Fig. 13, we

summarized the best-obtained results in terms of Boxplot, Violin plot, Radar plot, and Taylor diagram.

Conclusion

An efficient, hybrid machine learning algorithm is proposed in this study to forecast river stage (RS) with high precision and accuracy. In the algorithm, the impact of preprocessing signal decomposition using empirical mode decomposition (EMD) on RS is investigated. Moreover, as our modeling framework is based on linking the RS at time (t) to the RS at previous lag time, the ACF and the PACF were considered for selecting the relevant input variables, and it was found that the RS measured at $(t - 1)$ to $(t - 6)$ were the most significant input variables. In the second stage of our study, the selected six lag times were decomposed using the empirical mode decomposition (EMD), and the obtained intrinsic mode functions (IMF) were used as new input variables. Consequently, the in situ measured RS was estimated by the new hybrid models, and a comparison among the single models without the EMD was also done. The estimated daily RS was validated against in situ data collected at two USGS

Table 4 Effectiveness of various forecasting models at the USGS 14211499 station

Models	Training				Validation			
	R	NSE	RMSE	MAE	R	NSE	RMSE	MAE
ANN1	0.944	0.891	0.163	0.083	0.933	0.870	0.182	0.085
ANN1_EMD	0.954	0.910	0.148	0.080	0.938	0.879	0.175	0.084
ANN2	0.946	0.895	0.159	0.082	0.931	0.867	0.184	0.084
ANN2_EMD	0.964	0.930	0.130	0.072	0.948	0.899	0.160	0.079
ANN3	0.948	0.898	0.157	0.082	0.932	0.869	0.183	0.084
ANN3_EMD	0.968	0.935	0.125	0.063	0.947	0.895	0.164	0.071
ANN4	0.949	0.900	0.156	0.081	0.935	0.874	0.179	0.083
ANN4_EMD	0.968	0.936	0.125	0.068	0.955	0.913	0.149	0.066
ANN5	0.949	0.900	0.156	0.081	0.933	0.871	0.181	0.084
ANN5_EMD	0.970	0.942	0.119	0.066	0.955	0.912	0.150	0.077
ANN6	0.948	0.899	0.156	0.082	0.930	0.865	0.185	0.086
ANN6_EMD	0.968	0.936	0.125	0.072	0.952	0.905	0.156	0.080
RFR1	0.946	0.896	0.159	0.083	0.927	0.859	0.190	0.090
RFR1_EMD	0.975	0.949	0.111	0.051	0.917	0.835	0.205	0.120
RFR2	0.962	0.926	0.134	0.068	0.933	0.871	0.181	0.085
RFR2_EMD	0.978	0.956	0.103	0.046	0.935	0.870	0.182	0.104
RFR3	0.966	0.934	0.127	0.064	0.936	0.875	0.178	0.084
RFR3_EMD	0.980	0.960	0.098	0.043	0.940	0.879	0.175	0.099
RFR4	0.971	0.943	0.117	0.057	0.936	0.877	0.177	0.083
RFR4_EMD	0.981	0.962	0.096	0.042	0.940	0.881	0.174	0.096
RFR5	0.972	0.944	0.116	0.056	0.939	0.881	0.174	0.082
RFR5_EMD	0.981	0.962	0.096	0.041	0.939	0.881	0.174	0.096
RFR6	0.972	0.944	0.116	0.056	0.939	0.883	0.173	0.082
RFR6_EMD	0.981	0.962	0.095	0.041	0.941	0.883	0.172	0.096
Bagg1	0.947	0.896	0.159	0.083	0.927	0.859	0.190	0.090
Bagg1_EMD	0.975	0.949	0.111	0.051	0.920	0.841	0.201	0.118
Bagg2	0.962	0.926	0.134	0.068	0.933	0.871	0.181	0.086
Bagg2_EMD	0.978	0.957	0.103	0.046	0.936	0.872	0.181	0.103
Bagg3	0.966	0.933	0.127	0.064	0.936	0.876	0.178	0.084
Bagg3_EMD	0.980	0.960	0.099	0.043	0.940	0.881	0.174	0.098
Bagg4	0.971	0.943	0.118	0.057	0.937	0.878	0.177	0.083
Bagg4_EMD	0.981	0.961	0.097	0.042	0.939	0.879	0.175	0.098
Bagg5	0.972	0.944	0.117	0.056	0.939	0.881	0.174	0.082
Bagg5_EMD	0.981	0.962	0.095	0.041	0.940	0.881	0.174	0.096
Bagg6	0.972	0.944	0.116	0.056	0.940	0.883	0.173	0.082
Bagg6_EMD	0.982	0.963	0.095	0.041	0.940	0.882	0.173	0.096
AdaBoost1	0.942	0.887	0.166	0.084	0.914	0.834	0.206	0.091
AdaBoost1_EMD	0.954	0.908	0.149	0.076	0.920	0.845	0.199	0.107
AdaBoost2	0.943	0.889	0.164	0.083	0.913	0.831	0.207	0.092
AdaBoost2_EMD	0.964	0.928	0.132	0.067	0.918	0.836	0.204	0.116
AdaBoost3	0.943	0.890	0.164	0.083	0.913	0.829	0.208	0.093
AdaBoost3_EMD	0.965	0.930	0.130	0.065	0.921	0.845	0.198	0.116
AdaBoost4	0.945	0.892	0.162	0.083	0.914	0.832	0.207	0.093
AdaBoost4_EMD	0.965	0.931	0.130	0.064	0.912	0.830	0.208	0.122
AdaBoost5	0.945	0.892	0.162	0.083	0.914	0.832	0.207	0.094
AdaBoost5_EMD	0.966	0.933	0.128	0.062	0.922	0.849	0.196	0.112
AdaBoost6	0.945	0.893	0.161	0.083	0.914	0.833	0.206	0.094
AdaBoost6_EMD	0.967	0.933	0.128	0.062	0.910	0.821	0.214	0.120

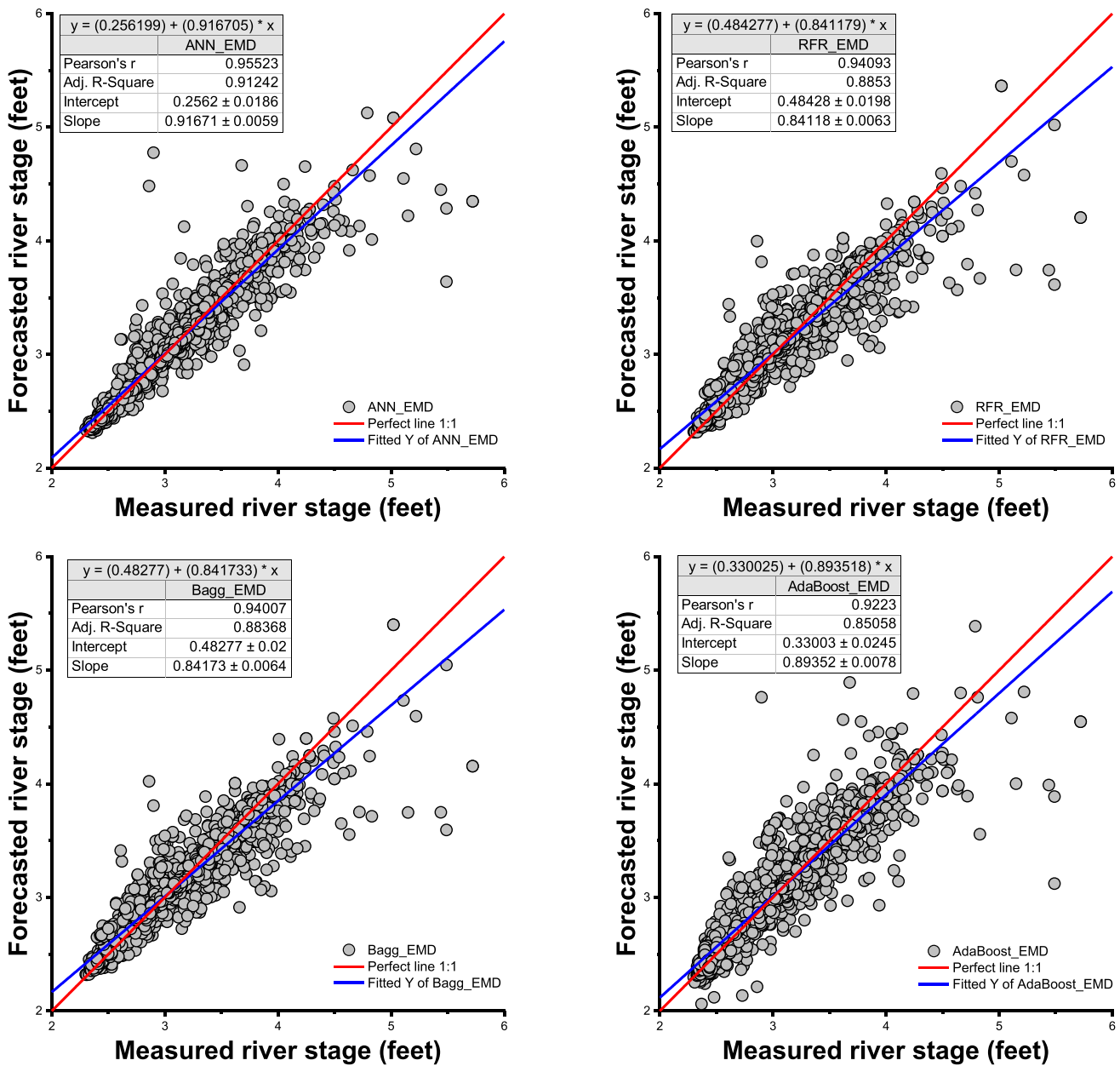


Fig. 11 Scatterplots of measured against forecasted daily river stage (RS) for the validation stage at the USGS 14211499

stations. The obtained results indicated that the newly proposed algorithm for RS retrieval could effectively retrieve RS with high accuracy and precision. The following conclusions are drawn:

- Using single models, i.e., ANN, RFR, Bagging, and AdaBoost without decomposition, i.e., without EMD, the four models were relatively equal with slightly supe-

riority in favor to the ANN model, and it was shown that beyond the third input combination, i.e., using only $Q(t - 1)$, $Q(t - 2)$, and $Q(t - 3)$ as input variables, improvement in models performances was negligible and marginal.

- Numerical results revealed that, for all models, the R, NSE, RMSE, and MAE ranged from 0.923 to 0.940, 0.852 to 0.883, 0.498 to 0.562, 0.216 to 0.250 for the

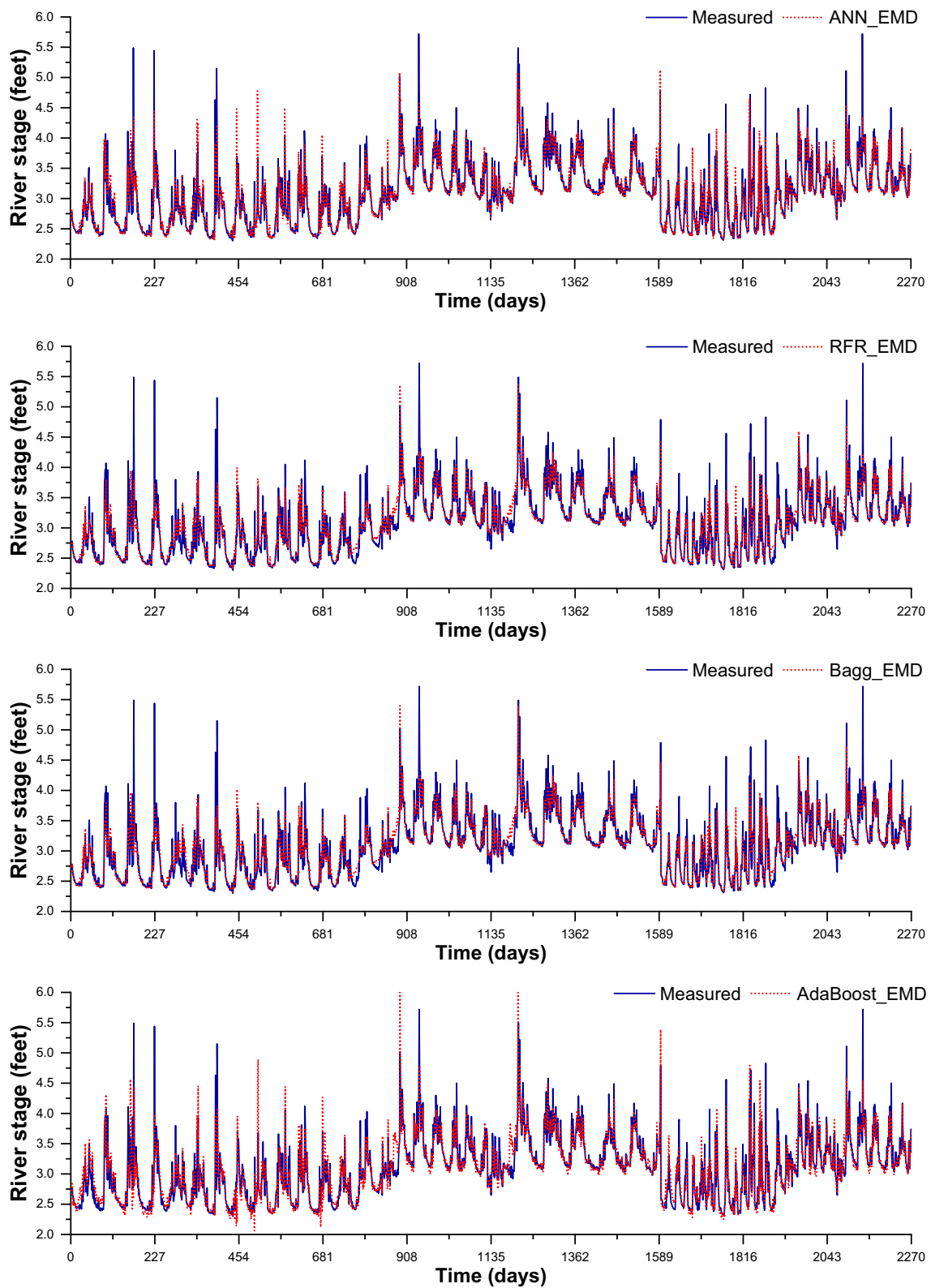


Fig. 12 Comparison between measured against forecasted daily river stage (*RS*) for the validation stage at the USGS 14211499

Fig. 13 Examples of graphs showing models performance for the best developed algorithms during the validation stage at the USGS 14211499: **a** Boxplot, **b** Violin plot, **c** Radar plot, and **d** Taylor diagram

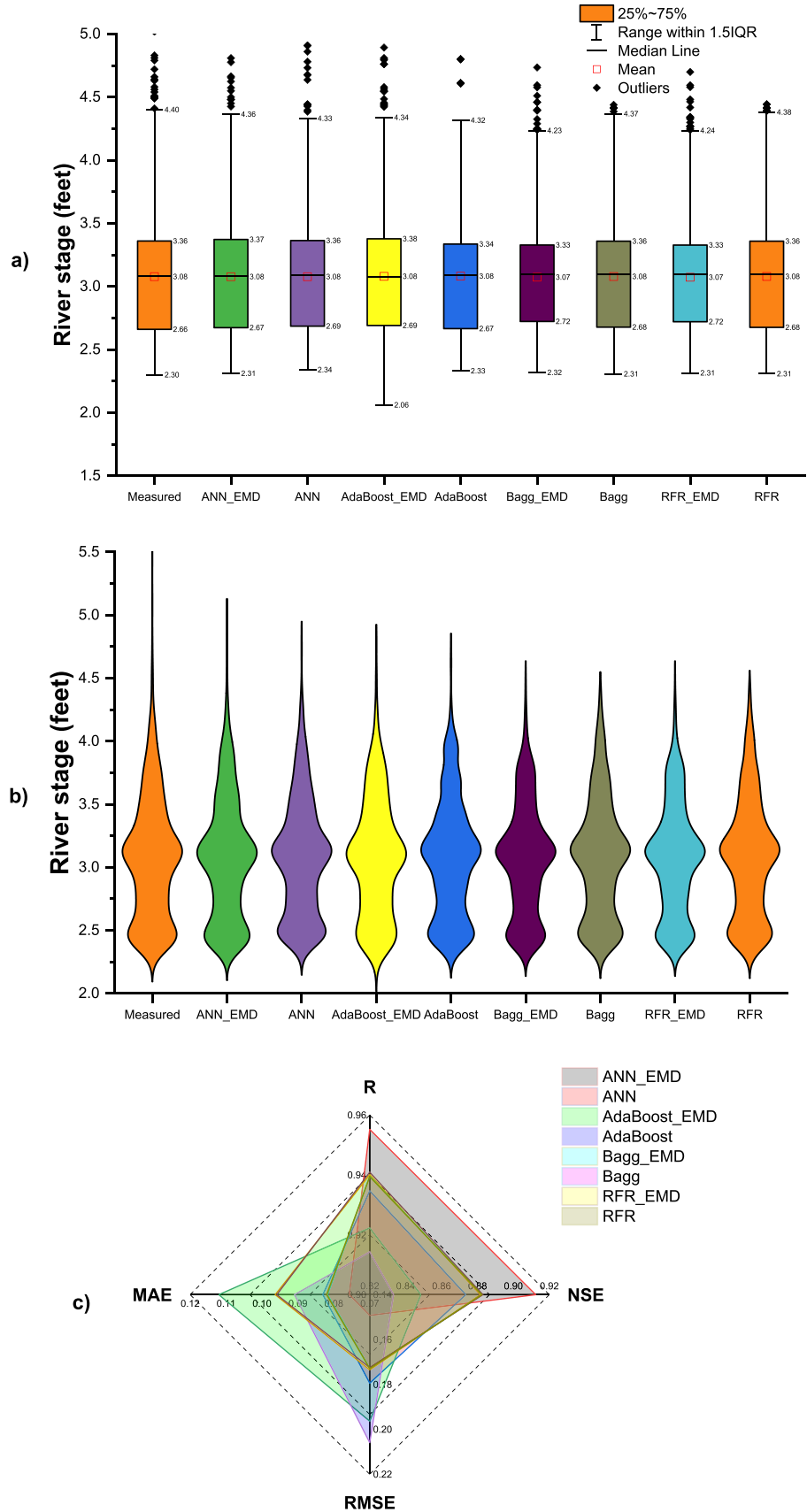
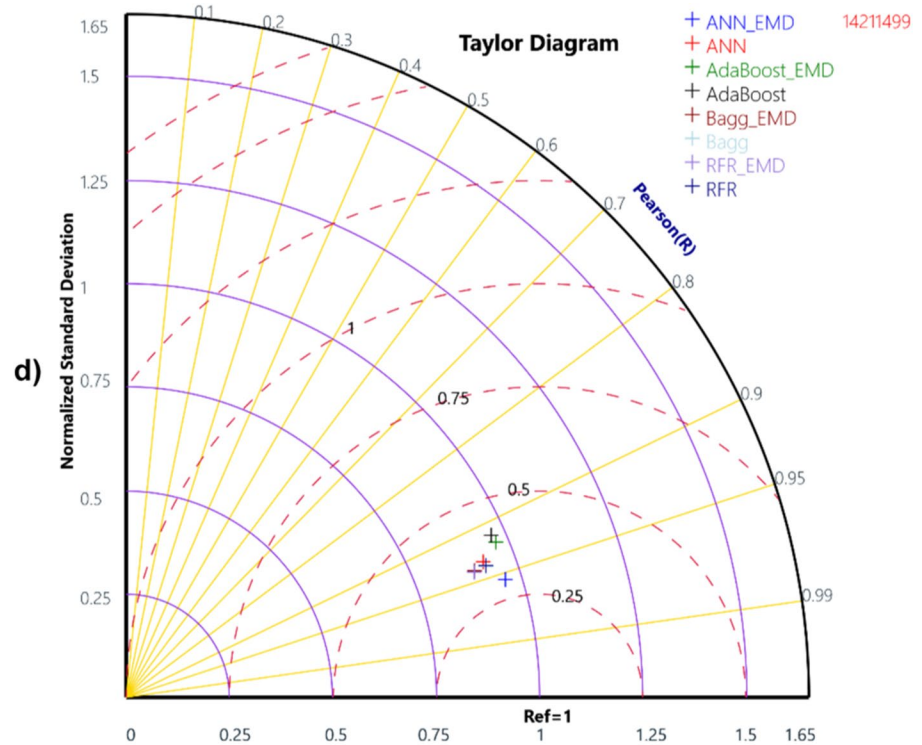


Fig. 13 (continued)



USGS 14210000 station, and from 0.913 to 0.940, 0.831 to 0.883, 0.173 to 0.208, and 0.082 to 0.094 for the USGS 14211499 station, respectively.

- Using the EMD as a preprocessing signal decomposition contributed to a high and significant improvement in models performances, for which we have concluded that; increasing the number of lag times from one to six led to an increase in all numerical values for all criteria, showing the R and NSE values reaching the maximal values of 0.974 and 0.949 at the USGS 14210000, and the maximal values of 0.955 and 0.913 at the USGS 14211499, respectively, and the $RMSE$ and MAE values were drastically decreased to their lowest values of 0.330 and 0.175 at the USGS 14210000, and to the values of 0.149 and 0.076 at the USGS 14211499, respectively.
- In fact, the obtained results in the present study appeared very consistent and very encouraging regardless of the considered period of record. However, extending the series of records and the forecasting horizon beyond the time (t), i.e., $t + 1$, $t + 2$, is more suitable for better conclusions.

Acknowledgements The authors extend their appreciation to the Deanship of Scientific Research, King Saud University for funding through the Vice Deanship of Scientific Research Chairs, Research Chair of Prince Sultan Bin Abdulaziz International Prize for Water.

Funding Open access funding provided by Lulea University of Technology. The authors thank the Deanship of Scientific Research, King Saud University, for funding through the Vice Deanship of Scientific Research Chairs, Research Chair of Prince Sultan Bin Abdulaziz International Prize for Water.

Declarations

Conflict of interest No conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Achite M, Elshaboury N, Jehanzaib M et al (2023) Performance of machine learning techniques for meteorological drought forecasting in the Wadi Mina basin. *Algeria Water* 15:765. <https://doi.org/10.3390/w15040765>
- Akay H (2021) Spatial modeling of snow avalanche susceptibility using hybrid and ensemble machine learning techniques. *CATENA* 206:105524. <https://doi.org/10.1016/j.catena.2021.105524>

- Alvisi S, Franchini M (2012) Grey neural networks for river stage forecasting with uncertainty. *Phys Chem Earth, Parts a/b/c* 42–44:108–118. <https://doi.org/10.1016/j.pce.2011.04.002>
- Bartlett P, Freund Y, Lee WS, Schapire RE (1998) Boosting the margin: a new explanation for the effectiveness of voting methods. *Ann Stat* 26:1651–1686. <https://doi.org/10.1214/aos/1024691352>
- Bhadoria RS, Pandey MK, Kundu P (2021) RVFR: Random vector forest regression model for integrated & enhanced approach in forest fires predictions. *Ecol Inform* 66:101471. <https://doi.org/10.1016/j.ecoinf.2021.101471>
- Breiman L (1996) Bagging predictors. *Mach Learn* 24:123–140. <https://doi.org/10.1007/BF00058655>
- Breiman L (2001) Random Forests. *Mach Learn* 45:5–32. <https://doi.org/10.1023/A:1010933404324>
- Busari GA, Lim DH (2021) Crude oil price prediction: a comparison between AdaBoost-LSTM and AdaBoost-GRU for improving forecasting performance. *Comput Chem Eng* 155:107513. <https://doi.org/10.1016/j.compchemeng.2021.107513>
- Chau KW (2006) Particle swarm optimization training algorithm for ANNs in stage prediction of Shing Mun River. *J Hydrol* 329:363–367. <https://doi.org/10.1016/j.jhydrol.2006.02.025>
- Chau KW (2007) A split-step particle swarm optimization algorithm in river stage forecasting. *J Hydrol* 346:131–135. <https://doi.org/10.1016/j.jhydrol.2007.09.004>
- Dou J, Yunus AP, Bui DT et al (2020) Improved landslide assessment using support vector machine with bagging, boosting, and stacking ensemble machine learning framework in a mountainous watershed, Japan. *Landslides* 17:641–658. <https://doi.org/10.1007/s10346-019-01286-5>
- 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. *Stoch Environ Res Risk Assess*. <https://doi.org/10.1007/s00477-022-02196-0>
- El-Bilali A, Taleb A, Bahlaoui MA, Brouziyne Y (2021) An integrated approach based on Gaussian noises-based data augmentation method and AdaBoost model to predict faecal coliforms in rivers with small dataset. *J Hydrol* 599:126510. <https://doi.org/10.1016/j.jhydrol.2021.126510>
- Freund Y, Schapire RE (1997) A decision-theoretic generalization of on-line learning and an application to boosting. *J Comput Syst Sci* 55:119–139. <https://doi.org/10.1006/jcss.1997.1504>
- Fu J-C, Huang H-Y, Jang J-H, Huang P-H (2019) River stage forecasting using multiple additive regression trees. *Water Resour Manag* 33:4491–4507. <https://doi.org/10.1007/s11269-019-02357-x>
- González S, García S, Del Ser J et al (2020) A practical tutorial on bagging and boosting based ensembles for machine learning: Algorithms, software tools, performance study, practical perspectives and opportunities. *Inf Fusion* 64:205–237. <https://doi.org/10.1016/j.inffus.2020.07.007>
- Gu Q, Zhang X, Chen L, Xiong N (2022) An improved bagging ensemble surrogate-assisted evolutionary algorithm for expensive many-objective optimization. *Appl Intell* 52:5949–5965. <https://doi.org/10.1007/s10489-021-02709-4>
- Hornik K (1991) Approximation capabilities of multilayer feedforward networks. *Neural Netw* 4:251–257. [https://doi.org/10.1016/0893-6080\(91\)90009-T](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](https://doi.org/10.1016/0893-6080(89)90020-8)
- Hsiao Y-H, Su C-T, Fu P-C (2020) Integrating MTS with bagging strategy for class imbalance problems. *Int J Mach Learn Cybern* 11:1217–1230. <https://doi.org/10.1007/s13042-019-01033-1>
- Hsu M-H, Lin S-H, Fu J-C et al (2010) Longitudinal stage profiles forecasting in rivers for flash floods. *J Hydrol* 388:426–437. <https://doi.org/10.1016/j.jhydrol.2010.05.028>
- Hu X, Huang C, Mei H, Zhang H (2021) Landslide susceptibility mapping using an ensemble model of Bagging scheme and random subspace-based naïve Bayes tree in Zigui County of the Three Gorges Reservoir Area, China. *Bull Eng Geol Environ* 80:5315–5329. <https://doi.org/10.1007/s10064-021-02275-6>
- Kawakita M, Minami M, Eguchi S, Lennert-Cody CE (2005) An introduction to the predictive technique AdaBoost with a comparison to generalized additive models. *Fish Res* 76:328–343. <https://doi.org/10.1016/j.fishres.2005.07.011>
- Khatibi R, Sivakumar B, Ghorbani MA et al (2012) Investigating chaos in river stage and discharge time series. *J Hydrol* 414–415:108–117. <https://doi.org/10.1016/j.jhydrol.2011.10.026>
- Kisi O (2011) Wavelet regression model as an alternative to neural networks for river stage forecasting. *Water Resour Manag* 25:579–600. <https://doi.org/10.1007/s11269-010-9715-8>
- Kotsiantis S (2011) Combining bagging, boosting, rotation forest and random subspace methods. *Artif Intell Rev* 35:223–240. <https://doi.org/10.1007/s10462-010-9192-8>
- Kumar D, Singh VK, Abed SA et al (2023) Multi-ahead electrical conductivity forecasting of surface water based on machine learning algorithms. *Appl Water Sci* 13:192. <https://doi.org/10.1007/s13201-023-02005-1>
- Lin J, Lu S, He X, Wang F (2021) Analyzing the impact of three-dimensional building structure on CO2 emissions based on random forest regression. *Energy* 236:121502. <https://doi.org/10.1016/j.energy.2021.121502>
- Liu Y, Wang H, Lei X, Wang H (2021) Real-time forecasting of river water level in urban based on radar rainfall: a case study in Fuzhou City. *J Hydrol* 603:126820. <https://doi.org/10.1016/j.jhydrol.2021.126820>
- Markuna S, Kumar P, Ali R et al (2023) Application of innovative machine learning techniques for long-term rainfall prediction. *Pure Appl Geophys* 180:335–363. <https://doi.org/10.1007/s00024-022-03189-4>
- Marques EAG, Silva Junior GC, Eger GZS et al (2020) Analysis of groundwater and river stage fluctuations and their relationship with water use and climate variation effects on Alto Grande watershed, Northeastern Brazil. *J South Am Earth Sci* 103:102723. <https://doi.org/10.1016/j.jsames.2020.102723>
- Panda RK, Pramanik N, Bala B (2010) Simulation of river stage using artificial neural network and MIKE 11 hydrodynamic model. *Comput Geosci* 36:735–745. <https://doi.org/10.1016/j.cageo.2009.07.012>
- Pham BT, Tien Bui D, Prakash I (2017) Landslide susceptibility assessment using bagging ensemble based alternating decision trees, logistic regression and j48 decision trees methods: a comparative study. *Geotech Geol Eng* 35:2597–2611. <https://doi.org/10.1007/s10706-017-0264-2>
- Qiao J, He Z, Du S (2020) Prediction of PM2.5 concentration based on weighted bagging and image contrast-sensitive features. *Stoch Environ Res Risk Assess* 34:561–573. <https://doi.org/10.1007/s00477-020-01787-z>
- Samantaray S, Sahoo A, Satapathy DP (2022) Prediction of groundwater-level using novel SVM-ALO, SVM-FOA, and SVM-FFA algorithms at Purba-Medinipur. *India Arab J Geosci* 15:723. <https://doi.org/10.1007/s12517-022-09900-y>
- Saroughi M, Mirzania E, Vishwakarma DK et al (2023) A novel hybrid algorithms for groundwater level prediction. *Iran J Sci Technol Trans Civ Eng*. <https://doi.org/10.1007/s40996-023-01068-z>
- Schapire RE (2003) The boosting approach to machine learning: An overview. In: Denison DD, Hansen MH, Holmes CC et al (eds) *Nonlinear estimation and classification*. Lecture notes in statistics, 107th edn. Springer, New York, pp 149–171
- Seo Y, Kim S (2016) River stage forecasting using wavelet packet decomposition and data-driven models. *Proc Eng* 154:1225–1230. <https://doi.org/10.1016/j.proeng.2016.07.439>

- Seo Y, Kim S, Kisi O et al (2016a) River stage forecasting using wavelet packet decomposition and machine learning models. *Water Resour Manag* 30:4011–4035. <https://doi.org/10.1007/s11269-016-1409-4>
- Seo Y, Kim S, Singh VP (2016b) Physical interpretation of river stage forecasting using soft computing and optimization algorithms. In: Kim JH, Geem ZW (eds) *Harmony search algorithm. Advances in intelligent systems and computing*, 382nd edn. Springer, Berlin, pp 259–266
- 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, New Jersey
- Strupczewski WG, Singh VP, Mitosek HT (2001) Non-stationary approach to at-site flood frequency modelling. III. Flood analysis of Polish rivers. *J Hydrol* 248:152–167. [https://doi.org/10.1016/S0022-1694\(01\)00399-7](https://doi.org/10.1016/S0022-1694(01)00399-7)
- Sun D, Xu J, Wen H, Wang D (2021) Assessment of landslide susceptibility mapping based on Bayesian hyperparameter optimization: a comparison between logistic regression and random forest. *Eng Geol* 281:105972. <https://doi.org/10.1016/j.enggeo.2020.105972>
- Tang D, Tang L, Dai R et al (2020) MF-Adaboost: LDoS attack detection based on multi-features and improved Adaboost. *Futur Gener Comput Syst* 106:347–359. <https://doi.org/10.1016/j.future.2019.12.034>
- Tien Bui D, Ho T-C, Pradhan B et al (2016) GIS-based modeling of rainfall-induced landslides using data mining-based functional trees classifier with AdaBoost, Bagging, and MultiBoost ensemble frameworks. *Environ Earth Sci* 75:1101. <https://doi.org/10.1007/s12665-016-5919-4>
- Vishwakarma DK, Kumar R, Abed SA et al (2023a) Modeling of soil moisture movement and wetting behavior under point-source trickle irrigation. *Sci Rep* 13:14981. <https://doi.org/10.1038/s41598-023-41435-4>
- Vishwakarma DK, Kumar R, Tomar AS, Kuriqi A (2023b) Eco-hydrological modeling of soil wetting pattern dimensions under drip irrigation systems. *Heliyon* 9:e18078. <https://doi.org/10.1016/j.heliyon.2023.e18078>
- Vishwakarma DK, Kuriqi A, Abed SA et al (2023c) Forecasting of stage-discharge in a non-perennial river using machine learning with gamma test. *Heliyon* 9:e16290. <https://doi.org/10.1016/j.heliyon.2023.e16290>
- Wang F, Wang Y, Zhang K et al (2021) Spatial heterogeneity modeling of water quality based on random forest regression and model interpretation. *Environ Res* 202:111660. <https://doi.org/10.1016/j.envres.2021.111660>
- Wu CL, Chau KW, Li YS (2008) River stage prediction based on a distributed support vector regression. *J Hydrol* 358:96–111. <https://doi.org/10.1016/j.jhydrol.2008.05.028>
- Xue L, Liu Y, Xiong Y et al (2021) A data-driven shale gas production forecasting method based on the multi-objective random forest regression. *J Pet Sci Eng* 196:107801. <https://doi.org/10.1016/j.petrol.2020.107801>
- Yang X, Wang Y, Byrne R et al (2019) Concepts of artificial intelligence for computer-assisted drug discovery. *Chem Rev* 119:10520–10594. <https://doi.org/10.1021/acs.chemrev.8b00728>
- Yariyan P, Janizadeh S, Van Phong T et al (2020) Improvement of best first decision trees using bagging and dagging ensembles for flood probability mapping. *Water Resour Manag* 34:3037–3053. <https://doi.org/10.1007/s11269-020-02603-7>
- Zounemat-Kermani M, Batelaan O, Fadaee M, Hinkelmann R (2021) Ensemble machine learning paradigms in hydrology: A review. *J Hydrol* 598:126266. <https://doi.org/10.1016/j.jhydrol.2021.126266>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Salim Heddami¹  · Dinesh Kumar Vishwakarma²  · Salwan Ali Abed³  · Pankaj Sharma⁴ · Nadhir Al-Ansari⁵  · Abed Alataway⁶ · Ahmed Z. Dewidar^{6,7} · Mohamed A. Mattar^{6,7,8} 

✉ Dinesh Kumar Vishwakarma
dinesh.vishwakarma4820@gmail.com

✉ Nadhir Al-Ansari
nadhira.alansari@itu.se

✉ Mohamed A. Mattar
mmattar@ksu.edu.sa

Salim Heddami
heddamisalim@yahoo.fr

Salwan Ali Abed
salwan.abed@qu.edu.iq

Pankaj Sharma
pankajsw122@gmail.com

Abed Alataway
aalataway@ksu.edu.sa

Ahmed Z. Dewidar
adewidar@ksu.edu.sa

¹ Faculty of Science, Agronomy Department, Hydraulics Division, Laboratory of Research in Biodiversity Interaction Ecosystem and Biotechnology, University 20 Aout 1955, Route El Hadaik, BP 26, Skikda, Algeria

² Department of Irrigation and Drainage Engineering, Govind Ballabh Pant University of Agriculture and Technology, Pantnagar, Uttarakhand 263145, India

³ Department of Environment, College of Science, University of Al-Qadisiyah, Al-Qadisiyah 58001, Iraq

⁴ Department of Soil and Water Engineering, Punjab Agricultural University, Ludhiana, Punjab 141027, India

- ⁵ Department of Civil, Environmental, and Natural Resources Engineering, Lulea University of Technology, 97187 Lulea, Sweden
- ⁶ Prince Sultan Bin Abdulaziz International Prize for Water Chair, Prince Sultan Institute for Environmental, Water and Desert Research, King Saud University, 11451 Riyadh, Saudi Arabia
- ⁷ Department of Agricultural Engineering, College of Food and Agriculture Sciences, King Saud University, 11451 Riyadh, Saudi Arabia
- ⁸ Agricultural Engineering Research Institute (AEnRI), Agricultural Research Centre, Giza 12618, Egypt