Stochastic Environmental Research and Risk Assessment (2022) 36:2753–2768
https://doi.org/10.1007/s00477-021-02159-x

ORIGINAL PAPER

Improving the performance of random forest for estimating monthly reservoir inflow via complete ensemble empirical mode decomposition and wavelet analysis

Farshad Ahmadi[1] · Saeid Mehdizadeh[2] · Vahid Nourani[3,4]

Accepted: 11 December 2021 / Published online: 29 January 2022

Abstract
Estimation of reservoir inflow is of particular importance in optimal planning and management of water resources, proper allocation of water to consumption sectors, hydrological studies, etc. This study aimed to estimate monthly inflow (Q) to the Maroon Dam reservoir located in Iran utilizing climatic data such as minimum, maximum, and mean air temperatures (T_{min}, T_{max}, T), reservoir evaporation (E), and rainfall (R). The impact of any of the mentioned variables was analyzed by the entropy-based pre-processing technique. The results of the pre-processing showed that the rainfall is the most important parameter affecting the reservoir inflow. Therefore, three types of input patterns were taken into consideration consisting the antecedent Q-based, antecedent R-based, and combined antecedent Q and R-based input combinations. To estimate the monthly reservoir inflow, a random forest (RF) was firstly employed as the standalone model. Then, two different types of hybrid models were proposed via coupling the RF on complete ensemble empirical mode decomposition (CEEMD) and wavelet analysis (W) in order to implement the coupled CEEMD-RF and W-RF models. It is worthwhile to mentioning that six mother wavelets were used in developing the hybrid W-RF models. Four error metrics including root mean square error (RMSE), mean absolute error (MAE), Kling-Gupta efficiency (KGE), and Willmott index (WI) were used to assess the accuracy of implemented models. The attained results indicated the superiority of proposed hybrid models over the classic RF for estimating the monthly reservoir inflow. The most precise model during the test phase was W-RF(3) utilizing the Sym(2) as the mother wavelet under a lagged Q-based pattern with error measures of RMSE = 15.011 m^3/s, MAE = 10.439 m^3/s, KGE = 0.832, WI = 0.773.

Keywords Complete ensemble empirical mode decomposition · Estimation · Hybrid models · Monthly reservoir inflow · Random forest · Wavelet analysis

1 Introduction

Water resources not only are essential for the human survival but also are a very important segment of socio-economic conservation (Chu and Huang 2020). Iran is located in arid and semi-arid regions of the world and therefore rainfall plays a significant role in meeting water demands. However, most of the rainfall events occur in the cold seasons of the year when the agricultural activities are in their lowest levels. Hence, there is a substantial need to store water in reservoir dams to supply water needs in the hot seasons (Khalili et al. 2016; Ahmadi et al. 2018; Pour et al. 2020; Salehi et al. 2020; Sharafi and Karim 2020).

An accurate estimation of inflows to dams is of particular importance for the short-term and long-term

✉ Saeid Mehdizadeh
saied.mehdizadeh@gmail.com

[1] Department of Hydrology and Water Resources Engineering, Shahid Chamran University of Ahvaz, Ahvaz, Iran

[2] Water Engineering Department, Urmia University, Urmia, Iran

[3] Center of Excellence in Hydroinformatics, Faculty of Civil Engineering, University of Tabriz, Tabriz, Iran

[4] Faculty of Civil and Environmental Engineering, Near East University, Near East Boulevard, via Mersin 10, 99138 Nicosia, Turkey

exploitation and plays a very important role in sustainable agriculture, floods and droughts management (Afan et al. 2020). For this purpose, many models have been proposed and a lot of research is being done for developing models to estimate complex hydrological phenomenon as accurately as possible (Rahmani-Rezaeieh et al. 2020). In this regard, the main problem is the involvement and impacts of different parameters like evaporation, rainfall, temperature, and other climatic factors, which should be taken into consideration in the hydrological studies (Nayak et al. 2004).

For modeling the inflows to the reservoirs, due to the non-linear nature, different perspectives have been proposed for the development and improvement of inflow predictive models (Rahmani-Rezaeieh et al. 2020). In general, two techniques including conceptual (white box) and systemic (black box) models have been recommended when modeling hydrological phenomena. The white box models are developed based on governing mathematical equations and existing physical parameters (Singh 2018). On the other side, it is not possible to present mathematical relationships in the black box models and the physical variables affecting the target parameter could not be easily recognized. The black box models include the potential of estimating the intended output by receiving the possible inputs and then performing a series of mathematical operations on them. The performance of black box models is significantly dependent on the quantity and quality of the data used (Mehr et al. 2017). Artificial intelligence (AI) model is a typical type of black box-based models that has been extensively used in recent years to solve various hydrological problems such as rainfall-runoff modeling (Vidyarthi et al. 2020; Adnan et al. 2021a; Herath et al. 2020; Molajou et al. 2021), estimating the rainfall (Nourani et al. 2019; Mehdizadeh 2020), river streamflow forecasting (Mehdizadeh and Sales 2018; Fathian et al. 2019; Mohammadi et al. 2020; Adnan et al. 2021b), and inflows to the dams reservoirs (Santos et al. 2019; Apaydin et al. 2020Lee et al. 2020).

One of the AI models is random forest (RF), which uses multiple iterative algorithms. It can be utilized as a powerful technique for evaluating the hydrological issues (Booker and Snelder 2012). The RF can learn complex patterns and consider the non-linear relationships between the independent and dependent variables. Besides, identifying the most effective input parameters influencing the target desired output is one of the important features of the RF. The aforementioned benefits have led to the use of RF when forecasting hydrological parameters (e.g., see Ali et al. 2020; Ghorbani et al. 2020; Hussain and Khan 2020; Pham et al. 2020; Tang et al. 2020).

In the application of AI-based models such as RF, determining the optimal input data always plays a major role in their final performance. Moreover, introducing the maximum number of inputs will not necessarily lead to achieving the highest accuracy of the relevant model. The Shannon's entropy theory is one of the approaches proposed in recent years for selecting the optimal inputs of the AI models (Ahmadi et al. 2021a). This theory shows that an event with a high probability of occurrence could provide less information; otherwise, if an event is less likely to occur, more information may be achieved (Saray et al. 2020). Indeed, the uncertainties are reduced through capturing the new information and the value of new information is equivalent to the amount of reduced uncertainty (Pei-Yue et al. 2010). Therefore, by weighting each of the inputs by the entropy method, the most effective ones can be selected and used in the modeling procedure. Such methodology has been already used in various studies when selecting the optimal input predictors (Darbandsari and Coulibaly 2020; Roy 2021; Ray and Chattopadhyay 2021).

Most of the recorded hydrological data have some noises so that they prevent the proper transfer of information to the models. Data pre-processing methods have been proposed to overcome this problem, which wavelet theory (W) and empirical mode decomposition (EMD) belong to such methods. The wavelet analysis is more sensitive to the proper choice of the mother wavelet type, but there is no such limitation in the EMD method and it can be therefore applied to the data without any special preconditions. EMD is a spectral analysis method, which was firstly introduced by Huang et al. (1998). After introducing the initial version (i.e., EMD), Wu and Huang (2009) proposed ensemble EMD (EEMD) due to the problem of mode composition existing in the EMD. Torres et. al. (2011) then introduced complete EEMD (i.e., CEEMD) to eliminate the imperfection of the previous versions (i.e., EMD and EEMD). Each of these methods has properties that make them suitable for decomposing the different original data. Data decomposition utilizing each of the EMD, CEMD and CEEND divides it into sections called as intrinsic modes, each of which contains parts of the same scale of data. Diverse coupled models have been proposed in literature to forecast hydrological parameters with the aim of this feature of EMD (e.g., see Chen and Dong 2020; Nazir et al. 2020; Ouarda et al. 2021).

As mentioned above, knowing the inflow time series to a dam reservoir could be of significant use for the optimal management and optimal allocation of water resources. The main objectives of present study are as follows: to (1) apply a pre-processing approach based on the entropy technique when implementing input patterns related to inflow estimation, (2) develop classic RF and then propose novel hybrid models through hybridizing the RF with the CEEMD and W, (3) evaluate the efficiency of six mother wavelets in developing the hybrid W-RF models, (4) compare the performance of whole the models proposed in the current study. According to the best knowledge of the authors, this study is the first try in the literature to propose the hybrid CEEMD-RF and compare its performance with the coupled W-RF ones when estimating the monthly reservoir inflow.

2 Materials and methods

2.1 Study area and data used description

The Maroon River originates in the Nil Mountains and springs in the foothills of the Sadat Mountains of the Zagros in Kohgoluyeh and Boyer-Ahmad Province in Iran. It reaches the Maroon Dam Lake after a distance of 120 km and enters the Behbahan plain through the Takab Strait. The Maroon Reservoir Dam is located 19 km northeast of Behbahan with a height of 165 m, a length of 345 m, a width of 15 m and a total volume of the reservoir up to 1200 million cubic meters. This dam is of sandy gravel type with clay core. The geographical position of study location is shown in Fig. 1.

Idanak hydrometric station, located in Idanak village and upstream of the Maroon Reservoir Dam, records the required data. The data sets applied in the current study were comprised of the minimum, maximum, and mean air temperatures (T_{min}, T_{max}, T), rainfall (R), reservoir evaporation (E), and reservoir inflow (Q) during 1982–2017 on a monthly time-scale. From whole the available data (i.e., 420 data), 300 data were used to train the models while 120 data were applied when testing the developed models. Figure 2 demonstrates the time series of monthly data used in this study during both the training and testing periods. Some of the statistical properties of the data used consisting of minimum (Min), maximum (Max), Average (Avg), standard deviation (SD), and coefficient of variation (CV) for both the train and test phases are summarized in Table 1.

3 Models applied overview

3.1 Entropy-based input selection

In modeling of an intended problem using the artificial intelligence-based approaches, defining the effective parameters as the models inputs plays a significant role in improving their performances. In addition, in the time series modeling of the hydrological phenomena, considering the effective lags of the investigated problem can lead to an acceptable result (Ahmadi et al. 2021a). The models inputs were discerned in this study through the Shannon's entropy measure. This method derived from the information theory was initially introduced by Shannon (1948). Entropy is a measure of disorder in a system and is also a measure of the amount of uncertainty expressed by a discrete probability distribution in information theory; so that, this uncertainty is greater if the frequency distribution is well distributed than when the frequency distribution is sharper (Bednarik et al. 2010). This technique requires a matrix based on criteria and options. If the decision matrix data are known, the entropy technique can be employed to evaluate the weights.

Here, the monthly minimum, maximum, and mean air temperatures, monthly rainfall, and monthly reservoir evaporation were considered as the possible inputs effective on the monthly reservoir inflow. The most important variables were then identified using the entropy method. In most of the previous studies, a systematic method is not provided to specify the optimal lags when modeling the intended problem. In the present study, the entropy technique was also applied to select the appropriate lags of the considered inputs.

3.2 Random forest

Random forest (RF) as a data-driven method is firstly proposed by Breiman (2001). Indeed, it is developed for solving problems based on the regression and clustering through the development of decision trees (Fathian et al. 2019). An RF is comprised of a collection of un-pruned trees in which each tree is obtained by a recursive segmentation algorithm. In other words, the RF is a combined form of some decision trees so that several self-organizing samples of data are involved in its construction (Friedman et al. 2001). To create a regression tree, recursive segmentation and multiple regressions are used. The decision process is repeated at each internal node of the root node according to the tree rule until the pre-determined stop condition is met (Breiman 2001).

In the RF, a random vector X_n is generated for the nth tree, which is independent of random vectors $X_1, X_2, ..., X_{n-1}$. Tree regression generates a set of trees utilizing the training dataset and achieved X_n as follows (Breiman 2001):

$$X_n = \{h_1(x), h_2(x), ..., h_n(x)\} \quad (1)$$

$$h_n = h(x, X_n), x = \{x_1, x_2, ..., x_p\} \quad (2)$$

The above P-dimensional vector forms a forest and the outputs for each tree are provided as (Breiman 2001):

$$y_1 = h_1(x), y_2 = h_2(x), ..., y_n = h_n(x) \quad (3)$$

where y_n denotes the output of nth tree.

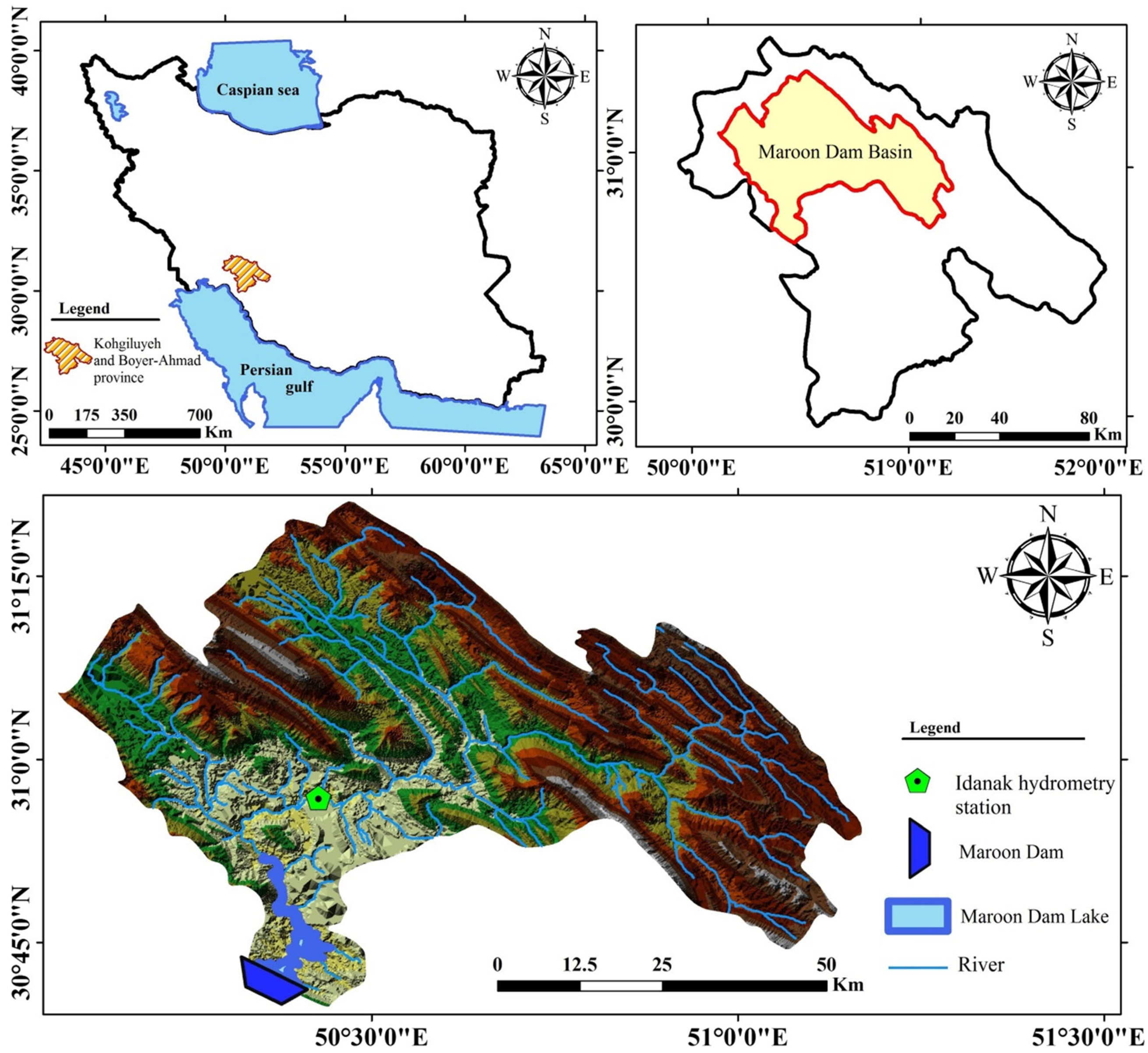

Fig. 1 Geographical position of study location

To obtain the final output, the average of predictions of all the tress is calculated (Breiman 2001). The prediction error is also computed according to Eq. (4) as (Breiman 2001):

$$MSE = \frac{\sum_{i=1}^{n} [y(x_i) - y_i]^2}{n} \tag{4}$$

where $y(x_i)$ illustrates the computational value, y_i denotes the observational value, n is the total number of observations, and MSE shows the mean square error rate between the observational and computational values.

3.3 Wavelet theory

A wavelet is a class of mathematical functions used to decompose a continuous signal into its frequency components. This method is a time-independent spectral analysis that separates time series in a time–frequency space in order to describe the time scale of processes and their relationships. Wavelet transform, like the Fourier transform, considers the time series as a linear combination of several base functions. One of the most important characteristics of the wavelet transform is its ability to obtain information in time, frequency, and position, simultaneously (Misiti et al. 1996). Continuous wavelet transform includes the capability to operate at any scale. However,

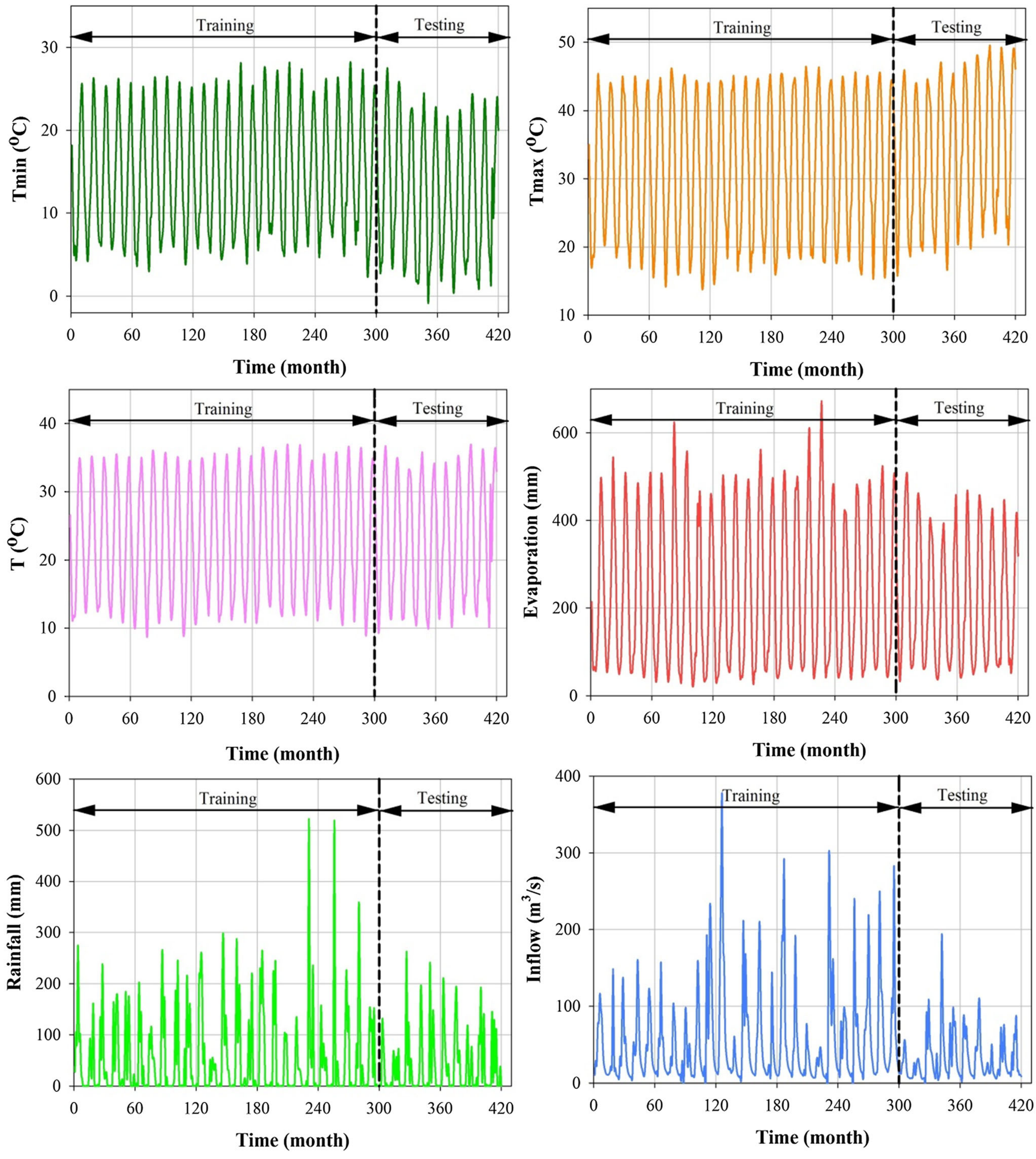

Fig. 2 Time series of the monthly climatic data as possible inputs and reservoir inflow as the target during the study period

the difficulty of calculating the wavelet coefficients as well as the need for high computational time and the production of large volumes of data are some of the problems of this type of wavelet transform. Discrete wavelet transform (DWT) method can be used to solve this problem (Chen et al. 1999).

To implement the DWT method, the Mallat algorithm or the Multi Resolution Analysis (MAR) method is presented (Mallat 2009). In this approach, the decomposed signal is passed through low-pass and high-pass filters. The low and high frequency contents of the signal are named as approximation and details, respectively (Mehdizadeh et al.

Table 1 Statistical parameters of the data used in this study

Parameters	Train					Test				
	Min	Max	Avg	SD	CV	Min	Max	Avg	SD	CV
T_{min}, °C	2.60	28.10	14.85	7.23	0.49	−0.70	27.50	12.44	7.66	0.62
T_{max}, °C	14.20	46.40	31.36	10.17	0.32	24.70	49.50	33.70	10.18	0.30
T, °C	9.10	36.80	23.11	8.64	0.37	9.30	36.80	23.25	8.63	0.37
R, mm	0.00	522.20	53.63	82.39	1.54	0.00	263.00	39.49	58.39	1.48
E, mm	24.20	672.30	237.42	171.13	0.72	32.60	505.10	217.44	142.55	0.66
Q, m^3/s	0.68	377.17	54.57	61.13	1.12	4.25	194.07	31.94	30.16	0.94

2020a; Ahmadi et al. 2021b). This filtering paradigm can be applied to obtain a time-scale display of a signal (Polikar 1999). In the DWT, the primary signal could be reconstructed via the synthesizing of the wavelet coefficients. This operation starts from the last level of decomposition and the original signal could be reconstructed through assembling the approximation and details series.

3.4 Complete ensemble empirical mode decomposition

Empirical mode decomposition (EMD) is a method of spectral data analysis, which was firstly proposed by Huang et al. (1998). This method has evolved several stages since its introduction. Wu and Huang (2009) then introduced ensemble EMD (EEMD) due to the problem of mode composition. Finally, Torres et al. (2011) solved the problem of imperfection of the EMD and EEMD methods by proposing the complete EEMD (CEEMD).

In the CEEMD method, intrinsic mode functions are displayed as $\overline{IMF}_k$. If we assume that the $E_j(.)$ operator provides the *j*th intrinsic mode computed by the EMD, ω^i is the white noise with standard deviation *N(0,1)*, *x* denotes the original data, and ε_0 illustrates an initial constant, the different steps of CEEMD are as follows:

The first intrinsic mode $x + \varepsilon_0\omega^i$ is calculated via the EMD and the first intrinsic mode of CEEMD is computed as shown in Eq. (5) (Torres et al. 2011):

$$\overline{IMF}_1 = \frac{1}{I}\sum_{i=1}^{I} IMF_1^i \tag{5}$$

The first residual value is then calculated from Eq. (6) as (Torres et al. 2011):

$$r_k = r_{k-1} - \overline{IMF}_k \tag{6}$$

In the next step, the second intrinsic mode function is obtained as (Torres et al. 2011):

$$\overline{IMF}_2 = \frac{1}{I}\sum_{i=1}^{I} E_1(r_1 + \varepsilon_1 E_1(\omega^i))\text{where } r_1 = r_{k-1} + \varepsilon_1 E_1(\omega^i) \text{ and } i = 1,\ldots,I. \tag{7}$$

The residual value is computed as the Eq. (6) for $k = 2,\ldots..k$.

The *(k + 1)*th intrinsic mode function is obtained from the following Eq. (8) as (Torres et al. 2011):

$$\overline{IMF}_{(k+1)} = \frac{1}{I}\sum_{i=1}^{I} E_1(r_k + \varepsilon_k E_k(\omega^i)) \tag{8}$$

where $i = 1,\ldots..,I$ As long as the residual has more than three extremes, the procedure of extracting the intrinsic mode functions continues.

3.5 Models development

Firstly, an entropy approach was used to discern the most important climatic data to apply them when defining the inputs of the models. This technique was also utilized to determine the appropriate lags of the most effective inputs.

After determining the inputs patterns, the single RF and then hybrid CEEMD-RF and W-RF models were implemented. Firstly, the classic RF models were implemented taking into consideration of the mean squared error obtained in training and testing datasets. The optimal number of trees was then used when modeling the intended parameter using the RF so that no change in the mean squared error was observed by increasing the number of trees (Shataee et al. 2012). Besides, the data decomposition through the wavelet functions and CEEMD technique was utilized to generate the hybrid models. For this aim, the selected inputs by the entropy method were processed (using five mother wavelet functions with appropriate decomposition levels and CEEMD approach) and then introduced as inputs to the RF model; thus, the coupled W-RF and CEEMD-RF models were developed.

3.6 Performance assessment metrics

This study used four evaluation metrics including root mean square error (RMSE), mean absolute error (MAE), Kling-Gupta efficiency (KGE), and Willmott index (WI) to investigate the estimation accuracy of single RF and coupled CEEMD-RF and W-RF models. These statistical metrics can be formulated as the following equations:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\left(Q_{o,i} - Q_{e,i}\right)^2}{N}} \tag{9}$$

$$MAE = \frac{\sum_{i=1}^{N}\left|Q_{o,i} - Q_{e,i}\right|}{N} \tag{10}$$

$$KGE = 1 - \sqrt{(CC-1)^2 + (\alpha-1)^2 + (\beta-1)^2} \tag{11}$$

$$WI = \left|1 - \left[\frac{\sum_{i=1}^{N}\left(Q_{o,i} - Q_{e,i}\right)^2}{\sum_{i=1}^{N}\left(\left|Q_{o,i} - \overline{Q_o}\right| + \left|Q_{e,i} - \overline{Q_o}\right|\right)^2}\right]\right|, 0 \le WI \le 1 \tag{12}$$

where $Q_{o,i}$ and $Q_{e,i}$ denote the ith observed and estimated monthly reservoir inflows, respectively, $\overline{Q_o}$ illustrates the mean of observed inflows, N is the total number of observational values, CC indicates the correlation coefficient among the observed and estimated monthly inflows, α is the standard deviation ration for the observed and estimated monthly inflows, and finally β shows the mean ratio for the observed and estimated monthly inflows. As it is apparent, lower amounts of the RMSE and MAE as well as higher amounts of KGE and WI metrics verify better performance of respective model in estimating the monthly inflow time series.

In addition to the evaluation statistical metrics mentioned above, scatter and violin plots were also provided to visually investigate the estimation accuracy of standalone RF and hybrid CEEMD-RF and W-RF models.

4 Results and discussion

In all models based on artificial intelligence, the correct choice of inputs plays a significant role in achieving the desired performance in order to estimate the target parameter (e.g., monthly reservoir inflow in this research). Therefore, before modeling, it is necessary to examine the importance of each of the input parameters affecting the output parameter by preprocessing methods. Here, an entropy-based pre-processing technique was employed. Possible input variables influencing the monthly reservoir inflow in this study were comprised of monthly minimum air temperature (T_{min}), monthly maximum air temperature (T_{max}), monthly mean air temperature (T), reservoir evaporation (E), and rainfall (R).

In the entropy method, a certain weight is assigned to each of the input variables, which indicates the impact factor and the importance of this parameter on the output target parameter. Figure 3 shows the values of the weights (in percent) assigned to the considered input parameters in the form of a radar chart. As it can be clearly seen, rainfall (R) is the most important parameter affecting the monthly reservoir inflow due to having the highest weight (64.71%). After R, the evaporation (E) parameter gained more weight (21.13%) while the air temperature parameters had the lowest assigned weight values. Hence, only the rainfall variable was chosen among the variables considered when defining the input patterns. In the present study, three different types of input scenarios were taken into consideration including antecedent Q-based, antecedent R-based, and combined antecedent Q and R-based patterns. The entropy approach was used again to determine the appropriate lags of rainfall and inflow. In this regard, five lags of rainfall and inflow were considered. The values of weights (in percent) assigned to the different lags of rainfall and inflow are depicted schematically in Fig. 4. As shown, the first three lags have the highest weights in both the rainfall and inflow variables, which indicates their greater impacts on the target parameter.

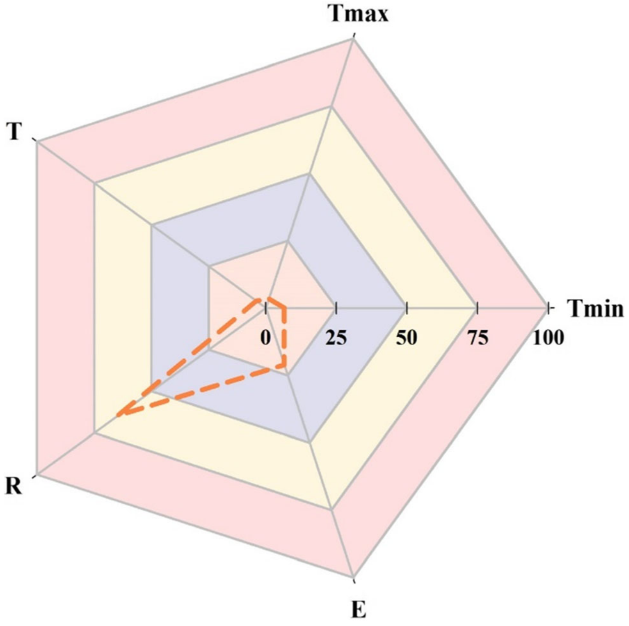

Fig. 3 Radar graph indicating the weights (in percent) assigned to each of the inputs

Initially, a classic RF was applied to estimate the monthly reservoir inflow of current month under the input patterns mentioned above. It is worth mentioning that the number of trees was selected in such a way that increasing the number of trees from the intended number had no significant effect on the performance RF-based models. The values of statistical metrics of RMSE, MAE, KGE, and WI computed for the single RF are summarized in Tables 2, 3, 4. As clear, the RF includes the potential of estimating the current month inflow as a function of intended inputs (i.e., antecedent Q in Table 2, antecedent R in Table 3, and combined antecedent Q and R in Table 4).

An attempt was then made in this study to enhance the accuracy of monthly reservoir inflow estimations via developing two types of coupled models. At first, a novel hybrid model was proposed by coupling the CEEMD on the classic RF. A performance comparison of the single RF and hybrid CEEMD-RF models in Tables 2–4 confirms the reliable potential of proposed coupled model compared to the classic RF. For an instance, considering the best hybrid model in Table 2 under the antecedent Q-based patterns during the test phase, it can be seen that the statistical measures of coupled CEEMD-RF3 are as RMSE = 16.723 m^3/s, MAE = 11.380 m^3/s, KGE = 0.434, WI = 0.752 while the mentioned error metrics of single RF3 were as RMSE = 39.152 m^3/s, MAE = 25.942 m^3/s, KGE = 0.354, WI = 0.435. This conclusion was also obtained for the other scenarios of this pattern as well as antecedent R-based and combined Q and R-based patterns (in Tables 3 and 4). The better estimation accuracy of hybrid CEEMD-RF models than the classical RF ones can be explained considering the fact that decomposing the original data via the CEEMD can provide the decomposed data so that they can be used successfully as the new inputs for improving the classic models performances.

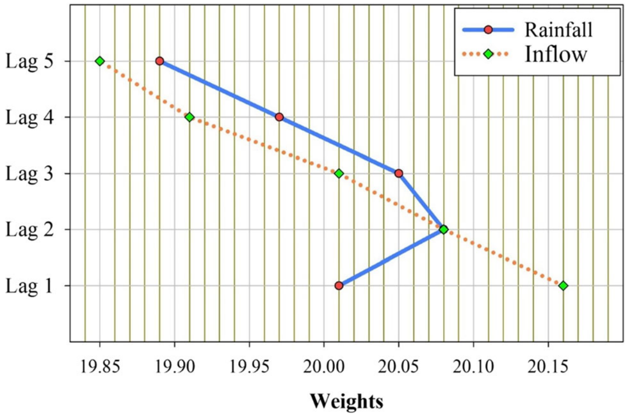

Fig. 4 The values of weights (in percent) assigned to the lagged rainfall and inflow data

In addition to proposing a new hybrid model called as CEEMD-RF, this study also developed another type of hybrid model using the hybridization of W theory and RF. Six various mother wavelets including Haar, Daubechies2 (db2), Daubechies4 (db4), Symlet (Sym), Coifflet (Coif), and Fejer-Korovkin (FK) were used during the development of coupled W-RF models. Based on the total number

Table 2 Statistical performance of conventional RF and hybrid CEEMD-RF models under the lagged Q-based patterns

Models	Train				Test			
	RMSE (m^3/s)	MAE (m^3/s)	KGE	WI	RMSE (m^3/s)	MAE (m^3/s)	KGE	WI
RF1	30.689	19.139	0.601	0.789	36.259	25.492	0.351	0.445
RF2	27.509	15.695	0.662	0.827	36.354	24.612	0.373	0.464
RF3	25.527	14.372	0.703	0.841	39.152	25.942	0.354	0.435
CEEMD-RF1	21.486	13.552	0.735	0.850	23.227	19.308	0.483	0.580
CEEMD-RF2	16.797	11.081	0.753	0.878	22.729	20.565	0.425	0.552
CEEMD-RF3	17.139	11.324	0.738	0.875	**16.723**	**11.380**	**0.434**	**0.752**

Bold values denote the statistical metrics of superior model in the test phase

Table 3 Statistical performance of conventional RF and hybrid CEEMD-RF models under the lagged R-based patterns

Models	Train				Test			
	RMSE (m^3/s)	MAE (m^3/s)	KGE	WI	RMSE (m^3/s)	MAE (m^3/s)	KGE	WI
RF1	31.643	20.249	0.588	0.776	33.302	23.510	0.478	0.488
RF2	25.252	15.540	0.717	0.828	32.525	21.618	0.485	0.529
RF3	24.211	14.339	0.727	0.842	30.225	20.128	0.532	0.562
CEEMD-RF1	20.316	13.041	0.754	0.856	31.632	25.360	0.508	0.448
CEEMD-RF2	18.733	11.405	0.752	0.874	**24.790**	**17.846**	**0.618**	**0.611**
CEEMD-RF3	19.658	12.005	0.743	0.851	30.110	23.934	0.533	0.479

Bold values denote the statistical metrics of superior model in the test phase

Table 4 Statistical performance of conventional RF and hybrid CEEMD-RF models under the lagged Q and R-based patterns

Models	Train				Test			
	RMSE (m^3/s)	MAE (m^3/s)	KGE	WI	RMSE (m^3/s)	MAE (m^3/s)	KGE	WI
RF1	25.917	14.283	0.718	0.842	32.455	21.579	0.497	0.530
RF2	24.232	13.336	0.735	0.853	30.462	21.622	0.526	0.529
RF3	22.909	12.670	0.733	0.860	32.480	22.710	0.489	0.505
CEEMD-RF1	16.869	10.105	0.786	0.888	22.112	17.918	0.638	0.610
CEEMD-RF2	16.652	10.033	0.764	0.889	18.569	14.133	0.688	0.692
CEEMD-RF3	17.067	10.507	0.744	0.884	**17.482**	**13.285**	**0.669**	**0.711**

Bold values denote the statistical metrics of superior model in the test phase

of observational data used for the modeling procedure (i.e., 420 data in the current study), certain levels of decomposed data should be used (Mehdizadeh et al. 2020a, 2020b). Here, two levels of data decomposition were taken into consideration ($Int[Log(420)] = 2$). The numbers in the parenthesis mentioned after the name of used mother wavelet in Tables 5, 6, 7 (i.e., 1 and 2) denote the level of decomposition applied when developing the coupled W-RF models. A comparative assessment of the classic RF with error metrics mentioned in Tables 2–4 and hybrid W-RF models with error metrics tabulated in Tables 5–7 clearly verifies that hybridizing the W and RF could lead to more accurate estimates of monthly reservoir inflow. As an example, the values of statistical metrics achieved for the single RF2 during the test period of Q and R-based patterns in Table 4 (i.e., RMSE = 30.462 m^3/s, MAE = 21.622 m^3/s, KGE = 0.526, WI = 0.529) were improved to RMSE = 15.418 m^3/s, MAE = 10.825 m^3/s, KGE = 0.806, WI = 0.764 in the hybrid W-RF2 model utilizing Sym(2) mother wavelet. The dependable performance of hybrid W-RF models than the classical RF can be justified by explaining the fact that the wavelet analysis provides useful subsets of the original observations series, which can increase the model's potential to estimate the desired target parameter by extracting suitable information produced by these new sub-series.

In a review paper, Nourani et al. (2014) evaluated the ability of the wavelet-artificial neural networks (W-ANN) hybrid model in various hydrological contexts (including rainfall-runoff) at short- and long-term time scales. They found out that due to the use of subsets resulting from the wavelet transform as the inputs of neural network models, the model performance increases significantly, which is completely consistent with the results of the present study.

A performance evaluation of six different mother wavelets when coupling them on the classic RF (Tables 5–7) clearly affirms that Sym and Coif are the best wavelets because of having lowest error values of the corresponding hybrid W-RF models; therefore, these wavelets could be suggested to be used as the suitable mother wavelets when estimating the monthly reservoir inflow through the hybrid W-RF technique. On the contrary, least-performing wavelets were the Haar and FK, which are not recommended. As mentioned above, two levels of decomposition were employed in the development of W-RF models. According to the values of statistical indicators mentioned in Tables 5–7, it can be clearly concluded that the estimation accuracy of coupled W-RF models was generally improved through applying the two decomposition levels in comparison to the use of one decomposition level. The wavelet transform by decomposing the original time series at higher decomposition levels helps to better interpret the structure of the original observational series and obtain useful information about its history; hence, this issue can be one of the reasons for improving the performance of W-RF models with increasing the level of data decomposition (Mehr et al. 2014).

Comparing the modeling accuracy of monthly reservoir inflow utilizing the hybrid CEEMD-RF and W-RF models demonstrates that CEEMD-RF models outperformed the W-RF ones for some cases and vice versa W-RF showed superior results than the CEEMD-RF for other cases. However, the W-RF models generally surpass the CEEMD-RF ones. The superior models for the estimation of monthly reservoir inflow time series of study location in the test stage were W-RF3 via Sym(2) wavelet under the antecedent Q-based patterns, W-RF2 through the Sym(2) wavelet under the antecedent R-based patterns, and W-RF2 via Sym(2) wavelet under the antecedent Q and R-based patterns. The values of statistical metrics for the mentioned superior models are bolded in Tables 5–7.

Regarding the ability of the intended input patterns, it can be seen from Tables 2–7 that the single RF and hybrid CEEMD-RF and W-RF models provided lower performances under the antecedent R-based input patterns. On the other side, using patterns based on the combined

Table 5 Statistical performance of hybrid W-RF models utilizing various mother wavelets under the lagged Q-based patterns

Models	Wavelet type	Train				Test			
		RMSE (m^3/s)	MAE (m^3/s)	KGE	WI	RMSE (m^3/s)	MAE(m^3/s)	KGE	WI
W-RF1	Haar(1)	25.644	14.104	0.728	0.844	30.832	20.085	0.444	0.563
	Haar(1)	24.236	13.663	0.744	0.849	21.944	15.091	0.709	0.671
	db2(1)	23.303	13.318	0.777	0.853	22.124	13.846	0.686	0.685
	db2(2)	19.244	10.885	0.824	0.880	17.133	10.975	0.734	0.761
	db4(1)	21.060	12.700	0.805	0.860	23.083	13.158	0.674	0.713
	db4(2)	17.700	10.847	0.819	0.880	21.547	13.401	0.699	0.708
	Sym(1)	24.957	13.896	0.752	0.841	22.733	14.788	0.679	0.678
	Sym(2)	20.366	11.222	0.801	0.854	18.523	11.321	0.701	0.742
	Coif(1)	21.858	13.629	0.758	0.849	21.249	14.602	0.637	0.682
	Coif(2)	20.961	13.029	0.779	0.856	17.038	12.007	0.739	0.739
	FK(1)	24.196	13.693	0.735	0.849	24.820	17.207	0.626	0.625
	FK(2)	20.117	11.023	0.803	0.878	24.460	17.159	0.630	0.626
W-RF2	Haar(1)	21.482	11.724	0.755	0.870	31.220	19.566	0.487	0.574
	Haar(1)	20.283	11.849	0.755	0.869	20.328	14.414	0.725	0.686
	db2(1)	17.399	9.785	0.829	0.892	19.125	11.895	0.735	0.702
	db2(2)	16.725	9.289	0.835	0.897	17.390	11.016	0.742	0.755
	db4(1)	15.543	8.809	0.849	0.903	18.686	10.950	0.759	0.762
	db4(2)	15.854	9.486	0.809	0.895	17.717	11.133	0.725	0.758
	Sym(1)	18.413	9.954	0.811	0.888	18.597	12.650	0.786	0.725
	Sym(2)	17.024	9.846	0.824	0.891	17.884	11.612	0.731	0.733
	Coif(1)	17.989	10.339	0.803	0.886	17.401	12.442	0.777	0.729
	Coif(2)	18.092	11.229	0.775	0.876	16.424	11.835	0.701	0.742
	FK(1)	20.380	11.313	0.765	0.875	21.780	15.669	0.707	0.659
	FK(2)	19.272	10.278	0.772	0.886	20.721	15.279	0.649	0.667
W-RF3	Haar(1)	21.086	11.268	0.742	0.876	30.475	19.959	0.510	0.565
	Haar(1)	18.644	11.102	0.759	0.877	20.348	14.645	0.718	0.681
	db2(1)	16.958	9.454	0.811	0.896	19.063	12.308	0.783	0.732
	db2(2)	16.553	9.428	0.812	0.896	18.441	12.194	0.809	0.748
	db4(1)	16.286	9.124	0.820	0.899	20.103	12.049	0.727	0.738
	db4(2)	16.677	10.051	0.787	0.889	18.076	11.540	0.702	0.749
	Sym(1)	17.013	9.496	0.810	0.896	17.145	11.046	0.794	0.762
	Sym(2)	16.313	9.395	0.815	0.896	**15.011**	**10.439**	**0.832**	**0.773**
	Coif(1)	17.478	9.844	0.795	0.891	18.236	13.126	0.754	0.714
	Coif(2)	18.619	11.554	0.751	0.872	16.755	11.952	0.684	0.740
	FK(1)	19.575	10.579	0.768	0.883	29.008	18.724	0.561	0.592
	FK(2)	19.498	10.461	0.762	0.884	20.128	15.145	0.629	0.670

Bold values denote the statistical metrics of superior model in the test phase

antecedent Q and R data is highly recommended to achieve the more accurate estimates of monthly reservoir inflow time series.

Besides the statistical error metrics used in the present study including the RMSE, MAE, KGE, and WI, two descriptive charts were also prepared and taken into consideration to visually evaluate the estimation accuracy of classic RF and coupled CEEMD-RF and W-RF models. In this context, scatter and violin diagrams were provided.

Figure 5 depicts the scatter plots of observed and estimated inflow data through the best models considering the different input patterns. According to this figure, it can be observed that the data dispersion around the dashed 1:1 line is significant for the single RF models, which indicates that

Table 6 Statistical performance of hybrid W-RF models utilizing various mother wavelets under the lagged R-based patterns

Models	Wavelet type	Train				Test			
		RMSE (m^3/s)	MAE (m^3/s)	KGE	WI	RMSE (m^3/s)	MAE (m^3/s)	KGE	WI
W-RF1	Haar(1)	24.202	15.092	0.768	0.833	28.613	20.925	0.532	0.544
	Haar(1)	24.160	14.868	0.750	0.836	29.237	20.294	0.550	0.558
	db2(1)	24.083	17.831	0.615	0.612	28.413	18.649	0.572	0.586
	db2(2)	28.595	19.528	0.558	0.575	20.424	15.547	0.702	0.682
	db4(1)	23.513	15.097	0.763	0.833	27.004	17.657	0.601	0.615
	db4(2)	20.884	13.375	0.787	0.852	27.268	17.850	0.586	0.611
	Sym(1)	23.513	15.097	0.763	0.833	27.001	18.134	0.596	0.605
	Sym(2)	21.613	12.936	0.797	0.857	19.351	14.020	0.719	0.695
	Coif(1)	23.768	14.939	0.739	0.835	25.990	17.891	0.621	0.610
	Coif(2)	22.680	13.710	0.750	0.849	29.846	18.783	0.542	0.591
	FK(1)	23.787	14.570	0.771	0.839	30.271	21.166	0.529	0.539
	FK(2)	22.679	14.020	0.775	0.845	29.179	20.435	0.546	0.555
W-RF2	Haar(1)	21.151	12.604	0.796	0.861	29.659	19.644	0.530	0.572
	Haar(1)	20.893	12.474	0.765	0.862	31.176	26.792	0.133	0.417
	db2(1)	20.898	11.832	0.785	0.869	28.021	18.846	0.546	0.587
	db2(2)	19.864	11.366	0.795	0.874	20.756	14.236	0.710	0.676
	db4(1)	20.125	11.619	0.793	0.872	26.709	17.628	0.593	0.616
	db4(2)	18.852	11.018	0.799	0.878	24.550	15.689	0.643	0.658
	Sym(1)	20.858	14.196	0.521	0.691	26.954	18.029	0.590	0.607
	Sym(2)	17.283	11.393	0.684	0.752	**19.616**	**13.878**	**0.729**	**0.698**
	Coif(1)	22.271	12.662	0.746	0.860	23.112	15.894	0.664	0.654
	Coif(2)	20.908	12.171	0.772	0.866	24.720	16.306	0.637	0.645
	FK(1)	21.693	12.462	0.790	0.862	30.007	20.401	0.528	0.556
	FK(2)	20.356	11.781	0.793	0.870	26.056	18.691	0.604	0.593
W-RF3	Haar(1)	20.365	11.704	0.788	0.871	26.605	18.535	0.584	0.596
	Haar(1)	20.190	11.950	0.765	0.868	26.309	17.756	0.611	0.613
	db2(1)	20.878	11.421	0.772	0.874	25.245	16.980	0.625	0.630
	db2(2)	19.580	11.170	0.784	0.877	23.541	16.421	0.674	0.654
	db4(1)	19.720	11.121	0.781	0.877	25.187	16.496	0.626	0.641
	db4(2)	18.642	10.935	0.793	0.879	22.778	14.951	0.676	0.674
	Sym(1)	20.007	11.324	0.738	0.877	24.746	15.811	0.637	0.643
	Sym(2)	19.314	11.044	0.805	0.880	21.702	15.255	0.692	0.668
	Coif(1)	19.812	11.244	0.774	0.876	22.236	15.457	0.682	0.663
	Coif(2)	19.730	11.624	0.774	0.872	23.096	15.759	0.668	0.657
	FK(1)	20.370	11.481	0.781	0.873	27.165	19.084	0.578	0.584
	FK(2)	42.252	27.010	0.652	0.702	25.441	18.464	0.613	0.598

Bold values denote the statistical metrics of superior model in the test phase

the classic RF could not perform well in estimating the observed monthly reservoir inflow data. However, coupling the RF with the CEEMD and W techniques has improved the accuracy of monthly inflow estimates. In this regard, W(Sym)(2)-RF3 hybrid model developed under the lagged Q-based pattern could present the highest convergence around the perfect 1:1 line.

One of the drawbacks of the scatter plot is that it does not provide any possibility to compare the distribution of estimated and observed data. In other words, through the scatter plot, it is not possible to find out whether the mean

Table 7 Statistical performance of hybrid W-RF models utilizing various mother wavelets under the lagged Q and R-based patterns

Models	Wavelet type	Train				Test			
		RMSE (m^3/s)	MAE (m^3/s)	KGE	WI	RMSE (m^3/s)	MAE (m^3/s)	KGE	WI
W-RF1	Haar(1)	21.471	11.225	0.802	0.876	23.742	16.820	0.650	0.634
	Haar(1)	21.354	11.733	0.791	0.870	19.896	13.630	0.766	0.703
	db2(1)	18.588	9.629	0.837	0.887	20.845	13.002	0.751	0.715
	db2(2)	18.094	9.502	0.842	0.890	16.845	11.482	0.802	0.723
	db4(1)	15.303	8.179	0.864	0.910	21.115	11.156	0.731	0.757
	db4(2)	14.897	8.636	0.850	0.905	17.540	9.994	0.773	0.782
	Sym(1)	18.223	9.350	0.848	0.897	20.430	12.970	0.753	0.718
	Sym(2)	17.331	9.332	0.851	0.897	15.726	10.482	0.824	0.772
	Coif(1)	17.615	9.951	0.825	0.890	18.375	11.782	0.754	0.743
	Coif(2)	17.177	10.218	0.826	0.887	16.844	10.800	0.776	0.765
	FK(1)	20.011	10.437	0.809	0.885	23.127	15.946	0.669	0.653
	FK(2)	18.450	9.961	0.819	0.890	21.134	15.308	0.711	0.667
W-RF2	Haar(1)	19.252	10.099	0.787	0.888	25.334	17.688	0.616	0.615
	Haar(1)	18.822	10.690	0.777	0.882	21.564	14.352	0.723	0.687
	db2(1)	17.730	9.342	0.800	0.877	20.124	13.211	0.751	0.712
	db2(2)	17.040	9.225	0.806	0.889	15.745	11.956	0.795	0.742
	db4(1)	15.161	8.205	0.840	0.909	20.038	11.664	0.740	0.746
	db4(2)	15.539	9.062	0.820	0.900	18.904	11.167	0.746	0.757
	Sym(1)	17.603	9.258	0.815	0.898	19.961	13.124	0.762	0.724
	Sym(2)	16.742	9.141	0.816	0.899	**15.418**	**10.825**	**0.806**	**0.764**
	Coif(1)	17.026	9.506	0.816	0.895	18.283	12.461	0.759	0.729
	Coif(2)	17.052	10.059	0.799	0.889	18.403	12.097	0.746	0.737
	FK(1)	19.212	10.036	0.794	0.889	22.066	15.990	0.680	0.652
	FK(2)	18.625	9.886	0.791	0.891	20.126	15.156	0.710	0.670
W-RF3	Haar(1)	19.819	10.418	0.769	0.885	25.690	18.581	0.602	0.595
	Haar(1)	18.582	10.614	0.767	0.883	21.078	16.846	0.618	0.633
	db2(1)	18.065	9.834	0.800	0.842	21.124	13.966	0.712	0.633
	db2(2)	16.997	9.874	0.838	0.841	15.955	12.112	0.642	0.712
	db4(1)	15.657	8.497	0.824	0.906	20.874	12.407	0.716	0.730
	db4(2)	15.637	9.160	0.812	0.899	18.095	12.449	0.460	0.729
	Sym(1)	17.521	9.236	0.806	0.892	20.680	13.840	0.730	0.699
	Sym(2)	16.886	9.128	0.814	0.896	15.884	11.999	0.802	0.739
	Coif(1)	17.374	9.687	0.797	0.893	18.729	12.884	0.747	0.719
	Coif(2)	17.168	10.198	0.784	0.887	18.799	14.289	0.360	0.689
	FK(1)	19.336	10.287	0.773	0.886	24.008	17.386	0.641	0.621
	FK(2)	19.085	10.252	0.770	0.887	19.720	15.025	0.502	0.673

Bold values denote the statistical metrics of superior model in the test phase

and variance of the observational data are correctly estimated by the developed models or not. To solve this problem, a violin diagram can be taken into consideration. It is better to mention that a violin diagram is another form of a box plot. Box plots only illustrate the minimum, maximum, mean, and quarters of the data; but, the violin diagram is used to visualize the data distribution and its possible density. The violin graphs for the optimal single and coupled models are given in Fig. 6. It can be seen that the single RF models with different inputs have not been able to estimate the maximum values well, but overestimation has occurred for the minimum and average data.

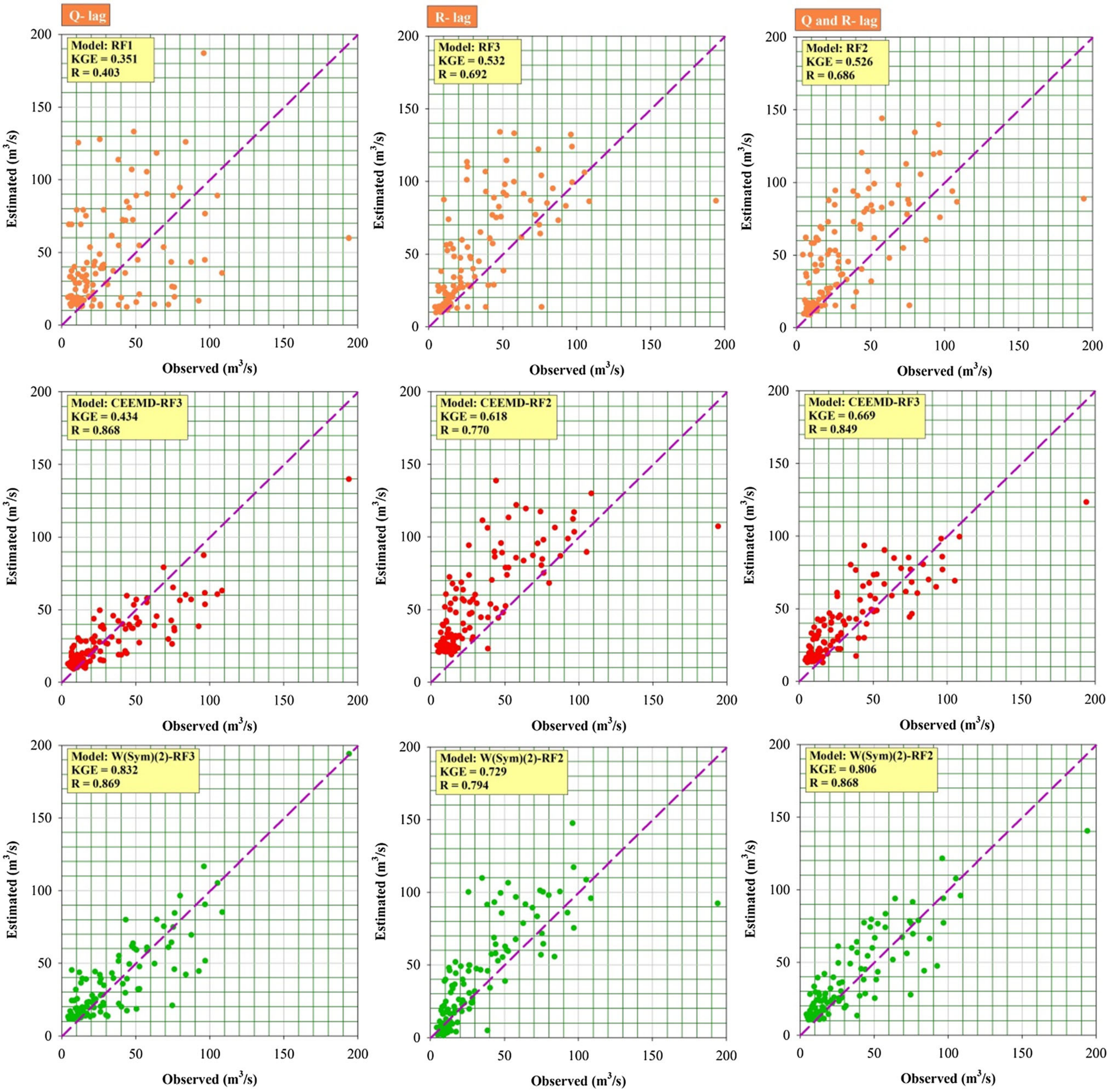

Fig. 5 Scatter plots of the observed and estimated monthly reservoir inflows via the best classic RF and hybrid CEEMD-RF and W-RF models for the considered input patterns during the test phase

The average of the estimated data is skewed. And therefore have a higher mean than the observational data. A comparison of the violin diagrams for the hybrid models of CEEMD-RF under the different input patterns shows that they could not be able to estimate the maximum values correctly. The hybrid W(Sym)(2)-RF3 model implemented under the lagged Q data-based pattern illustrated the best performance in estimating the observational inflow data so that the minimum and maximum values are estimated proportionally and the average of the estimated data is very close to the average of the observed values.

5 Conclusion

In the present study, improved models of RF were developed and proposed for the estimation of monthly reservoir inflow time series. To reach this goal, CEEMD and W were hybridized with the classic RF (i.e., CEEMD-RF and

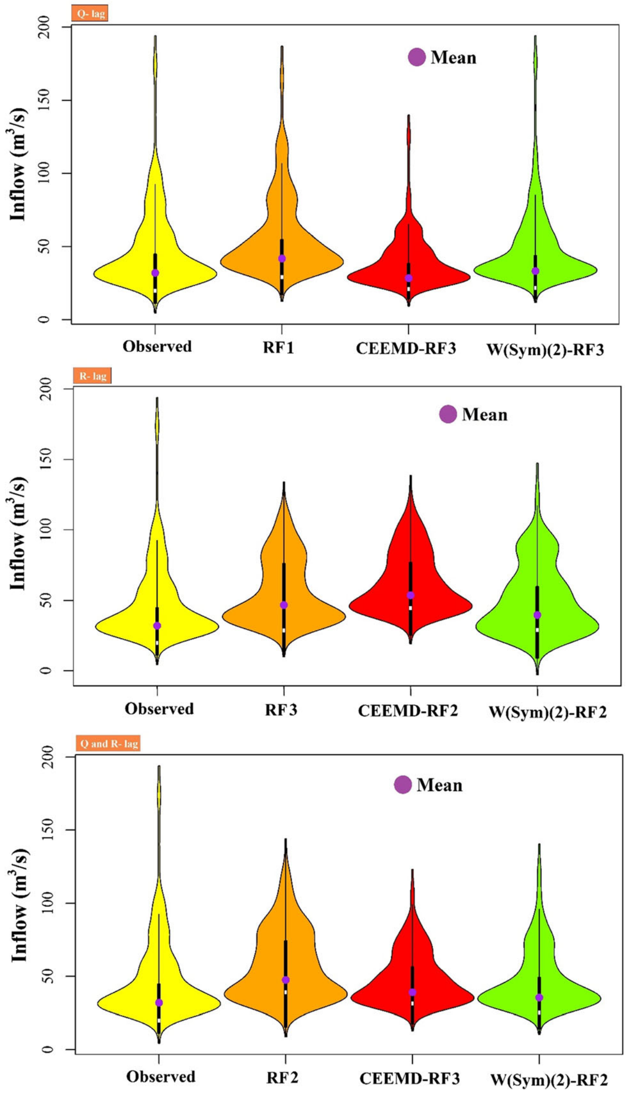

Fig. 6 Violin plots of the observed and estimated monthly reservoir inflows via the best classic RF and hybrid CEEMD-RF and W-RF models for the considered input patterns during the test phase

W-RF coupled models). To implement the hybrid W-RF, six various mother wavelets were employed under two decomposition levels. It is worthy to mention that an entropy-based pre-processing technique was used to determine the input patterns. The attained outcomes can be summarized as follows:

- Results of entropy approach revealed that the rainfall was the most important variable influencing the monthly inflow time series.
- Among the three different input patterns intended for the development of simple and hybrid models (i.e., antecedent Q-based, antecedent R-based, and combined antecedent Q and R-based patterns), whole the models developed via the application of combined antecedent Q and R data generally illustrated the better performance.
- Hybridizing the CEEMD and W techniques on the RF led to better estimations of the monthly inflow time series compared with the classic RF. Among the best-performing hybrid models of CEEMD-RF and W-RF, the best W-RF models demonstrated superior performances than the other hybrid ones.
- Testing the six different mother wavelets to couple them on the classic RF showed that Sym and Coif were generally the suitable wavelets to improve the estimation accuracy of monthly inflow through the hybrid W-RF models. On the other hand, Haar and FK wavelets were the least-performing wavelets.
- It was concluded that the estimation accuracy of W-RF models was significantly improved through increasing the decomposition levels from one to two when decomposing the input data.

This study applied the developed hybrid models for estimating the monthly reservoir inflow. It is recommended that the proposed hybrid models, specifically the new hybrid CEEMD-RF one, could be of use and tested for modeling the other hydrological phenomena like rainfall, river streamflow, evaporation, drought, etc. Besides the hybrid CEEMD-RF and W-RF models proposed in the current study, more efforts could be made to introduce other types of coupled techniques via hybridizing the artificial intelligence models with the time series analysis and nature-inspired optimization algorithms.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

Adnan RM, Petroselli A, Heddam S, Santos CAG, Kisi O (2021a) Short term rainfall-runoff modelling using several machine learning methods and a conceptual event-based model. Stoch Environ Res Risk Assess 35(3):597–616

Adnan RM, Liang Z, Parmar KS, Soni K, Kisi O (2021b) Modeling monthly streamflow in mountainous basin by MARS, GMDH-NN and DENFIS using hydroclimatic data. Neural Comput Applic 33(7):2853–2871

Afan HA, Allawi MF, El-Shafie A, Yaseen ZM, Ahmed AN, Malek MA, El-Shafie A (2020) Input attributes optimization using the feasibility of genetic nature inspired algorithm: application of river flow forecasting. Sci Reports 10(1):1–15

Ahmadi F, Mehdizadeh S, Mohammadi B, Pham QB, Doan TNC, Vo ND (2021a) Application of an artificial intelligence technique enhanced with intelligent water drops for monthly reference evapotranspiration estimation. Agric Water Manage 244:106622

Ahmadi F, Mehdizadeh S, Mohammadi B (2021b) Development of bio-inspired- and wavelet-based hybrid models for reconnaissance drought index modeling. Water Resour Manage 35(12):4127–4147

Ahmadi F, Nazeri Tahroudi M, Mirabbasi R, Khalili K, Jhajharia D (2018) Spatiotemporal trend and abrupt change analysis of temperature in Iran. Meteorol Appl 25(2):314–321

Ali M, Prasad R, Xiang Y, Yaseen ZM (2020) Complete ensemble empirical mode decomposition hybridized with random forest and kernel ridge regression model for monthly rainfall forecasts. J Hydrol 584:124647

Apaydin H, Feizi H, Sattari MT, Colak MS, Shamshirband S, Chau KW (2020) Comparative analysis of recurrent neural network architectures for reservoir inflow forecasting. Water 12(5):1500

Bednarik M, Magulová B, Matys M, Marschalko M (2010) Landslide susceptibility assessment of the Kral'ovany-Liptovský Mikuláš railway case study. Phys Chem Earth Parts a/b/c 35(3–5):162–171

Booker DJ, Snelder TH (2012) Comparing methods for estimating flow duration curves at ungauged sites. J Hydrol 434:78–94

Breiman L (2001) Random Forests. Mach Learn 45(1):5–32

Chen BH, Wang XZ, Yang SH, McGreavy C (1999) Application of wavelets and neural networks to diagnostic system development, 1, feature extraction. Comput Chem Eng 23(7):899–906

Chen S, Dong S (2020) A sequential structure for water inflow forecasting in coal mines integrating feature selection and multi-objective optimization. IEEE Access 8:183619–183632

Chu TY, Huang WC (2020) Application of empirical mode decomposition method to synthesize flow data: A case study of Hushan Reservoir in Taiwan. Water 12(4):927

Darbandsari P, Coulibaly P (2020) Introducing entropy-based Bayesian model averaging for streamflow forecast. J Hydrol 591:125577

Fathian F, Mehdizadeh S, Sales AK, Safari MJS (2019) Hybrid models to improve the monthly river flow prediction: Integrating artificial intelligence and non-linear time series models. J Hydrol 575:1200–1213

Friedman J, Hastie T, Tibshirani R (2001) The elements of statistical learning (Vol. 1, No. 10). Springer series in statisti, New York

Ghorbani MA, Deo RC, Kim S, Kashani MH, Karimi V, Izadkhah M (2020) Development and evaluation of the cascade correlation neural network and the random forest models for river stage and river flow prediction in Australia. Soft Comput 24:12079–12090

Herath HMVV, Chadalawada J, Babovic V (2020) Hydrologically informed machine learning for rainfall-runoff modelling: Towards distributed modelling. Hydrol Earth Syst Sci Discussions, pp 1–42

Huang NE, Shen Z, Long SR, Wu MC, Shih HH, Zheng Q, Yen NC, Tong CC, Liu H (1998) The empirical mode decomposition and Hilbert spectrum for nonlinear and nonstationary time series analysis. Procee Royal Soci A 545(1971):903–995

Hussain D, Khan AA (2020) Machine learning techniques for monthly river flow forecasting of Hunza River, Pakistan. Earth Sci Inform 13:939–949

Khalili K, Nazeri Tahoudi M, Mirabbasi R, Ahmadi F (2016) Investigation of spatial and temporal variability of precipitation in Iran over the last half century. Stoch Environ Res Risk Assess 30(4):1205–1221

Lee D, Kim H, Jung I, Yoon J (202) Monthly reservoir inflow forecasting for dry period using teleconnection indices: a statistical ensemble approach. Appl Sci 10(10):3470.

Mallat SG (2009) A theory for multiresolution signal decomposition: the wavelet representation. In Fundamental Papers in Wavelet Theory (pp. 494–513). Princeton University Press.

Mehdizadeh S (2020) Using AR, MA, and ARMA time series models to improve the performance of MARS and KNN approaches in monthly precipitation modeling under limited climatic data. Water Resour Manage 34(1):263–282

Mehdizadeh S, Ahmadi F, Mehr AD, Safari MJS (2020a) Drought modeling using classic time series and hybrid wavelet-gene expression programming models. J Hydrol 587:125017

Mehdizadeh S, Ahmadi F, Sales AK (2020b) Modelling daily soil temperature at different depths via the classical and hybrid models. Meteorol Appl 27(4):e1941

Mehdizadeh S, Sales AK (2018) A comparative study of autoregressive, autoregressive moving average, gene expression programming and Bayesian networks for estimating monthly streamflow. Water Resour Manage 32(9):3001–3022

Mehr AD, Kahya E, Bagheri F, Deliktas E (2014) Successive-station monthly streamflow prediction using neuro-wavelet technique. Earth Sci Inform 7(4):217–229

Mehr AD, Nourani V, Hrnjica B, Molajou A (2017) A binary genetic programing model for teleconnection identification between global sea surface temperature and local maximum monthly rainfall events. J Hydrol 555:397–406

Misiti M, Misiti Y, Oppenheim G, Poggi JM (1996) Wavelet Toolbox for Use with Matlab. The Mathworks Inc, Natick, Massachusetts, USA

Mohammadi B, Ahmadi F, Mehdizadeh S, Guan Y, Pham QB, Linh NTT, Tri DQ (2020) Developing novel robust models to improve the accuracy of daily streamflow modeling. Water Resour Manage 34(10):3387–3409

Molajou A, Nourani V, Afshar A, Khosravi M, Brysiewicz A (2021) Optimal design and feature selection by genetic algorithm for emotional artificial neural network (EANN) in rainfall-runoff modeling. Water Resour Manage. https://doi.org/10.1007/s11269-021-02818-2

Nayak PC, Sudheer KP, Rangan DM, Ramasastri KS (2004) A neuro-fuzzy computing technique for modeling hydrological time series. J Hydrol 291(1–2):52–66

Nazir HM, Hussain I, Faisal M, Shoukry AM, Sharkawy MAW, Al-Deek FF, Ismail M (2020) Dependence structure analysis of multisite river inflow data using vine copula-CEEMDAN based hybrid model. PeerJ 8:e10285

Nourani V, Baghanam AH, Adamowski J, Kisi O (2014) Applications of hybrid wavelet–artificial intelligence models in hydrology: a review. J Hydrol 514:358–377

Nourani V, Molajou A, Uzelaltinbulat S, Sadikoglu F (2019) Emotional artificial neural networks (EANNs) for multi-step ahead prediction of monthly precipitation; case study: northern Cyprus. Theor Appl Climatol 138:1419–1434

Ouarda TB, Charron C, Mahdi S, Yousef LA (2021) Climate teleconnections, interannual variability, and evolution of the rainfall regime in a tropical Caribbean island: case study of Barbados. Theor Appl Climatol. https://doi.org/10.1007/s00704-021-03653-6

Pei-Yue L, Hui Q, Jian-Hua W (2010) Groundwater quality assessment based on improved water quality index in Pengyang County, Ningxia. Northwest China J Chem 7(S1):209–216

Pham LT, Luo L, Finley AO (2020) Evaluation of random forest for short-term daily streamflow forecast in rainfall and snowmelt driven watersheds. Hydrol Earth Syst Sci Discussions, pp 1–33

Polikar R (1999) Fundamental concepts and overview of the wavelet theory: the wavelet tutorial–part I.

Pour SH, Abd Wahab AK, Shahid S (2020) Spatiotemporal changes in precipitation indicators related to bioclimate in Iran. Theor Appl Climatol 141(1):99–115

Rahmani-Rezaeieh A, Mohammadi M, Mehr AD (2020) Ensemble gene expression programming: a new approach for evolution of parsimonious streamflow forecasting model. Theor Appl Climatol 139(1–2):549–564

Ray SN, Chattopadhyay S (2021) Analyzing surface air temperature and rainfall in univariate framework, quantifying uncertainty through Shannon entropy and prediction through artificial neural network. Earth Sci Inform 14(1):485–503

Roy DK (2021) Long short-term memory networks to predict one-step ahead reference evapotranspiration in a subtropical climatic zone. Environ Proc 8(2):911–941

Salehi S, Dehghani M, Mortazavi SM, Singh VP (2020) Trend analysis and change point detection of seasonal and annual precipitation in Iran. Int J Climatol 40(1):308–323

Santos CA, Freire PK, Silva RMD, Akrami SA (2019) Hybrid wavelet neural network approach for daily inflow forecasting using tropical rainfall measuring mission data. J Hydrol Eng 24(2):04018062

Saray MH, Eslamian SS, Klöve B, Gohari A (2020) Regionalization of potential evapotranspiration using a modified region of influence. Theor Appl Climatol 140(1):115–127

Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27(3):379–423

Sharafi S, Karim NM (2020) Investigating trend changes of annual mean temperature and precipitation in Iran. Arab J Geosci 13(16):1–11

Shataee S, Kalbi S, Fallah A, Pelz D (2012) Forest attribute imputation using machine-learning methods and ASTER data: comparison of k-NN, SVR and random forest regression algorithms. Int J Remote Sens 33(19):6254–6280

Singh VP (2018) Hydrologic modeling: progress and future directions. Geosci Lett 5(1):1–18

Tang T, Liang Z, Hu Y, Li B, Wang J (2020) Research on flood forecasting based on flood hydrograph generalization and random forest in Qiushui River basin. China J Hydroinform 22(6):1588–1602

Torres ME, Colominas MA, Schlotthauer G, Flandrin P (2011) A complete ensemble empirical mode decomposition with adaptive noise. In 2011 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 4144–4147). IEEE.

Vidyarthi VK, Jain A, Chourasiya S (2020) Modeling rainfall-runoff process using artificial neural network with emphasis on parameter sensitivity. Model Earth Syst Environ 6:2177–2188

Wang J, Wang X, hui Lei X, Wang H, hua Zhang X, jun You J, lian Liu X (2020) Teleconnection analysis of monthly streamflow using ensemble empirical mode decomposition. J Hydrol 582:124411

Wu Z, Huang NE (2009) Ensemble empirical mode decomposition: a noise-assisted data analysis method. Adv Adapt Data Anal 1(01):1–41