Optimum Application of Hybrid Data Driven Models With Two Step Verification Method at Mangla Watershed, Pakistan

Muhammad Tayyab (mtayyab@ctgu.edu.cn)
China Three Gorges University https://orcid.org/0000-0002-0638-3044
Dong Xiaohua
China Three Gorges University
Muhammad Sibtain
China Three Gorges University
Ijaz Ahmad
University of Engineering and Technology
Aqeela Zahra
Wuhan University of Technology - Mafangshan Campus: Wuhan University of Technology
Muhammad Imran Azam
China Three Gorges University

Research Article

Keywords: streamflow prediction, data driven methods, decomposition methods, input variables, extreme value analysis, least value analysis

Posted Date: October 21st, 2021

DOI: https://doi.org/10.21203/rs.3.rs-989802/v1

Optimum Application of Hybrid Data Driven Models with Two Step Verification Method at Mangla watershed, Pakistan

Muhammad Tayyab[a], Dong Xiaohua[a,b*], Muhammad Sibtain[c], Ijaz Ahmad[d], Aqeela Zahra[e], Muhammad Imran Azam[a]

[a] College of Hydraulic and Environmental Engineering, China Three Gorges University, Yichang 443002, China; mtayyab@ctgu.edu.cn; imranazam@ctgu.edu.cn; xhdong@ctgu.edu.cn*

[b] Hubei Provincial Collaborative Innovation Center for Water Security, Wuhan, 430070, China; xhdong@ctgu.edu.cn

[c] Laboratory for Operation and Control of Cascaded Hydropower Station, China Three Gorges, University Yichang 443002, China; sibtain92@ctgu.edu.cn

[d] Centre of Excellence in Water Resources Engineering, University of Engineering and Technology, Lahore 54890, Pakistan; ijaz.ahmad@cewre.edu.pk

[e] School of Chemistry, Chemical Engineering and Life Sciences, Wuhan University of Technology, Wuhan, China 430070; zahra@whut.edu.cn

Abstract:

In this study, 03 ensemble and decomposition methods (DMs) i.e., empirical mode decomposition (EMD), ensemble empirical mode decomposition (EEMD) and improved complete ensemble empirical mode decomposition with additive noise (ICEEMDAN) were coupled artificial intelligence and machine learning based method AI-ML, i.e., multilayer perceptron (MLP), support vector regression (SVR) to develop 06 fundamental hybrid models to predict streamflow with one-month lead time. Developed models in this study were categorized into runoff models (RMs) and rainfall-runoff models (RRMs). Results indicated that (i) among standalone models (SMs), support vector regression (SVR) performs better than multilayer perceptron (MLP), (ii) decomposition methods (DMs) have ability to improve the accuracy rate of the standalone models (SMs) and, (iii) rainfall runoff models (RRMs) have shown great accuracy throughout the investigation as compared to runoff models (RMs). To compare model performances flow-hydrographs (FHG) were generated, 05 performance evaluation criteria (PEC) were used to quantify the model precision. Two step verification method i.e., extreme value analysis (EVA) and least value analysis (LVA) approaches were proposed to verify the performances. Among all developed hybrid models (HMs), i.e., EMD- (MLP, SVR), EEMD- (MLP, SVR) and ICEEMDAN- (MLP, SVR), rainfall-runoff ICEEMDAN-(SVR) model was selected as optimal model with MAE (59.56), RMSE (91.82), R (0.97) MAPE (8.75), and NSEC (0.97) for Mangla watershed, Pakistan.

Keywords: streamflow prediction, data driven methods, decomposition methods, input variables, extreme value analysis, least value analysis

1. Introduction

Accurate forecasting of streamflow is vital for proper water control and preventing monetary or fiscal disruptions in the long run. As a result, streamflow forecasting has

acquired a great deal of popularity. In water research sectors, three subsets of AI have been extensively used: (1) Evolutionary computation (2) Fuzzy logic and (3) Machine learning methods and classifiers (Vapnik 2000). Artificial neural networks (ANNs) are used effectively in various fields, including hydrology and water resource management, with a good nonlinear mapping capability (Huo et al. 2012). They affirm that the multilayer perceptron (MLP) is one of the hydrological Models of ANN most commonly studied. It was established through experimentation that the MLP is an approximately fulfills the universal description (Cigizoglu 2009). The verification of this very significant property has been verified by several hydrological studies (Ahmed and Sarma 2006; Mutlu et al. 2008).

Additionally, however, ANNs have issues such as slow learning speed, dimension shifting, and local minima being addictive, and they also tend to overfit to the data (Shamseldin et al. 2002). Over the last few years, there have been notable developments in the application of the Support vector machine (SVM) in the area of hydrology (Shrestha et al. 2005). Support vector regression (SVR) is a powerful tool based on SVM for solving regression, nonlinear classification, and time series prediction problems (Yonaba et al. 2010). When compared to ANNs, SVRs are better at learning when dealing with sparse and incomplete hydrologic data (Toth and Brath 2007; Aggarwal et al. 2012; Bhagwat and Maity 2012; Yu et al. 2012; SM and N 2014; Dehghani et al. 2020).

In recent years, hybrid ANN systems have been commonly included in long and short-term flood prediction. To improve prediction accuracy, empirical mode decomposition (EMD) (Karthikeyan and Nagesh Kumar 2013) and Ensemble Empirical Mode Decomposition (EEMD) (Zhao et al. 2017) are commonly used for decomposition of hydrological runoff time series which are based on ‘decomposition and ensemble’ (Zhao et al. 2017). To address the limitations, in EMD, EEMD is lately applied by (Wu et al. 2009). A new noise-assisted data analysis approach that can resolve the mode mixing disadvantage of the original EMD, which was first introduced by (Huang et al. 1998; Wu and Huang 2004). ICEEMDAN uses the observed signal for IMF sifting to solve EEMD and its variant problems has been studied in the literature (Zhang et al. 2018; Xu and Ren 2019).

However, to the best of the author’s knowledge no study has focused on the streamflow prediction by using data driven and decomposition techniques at selected study area (Mangla watershed) which is considered vital in the planning of water release patterns for sustainable agricultural development and hydropower generation. Therefore, present study aims to develop numerous hybrid models by combining the traditional AI-ML methods such as MLP and SVR models with DMs such as, EMD, EEMD and ICEEMDAN to simulate the monthly streamflow in Mangla watershed, Pakistan. For the selection of input parameters, autocorrelation (ACF) was used and the partial auto-correlation function (PACF) of stream, precipitation, and stream temperature were combined to obtain their correlation output. On the basis of input selection i.e. single input variable (SIV) or multiple input variable (MIV) and decomposition approaches following models were developed in this study: (EMD-MLP, EMD-SVR) with SIV, (EEMD-MLP, EEMD-SVR) with SIV, (ICEEMDAN-MLP, ICEEMDAN-SVR) with SIV, (EMD-MLP, EMD-SVR) with MIV, (EEMD-MLP, EEMD-SVR) with MIV, and (ICEEMDAN-MLP, ICEEMDAN-SVR) with

MIV. Developed models with SIV and MIV can be categorized as runoff models (RMs) and rainfall runoff models (RRMs), respectively. Furthermore, extension of extreme value analysis (EVA) and least value analysis (LVA) approach were adopted to ensure the findings for each model in both calibration and validation phase (CVP). The accuracies of the forecasting models in this study are measured using five-error metrics (MAE, RMSE, MAPE, R, and NSEC). The constructed models were verified for predicting the streamflow with a 1-month time lag at one hydrological station and eleven meteorological stations over Mangla watershed, Pakistan.

The below is a breakdown of the paper's structure: Segment 2 presents overview to the methodology discussed above, which includes AI-ML based models, i.e., MLP and SVR DMs, i.e., EMD, EEMD, Improved CEEMDAN, Segment 3 explains the study area and data in detail, Segment 4 calls and inspects the case study outcomes, and Section 5 draws conclusions of this study.

2. Study Area

2.1 Mangla watershed

The research region is Pakistan's primary source of fresh water as well as power generation and plays an important role in the country's long-term economic growth. The Mangla Watershed is located in the western part of the Himalaya and northeastern part of Pakistan (Khan et al. 2020). Mangla Watershed extends from 73°55’ to 75°35’ east of longitude and 33°25’ to 34°40’ north of latitude as shown in Figure 1. The Jhelum River is one of the major streams that feed into the Indus River, and the Mangla Dam is located on its watershed. The Mangla basin covers about 33,490 km2. The highest point in the Mangla basin is about 5840 m, while the lowest is around 182 m above sea level (a.m.s.l.). The Mangla Dam, built in 1967, has a water capacity of 6.5 million cubic meters. The first dam built to remedy this shortcoming was the Mangla Dam on the Jhelum River, and the other was the Tarbela Dam on the Indus River in Khyber Pakhtunkhana. The inflow into the dam is 1699.1 m^3/s, while the drainage is 566.3 m^3/s. When it comes to hydro-power production, the capability of the Mangla Dam is estimated to be 1310 MW. In the mountainous and cultivated areas of the watershed, population density ranges from 350 to 1000 inhabitants per square kilometer.

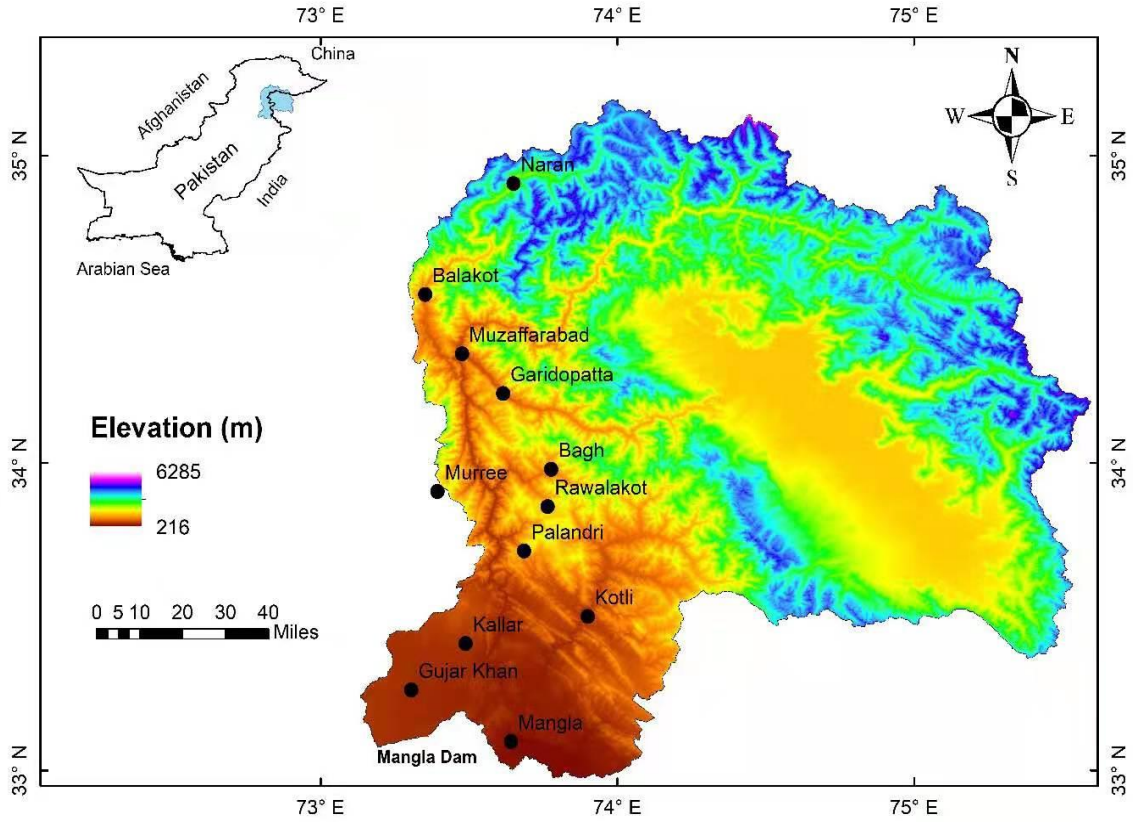


Fig 1 Study area and location of climatic stations at Mangla watershed, Pakistan

2.2 Data

Owing to a lack of data in India, the research region was limited to a catchment that ran along the Pakistan border. Pakistan Meteorological Department (PMD) and Water and Power Development Authority (WAPDA) provided observed data including streamflow (Q), mean monthly temperature (T) and mean monthly precipitation (P). Selected stations and their corresponding statistical details are available in *supplementary Table 1*. Thus, with typical ANN and DM, the entire dataset is employed in training and then the model is used to provide verification on the correctness or predictive power for an applied training set. Total 43 years (1971–2013) of data was divided into two sets first 34 years (1971–2004, 80 %) and 09 years (2004– 2013, 20 %), first set of data were used for the calibration purpose and the second set of data were used for validation.

3. Materials and Methods

3.1 AL-ML methods

3.1.1 Multilayer Perceptron

The MLP is the most common form of ANN used to model hydrological runoff results. MLP is a feedforward neural network type and both integrable and continuous functions can be approximated using this method (Wu et al. 2009). While MLP is a one structure, it is made up of groups of neurons, in layers (Ali et al. 2017). Although MLP employs all input nodes, it is further decoupled into several hidden layers, and in contrast to other architectures, the decision to use these nodes is derived solely on the issue at the top, while in MLP there are many hidden layers of one or two, depending on the problem being examined, but there are no guidelines governing which ones are selected and which ones remain hidden shown in supplementary Figure 1. The value of each node is equivalent to the weighted number of all nodes in the preceding layer.

$$x_{ij=f(W_i X_{i-1}+b_{i-1})} \quad (1)$$

Where f indicates the activation function, and W_i represents weight vector. The $i-1$ layer is X_{i-1}displays neurons rate vector, where x_{ij} demonstrate j neuron rate at i layer and $i-1$ characterized b_{i-1} as bias of the layer.

The architecture and network activation function determine the overall generalization capability of neural network. is a mathematical procedure that, commonly referred to as the backward propagation (BP), which is utilized to modify and, respectively, the MLP and minimize the cost function. The two most commonly used activation functions are linear and nonlinear. The Sigmoid nonlinear activation is a single-function perceptron in the case of linear activation (Wang et al. 1994).

$$f(x) = sigmoid(x) = \frac{1}{1+e^x} \quad (2)$$

The loss function of the real value and the optimal production can be represented as follows (Lange et al. 2016):

$$J(W, b; x, z) = \frac{1}{2}\left\|h_{W,b(x)} - z\right\|^2 \quad (3)$$

Where z denotes the actual value, the output value is given by h, distance norm is shown by $\|.\|$.

3.1.2 Support vector regression

In 1990s Vapnik (Vapnik 1999) developed the support vector machine, which is also regarded as classification and has been applied to regression. SVMs is designed for binary classification, however with the addition of such a loss function, they can also be used to solve regression problems (Hearst et al. 1998). SVM is a nonlinear algorithm that

reduced the problem of over-fitting, minimizes predicted errors in ML shown in supplementary Figure 2.

Based on the preparation, SVM uses a forward quantity to forecast the future (Vapnik and Mukherjee 1999). SVM outperforms ANN networks in the case of correctly determined kernel filters and help vectors (Wu et al. 2008). When using a radial base function:

$$K(x, x_j) = exp\left\{\left\|x_j - x\right\|^2 \delta^2\right\} \quad (4)$$

Support vector regression (SVR) is utilized to perform SVM and many open literatures include detailed descriptions of SVR (Kang et al. 2016). SVR is a viable approach for dealing about forecasting issues, since it is built on systemic risk minimization theory and the Vapnik dimension model.

$$y = f(x) = \left(w.\emptyset(x)\right) + b \quad (5)$$

By lowering the risk function R(C), the coefficients w and b become predictable. Once the performance and period of a training set are fixed, three variables control the performance of an SVR network: ϵ is epsilon, controlling epsilon tube width in the training loss function, σ controlling kernel width of the Gauussian function, and C is the regulatory limitation and objectively regulates SVR peril degree.

3.2 Decomposition Techniques

3.2.1 Empirical Mode Decomposition

By combining EMD with Hilbert spectral analysis, the Hilbert– Huang transform is generated (Huang et al. 1998). HHT functions in a similar manner to wavelet analysis, with the distinction that it is posteriori and requires an analytical theoretical base (Zhou et al. 2017). The core of EMD lies in analyzing feature time scales and empirically recognizing the internal oscillatory modes and eventually decomposing the time series into a sum of various time modes (Huang and Wu 2008). The empirical decomposition mode (EMD) can be used in nonlinear and random time series analysis and can break down the original series into multiple intrinsic mode functions (IMF) and a single signal or residue. Any IMF shall meet two conditions: one is that the number of extremes must be equal to or different by one over the entire time series, while the second is that the upper and lower average values must always be zero. Based on the above conditions, a time series x(t) decomposing by EMD method can be expressed as follows:

$$x(t) = \sum_{i=1}^{m} h_i(t) + r(t) \quad (6)$$

Where IMFs are expressed as $h_i(t)$, number of IMF is denoted as m and the characterizes residual series $r(t)$.

3.2.2 Ensemble empirical mode decomposition

A new technique, ensemble empirical mode decomposition (EEMD), has been developed to help to address the mode issue of mode mix (Wang et al. 2006; Ouyang et al. 2016). The

purpose of the EEMD approach is just to support the EMD method by introducing white noise that drops evenly through the whole time-frequency space, allowing the frequency scales to naturally separate and, as a consequence, the occurrence of mode mixing is reduced (Zhu et al. 2016; Tayyab et al. 2018). The EEMD approach's entire protocol can be summarized as follows:

1. To begin, set the ensemble number (En) and the extra white noise amplitude.
2. Second, you can apply a white noise sequence $n(t)$ to the original time series $x(t)$ to create a new time series $x'(t)$.
3. Third, using the conventional EMD process, decompose the new time series into several IMFs.
4. Fourth, decompose the new time series into many IMFs using the traditional EMD form.

$$x_j'(t) = \sum_{i=1}^{m} h_{i,j}(t) + r_j(t) \qquad (7)$$

5. In last, repeat steps two and three as many times as necessary, each time adding a different white noise sequence. Finally, as the final result, calculate the average ensemble value of all the components of the IMF and of residues.

$$h_i(t) = \frac{1}{En}\sum_{j=1}^{En} h_{i,j}(t) \qquad (8)$$

where $h_{i,j}(t)$ indicates j-th IMF and $r_j(t)$ indicates residual series in i-th test.

3.3.3 Improved complete ensemble empirical mode decomposition with additive noise

ICEEMDAN was proposed to resolve the issues of false modes and frequency aliasing that plague other EMD-based technologies (Niu et al. 2016). ICEEMDAN adds white ringing and understands the frequency continuity between the adjacent dimensions, thus weakening the aliasing effect of frequencies (Ghasempour et al. 2021). The ICEEMDAN calculation technique is presented below:

a. Assign a fixed amount of white noise to the initial signal x, as shown below:

$$x^i = x + \beta_0 E_1(w^i) \qquad (9)$$

Here i represents noise number which is the added, decomposed signal is represented by x^i , the white noiseis donated by ω^i and $E_1(w^{(i)})$ describes the white noise's first EMD portion.

b. The first residue (R_1) can then be obtained as follows:

$$R_1 = \langle M(x^i)\rangle \qquad (10)$$

where $M(.)$ is the local mean of the envelope that meets IMF's sifting verge after decomposition of N signals:

c. The first IMF can be obtained by using EMD as follows:

$$IMF_1 = x - R_1 \quad (11)$$

d. To measure the second residue and mode, follow the steps below:

$$R_2 = \langle M\left(R_1 + \beta_1 E_2\left(w^i\right)\right)\rangle \quad (12)$$

$$IMF_2 = R_1 - R_2$$

e. Using this expression, find the kth residue and mode:

$$R_k = \langle M\left(R_{k-1} + \beta_{k-1} E_k\left(w^i\right)\right)\rangle \quad (13)$$

$$IMF_k = R_{k-1} - R_k$$

f. Repetition of above step leads to the next k stages.

4. Results and Discussions

4.1 Data Classifications and Inputs selections

During this study, a total of 42 years of observation, a minimum of 34 years (1975 to 2008, 80%) of the data was used for model calibration, while a substantial (or residual) data 09 years (2009– 2017, 20 %) was used for model validation. Until normalization was applied, all datasets were set to be between 0 and 1 to remove any variation between them, the dataset's comparative value was to any external information (Tayyab et al. 2017).

$$\text{normalized x} = \left(\frac{q - q_{min}}{q_{max} - q_{min}}\right) \quad (14)$$

Where q is original streamflow data set, q_{min} represents minimum value in original data set and q_{max} represents maximum value in original data set.

To first deal with input selection problem, a suitable input vector for hybrid models is to be calculated by ACF and PACF as shown in **Figure 2**. Based on ACF and the PACF results of Q, T and P inputs were selected for this study and total sixteen models were established, out which eight as RMs which includes Q as SIV where eight were RRMs which includes Q, T and P as MIV. List of all developed models and corresponding inputs for both RMs and RRMs shown in *supplementary Table 2 and Table 3* respectively.

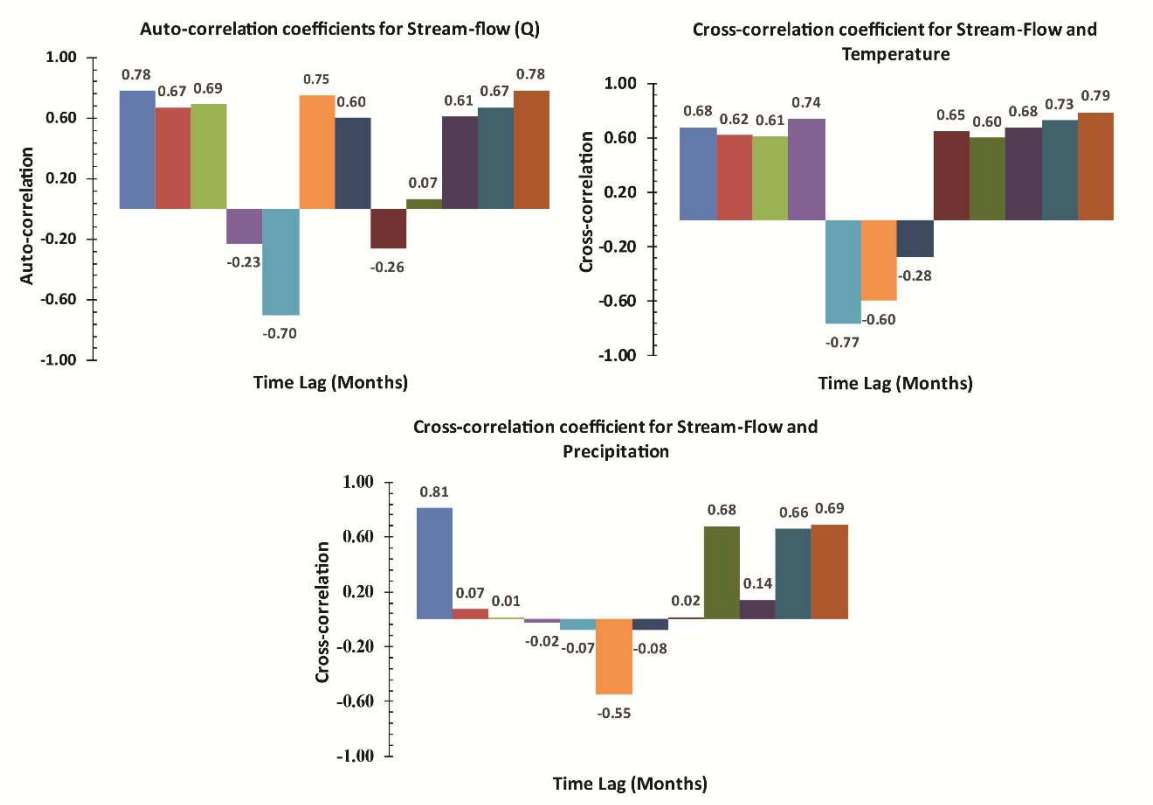


Fig 2 (a) ACF graph for streamflow (b) PACF graph for stream flow and temperature and (c) PACF graph for streamflow and precipitation

4.3 Proposed Hybrid Models

The aim of this research is to create a hybrid data-driven models based on decomposition methods that can best represent the nonlinear behaviour of hydrological series and enhance monthly streamflow predictive performance at Mangla watershed, Pakistan. Three DMs, EMD, EEMD, and ICEEMD and two AI-ML based models MLP and SVR discussed in (section 3) were combined into six basic hybrid models (HMs) in this study including EMD-MLP, EMD-SVR, EEMD-MLP, EEMD-SVR, ICEEMD-MLP and ICEEMD-SVR as shown in **Figure 3**. To obtain further characteristics of varying resolutions for random or non-linear the original series of available data which contains Q, T and P were disintegrated into several IMFs and a residual component (RC) of various frequencies using the EMD-DM. Q time series was disintegrated as (IMF1- IMF6) and a RC, T time series was disintegrated as (IMF1- IMF6) and a RC and P time series was disintegrated as (IMF1- IMF7) and a RC as shown in *supplementary Figure 3*. Similar to EMD to begin, using the EEMD process, the original observed data (Q, T and P) is disintegrated into some IMFs components and a RC. Q observed data was disintegrated as (IMF1- IMF8) and a RC, T observed data was disintegrated as (IMF1- IMF7) and a RC and P observed data were disintegrated as (IMF1- IMF9) and a RC as shown in *supplementary Figure 4*. Second, established AI-ML based models were applied as a predicting technique for all IMFs and RC. Initially, ICEEMDAN decomposes the original observed data including (Q,

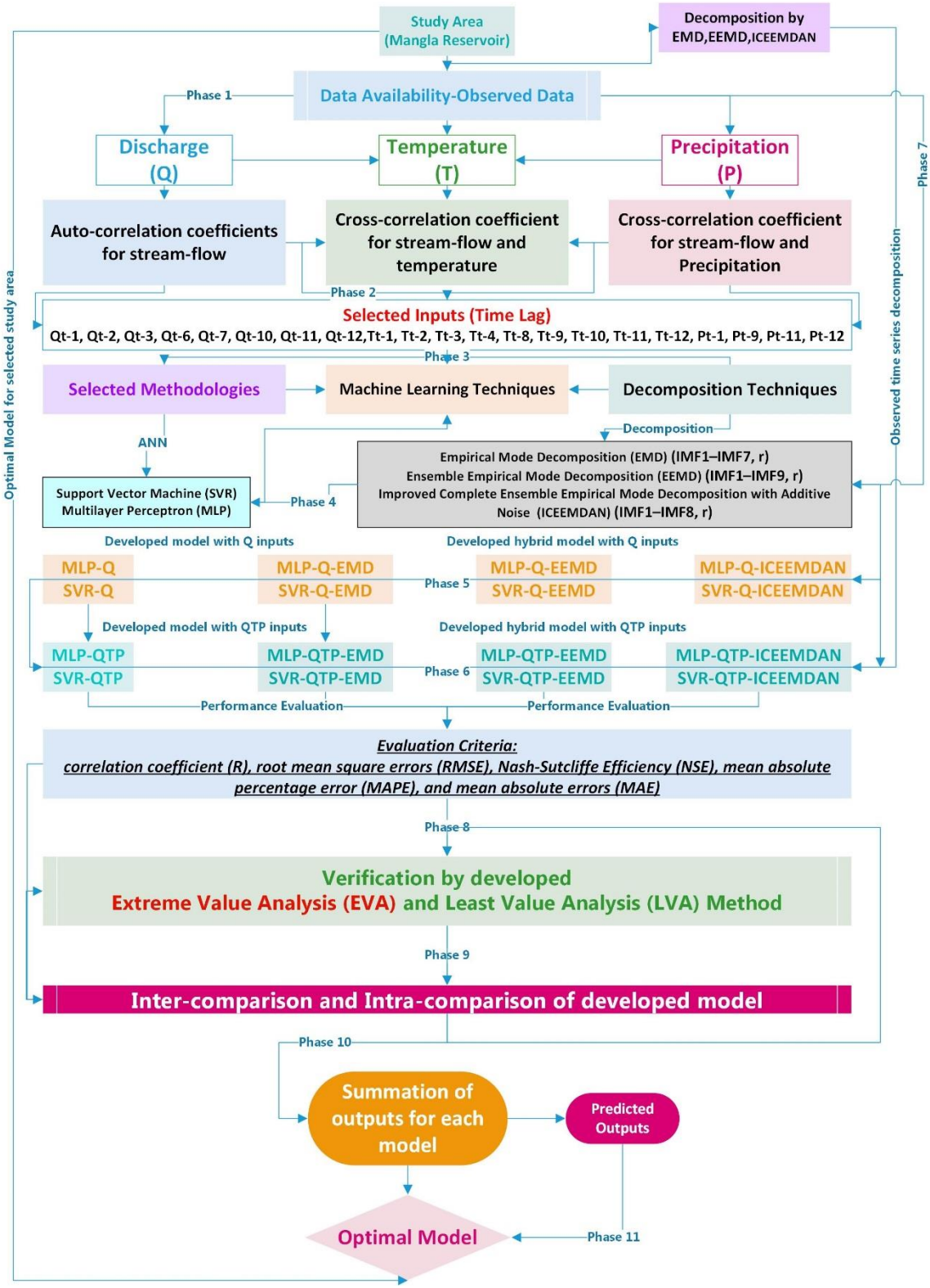


Fig 3 Development of EMD, EEMD and ICEEMDAN based hybrid models

T and P) into multiple IMFs and a RC. Q observed data was disintegrated (IMF1- IMF7) and a RC, T observed data was disintegrated as (IMF1- IMF7) and a RC and P observed

data was disintegrated as (IMF1- IMF8) and a RC *supplementary Figure 5*. By applying DMs total twelve HMs were developed in this study which includes: EMD-MLP-Q, EMD-SVR-Q, EMD-MLP-QTP, EMD-SVR-QTP, EEMD-MLP-Q, EEMD-SVR-Q, EEMD-MLP-QTP, EEMD-SVR-QTP, ICEEMDAN-MLP-Q, ICEEMDAN -SVR-Q, ICEEMDAN-MLP-QTP and ICEEMDAN-SVR-QTP.

4.4 Model performance evaluation Criteria

Five error metrics are used to assess the precisions of the developed predicting models in this analysis. MAE, RMSE, and MAPE with decreasing trends while R and NSEC with increasing trends, indicate good efficacy rate. The following is a list of their definitions shown in **Table 1**.

Table 1. List of performance evaluation indices

No	**Description**	**Abbreviation**	**Represented Equations**
1	mean absolute errors	MAE	$\mathrm{MAE} = \frac{1}{TN}\sum_{i=1}^{TN}\left\|Q_{o,i} - Q_{c,i}\right\|$
2	root mean square errors	RMSE	$\mathrm{RMSE} = \sqrt{\frac{1}{TN}\sum_{i=1}^{n}(Q_{o,i} - Q_{c,i})^2}$
3	mean absolute percentage error	MAPE	$\mathrm{MAPE} = \frac{1}{TN}\sum_{i=1}^{TN}\left\|\frac{Q_{o,i} - Q_{c,i}}{Q_{o,i}}\right\|$
4	correlation coefficient	R	$\mathrm{R} = \frac{\sum_{i=1}^{TN}(Q_{s,i} - \bar{Q}_{s,i})(Q_{o,i} - \bar{Q}_{o,i})}{\sqrt{\sum_{i=1}^{TN}(Q_{c,i} - \bar{Q}_{c,i})^2}\sqrt{\sum_{i=1}^{TN}(Q_{o,i} - \bar{Q}_{o,i})^2}}$
5	Nash-Sutcliffe Efficiency	NSCE	$\mathrm{NSE} = 1 - [\sum_{i=1}^{tn}(Q_{o,i} - Q_{c,i})^2] / \sum_{i=1}^{tn}(Q_{o,i} - \bar{Q}_{o,i})^2]$

where $Q_{o,i}$ and $Q_{c,i}$ characterize the original data and simulated data, $\bar{Q}_{o,i}$ indicates mean observed value and TN indicates sample number in total.

4.5 Results analysis

4.5.1 Standalone Models Performances (Intra-comparison)

To begin, simple AI-ML-based models such as the MLP and SVR were created without using any data pre-processing (decomposition and ensemble) on the original time series to predict the runoff and rainfall-runoff transformation phase of the designated Mangla catchment. FHGs of MLP-Q and MLP-QTP as well as for SVR-Q and SVR-QTP between observed time series and simulated outcomes for both CVP are shown in *supplementary* Figure *6* and **Figure 4** respectively. AI-based models MLP-Q, SVR-Q and MLP-QTP,

SVR-QTP captures the lower obsereved values almost identlically on other hand MLP-QTP, SVR-QTP holds better ground for higher values in both CVP. FHG among observed data verses SVR-Q and SVR-QTP outcomes displays strong evidence in favor of SVR-QTP, that it has ability to overcome the shortcomings of SVR-Q with SIV as shown in *supplementary Figure 6* (a, b) for CP and **Figure 4 (a, b)** VP. Results for MLP-QTP superiority as compared to MLP-Q were justified by the applied PEC i.e., MAE (135.81<195.43), RMSE (217.41<253.06), R (0.84>0.79) MAPE (19.01< 29.55), and NSEC (0.82<0.76) as shown in **Table 2** for both CVP. Similarly, **Table 2** results of PEC meant for SVR-Q and SVR-QTP at VP indicates the dominance of RRM with MIV. SVR-QTP presented improved PEC results i.e., MAE (129.74, 120.43), RMSE (229.94, 200.81), R (0.88, 0.87) MAPE (18.08, 17.21), and NSEC (0.88, 0.85) in CVP as compared to SVR-Q PEC results i.e., MAE (184.14, 147.77), RMSE (284.26, 226.17), R (0.81, 0.81) MAPE (26.94, 19.63), and NSEC (0.81, 0.81). It is concluded here that MLP-QTP> MLP-Q and SVR-QTP>SVR-Q on the basis of FHG and applied PEC results.

4.5.2 Established Hybrid models Performances (Intra-comparison)

To address the shortcomings of standalone models such as MLP and SVR, three decomposition methods (EMD), (EEMD), and ICEEMDAN are used to develop six fundamental hybrid models for monthly streamflow prediction on the Mangla watershed in Pakistan, including MLP-EMD, SVR-EMD, MLP-EEMD, SVR-EEMD, and ICEEMDAN-MLP, ICEEMDAN-SVR.

The FHG results for EMD-MLP-Q and EMD-MLP-QTP shows that model with MIV accurately simulate the observed data than model with SIV. Validation FHG presented the clear picture of EMD-MLP-QTP simulating higher and lower values superiority over the MLP-Q. Likewise, other developed hybrid models were EMD-SVR-Q and EMD-SVR-QTP and EMD-SVR-QTP has improved the accuracy of EMD-SVR-Q. Confirmation of the above stated outcomes can be seen in the FHG *supplementary Figure 6* (c, d) for CP and **Figure 4 (c, d)** VP. It is concluded here that EMD-MLP-QTP> EMD-MLP-Q and EMD-SVR-QTP>EMD-SVR-Q on the basis of FHG and applied PEC results as shown in **Table 2**.

VP PEC results for EMD-MLP based also specifies that (136.63 for RM >108.78 for RRM), RMSE (184.37for RM >143.07 for RRM), R (0.88 for RM < 0.95 for RRM) MAPE (20.45 for RM >17.87 for RRM), and NSEC (0.87 for RM <0.92for RRM) and EEMD-SVR-QTP PEC results for VP can be seen in **Table 2**. It is concluded here that EEMD-MLP-QTP > EEMD-MLP-Q and EEMD-SVR-QTP >EEMD-SVR-Q on the basis of FHG as shown in *supplementary Figure 6* (e, f) for CP and **Figure 4 (e, f)** VP and applied PEC results. In VP of observed data vs both RMs and RRMs the simulation capability of RRMs outperforms RMs with SIV.

To fix the uncertainty factor, the ICEEMDAN approach was used to build further hybrid models, including ICEEMDAN-MLP and ICEEMDAV-SVR. FHG between observed data and simulated outcomes from ICEEMDAN-MLP and ICEEMDAN-SVR based RMs and RRMs verifies the above discussed results that RRMs have shown better performance as compared to RMs as shown in *supplementary Figure 6* (g, h) for CP and **Figure 4 (g, h)** VP. VP PEC results between ICEEMDAN-MLP based RMs, RRMs and ICEEMDAN-

SVR based RMs and RRMs indicates ICEEMDAN-MLP-QTP > ICEEMDAN -MLP-Q and ICEEMDAN-SVR-QTP > ICEEMDAN-SVR-Q as shown in **Table 2.**

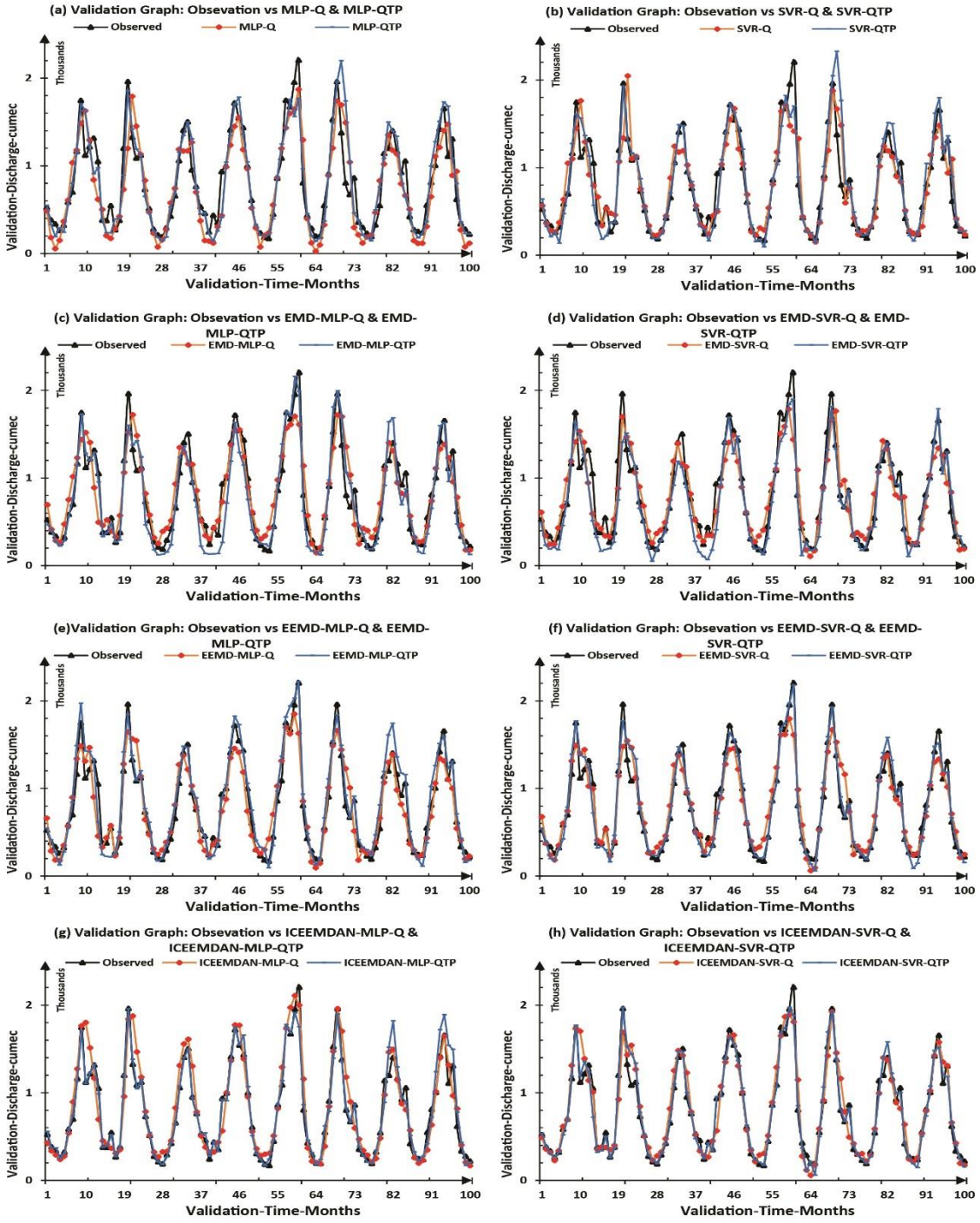


Fig 4 FHGs for observed vs all developed RMs and RRMs models during validation phase

Table 2. PEC results for MLP and SVR based RMs, RRMs with SIV and MIV respectively

S.no	Model	Variable	Input	MAE	RMSE	R^2	MAPE	NSCE	Conclusion
1	MLP	Q	SIV	195.43	253.06	0.79	29.55	0.76	
2	MLP	QTP	MIV	135.81	217.41	0.84	19.01	0.82	
3	SVR	Q	SIV	147.77	226.17	0.81	19.63	0.81	
4	SVR	QTP	MIV	120.43	200.81	0.87	17.21	0.85	
5	EMD-MLP	Q	SIV	163.14	208.68	0.84	25.62	0.84	
6	EMD-MLP	QTP	MIV	135.34	178.05	0.91	20.99	0.88	
7	EMD-SVR	Q	SIV	156.69	201.72	0.85	23.95	0.85	
8	EMD-SVR	QTP	MIV	120.04	176.53	0.91	18.4	0.88	
9	EEMD-MLP	Q	SIV	136.63	184.37	0.88	20.45	0.87	
10	EEMD-MLP	QTP	MIV	108.78	143.07	0.95	17.87	0.92	
11	EEMD-SVR	Q	SIV	127.56	171.42	0.89	20.11	0.89	
12	EEMD-SVR	QTP	MIV	77.03	107.94	0.96	12.5	0.96	
13	ICEEMD-MLP	Q	SIV	125.37	176.99	0.9	18.48	0.88	
14	ICEEMD-MLP	QTP	MIV	63.07	112.43	0.96	8.39	0.95	
15	ICEEMD-SVR	Q	SIV	105.14	152.17	0.92	16.48	0.91	
16	ICEEMD-SVR	QTP	MIV	59.56	91.82	0.97	8.75	0.97	RRMs > RMs Q<QTP

4.6 Extreme Value Analysis and Least Value Analysis

It's crucial to look into the outcomes of the built models in terms of their ability to simulate extreme and least values within the observed time series. In the Mangla watershed, flood season (extreme streamflow values) greater than 830 cumec is considered from May to October, and drought from November to April (least streamflow values) is considered less than 830 cumec. In this study 20% is consider is acceptable relative percentile error, for both calibration and validation phase while applying all developed models.

EVA and LVA for AI-ML based RMs, RRMs with SIV and MIV has been applied and results at VP for all developed models are shown in **Figure 5 and Figure 6** respectively.

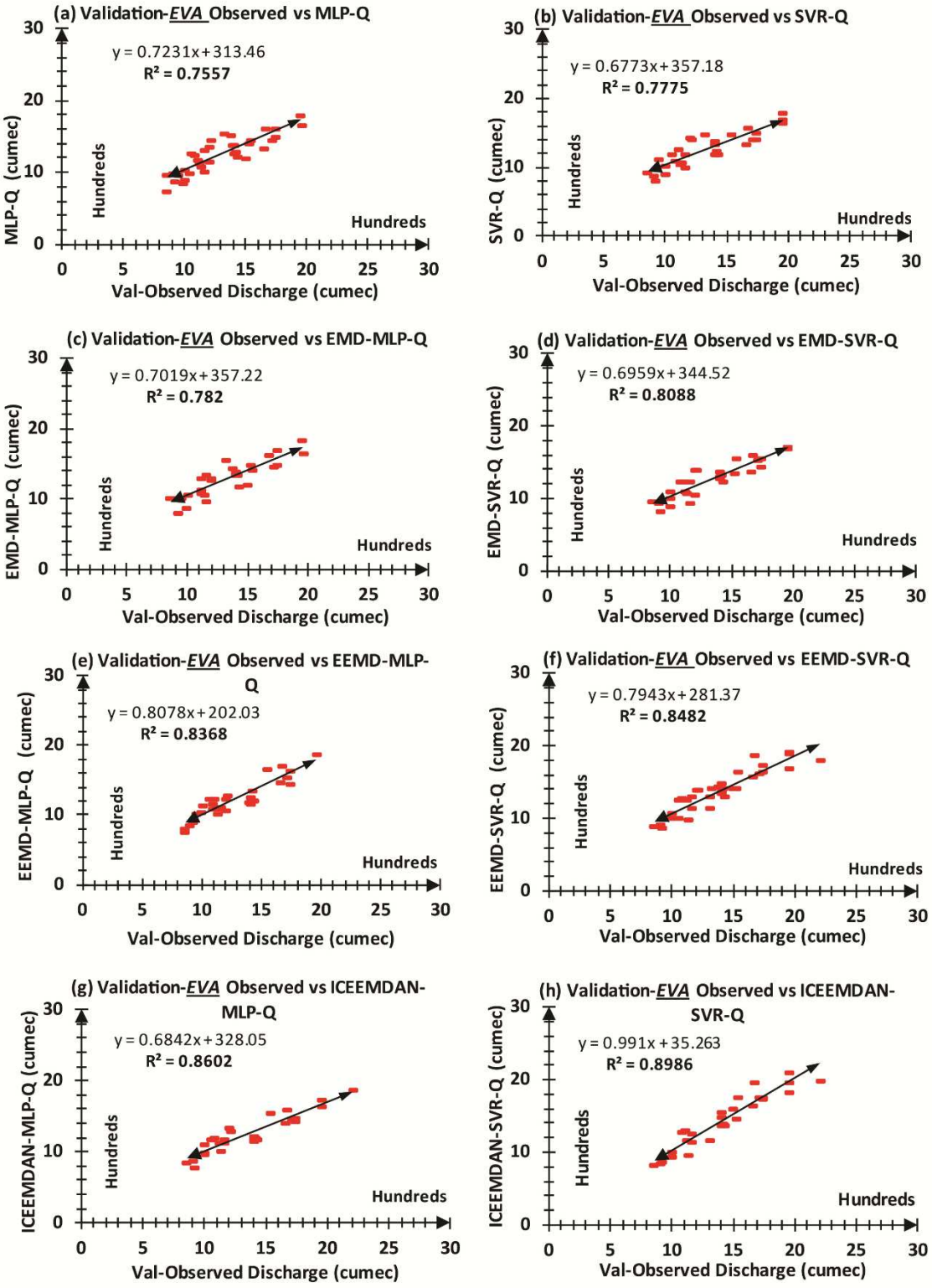
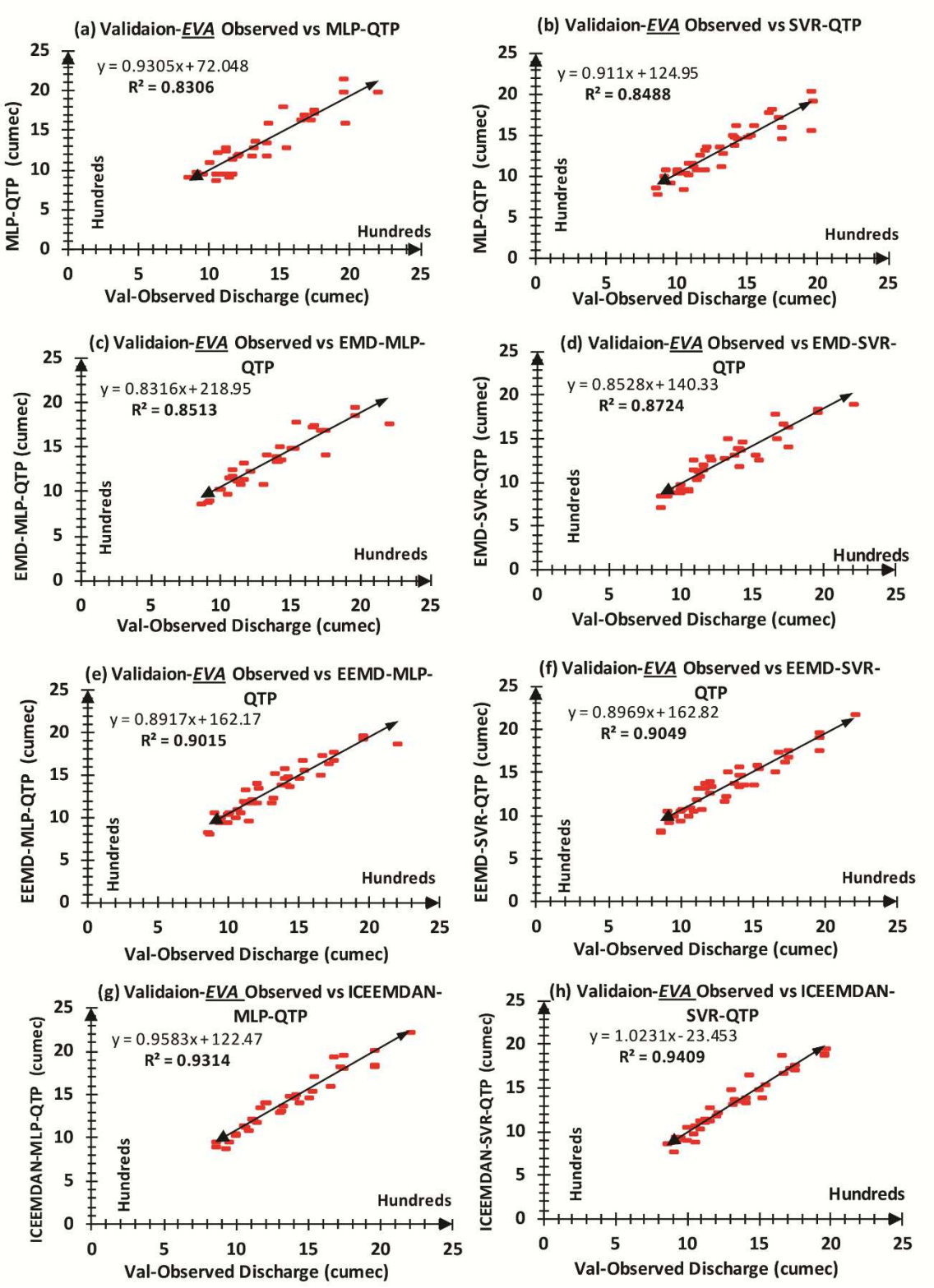


Fig 5 Extreme value analysis (EVA) for all runoff models (RMs) during validation period



371

372 **Fig 6** Extreme value analysis (EVA) for all rainfall-runoff models (RRMs) during
373 validation period

EVA and LVA R^2 indicator for hybrid RMs with SIV and MIV at VP shows the increasing trend of correlation coefficient. Based on EVA and LVA graphs MLP RMs and RRM the trend of superiority will be arranged as MLP-Q < MLP-QTP, EMD-MLP-Q < EMD-MLP-QTP, EEMD-MLP-Q < EEMD-MLP-QTP, ICEEMDAN-MLP-Q < ICEEMDAN-MLP-QTP and for SVR RMs and RRM the trend of superiority will be arranged as SVR-Q < SVR-QTP, EMD- SVR-Q < EMD-SVR-QTP, EEMD-SVR-Q < EEMD-SVR-QTP, ICEEMDAN-SVR-Q < ICEEMDAN-SVR-QTP. EVA and LVA AP% of all developed RMs and RRMs at both CVP are shown in **Table 3.**

Table 3. EVA and LVA accuracy% rate for all RMs and RRMs at CVP

Model	**Calibration**	**Validation**	**Calibration**	**Validation**
Method	**EVA**	**EVA**	**LVA**	**LVA**
Input	**Q(SIV)/ QTP(MIV)**	**Q(SIV)/ QTP(MIV)**	**Q(SIV)/ QTP(MIV)**	**Q(SIV)/ QTP(MIV)**
Accuracy%	**AP%**	**AP%**	**AP%**	**AP%**
MLP	63/79	75/83	91/92	90/93
SVR	66/83	77/84	93/92	91/93
EMD-MLP	68/83	78/85	93/92	92/94
EMD-SVR	68/85	80/87	93/94	92/96
EEMD-MLP	71/84	83/90	94/94	93/96
EEMD-SVR	74/86	84/90	94/95	93/96
ICEEMDAN-MLP	77/87	86/93	95/96	94/97
ICEEMDAN-SVR	79/88	89/94	96/97	96/97

4.7 Inter comparison of all developed models

In this section we will do inter-comparison of all developed models which means AI-based models will be compared with ML-based models. FHGs for VP between AI-ML based models MLP-Q, MLP-QTP, SVR-Q and SVR-QTP showed that SVR-Q is showing better results as compared to MLP-Q but by comparing SVR-Q with MLP-QTP, MLP-QTP is performing better than SVR-Q which shows the effect of input variable selection. Among all these the SVR-QTP has shown great deal of accuracy and ability to improve the results of MLP-Q, MLP-QTP, SVR-Q. SVR-QTP PEC results for VP i.e., MAE (120.43), RMSE (200.81), R (0.87) MAPE (17.21), and NSEC (0.85). During VP AP% in EVA for RMs are (75% and 77%) which is < AP% in EVA for RRMs are (83% and 84%). Based on above results SVR-QTP stands on top among standalone models.

During this inter comparison it is revealed that AI-ML based standalone models i.e. MLP and SVR were improved by all DMs i.e. EMD, EEMD and ICEEMDAN. Among MLP and SVR based models SVR based RMs and RRMs performed better. Results for inter comparison FHGs are shown in **Figure 7**, and EVA and LVA results for ICEEMDAN-

MLP-Q, ICEEMDAN-MLP-QTP, ICEEMDAN-SVR-Q and ICEEMDAN-SVR-QTP during VP can be seen in *supplementary Figure 7*. *Supplementary Table 4* holds PEC results for VP of best ranked models SVR-QTP, EMD-SVR-QTP, EEMD-SVR-QTP, ICEEMDAN-SVR-QTP. After doing intercomparing of each applied and developed models the ranking of best models can be represented as SVR-QTP < EMD-SVR-QTP < EEMD-SVR-QTP < ICEEMDAN-SVR-QTP.

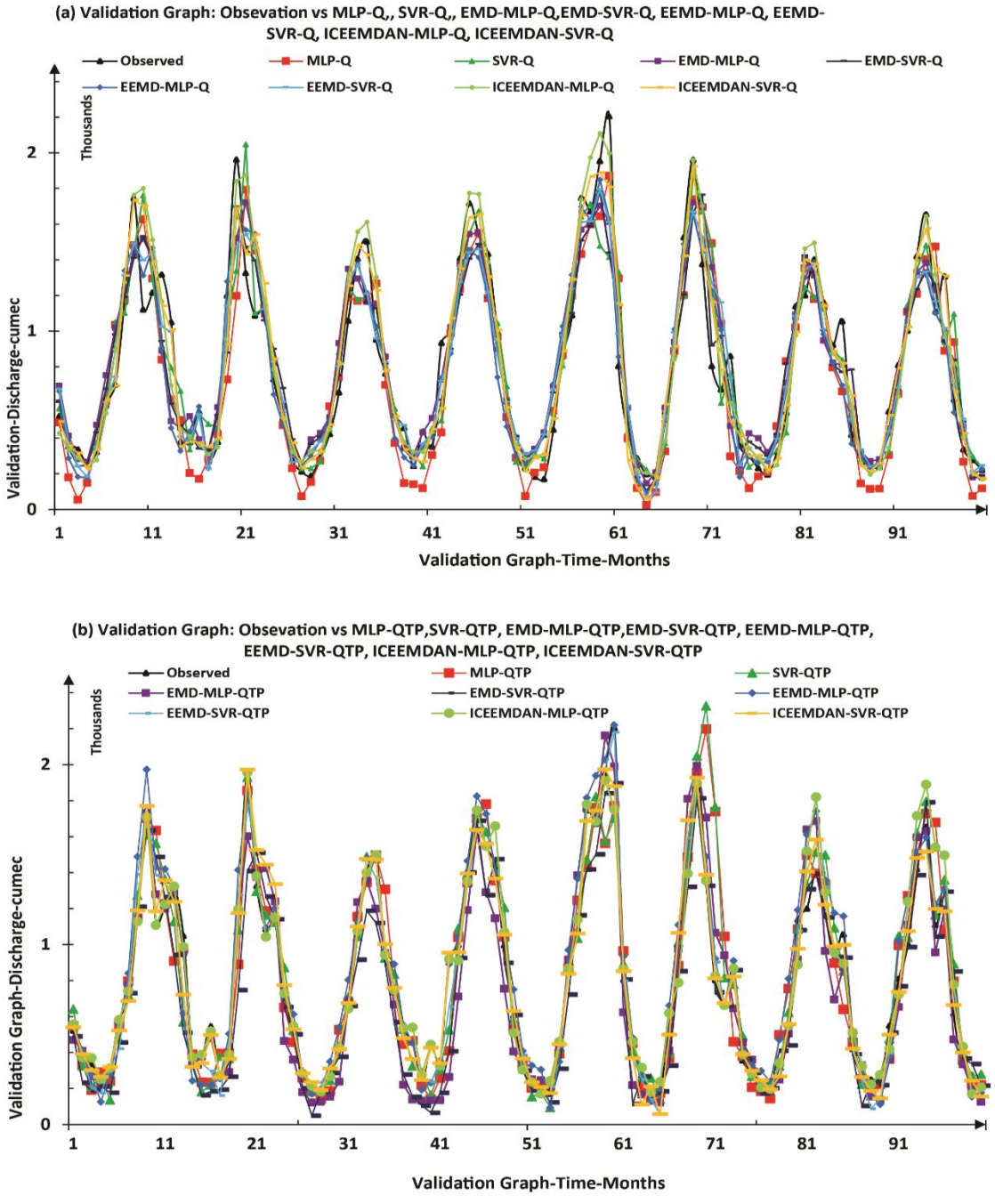


Fig 7 Inter comparison FHGs between observed all RRMs during (a) calibration (CP) phase and (b) validation Phase (VP).

5. Conclusion

Mangla watershed which is one of the biggest fresh water resource for Pakistan selected in this research to predict monthly streamflow with one-month lead time. For this purpose, MLP and SVR models are coupled with EMD, EEMD and ICEEMDAN to develop six basic hybrid models including MLP EMD, SVR-EMD, MLP-EEMD, SVR-EEMD, and ICEEMDAN-MLP, ICEEMDAN-SVR. On the basis of input selection models were further divided into two categories runoff (which includes only monthly streamflow as input variable) and rainfall runoff models (which include monthly streamflow, mean monthly temperature and mean monthly precipitation as inputs variables). Results were analyzed on the basis of hydrograph between observed time series and simulated time series for each model, performance indices (MAE, RMSE, MAPE, R and NSEC) extreme value analysis and least value analysis. Based on this research it is concluded:

1. Standalone developed model (MLP and SVR) shown good results but due to their limitations and shortcoming results a huge gap were found which were filled later by applying decomposition methods (EMD, EEMD and ICEEMDAN).
2. Input selection is the most important part to develop data driven models. For that in this study two set of input variables were used to develop models. Runoff models (RMs) which contains only streamflow as input variable and rainfall-runoff models (RRMs) which contain three streamflow, temperature and precipitation as input.
3. Total sixteen models were developed in this study (MLP-Q, SVR-Q, EMD-MLP-Q, EMD-SVR-Q, EEMD-MLP-Q, EEMD-SVR-Q, ICEEMDAN-MLP-Q, ICEEMDAN-SVR-Q, MLP-QTP, SVR-QTP, EMD-MLP-QTP, EMD-SVR-QTP, EEMD-MLP-QTP, EEMD-SVR-QTP, ICEEMDAN-MLP-QTP).
4. Overall, models with MLP are inferior as compared to SVR based models.
5. Performance indices (PEC) and flow hydrographs (FHGs) have shown that the results have been improved significantly by applying all decomposition methods (EMD, EEMD, ICEEMDAN) but ICEEMDAN has made huge impact on the MLP and SVR standalone results as well as to other hybrid models.
6. The extreme value analysis (EVA) and least value analysis (LVA) technique was introduced to verify the outcome of each developed models for both calibration and validation phase.
7. Based on this research it is revealed that among all proposed models' hybridization of model SVR and ICEEMDAN with QTP has more tendency to predict monthly streamflow at Mangla watershed Pakistan.

Data Availability

All the data were analyzed using MATLAB. The data used to support the findings of this study are available from the corresponding author upon request.

Declaration of competing interest

There is no conflict of Interest for us to publish the finding of this manuscript in the "Meteorology and Atmospheric Physics" Journal.

Acknowledgments

This work is completed with support of the National Natural Science Foundation of China and Funding for post-doctoral work by department of human resources and social security of Hubei Province and by the State Key Program of National Natural Science of China [No. 51239004] and the National Natural Science Foundation of China [No. 51309105].

Contributions

Muhammad Tayyab: Formal analysis, Conceptualization, Methodology, Writing - Original Draft, Visualization. Muhammad Sibtain: Software. Ijaz Ahmed: Writing- Reviewing and Editing. Aqeela Zahra: Writing - Review & Editing. Dong Xiaohua: Writing - Project administration, Supervision. Muhammad Imran Azam: Investigation

Ethics declarations

Ethics Approval

Compliance with Ethical Standards.

Consent to Participate

Not applicable.

Consent for Publication

All Authors declare that they agreed with the content and give explicit consent for publication.

Conflicts of Interest

All Authors declare that they have no conflict of interest.

References

Aggarwal SK, Goel A, Singh VP (2012) Stage and Discharge Forecasting by SVM and ANN Techniques. Water Resour Manag 2012 2613 26:3705–3724. https://doi.org/10.1007/S11269-012-0098-X

Ahmed JA, Sarma AK (2006) Artificial neural network model for synthetic streamflow generation. Water Resour Manag 2006 216 21:1015–1029. https://doi.org/10.1007/S11269-006-9070-Y

Ali Z, Hussain I, Faisal M, et al (2017) Forecasting Drought Using Multilayer Perceptron Artificial Neural Network Model. Adv Meteorol 2017:. https://doi.org/10.1155/2017/5681308

Bhagwat PP, Maity R (2012) Multistep-ahead River Flow Prediction using LS-SVR at Daily Scale. J Water Resour Prot 4:528–539. https://doi.org/10.4236/JWARP.2012.47062

Cigizoglu HK (2009) Estimation, forecasting and extrapolation of river flows by artificial neural networks. https://doi.org/101623/hysj48334945288 48:349–361. https://doi.org/10.1623/HYSJ.48.3.349.45288

Dehghani R, Torabi Poudeh H, Younesi H, Shahinejad B (2020) Daily streamflow prediction using support vector machine-artificial flora (SVM-AF) hybrid model. Acta Geophys 2020 686 68:1763–1778. https://doi.org/10.1007/S11600-020-00472-7

Ghasempour R, Azamathulla HM, Roushangar K (2021) EEMD- and VMD-based hybrid GPR models for river streamflow point and interval predictions. Water Supply. https://doi.org/10.2166/WS.2021.151

Hearst MA, Dumais ; S.T., Osuna E, et al (1998) Support Vector Machines. IEEE Intell Syst 13:18–21. https://doi.org/10.1109/5254.708428

Huang NE, Shen Z, Long SR, et al (1998) The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proc R Soc London Ser A Math Phys Eng Sci 454:903–995. https://doi.org/10.1098/RSPA.1998.0193

Huang NE, Wu Z (2008) A review on Hilbert-Huang transform: Method and its applications to geophysical studies. Rev Geophys 46:. https://doi.org/10.1029/2007RG000228

Huo Z, Feng S, Kang S, et al (2012) Integrated neural networks for monthly river flow estimation in arid inland basin of Northwest China. J Hydrol 420–421:159–170. https://doi.org/10.1016/J.JHYDROL.2011.11.054

Kang P, Kim D, Cho S (2016) Semi-supervised support vector regression based on self-training with label uncertainty: An application to virtual metrology in semiconductor manufacturing. Expert Syst Appl 51:85–106. https://doi.org/10.1016/J.ESWA.2015.12.027

Karthikeyan L, Nagesh Kumar D (2013) Predictability of nonstationary time series using wavelet and EMD based ARMA models. J Hydrol 502:103–119. https://doi.org/10.1016/J.JHYDROL.2013.08.030

Khan MA, Stamm J, Haider S (2020) Simulating the Impact of Climate Change with Different Reservoir Operating Strategies on Sedimentation of the Mangla Reservoir, Northern Pakistan. Water 2020, Vol 12, Page 2736 12:2736. https://doi.org/10.3390/W12102736

Lange S, Gehmlich K, Lun AS, et al (2016) MLP and CARP are linked to chronic PKCα signalling in dilated cardiomyopathy. Nat Commun 2016 71 7:1–11. https://doi.org/10.1038/ncomms12120

Mutlu E, Chaubey I, Hexmoor H, Bajwa SG (2008) Comparison of artificial neural network models for hydrologic predictions at multiple gauging stations in an agricultural watershed. Hydrol Process 22:5097–5106. https://doi.org/10.1002/HYP.7136

Niu M, Wang Y, Sun S, Li Y (2016) A novel hybrid decomposition-and-ensemble model based on CEEMD and GWO for short-term PM2.5 concentration forecasting. Atmos Environ 134:168–180. https://doi.org/10.1016/J.ATMOSENV.2016.03.056

Ouyang Q, Lu W, Xin X, et al (2016) Monthly Rainfall Forecasting Using EEMD-SVR Based on Phase-Space Reconstruction. Water Resour Manag 2016 307 30:2311–2325. https://doi.org/10.1007/S11269-016-1288-8

Shamseldin AY, Nasr AE, O'Connor KM (2002) Comparison of different forms of the Multi-layer Feed-Forward Neural Network method used for river flow forecasting. Hydrol Earth Syst Sci 6:671–684. https://doi.org/10.5194/HESS-6-671-2002

Shrestha RR, Theobald S, Nestmann F (2005) Simulation of flood flow in a river system using artificial neural networks. Hydrol Earth Syst Sci 9:313–321. https://doi.org/10.5194/HESS-9-313-2005

SM H, N M (2014) Developing a fuzzy neural network-based support vector regression (FNN-SVR) for regionalizing nitrate concentration in groundwater. Environ Monit Assess 186:3685–3699. https://doi.org/10.1007/S10661-014-3650-8

Tayyab M, Ahmad I, Sun N, et al (2018) Application of Integrated Artificial Neural Networks Based on Decomposition Methods to Predict Streamflow at Upper Indus Basin, Pakistan. Atmos 2018, Vol 9, Page 494 9:494. https://doi.org/10.3390/ATMOS9120494

Tayyab M, Zhou J, Dong X, et al (2017) Rainfall-runoff modeling at Jinsha River basin by integrated neural network with discrete wavelet transform. Meteorol Atmos Phys. https://doi.org/10.1007/s00703-017-0546-5

Toth E, Brath A (2007) Multistep ahead streamflow forecasting: Role of calibration data in conceptual and neural network modeling. Water Resour Res 43:. https://doi.org/10.1029/2006WR005383

Vapnik VN (2000) Introduction: Four Periods in the Research of the Learning Problem. Nat Stat Learn Theory 1–15. https://doi.org/10.1007/978-1-4757-3264-1_1

Vapnik VN (1999) An overview of statistical learning theory. IEEE Trans Neural Networks 10:988–999. https://doi.org/10.1109/72.788640

Vapnik VN, Mukherjee S (1999) Support vector method for multivariate density estimation. Proc 12th Int Conf Neural Inf Process Syst 76:659–665

Wang W, Gelder PHAJMV, Vrijling JK, Ma J (2006) Forecasting daily streamflow using hybrid ANN models. J Hydrol 324:383–399. https://doi.org/10.1016/J.JHYDROL.2005.09.032

Wang Z, Di Massimo C, Tham MT, Julian Morris A (1994) A procedure for determining the topology of multilayer feedforward neural networks. Neural Networks 7:291–300. https://doi.org/10.1016/0893-6080(94)90023-X

Wu CL, Chau KW, Li YS (2009) Predicting monthly streamflow using data-driven models coupled with data-preprocessing techniques. Water Resour Res 45:. https://doi.org/10.1029/2007WR006737

Wu W, Wang X, Xie D, Liu H (2008) Soil water content forecasting by support vector machine in purple hilly region. IFIP Int Fed Inf Process 258:223–230.

https://doi.org/10.1007/978-0-387-77251-6_25

Wu Z, Huang NE (2004) A study of the characteristics of white noise using the empirical mode decomposition method. Proc R Soc London Ser A Math Phys Eng Sci 460:1597–1611. https://doi.org/10.1098/RSPA.2003.1221

Xu X, Ren W (2019) A Hybrid Model Based on a Two-Layer Decomposition Approach and an Optimized Neural Network for Chaotic Time Series Prediction. Symmetry 2019, Vol 11, Page 610 11:610. https://doi.org/10.3390/SYM11050610

Yonaba H, Anctil F, Fortin V (2010) Comparing Sigmoid Transfer Functions for Neural Network Multistep Ahead Streamflow Forecasting. J Hydrol Eng 15:275–283. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000188

Yu Z, Liu D, Lü H, et al (2012) A multi-layer soil moisture data assimilation using support vector machines and ensemble particle filter. J Hydrol Complete:53–64. https://doi.org/10.1016/J.JHYDROL.2012.08.034

Zhang J, Guo Y, Shen Y, et al (2018) Improved CEEMDAN-wavelet transform de-noising method and its application in well logging noise reduction. JGE 15:775. https://doi.org/10.1088/1742-2140/AAA076

Zhao X, Chen X, Xu Y, et al (2017) An EMD-Based Chaotic Least Squares Support Vector Machine Hybrid Model for Annual Runoff Forecasting. Water 2017, Vol 9, Page 153 9:153. https://doi.org/10.3390/W9030153

Zhou X, Liu W, Jiang Z, Ma F (2017) An Improved Hilbert-Huang Transform Method and Its Application. Gongcheng Kexue Yu Jishu/Advanced Eng Sci 49:196–204. https://doi.org/10.15961/J.JSUESE.201600994

Zhu S, Zhou J, Ye L, Meng C (2016) Streamflow estimation by support vector machine coupled with different methods of time series decomposition in the upper reaches of Yangtze River, China. Environ Earth Sci 2016 756 75:1–12. https://doi.org/10.1007/S12665-016-5337-7

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementarydata.docx