A novel hybrid framework to model the relationship of daily river discharge with meteorological variables

Maha Shabbir, Sohail Chand
University of the Punjab, Pakistan
Farhat Iqbal
Imam Abulrahman Bin Faisal University, Saudi Arabia

Abstract

River discharge is affected by many factors, such as water level, rainfall, and precipitation. This study proposes a new hybrid framework named LAES (LASSO-ANN-EMD-SVM) to model the relationship of daily river discharge with meteorological variables. This hybrid framework is a composite of the least absolute shrinkage and selection operator (LASSO), an artificial neural network (ANN), and an error correction method. In the first stage, LASSO identifies meteorological variables that have a significant influence on the generation of river discharge. Next, the ANN model is used to predict river discharge using meteorological variables selected by LASSO, and the error series is determined. The error series is decomposed into intrinsic mode functions and residuals using empirical mode decomposition (EMD). The EMD components are modeled using the support vector machine (SVM) model, and the error predictions are aggregated. In the last stage, the LASSO-ANN predictions and the predicted error series are aggregated as the final discharge prediction. The proposed hybrid framework is illustrated on the Kabul River of Pakistan. The performance of the proposed hybrid framework is compared with six models using various performance measures and the Diebold-Mariano test. These models include multiple linear regression (MLR), SVM, ANN, LASSO-MLR, LASSO-SVM, and LASSO-ANN models. The findings reveal that the proposed hybrid model outperforms all other models considered in the study. In the testing phase, the root mean square error (*RMSE*), mean absolute percentage error (*MAPE*), and mean absolute error (*MAE*) of the proposed LAES hybrid model are 337.143 m^3/s, 32.354%, and 218.353 m^3/s which are smaller than all other models compared in the study. Our proposed hybrid system is an efficient model for river discharge prediction that will be helpful in water management and protection against floods. Long-term prediction can help to identify the major effects of climate change and to make evidence-based environmental policies.

Keywords

LASSO, river discharge, ANN, SVM, EMD.

Submitted 5 September 2023, revised 22 April 2024, accepted 24 April 2024
DOI: 10.26491/mhwm/187899

1. Introduction

Water is necessary for the survival of all living organisms in the world. Water is life, and demand for it is increasing due to rapid increases in population, urbanization, and industrialization. Moreover, water is a primary need for domestic, industrial, and agricultural activities (Mehta et al. 2022). Thus, it is essential to carefully manage and plan water resources to reduce loss of life and property damage caused by drought, floods, or heat waves (Ali, Shahbaz 2020; Mangukiya et al. 2022). Climate changes influence the hydrological cycle globally; the resulting variations in weather and climate have increased the risks of drought and floods because weather changes, variations in precipitation, peak flows, and extreme temperatures have impacts on river discharge (Mehmood et al. 2021). The amount of discharge generated from a catchment depends on various factors such as duration, meteorological variables, velocity, and water level (Gleason et

al. 2014; Saidi et al. 2018; Malik et al. 2020). Therefore, it is necessary to model river discharge using information on the weather at the relevant hydrological station (Dariane, Azimi 2018).

In the past thirty years, stochastic, physical, black box (machine learning and statistical), and conceptual models have been widely applied in hydrological studies. Physical models have been used for hydrological modeling, but their successful application is bound to the complexity of governing equations and the difficulty in measuring the parameters involved (Yousuf et al. 2017). Statistical models try to determine the relationships within the actual data. Their application is limited when data have unique and complex characteristics such as non-linearity, multicollinearity, volatility, irregularities, noise, outliers, and more. In the past two decades, machine learning models have gained importance in hydrology due to their flexibility in handling datasets with unique characteristics (Ravindran et al. 2021; Elbeltagi et al. 2022). Rasouli et al. (2012) applied a support vector machine (SVM), Bayesian neural network, and Gaussian process to predict non-linear river discharge in North America using climate and weather variables. Ali and Shahbaz (2020) applied an artificial neural network (ANN) to predict river discharge in the upper Jhelum River basin of Pakistan.

Although data-driven (statistical and machine learning) models are applied to predict river discharge, there is no single model that can predict river discharge without bias or with utmost certainty (Mehmood et al. 2021). Literature shows that researchers have developed hybrid models by combining two or more techniques to improve the prediction ability of the models (Shabbir et al. 2024). Wang and Li (2018) introduced a hybrid framework based on an error correction approach using the generalized autoregressive conditionally heteroscedastic (GARCH) model when inherent correction and heteroscedasticity of errors cannot be ignored. Zhang et al. (2018a) developed an error-correction-based hybrid framework using an autoregressive (AR) model to predict water levels with improved accuracy. Luo et al. (2019) suggested a hybrid framework based on a composition of factor analysis, decomposition of time series, data regression, and error suppression to predict river discharge. Yan et al. (2020) combined a generalized additive model (GAM) with principal component analysis (PCA) to model the relationship between water level and macroinvertebrate diversity index in the Baiyandian Lake of China. Mehr and Gandomi (2021) suggested a hybrid model by integrating a multi-stage genetic programming (MSGP) model with the least absolute shrinkage and selection operator (LASSO) for improved prediction of river flow. Emadi et al. (2022) modeled river water using a hybrid evolutionary data-driven approach.

River discharge estimation is challenging in hydrological studies because its generation depends on various factors such as rainfall patterns, spatial-temporal irregularities, climatic changes, and many more (Cheng et al. 2019; Hu et al. 2022). In literature, much discussion is on the time series prediction of river discharge (see Luo et al. 2019; Mehr, Gandomi 2021; Adnan et al. 2022). There is an essential need to develop new methods to evaluate the possible influence of different factors on the generation of river discharge. Keeping in view this gap, this study aims to develop a new hybrid approach to examine the relationship between river discharge and meteorological variables.

A new hybrid framework named LAES (LASSO-ANN-EMD-SVM) is proposed in this study based on a combination of feature selection, an ANN model, and an error correction method. In the first stage, LASSO is used to identify meteorological variables that have significant relationships with river discharge. The variables identified by LASSO are then used as input variables to the ANN model to obtain the discharge predictions, and then the error series is computed. Further, the empirical mode decomposition (EMD) technique is used to decompose error series into intrinsic mode functions and residuals. These components are modeled using the SVM model, and their predictions are aggregated. The final discharge prediction is obtained by adding the LASSO-ANN discharge predictions with EMD-SVM error predictions. Application of the proposed LAES hybrid framework is demonstrated for the Kabul River of Pakistan, and its prediction performance is compared with different models.

The proposed hybrid framework is novel as it efficiently predicts river discharge by considering the influence of meteorological variables that have a significant impact on river discharge using LASSO. In addition, the error correction approach in the proposed LAES hybrid model helps to enhance the prediction of discharge by capturing the randomness and volatility of the error series. It provides reliable estimates of river discharge and can be helpful in the management of water supply and flood control.

2. Methods

2.1. Multiple linear regression

The multiple linear regression (MLR) model is a simple and widely used modeling technique. The MLR model is given as:

$$y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \cdots + \beta_p x_{pj} + u_j, \quad j = 1,2,\ldots,n \tag{1}$$

where y_j is the dependent (output) variable, β_j are the regression coefficients, x_j are the independent (input) variable, n is the number of observations, p is the number of independent variables, and u_j is the residual term.

2.2. Least absolute shrinkage and selection operator

Tibshirani (1996) introduced the least absolute shrinkage and selection operator (LASSO) as a variable-selection approach for regression models. The method minimizes the residual sum of squares subject to the absolute values of the regression coefficients. LASSO performs variable selection and regularization simultaneously to enhance the interpretability and precision of statistical models (Tibshirani 1996). This study applies LASSO to determine important meteorological variables for predicting river discharge.

Assuming a sample contains M events where each event has p number of independent variables and one dependent variable, let $\mathbf{y_i}$ be the dependent (output) variable, and $\mathbf{x_i} = \left(x_1, x_2, \ldots, x_p\right)^T$ be the vector of i^{th} independent (input) variables, then the objective function of LASSO is:

For all $\sum_{j=1}^{p}|\beta_j| \leq \lambda$, find the $\min_{\beta} \frac{1}{M}\sum_{i=1}^{M}\left(\boldsymbol{y_i} - \mathbf{x_i^T}\boldsymbol{\beta}\right)^2$ (2)

where λ is a pre-determined parameter that determines the regularization degree and $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)$ is the vector of regression coefficients. Let $\mathbf{X}$ be the matrix of independent variables, i.e. $\mathbf{X_{ij}} = (x_i)_j$, where $i = 1, 2, \ldots, M, j = 1, 2, \ldots, p$ and x_i^T is the i^{th} row of $\mathbf{X}$. Then, the above formula in a compact form can be written as:

For all $\|\boldsymbol{\beta}\|_1 \leq \lambda$, calculate $\min_{\beta}\left\{\frac{1}{M}\|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2\right\}$ (3)

where $\|\boldsymbol{\beta}\|_p = \left(\sum_{i=1}^{M}|\beta_i|_p\right)^{1/p}$ is the standard L_p norm, $\mathbf{1_M}$ is a column vector of M dimensions with entries 1. In this study, LASSO is employed using the optimal glmnet library in R language and the optimal value of the LASSO parameter using this library is obtained using a 5-fold cross-validation approach.

2.3. Artificial neural network

The artificial neural network (ANN) is a robust modeling tool in which information processing is a representation of biological systems (Kachrimanis et al. 2003). The network is constructed from interconnected neurons, which can determine values from the inputs through network processing. The neuron receives input signals and provides the output signal that mainly depends on the neuron processing function. The ANN architecture consists of a series of interlinked neuron layers. Every layer is linked with another layer through neurons, which transfer information between these layers. Through this processing, the information reaches the output (dependent variable) layer. The ANN mechanism follows four assumptions:

a) Inputs are handled by neurons.
b) Through the connection of neurons, the information of inputs is passed on to the adjacent layers.
c) Each neuron has a weight, and the output from the neuron is the product of its input and its associated weight.
d) The transmitted inputs are passed via the activation of neurons to obtain the output.

Figure 1a shows the architecture of the ANN model, and Figure 1b presents the structure of a neuron where every input (independent variable) comes from other neurons and are multiplied by their weights $(w_j; j = 1,2, \ldots, n)$ respectively and then aggregated with the bias ($\mathbf{b}$) vector. This aggregated input (s) is passed using the transfer or activation function (f) to obtain the output (a) of a specific neuron. Letting $\mathbf{x}$ be the vector of independent (input) variables, the neural network maps into another output vector $\mathbf{a}$ through:

$$\mathbf{a} = f(\mathbf{x}.w + \mathbf{b}) \quad (4)$$

The mean squared error (*MSE*) is computed and using the back-propagation process, the weights of the entire network are modified in the training process. The accuracy of the ANN depends on the quality and amount of data in training.

In this study, the ANN algorithm is trained by a back-propagation technique where the output and input variables are applied in the network. The Broyden-Fletcher-Goldfarb-Shanno (BFGS) optimization is employed in a three-hidden-layer network. In the input layer of the ANN algorithm, the activation function is applied with 1000 iterations in the hidden layers. In this study, the ANN algorithm is applied using the validant library in the R programming language.

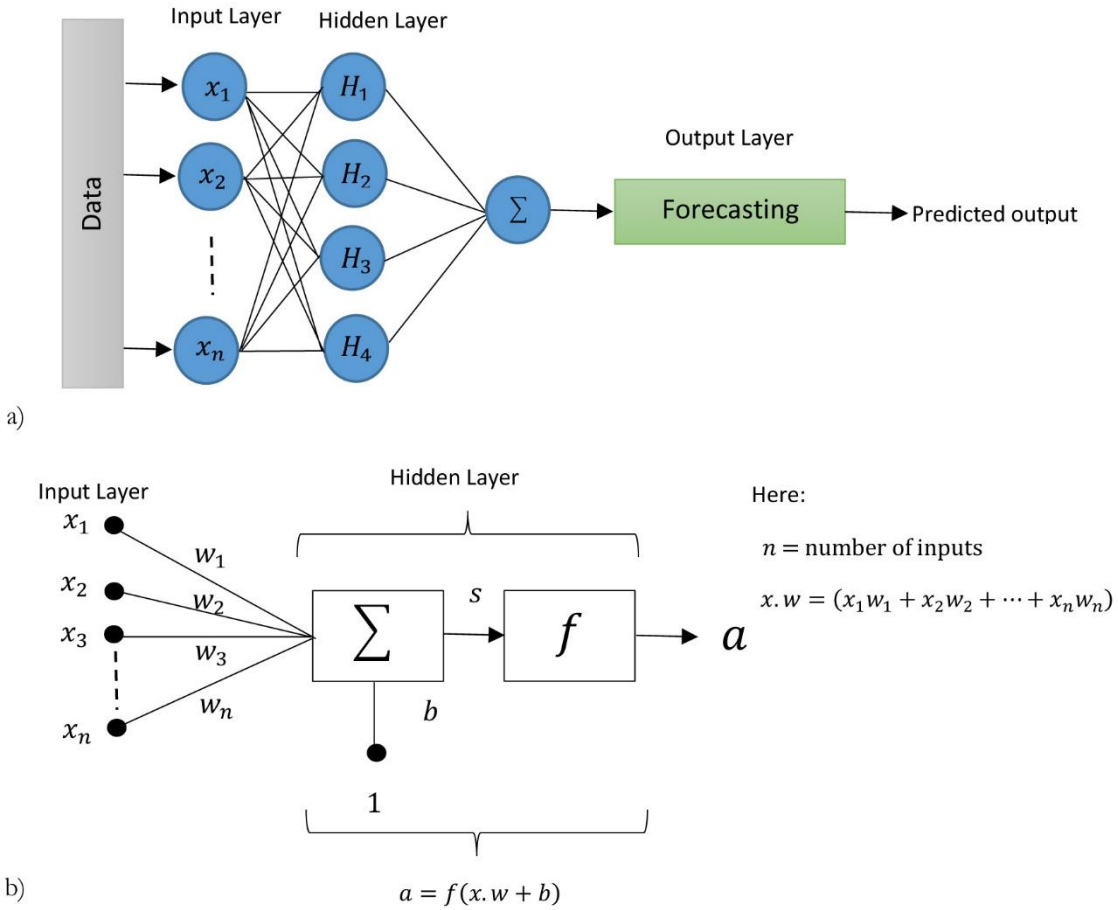


Fig. 1. The mathematical model of ANN (a) and systematic representation of a neuron (b).

2.4. Empirical mode decomposition

Huang et al. (1998) introduced empirical mode decomposition (EMD) as an adaptive method for signal analysis. The EMD is designed to analyze non-linear series. The EMD approach assumes that a signal contains different intrinsic mode functions ($IMFs$) of oscillations. Every mode has the same number of extrema and zero-crossings. There is a single extremum between successive zero-crossings. In this way, the signal is decomposed into different $IMFs$ and residuals. A component is an IMF if it satisfies two conditions: (i) the number of extrema and the number of zero-crossings must be equal to one or differ at most by one, and (ii) at any point, the average of the envelope is zero (Huang et al. 1998). Any original signal $y(t)$ can be decomposed using the EMD algorithm as follows (Lei et al. 2003; Jungsheng et al. 2006):

a) Find the local minima and maximum through the cubic spline line as the upper envelope and lower envelope, respectively.

b) Find the mean (m_1) of upper and lower envelopes.

c) The difference between the $y(t)$ and the 1^{st} component m_1 is the first component denoted as h_1 i.e. $h_1 = y(t) - m_1$. If h_1 is an IMF, then it is said to be the first IMF component of $y(t)$.

d) If h_1 is not an IMF, then it is treated as an original signal, and the steps (a)-(c) are repeated, then $h_1 - m_{11} = h_{11}$.

After repeating the sifting process k times, h_{1k} becomes an IMF , i.e. $h_{1(k-1)} - m_{1k} = h_{1k}$, then it is termed as:

$$c_1 = h_{1k} \tag{5}$$

The first IMF component from the data.

e) Next, subtract c_1 from $y(t)$ to obtain $u_1 = y(t) - c_1$ where u_1 denotes the treated data, and the process is repeated n times to get n $IMFs$ of $y(t)$. Then,

$$\left.\begin{array}{c} u_1 - c_2 = u_2 \\ \vdots \\ u_{n-1} - c_n = u_n \end{array}\right\} \tag{6}$$

At the end of the process, we have $IMFs$ $(c_j; j = 1,2, \dots, n)$ and residual (u_j). By summation of all the components, the original signal $y(t)$ can be obtained as:

$$y(t) = \sum_{j=1}^{n} c_j + u_n \tag{7}$$

The EMD method is implemented using the *EMD* library in R language in this study.

2.5. Support vector machine

Support vector machine (SVM) is a popular modeling technique for classification and regression problems. The SVM algorithm maps complex high-dimensional data into high-feature space (Vapnik 1995). We assume a training set with n observations, $\{x_d, y_d\}, d = 1,2, \dots, n, x_d \in R, y_d \in R$, where y_d denotes the estimated value of the dependent (output) variable, x_d is the corresponding lagged values of the dependent variable, and n is the sample size. Then, the SVM is developed as:

$$f(x) = \boldsymbol{\omega}^{\mathbf{T}} \varphi(x) + b \tag{8}$$

where $f(x)$ is the estimated dependent variable, $b \in R$ is the bias, and $\boldsymbol{\omega} \in R$ represents the vector of weights. The transfer function $\varphi(x)$ maps input data into high-dimensional space. The Eq. (8) is solved by risk minimization as follows:

$$\text{Minimum: } \left(\frac{\|\boldsymbol{\omega}^2\|}{2} + c\sum_{d=1}^{n}(\xi^* + \xi)\right) \text{ subject to: } \begin{cases} f(x_d) - y_d \leq \varepsilon + \xi^* \\ y_d - f(x_d) \leq \varepsilon + \xi \\ \xi, \xi^* \geq 0 \end{cases} \quad (9)$$

where $c > 0$ represents the penalty parameter, ξ and $\xi^*$are slack variables that show the upper and lower constraint of $f(x)$, and ε denotes the insensitive loss function. Further, the Lagrangian function is used as the non-linear regression function, which replaces $\varphi(x)$ and $\boldsymbol{\omega}$ in Eq. (8) as:

$$f(x_d) = \sum_{d=1}^{n}(\alpha_d - \alpha_d^*)k(x, x_d) + b \quad (10)$$

where $k(x, x_d) = \langle \varphi(x), \varphi(x_d) \rangle$ is the kernel function. The α_d^* and α_d represents the Lagrange coefficients.

In this study, SVM is applied to capture the features of the error series using the radial basis function (RBF) kernel, i.e. $k(x, x_d) = e^{\frac{\|x - x_d\|}{2g^2}}$, where g is the width of RBF (Baydaroğlu et al. 2018). The SVM algorithm is applied in this study using the R language e1071 library with unit cost and $g = 1/m$ where m is the number of input variables.

3. Proposed hybrid framework

In this paper, we propose a novel LASSO-ANN-EMD-SVM (LAES) hybrid framework to predict daily river discharge based on its relationship with the meteorological variables. The proposed LAES hybrid framework is displayed in Figure 2.

The steps of the LAES framework are:

a) LASSO is applied for the selection of meteorological variables that influence discharge (y) of the river.

b) Next, the ANN model is employed to model river discharge using meteorological variables as independent variables and the predictions of river discharge $(\hat{y}_{LA})$ are obtained. Further, the error (i.e. $\hat{e} = y - \hat{y}_{LA}$) is computed.

c) Using EMD, the error is decomposed into sub-series, and then the SVM model is used to predict each sub-series. By aggregating them, the predicted error $(\hat{e}_{ES})$ is obtained.

d) The final river discharge prediction is obtained using the predicted error series to correct the predicted river discharge in stage II $(\text{i.e. } \hat{y}_{LAES} = \hat{y}_{LA} + \hat{e}_{ES})$.

The proposed LAES hybrid method is a unique combination of the feature selection method with the ANN model and error correction approach. To the best of our knowledge, there is no hybrid model in the literature that integrates LASSO with an error correction approach for modeling non-linear and high-dimensional data sets.

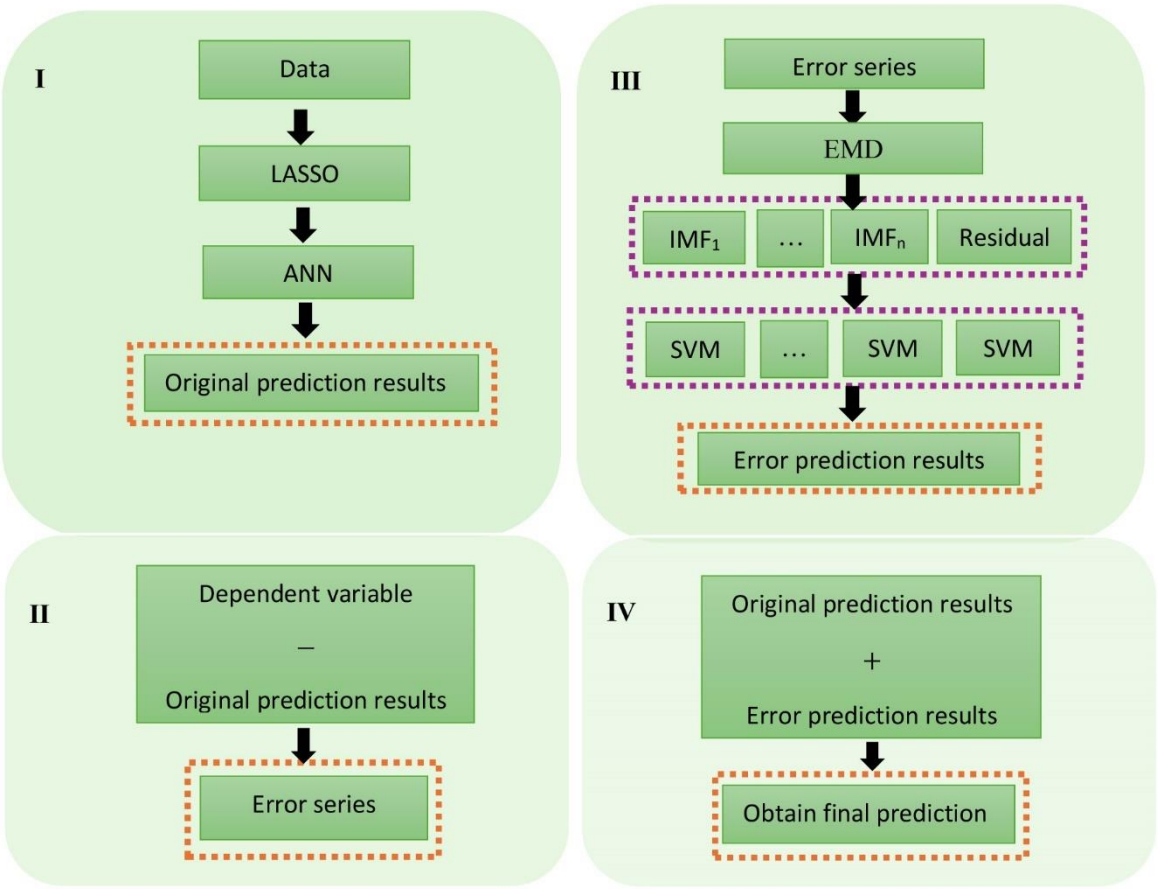


Fig. 2. The proposed LAES hybrid framework.

3.1. Limitations of LAES hybrid framework

The efficiency of the LAES hybrid framework depends on the optimal choice of parameters of the LASSO approach. This framework works efficiently when the independent variables are selected using the optimal value of the LASSO parameter and the information loss by dropping variables is minimal. A high value of the LASSO parameter can contribute toward a loss of information, which may result in poor model fit. Secondly, the performance of the proposed hybrid method depends on the availability of data variables that may vary in different regions of the world due to differences in weather characteristics. The performance of the LAES hybrid model may vary with respect to changes in region (or location) of study and climatic conditions.

4. Application

Data and performance measures are described in this section. The codes of this study were written in R language version 4.1.0. The complete analysis is performed on a personal computer with an Intel Core i9-9900 CPU (32GB RAM).

4.1. Description of data

The Khyber Pakhtunkhwa province is a mountainous region, including the Tirich Mir, Lalazar, Hindu Kush, and some other mountain ranges. The changing climate of this region affects air temperature, water

flows, precipitation, and groundwater resources for irrigation systems and domestic use. These conditions make the northern area of Pakistan prone to drought or flooding due to changing environment and weather conditions.

The Kabul River begins at the Unai pass base from the Hindu Kush mountains in Afghanistan, flowing toward the east and spanning 700 km to drain into the Indus River of Pakistan (Mehmood et al. 2021). The Kabul River at Nowshera station is located at a latitude of 34°0'25"N and longitude of 71°58'50"E. The hydrometeorological regime is characterized by rain in the spring and snow in the winter. The melting of glaciers in summer is increasing each year due to high temperatures, leading to rising water levels in the river (Rasouli 2022). In addition, rainfall in the monsoon season also affects water levels in the river. The Kabul River is influenced by varying climatic conditions, which may lead to hydrometeorological hazards (i.e., heatwaves, floods or drought).

Figure 3 shows the location of the Kabul River in Pakistan. Kabul River data was collected from the Surface Water Hydrology Project (SWHP) Department of the Water and Power Development Authority of Pakistan (WAPDA) from 1st January 2005 to 31st December 2017. The data contain river discharge and meteorological variables. The meteorological variables include air temperature (minimum and maximum), pan water (minimum and maximum), relative humidity (8 AM and 5 PM), dew point (8 AM and 5 PM), evapotranspiration, and wind speed. Average temperature and precipitation have high variability across the basin. River flow has been high during the monsoon period in Pakistan, particularly in July and August. In the midst of 2005, 2010, and 2015, there was extensive flooding due to high temperatures and heavy rainfall in the region. The discharge had some missing values, which were replaced with the monthly average (mean) value. Outliers present in the data were also replaced by median of the respective month. The number of observations for each variable is 4748, approximately 365 daily values for 13 years.

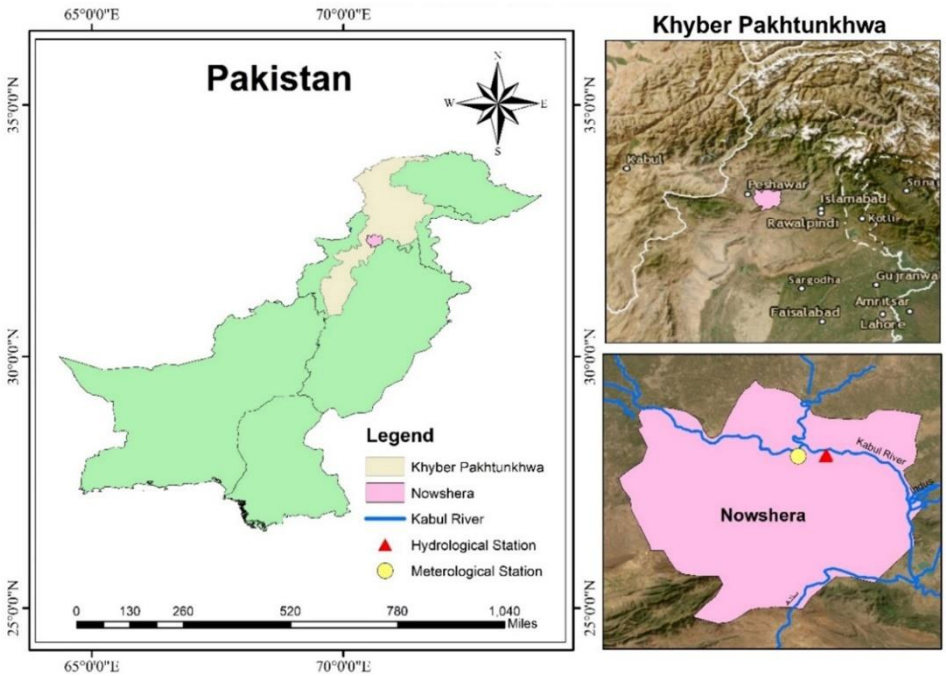


Fig. 3. Location of Kabul River in Pakistan.

Table 1 shows summary descriptions of all the variables of the Kabul River data. The air temperature (maximum), air temperature (minimum), pan water (maximum), pan water (minimum), dew point (8 AM and 5 PM), relative humidity (8 AM and 5 PM) have negatively skewed distributions, while river discharge, wind speed, evapotranspiration, precipitation and rainfall have positively skewed distributions. The average discharge in the Kabul River is 871.8 m^3/s. Figure 4 shows the Kabul River discharge series. It shows that there are non-linear relationships between river discharge and all meteorological variables.

Table 1. Descriptive summary of variables.

Variables	Units	Variables	Mean	Minimum	Maximum	Standard Deviation	Skewness
River discharge	m^3/s	y	871.8	68.7	4724.0	750.7	1.4
Air Temperature Maximum	°F	x_1	85.0	5.0	122.0	15.4	–0.3
Air Temperature Minimum	°F	x_2	64.0	5.0	110.0	13.9	–0.1
Pan Water Maximum	°F	x_3	79.9	8.0	112.0	14.3	–0.3
Pan Water Minimum	°F	x_4	72.8	16.0	106.0	13.1	–0.1
Dew point 8 AM	°F	x_5	61.2	–9.0	93.0	13.6	–0.1
Dew point 5 PM	°F	x_6	70.1	12.0	110.0	15.5	–0.1
Relative Humidity 8 AM	%	x_7	81.7	4.0	100.0	14.4	–1.7
Relative Humidity 5 PM	%	x_8	70.9	1.0	100.0	15.9	–0.9
Wind Speed	mph	x_9	30.6	0.0	170.0	24.3	1.3
Evapotranspiration	mm d^{-1}	x_{10}	5.1	0.0	27.9	5.1	0.9
Precipitation	mm d^{-1}	x_{11}	2.7	0.0	91.0	8.2	4.7
Rainfall	mm d^{-1}	x_{12}	3.3	0.0	161.0	11.5	6.1

The data variables were normalized using the following (Duan et al. 2021):

$$z_{normal} = \frac{z - z_{min}}{z_{max} - z_{min}} \tag{11}$$

where z is the original data variable, z_{normal} is the normalized data variable, z_{min} is the minimum value, and z_{max} is the maximum value of the original data variable. After normalization, the dataset is divided into two parts, where 80% of the data are used for training and the remaining 20% for testing (Kisi et al. 2021; Shabbir et al. 2022). The performance of models is evaluated by 5-fold cross-validation using different performance evaluation measures and the average results of these indicators for training and testing data.

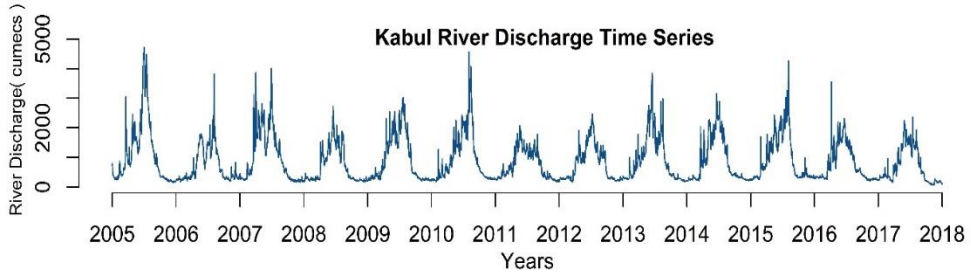


Fig. 4. Kabul River discharge series.

4.2. Performance evaluation measures

The prediction performance of the proposed hybrid framework is evaluated on both training and testing datasets. A 5-fold cross-validation approach and different goodness-of-fit measures are selected to assess the performance of models. These measures include root mean square error (***RMSE***), mean absolute percentage error (***MAPE***), root-relative square error (***RRSE***), mean absolute error (***MAE***) and coefficient of determination (R^2). These measures are given as follows (Zeinali et al. 2020; Shabbir et al. 2023):

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(y_j - \hat{y}_j\right)^2} \tag{12}$$

$$MAPE = \frac{100}{n}\sum_{j=1}^{n}\left|\frac{y_j - \hat{y}_j}{y_j}\right| \tag{13}$$

$$RRSE = \sqrt{\frac{\sum_{j=1}^{n}\left(y_j - \hat{y}_j\right)^2}{\sum_{j=1}^{n}\left(\hat{y}_j - \bar{\hat{y}}\right)^2}} \tag{14}$$

$$MAE = \frac{1}{n}\sum_{j=1}^{n}\left|y_j - \hat{y}_j\right| \tag{15}$$

$$R^2 = 1 - \left(\frac{\sum_{j=1}^{n}\left(y_j - \hat{y}_j\right)^2}{\sum_{j=1}^{n}\left(y_j - \bar{y}\right)^2}\right) \tag{16}$$

where n denotes the total number of observations, y_j denotes the actual observation and $\hat{y}_j$ denotes the predicted values. The terms $\bar{y}$ and $\bar{\hat{y}}$ denote the average of observed and predicted values, respectively.

To compare the performance of the different models for river discharge prediction, the improvement percentages of ***RMSE***, ***MAPE***, ***RRSE***, and ***MAE*** are also used and are given as:

$$P_{RMSE} = \frac{\left(RMSE_i - RMSE_j\right)}{RMSE_i} \times 100 \tag{17}$$

$$P_{MAPE} = \frac{\left(MAPE_i - MAPE_j\right)}{MAPE_i} \times 100 \tag{18}$$

$$P_{RRSE} = \frac{\left(RRSE_i - RRSE_j\right)}{RRSE_i} \times 100 \tag{19}$$

$$P_{MAE} = \frac{\left(MAE_i - MAE_j\right)}{MAE_i} \times 100 \tag{20}$$

$$P_{R^2} = \frac{\left(R_i^2 - R_j^2\right)}{R_i^2} \times 100 \tag{21}$$

where subscript i denotes the competing model and subscript j indicates the proposed LAES hybrid model. These quantities indicate the degree of improvement in the prediction performance of one model relative to another model (Duan et al. 2021).

The Diebold-Mariano (DM) test has been widely used in literature to compare the forecast accuracy of two models (Silva et al. 2021; Shabbir et al. 2022). The null and alternative hypotheses are:

$$H_0: E[d_t] \geq 0 \quad (22)$$

$$H_1: E[d_t] < 0$$

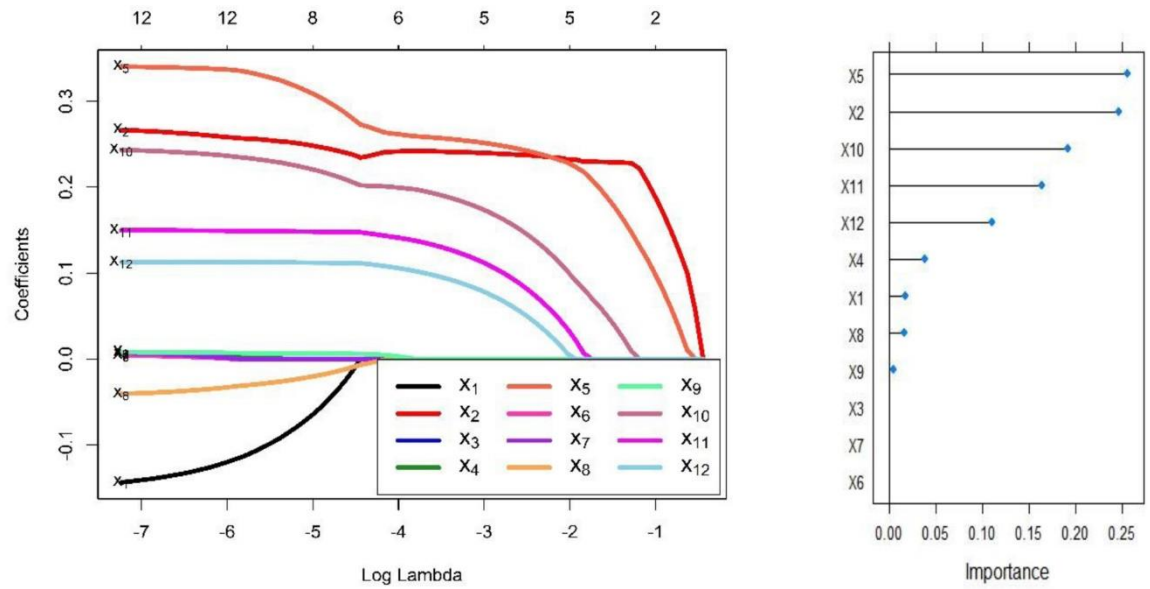
where d_t is the difference loss function, i.e., $d_t = e_{1t} - e_{2t}$, e_{1t} and e_{2t} denotes the set of prediction errors of two competing models. The test statistic is $DM = \frac{\bar{d}}{\left(2\pi \hat{f}_d(0)/m\right)^{1/2}}$, where m is the length of prediction errors, $\bar{d} = \frac{1}{m}\sum_{t=1}^{m}(d_t)$ is the average loss differential between two predictions, The DM statistic follows the standard normal distribution (i.e. $N(0,1)$) and $\hat{f}_d(0)$ is the spectral density. The $2\pi\hat{f}_d(0)$ is the consistent estimator of the asymptotic variance. The null hypothesis (H_0) is rejected if $DM < -Z_\alpha$, where Z is the standardized normal percentile with probability α.

In this study, a one-sided DM test is used to compare the prediction accuracy of the LAES model with six models. This test uses subscript 1 for the proposed LAES model and subscript 2 for the competing models. This test is applied using the squared differences loss function to compare models at a 1% significance level. If $DM < -2.326$, we will reject the null hypothesis. The proposed LAES hybrid model is compared with MLR, SVM, ANN, LASSO-MLR, LASSO-SVM and LASSO-ANN models in this study.

5. Results and discussion

In the proposed hybrid framework, LASSO is employed to choose meteorological variables that have significant roles in predicting Kabul River discharge. This step eliminates insignificant variables and constructs a better prediction model. Using LASSO, we retain only important input variables that influence the river discharge of the Kabul River. The results of the LASSO using $\lambda = 0.010$ are shown in Figure 5a. LASSO eliminates three meteorological variables, i.e., pan water (maximum), relative humidity (8 AM) and relative humidity (5 PM). The air temperature (minimum and maximum), dew point (8 AM), relative humidity (5 PM), rainfall, precipitation, wind speed, and evapotranspiration are significant variables for prediction of river discharge. These variables. $\{x_1, x_2, x_4, x_5, x_8, x_9, x_{10}, x_{11}, x_{12}\}$ are used as inputs to LASSO-based models. Bui et al. (2019) stated that dew point is a component of the temperature variable. The precipitation and rainfall factors are dependent on the air temperature and are indirectly associated with the dew point.

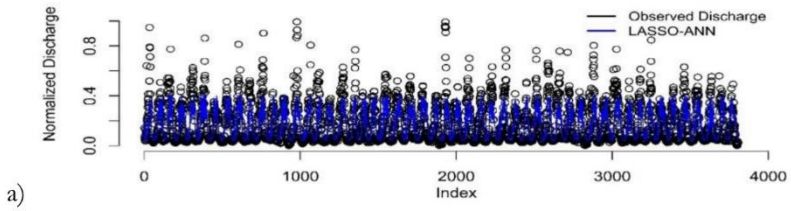
Figure 5b shows that dew point (8 AM) is the most significant variable for predicting river discharge. These variables selected by LASSO are used as inputs to the ANN model in the proposed hybrid framework. The prediction results by LASSO-ANN in the first round of the training phase are demonstrated in Figure 6a. The results of the remaining rounds are given in supplementary materials.



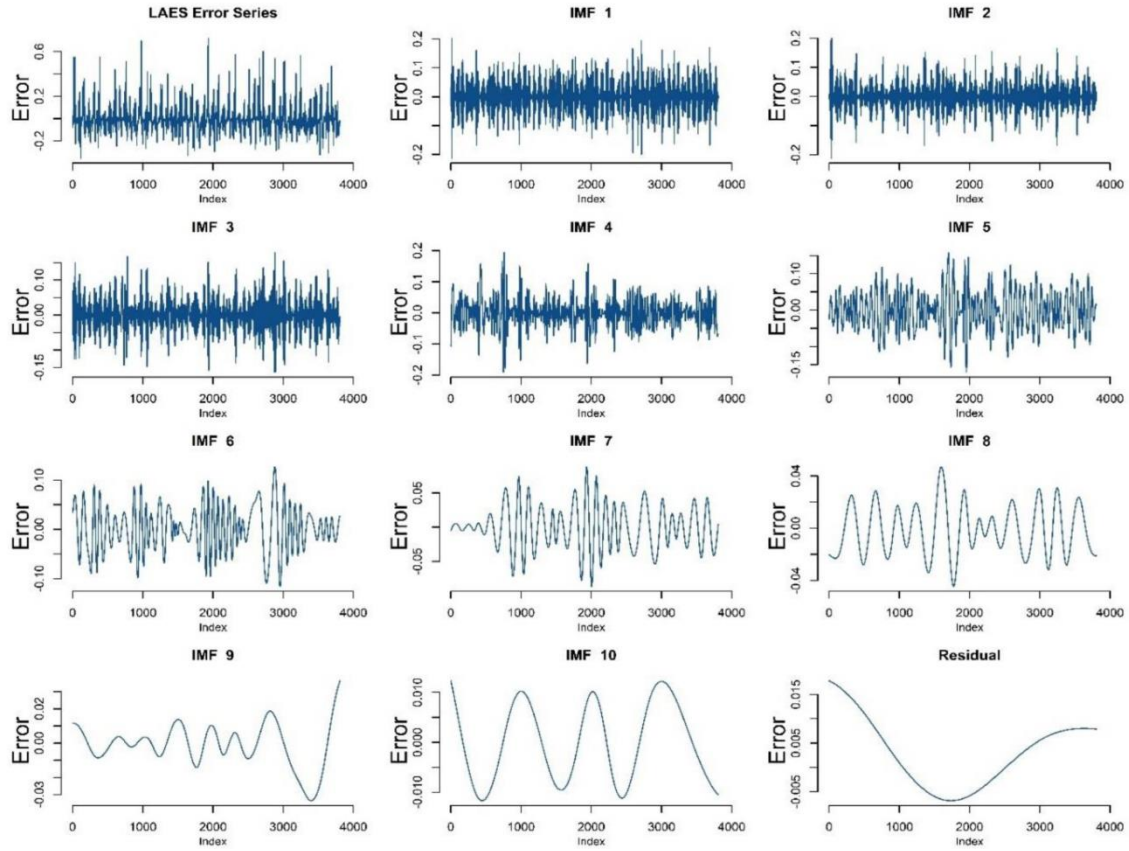
a) b)

Fig. 5. The variable screening (a) and variable importance (b) results from LASSO on Kabul River data.

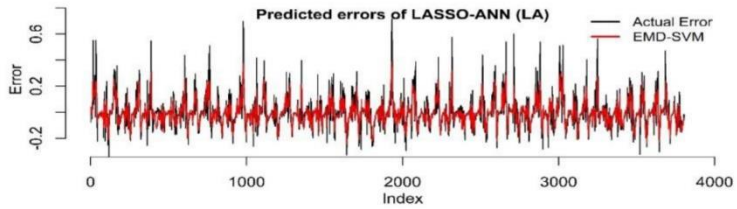
After ANN model training, the predictions and error series are obtained. Stationarity of the error series is checked using an augmented Duckey-Fuller (ADF) test. The Dickey-Fuller statistic is –3.3875, indicating that the error series in the first round is non-stationary at the 5% level of significance. The results of ADF tests of the remaining rounds are provided in the supplementary materials. Next, the EMD decomposes the error series into $IMFs$ and residuals as shown in Figure 6b. Then, the SVM is applied to model each component of the decomposed error series. The sub-series predictions are obtained and aggregated as the final error prediction shown in Figure 6c. The final prediction of river discharge is computed by adding the predicted errors and predicted river discharge. Lastly, the actual predicted values of river discharge are obtained by anti-normalization using Eq. 11. Figure 7 shows the predicted discharge plot in the testing phase in the first round. It reveals that the proposed LAES hybrid models have the closest predictions to the observed river discharge.



a)



b)



c)

Fig. 6. Prediction results of LASSO-ANN: (a) error decomposition using EMD; (b) modeling of decomposed components (c) in the first round of training the phase for the Kabul River.

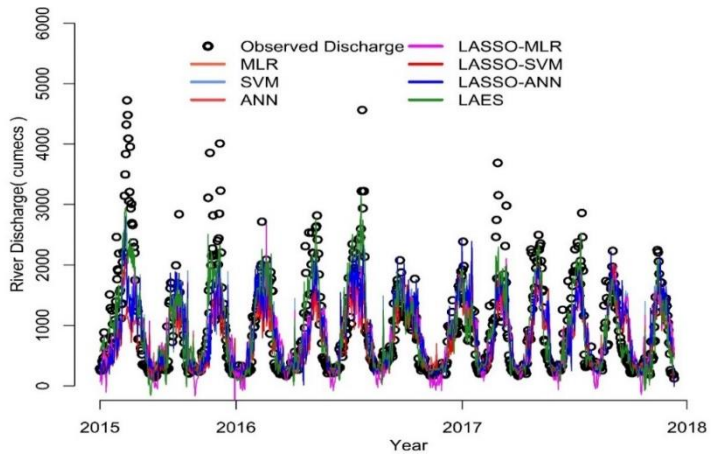


Fig. 7. Prediction plot of Kabul River discharge on test data of first round.

5.1. Comparison of model accuracy

The daily river discharge was estimated against various meteorological variables. Table 2 presents the training and testing phase results of daily river discharge prediction. In the training phase, the MLR model is the worst performer among all models (*RMSE* = 533.822 m^3/s, *MAE* = 378.003 m^3/s, *RRSE* = 0.711, *MAPE* = 66.786% and R^2 = 49.4%). However, the SVM and ANN models performed relatively better than the MLR model. For example, in the training phase, the *RMSE* for MLR, SVM and ANN models is 533.822 m^3/s, 511.262 m^3/s and 507.015 m^3/s, respectively. Similar to this study, Zhang et al. (2018b) found that the MLR model is the worst performer for predicting river discharge in the East River basin of China. Some other studies found that the non-linear features of river discharge are captured well by SVM and ANN models (see Poul et al. 2019 and Meng et al. 2021).

Comparing the performance of models based on meteorological variables selected by LASSO, we found that the performance of all models is improved in most of the instances. The performance of the LASSO-MLR model is better than the MLR model in the testing phase (*RMSE* = 543.559 m^3/s, *MAE* = 381.889 m^3/s, *RRSE* = 0.725, *MAPE* = 67.758% and R^2 = 47.4%). However contrary results are obtained in the training phase, in which the LASSO-MLR model has a similar fit to the MLR model. The prediction ability of LASSO-ANN and LASSO-SVM is better than ANN and SVM models respectively. Mehr and Gandomi (2021) found that LASSO improved the predictive ability of a multi-stage genetic programming model by reducing the number of genes for predicting river discharge in the Sedre River of Turkey. In the training phase, the proposed LAES hybrid model has the best fit for river discharge data based on various performance criteria (*RMSE* = 302.952 m^3/s, *MAE* = 201.022 m^3/s, *RRSE* = 0.404, *MAPE* = 30.494% and R^2 = 83.7%).

Comparing the results in the testing phase, the MLR model has the poorest performance when all the meteorological variables were used as inputs (*RMSE* = 554.277 m^3/s, *MAE* = 383.541 m^3/s, *RRSE* = 0.739, *MAPE* = 68.134% and R^2 = 45.3%). The use of LASSO for dimension reduction enhanced the performance of MLR, SVM, and ANN models in the testing phase. Judging by *RMSE*, *RRSE* and R^2, the

LASSO-ANN model is a better performer than the LASSO-SVM and LASSO-MLR models. However, comparing *MAE* and *MAPE*, the LASSO-SVM model performs better than the LASSO-MLR and LASSO-ANN hybrid models (*MAE* = 307.124 m^3/s and *MAPE* = 39.394%). The proposed LAES model outperforms all competing models in the testing phase (i.e., *RMSE* = 337.143 m^3/s, *MAE* = 218.353 m^3/s, *RRSE* = 0.449, *MAPE* = 32.354% and R^2 = 79.8%). Overall, the proposed LAES hybrid model has higher prediction accuracy than single and LASSO-based ANN, SVM, and MLR models.

Figure 8a presents the goodness-of-fit measure values of all the models considered in the study in both training and testing data. It shows that the proposed LAES hybrid model has the highest accuracy among all models considered in the study. The Taylor diagram in Figure 8b shows that the proposed LAES model is the most efficient among all models considered in predicting daily river discharge based on its relationship with meteorological variables.

Table 2. Performance analysis of the proposed model with different models.

Models	*RMSE* (m^3/s)	*MAE* (m^3/s)	*RRSE*	*MAPE* (%)	R^2
	Training				
MLR	533.822	378.003	0.711	66.786	0.494
SVM	511.262	309.783	0.681	39.372	0.536
ANN	507.015	334.263	0.676	50.508	0.542
LASSO-MLR	534.091	378.263	0.712	66.878	0.494
LASSO-SVM	469.381	280.664	0.625	35.972	0.609
LASSO-ANN	456.981	302.596	0.609	45.686	0.629
LAES	**302.952**	**201.022**	**0.404**	**30.494**	**0.837**
	Testing				
MLR	554.277	383.541	0.739	68.134	0.453
SVM	527.427	324.443	0.702	41.814	0.505
ANN	524.117	342.108	0.699	51.618	0.511
LASSO-MLR	543.559	381.889	0.725	67.758	0.474
LASSO-SVM	499.947	307.124	0.666	39.394	0.556
LASSO-ANN	497.256	324.178	0.664	48.056	0.559
LAES	**337.143**	**218.353**	**0.449**	**32.354**	**0.798**
Note: Bold values represent minimum values in each column					

The improvements of the proposed LAES hybrid model are shown in Table 3 in terms of P_{RMSE}, P_{MAE}, P_{RRSE}, P_{MAPE} and P_{R^2} for both training and testing phases. The proposed LAES hybrid model has 43.3%, 40.7% and 40.3% lower *RMSE* than the MLR, SVM, and ANN models, respectively, in the training phase. The findings indicate that the MLR model is least efficient for non-linear data, consistent with the findings of Zhang et al. (2018b).

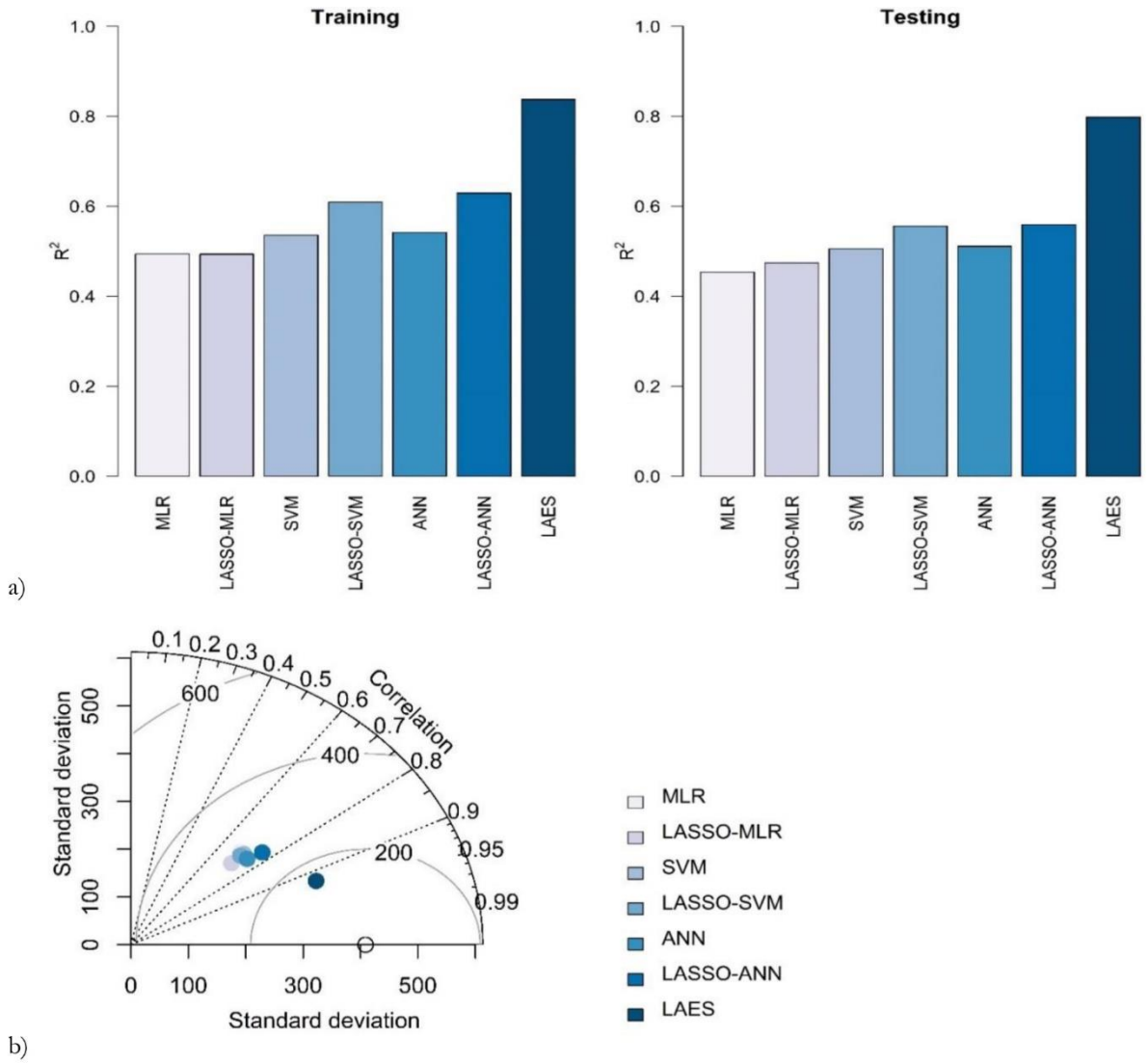


Fig. 8. Prediction results of models in training and testing phase (a) and Taylor diagram (b) for the Kabul River.

Comparing the LAES model to LASSO-based models, we found that their promoting improvements were lower compared to single MLR, SVM, and ANN models in the majority of the scenarios. During testing, the reduction in *RMSE* by LASSO-MLR and MLR models is 38% and 39.2%, respectively. Similarly, the improvements by the LAES model vs. the SVM model (36.1%) are higher than the LAES model vs. the SVM model (32.6%). The proposed LAES hybrid model has 68.2%, 43.6%, and 42.7% better prediction accuracy than the LASSO-MLR, LASSO-SVM, and LASSO-ANN models. Kang et al. (2023) also stated that LASSO helps enhance the predictive performance of monthly run-off, which is influenced by meteorological events.

Generally, the proposed LAES hybrid model has promising predictions compared to all six models. During the training phase, the MAE of LAES compared to MLR, SVM, ANN, LASSO-MLR, LASSO-SVM, and LASSO-ANN decreased by 46.8%, 35.1%, 39.9%, 46.9%, 28.4%, and 33.6% respectively. These results are in agreement with the findings of Duan et al. (2021). They reported that the decomposition-based error correction approach significantly improves the accuracy of models.

Table 3. Improved percentage (%) of proposed model versus other models.

Models	Training					Testing				
	P_{RMSE}	P_{MAE}	P_{RRSE}	P_{MAPE}	P_{R^2}	P_{RMSE}	P_{MAE}	P_{RRSE}	P_{MAPE}	P_{R^2}
LAES vs. MLR	43.3	46.8	43.3	54.3	–69.4	39.2	43.1	39.2	52.5	–76.1
LAES vs. SVM	40.7	35.1	40.8	22.6	–56.2	36.1	32.7	36.0	22.6	–57.9
LAES vs. ANN	40.3	39.9	40.3	39.6	–54.5	35.7	36.2	35.7	37.3	–56.1
LAES vs. LASSO-MLR	43.3	46.9	43.3	54.4	–69.6	38.0	42.8	38.0	52.3	–68.2
LAES vs. LASSO-SVM	35.5	28.4	35.5	15.2	–37.5	32.6	28.9	32.6	17.9	–43.6
LAES vs. LASSO-ANN	33.7	33.6	33.7	33.3	–33.0	32.2	32.6	32.3	32.7	–42.7

The DM test results on the testing data of Kabul River discharge are given in Table 4. The null hypothesis for all competing models is rejected at a 1% significance level. Thus, the prediction accuracy of the proposed hybrid LAES model is higher than the six benchmark models. Therefore, the DM test confirms that the proposed LAES hybrid model has higher prediction accuracy than the competing models in predicting river discharge.

Table 4. DM test of proposed hybrid model versus different models on the testing dataset.

Model	MLR	SVM	ANN	LASSO-MLR	LASSO-SVM	LASSO-ANN
DM-value	–9.118***	–8.688***	–10.256***	–10.702***	–8.434***	–8.299***
*** significant at a 1% significance level						

6. Conclusion

In this study, a new hybrid framework named LAES (LASSO-ANN-EMD-SVM) is introduced for modeling river discharge using information from several meteorological variables. The proposed hybrid model is a composite of a variable selection approach with an artificial neural network and error correction method. The application of the LAES hybrid framework is illustrated using the data from the Kabul River in Pakistan. The effectiveness and predictive ability of the proposed framework are compared with six models using different performance measures. The numerical findings reveal that the LAES hybrid model has better prediction performance than the single and LASSO-based MLR, SVM, and ANN models. Judging by *RRSE*, the LAES hybrid model has 43.3%, 40.8%, 40.3%, 43.3%, 35.5%, and 33.7% lower prediction errors than MLR, SVM, ANN, LASSO-MLR, LASSO-SVM and LASSO-ANN models respectively. The Diebold-Mariano test shows that the proposed LAES model has higher prediction accuracy than all competing models in the study. The proposed LAES model can serve as a successful tool for river discharge prediction by considering the impact of meteorological variables. In this study, we have used the LAES hybrid model for regression modeling only, but it can be applied for time series prediction of hydrological variables (such as river inflow and monthly run-off). For future research, new hybrid models can be developed by considering (i) relevance vector machine (RVM) or deep learning models such as multilayer perceptron (MLP) in modeling; and (iii) using decomposition techniques such as ensemble EMD, complete EEMD (CEEMD), and variational mode decomposition (VMD) methods in the error correction stage. The proposed LAES model can serve as a successful tool for river discharge prediction of catchment areas of different areas of the world for efficient planning of water resources.

Acknowledgments

We acknowledge the SWHP department of WAPDA, Pakistan, for providing the data required for this research work.

References

Adnan R.M., Mostafa R.R., Elbeltagi A., Yaseen Z.M., Shahid S., Kisi O., 2022, Development of new machine learning model for streamflow prediction: case studies in Pakistan, Stochastic Environmental Research and Risk Assessment, 36, 999-1033, DOI: 10.1007/s00477-021-02111-z.

Ali S., Shahbaz M., 2020, Streamflow forecasting by modeling the rainfall–streamflow relationship using artificial neural networks, Modeling Earth Systems and Environment, 6, 1645-1656, DOI: 10.1007/s40808-020-00780-3.

Baydaroğlu Ö., Koçak K., Duran K., 2018, River flow prediction using hybrid models of support vector regression with the wavelet transform, singular spectrum analysis and chaotic approach, Meteorology and Atmospheric Physics, 130, 349-359, DOI: 10.1007/s00703-017-0518-9.

Bui A., Johnson F., Wasko C., 2019, The relationship of atmospheric air temperature and dew point temperature to extreme rainfall, Environmental Research Letters, 14 (7), DOI: 10.1088/1748-9326/ab2a26.

Cheng K., Wei S., Fu Q., Li T., 2019, Adaptive management of water resources based on an advanced entropy method to quantify agent information, Journal of Hydroinformatics, 21 (3), 381-396, DOI: 10.2166/hydro.2019.007.

Dariane A., Azimi S., 2018, Streamflow forecasting by combining neural networks and fuzzy models using advanced methods of input variable selection, Journal of Hydroinformatics, 20 (2), 520-532. DOI: 10.2166/hydro.2017.076.

Duan J., Zuo H., Bai Y., Duan J., Chang M., Chen B., 2021, Short-term wind speed forecasting using recurrent neural networks with error correction, Energy, 217, DOI: 10.1016/j.energy.2020.119397.

Elbeltagi A., Nunno F.D., Kushwaha N.L., Marinis G.D., Granata F., 2022, River flow rate prediction in the Des Moines watershed (Iowa, USA): a machine learning approach, Stochastic Environmental Research and Risk Assessment, 36, 3835-3855, DOI: 10.1007/s00477-022-02228-9.

Emadi A., Sobhani R., Ahmadi H., Boroomandnia A., Zamanzad-Ghavidel S., Azamathulla H.M., 2022, Multivariate modeling of river water withdrawal using a hybrid evolutionary data-driven method, Water Supply, 22 (1), 957-980, DOI: 10.2166/ws.2021.224.

Gleason C.J., Smith L.C., Lee J., 2014, Retrieval of river discharge solely from satellite imagery and at-many-stations hydraulic geometry: Sensitivity to river form and optimization parameters, Water Resources Research, 50 (12), 9604-9619, DOI: 10.1002/2014WR016109.

Hu J., Wu Y., Sun P., Zhao F., Sun K., Li T., Sivakumar B., Qiu L., Sun Y., Jin Z., 2022, Predicting long-term hydrological change caused by climate shifting in the 21st century in the headwater area of the Yellow River basin, Stochastic Environmental Research and Risk Assessment, 36, 1651-1668, DOI: 10.1007/s00477-021-02099-6.

Huang N.E., Shen Z., Long S.R., Wu M.C., Shih H.H., Zheng Q., Yen N.C., Tung C.C., Liu H.H., 1998, The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proceedings of the Royal Society A. London, 454 (1971), 903-995, DOI: 10.1098/rspa.1998.0193.

Jungsheng C., Dejie Y., Yu Y., 2006, A fault diagnosis approach for roller bearings based on EMD method and AR model, Mechanical Systems Signal Processing, 20 (2), 350-362, DOI: 10.1016/j.ymssp.2004.11.002.

Kachrimanis K., Kamaryan V., Malamataris S., 2003, Artificial neural networks (ANNs) and modeling of powder flow, International Journal of Pharmaceutics, 250 (1), 13-23, DOI: 10.1016/S0378-5173(02)00528-8.

Kang Y., Cheng X., Chen P., Zhang S., Yang Q., 2023, Monthly runoff prediction by a multivariate hybrid model based on decomposition-normality and Lasso regression, Environmental Science and Pollution Research, 30, 27743-27762, DOI: 10.1007/s11356-022-23990-x.

Kisi O., Alizamir M., Shiri J., 2021, Conjunction model design for intermittent streamflow forecasts: extreme learning machine with discrete wavelet transform, [in:] Intelligent Data Analytics for Decision-Support Systems in Hazard Mitigation, Springer, Singapore, 171-181, DOI: 10.1007/978-981-15-5772-9_9.

Lei Y., He Z., Zi Y., 2003, Application of the EEMD method to rotor fault diagnosis of rotating machinery, Mechanical Systems and Signal Processing, 23 (4), 1327-1338, DOI: 10.1016/j.ymssp.2008.11.005.

Luo X., Xiaohui Y., Zhu S., Xu Z., Meng L., Peng J., 2019, A hybrid support vector regression framework for streamflow forecast, Journal of Hydrology, 568, 184-193, DOI: 10.1016/j.jhydrol.2018.10.064.

Malik A., Tikhamarine Y., Souag-Gamane D., Kisi O., Pham Q.B., 2020, Support vector regression optimized by meta-heuristic algorithms for daily streamflow prediction, Stochastic Environmental Research and Risk Assessment, 34, 1755-1773, DOI: 10.1007/s00477-020-01874-1.

Mangukiya N.K., Mehta D.J., Jariwala R., 2022, Flood frequency analysis and inundation mapping for lower Narmada basin, India, Water Practice and Technology, 17(2), 612-622, DOI: 10.2166/wpt.2022.009.

Mehmood A., Jia S., Lv A., Zhu W., Mehmood R., Saifullah M., Adnan M.R., 2021, Detection of spatial shift in flood regime of the Kabul river basin in Pakistan, causes, challenges, and opportunities, Water, 13 (9), 1296-1301, DOI: 10.3390/w13091276.

Mehr A.D., Gandomi A.H., 2021, MSGP-LASSO: An improved multi-stage genetic programming model for streamflow prediction, Information Sciences, 561, 181-195, DOI: 10.1016/j.ins.2021.02.011.

Mehta D.J., Eslamian S., Prajapati K., 2022, Flood modelling for a data-scare semi-arid region using 1-D hydrodynamic model: a case study of Navsari Region, Modeling Earth Systems and Environment, 8 (2), 2675-2685, DOI: 10.1007/s40808-021-01259-5.

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 Resources Management, 35, 1321-1337, DOI: 10.1007/s11269-021-02786-7.

Poul A.K., Shourian M., Ebrahimi H., 2019, A comparative study of MLR, KNN, ANN and ANFIS models with wavelet transform in monthly stream flow prediction, Water Resources Management, 33, 2907-2923, DOI: 10.1007/s11269-019-02273-0.

R Core Team, 2022, R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/.

Rasouli H., 2022, Climate change impacts on water resource and air pollution in Kabul sub-basins, Afghanistan, Advances in Geological and Geotechnical Engineering Research, 4 (1), 11-27, DOI: 10.30564/agger.v4i1.4312.

Rasouli K., Hsieh W.W., Cannon A.J., 2012, Daily streamflow forecasting by machine learning methods with weather and climate inputs, Journal of Hydrology, 414-415, 284-293, DOI: 10.1016/j.jhydrol.2011.10.039.

Ravindran S.M., Bhaskaran S.K., Ambat S.K., 2021, A deep neural network architecture to model reference evapotranspiration using a single input meteorological parameter, Environmental Processes, 8, 1567-1599, DOI: 10.1007/s40710-021-00543-x.

Saidi H., Dresti C., Manca D., Ciampittiello M., 2018, Quantifying impacts of climate variability and human activities on the streamflow of an Alpine river, Environmental Earth Sciences, 77, DOI: 10.1007/s12665-018-7870-z.

Shabbir M., Chand S., Iqbal F., 2022, A Novel Hybrid Method for River Discharge Prediction. Water Resources Management, 36, 253-272. DOI: https://doi.org/10.1007/s11269-021-03026-8.

Shabbir M., Chand S., Iqbal F., 2023, Prediction of river inflow of the major tributaries of Indus river basin using hybrids of EEMD and LMD methods, Arabian Journal of Geosciences, 16, 257, DOI: 10.1007/s12517-023-11351-y.

Shabbir M., Chand S., Iqbal F., 2024, Novel hybrid and weighted ensemble models to predict river discharge series with outliers, Kuwait Journal of Science, 51 (2), DOI: 10.1016/j.kjs.2024.100188.

Silva R.G., Ribeiro M.H., Moreno S.R., Mariani V.C., Coelho L.D., 2021, A novel decomposition-ensemble learning framework for multi-step ahead wind energy forecasting, Energy, 216, DOI: 10.1016/j.energy.2020.119174.

Tibshirani R., 1996, Regression Shrinkage and Selection via the Lasso, Journal of Royal Statistical Society: Series B (Methodological), 58 (1), 267-288, DOI: 10.1111/j.2517-6161.1996.tb02080.x.

Vapnik V., 1995, The nature of statistical learning theory, Springer, New York, 188 pp., DOI: 10.1007/978-1-4757-2440-0.

Wang J., Li Y., 2018, Multi-step ahead wind speed prediction based on optimal feature extraction, long short term memory neural network and error correction strategy, Applied Energy, 230, 429-443, DOI: 10.1016/j.apenergy.2018.08.114.

Yan S., Wang X., Zhang Y., Liu D., Yi Y., Li C., Liu Q., Yang Z., 2020, A hybrid PCA-GAM model for investigating the spatiotemporal impacts of water level fluctuations on the diversity of benthic macroinvertebrates in Baiyangdian Lake, North China, Ecological Indicators, 116, DOI: 10.1016/j.ecolind.2020.106459.

Yousuf I., Ghumman A.R., Hashmi H.N., 2017, Optimally sizing small hydropower project under future projected flows, KSCE Journal of Civil Engineering, 21, 1964-1978, DOI: 10.1007/s12205-016-1043-y.

Zeinali M., Azari A., Heidari M., 2020, Multiobjective optimization for water resource management in low-flow areas based on a coupled surface water–groundwater model, Journal of Water Resources Planning and Management, 146 (5), DOI: 10.1061/(ASCE)WR.1943-5452.0001189.

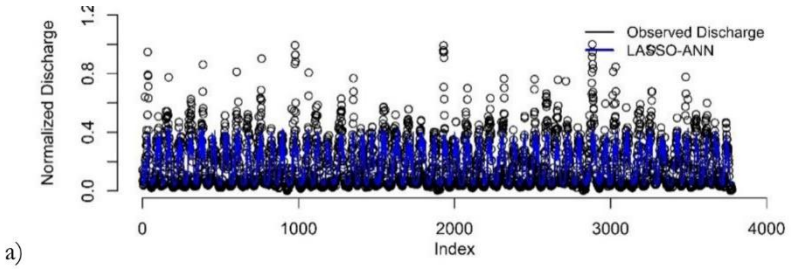
Zhang X., Liu P., Zhao Y., Deng C., Li Z., Xiong M., 2018a, Error correction-based forecasting of reservoir water levels: Improving accuracy over multiple lead times, Environmental Modelling and Software, 104, 27-39, DOI: 10.1016/j.envsoft.2018.02.017.

Zhang Z., Zhang Q., Singh V.P., 2018b, Univariate streamflow forecasting using commonly used data-driven models: Literature review and case study, Hydrological Sciences Journal, 63 (7), 1091-1111, DOI: 10.1080/02626667.2018.1469756.

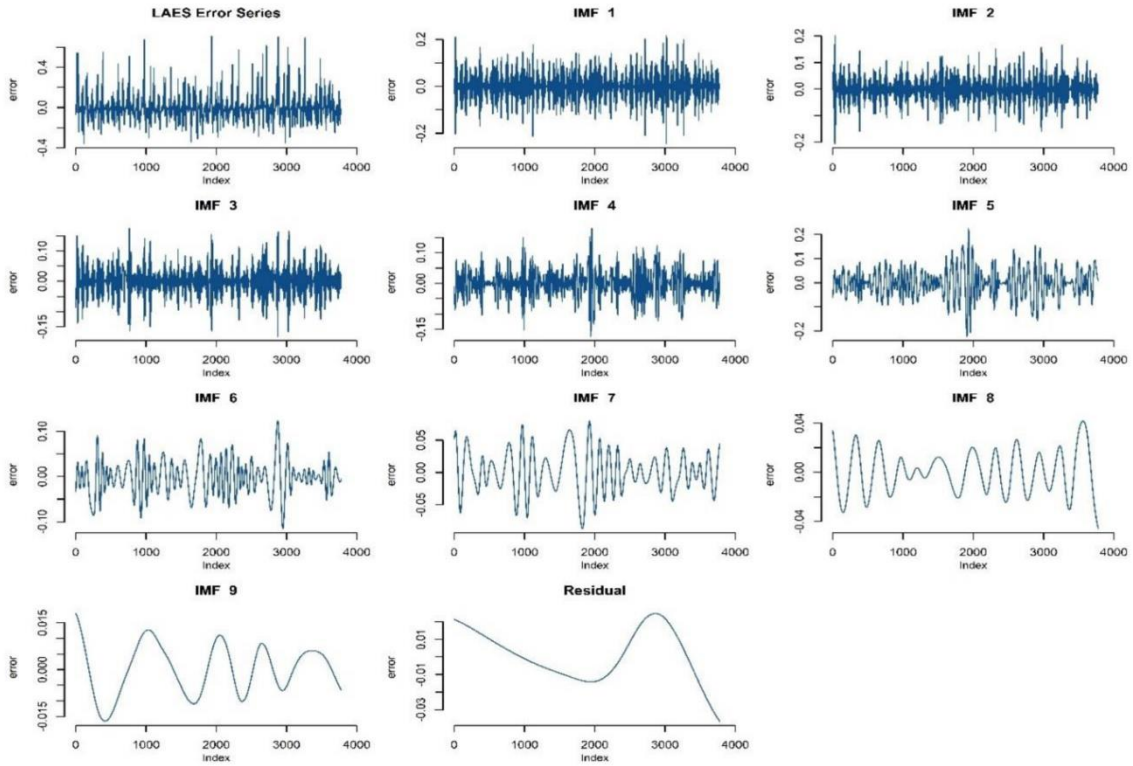
Appendix

ADF test results. H_0 – the time series contains unit root and is non-stationary; H_1 – the time series is stationary.

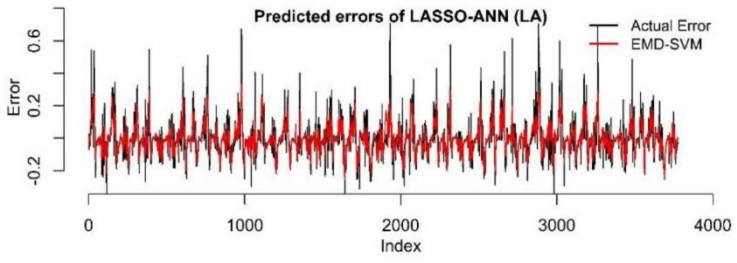
Fold	2	3	4	5
Duckey-Fuller Statistic	–3.2011*	–3.1671*	–3.0869*	–3.3423*
p-value	0.08769	0.0935	0.1327	0.06334



a)

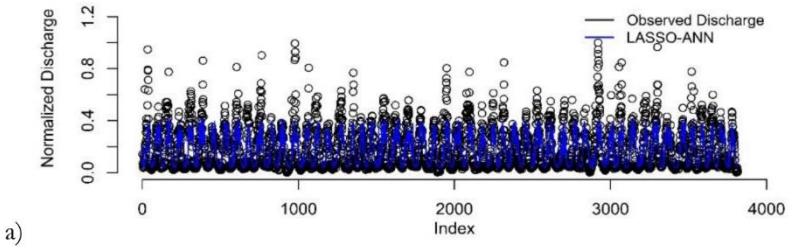


b)

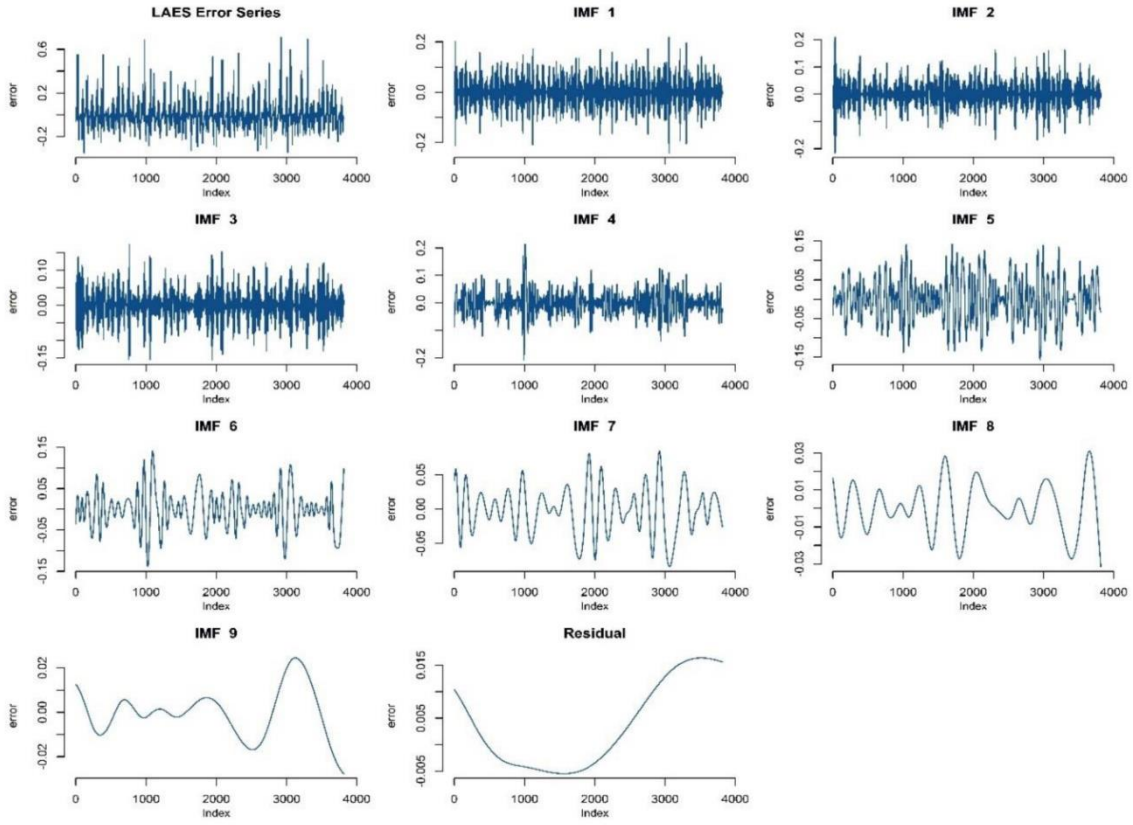


c)

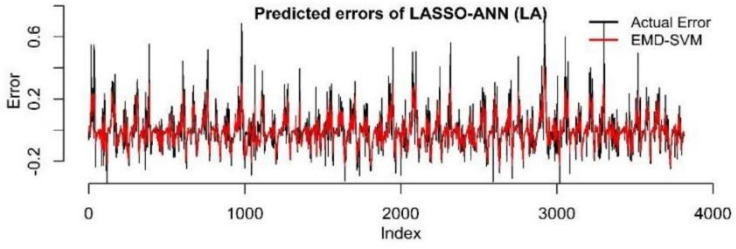
Fig. S1. Prediction results of LASSO-ANN (a) Error decomposition using EMD (b), modeling of decomposed components (c) in the second fold of training phase of Kabul River.



a)

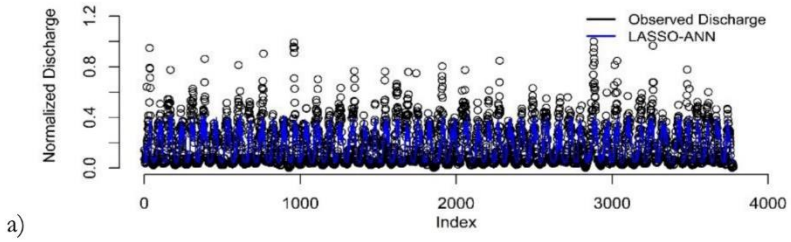


b)

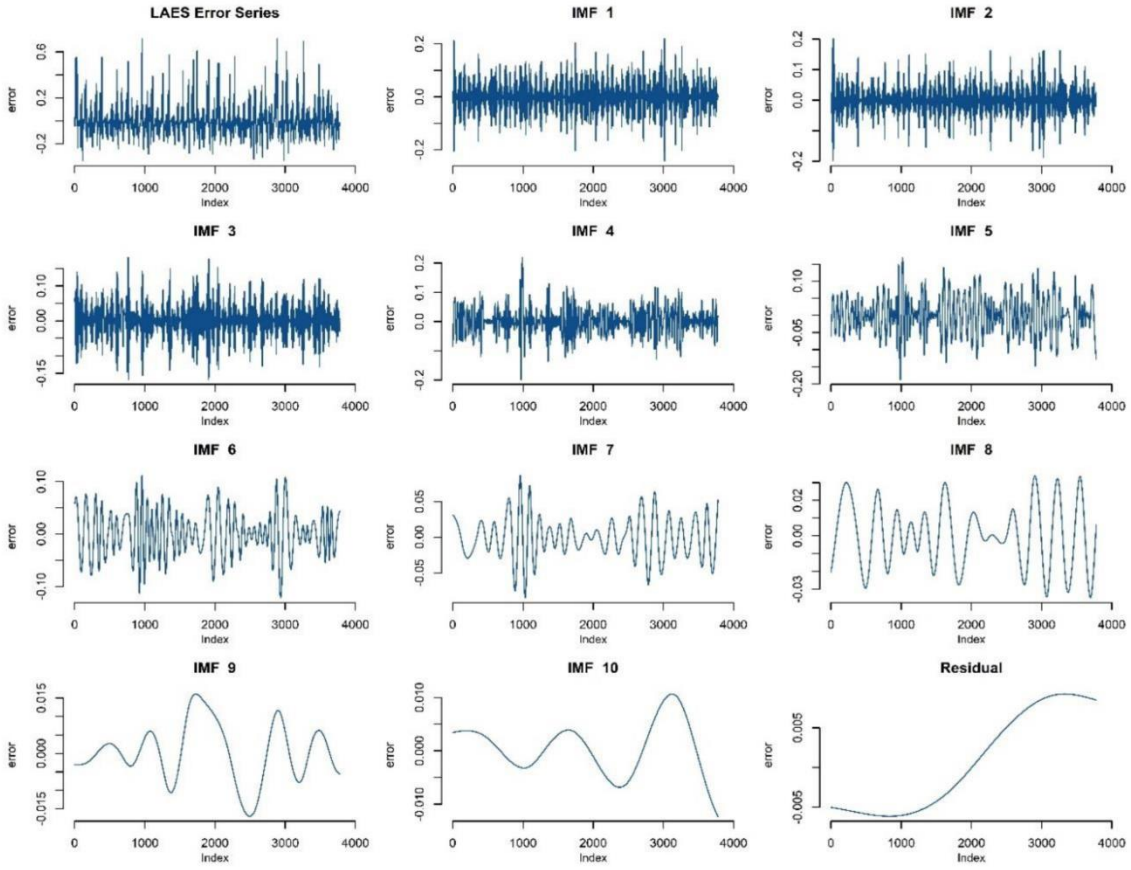


c)

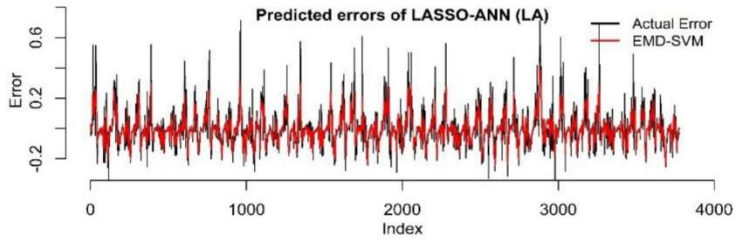
Fig. S2. Prediction results of LASSO-ANN (a) Error decomposition using EMD (b), modeling of decomposed components (c) in the third fold of training phase of Kabul River.



a)

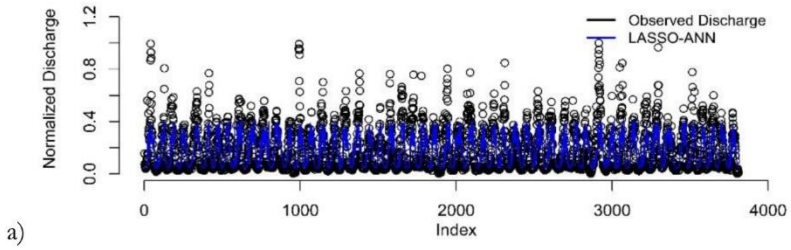


b)

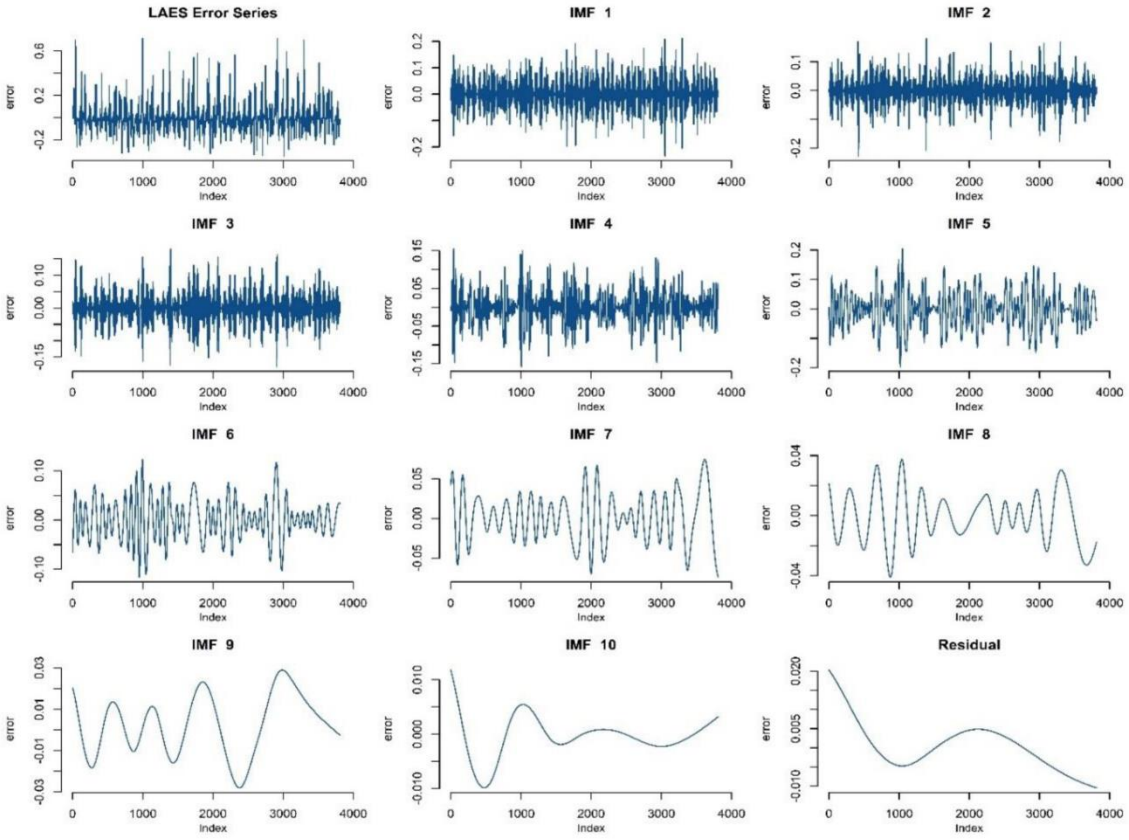


c)

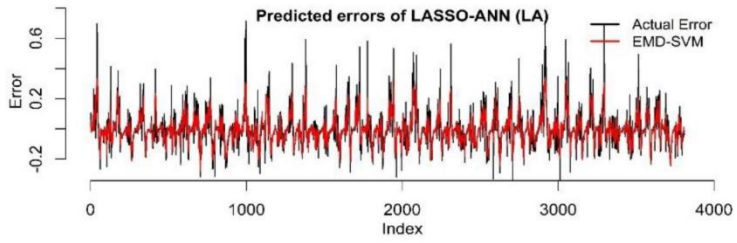
Fig. S3. Prediction results of LASSO-ANN (a) Error decomposition using EMD (b), modeling of decomposed components (c) in the fourth fold of training phase of Kabul River.



a)



b)



c)

Fig. S4. Prediction results of LASSO-ANN (a) Error decomposition using EMD (b), modeling of decomposed components (c) in the fifth fold of training phase of Kabul River.