Developing stage–discharge relationships using multivariate empirical mode decomposition-based hybrid modeling

S. Adarsh¹ · Ajin P. John¹ · R. N. Anagha^{1,2} · Abi Abraham^{1,3} · M. P. Afya¹ · K. K. Arathi¹ · Aaliya Azeem¹

Received: 28 March 2018 / Accepted: 7 November 2018 / Published online: 17 November 2018
© The Author(s) 2018

Abstract

This paper proposes an alternative method for modeling stage–discharge relationships by accounting significant information from different process scales employing the multivariate empirical mode decomposition (MEMD). First, the multivariate dataset comprising the discharge of current time step and appropriate lagged inputs of stage and discharge are decomposed using MEMD. Then, genetic programming (GP)-based models are developed for each sub-series to predict discharge of current time step, considering the sub-series of predictors at the corresponding timescale as inputs. Finally, the predicted sub-series are recombined to obtain the discharge of current time step. The method is applied for the prediction of the daily discharge from Pattazhy station of Kallada River in the state of Kerala, India. Statistical evaluation based on different performance criteria revealed the substantial improvement in performance of the proposed MEMD-based hybrid model over the popular nonlinear models like GP and M5 model trees. The application of the method for three more stations Ayilam, Thumpamon and Malakkara falling in different rivers in southern Kerala confirmed the robustness of the approach even in handling zero values and extreme flows in addition to the overall improvement in predictability of daily streamflow. The superiority of the proposed method is attributed to its capability in handling multiple causative inputs and capturing significant information from different timescales common to these input variables.

Keywords Streamflow · Stage · MEMD · Genetic programming · Scale

Introduction

Measurement of flood discharges in rivers is a challenging task for the hydrologists. To overcome this challenge, hydrologists and researchers have been following mathematical modeling of river discharge by developing stage–discharge relationships. For the accurate estimation of discharge, time series analysis is used as a simple and widely accepted technique based on the inherent assumption that the lagged values of input variables fairly represent the factors deciding the fate of hydrologic variable of present time step. In most of the discharge prediction studies, the lagged discharge value alone or present and lagged values of river stage were used as potential inputs. However, the role

of stage and discharge values in providing an understanding into the hydraulic behavior of river systems is different (Ghorbani et al. 2016). Stage values in river hydraulics reflect local effects such as roughness, channel geomorphology and interactions between the main channel and floodplain flows, whereas discharge values provide an overview of general river hydraulics in terms of responsiveness and attenuation (Ghorbani et al. 2016). Hence, it is more appropriate to consider both stage and discharge values in combination as promising inputs in modeling discharge. Due to the complexity of the problem of river flow prediction, the artificial intelligence techniques have gained wide attention in the prediction of discharge (Yazeen et al. 2015). The appropriate combination of river stage and discharge of past time steps is used as inputs for modeling river flow in a time series perspective. The techniques such as artificial neural networks (ANNs), support vector machine (SVM), M5 model trees, fuzzy inference system /adaptive neuro-fuzzy inference system (FIS/ANFIS) are used in popular tools for discharge predictions considering lagged stage and/or discharge values (Deka and Chandramouli 2003; Sudheer and

✉ S. Adarsh
adarsh_lce@yhaoo.co.in

¹ TKM College of Engineering, Kollam 691005, India

² IISc Bangalore, Bangalore, India

³ IIT Guwahati, Guwahati, India

Jain 2003; Sivapragasam and Muttill 2005; Bhattacharya and Solomatine 2005; Kisi and Cobaner 2009; Aggarwal et al. 2012; Al-Abadi 2016, to name a few). Later on, the hybrid soft computing models based on wavelet transforms became popular for discharge prediction (Wei et al. 2012; Tiwari and Chatterjee 2011; Shiri and Kisi 2010; Seghal et al. 2014). Empirical mode decomposition (EMD) proposed by Huang et al. (1998) is a data adaptive decomposition technique, which solves some of the challenging issues related to wavelet transforms such as fixation of most appropriate decomposition levels and selection of appropriate basis function. EMD-integrated data-driven methods are reported to be alternatives for improved predictions of discharge (Napolitano et al. 2011; Karthikeyan and Nagesh Kumar 2013; Huang et al. 2014; Zhu et al. 2016). In most of the EMD-based hybrid models, first the decomposition of discharge time series is performed and it is followed by the predictions of sub-series considering the appropriate lagged values of sub-series as inputs. The final summation of the predicted series at different timescales will provide the discharge at the observation scale (e.g., daily or monthly). As the input variables possess multiscaling behavior, it is more appropriate to consider the potential predictors at each timescale. This may improve the modeling efforts of discharge, as the low and high flow modeling has of particular importance in hydrological applications. It is believed that identifying the most appropriate predictors and excluding the less contributing (less significant) components at different timescales may improve the accuracy of discharge predictions. Investigating in this direction, this paper proposes an alternative method for modeling discharge based on multivariate empirical mode decomposition (MEMD). Recently, few researchers used MEMD-based hybrid modeling for the simulation of soil water, rainfall and evapotranspiration (Hu and Si 2013; Adarsh and Janga Reddy 2017; Adarsh et al. 2017).

The specific objectives of the paper are: (1) to propose a hybrid method for discharge prediction based on MEMD and genetic programming (GP); and (2) to compare the performance of the proposed method with that by the M5 model tree (MT) and GP methods for overall and extreme flow predictions.

Materials and methods

This section presents the theoretical background of different methods used in the study.

Empirical mode decomposition and multivariate extension

Empirical mode decomposition (EMD) is a signal decomposition method proposed by Huang et al. (1998), which

decomposes a time series signal into different oscillatory modes having specific periodicity in purely empirical and data adaptive manner.

The different steps involved in the process are:

1. Identify all extrema (maxima and minima) of the signal $X(t)$
2. Connect these maxima with any interpolation function (for example, cubic spline) to construct an upper envelope ($E_{\max}(t)$); use the same procedure for minima to construct a lower envelope ($E_{\min}(t)$)
3. Compute the mean of the upper and lower envelope, $m(t)$
4. Calculate the difference time series $d(t) = X(t) - m(t)$
5. Let $d(t)$ be the new signal and repeat steps (1)–(4) until $d(t)$ becomes a zero-mean series with no riding waves (i.e., there are no negative local maxima and positive local minima) with smoothened amplitudes. Such an oscillatory signal is called an intrinsic mode function (IMF). To satisfy this step, an appropriate criterion is to be applied in the evolution of an IMF and the modified Cauchy-type stopping criterion is one of the popular ones (Huang and Wu 2008).
6. On satisfying the zero-mean condition, $d(t)$ can be designated as the first intrinsic mode function IMF1
7. Compute the residue $R_1(t)$ by subtracting IMF1 from original signal (i.e., $R_1(t) = X(t) - \text{IMF1}(t)$) is used as new signal. The iterations (so-called sifting) are to be repeated upon this signal to obtain IMF2.
8. The higher oscillatory modes are obtained by treating the residue ($R_k(t)$) as the signal ($X(t)$), iteratively.

The k th residue is defined as

$$R_k(t) = X(t) - \sum_{j=1}^k \text{IMF}_j(t) \quad (1)$$

The process will be continued till the resulting residue is a monotonic function or a function having only one extrema. The final component is called 'Residue' which indicates the long term inherent trend within the time series.

Multivariate EMD method proposed by Rehman and Mandic (2010) is an extension of the traditional EMD, which decomposes multiple time series simultaneously after identifying the common scales inherent in different time series of concern. In this method, multiple envelopes are produced by taking projections of multiple inputs along different directions in an m -dimensional space. Assuming $V(t) = \{v_1(t), v_2(t) \dots v_m(t)\}$ being the m vectors as a function of time t and $X^{\phi_k} = \{x_1^k, x_2^k, \dots, x_m^k\}$ denoting the direction vector along different directions given by angles $\phi_k = \{\phi_1^k, \phi_2^k, \dots, \phi_{m-1}^k\}$ in a direction set X ($k=1, 2, 3, \dots, K$, K is the total number of direction). It can be noted that the

rotational modes appear as the counterparts of the oscillatory modes in EMD. The IMFs of m temporal datasets can be obtained by the following algorithm:

1. Generate a suitable set of direction vectors by sampling on a $(m-1)$ unit hypersphere
2. Calculate the projection $p^{\phi_k}(t)$ of the datasets $V(t)$ along the direction vector X^{ϕ_k} for all k
3. Find temporal instants $t_i^{\phi_k}$ corresponding to the maxima of projection for all k
4. Interpolate $[t_i^{\phi_k}, V(t_i^{\phi_k})]$ to obtain multivariate envelope curves $e^{\phi_k}(t)$ for all k
5. The mean of envelope curves ($M(t)$) is calculated by $M(t) = \frac{1}{K} \sum_{k=1}^K e^{\phi_k}(t)$
6. Extract the 'detail' $D(t)$ using $D(t) = V(t) - M(t)$. If $D(t)$ fulfills the stoppage criterion for a multivariate IMF, apply the above procedure from step (1) onward upon the residue series (i.e., $V(t) - D(t)$). Otherwise, repeat the steps from (2) onward upon the series $D(t)$.

Hammersley sampling sequence can be used for the generation of direction vectors, and the Cauchy-type stoppage criteria reported in the literature (Huang and Wu 2008; Huang et al. 2016) can be used in the implementation of MEMD.

In the present study MEMD is invoked for simultaneous decomposition of multivariate datasets required for developing the regression model. The proposed methodology involves the following steps:

1. Decompose the multivariate dataset comprising discharge at current time step and the lagged inputs using MEMD to obtain different orthogonal modes (sub-series) called intrinsic mode functions (IMFs), each with specific timescale of variability
2. Build genetic programming models to predict each mode as a function of the corresponding mode of different input variables
3. Predict the modes of discharge (at current time step) at different timescales by the refined models
4. Combine the predicted modes to obtain the predicted discharge at observation scale

The above built model is designated as MEMD-GP model in this paper. The MEMD-GP model takes general form

$$OM_{Qi} = \sum_{i=1}^{NP} r_i OM_{PVi} \quad (2)$$

and

$$Q = \sum_{i=1}^M OM_{Qi}$$

where OM denotes an orthogonal mode (an IMF or the final component namely the residue), M is the total number of decomposed modes, NP is the number of predictor variables, PV is the predictor variable, r_i is the regression coefficient, and Q is the discharge. The concept of multivariate dataset

Fig. 1 Concept of multivariate dataset in MEMD, comprising of four inputs (Q_t , S_t , S_{t-1} and Q_{t-1}). Here Q_t and S_t are discharge and stage of current time step, respectively; Q_{t-1} and S_{t-1} are lag-1 discharge and stage values

Fig. 2 Flowchart of MEMD–GP hybrid model. OM orthogonal mode; var1, var2, ... varM—variables

is depicted in Fig. 1, and the flowchart of proposed method is provided in Fig. 2.

Brief theoretical background of GP and M5 model tree is presented in the following subsections.

Genetic programming (GP)

Genetic programming (GP) proposed by Koza (1992) is an automatic programming technique for evolving computer programs to approximately solve different type of optimization or regression problems. The basic principle of GP lies in optimization, which can also be effectively applied for genetic symbolic regression (GSR). In the traditional numerical regression, one predetermines the functional form (like linear, polynomial, exponential) and the task to determine the coefficients is involved. GSR involves finding a mathematical expression in symbolic form relating finite values of set of independent variables (x_i) and a set of dependent variables (y_i). Like other evolutionary algorithms, the solution is started with a random set of population. The solution set in each iteration is collectively called as generation. The subsequent generations are formed through genetic operations such as ‘mutation’ and ‘crossover.’ Each possible solution set can be visualized as a ‘parse tree.’ A parse tree is built up from a terminal set (the independent variables and constants in the problem) and a function set (the basic operators used to form the function like +, −, /, *, log). As the new solution evolves through basic genetic operations, this solution set replaces

the previous set and are supposed to be perform better. The solution set in a population associated with the ‘best fit’ individuals will be reproduced more often than the less fit solution sets. The ‘fitness’ is a measure of how closely the population will solve the problem. The above-stated fitness-based betterment of the solution can be termed as Darwin’s principle of ‘survival of the fittest.’

The basic procedure of GP can be summarized as follows:

1. Generate the set of initial population
2. Evaluate the parse tree and assign the fitness
3. Form temporary population by selecting candidates according to their fitness. This temporary population is called as ‘mating pool.’ Candidates with higher fitness are given as greater probability to mate, to produce children or offspring.
4. Choose pairs of parse trees from the temporary mating pool randomly for mating and apply the genetic operator called crossover. Crossover is the exchange of genetic material (such as fitness between two selected candidates).
5. Select a crossover site where the material will be exchanged randomly thereby resulting in the creation of an offspring
6. Apply another genetic operator called as ‘mutation’ which randomly changes the genetic information of a candidate.
7. Copy the resulting chromosome into the new population
8. Evaluate the performance of new population

9. Steps 3 to 8 are to be repeated until a pre-specified termination criterion is reached.

M5 model tree

Model tree is a machine-learning technique proposed by Quinlan (1992) which combines the features of classification and regression. The algorithm splits the parameter space into sub-areas and builds a linear regression model in each of them. So, they are analogous to piecewise linear functions. The splitting in the MT follows the idea of a decision tree, but instead of class labels, it has linear regression functions at the leaves. The splitting criterion is used to determine which attribute is the best to split the portion T (set of examples that reach the node) of the training data that reaches a particular node. It is based on treating the standard deviation of the class values in T as measure of the error at that node, and calculating the expected error as a result of testing each attribute at that node. The formula to compute the standard deviation reduction (SDR) is

$$\text{SDR} = \text{sd}(T) = \sum_i \frac{|T_i|}{|T|} \text{sd}(T_i)$$

where T = set of examples that reach the node; T_i = subset of examples that have the i th outcome of the potential set; and sd = standard deviation. The variant of model tree which uses SDR as the splitting criteria is called as M5 model tree. When a model tree is used to predict the value for a test instance, the tree is followed down to a leaf in the normal way, using the instances attribute values to make the routing decisions at each node. The leaf will contain a linear model based on some of the attribute values, and is evaluated for the test instance to yield a raw predicted value. After the initial tree has grown, there are several steps that have to be taken, such as calculation of error estimates, generation of linear models, simplification of linear models and pruning and smoothing. Once the linear model is in place for each interior node, the tree is pruned back from the leaves so

long as the expected estimated error decreases. Smoothing is a process used to compensate for the sharp discontinuities that will inevitably occur between adjacent linear models at the leaves of the pruned trees. More details on the smoothing and pruning operations can be algorithm can be found elsewhere (Witten et al. 2005; Jothiprakash and Kote 2011).

Study area and database

The Central Water Commission (CWC) defined 22 major river basins across India. The west-flowing rivers from Tadri to Kanyakumari are located in the southwest coast of India, which is bounded on eastern side by the Western Ghats, western side by the Arabian Sea and southern side by the Indian Ocean. Ayilam gauging station is the southernmost one operated by CWC located at Vamanapuram River having a length of 88 km and drains an area of 687 km². Pattazhy station is located in Kallada River with the length of 121 km and drains an area of 1700 km². Achankovil River has a drainage area of 810 km², and the CWC hydrological station Thumpamon is located in this river. Malakkara station is located in Pamba River, the third longest one in the state of Kerala (in which the major part of the basin falls) with 176 km length and drains 22235 km². The data used in this study include daily discharge and river stage measurements of the CWC India. Six years of data from May 31, 2007, till May 31, 2013, of each station has been collected and used for the study. First 4 years of data (1461 data points) from May 31, 2007, to May 30, 2011, has been used for model calibration, while that from May 31, 2011, to May 31, 2013 (732 data points), has been used for validation. The basic statistical properties of datasets are given in Table 1. The daily stage and discharge observations are downloaded from the official website of Water Resource Information System (www.india-wris.nrsc.gov). The location map of the CWC stations is given in Fig. 3.

Table 1 Statistical properties of datasets

Station	Statistical property	Calibration	Validation	Station	Statistical property	Calibration	Validation
Pattazhy	Mean	40.559	31.486	Malakkara	Mean	115.846	84.190
	SD	47.796	29.553		SD	151.693	126.158
	Max.	502.61	174.905		Max.	1127.00	920.251
	Min.	0.449	0.524		Min.	0.000	0.000
Ayilam	Mean	22.718	8.616	Thumpamon	Mean	32.876	21.883
	SD	39.946	20.535		SD	54.242	33.579
	Max.	521.96	252.58		Max.	578.628	225.243
	Min.	0.000	0.000		Min.	0.000	0.000

Min. minimum, *Max* maximum, *SD* standard deviation

Fig. 3 Map of the four hydro-logic observation stations.

Source: www.india-wris.nrsc.gov

Results and discussion

Lagged time series values of discharge and stage are the integral effect of different factors influencing with river flow, and on accounting the fact, the lagged values of discharge and stage are considered as inputs to model the discharge at the current time step. Considering different combinations of inputs (considering a maximum of 2-day

lag), 10 models are prepared in this study, which are listed in Table 2. A typical model with current time step stage, previous two lagged values of stage and two lagged values of discharge can be represented in the following form: $Q_t = f(S_t, Q_{t-1}, Q_{t-2}, S_{t-1}, S_{t-2})$

All the inputs and output together will constitute the multivariate dataset for preparing the MEMD–GP model. The decomposition process should be employed in such a way

Table 2 Description of models and omitted modes in developed GP models at different timescales

Model no.	Input combination	Input variable (IMF of)	Mode number
1	$S_t Q_{t-1}$	S_t	7
2	$Q_{t-1} Q_{t-2}$	—	—
3	$S_t S_{t-1}$	S_{t-1}	6
4	$Q_{t-1} S_{t-1}$	Q_{t-1}	2
		S_{t-1}	4
5	$Q_{t-1} S_t S_{t-1}$	—	—
6	$Q_{t-1} Q_{t-2} S_t$	S_t	6
7	$S_t S_{t-1} S_{t-2}$	S_{t-2}	1, 2, 3, 11
		S_{t-1}	11
8	$Q_{t-1} S_t S_{t-1} S_{t-2}$	S_{t-2}	1, 2, 5, 8, 11
		S_{t-1}	11
9	$Q_{t-1} Q_{t-2} S_t S_{t-1}$	—	—
10	$Q_{t-1} Q_{t-2} S_t S_{t-1} S_{t-2}$	S_{t-2}	9, 11

The bold figures refer the best performing model

that it should not lead to over-iteration and over-decomposition, for which certain threshold parameters are used to control the magnitude of fluctuations of the series and the number of sifting iterations (Rilling et al. 2003). The lower and upper threshold parameters $\theta_{\min}$ and $\theta_{\max}$ evaluate the relative magnitude of amplitude of the mean in comparison with the amplitude of the corresponding modes. The upper and lower bounds of control parameters of MEMD are fixed as 0.075 and 0.75 and the fraction which controls sifting iterations is taken as 0.5, following the past studies (Rilling et al. 2003; Hu and Si 2013). The decomposition resulted in 10–11 orthogonal modes for the dataset of Pattazhy station. This decomposition can be considered to be accurate (i.e., not subjected to over-decomposition), as the maximum expected number of modes is $\log_2(N)$, N is the data length. Then the GP models are prepared by considering the orthogonal mode of discharge (at a particular timescale) at current time step as dependent variable and orthogonal modes of lagged value of predictors at the same scale as independent variables. For different models, the less significant variables (whose regression coefficients are zero) are summarized in Table 2. In Models 2, 5 and 9, all variables are included in the modeling at all timescales.

Finally, the summation of all predicted modes is used to rebuild the discharge at current time step. The GP models prepared at different timescales are used to predict the discharge of validation period (2011–2013). The discharge predictions for calibration data and validation data were also made using GP (a purely nonlinear method) and M5 model tree (a semi-nonlinear method) methods for comparison of the results. The performance of the proposed method is compared using the most popular performance evaluation

measures of correlation coefficient (R) and root mean square error (RMSE) and formulae of computation of these statistics are available elsewhere (Napolitano et al. 2011). The time series plots and scatter plots of predictions by all the three methods (for the best performing model, Model 10) are presented in Fig. 4.

Performance statistics of prediction of daily streamflow of Pattazhy station by the ten models is provided in Table 3. From Table 3, it is noted that the models which consider both stage and discharge and with more than two variables show better predictive capabilities (Models 5, 6, 8, 9, 10, etc.). Here in all such cases, the MEMD–GP model results in better prediction (with less RMSE and high R value for validation datasets) when compared with MT and GP models. From Table 3, it can also be noted that on considering the stage alone as input, performance of predictions is found to be poor in both cases (Model 3 and Model 7). It can be noted that overall predictions made by GP or MT found to be better than the hybrid MEMD–GP model in such cases. It is quite anticipated as the scale separation is not helpful when the stage and its lagged value alone are considered as input. It is noticed that the performance measures of Model 10 (the combination with current time stage, stages of two time lags, discharge with two time lags) are performing best than any other combination model considered in this study. Here there is an improvement of 48% in the RMSE statistics of predictions by MEMD–GP model when compared with that by GP model for the validation set (RMSE of 3.26 and 6.18, respectively, for MEMD–GP and the GP model).

Similarly, the three different methods (MEMD–GP, GP and MT) were invoked again for the prediction of daily discharge of Ayilam, Malakkara and Thumpamon stations considering the 10 different input combinations. The sensitivity analysis of different input combinations revealed that the model with current and lag-1 time stage and two previous discharge values $Q_t = f(S_t, Q_{t-1}, Q_{t-2}, S_{t-1})$ is found to be performing best among the ten models investigated in this study. The results of the performance measures of daily discharge predictions of the three stations by different methods are summarized in Table 4. Further, time series plots and scatter plots of predictions are provided in Figs. 5, 6 and 7, respectively, for Ayilam, Malakkara and Thumpamon stations. Figure 8 presents the developed stage discharge curves of different stations. It is worth mentioning that in the hybrid MEMD–GP models as scale separation and different models for different timescales (the predictions of which are recombined later) a unique single expression connecting discharge of current time step with the lagged stage or discharge values are not expressible.

On handling the daily datasets of perennial rivers, it is quite logical to expect consecutive zero discharge values and it will have significant impact on the relationships. In the selected stations of Pattazhy, no zero discharges are

Fig. 4 Daily discharge predictions for Pattazhy station by different methods. Upper panels show time series plots; lower panels show scatter plots

Table 3 Results of performance evaluation of streamflow prediction models with different input model combinations

Model no.	MEMD-GP				GP				MT			
	C		V		C		V		C		V	
	R	RMSE	R	RMSE	R	RMSE	R	RMSE	R	RMSE	R	RMSE
1	0.98	9.03	0.97	6.62	0.96	13.3	0.90	12.8	0.92	19.1	0.94	11.3
2	0.97	10.2	0.97	6.63	0.86	23.8	0.86	14.8	0.88	22.3	0.87	14.8
3	0.93	17.2	0.96	18.5	0.88	21.9	0.95	13.7	0.84	26.2	0.97	11.9
4	0.93	17.4	0.92	11.4	0.88	22.8	0.86	15.2	0.86	24.0	0.86	15.1
5	0.99	6.57	0.98	5.27	0.99	6.39	0.97	6.18	0.98	9.60	0.97	6.98
6	0.99	6.78	0.98	4.85	0.96	12.7	0.89	12.9	0.92	18.8	0.94	11.2
7	0.95	14.6	0.86	23.7	0.89	21.2	0.95	13.8	0.84	26.1	0.97	11.9
8	0.99	6.57	0.98	5.20	0.99	6.43	0.98	6.18	0.98	9.56	0.97	6.98
9	0.99	5.69	0.98	4.32	0.99	6.23	0.98	6.17	0.98	9.44	0.97	6.94
10	0.99	4.27	0.99	3.26	0.99	5.78	0.98	6.51	0.98	9.34	0.97	6.92

The bold figures refer the performance statistics of best model

C calibration, V validation

present in the datasets owing to the supplemented flows to the river at the upstream of the station. In the dataset of Ayilam station, 40% zero data values are present in the calibration set, while 60.52% zero data values are present in the validation data. In the discharge data of Thumpamon

station, 30.4% are zero values are present in calibration and 37% are zero values. In the discharge dataset of Malakkara station, 24% are zero values in calibration set and 44% are zeros in the validation set. From Tables 3 and 4, it is evident that the MEMD-GP model of Pattazhy

Table 4 Performance evaluation of best models for daily streamflow predictions of the three stations by three different methods

Station	MEMD-GP				GP				MT			
	Calibration		Validation		Calibration		Validation		Calibration		Validation	
	RMSE	R	RMSE	R	RMSE	R	RMSE	R	RMSE	R	RMSE	R
Ayilam	4.53	0.994	4.55	0.979	6.29	0.988	6.12	0.96	3.55	0.99	10.0	0.94
Malakkara	13.59	0.996	13.7	0.994	22.21	0.989	20.29	0.99	5.68	0.99	54.42	0.95
Thumpamon	8.97	0.988	4.15	0.995	9.61	0.979	6.11	0.98	2.05	0.99	5.23	0.99

Fig. 5 Daily discharge predictions for Ayilam station by different methods. Upper panels show time series plots; lower panels show scatter plots

station performs better (RMSE of 3.26 for validation data) when compared with that of other stations (RMSE values of 4.55, 13.7 and 4.15 for validation data of Ayilam, Malakkara and Thumpamon stations). However, it is consistent that in even these three stations, MEMD-GP hybrid method performed well when compared with traditional GP or model tree approaches. For example, at Ayilam station (with 60% zero values for validation data), the MEMD-GP model results displayed an RMSE of 4.55 against the RMSE of GP and model tree-based results (6.12 and 10, respectively). Identical observations can be made for other stations also, and hence, it can be concluded that the proposed MEMD-GP model gets influenced least by the presence of zero discharges—a typical

characteristics of daily discharge datasets of perennial rivers, when compared with the GP or model tree approaches.

Similar to the presence of the zero discharges, extreme high flows also may be present in daily discharge datasets of tropical river basins in India. For obtaining a better understanding on the quality of the predictions, the datasets are split into high, low and medium and the predictions were examined for two stations of distinctly different characters—Pattazhy and Ayilam. For these stations, the discharge values exceeding (mean + standard deviation) are categorized as high flow, those values less than ‘mean’ are classified as low flows, and the rest are considered as medium flow. The correlation coefficients and RMSE of the segregated datasets are computed and presented in Table 5. The results in

Fig. 6 Daily discharge predictions for Malakkara station by different methods. Upper panels show time series plots; lower panels show scatter plots

Table 5 also show that the extreme flows are best captured best by the MEMD–GP method when compared with GP method and more inconsistencies for the M5 model tree-based approach.

MEMD method is capable to provide an equal number of modes for different time series of a multivariate dataset, which is highly advantageous and convenient in building the separate regression models at different timescales in regression modeling. Thus, it facilitates the inclusion of multiple causal inputs in regression modeling, which makes the MEMD a superior alternative over the classical EMD. Decomposition into oscillatory modes makes it possible to analyze the contributions in different timescales separately, and this may be the reason for better performance of the proposed method on considering datasets of distinctly different characteristics. The results of this study show that the model is able to capture the high and low flow states satisfactorily in rivers of two distinctly different data characteristics (Pattazhy and Ayilam, in the former it is with no zero discharge values, while in the latter it is with many zero values). It is to be noted that the methodology proposed in this paper being general one, it can be applied for developing stage–discharge relationships in any river irrespective

of their origin and water flow considering any set of input combinations. The time series methods of discharge prediction are framed on the basic assumption that the governing physical processes are embedded in the time series itself and more input variables can be included in the modeling, if they are found to be influential. The proposed multivariate framework is flexible in including more causal variables (such as rainfall) as input as and when required for modeling daily discharge. Also, the proposed methodology can be adopted for real-time discharge predictions by using adaptive variants of the proposed hybrid model (Tan et al. 2018). In short, it is evident that the hybrid modeling approach involving MEMD and GP is a robust alternative for modeling stage–discharge relationships of rivers of distinctly different characteristics by accounting multiple causal variables.

Conclusions

This study proposed a new hybrid modeling framework for developing stage–discharge relationships by utilizing the scaling property of causal variables. The proposed method employed multivariate empirical mode decomposition

Fig. 7 Daily discharge predictions for Thumpamon station by different methods. Upper panels show time series plots; lower panels show scatter plots

Fig. 8 Stage-discharge relationships developed for the four stations **a** Pattazhy; **b** Ayilam; **c** Malakkara; **d** Thumpamon using MEMD-GP hybrid model

Table 5 Performance evaluation of high, low and medium states of daily discharge predictions using different methods

Station	Flow state	MEMD–GP				GP				MT			
		Training		Validation		Training		Validation		Training		Validation	
		RMSE	R	RMSE	R	RMSE	R	RMSE	R	RMSE	R	RMSE	R
Pattazhy	HF	12.69	0.99	7.09	0.97	21.18	0.96	9.08	0.94	16.09	0.98	30.93	0.36
	MF	4.81	0.59	1.57	0.95	5.76	0.52	1.67	0.20	1.59	0.85	5.54	0.00
	LF	6.40	0.95	6.28	0.89	8.56	0.90	8.13	0.82	5.41	0.96	14.14	0.68
Ayilam	HF	9.478	0.99	5.97	0.99	15.52	0.98	9.24	0.98	3.147	0.99	20.775	0.954
	MF	13.37	0.22	2.67	0.43	4.26	0.26	4.16	0.37	0.551	0.93	25.213	−0.221
	LF	3.522	0.98	3.56	0.89	4.25	0.97	5.74	0.79	1.059	0.99	23.562	0.192

HF high flow, LF low flow, MF medium flow

(MEMD) for the decomposition of multivariate dataset, and then separate models are developed to predict the components of discharge at different process scales using the genetic programming (GP) method. The proposed method is demonstrated by applying it for developing stage–discharge relationships at four different locations in Southern Kerala, India. Major conclusions drawn from this study are:

- MEMD–GP coupled model displayed superior performance when compared with conventional GP and M5 model tree in establishing stage–discharge relationships owing to its ability to account the contributions of different input variables from different process scales
- As the ratio of number of stage parameters to discharge parameters in the inputs increases, the models show poor performance, which is particularly evident in datasets with more zero values
- Zero discharge values affect the performance of stage–discharge models, and it has least influence in case of MEMD–GP model when compared with traditional M5 model tree or GP based models. This is evident from the results of Ayilam station with 60% zero discharge values.
- The method displayed better performance in capturing extreme low flows and high flows when applied to the datasets of Pattazhy and Ayilam stations, where the discharge characteristics of the corresponding rivers are distinctly different
- The framework proposed in this study is general; it can be applied for developing stage–discharge relationships of any river irrespective of their origin, geography and discharge characteristics, considering any set of input combinations.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

- Adarsh S, Janga Reddy M (2017) Multiscale characterization and prediction of monsoon rainfall in India using Hilbert–Huang transform and time dependent intrinsic correlation analysis. *Meteorol Atmos Phys.* <https://doi.org/10.1007/s00703-017-0545-6>
- Adarsh S, Sulaiman Sanah, Murshida KK, Nooramol P (2017) Scale-dependent prediction of reference evapotranspiration based on multivariate empirical mode decomposition. *Ain Shams Eng J.* <https://doi.org/10.1016/j.asej.2016.10.104> (in press)
- Aggarwal SK, Goel A, Singh VP (2012) Stage and discharge forecasting by SVM and ANN techniques. *Water Resour Manag* 26(13):3705–3724
- Al-Abadi AM (2016) Modeling of stage–discharge relationship for Gharraf River, southern Iraq using backpropagation artificial neural networks, M5 decision trees, and Takagi–Sugeno inference system technique: a comparative study. *Appl Water Sci* 6(4):407–420
- Bhattacharya B, Solomatine DP (2005) Neural networks and M5 model trees in modeling water level–discharge relationship. *Neurocomputing* 63(2005):381–396
- Deka PC, Chandramouli V (2003) A fuzzy neural network model for deriving the river stage discharge relationship. *Hydrol Sci J* 48(2003):197–209
- Ghorbani MA, Khatibi R, Goel A, Fazeli Fard MH, Azani A (2016) Modeling river discharge time series using support vector machine and artificial neural networks. *Environ Earth Sci.* <https://doi.org/10.1007/s12665-016-5435-6>
- Hu W, Si BC (2013) Soil water prediction based on its scale-specific control using multivariate empirical mode decomposition. *Geoderma* 193–194:180–188
- Huang NE, Shen Z, Long SR, Wu MLC, Shih HH, Zheng QN (1998) The empirical mode decomposition method and Hilbert spectrum for non-stationary time series analysis. *Proc R Soc* 454(1998):903–995
- Huang NE, Wu Z (2008) A review on Hilbert Huang Transform: Method and its applications to geophysical studies. *Rev Geophys.* <https://doi.org/10.1029/2007RG000228>
- Huang S, Cheng J, Huang Q, Chen Y (2014) Monthly streamflow prediction using modified EMD-based support vector machine. *J Hydrol* 511:764–775
- Huang G, Su Y, Kareem A, Liao H (2016) Time-frequency analysis of non-stationary process based on multivariate empirical mode decomposition. *J Eng Mech* 142(2016):10.1061/(ASCE)EM.1943-7889.0000975

- Jothiprakash V, Kote AS (2011) Effect of pruning and smoothing while using M5 model tree technique for reservoir inflow prediction. *J Hydrol Eng* 16(7):563–574
- Karthikeyan L, Nagesh Kumar D (2013) Predictability of non-stationary time series using wavelet and EMD based ARMA models. *J Hydrol* 502(2013):103–119
- Kisi O, Cobaner M (2009) Modeling river stage–discharge relationships using different neural network computing techniques. *Clean Soil Air Water* 37(2009):160–169
- Koza JR (1992) Genetic programming: on the programming of computers by means of natural selection. MIT Press, Cambridge
- Napolitano G, Serinaldi F, See L (2011) Impact of EMD decomposition and random initialization of weights in ANN hindcasting of daily streamflow series: an empirical examination. *J Hydrol* 406(3–4):199–214
- Quinlan JR (1992) Learning with continuous classes. In: *Proceedings of Australian joint conference on artificial intelligence*. World Scientific Press, Singapore, pp 343–348
- Rehman N, Mandic DP (2010) Multivariate empirical mode decomposition. *Proc R Soc* 466:1291–1302
- Rilling G, Flandrin P, Goncalves P (2003) On empirical mode decomposition and its algorithms. In: *Proceedings of IEEE-EURASIP workshop on nonlinear signal and image processing NSIP-03*, Grado, Italy, pp 8–11
- Seghal V, Sahay RR, Chatterjee C (2014) Effect of utilization of discrete wavelet components on flood forecasting performance of wavelet based ANFIS models. *Water Resour Manag* 28(2014):1733–1749
- Shiri J, Kisi O (2010) Short-term and long-term streamflow forecasting using a wavelet and neuro-fuzzy conjunction model. *J Hydrol* 394(2010):486–493
- Sivapragasam C, Muttill N (2005) Discharge rating curve extension—a new approach. *Water Resour Manag* 19(2005):505–520
- Sudheer KP, Jain SK (2003) Radial basis function neural network for modeling rating curves. *J Hydrol Eng* 8(2003):161–164
- Tan Q-F, Lei X-H, Wang X, Wang H, Wen X, Ji Y, Kang A-Q (2018) An adaptive middle and long-term runoff forecast model using EEMD–ANN hybrid approach. *J Hydrol*. <https://doi.org/10.1016/j.jhydrol.2018.01.015> (in press)
- Tiwari MK, Chatterjee C (2011) A new wavelet–bootstrap–ANN hybrid model for daily discharge forecasting. *J Hydroinf* 13(1):500–519
- Wei S, Song J, Khan NI (2012) Simulating and predicting river discharge time series using a wavelet–neural network hybrid modeling approach. *Hydrol Process* 26:281–296
- Witten IH, Frank E, Hall MA (2005) *Data mining: practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco
- Yazeen ZM, El-shafie A, Jaafar O, Afan HA, Sayl KN (2015) Artificial intelligence based models for streamflow forecasting: 2000–2015. *J Hydrol* 530:829–844
- Zhu SJ, Zhou 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* 75:531. <https://doi.org/10.1007/s12665-016-5337-7>

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