



Streamflow prediction using a hybrid methodology based on variational mode decomposition (VMD) and machine learning approaches

Farshad Ahmadi¹ · Mansour Tohidi² · Meysam Sadriazade³

Received: 13 February 2023 / Accepted: 8 May 2023 / Published online: 17 May 2023
© The Author(s) 2023

Abstract

The optimal management of water resources depends on accurate and reliable streamflow prediction. Therefore, researchers have become interested in the development of hybrid approaches in recent years to enhance the performance of modeling techniques for predicting hydrological variables. In this study, hybrid models based on variational mode decomposition (VMD) and machine learning models such as random forest (RF) and *K*-star algorithm (KS) were developed to improve the accuracy of streamflow forecasting. The monthly data obtained between 1956 and 2017 at the Iranian Bibijan Abad station on the Zohreh River were used for this purpose. The streamflow data were initially decomposed into intrinsic modes functions (IMFs) using the VMD approach up to level eight to develop the hybrid models. The following step models the IMFs obtained by the VMD approach using the RF and KS methods. The ensemble forecasting result is then accomplished by adding the IMFs' forecasting outputs. Other hybrid models, such as EDM-RF, EMD-KS, CEEMD-RF, and CEEMD-KS, were also developed in this research in order to assess the performance of VMD-RF and VMD-KS hybrid models. The findings demonstrated that data preprocessing enhanced standalone models' performance, and those hybrid models developed based on VMD performed best in terms of increasing the accuracy of monthly streamflow predictions. The VMD-RF model is proposed as a superior method based on root mean square error (RMSE = 13.79), mean absolute error (MAE = 8.35), and Kling–Gupta (KGE = 0.89) indices.

Keywords Central frequency · Decomposition level · Noise · Sub-series

Introduction

Streamflow forecasting is an important issue for water resource management, since it is necessary to develop flood warning systems, the optimal operation of dam reservoirs, and hydropower generation (Lin et al. 2021). Various equations and models for forecasting streamflow have been developed so far, including conceptual rainfall-runoff methods, time-series models, and hybrid techniques, all of which can be classified into two categories: physical models and data-driven models (Kartzert et al. 2018; Lin et al. 2021). Physical models are created from field data and are based on pre-existing mathematical correlations between various hydrological processes. Soil texture, land use, and vegetation cover in the basin are only a few data points (Beven 2020). As can be recognized in physical models, there are several aspects that influence the model's outputs. Therefore, modeling is quite expensive. Furthermore, one of these models' shortcomings is the requirement for extensive hydrological

✉ Farshad Ahmadi
f.ahmadi@scu.ac.ir

Mansour Tohidi
mt_tohidi_ar@yahoo.com

Meysam Sadriazade
m.sadriazade@gmail.com

¹ Department of Hydrology & Water Resources Engineering, Shahid Chamran University of Ahvaz, Ahvaz, Iran

² Department of Civil Engineering, Ahvaz Branch, Islamic Azad University, Ahvaz, Iran

³ Department of Civil Engineering-Water Resources Engineering and Management, Shoushtar Branch, Islamic Azad University, Shoushtar, Iran

data. In other words, the presence of high uncertainty in the data required by physical models can lead to inaccurate predictions of complicated variables like streamflow (Lin et al. 2021; Biondi and De Luca 2013).

In recent years, data-driven strategies have made tremendous progress in overcoming the limitations of physical models. These techniques require less inputs, are less parameter-dependent, simpler to debug, and practical (Meng et al. 2021; Fang et al. 2019; Chen et al. 2018). Time-series models are one type of data-driven strategy that has been widely utilized to forecast river flow in many regions of the world (Adnan et al. 2017; Wen et al. 2019; Mehdizadeh et al. 2019). Autoregressive (AR), moving average (MA), autoregressive moving average (ARMA), and autoregressive vector (VAR) models are popular types of these models. In streamflow forecasting, time-series models imply a linear connection between inputs and outputs, which is harder to accomplish. In addition, the stationarity of the recorded data is a fundamental assumption in their implementation. Due to the trend, climate change, and other elements impacting streamflow, the stationarity criteria are critical to satisfying (Ghimire et al. 2021).

Unlike time-series models, machine learning approaches are better at recognizing complex relationships among the components of phenomenon and have been proposed as a feasible alternative method for estimating streamflow. (Meng et al. 2021; Lin et al. 2021). The following are some of the machine learning algorithms that have been frequently employed in hydrological modeling. Support vector machine (SVM) (Essam et al. 2022; Samantaray et al. 2022), gene expression programming (GEP) (Mehdizadeh et al. 2018; Esmaili-Gisavandani et al. 2021), Bayesian regression (BR) (Wagena et al. 2020; Achieng and Zhu 2019), random forests (RF) (Ahmadi et al. 2022) and the K-star algorithm (KS) (Salih et al. 2020).

Human activities and climate changes are among the challenges that have a significant impact on machine learning methods performance with increasing non-stationarity in hydrological data (Meng et al. 2021). Machine learning models based on fundamental decomposition approaches have been developed to overcome this challenge. Streamflow series could be divided into multiple sub-series using these approaches. In this case, the characteristics of the data, including periodicity, trends and noises, are identified and directly provided to the models.

The performance of models is considerably enhanced when sub-series are added (Yilmaz et al. 2022; Ahmadi et al. 2022; Meng et al. 2021; Fang et al. 2019). Wavelet transform is a strong technique for detecting non-stationarity in dataset. By dividing the initial series into high and low frequencies, this strategy allows the model to identify data features. One of the most significant drawbacks of the wavelet transform approach is that it is dependent on the mother wavelet

function, which means that each mother wavelet function generates different outputs for a series, and determining the appropriate mother wavelet function necessitates a try-and-error procedure to identify (Ahmadi et al. 2022).

Empirical mode decomposition (EMD) is a signal analysis technique created by Huang et al. (1998) that has undergone several improvements. When oscillations with the same time scale are stored in various intrinsic mode functions (IMFs), EMD runs into the problem of mode mixing. As a result of the influence of mixing modes, the EMD approach is rendered useless. The ensemble empirical mode decomposition (EEMD) approach is used to compensate for EMD's flaws (Wu and Huang 2009). Trials with the addition of white noise to the signal reveal that the EEMD approach scales better than the EMD, and the resultant IMF does not exhibit linkages with other IMFs. The EEMD approach contains flaws, such as residual noise in reconstructed data. Therefore, Torres et al. (2011) developed the complete ensemble empirical mode decomposition (CEEMD). White noise is introduced to the signal numerous times in this approach, with the exception that the first IMF is extracted while adding white noise, and the steps are repeated by adding noise again to extract the remaining IMFs.

For signal analysis, Dragomiretskiy and Zosso (2013) presented the variational mode decomposition (VMD) approach. The VMD can decompose complicated series with more efficiency (Lahmiri 2015). He et al. (2020) found that using VMD rather than EEMD increases the performance of the GBRT model in runoff prediction. Sun et al. (2022) reported that the hybrid models using VMD is outperformed to other signal analysis approaches in estimating the daily flow of the Han River in China.

The literature review indicates that the VMD approach has received less attention in streamflow forecast research than other signal analysis methods. Therefore, in the present study, we evaluate the capability of the VMD-RF and VMD-KS hybrid models in predicting monthly river flow. In addition, some hybrid models are developed, and the performance of VMD-based hybrid models is compared with EMD-based and CEEMD-based hybrid models.

Methodology

Variational mode decomposition (VMD)

VMD is an emerging and powerful decomposition model employed for decomposing a stationary and non-stationary time series into a specified number of intrinsic mode functions (IMFs). Hence, modes are compact around its central frequency. Therefore, it can be ensured that the IMFs can reconstruct the Initial data series with the utmost accuracy (Dragomiretskiy and Zosso 2013). A restricted variational

issue can be written as follows to achieve each mode and its central frequency (He et al. 2019):

$$\left\{ \begin{array}{l} \min_{\{u_k(t)\}, \{\omega_k(t)\}} \left\{ \sum_k \left\| \partial_t \left[\left(\delta(t) + \frac{j}{\pi t} \right) \otimes u_k(t) \right] e^{-j\omega_k t} \right\|_2^2 \right\} \\ s.t. \sum_k u_k(t) = f(t) \end{array} \right. \quad (1)$$

where t denotes time step, $\delta(t)$ is the Dirac distribution, $u_k(t)$ stands for the k th mode, $\omega_k(t)$ shows corresponding center frequency, and $f(t)$ indicates the t th data of the considered signal. Furthermore, the Hilbert transform of $u_k(t)$ is expressed by $(\delta(t) + \frac{j}{\pi t}) \otimes u_k(t)$, which can convert $u_k(t)$ into analytical data to create a one-sided frequency spectrum with only positive frequencies. Therefore, according to the index term $e^{-j\omega_k t}$, the spectrum of modes can be transfer to a baseband. To convert the given optimization problem into a non-objective optimization term, two aspects should be considered: Lagrange multipliers λ and the quadratic penalty parameter. The following is the equation for the augmented Lagrange function (Li et al. 2022):

$$L(u_k, \omega_k, \lambda) = \alpha \sum_{k=1}^K \left\| \partial_t \left[\left(\delta(t) + \frac{j}{\pi t} \right) \otimes u_k(t) \right] e^{-j\omega_k t} \right\|_2^2 + \left\| f(t) - \sum_{k=1}^K u_k(t) \right\|_2^2 + \left\langle \lambda(t) f(t) - \sum_{k=1}^K u_k(t) \right\rangle \quad (2)$$

where $\left\| f(t) - \sum_{k=1}^K u_k(t) \right\|_2^2$ denotes a quadratic penalty function to reduce convergence time. The alternate direction method of multipliers (ADMM) is an optimization procedure that allows $u_k(t)$ and $\omega_k(t)$ to be updated in two different points to complete the VMD method. As a result, the modified equations are known as (He et al. 2019):

$$u_k^{n+1} = \frac{f(\omega) - \sum_{i \neq k} u_i(\omega) + \frac{\lambda(\omega)}{2}}{1 + 2\alpha(\omega - \omega_k)} \quad (3)$$

$$\omega_k^{n+1} = \frac{\int_0^\infty \omega |u_k(\omega)|^2 d\omega}{\int_0^\infty |u_k(\omega)|^2 d\omega} \quad (4)$$

$$\lambda^{n+1}(\omega) = \lambda^n(\omega) + \eta \left(f(\omega) - \sum_k u_k^{n+1} \right) \quad (5)$$

where n is the number of repetition, $\lambda(\omega)$, $f(\omega)$ and $u_i(\omega)$ are the Fourier transform parameters, and η denotes the iterative

factor. More details about the mathematical logic and algorithm of VMD can be found in Dragomiretskiy and Zosso (2013).

Empirical mode decomposition (EMD)

The EMD decomposes natural, or synthetic, data into a number of distinct oscillating trends. Except for the last extracted subcomponent, all subcomponents are denoted as intrinsic mode functions (IMFs), with the remnant being the final subcomponent that depicts the general trend of the data (Barge and Sharif 2016). The initial data are equal to the summation of all IMFs and the final residual subcomponent. Individual sub-series are only designated as IMFs if they satisfy the next two conditions: i) The number of extreme (minima and maxima) must be equal to the number of zero crossings or differ by no more than one, and ii) the average of the cubic spline interpolated envelope, specified by the local minima and maxima must have a value equal or near to zero (Lee and Ouarda 2012). EMD can decompose any complex data set ($x(t)$) into a

confined number of IMFs through the sieving process. The sieving procedure for obtaining IFMs from a time series is described below.

- (1). To get the upper (lower) envelope, find all of the local extrema and use a smoothing approach to connect all of the local maxima (minimas). A cubic spline interpolation is a regularly used method (Lee and Ouarda 2012).
- (2). Use the following equation to find the mean value of the upper (e_{upper}) and lower (e_{lower}) envelopes:
- (3). $m(t) = \left[\frac{e_{upper}(t) + e_{lower}(t)}{2} \right]$ (6)
- (4). Take the difference ($d(t)$) between the computed mean value and the time series ($x(t)$):
- (5). $d(t) = x(t) - m(t)$ (7)
- (6). Examine $d(t)$ to check if it satisfies the IMF's specified conditions. If all of the conditions are met, $d(t)$ becomes the i th IMF, indicated as $C_i(t)$, and the residual $x(t)$ becomes the remaining residue $r(t)$, represented as $r(t) = x(t) - C_i(t)$, allowing the process to continue and the next IMF to be obtained. If the con-

ditions are not met, the process is repeated with $d(t)$ substituting $x(t)$ in the above mentioned equation and all steps applied to the remaining $d(t)$ until $C_i(t)$ is found.

- (7). Repeat stages 1–4 until either $r(t)$ becomes a uniform function or the number of extrema becomes less than one or equal to zero.
- (8). Finally, the initial time series obtained from summation of the extracted IMFs as:
- (9). $x(t) = \sum_{i=1}^n C_i(t) + r(t)$ (8)

Complete ensemble empirical mode decomposition (CEEMD)

CEEMD is known as an improved version of EMD. CEEMD technique was applied in this study to decompose the monthly streamflow data into a series of relatively stationary components. In this method, the Gaussian white noise is added to the source data during the decomposition process to provide a uniform reference frame, which is then eliminated by averaging the relevant IMFs and residue. The decomposition process in CEEMD is detailed here.

- (1). To the time series $x(t)$, add some white noise $\varepsilon(t)$ as follows:
- (2). $x'(t) = x(t) + \varepsilon(t)$ (9)
- (3). To extract the first component, decompose $x'(t)$ using the EMD approach.
- (4). Steps 1 and 2 should be repeated with a different number of substantiation. This procedure is frequently repeated with τ number of times with each realization being distinct.
- (5). Then, as shown below, compute the mean of all IMF_1 :
- (6). $IMF_1(t) = \frac{1}{\tau} \sum_{k=1}^{\tau} \psi_1(x(t) + \eta_k \cdot \varepsilon(t))$ (10)
- (7). where ψ_1 denotes the k th IMF estimation and indicates the amplitude adjustment required to obtain an adequate signal.
- (8). As a result, the first residual component, also known as the residual component, is as follows:
- (9). $R_1(t) = x(t) - IMF_1(t)$ (11)
- (10). Following that, IMF_2 can be derived from the original signal as below:
- (11). $IMF_2(t) = \frac{1}{\tau} \sum_{k=1}^{\tau} \psi_1[x(t) + \psi_1 \cdot \varepsilon(t)]$ (12)
- (12). To obtain the $IMF_{\beta+1}(t)$ factor, the above-mentioned process is repeated.

$$(13). IMF_{\beta+1}(t) = \frac{1}{\tau} \sum_{k=1}^{\tau} \psi_1 [x(t) + \psi_{\beta} \cdot \varepsilon(t)] \quad (13)$$

- (14). Finally, the residuals are averaged, revealing a gradual variation around the long-term average, which represents the general trend.

Random forests (RF)

The random forests (RF) algorithm is a popular machine learning algorithm from the field of artificial intelligence that belongs to supervised learning techniques. It can be used for classification and regression problems (Breiman 2001). It is based on the idea of group learning, which entails combining numerous classifiers to solve a complex problem and improve model performance. As the name implies, the random forest algorithm is a classifier that uses a variety of decision trees in different subsets of the data set (Cutler et al. 2012). Instead of using a decision tree, random forest forecasts each tree's prediction based on the majority of votes and uses the final result as the output. With more trees in the forest, accuracy improves and overfitting is avoided (Breiman 2001). The following is a description of how the RF algorithm works (Ali et al. 2020).

- Step 1: Start by randomly selecting samples from a data set.
- Step 2: For each instance, RF generates a decision tree. The forecast result from each decision tree is then obtained.
- Step 3—Each expected result is voted on in this step.
- Step 4—As the final prediction result, choose the maximum forecast result.

K-star algorithm (KS)

The KS algorithm is an instance-based classification method that performs classification based on similar training cases and achieves the desired results compared to machine learning algorithms (Hernández 2015). Unlike other data mining methods that classify based on the entropy-based distance function, this method uses the similarity function to estimate different variables. The essential principle of instance classification is that cases belong to the same category. This algorithm considers K points at random first, resulting in K clusters with random points at their centers. Following an evaluation of the data, the data that are closest to each center are assigned to the appropriate cluster. After that, an average

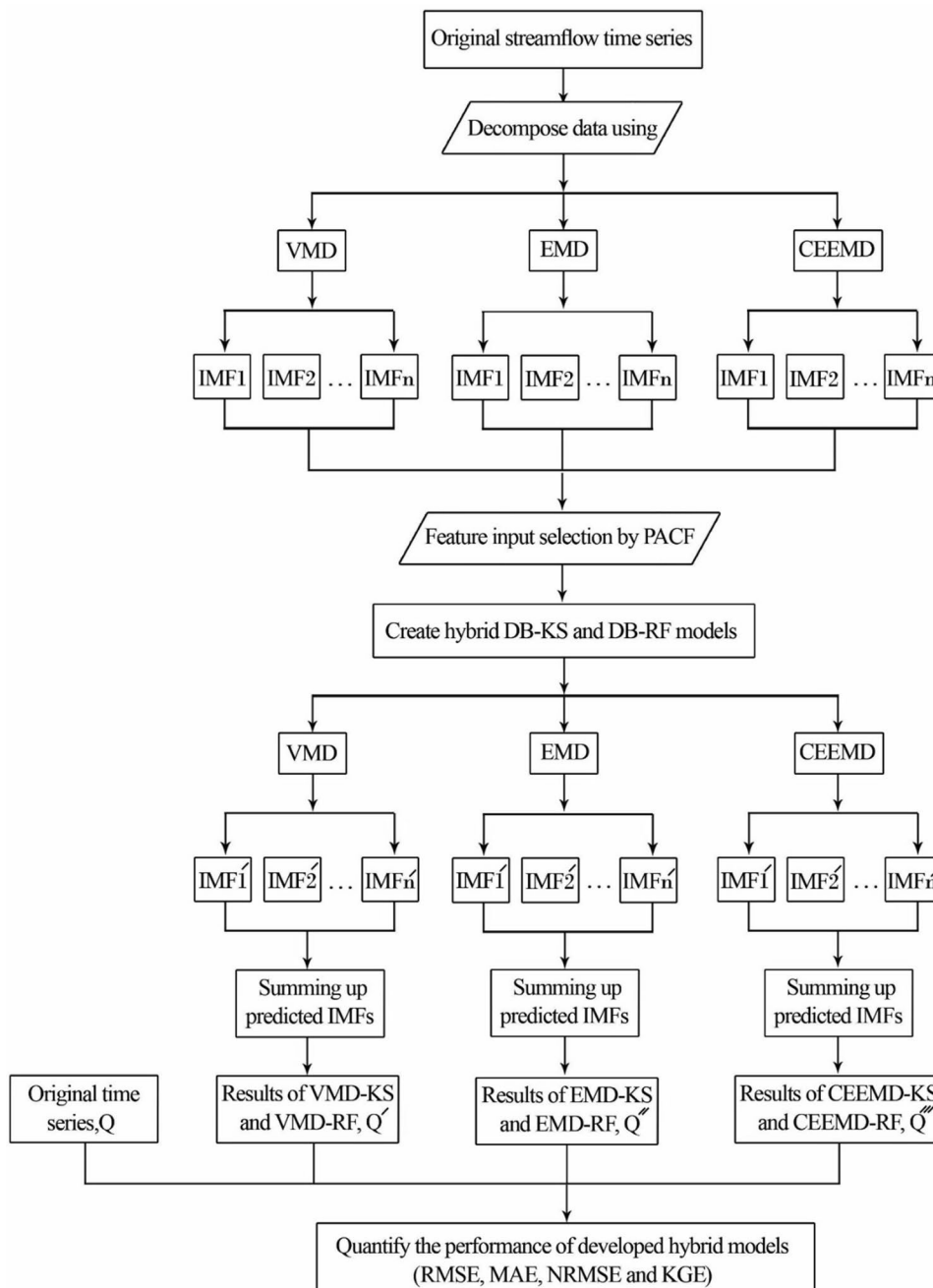
is taken from each cluster; this time, the averages are the centers of the categories, and any data that are farther from this average and closer to another cluster shift the cluster (Cleary and Trigg 1995). This cycle continues until all of the clusters are stable and data changes are no longer conceivable. Any digit can be used as K. This algorithm splits the data into the most similar categories and runs it multiple

times with varied beginning points to achieve the optimal outcome, resulting in better integrated cluster components (Cleary and Trigg 1995).

Model development

One of the most difficult aspects of modeling streamflow is selecting the appropriate inputs for the models. In order to

Fig. 1 Flowchart of the proposed hybrid models for streamflow modeling



establish suitable relationships between phenomena, models require inputs. As a result, the decomposed base (DB) method was employed to create hybrid models in this study. River flow data were initially decomposed into different IMFs using VMD, EMD, and CEEMD algorithms for this purpose. The following step is to choose the optimal delay for each IMF. The PACF approach is one of the statistical methods that has been utilized in many research to identify the optimum delay (He et al. 2019; He et al. 2020; Hu et al. 2021). Assume that the $IMF(t)$ is to be estimated and is the model's output; therefore, the $IMF(t-i)$ will be considered as an input by the PACF technique during the delay i if it exceeds the 95 percent confidence interval. After determining the appropriate delays for each IMF, DB-RF and DB-KS hybrid methods were implemented. Finally, all estimated IMFs were combined to obtain the predicted river flow from

the hybrid models. Figure 1 presents the development structure of the hybrid models in this study.

Data and study area

Monthly flow data from the Zohreh River were utilized to create hybrid models and assess their performance in this study. For this purpose, 10 operational hydrometric stations in the study area were surveyed. Meanwhile, the Bibijan hydrometric station was chosen to continue the investigation since it has a lengthy statistical period (1956 to 2017) and reliable data recording quality. The Zohreh River basin is located in Iran's southwest (Fig. 2). According to the kind of terrain, the elevation of the Zohreh River basin can be divided into three categories: i) the basin's high and mountainous sections in the north and north-east, ii) in the middle of the basin, there are submontane

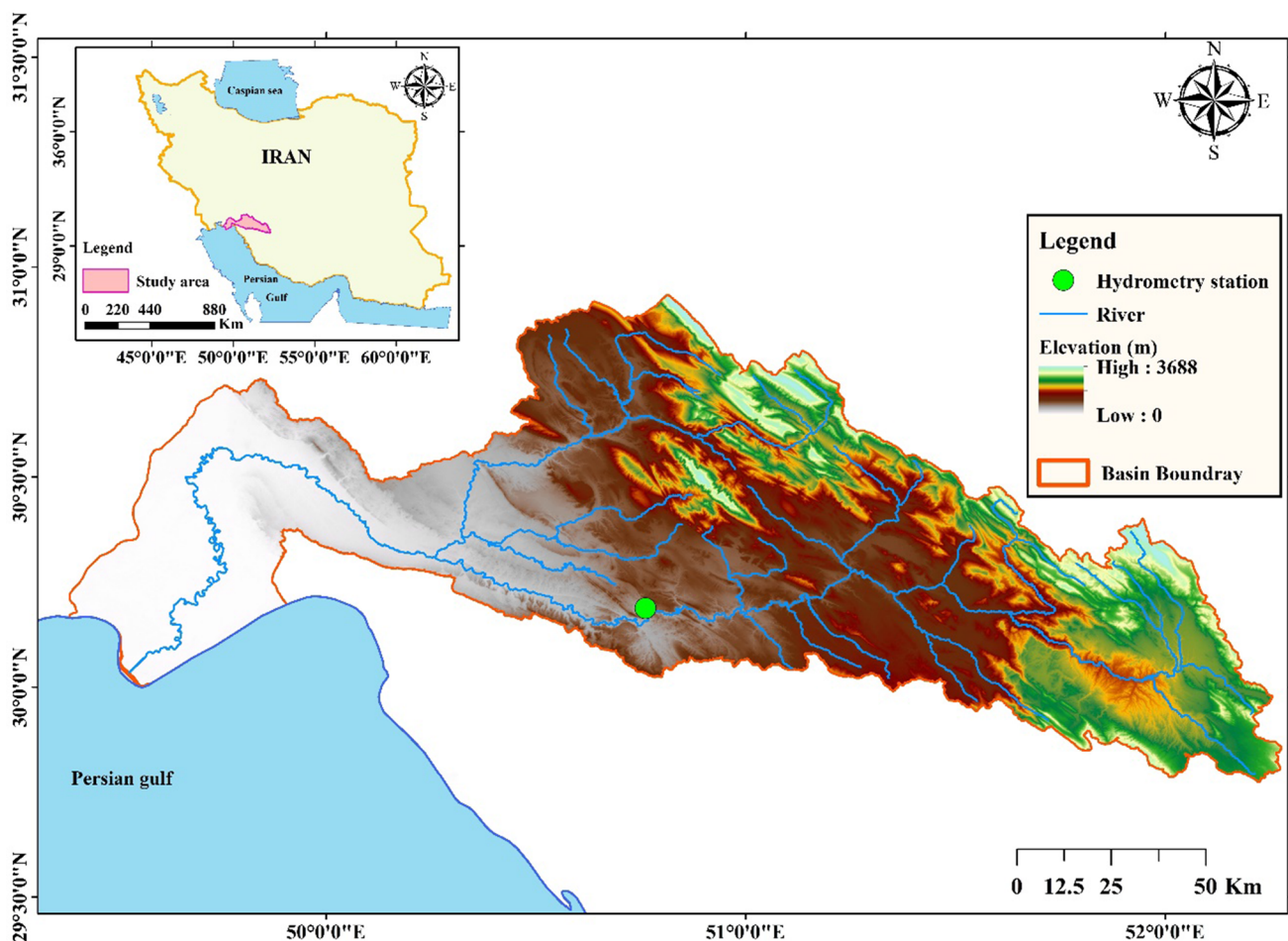


Fig. 2 Locations of the Zohreh River basin and considered hydrometric station in Iran

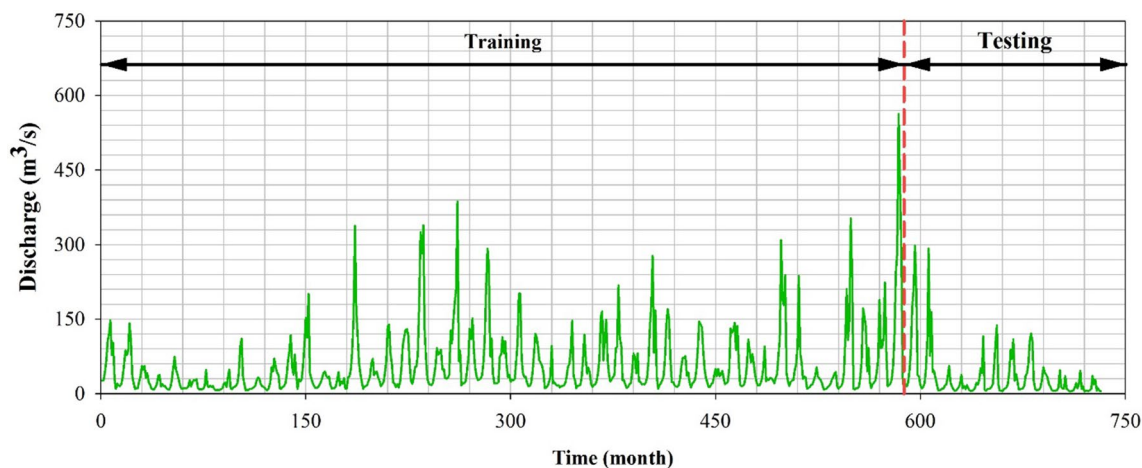


Fig. 3 Time series of the observed monthly streamflow data at Bibijan (BJ) station

sections and semi-high hilly lands and iii) in the southern and southwestern parts of the basin, there are flat plains and lowlands. The basin has a maximum height of 3688 m and a minimum height of zero meters. With an average annual volume of 1655 million cubic meters, the Zohreh River is one of the most water-rich rivers in southwestern Iran, and it plays a vital role in the region’s economy and agricultural operations. In order to estimate the monthly streamflow of the Zohreh River, the data were divided into training and testing phase. Out of the 732 months recorded at the Bibijan (BJ) hydrometric station, 588 months (80% of data) were used for training and 144 months (20%) for model testing. Figure 3 depicts the time-series graphs of the observed monthly streamflow data at BJ station.

Performance evaluation metrics

The root mean square error (RMSE), normalized root mean square error (NRMSE), mean absolute error (MAE), and Kling–Gupta efficiency (KGE) were implemented to assess the performance of the developed hybrid models in modeling the monthly streamflow. These four performance metrics are specified as follows:

$$RMSE = \left(\frac{\sum_{i=1}^n (O_i - P_i)^2}{n} \right)^{0.5} \tag{14}$$

$$NRMSE = 100 \times \frac{\sqrt{\frac{1}{n} \sum_{i=1}^n (O_i - P_i)^2}}{O_{\max} - O_{\min}} \tag{15}$$

$$MAE = \frac{\sum_{i=1}^n (O_i - P_i)^2}{n} \tag{16}$$

$$KGE = 1 - \sqrt{(CC - 1)^2 + \left(\frac{P_{SD}}{O_{SD}} - 1\right)^2 + \left(\frac{\bar{P}}{\bar{O}} - 1\right)^2} \tag{17}$$

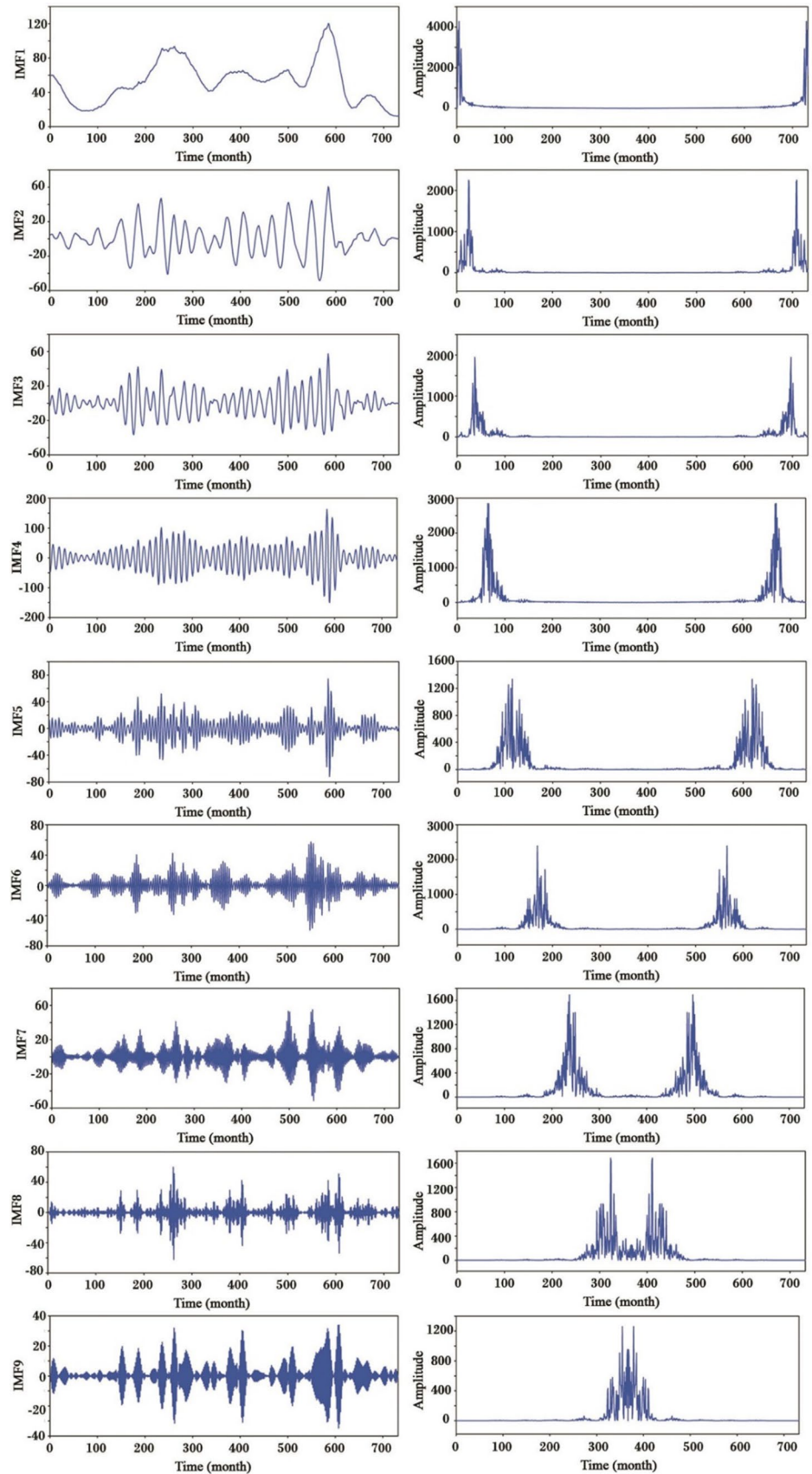
where O_i are observed values, P_i are estimated values, \bar{O} and O_{SD} represent the average and standard deviation of recorded monthly streamflow, \bar{P} and P_{SD} denote the average and standard deviation of estimated streamflow, O_{\max} and O_{\min} are maximum and minimum of observed data, CC is the correlation coefficient between observed and estimated streamflow and, n is the number of data. The developed hybrid model is selected as the most appropriate option for monthly streamflow prediction that has the lowest (highest) values of RMSE, NRMSE, and MAE (KGE).

Results

Data decomposition

The VMD, EMD, and CEEMD methods were used in this study to analyze streamflow time series. The proper decomposition level K and the penalty parameter for balancing the

Fig. 4 Results of monthly streamflow data decomposition using the VMD method



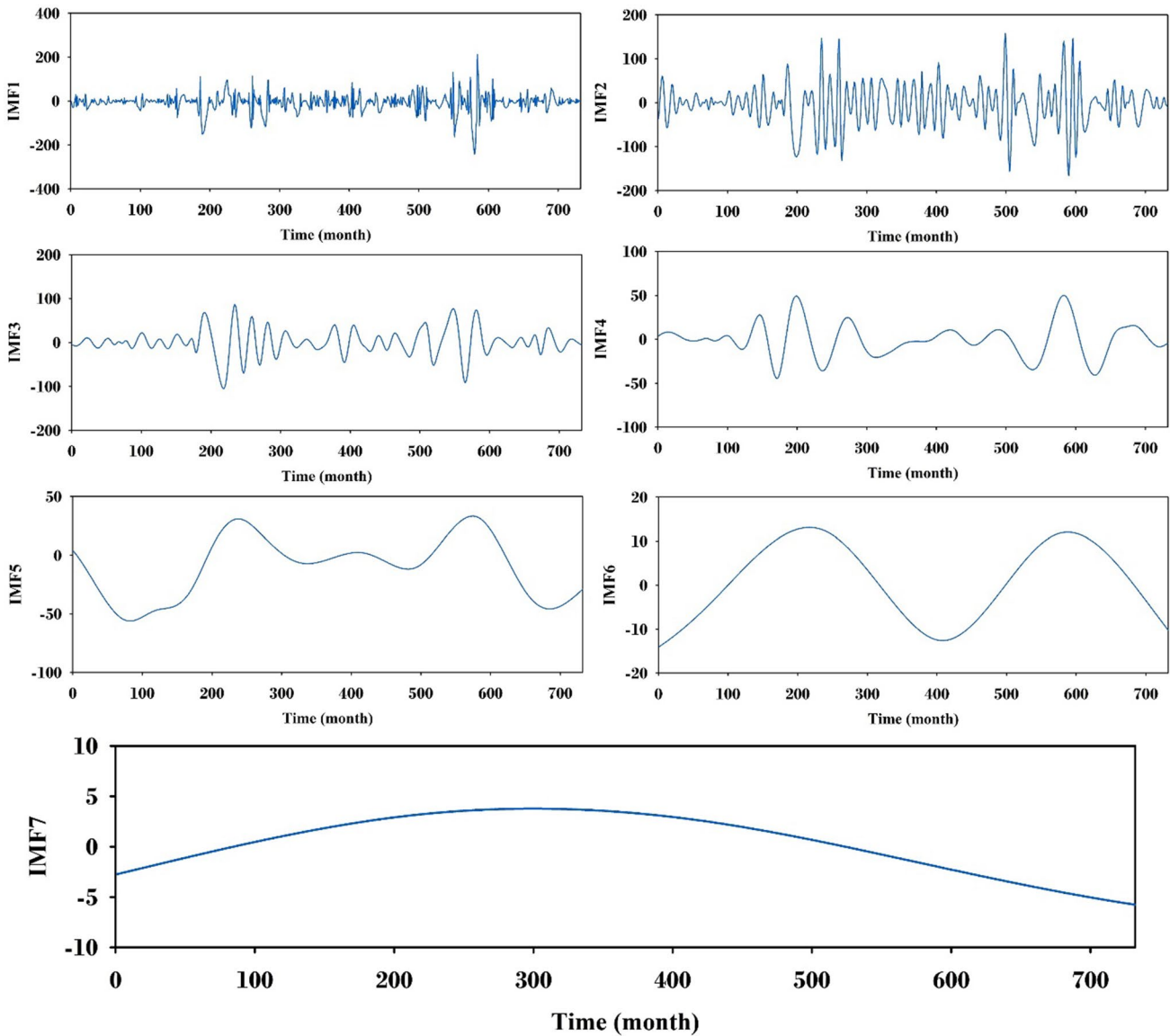


Fig. 5 Results of monthly streamflow data decomposition using the EMD method

data-fidelity requirement α are two crucial parameters that must be set for the VMD approach (Meng et al. 2021). The bandwidth of IMFs is specified by α . Wider bandwidth is produced by a higher value of α , which is discovered through trial and error (Li et al. 2022). In this study, various values between 10 and 2000 were investigated, and ultimately $\alpha = 1000$ produced the best results when monthly streamflow data were analyzed.

The number of VMD implementation steps is determined by choosing the appropriate decomposition level (K). The performance of the models is compromised if only a few

decomposition levels are chosen since not all the data from the time series are retrieved. However, if a high level of decomposition is used, the original time series will be over-decomposed, and several IMFs will share the same information. The suitable K was chosen in this study by decomposing streamflow data from two to nine levels using VMD. Then, for each IMF, central frequency diagrams were plotted. Figure 4 illustrates the outcomes of data analysis at various decomposition levels and their central frequencies. This figure shows that at level eight of decomposition, the center frequencies are close to one another, and there are no modal disruptions. Nevertheless, at the level of decomposition nine,

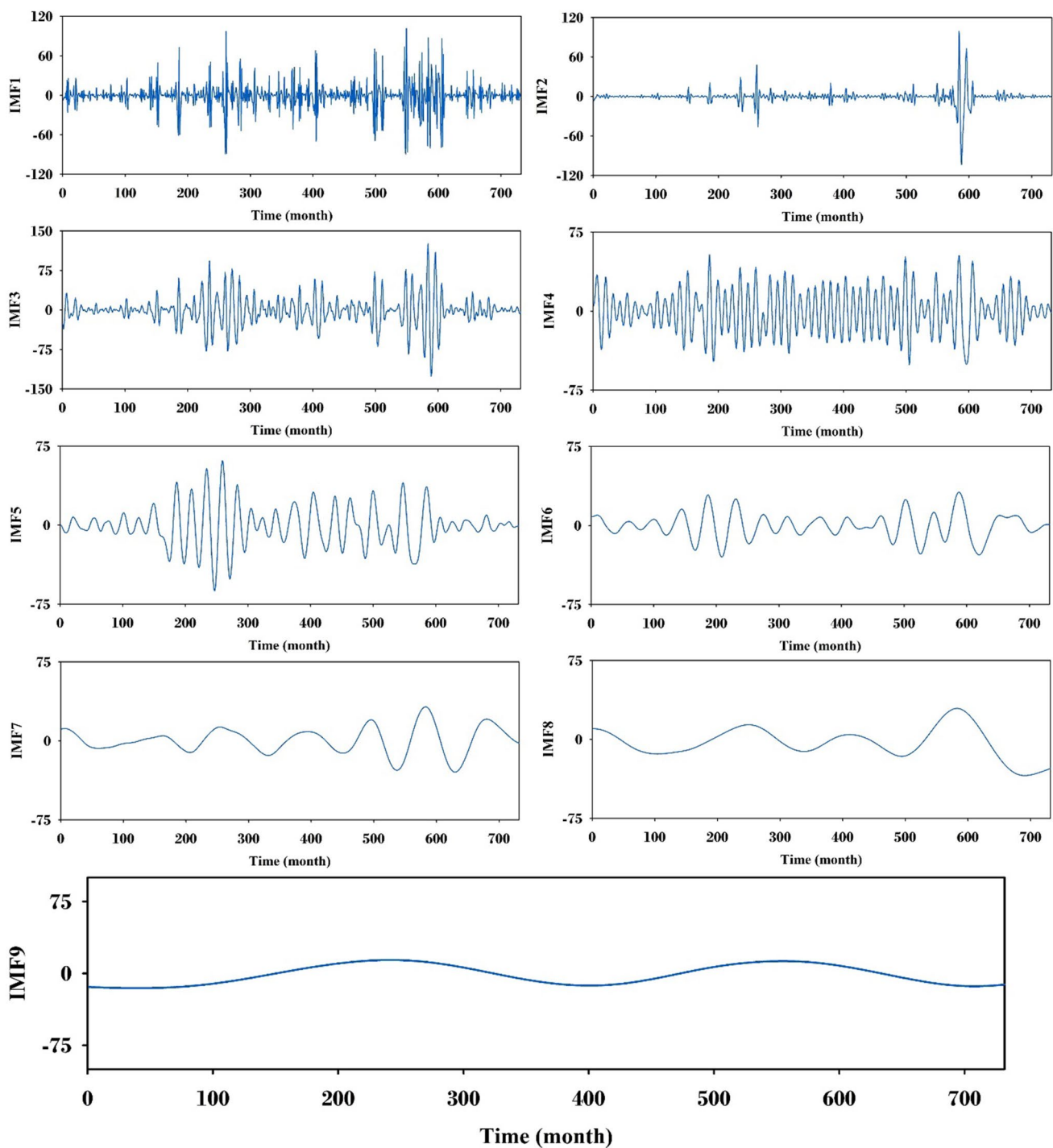


Fig. 6 Results of monthly streamflow data decomposition using the CEEMD method

mode mixing takes place. As a result, K should be set to eight ($K=8$) for assessing the monthly streamflow data of the Zohreh River basin. Figures 5 and 6 show the outcomes of data decomposition using EMD and CEEMD methods, respectively.

Input variables selection

Forecasting results are directly impacted by the kind and quality of input data. As a result, choosing the right approach for evaluating input data is crucial. The PACF

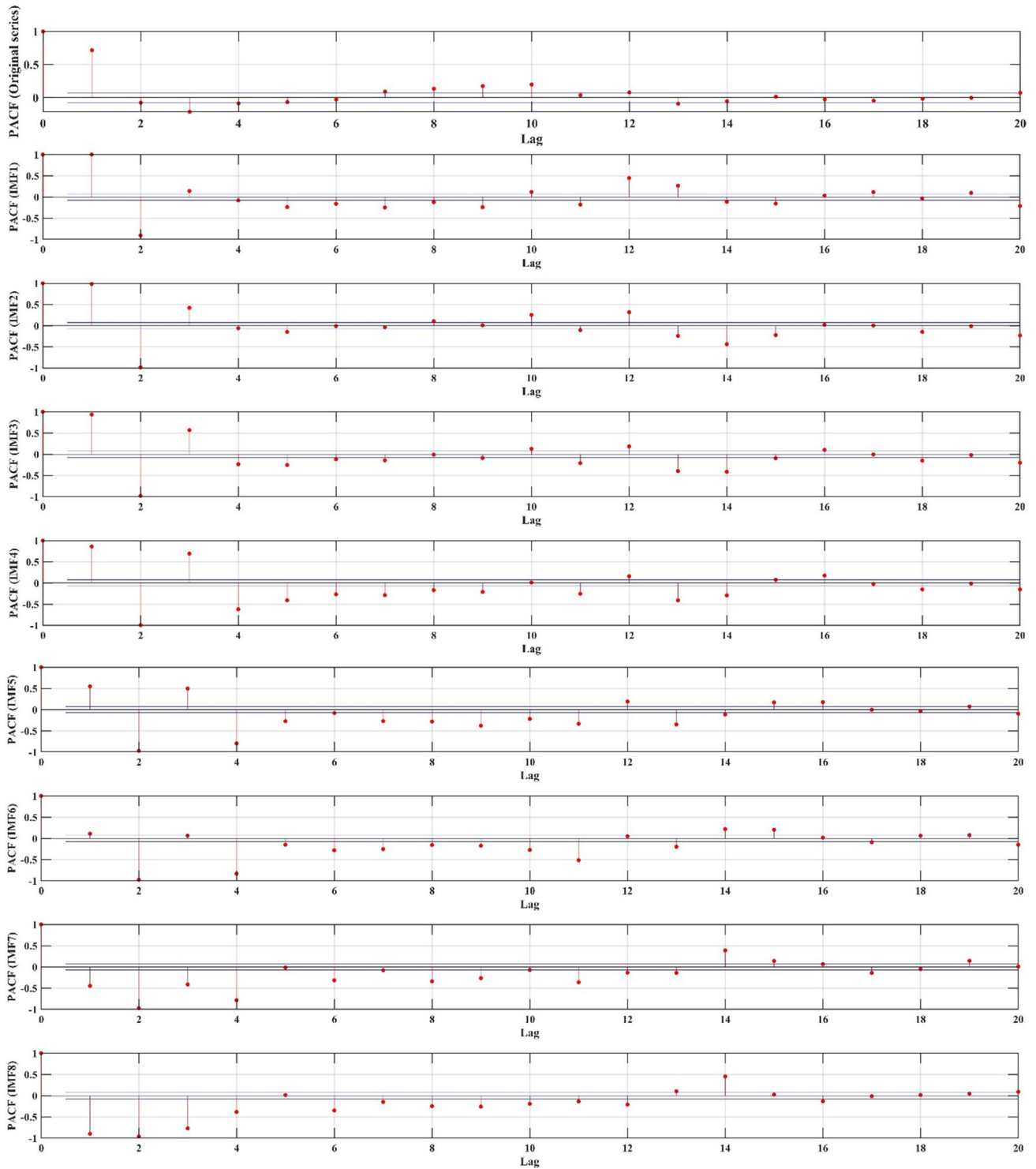


Fig. 7 PACFs of the original streamflow data and the IMF series obtained by VMD at the Bibijan station

approach was employed as a recognized technique to determine the optimum number of inputs for the RF and KS models (He et al. 2019). The PACF diagrams for the IMFs of the VMD approach are shown in Fig. 7. The suitable lag, as shown by this figure, begins at tow delays and go up

to seven lags. Also, PACF diagrams were drawn for IMFs obtained by CEEMD, and EMD methods and the appropriate number of inputs were selected. The final results of determining the number of optimal inputs for the original

Table 1 Input features and output targets for VMD, CEEMD, and EMD decomposition sequences of raw data and their decomposition sequences at the Bibijan station

Decomposer	Feature term	Input variables	Numbers of input	Outputs
–	Original series	$X(t-1), \dots, X(t-3)$	3	X(t)
VMD	IMF1	$IMF1(t-1), \dots, IMF1(t-3)$	3	IMF1(t)
	IMF2	$IMF2(t-1), \dots, IMF2(t-3)$	3	IMF2(t)
	IMF3	$IMF3(t-1), \dots, IMF3(t-7)$	7	IMF3(t)
	IMF4	$IMF4(t-1), IMF4(t-2)$	2	IMF4(t)
	IMF5	$IMF5(t-1), \dots, IMF5(t-4)$	4	IMF5(t)
	IMF6	$IMF6(t-1), IMF6(t-2)$	2	IMF6(t)
	IMF7	$IMF7(t-1), \dots, IMF7(t-4)$	4	IMF7(t)
	IMF8	$IMF8(t-1), \dots, IMF8(t-4)$	4	IMF8(t)
CEEMD	IMF1	$IMF1(t-1), \dots, IMF1(t-3)$	3	IMF1(t)
	IMF2	$IMF2(t-1), \dots, IMF2(t-6)$	6	IMF2(t)
	IMF3	$IMF3(t-1), \dots, IMF3(t-5)$	5	IMF3(t)
	IMF4	$IMF4(t-1), \dots, IMF4(t-6)$	6	IMF4(t)
	IMF5	$IMF5(t-1), \dots, IMF5(t-6)$	6	IMF5(t)
	IMF6	$IMF6(t-1), \dots, IMF6(t-6)$	6	IMF6(t)
	IMF7	$IMF7(t-1), \dots, IMF7(t-7)$	7	IMF7(t)
	IMF8	$IMF8(t-1), \dots, IMF8(t-7)$	7	IMF8(t)
EMD	IMF1	$IMF1(t-1)$	1	IMF1(t)
	IMF2	$IMF2(t-1), \dots, IMF2(t-8)$	8	IMF2(t)
	IMF3	$IMF3(t-1), \dots, IMF3(t-6)$	6	IMF3(t)
	IMF4	$IMF4(t-1), \dots, IMF4(t-6)$	6	IMF4(t)
	IMF5	$IMF5(t-1), \dots, IMF5(t-5)$	5	IMF5(t)
	IMF6	$IMF6(t-1), \dots, IMF6(t-9)$	9	IMF6(t)
	IMF7	$IMF7(t-1), \dots, IMF7(t-5)$	5	IMF7(t)

Table 2 The performance of various estimation models during the testing period for the Bibijan station in ZR basin. The best model, KS(3), is shown in bold

Model	RMSE (m^3/s)	MAE (m^3/s)	KGE	NRMSE
RF(1)	35.33	19.24	0.71	12.05
RF(2)	36.39	19.05	0.70	12.41
RF(3)	34.56	18.32	0.71	11.78
RF(4)	34.94	18.40	0.66	11.92
RF(5)	36.51	19.02	0.66	12.45
KS(1)	35.40	20.26	0.41	12.05
KS(2)	34.29	19.04	0.60	12.41
KS(3)	32.18	17.84	0.68	11.78
KS(4)	32.97	18.79	0.55	11.92
KS(5)	32.50	18.05	0.52	12.45

time series of monthly streamflow and IMFs obtained from the decomposition methods are presented in Table 1.

Results of standalone and hybrid models

The appropriate lag for the initial streamflow time-series modeling is equal to three, as shown in Table 1. The RF and KS models were subsequently examined with up to five delays. Table 2 displays the statistical results of the monthly streamflow forecast utilizing standalone models. The RMSE index in the test phase decreases with increasing lag in this table. In the fourth and fifth lags, the RMSE increases. These results demonstrate that the PACF approach performs satisfactorily in identifying the proper inputs. The best performance was achieved by the RF and KS models with the three lags input. The RMSE values for the KS and the RF method in the test phase may be compared, and it can be seen that they are 32.18 (m^3/s) and 34.56 (m^3/s), respectively. Additionally, when considering additional evaluation metrics, it is evident that the KS model performed better than the RF.

As can be seen from Table 2, the standalone approaches have a very large modeling error. In this study, a hybrid technique based on signal preprocessing was developed to increase the modeling accuracy. The data were initially decomposed for this purpose using VMD, EMD, and CEEMD methods. The RF and KS approaches were then used to model the obtained IMFs. The appropriate inputs were utilized to model IMFs using PACF. Tables 3 presents the statistical findings of the hybrid technique for estimating IMFs using RF and KS methods. This table shows that the RF model outperforms the KS model at predicting IMFs obtained using the VMD technique. For instance, in IMF4, the RF model’s RMSE value is 9.55, while the KS model’s RMSE value is 12.45. Additionally, IMFs four and five are the subjects of the most significant modeling error. This demonstrates that the performance of RF and KS models

Table 3 Statistical results of RF and KS models in estimating IMFs obtained from VMD, CEEMD, and EMD methods for monthly streamflow in the test phase

Decomposer	Feature term	RF				KS			
		RMSE	MAE	KGE	NRMSE	RMSE	MAE	KGE	NRMSE
VMD	IMF1	3.50	2.39	0.93	3.51	3.62	2.73	0.94	3.64
	IMF2	0.83	0.59	0.90	1.58	1.21	0.83	0.84	2.28
	IMF3	2.24	1.69	0.00	4.62	2.59	1.92	-0.01	5.33
	IMF4	9.55	5.73	-0.11	3.35	12.45	6.64	0.54	4.37
	IMF5	4.04	2.74	-0.29	4.14	4.76	2.86	-0.32	4.88
	IMF6	2.86	2.06	0.44	5.77	2.94	1.99	0.70	5.92
	IMF7	3.48	2.30	-0.33	4.61	4.30	2.59	0.03	5.70
	IMF8	2.78	1.68	0.28	2.66	5.15	2.40	0.16	4.92
CEEMD	IMF1	20.89	11.21	-0.23	12.65	16.36	9.78	-0.52	9.91
	IMF2	16.48	5.40	-3.37	10.43	14.54	5.02	-1.52	9.21
	IMF3	9.22	5.24	0.51	3.91	10.67	5.63	0.60	4.52
	IMF4	8.70	4.32	-0.08	8.87	10.04	4.89	-0.01	10.24
	IMF5	1.06	0.84	0.74	2.55	1.50	1.12	0.49	3.62
	IMF6	0.59	0.43	0.90	0.99	1.14	0.86	0.74	1.93
	IMF7	2.34	1.39	0.55	3.95	1.93	1.26	0.67	3.27
	IMF8	29.79	27.56	-0.24	47.21	11.58	9.49	-0.32	18.35
	IMF9	0.36	0.34	0.93	1.54	0.47	0.45	0.90	1.99
EMD	IMF1	31.78	20.07	-0.14	16.51	26.01	17.23	-0.06	13.51
	IMF2	15.65	10.62	0.65	5.01	18.45	12.80	0.74	5.90
	IMF3	1.92	1.41	0.84	3.07	3.14	2.36	0.67	5.02
	IMF4	0.79	0.65	0.81	0.92	1.92	1.53	0.81	2.24
	IMF5	0.90	0.75	0.97	1.20	1.65	1.45	0.94	2.20
	IMF6	0.26	0.25	0.93	1.16	0.19	0.17	0.95	0.84
	IMF7	1.83	1.53	-4.19	48.13	1.95	1.65	-6.19	51.22

is most affected by the information gleaned from these IMFs. Therefore, if the original data are not decomposed, the machine learning models will receive less usable data.

A radar diagram was utilized to more effectively assess how well the RF and KS models performed in estimating IMFs obtained using various decomposition techniques (Fig. 8). The RMSE index is depicted in Fig. 8. As can be seen, the RMSE index in estimating the IMFs of the VMD method varies between 0.83 and 12.45. However, the modeling error (RMSE index) for the EMD and CEEMD approaches varies from 0.26 to 31.78. Additionally, the RF model outperformed the KS technique in estimating the IMFs of the VMD method. In contrast, the KS model had a significant advantage over the RF method in estimating the IMFs obtained from EMD and CEEMD methods.

After modeling the IMFs using RF and KS models, the estimated series were combined, and the predicted monthly streamflow time series was reconstructed. Table 4 illustrates the statistical outcomes of hybrid and standalone models. This table shows that when compared to the standalone RF and KS models, the decomposition-based hybrid models are more accurate and less error-prone. Performance of standalone models and EMD-based method is pretty similar. In other words, using the EMD method for data analysis has

not improved the information that the RF and KS models receive and has had almost no positive impact on increasing prediction accuracy. This is the result of EMD’s improper data decomposition. However, the performance and accuracy of RF and KS models have significantly improved with the adoption of CEEMD, and VMD approaches. For instance, the VMD-RF technique reduces the RMSE of the RF model from 34.56 to 13.76 (m^3/s). At the same time, the RMSE is decreased from 32.18 to 14.76 (m^3/s) for the VMD-KS hybrid model. Compared to VMD-RF and VMD-KS models, the performance increase for CEEMD-RF and CEEMD-KS hybrid models was less. The CEEMD-RF technique can only cut the RMSE index by 1.2 (m^3/s), whereas the CEEMD-KS reduced RMSE by 5.37 (m^3/s).

Figure 9 shows the time series and scatter plots for the best case of standalone (KS(3) and RF(3)) and hybrid (VMD-RF and VMD-KS) models. This figure demonstrates that the standalone RF(3) model performs better than the KS(3) technique in estimating the maximum values. Additionally, a significant dispersion is visible around the one-to-one line in the scatter plots of both standalone models. Monthly streamflow estimation has been improved when using the hybrid technique and VMD data decomposition method. In the scatter plot, the observed streamflow and



Fig. 8 The radar graph of the RMSE index for estimating IMFs obtained from VMD, EMD, and CEEMD methods using RF and KS models in the test phase

Table 4 Statistical results of hybrid and standalone models in predicting monthly streamflow in the test phase

Approach	Model	RMSE (m3/s)	MAE (m3/s)	KGE	NRMSE
Standalone models	RF(3)	34.56	18.32	0.71	11.78
	KS(3)	32.18	17.84	0.55	10.97
Decomposed based hybrid models	VMD-RF	13.79	8.35	0.89	4.70
	EMD-RF	35.40	18.67	0.67	12.07
	CEEMD-RF	33.36	17.78	0.65	11.38
	VMD-KS	14.76	9.36	0.86	5.03
	EMD-KS	32.36	17.95	0.68	11.03
	CEEMD-KS	26.81	16.88	0.68	9.14

data from the VMD-RF model are close to the one-to-one line. Furthermore, hybrid VMD-RF and VMD-KS models are significantly better at estimating maximum values than single models. When the performance of the hybrid models is compared, it is clear that the VMD-KS technique

still performs poorly at estimating maximum values and is unable to make reliable predictions. The distribution of the predicted and observed data cannot be fully understood from time series and scatter plots. Additionally, statistical

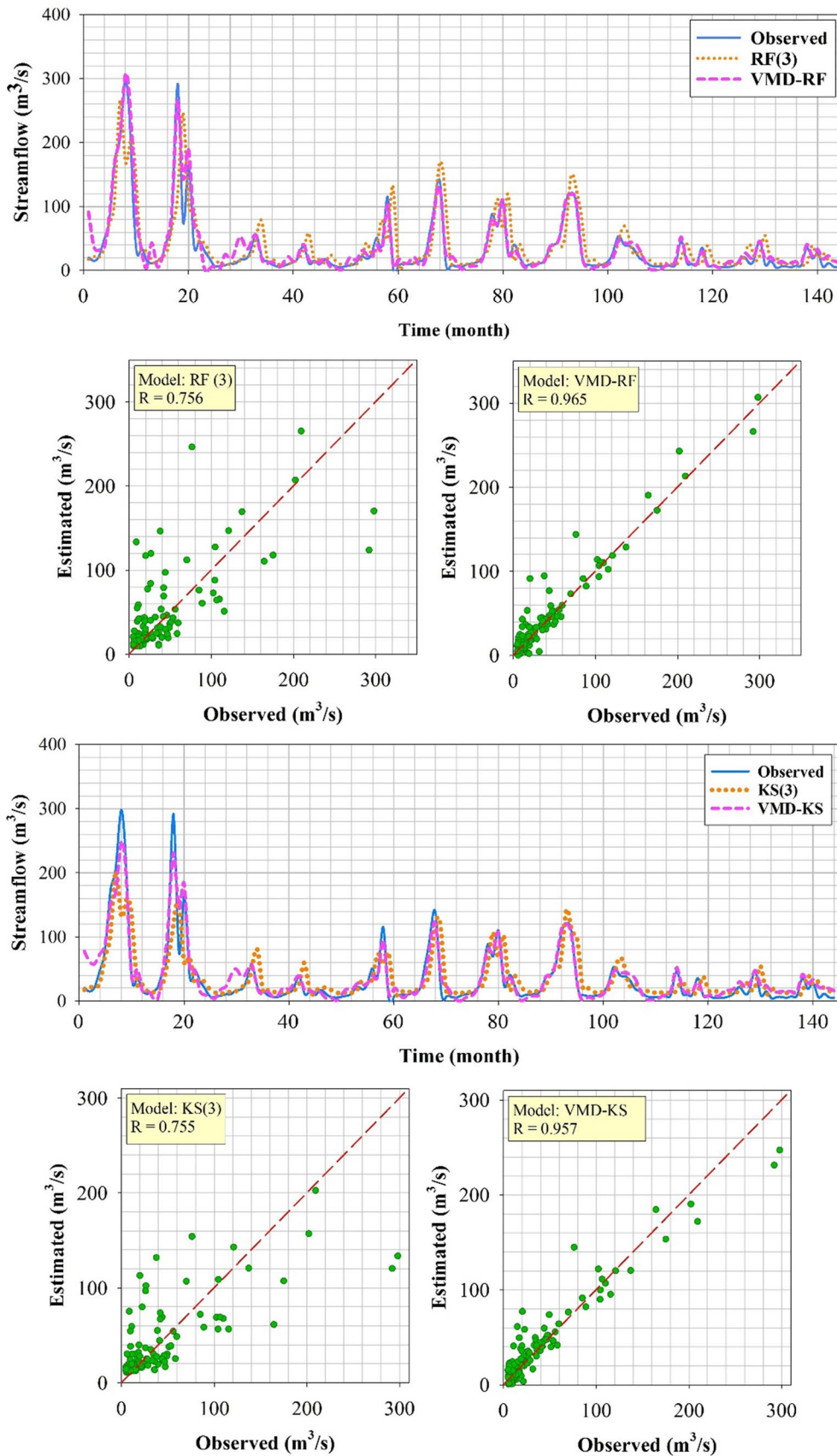


Fig. 9 Time-series and scatter plots of the observed against the modeled streamflow of the BJ station for the best hybrid and standalone models during the test phase

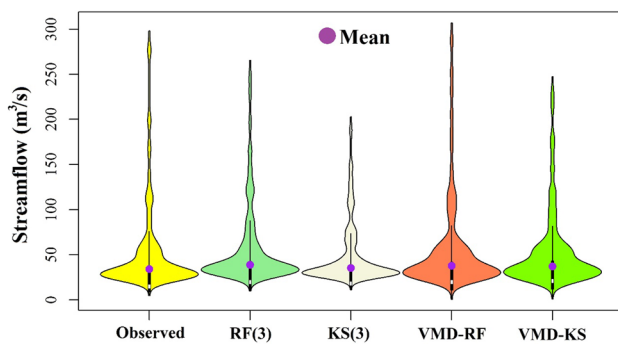


Fig. 10 Violin plot of the best hybrid and standalone models in the test phase to predict the monthly streamflow

indicators only represent model performance in numerical terms and offer no interpretation of the data's mean or variation. Therefore, it is essential to examine the distribution of predicted values in order to better comprehend the performance of the studied models. The performance of the models can be assessed by comparing the distribution of observed and estimated data.

One technique for displaying the distribution of data and covering the shortcomings of scatter plots and statistical indicators is using violin plots. Another type of box plot is the violin plot. The violin plot is used to demonstrate the data distribution and probability density, while the box plot only displays the lowest, maximum, mean, and quarters of the data. Figure 10 shows the violin diagram for the best standalone and hybrid models. According to this figure, it is observed that single models have estimated the maximum values in the test phase more negligible than the observational values, which is a drawback of single models. Additionally, the KS(3) model performs the worst when estimating maximal values. It is clear by comparing the VMD-RF and VMD-KS violins that the estimation of maximum values in both models has significantly improved. However, the VMD-RF approach has been far more effective in this regard. Thus, it can be inferred from the justifications given that the VMD-RF model is the best suitable technique for forecasting the monthly streamflow.

Discussion

In this research, hybrid models were constructed using the VMD, CEEMD, and EMD approaches. The outcomes demonstrated that RF and KS model performance could be enhanced by data preprocessing in general. Additionally, the findings indicate that, compared to other hybrid models

developed in this study, the VMD-RF and VMD-KS models were more effective in estimating monthly streamflow. Similar findings have been reported by He et al. (2020), Hu et al. (2021), and Meng et al. (2021). The studies mentioned above demonstrate how hybrid models developed using the VMD approach have improved maximum values prediction. He et al. (2019) developed hybrid models based on EMD, EEMD, and VMD to forecast the daily streamflow of the Jing River in China. They stated that while estimating daily streamflow, the VMD-based model performed better. In other words, it can be concluded that, regardless of time scale, combining data-driven models with the VMD technique can provide better results. The findings of this study demonstrate that, in comparison with EMD and CEEMD, the VMD technique offers data-driven models better information.

One of the most widely employed techniques for developing hybrid models is the wavelet transform (Rahmati et al. 2020; Song et al. 2021). Considering studies like Saraiva et al. (2021), Ahmadi et al. (2022), Momenh and Nourani (2022), and Yilmaz et al. (2022) as a comparison, it can be seen that wavelet transform is also quite effective in enhancing the performance of data-driven models. However, while applying the wavelet transform, the various mother wavelet functions must be compared, and the best ones for the data must be chosen. Due to the increase in processing steps, this is a significant restriction on using the wavelet transform method. In addition, selecting of the appropriate decomposition level in the wavelet transform method depends only on the number of data and no other variables are considered. In contrast, the VMD method does not require external functions to decompose the data, and its appropriate decomposition level can be easily identified based on the central frequency of the IMFs.

The non-stationary of the observed data is one of the substantial challenges that affects the performance of machine learning techniques. Decomposed based methods such as VMD can significantly reduce the non-stationary of the time series by smoothing the data (Lin et al. 2021). However, if the time-series decomposition approach is directly applied to the entire data set, the details related to the verification data will be included in the training data. In other words, using overall decomposition methods to train the model might be influenced by future knowledge, which would create an issue with backward induction and produce incorrect conclusions (Zhang et al. 2015). In the present study, this issue was resolved using an ensemble method for training models. The ensemble modeling method has been used and successfully tested in the investigations of He et al. (2019) and He et al. (2020), which is compatible with the findings of this study.

Conclusion

The RF and KS models have been used extensively in streamflow prediction. However, earlier research seldom took the data features into account while building the model, and prediction accuracy might still be increased. In this study, a hybrid monthly streamflow prediction approach based on data decomposition principles was presented to enhance the performance of monthly streamflow prediction, reduce the non-stationarity of the time series, and fully mine the embedded knowledge of the hydrological data. This method combines VMD, EMD, CEEMD, RF, and KS. The four phases for developing the hybrid models are as follows: (1) decomposing original monthly streamflow with VMD, EMD and CEEMD; (2) determination of the input variable using PACF; (3) developing hybrid decomposition based RF and KS models; and (4) ensemble predicted IMFs.

The findings demonstrated that the proposed model effectively extracts the characteristic data about the streamflow series while minimizing the non-stationarity and complexity of the original runoff series as well as its prediction difficulty. The proposed model uses VMD to decompose the original monthly streamflow series into some IMF components with low complexity and strong periodicity.

The VMD-RF and VMD-KS models are compared with four different models, namely RF, KS, EMD-RF, EMD-KS, CEEMD-RF, and CEEMD-KS. When comparing the studied models, hybrid techniques outperformed single models. EMD-based hybrid models, meanwhile, demonstrated poor performance. The VMD-RF model showed the best performance in predicting monthly streamflow. In comparison with the random forest model, this model was able to reduce the RMSE index by 60% and produce the most accurate forecast. In general, the accuracy has been much improved by combining the RF and KS models with the VMD approach.

The hybrid models developed in this study integrate data preprocessing techniques with machine learning methods to create a streamflow forecasting model that is more suited to being a practical and effective soft computing model to forecast streamflow series. It is recommended that VMD-based hybrid models be applied to forecast more hydrological parameters in subsequent studies. It is also advised to carry out more research on predicting streamflow using additional decomposed hydrological data. Future researches can also examine the impact of statistical period length on the VMD method's analytical accuracy.

Acknowledgements We are very grateful to the Research Council of Shahid Chamran University of Ahvaz and Ati Rah Consulting Engineers Company for their genuine support in conducting this research.

Funding This paper has not received any funding.

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

Declarations

Conflict of interest The authors declare no conflicts 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

- Achieng KO, Zhu J (2019) Application of Bayesian framework for evaluation of streamflow simulations using multiple climate models. *J Hydrol* 574:1110–1128
- Adnan RM, Yuan X, Kisi O, Yuan Y (2017) Streamflow forecasting of Astore River with seasonal autoregressive integrated moving average model. *Eur Sci J* 13(12):145–156
- Ahmadi F, Mehdizadeh S, Nourani V (2022) Improving the performance of random forest for estimating monthly reservoir inflow via complete ensemble empirical mode decomposition and wavelet analysis. *Stoch Environ Res Risk Assess* 36:2753–2768
- 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
- Barge JT, Sharif HO (2016) An ensemble empirical mode decomposition, self-organizing map and linear genetic programming approach for forecasting river streamflow. *Water* 8(6):247
- Beven K (2020) Deep learning hydrological processes and the uniqueness of place. *Hydrol Process* 34(16):3608–3613
- Biondi D, De Luca DL (2013) Performance assessment of a Bayesian Forecasting system (BFS) for real-time flood forecasting. *J Hydrol* 479:51–63
- Breiman L (2001) Random forests. *Mach Learn* 45(1):5–32
- Chen IT, Chang LC, Chang FJ (2018) Exploring the spatio-temporal interrelation between groundwater and surface water by using the self-organizing maps. *J Hydrol* 556:131–142
- JG Cleary LE Trigg 1995 K*: An instance-based learner using an entropic distance measure. In *Machine Learning Proceedings 1995* 108–114
- Cutler A, Cutler DR, Stevens JR (2012) Random forests. *Ensemble machine learning*. Springer, Boston, MA, pp 157–175
- Dragomiretskiy K, Zosso D (2013) Variational mode decomposition. *IEEE Trans Signal Process* 62(3):531–544
- Esmaeili-Gisavandani H, Lotfifard M, Sofla MSD, Ashrafzadeh A (2021) Improving the performance of rainfall-runoff models using the gene expression programming approach. *J Water Climate Change* 12(7):3308–3329
- Essam Y, Huang YF, Ng JL, Birima AH, Ahmed AN, El-Shafie A (2022) Predicting streamflow in Peninsular Malaysia using

- support vector machine and deep learning algorithms. *Sci Rep* 12(1):1–26
- Fang W, Huang S, Ren K, Huang Q, Huang G, Cheng G, Li K (2019) Examining the applicability of different sampling techniques in the development of decomposition-based streamflow forecasting models. *J Hydrol* 568:534–550
- Ghimire S, Yaseen ZM, Farooque AA, Deo RC, Zhang J, Tao X (2021) Streamflow prediction using an integrated methodology based on convolutional neural network and long short-term memory networks. *Sci Rep* 11(1):1–26
- He X, Luo J, Zuo G, Xie J (2019) Daily runoff forecasting using a hybrid model based on variational mode decomposition and deep neural networks. *Water Resour Manage* 33(4):1571–1590
- He X, Luo J, Li P, Zuo G, Xie J (2020) A hybrid model based on variational mode decomposition and gradient boosting regression tree for monthly runoff forecasting. *Water Resour Manage* 34(2):865–884
- Hernández DCT (2015) An experimental study of K* algorithm. *Int J Inform Eng Electron Bus* 7(2):14–19
- Hu H, Zhang J, Li T (2021) A novel hybrid decompose-ensemble strategy with a VMD-BPNN approach for daily streamflow estimating. *Water Resour Manage* 35(15):5119–5138
- Huang NE, Shen Z, Long SR, Wu MC, Shih HH, Zheng Q, Yen NC, Tung CC, Liu HH (1998). The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society of London. Series A: mathematical, physical and engineering sciences*, 454(1971), 903–995.
- Kaufmann M, Hernández DCT (2015) An experimental study of K* algorithm. *IJ Inform Eng Electronic Business* 7(2):14–19
- Kratzert F, Klotz D, Brenner C, Schulz K, Herrnegger M (2018) Rainfall–runoff modelling using long short-term memory (LSTM) networks. *Hydrol Earth Syst Sci* 22(11):6005–6022
- Lahmiri S (2015) Long memory in international financial markets trends and short movements during 2008 financial crisis based on variational mode decomposition and detrended fluctuation analysis. *Physica A* 437:130–138
- Lee T, & Ouarda TB (2012). Stochastic simulation of nonstationary oscillation hydroclimatic processes using empirical mode decomposition. *Water Resources Research*, 48(2)
- Li BJ, Sun GL, Liu Y, Wang WC, Huang XD (2022) Monthly runoff forecasting using variational mode decomposition coupled with gray wolf optimizer-based long short-term memory neural networks. *Water Resour Manage* 36(6):2095–2115
- Lin Y, Wang D, Wang G, Qiu J, Long K, Du Y, Xie H, Wei Z, Shang-guan W, Dai Y (2021) A hybrid deep learning algorithm and its application to streamflow prediction. *J Hydrol* 601:126636
- Mehdizadeh S, Kozekalani Sales A (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
- Mehdizadeh S, Fathian F, Adamowski JF (2019) Hybrid artificial intelligence-time series models for monthly streamflow modeling. *Appl Soft Comput* 80:873–887
- Meng E, Huang S, Huang Q, Fang W, Wang H, Leng G, Wang L, Liang H (2021) A Hybrid VMD-SVM model for practical streamflow prediction using an innovative input selection framework. *Water Resour Manage* 35(4):1321–1337
- Momeneh S, Nourani V (2022) Application of a novel technique of the multi-discrete wavelet transforms in hybrid with artificial neural network to forecast the daily and monthly streamflow. *Model Earth Syst Environ* 8(4):4629–4648
- Rahmati O, Darabi H, Panahi M, Kalantari Z, Naghibi SA, Ferreira CSS, Haghighi AT (2020) Development of novel hybridized models for urban flood susceptibility mapping. *Sci Rep* 10(1):1–19
- Salih SQ, Sharafati A, Khosravi K, Faris H, Kisi O, Tao H, Ali M, Yaseen ZM (2020) River suspended sediment load prediction based on river discharge information: application of newly developed data mining models. *Hydrol Sci J* 65(4):624–637
- Samantaray S, Das SS, Sahoo A, Satapathy DP (2022) Monthly runoff prediction at Baitarani river basin by support vector machine based on Salp swarm algorithm. *Ain Shams Engineering Journal* 13(5):101732
- Saraiva SV, de Oliveira Carvalho F, Santos CAG, Barreto LC, Freire PKDMM (2021) Daily streamflow forecasting in Sobradinho Reservoir using machine learning models coupled with wavelet transform and bootstrapping. *Appl Soft Comput* 102:107081
- Song Y, Shen Z, Wu P, Viscarra Rossel RA (2021) Wavelet geographically weighted regression for spectroscopic modelling of soil properties. *Sci Rep* 11(1):1–11
- Sun X, Zhang H, Wang J, Shi C, Hua D, Li J (2022) Ensemble streamflow forecasting based on variational mode decomposition and long short term memory. *Sci Rep* 12(1):1–19
- Torres, M. E., Colominas, M. A., 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.
- Wagena MB, Goering D, Collick AS, Bock E, Fuka DR, Buda A, Easton ZM (2020) Comparison of short-term streamflow forecasting using stochastic time series neural networks process-based and Bayesian models. *Environ Model Softw* 126:104669
- Wen X, Feng Q, Deo RC, Wu M, Yin Z, Yang L, Singh VP (2019) Two-phase extreme learning machines integrated with the complete ensemble empirical mode decomposition with adaptive noise algorithm for multi-scale runoff prediction problems. *J Hydrol* 570:167–184
- Wu Z, Huang NE (2009) Ensemble empirical mode decomposition: a noise-assisted data analysis method. *Adv Adapt Data Anal* 1(01):1–41
- Yilmaz M, Tosunoğlu F, Kaplan NH, Üneş F, Hanay YS (2022) Predicting monthly streamflow using artificial neural networks and wavelet neural networks models. *Model Earth Syst and Environ* 8(4):5547–5563
- Zhang X, Peng Y, Zhang C, Wang B (2015) Are hybrid models integrated with data preprocessing techniques suitable for monthly streamflow forecasting? Some experiment evidences. *J Hydrol* 530:137–152

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