Water Resources Management (2022) 36:253–272
https://doi.org/10.1007/s11269-021-03026-8

A Novel Hybrid Method for River Discharge Prediction

Maha Shabbir[1] · **Sohail Chand**[1] · **Farhat Iqbal**[2]

Received: 26 July 2021 / Accepted: 8 November 2021 / Published online: 27 November 2021

Abstract
Accurate prediction of river discharge is essential for the planning and management of water resources. This study proposes a novel hybrid method named HD-SKA by integrating two decomposition techniques (termed as HD) with support vector regression (SVR), K-nearest neighbor (KNN) and ARIMA models (combined as SKA) respectively. Firstly, the proposed method utilizes local mean decomposition (LMD) to decompose the original river discharge series into sub-series. Next, ensemble empirical mode decomposition (EEMD) is employed to further decompose the LMD-based sub-series into intrinsic mode functions. Further, the EEMD decomposed components are used as inputs in three data-driven models to predict river discharge respectively. The prediction of all components is then aggregated to obtain the results of HD-SVR, HD-KNN and HD-ARIMA models. The final prediction is obtained by taking the average prediction of these models. The proposed method is illustrated using five rivers in Indus Basin System. In five case studies, six models were built to compare the performance of the proposed HD-SKA model. The data analysis results show that the HD-SKA model performs better than all other considered models. The Diebold-Mariano test confirms the superiority of the proposed HD-SKA model over ARIMA, SVR, KNN, EEMD-ARIMA, EEMD-KNN, and EEMD-SVR models.

Keywords River discharge · Hybrid model · Indus basin · Decomposition · ARIMA

1 Introduction

The prediction of river discharge is essential for the planning and management of water resources. Hydrological data prediction provides critical information on the impending drought, heatwaves, or floods which brings devastation due to its delayed and inaccurate estimation (Sehgal et al. 2014). River discharge prediction has gained attention to handle extreme events as an outcome of climate changes. Different studies on river discharge have been studied in China (Wei et al. 2013), the USA (Meshram et al. 2019), Iran (Dehghani et al. 2021) and North America (Alizadeh et al. 2021). Thus, precise river discharge estimation is necessary to

✉ Sohail Chand
sohail.stat@pu.edu.pk

[1] College of Statistical and Actuarial Sciences, University of the Punjab, Lahore, Pakistan

[2] Department of Statistics, University of Balochistan, Quetta, Pakistan

construct warning systems for water management to deal with extreme events (Wang et al. 2018; Adnan et al. 2021).

Data-driven models are used to predict hydrological variables such as inflow, out-flow and discharge. These include both statistical and artificial intelligence (AI) models. Statistical models for time series include Moving Average (MA), Autoregressive (AR), ARMA and Autoregressive Integrated Moving Average (ARIMA) models and others (Bayazit 2015; Fashae et al. 2019; Bonakdari et al. 2020; Musarat et al. 2021; Aghelpour et al. 2021). However, these models are unable to capture time changes with sufficient accuracy because of linear analysis phenomena. Nowadays, AI models are widely used to model complex and nonlinear time series data. For example, K-nearest Neighbor (KNN), Support Vector Regression (SVR) and Artificial Neural Network (ANN) are popular models implemented in hydrological studies (Wu et al. 2008; Nikolic and Simonovic 2015; Poul et al. 2019; Sharghi et al. 2019; Riahi-Madvar et al. 2021). However, the drawback of AI models is that they do not incorporate noise and ignore the complex multi-scale structure of hydrological variables (Rezaie-Balf et al. 2019). Therefore, using a single model to predict river discharge is challenging. Researchers in the field of hydrology have introduced hybrid methods to improve the prediction accuracy of models. This study aims to overcome the limitations on the models application and proposes a new hybrid method that helps in the precise estimation of river discharge.

Many pre-processing techniques are integrated with data-driven models to create hybrid models for predicting hydrological variables. Some pre-processing methods such as Empirical Mode Decomposition (EMD), Local Mean Decomposition (LMD) and Ensemble EMD (EEMD) are combined with data-driven models to predict river discharge (Rezaie-Balf et al. 2019; Silva et al. 2021; Vidya and Janani 2021). In this study, the focus is on improving the prediction of river discharge by using hybrid pre-processing techniques.

In this study, a novel hybrid method is proposed by combining two pre-processing techniques with data-driven models. The two pre-processing techniques used are LMD and EEMD (abbreviated as HD) and three models applied are SVR, KNN, and ARIMA models (termed as SKA). Our proposed hybrid method uses LMD to decompose the original river discharge series into sub-series. Then, EEMD is applied to decompose the obtained sub-series into different components. The HD components are used as the inputs to SVR, KNN and ARIMA models. The predictions of these components are aggregated for HD-SVR, HD-KNN and HD-ARIMA models respectively. The final forecast of HD-SKA model is obtained by taking the average of these predictions. The effectiveness of the proposed hybrid method is illustrated on the discharge data of five rivers of Indus Basin System, Pakistan. The rivers include Kabul River, Kanshi River, Kunhar River, Jhelum River (Domel station) and Jhelum River (Chattar Kallas station). The superiority of the HD-SKA method is successfully presented on five data sets where six benchmark models are used to verify the performance of the proposed hybrid model. The HD-SKA method is novel based on the combination of hybrid decomposition with data-driven models which efficiently decomposes discharge series and capture its complex features with higher prediction accuracy.

2 Methodologies

2.1 Local Mean Decomposition (LMD)

Smith (2005) introduced LMD as a tool for analyzing the time–frequency of the electroencephalogram signal. LMD is a self-adaptive time–frequency approach that is useful in

capturing non-linear features of data (Liu and Han 2014; Huynh et al. 2021). LMD decomposes a signal into several product functions (*PFs*) that have a physical meaning. The time–frequency distribution of the signal is obtained by assembling the instantaneous frequency and instantaneous amplitude of all the obtained *PFs*. Given the original signal $y(t)$, it is decomposed using the following steps:

(i) Obtain the local extrema (n_i) from $y(t)$ and then compute the average (m_i) of the two successive extrema as follows:

$$m_i = \frac{1}{2}\left(n_i + n_{i+1}\right) \tag{1}$$

All the mean values ($m_i's$) are connected through the straight lines and the local mean function $m_{11}(t)$ is formed by moving averaging which smooth the $m_i's$.

(ii) The envelope estimate (a_i) is defined as follows:

$$a_i = \frac{1}{2}|n_i + n_{i+1}|. \tag{2}$$

The estimates of the local envelope are smoothed in similar ways as the local means to derive the envelope function $a_{11}(t)$.

(iii) The $m_{11}(t)$ is subtracted from $y(t)$ which forms the resulting signal as $h_{11}(t)$:

$$h_{11}(t) = y(t) - m_{11}(t). \tag{3}$$

(iv) $h_{11}(t)$ is amplitude demodulated by dividing it by envelope function $a_{11}(t)$ which forms $s_{11}(t)$ given as:

$$s_{11}(t) = \frac{h_{11}(t)}{a_{11}(t)}. \tag{4}$$

The function $s_{11}(t)$ is the purely frequency modulated signal known as the envelope function $a_{12}(t)$ of $s_{11}(t)$ should satisfy the condition $a_{12}(t) = 1$. If this condition is not satisfied, then $s_{11}(t)$ is regarded as the original signal and the above process is repeated until the purely frequency modulated signal ($s_{1n}(t)$) is derived that satisfies $-1 \le s_{1n}(t) \le 1$. Therefore,

$$\begin{cases} h_{11}(t) = y(t) - m_{11}(t) \\ h_{12}(t) = s_{11}(t) - m_{12}(t) \\ \quad\vdots \\ h_{1n}(t) = s_{1(n-1)}(t) - m_{1n}(t) \end{cases}, \tag{5}$$

$$\text{where} \begin{cases} s_{11}(t) = h_{11}(t)/a_{11}(t) \\ s_{12}(t) = h_{12}(t)/a_{12}(t) \\ \quad\vdots \\ s_{1n}(t) = h_{1n}(t)/a_{1n}(t) \end{cases}.$$

(v) The envelope signal $a_1(t)$ known as instantaneous amplitude function is obtained by taking the product of the functions of successive envelope estimate which are obtained during the iterative procedure discussed above.

$$a_1(t) = a_{11}(t)a_{12}(t)\ldots a_{1n}(t) = \prod_{l=1}^{n} a_{1l}(t), \tag{6}$$

where l denoted the times of iterative procedure.

(vi) The first product function (PF_1) is obtained by taking the product of $a_1(t)$ with the purely $s_{1n}(t)$ i.e., $PF_1 = a_1(t)s_{1n}(t)$. The instantaneous amplitude of PF_1 is $a_1(t)$ and the instantaneous frequency of PF_1 can be obtained from $s_{1n}(t)$ as:

$$f_1(t) = \frac{1}{2\pi}.\frac{d\left[\arccos\left(s_{1n}(t)\right)\right]}{dt}. \tag{7}$$

(vii) The difference of $y(t)$ from PF_1 is obtained as the new resulting signal. The whole process is repeated k times until $u_k(t)$ is monotonic or constant.

$$\begin{cases} u_1(t) = y(t) - PF_1(t) \\ u_2(t) = u_1(t) - PF_2(t) \\ \vdots \\ u_k(t) = u_{(k-1)}(t) - PF_k(t) \end{cases}. \tag{8}$$

Up to this point. However, the original signal can be obtained by

$$y(t) = \sum_{q=1}^{k} PF_q(t) + u_k(t), \tag{9}$$

where q is the number of PF and $u_k(t)$ represents the residual term. In this study, the LMD is applied in Python through Jupyter Notebook in Anaconda Navigator using "pyLMD" package.

2.2 Ensemble Empirical Mode Decomposition

In 1998, Huang introduced empirical mode decomposition (EMD) to analyze non-linear signals. In EMD, a signal is decomposed into intrinsic mode functions (*IMFs*). To become an *IMF*, a signal needs to satisfy two conditions (a) the average of lower and upper envelopes is zero everywhere and (b) The number of extremes and zero-crossing should be equal or differ at most by 1. However, EMD has a major issue of mode-mixing (Huang et al. 1998). To overcome the drawback of EMD, Wu and Huang (2009) introduced Ensemble empirical mode decomposition (EEMD) which has a noise-assisted system. The EEMD algorithm is given as follows:

(i) Initialize the amplitude of added white noise ($wn_j(t)$) and ensemble number N to the observed series $y(t)$. The j^{th} noise-added signal is $y_j(t) = y(t) + wn_j(t)$.
(ii) All the local minima and maxima of $y_j(t)$ are identified as lower and upper envelopes obtained by cubic spline functions.
(iii) Compute the average of $m_1(t)$ lower and upper envelopes.
(iv) Compute difference of $y_j(t)$ and average value $m_j(t)$ i.e., $h_1(t) = y_j(t) - m_1(t)$.
(v) Check if $h_1(t)$ is the first *IMF* component of the signal i.e., $h_1(t) = c_1(t)$, $r_1(t)$ is defined from the remaining data by $r_1(t) = y_j(t) - c_1(t)$. Otherwise, steps (ii)-(v) are repeated.

To determine the residue ($r_1(t)$) as a new signal, step (ii)-(v) should be repeated n times until the sift out all the *IMFs* till the stopping criterion is met. The stopping criteria occurs when the *IMF* component or residue becomes so small that is smaller than the predetermined value. After the sifting processing, the original signal $y_j(t)$ can be determined as the sum of all the *IMFs* and the residual error as follows:

$$y_j(t) = \sum_{k=1}^{n} c_k(t) + r_n(t), \tag{10}$$

where n is the number of *IMFs*, $c_k(t)$ denotes the k^{th} *IMF* when adding j^{th} noise and $r_n(t)$ denotes the final residual error.

2.3 Support Vector Regression (SVR)

The SVR is a widely applied supervised learning algorithm for regression and classification problems. SVR algorithm estimates the relationship between the output and input of a system using an existing sample (Vapnik 1995). Therefore, SVR is applied to model the non-linear and complex features of the river discharge series.

Suppose we have the training set with n observations $\{x_j, y_j\}$, where y_j represents the estimated output value of the data, x_j is the corresponding lagged input vector. Then, the SVR is developed as follows:

$$g(x) = w^T \Phi(x) + b, \tag{11}$$

where w is the weights vector, b is a constant which represents the bias and $\Phi(x)$ is the non-linear transfer function applied to project the input data into high dimensional space. Using the structural approach of risk minimization, Eq. (11) is solved as follows:

$$\text{Minimize}: \left(\frac{||w^2||}{2} + C \sum_{j=1}^{n} (\zeta + \zeta^*) \right) \text{subject to}: \begin{cases} g(x_j) - y_j \le \varepsilon + \zeta^* \\ y_j - g(x_j) \le \varepsilon + \zeta, \\ \zeta, \zeta^* \ge 0 \end{cases} \tag{12}$$

where $C > 0$ denotes the penalty parameter, ζ^* and ζ represent the slack variables which denotes the lower and upper constraint of $g(x)$ and ε represents the insensitive loss function. The Lagrangian function is further implemented which uses the regression function to replace the $\Phi(x)$ and weight vector given in the Eq. (11) as:

$$g(x_j) = \sum_{j=1}^{n} \left(\alpha_j - \alpha_j^* \right) k(x, x_j) + b, \tag{13}$$

where α_j and α_j^* are the Lagrange coefficients and $k(x, x_j) = \langle \Phi(x), \Phi(x_j) \rangle$ denotes the kernel function. There are different forms of kernel functions such as linear and radial basis. However, the choice of kernel depends upon the nature of the data. In this study, the SVR algorithm is implemented in R programming language using "e1071" package.

2.4 K-Nearest Neighbor (KNN)

Cover and Hart (1967) introduced the KNN algorithm as a non-parametric method for pattern recognition work. The KNN algorithm also has a good approximation ability for non-linear dynamics and is used for time series prediction (Martinez et al. 2018; Al-Juboor 2021). Thus, it is utilized as an efficient tool to capture non-linear dynamics of river discharge in this study.

The KNN algorithm calculates the similarity (neighborhood) of the input variables $X_o = \{x_{1o}, x_{2o}, \ldots, x_{no}\}$ with historical observations of the input variable $X_t = \{x_{1t}, x_{2t}, \ldots, x_{nt}\}$ using Euclidean distance function (D_{ot}) which is given by (Araghinejad 2013):

$$D_{ot} = \sqrt{\sum_{j=1}^{n} \left(x_{jo} - x_{jt}\right)^2}, t = 1, 2, \ldots, n. \tag{14}$$

The predicted variable (Y_o) is computed by using the probabilistic function of the observed values of discharge series (T_p):

$$Y_o = \sum_{p=1}^{K} f\left(D_{op}\right) \times T_p, \tag{15}$$

where $f\left(D_{op}\right)$ denotes the kernel function of the KNN computed by using the distance (D_{op}):

$$f\left(D_{op}\right) = \frac{1/D_{op}}{\sum_{p=1}^{K} \left(1/D_{op}\right)}. \tag{16}$$

The main parameter of KNN algorithm is K, finding an optimal value of K is a practical problem. In this study, the optimal value of K is computed by using metrics through the "caret" package in R.

2.5 Autoregressive Integrated Moving Average (ARIMA) Model

ARIMA model is a simple method for predicting a time series. ARIMA model is applicable for both non-stationary and stationary time series, which makes it an efficient model for predicting the uncertainty of river discharge (Wang et al. 2018). Thus, ARIMA model is used to predict river discharge which has uncertainty with time changes. Let y_t be the time series and ε_t represent the random error at time t. Then y_t is considered to be the linear function of past p observations ($y_{t-1}, y_{t-2}, \ldots, y_{t-p}$) and q random errors ($\varepsilon_t, \varepsilon_{t-1}, \ldots, \varepsilon_{t-q}$). The corresponding ARIMA model is given as:

$$y_t = \theta_1 y_{t-1} + \theta_2 y_{t-2} + \cdots + \theta_p y_{t-p} + \varepsilon_t - \beta_1 \varepsilon_{t-1} - \beta_2 \varepsilon_{t-2} - \cdots - \beta_q \varepsilon_{t-q}, \tag{17}$$

where $\theta_j (j = 0, 1, 2, \ldots, p)$ are the autoregressive coefficients, $\beta_j (j = 0, 1, 2, \ldots, q)$ are the moving average coefficients and ε_t is identically distributed with zero mean and constant variance. Similar to parameter d, the coefficients q and p are referred as the order of the ARIMA model. The main issue of ARIMA model is the determination of the appropriate order of (p, d, q) which is determined using correlation tools i.e., autocorrelation function (ACF) and partial ACF (Box et al. 2008).

3 Proposed Hybrid Modelling Method

In this study, a novel hybrid method is proposed to enhance the forecasting accuracy of hydrological data. The hybrid method proposed in this study is shown in Fig. 1.

The steps of the proposed hybrid method are described as follows:

(i) The LMD algorithm is applied to decompose the original river discharge series into several product functions (*PFs*) and residual.
(ii) The EEMD is employed to decompose the sub-series obtained in step (i). In this step, each component obtained by LMD is further decomposed into *IMFs* and residual.

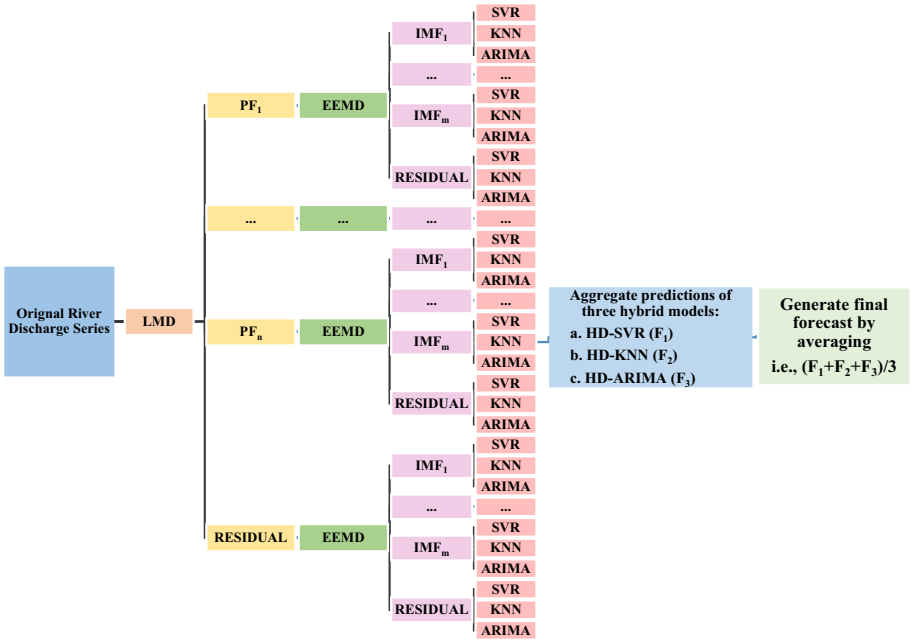


Fig. 1 Proposed Hybrid Method (HD-SKA)

(iii) The SVR, KNN and ARIMA models are applied to predict each *IMF* and residual component obtained in step (ii).
(iv) The predictions of all components is aggregated for HD-SVR, HD-KNN and HD-ARIMA models.
(v) The average of the predictions obtained from HD-SVR, HD-KNN and HD-ARIMA models is the final prediction of the river discharge series.

4 Case Studies

In this section, the description of data and the index measures used for evaluation are provided. Program codes were written in R and Python programming language version 4.0.3 and 3.7, respectively. All the analyses were performed on a personal computer with Intel Core i9-9900 CPU and 32.0 GB of RAM.

4.1 Data Description

In this study, the discharge data of different rivers in Indus Basin System has been collected which is used for the thorough evaluation of the performance of the proposed hybrid method. The data is collected for five rivers i.e., Jhelum River (Chattar Kallas station), Jhelum River (Domel station), Kabul River (Nowshera station), Kunhar river (Talhata station) and Kanshi River (Palotte station). The daily river discharge data is obtained from the Surface Water Hydrology Project agency of Water and Power Development Authority (WAPDA) Pakistan with different hydrological periods given in Table 1.

The river discharge series (m^3/s) is denoted by S_i for i = 1, 2, 3, 4 and 5 for Jhelum River (Chattar Kallas station), Jhelum River (Domel station), Kabul River, Kunhar river and Kanshi river respectively. Initially, the data cleaning procedure was performed on the five data series and the missing values were replaced by the average discharge of the month. In Table 1, the descriptive summary of the river discharge is given. It shows that on average the discharge was highest in S_1. The standard deviation (SD) of discharge is 564.13 m^3/s, 210.66 m^3/s, 750.68 m^3/s, 85.18 m^3/s and 9.31 m^3/s in S_1, S_2, S_3, S_4 and S_5, respectively. The shape of discharge of all rivers is positively skewed.

Figure 2b represents the box plot of discharge of five rivers. There is an evident difference between the discharge of rivers. The discharge of S_5 is relatively low as compared to other rivers (S_1-S_4). In Fig. 2b, some unusual discharge records are in all rivers that indicate a sign of water overflow in these rivers that is a hydrological hazard and needs proper exploration to avoid floods.

Figure 3 represents the time series plot of river discharge. It shows that the river discharge in all cases is non-linear, volatile and has huge variability during the given years. The discharge series (S_1-S_5) is divided into the training and testing sets. The first 80% observations are used as the training set and the last 20% observations are used as the testing set.

4.2 Evaluation Indexes

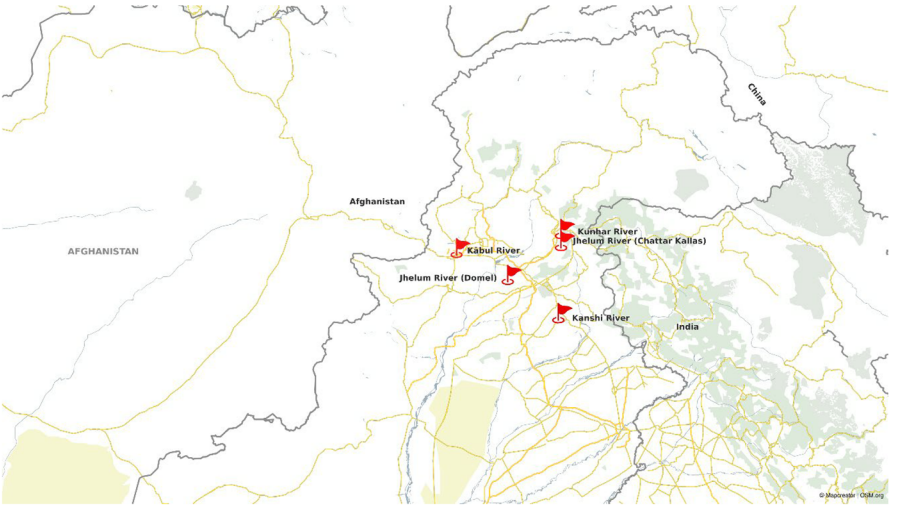
The root mean square error (RMSE), mean absolute error (MAE) and root-relative square error (RRSE) are adopted to evaluate the prediction performance of models. Several researchers have used these indicators e.g., Rezaie-Balf et al. (2019). These measures are defined as follows:

$$RMSE = \sqrt{\frac{1}{m}\sum_{j=1}^{m}\left(y_j - \hat{y}_j\right)^2},$$

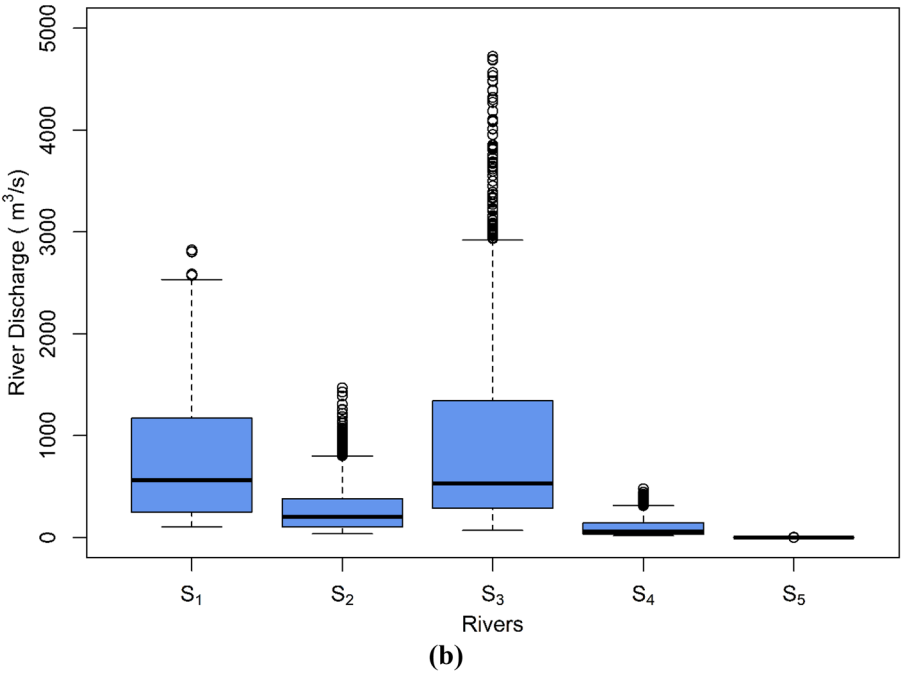
$$MAE = \frac{1}{m}\sum_{j=1}^{m}\left|y_j - \hat{y}_j\right|,$$

Table 1 Hydrological stations details and descriptive summary

	Jhelum River	Jhelum River	Kabul River	Kunhar River	Kanshi River
Stations	Chattar Kallas	Domel	Nowshera	Talhata	Palotte
Abbreviation	S_1	S_2	S_3	S_4	S_5
Period	2003–2017	2000–2017	2005–2017	2005–2017	2003–2017
Years	15	18	13	13	15
Observations	5479	6575	4748	4748	5479
Mean (m^3/s)	748.74	272.5	871.79	96.86	2.39
SD (m^3/s)	564.13	210.66	750.68	85.18	9.31
Minimum (m^3/s)	104.6	35.28	68.7	15.75	0.09
Median (m^3/s)	560.2	200.6	531.6	58.09	1.09
Maximum (m^3/s)	2823	1469	4724	478.2	230.6
Skewness	0.82	1.25	1.41	1.24	13.49



(a)



(b)

Fig. 2 (**a**) Pakistan Rivers Network (**b**) Box plot of original river discharge series (S_1-S_5)

$$RRSE = \sqrt{\frac{\sum_{j=1}^{m}\left(y_j - \hat{y}_j\right)^2}{\sum_{j=1}^{m}\left(\bar{\hat{y}} - \hat{y}_j\right)^2}},$$

where y_j represents the actual value, $\hat{y}_j$ represent the predicted value, $\bar{\hat{y}}$ is the mean of predicted values and m is the total number of observations in the considered case. To further

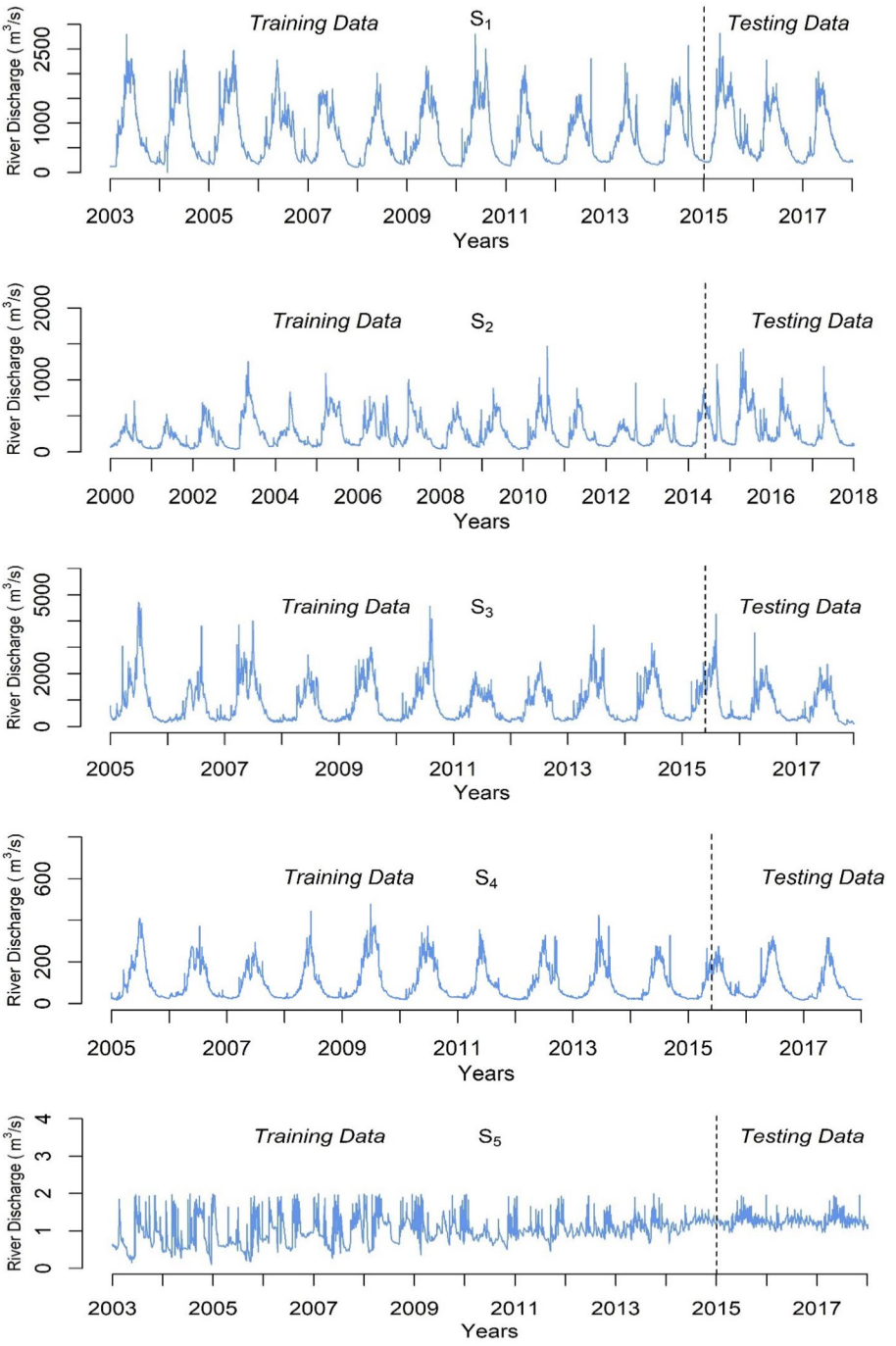


Fig. 3 Original River Discharge series (S_1-S_5)

evaluate the performance of models, the improved percentage indicators of RMSE, MAE and RRSE are used in this study. These are defined as follows:

$$P_{RMSE} = \frac{RMSE_1 - RMSE_2}{RMSE_1} \times 100\%,$$

$$P_{MAE} = \frac{MAE_1 - MAE_2}{MAE_1} \times 100\%,$$

$$P_{RRSE} = \frac{RRSE_1 - RRSE_2}{RRSE_1} \times 100\%.$$

5 Results and Discussion

In the proposed hybrid method, LMD and EEMD algorithms are used to decompose the original river discharge series. The training set of the original discharge series of S_1 is decomposed by LMD and the resulting *PFs* are given in Fig. 4. Then, using EEMD each *PF* and residual is further decomposed into *IMFs*. The EEMD decomposition results of PF_1 to PF_4 are presented in Fig. 5. Similarly, all the other series are decomposed. The remaining decomposition results of S_1-S_5 are provided in the supplementary materials. Further, the EEMD decomposed components were fed to SVR, ARIMA and KNN models individually. The predictions for all components is obtained and aggregated respectively to obtain the final forecasts of the HD-SVR, HD-KNN and HD-ARIMA models. Lastly, the average prediction of these three models is computed as the final prediction of the HD-SKA model.

Discharge data of five rivers is used to verify the prediction performance of the proposed HD-SKA model. The six benchmark models compared to the HD-SKA model are ARIMA, SVR, KNN, EEMD-ARIMA, EEMD-SVR and EEMD-KNN models. Lastly, the proposed hybrid model (HD-SKA) is applied to predict the daily river discharge of S_1-S_5.

5.1 Prediction Results

In this section, the prediction results of seven models on the S_1-S_5 series are compared and discussed. The prediction results of the training and testing phase are represented in Table 2. The detailed results are summarized as follows:

i. The SVR, KNN and ARIMA models are close competitors of each other in predicting daily river discharge of S_1-S_5.
ii. The EEMD-based models have better performance than single SVR, ARIMA, and KNN models except in some instances. For example, in S_4 training phase, the MAE of the EEMD-KNN model is more than the KNN model. The use of the decomposition technique improves the prediction accuracy of models for hydrological time series (Huynh et al. 2021). Therefore, EEMD has improved the prediction accuracy of SVR, ARIMA and KNN models in all five cases.
iii. For S_1-S_5, the RMSE, MAE and RRSE of the proposed HD-SKA model are smaller compared to the other considered models. However, there are certain situations where

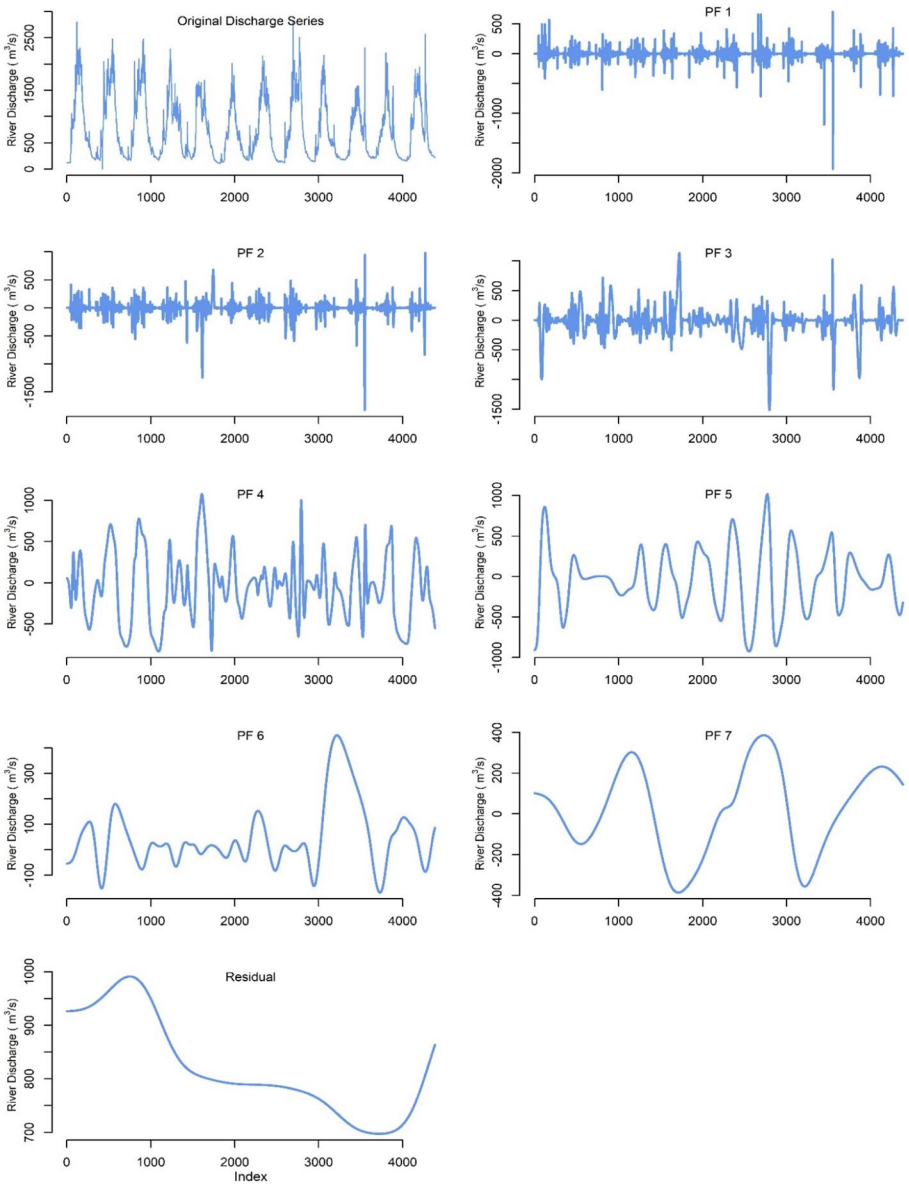


Fig. 4 LMD results for the original river discharge series of S_1

the MAE of the proposed HD-SKA model is slightly higher than EEMD-based models but this difference is negligible. The proposed HD-SKA model has better predictive accuracy than EEMD-based ARIMA, KNN and SVR models.

iv. Overall, the proposed HD-SKA model has better performance than all considered models in the study. The implementation of hybrid decomposition in the proposed

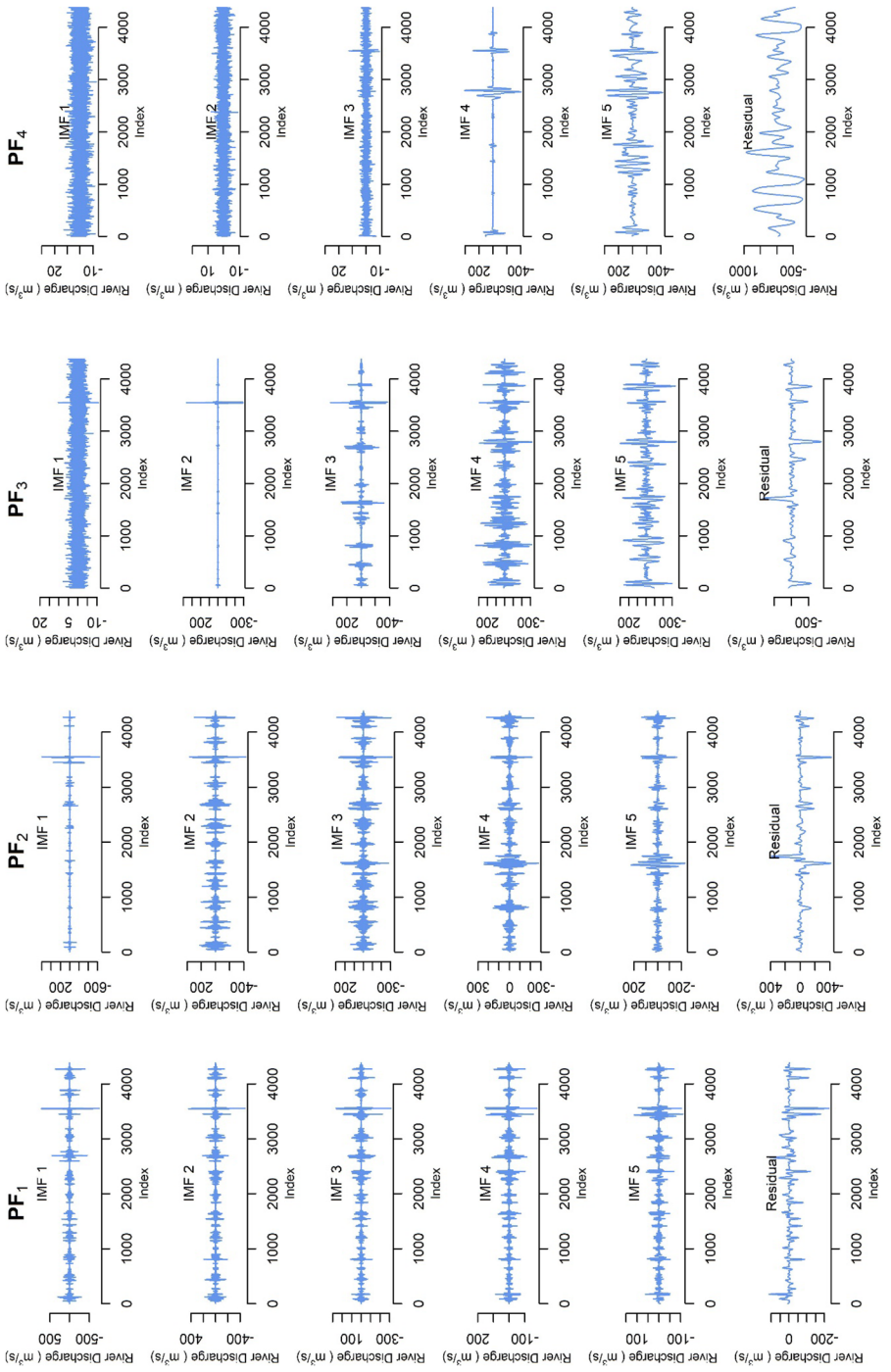


Fig. 5 Decomposition of PF_1-PF_4 obtained from LMD using EEMD (S_1)

Table 2 Performance evaluation of training and testing prediction error of proposed model with different models for S_1- S_5 river discharge series

Models	Index	Training Phase					Testing Phase				
		S_1	S_2	S_3	S_4	S_5	S_1	S_2	S_3	S_4	S_5
ARIMA	RMSE	81.4272	38.5317	131.0344	15.9721	0.1784	84.3583	52.6230	131.0225	11.6883	0.0969
SVR	(m^3/s)	80.9997	38.9483	140.0169	15.6500	0.1872	82.8488	52.8298	139.3558	11.7012	0.1053
KNN		79.7532	36.2794	138.5283	15.9214	0.1628	82.1642	52.7058	140.4319	11.6005	0.0970
EEMD-ARIMA		82.2954	37.4640	107.8559	11.6406	0.2340	73.7860	52.5147	105.7624	8.7127	0.0751
EEMD-KNN		76.7171	34.8272	131.3748	15.1920	0.1207	73.6116	44.1823	114.2192	10.9026	0.0842
EEMD-SVR		62.3850	27.5359	114.1717	11.4621	0.1281	63.3054	36.1627	128.5893	9.8583	0.0707
HD-SKA		**35.0683**	**21.4559**	**70.1673**	**6.6285**	**0.0987**	**52.7986**	**31.3009**	**91.2971**	**7.6326**	**0.0660**
ARIMA	MAE	40.9198	15.9683	67.7828	7.8752	0.0857	41.4753	22.6854	64.1092	6.5594	0.0536
SVR	(m^3/s)	42.5782	15.4375	73.6576	7.9713	0.0808	42.9115	22.4639	68.3356	6.8039	0.0554
KNN		40.9502	15.5878	72.9942	7.5801	0.0783	40.8584	23.2362	72.4327	6.2920	0.0576
EEMD-ARIMA		56.6170	24.2029	59.4547	6.1254	0.1773	50.7328	34.1026	53.9780	5.0311	0.0559
EEMD-KNN		40.8936	15.8171	62.8141	7.6403	**0.0652**	39.5701	22.1991	60.3089	6.1340	0.0439
EEMD-SVR		33.3988	**12.4023**	61.6146	5.7372	0.0653	**28.0107**	15.8208	78.6083	6.6263	**0.0376**
HD-SKA		**22.0289**	15.2136	**52.5121**	**3.7020**	0.0661	31.8084	**15.4365**	**51.5411**	**4.6620**	0.0397
ARIMA	RRSE	0.1433	0.1951	0.1713	0.1841	0.5123	0.1545	0.2138	0.1910	0.1490	0.7464
SVR	(m^3/s)	0.1426	0.1972	0.1830	0.1804	0.5375	0.1517	0.2146	0.2031	0.1492	0.8111
KNN		0.1404	0.1837	0.1811	0.1835	0.4676	0.1504	0.2141	0.2047	0.1479	0.7467
EEMD-ARIMA		0.1449	0.1897	0.1410	0.1342	0.6719	0.1351	0.2133	0.1542	0.1111	0.5786
EEMD-KNN		0.1350	0.1764	0.1717	0.1751	0.3466	0.1348	0.1795	0.1665	0.1390	0.6486
EEMD-SVR		0.1098	0.1395	0.1492	0.1321	0.3677	0.1159	0.1469	0.1874	0.1257	0.5446
HD-SKA		**0.0617**	**0.1087**	**0.0918**	**0.0764**	**0.2833**	**0.0967**	**0.1272**	**0.1331**	**0.0973**	**0.5085**

Bold values present minimum values in each column with respect to the river discharge series

method has enhanced the prediction capability of models. The HD-SKA model produces reliable river discharge prediction.

Figure 6a represents the prediction performance of the HD-SKA model compared to EEMD-based models in the testing phase of the S_1 and S_2 series. It can be observed that the proposed HD-SKA model precisely predicts the S_1 and S_2 series and is quite near to the original daily discharge. The remaining prediction plots for the S_3-S_5 series are provided in the supplementary material. For further verification of HD-SKA model performance, the coefficient of determination (R^2) of all the seven models is represented in Fig. 6b. It is observed that the proposed HD-SKA model yields highest R^2 value and has produces better prediction results compared to all other considered models in both training and testing phase.

5.2 Improvements by the Proposed Hybrid Model

The effectiveness of hybrid decomposition in the proposed hybrid model is illustrated in Table 3 through improved percentage indexes of RMSE, MAE and RRSE. By comparing the percentage improvements of the HD-SKA model to ARIMA, SVR and KNN models, it is observed that P_{RMSE}, P_{MAE}, and P_{RRSE} are all positive. The performance of the proposed HD-SKA model is superior to ARIMA, SVR, and KNN models in all five cases.

Further, the comparison of percentage improvements of the proposed HD-SKA model with EEMD-based models indicates that the HD-SKA model performs better than EEMD-ARIMA, EEMD-SVR and EEMD-KNN in all case studies. The P_{RMSE}, P_{MAE}, and P_{RRSE} are mostly positive except P_{MAE} index in some situations. For example, the MAE of the HD-SKA model compared to the EEMD-KNN model is increased by 1.38% in the training phase of the S_5 series. However, it can be observed that both models have similar performance and the difference is not much.

Overall, the proposed HD-SKA model has better performance than all other considered models. The hybrid decomposition-based models have better predictive performance than single decomposition-based models as studied in the literature see Vidya and Janani (2021). It may be deduced that the implementation of LMD with EEMD in the proposed hybrid model efficiently reduces the complexity and randomness of river discharge.

The proposed HD-SKA model in this study can serve as a helpful tool for the accurate prediction of river discharge.

5.3 Diebold-Mariano Test

The Diebold Mariano (DM) test is a well-known statistical hypothesis testing approach that helps in the identification of the degree of discrepancy among the proposed model and the compared models (Silva et al. 2021). The null hypothesis of the DM test is that the two models have similar prediction accuracy against the alternative that model 2 has lower prediction accuracy than model 1. Symbolically,

$$H_0 : E[L(e_t^1)] \geq E[L(e_t^2)],$$

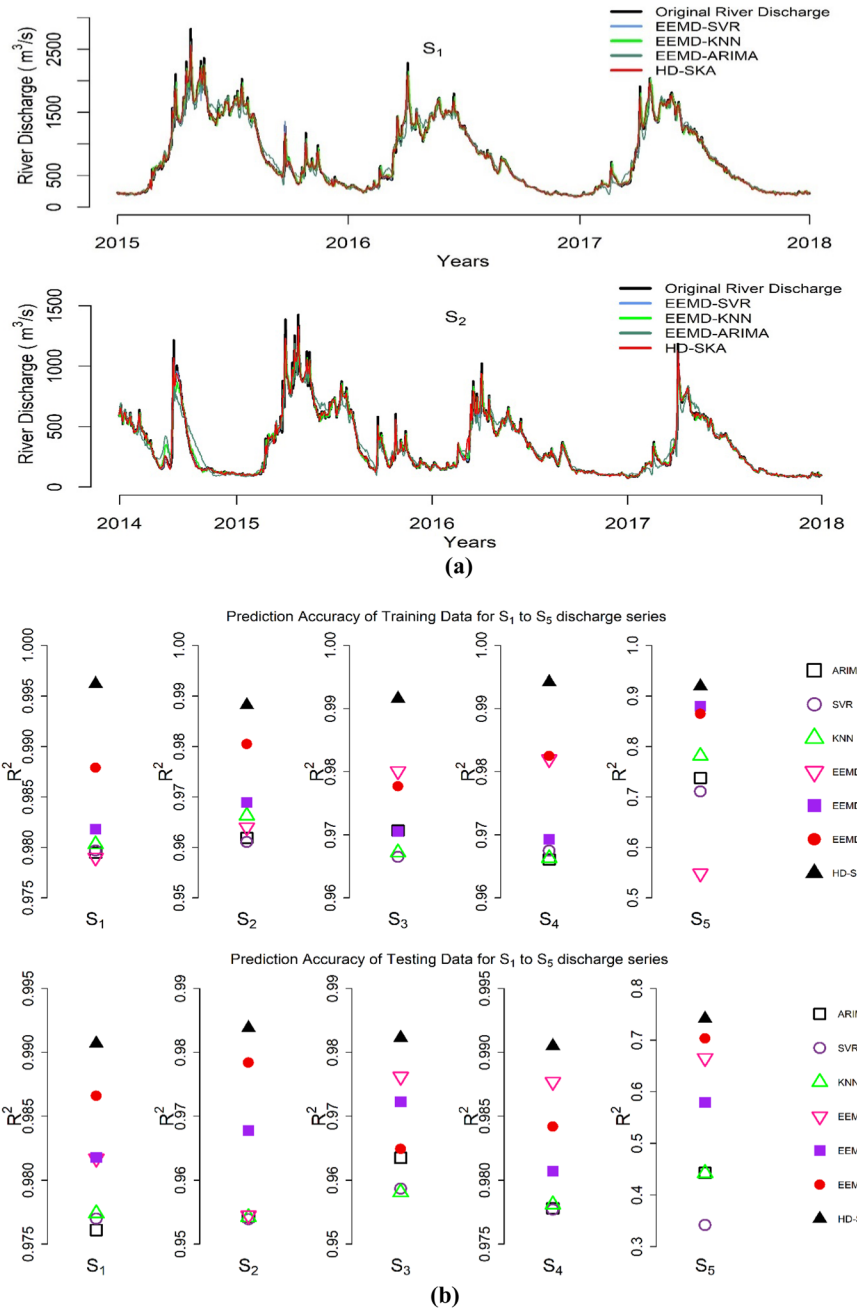


Fig. 6 Models analysis (**a**) Prediction performance of HD-SKA model on testing data of S_1 and S_2 series (**b**) Prediction accuracy plot for of seven models for S_1-S_5

Table 3 Improved percentage of proposed model among different models for S_1-S_5

Models	Index	Training Phase					Testing Phase				
		S_1	S_2	S_3	S_4	S_5	S_1	S_2	S_3	S_4	S_5
HD-SKA vs. ARIMA	P_{RMSE}	56.93%	44.32%	46.45%	58.50%	44.67%	37.41%	40.52%	30.32%	34.70%	31.89%
	P_{MAE}	46.17%	4.73%	22.53%	52.99%	22.87%	23.31%	31.95%	19.60%	28.93%	25.93%
	P_{RRSE}	56.94%	44.28%	46.41%	58.50%	44.70%	37.41%	40.51%	30.31%	34.70%	31.87%
HD-SKA vs. SVR	P_{RMSE}	56.71%	44.91%	49.89%	57.65%	47.28%	36.27%	40.75%	34.49%	34.77%	37.32%
	P_{MAE}	48.26%	1.45%	28.71%	53.56%	18.19%	25.87%	31.28%	24.58%	31.48%	28.34%
	P_{RRSE}	56.73%	44.88%	49.84%	57.65%	47.29%	36.26%	40.73%	34.47%	34.79%	37.31%
HD-SKA vs. KNN	P_{RMSE}	56.03%	40.86%	49.35%	58.37%	39.37%	35.74%	40.61%	34.99%	34.20%	31.96%
	P_{MAE}	46.21%	2.40%	28.06%	51.16%	15.58%	22.15%	33.57%	28.84%	25.91%	31.08%
	P_{RRSE}	56.05%	40.83%	49.31%	58.37%	39.41%	35.70%	40.59%	34.98%	34.21%	31.90%
HD-SKA vs. EEMD-ARIMA	P_{RMSE}	57.39%	42.73%	34.94%	43.06%	57.82%	28.44%	40.40%	13.68%	12.40%	12.12%
	P_{MAE}	61.09%	37.14%	11.68%	39.56%	62.72%	37.30%	54.74%	4.51%	7.34%	28.98%
	P_{RRSE}	57.42%	42.70%	34.89%	43.07%	57.84%	28.42%	40.37%	13.68%	12.42%	12.12%
HD-SKA vs. EEEMD-KNN	P_{RMSE}	54.29%	38.39%	46.59%	56.37%	18.23%	28.27%	29.16%	20.07%	29.99%	21.62%
	P_{MAE}	46.13%	3.82%	16.40%	51.55%	–1.38%	19.62%	30.46%	14.54%	24.00%	9.57%
	P_{RRSE}	54.30%	38.38%	46.53%	56.37%	18.26%	28.26%	29.14%	20.06%	30.00%	21.60%
HD-SKA vs. EEMD-SVR	P_{RMSE}	43.79%	22.08%	38.54%	42.17%	22.95%	16.60%	13.44%	29.00%	22.58%	6.65%
	P_{MAE}	34.04%	–22.67%	14.77%	35.47%	–1.23%	–13.56%	2.43%	34.43%	29.64%	–5.59%
	P_{RRSE}	43.81%	22.08%	38.47%	42.17%	22.95%	16.57%	13.41%	28.98%	22.59%	6.63%

Table 4 Diebold-Mariano test of different models in testing phase of S_1-S_5

Different Models	S_1	S_2	S_3	S_4	S_5
ARIMA	–4.2243***	–3.7329***	–2.7271***	–4.1192***	–6.6506***
SVR	–5.4130***	–4.5160***	–3.4066***	–4.5478***	–5.6435***
KNN	–5.2791***	–4.3066***	–3.9809***	–4.3201***	–6.1030***
EEMD-ARIMA	–7.0139***	–8.5430***	–1.3390*	–1.8785**	–3.3140***
EEMD-KNN	–3.7729***	–3.5380***	–1.8044**	–3.9566***	–5.8233***
EEMD-SVR	–1.9166**	–2.0631**	–3.5239***	–3.9858***	–1.2024

* is at 10%; ** is at 5%; *** is at 1% significance level

$$H_1 : E[L(e_t^1)] < E[L(e_t^2)],$$

where e_t^1 and e_t^2 are the prediction errors of two comparison models and L denotes the loss function of the prediction errors. The DM statistic is defined as follows:

$$DM = \frac{\bar{d}}{\sqrt{2\pi\hat{f}_d(0)/m}} \rightarrow N(0,1), \tag{18}$$

where $\hat{f}_d(0)$ represents the spectral density, m is the length of prediction results, $2\pi\hat{f}_d(0)$ denotes the consistent estimator of the asymptotic variance and $\bar{d} = \frac{1}{m}\sum_{t=1}^{m}\left(L(e_t^1) - L(e_t^2)\right)$. In this study, the loss function used is mean squared error (MSE). The null hypothesis will be rejected if $DM < -Z_{\alpha/2}$ where α is the level of significance.

In this study, the DM test is applied to verify the prediction results of models at 1%, 5%, and 10% levels of significance. The null hypothesis will be rejected for S_1-S_5 if the DM statistic value is smaller than –2.58, –1.96, and –1.64 for 1%, 5% and 10% levels of significance respectively. The DM test is applied using the proposed HD-SKA model as "model 1" and benchmark models as "model 2" respectively.

In Table 4, the DM test statistic values on predictions of testing data sets are given for S_1-S_5 series. The null hypothesis is rejected for all cases except for the EEMD-SVR model in S_5. The DM test results reveals that the performance of the proposed HD-SKA model for the S_1-S_5 series is markedly diverse and superior to all considered models. However, in S_5 the EEMD-SVR model and HD-SKA model have similar prediction accuracy. Thus, the proposed HD-SKA model has higher prediction accuracy among all considered models for river discharge prediction.

6 Conclusion

In this study, a novel hybrid method is proposed to predict daily river discharge. The application of the proposed method is illustrated using discharge data of five rivers in the Indus Basin System, Pakistan. The performance of the proposed HD-SKA model is compared with six benchmark models. The models' performance is evaluated using different performance indicators and the Diebold-Mariano test. The data analysis results show that the proposed HD-SKA model outperforms all the models considered in the study. The results of the Diebold-Mariano test revealed that the HD-SKA model possesses a higher predictive ability than benchmark models. Overall, the proposed hybrid method can be a successful

tool to predict daily river discharge. The proposed method has greater accuracy when the number of components identified at the second decomposition stage is not too large. Moreover, the application of the proposed method may be checked on large data sets in future research work.

Acknowledgements We would also like to acknowledge the Surface Water Hydrology Project (SWHP) Agency of Water and Power Development Authority (WAPDA), Pakistan for providing the river discharge data for this research work. We are thankful to the Editor and anonymous reviewers for teir comments on earlier versions of the manuscript which improved the paper.

Author Contribution All the authors jointly worked on the idea. MS collected the data. SC and FI worked on the methodology. MS performed the analysis and prepared the initial draft. SC and FI reviewed and revised the draft.

Funding No funding for this research work.

Availability of Data and Material Will be provided on request.

Code Availability Will be provided on request.

References

Adnan RM, Petroselli A, Heddam S, Santos C, Kisi O (2021) Short term rainfall-runoff modelling using several machine learning methods and a conceptual event-based model. Stoch Environ Res Risk Assess 35:597–616. https://doi.org/10.1007/s00477-020-01910-0

Aghelpour P, Bahrami-Pichaghchi H, Varshavian V (2021) Hydrological drought forecasting using multi-scalar streamflow drought index, stochastic models and machine learning approaches, in northern Iran. Stoch Environ Res Risk Assess 35:1615–1635. https://doi.org/10.1007/s00477-020-01949-z

Alizadeh F, Gharamaleki AF, Jalilzadeh R (2021) A two-stage multiple-point conceptual model to predict river stage-discharge process using machine learning approaches. J Water Clim Change 12:278–295. https://doi.org/10.2166/wcc.2020.006

Al-Juboor AM (2021) A hybrid model to predict monthly streamflow using neighboring rivers annual flows. Water Resour Manage 35:729–743. https://doi.org/10.1007/s11269-020-02757-4

Araghinejad S (2013) Data-driven modeling: Using MATLAB in water resources and environmental. Springer Science & Business Media, Berlin

Bayazit M (2015) Nonstationarity of hydrological records and recent trends in trend analysis: a state-of-the-art review. Environ Process 2:527–542. https://doi.org/10.1007/s40710-015-0081-7

Bonakdari H, Binns AD, Gharabaghi B (2020) A comparative study of linear stochastic with nonlinear daily river discharge forecast models. Water Resour Manage 34:3689–3708. https://doi.org/10.1007/s11269-020-02644-y

Box GE, Jenkins GM, Reinsel GC (2008) Operational Research Quarterly. Time Series Analysis: Forecasting and Control, 4th edn. John Wiley & Sons Inc, New York, pp 137–191

Cover T, Hart P (1967) Nearest neighbor pattern classification. IEEE Trans Inf Theory 13:21–27. https://doi.org/10.1109/TIT.1967.1053964

Dehghani R, Torabi H, Younesi H, Shahinejad B (2021) Application of wavelet support vector machine (WSVM) model in predicting river flow (Case study: Dez basin). Watershed Eng Manage 13:98–110. https://doi.org/10.22092/IJWMSE.2020.128735.1748

Fashae O, Olusola A, Ndubuisi I, Udomboso C (2019) Comparing ANN and ARIMA model in predicting the discharge of River Opeki from 2010 to 2020. River Res Appl 35:169–177. https://doi.org/10.1002/rra.3391

Huang N, Shen Z, Long S, Wu M, Shih H, Zheng Q, Yen N, Tung CC, Liu H (1998) The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proc R Soc Lond a: Math Phys Eng Sci 454:903–995. https://doi.org/10.1098/rspa.1998.0193

Huynh AN, Deo RC, Ali M, Abdulla S, Raj N (2021) Novel short-term solar radiation hybrid model: Long short-term memory network integrated with robust local mean decomposition. Appl Energy 298:117193. https://doi.org/10.1016/j.apenergy.2021.117193

Liu H, Han M (2014) A fault diagnosis method based on local mean decomposition and multi-scale entropy for roller bearings. Mech Mach Theory 75:67–78. https://doi.org/10.1016/j.mechmachtheory.2014.01.011

Martinez F, Frias MP, Perez-Godoy MD, Rivera AJ (2018) Dealing with seasonality by narrowing the training set in time series forecasting with kNN. Expert Syst Appl 103:38–48. https://doi.org/10.1016/j.eswa.2018.03.005

Meshram SG, Ghorbani MA, Shamshirband S, Karimi V, Meshram C (2019) River flow prediction using hybrid PSOGSA algorithm based on feed-forward neural network. Soft Comput 23:10429–10438. https://doi.org/10.1007/s00500-018-3598-7

Musarat MA, Alaloul WS, Rabbani MB, Ali M, Altaf M, Fediuk R, Vatin N, Klyuev S, Bukhari H, Sadiq A, Rafiq W, Farooq W (2021) Kabul river flow prediction using automated ARIMA forecasting: a machine learning approach. Sustainability 13:10720–10746. https://doi.org/10.3390/su131910720

Nikolic VV, Simonovic SP (2015) Multi-method modeling framework for support of integrated water resources management. Environ Process 2:461–483. https://doi.org/10.1007/s40710-015-0082-6

Poul A, Shourian M, Ebrahimi H (2019) A comparative study of MLR, KNN, ANN and ANFIS models with wavelet transform in monthly stream flow prediction. Water Resour Manage 33:2907–2923. https://doi.org/10.1007/s11269-019-02273-0

Rezaie-Balf M, Fani Nowbandegani S, Samadi S, Fallah H, Alaghmand S (2019) An ensemble decomposition-based artificial intelligence approach for daily streamflow prediction. Water 11:709–738. https://doi.org/10.3390/w11040709

Riahi-Madvar H, Dehghani M, Memarzadeh R, Gharabaghi B (2021) Short to long-term forecasting of river flows by Heuristic optimization algorithms hybridized with ANFIS. Water Resour Manage 35:1149–1166. https://doi.org/10.1007/s11269-020-02756-5

Sehgal V, Tiwari MK, Chatterjee C (2014) Wavelet bootstrap multiple linear regression based hybrid modeling for daily river discharge forecasting. Water Resour Manag 28:2793–2811. https://doi.org/10.1007/s11269-014-0638-7

Sharghi E, Nourani V, Najafi H, Soleimani S (2019) Wavelet-exponential smoothing: a new hybrid method for suspended sediment load modeling. Environ Process 6:191–218. https://doi.org/10.1007/s40710-019-00363-0

Silva RG, Ribeiro MH, Moreno SR, Mariani VC, Coelho LDS (2021) A novel decomposition-ensemble learning framework for multi-step ahead wind energy forecasting. Energy 216:119174. https://doi.org/10.1016/j.energy.2020.119174

Smith JS (2005) The local mean decomposition and its application to EEG perception data. J R Soc Interface 2:443–454. https://doi.org/10.1098/rsif.2005.0058

Vapnik V (1995) The nature of statistical learning theory. Springer, New York

Vidya S, Janani SV (2021) Wind speed multistep forecasting model using a hybrid decomposition technique and a selfish herd optimizer-based deep neural network. Soft Comput 25:6237–6270. https://doi.org/10.1007/s00500-021-05608-5

Wang ZY, Qiu J, Li FF (2018) Hybrid models combining EMD/EEMD and ARIMA for Long-term streamflow forecasting. Water 10:853–866. https://doi.org/10.3390/w10070853

Wei S, Yang H, Song J, Abbaspour K, Xu Z (2013) A wavelet-neural network hybrid modelling approach for estimating and predicting river monthly flows. Hydrol Sci J 58:374–389. https://doi.org/10.1080/02626667.2012.754102

Wu C, Chau K, Li Y (2008) River stage prediction based on a distributed support vector regression. J Hydrol 358:96–111. https://doi.org/10.1016/j.jhydrol.2008.05.028

Wu Z, Huang NE (2009) Ensemble empirical mode decomposition: a noise-assisted data analysis method. Adv Adapt Data Anal (AADA) 1:1–41. https://doi.org/10.1142/S1793536909000047