ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645
Indian Journal of Science and Technology, *Vol 12(39), DOI: 10.17485/ijst/2019/v12i39/147941, October 2019*

Forecasting Analysis of GMDH Model with LSSVM and MARS Models for Hydrological Datasets (Case Study)

W. A. Shaikh[1,*], S. F. Shah[2], M. A. Solangi[2] and Siraj Muhammed Pandhiani[3]

[1]Department of Mathematics and Statistics, Quaid-e-Awam University of Engineering, Science & Technology, Nawabshah, Sindh, Pakistan; wajid@quest.edu.pk
[2]Department of Basic Sciences & Related Studies, Mehran University of Engineering & Technology, Jamshoro, Sindh, Pakistan; feroz.shah@faculty.muet.edu.pk, anwar.solangi@faculty.muet.edu.pk
[3]Department of General Studies, Jubail University College, Kingdom of Saudi

Abstract

Objectives: To forecast hydrological datasets using time-series forecasting model, namely, Group Method and Data Handling (GMDH). **Methods/statistical analysis:** The monthly streamflow datasets covering a period of 485 and 550 months have been collected from two well-known rivers of Pakistan, the Indus and the Chenab, respectively, for the endorsement of the GMDH model. Computed results are compared with two other forecasting models: Least Square Support Vector Machine (LSSVM) and Multivariate Adaptive Regression Splines (MARS). The accuracy of the model has been verified by the following three statistical estimations: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Correlation Coefficient (CE). **Findings:** The GMDH model has the potential to estimate with high precision the forecast real value of the hydrological datasets compared to the other models discussed in the present article. Findings show that the GMDH forecasting model is more robust than the other models discussed here. **Applications/improvements:** The novelty of this study is that it provides a trustable forecast of streamflow of the rivers.

Keywords: GMDH, LSSVM, MARS, MAE, RMSE, CE

1. Introduction

The Islamic Republic of Pakistan is bestowed with the largest irrigation system. The river system comprises 61,000 km of canals and 105,000 watercourses and irrigates around 35 million acres of land.[1] Pakistan's economy is highly dependent on agriculture. The Indus River (locally called Sindhu) runs through the entire length of country, and Chenab River (originates from Bara Lacha Pass) plays a crucial role in the irrigation system. The country's increasing agricultural growth and subsequent new stresses on limited water resources necessitate well-organised management of existing water resources rather than building new amenities to meet the challenges.[2] In water management societies, Pakistan is well-acknowledged in term of efforts made to maximise the efficiency of water management based on streamflow forecasting, a methodology that plays a vital role in helping the government's water management efforts to tackle water shortages. Therefore, forecasting river streamflow plays a significant role in the planning and operating of water management, whereas streamflow predicting is essential for improving water management efficiency and useful for irrigation, hydropower generation, recreation, and ecological and other purposes. The superiority of streamflow forecasting can be assessed in terms of

*Author for correspondence

lead-time series data and accuracy. Lead-time refers to the time interval between the forecasting date and the rate of the forecasted flow that is happening. The accuracy of forecasts is an essential requirement to improve the operational, managerial, and strategic executive process.

GMDH was introduced by Ukrainian scientist A. G. Ivakhnenko and colleagues in the late 1960s; it used a multivariate analysis to study nonlinear relations between input and output unknowns and multilayered system of modeling. The GMDH model is ideal for use in multilayered, unstructured systems. Its predictions are useful in data mining, optimisation, and pattern acknowledgement in many areas. The concept of GMDH as a forecasting model for regression estimation was used to develop and determine an analytically based quadratic node transferal function (TF) in a feed-forward network.[3,4]

Several models for estimating and predicting time arrangement have been discussed in the literature. The LSSVM model and MARS are considered the foremost among dominant models in customary time-arrangement, predicting, and are commonly used for divergence and comparison. The LSSVM and MARS show the classifications of the linear models and their potential to increase the linear component of information regarding time-series forecasting. Subsequently, many researchers have attempted to integrate the different time-series models to enhance the precision of forecasting.[5,6] New methods have been developed and are being used in these models in relation to forecasting and the tasks accomplished by them are way more complex than what the previous models could accomplish.[7,8]

This study focuses on estimating the monthly river streamflow forecasting performance of GMDH model and compares the computed forecasting results generated from using the GMDH model with LSSVM and MARS models.

2. Group Method of Data Handling (GMDH) Model

The GMDH model is a group of PC-based scientific calculations of multi-parametric datasets that highlight involuntary mechanical and parametric improvements. GMDH algorithm gives the opportunity to identify and obtain the inevitable interrelations in data and choose any optimal model to enhance the accuracy of intact present algorithms.[9]

The GMDH model has been recognised for its ability to display the complex nonlinear framework by utilising a transfer function (TF) to communicate the connection between datasets of input and output structures as expressed in the Volterra Functional Series, more commonly known as the KGP (Kolmogorov-Gabor polynomial), which is defined as

$$y = a_0 + \sum_{i=1}^{M} a_i v_i + \sum_{i=1}^{M}\sum_{j=1}^{M} a_{ij} v_i v_j + \sum_{i=1}^{M}\sum_{j=1}^{M}\sum_{k=1}^{M} a_{ijk} v_i v_j v_k + \ldots \quad (1)$$

This algebraic series is expressed by a system of TF comprising two unknowns (Neurons) defined as follows:

$$h = f(v_i, v_j) = a_0 + a_1 v_i + a_2 v_j + a_3 v_i^2 + a_4 v_j^2 + a_5 v_i v_j \quad (2)$$

Let us consider the coefficients $\{a_0, a_1, a_2, a_3, a_4, a_5,\}$ determined to expand the least square method. The input unknowns to the system (observed variable) set to x and output unknowns (predict variable) to h.

The following iterative structure was observed for GMDH model:

[Step 1]: Let $N_1 = n$ neurons from of input vector $X = (v_i, v_i, \ldots, v_n)$ in the first layer, where n is the inputs unknowns; $k = 1$ is the threshold value.

[Step 2]: For all independent unknowns, set $M = N_k$ $(N_k - 1)/2$ as new unknowns $h_1, h_2, \ldots, h_M$ in the dataset. Now construct the hypothesis for the partial quadratic polynomials (PQP) by

$$h_k = f(v_i, v_j) = a_0 + a_1 v_i + a_2 v_j + a_3 v_i^2 + a_4 v_j^2 + a_5 v_i v_j \quad (k = 1, 2, \ldots, m)$$

[Step 3]: Coefficients of TF determined by SME are in the form of $A_i = \left(X_i^T X_i\right)^{-1} X_i Y$. Here A = $\{a_0, a_1, \ldots, a_5\}$ is the unknown coefficients vector, $Y = \{y_1, y_2, \ldots, y_M\}^T$ is the output value vector from observation, and

$$X = \begin{bmatrix} 1 & v_{1i} & v_{1j} & v_{1i}v_{1j} & v_{1i}^2 & v_{1j}^2 \\ 1 & v_{2i} & v_{2j} & v_{2i}v_{2j} & v_{2i}^2 & v_{2j}^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & v_{mi} & v_{mj} & v_{mi}v_{mj} & v_{mi}^2 & v_{mj}^2 \end{bmatrix}$$

[Step 4]: Select the optimal factors and eradicate the weakest variable. The determination of the parameter of the optimal factors depends on the three performances indexes that express

how h_m values follow the output y. Column approach of h_1, h_2, …, h_n is replaced by the retained columns of h_1, h_2, …, h_m, where m is the number of columns retained. In different models, the optimal neuron of these m neurons is added to columns v_1, v_2, …, v_n for forming a new set of input factors.

[Step 5]: To test the model, the set of equations may or may not need to be improved. The lowest MSE value in the current layer was compared to the minimum value in the previous layer. Repeat step 1 and 2 if model improvement is not achieved. Otherwise, we can conclude that the iteration has ended and the network has been realised. Figures 1 and 2 show the basic architecture of the GMDH structure.

3. Least-Squares Support Vector Machine (LSSVM) Formulation

The least-squares version of the support vector machine (SVM) classifier determines the problem of minimisation by using re-manipulation, which is represented as follows:

$$Min\, J(\omega, b, e) = \frac{\mu}{2}\omega^T\omega + \frac{\zeta}{2}\sum_{i=1}^{N} e_i^2 \tag{3}$$

Subject to constraints on equality

$$y_i\left[\omega^T\phi(x_i)+b\right] = 1 - e_i \qquad (i = 1, 2, \ldots, N) \tag{4}$$

The above manipulation of the LSSVM classifier is implicitly consistent with the clarification of regression with binary objectives $y_i = \pm 1$. Applying $y_i^2 = 1$, we get

$$\sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N}(y_i e_i)^2 = \sum_{i=1}^{N}\left[y_i - \left(\omega^T\phi(x_i)+b\right)\right]^2 \qquad \left(\because e_i = y_i - \left(\omega^T\phi(x_i)+b\right)\right) \tag{5}$$

Thus, e_i develops a sense for LS data-fitting. Therefore, LSSVM classifier development is equivalent to

$$J(\omega, b, e) = \mu E_w + \zeta E_D \qquad \left(\because E_w = \frac{1}{2}\omega^T\omega \quad and \quad E_D = \frac{1}{2}\sum_{i=1}^{N}\left[y_i - \left(\omega^T\phi(x_i)+b\right)\right]^2\right) \tag{6}$$

where μ and ζ are considered as hyper-parameters that can be used to adjust the amount of regularisation versus the sum square error. Therefore, the solution depends on the ratio $\gamma = \zeta/\mu$, and the original development provides γ as a tuning parameter. Apply both parameters μ and ζ to use a Bayesian definition to LSSVM. After developing the following Lagrangian function, we obtained the solution of LSSVM model:

$$\begin{aligned}\left\{L(\omega, b, e, \alpha) = J(\omega, e) - \sum_{i=1}^{N}\alpha_i\left[\left\{\omega^T\phi(x_i)+b\right\} + e_i - y_i\right]\right. \\ \left. = \frac{\omega^T\omega}{2} + \frac{\gamma}{2}\sum_{i=1}^{N} e_i^2 - \sum_{i=1}^{N}\alpha_i\left[\left\{\omega^T\phi(x_i)+b\right\} + e_i - y_i\right]\right.\end{aligned} \tag{7}$$

Here, $\alpha_i \in \mathbb{R}$ and α_i are Lagrange multipliers. The following formations are optimal for LSSVM model.

$$\begin{cases} \dfrac{\partial L}{\partial \omega} = 0 \;\rightarrow\; \omega = \displaystyle\sum_{i=1}^{N}\alpha_i\phi(x_i); \\ \dfrac{\partial L}{\partial b} = 0 \;\rightarrow\; \displaystyle\sum_{i=1}^{N}\alpha_i = 0; \\ \dfrac{\partial L}{\partial e_i} = 0 \;\rightarrow\; \alpha_i = \gamma e_i; \quad (\because i = 1, 2, \ldots, N) \\ \dfrac{\partial L}{\partial \alpha_i} = 0 \;\rightarrow\; y_i = \omega^T\phi(x_i) + b + e_i; \quad (\because i = 1, 2, \ldots, N) \end{cases} \tag{8}$$

After eliminating ω and e a linear system was arrived at in the place of a quadratic programming (QP) problem.

$$\begin{bmatrix} 0 & 1_N^T \\ 1_N & \Omega + \gamma^{-1} I_N \end{bmatrix}\begin{bmatrix} b \\ a \end{bmatrix} = \begin{bmatrix} o \\ Y \end{bmatrix} \tag{9}$$

$\because 1_N = [1, \ldots, 1]^T$, I_N is N^{th} order identity matrix; $\Omega \in \mathbb{R}^{N\times N}$ is kernel matrix and defined as $\Omega_{ij} = K(x_i, x_j) = \phi(x_i)^T\phi(x_j)$; and $Y = [y_1, \ldots, y_N]$.

Here, choose $K(x_i, x_j) = e^{\frac{-\|x - x_i\|^2}{\sigma^2}}$ as a kernel function, where $\sigma \in \mathbb{R}^+$ is a scale parameter, and determine the scaling of inputs in the RBF kernel.[10,11]

4. Multivariate Adaptive Regression Splines (MARS)

The MARS model is appropriate for forecasting continuous datasets outcomes and is implemented in two stages: forward and backward stepwise techniques. The forward stepwise technique uses a large set of input variables (basis function) with different knots; however, the use of this technique might result in a complex and

multilayered model.[12] Nevertheless, such a model also cannot guarantee a strong forecast as it in fact has been found to have a weak forecasting ability. For increasing the accuracy of forecast, the backward stepwise technique was thus preferred, and it was found to have the capacity to eradicate pointless variables among the chosen datasets; this may have had a weaker effect on the approximation procedure that was pruned by the MARS. The projection of x, input variable, onto a novel, y, output variable, was carried out using the technique of appropriation, a basic function that defines the point of inflection along the input range[13]:

$$y = \begin{cases} \max(0, x - c) \\ \max(0, c - x) \end{cases} \quad (10)$$

In these y functions, x is the input, and c is the threshold value, which is said to be the knot. The function is useful in forward-backward stepwise techniques used for each input unknown in order to classify the position of knots, where the value of the function changes.[14] These y functions are called Spline functions, which is indicated by a c-knot reflected pair. The following is the common formation of the MARS model.[15]

$$y = f(x) = c_0 + \sum_{i=1}^{M} c_i B_i(x) \quad (11)$$

where, output variable y is estimated by MARS model, c_0 is constant, c_i is the ith basic function coefficient determined by minimising the Root Mean Squared Errors (RMSE), and $B_i(x)$ is the ith basic function. The optimal MARS model scheme is designated based on the Generalised Cross-Validation (GCV) principle's smallest value. The GCV is defined as follows:

$$GCV(M) = \frac{\sum_{i=1}^{n} [y_i - f(x_i)]^2}{n\left(1 - \frac{C(M)}{n}\right)^2} \quad (12)$$

where y_i is the objective of output, $f(x_i)$ is the projected output, n is the number of inputs, and $C(M)$ is a penalty which is further expressed as:

$$C(M) = d \times M + M + 1 \quad (13)$$

where d is the penalty for each basic function assessed by the model, and M denotes the number of basic functions.

5. Results and Discussion

The chief aim of the GMDH model used for the present study was to analyse the input time-series hydrological data collected from Indus and Chenab rivers and arrive at accurate real values as has been discussed in the opening paragraphs of this article. Specimens of six distinct input data combinations prepared for this scheme are shown in Table 1.

The combination of M1–M6 input models was used in the training and testing phases for forecasting models. Among the combinations of input models, M1-M6 represents the number of unknowns selected on the basis of previous analyses of monthly river streamflows.

The computed results for GMDH, LSSVM, and MARS models are illustrated in Tables 2, 3, and 4, respectively. Table 2 presents the details of analysis carried out using the GMDH model, where it can be seen that M5 and M6 models perform better for Chenab and Indus Rivers, respectively. Similarly, Table 3 presents the details of the analysis carried out using the LSSVM model, and the findings of the analysis revealed that M6 model works better for both Indus and Chenab Rivers. Moreover, from the details presented in Table 4 for the analysis carried using the MARS model, it can be seen that both M3 and M6 models perform better for Indus and Chenab Rivers, respectively, which is a crucial finding. The computed values have been assessed using statistical estimations. Accuracy of the said models is shown in Table 5.

In Table 5, it can be seen that in terms of the results of statistical tools used for measuring the accuracy of estimation, such as MAE and RMSE, the small error has been achieved in the case for GMDH model compared to LSSVM and MARS models. It is the evidence that GMDH model is better than the other two models. Furthermore, in regards to the robustness of the model, that large value of CE observed in the GMDH model is

Table 1. Model structure with different combinations

Model	Input combination
M1	$y_t = f(z_{t-1}, z_{t-2})$
M2	$y_t = f(z_{t-1}, z_{t-2}, z_{t-3}, z_{t-4})$
M3	$y_t = f(z_{t-1}, z_{t-2}, z_{t-3}, z_{t-4}, z_{t-5}, z_{t-6})$
M4	$y_t = f(z_{t-1}, z_{t-2}, z_{t-3}, z_{t-4}, z_{t-5}, z_{t-6}, z_{t-7}, z_{t-8})$
M5	$y_t = f(z_{t-1}, z_{t-2}, z_{t-3}, z_{t-4}, z_{t-5}, z_{t-6}, z_{t-7}, z_{t-8}, z_{t-9}, z_{t-10})$
M6	$y_t = f(z_{t-1}, z_{t-2}, z_{t-3}, z_{t-4}, z_{t-5}, z_{t-6}, z_{t-7}, z_{t-8}, z_{t-9}, z_{t-10}, z_{t-11}, z_{t-12})$

Table 2. Forecast results by GMDH

	Indus River GMDH						Chenab River GMDH					
	Training Data Set			Testing Data Set			Training Data Set			Testing Data Set		
	CE	MAE	MSE	CE	MAE	MSE	CE	MAE	MSE	CE	MAE	MSE
M1	0.6775	0.1071	0.0198	0.5355	0.0969	0.0151	0.9450	0.0357	0.0027	0.9193	0.0419	0.0029
M2	0.9688	0.0251	0.0013	0.9518	0.0304	0.0016	0.9578	0.0290	0.0019	0.9385	0.0360	0.0025
M3	0.9761	0.0185	0.0008	0.9630	0.0251	0.0011	0.9662	0.0263	0.0016	0.9455	0.0335	0.0020
M4	0.9725	0.0204	0.0011	0.9578	0.0279	0.0014	0.9796	0.0223	0.0012	0.9498	0.0317	0.0019
M5	0.9704	0.0227	0.0012	0.9553	0.0276	0.0014	0.9891	0.0182	0.0007	**0.9547**	**0.0290**	**0.0015**
M6	0.9780	0.0172	0.0007	**0.9635**	**0.0247**	**0.0011**	0.9743	0.0253	0.0015	0.9545	0.0324	0.0017

Table 3. Forecast results by LSSVM

	Indus River LSSVM						Chenab River LSSVM					
	Training Data Set			Testing Data Set			Training Data Set			Testing Data Set		
	CE	MAE	MSE	CE	MAE	MSE	CE	MAE	MSE	CE	MAE	MSE
M1	0.9026	0.0086	0.00025	0.8680	0.0318	0.0109	0.9172	0.0421	0.0027	0.9041	0.0437	0.0042
M2	0.9032	0.0086	0.00026	0.7357	0.0279	0.0097	0.9266	0.0401	0.0040	0.9126	0.0445	0.0040
M3	0.9049	0.0086	0.00025	0.9161	0.0313	0.0109	0.9323	0.0393	0.0039	0.9196	0.0430	0.0036
M4	0.9070	0.0084	0.00027	0.8930	0.0323	0.0115	0.9342	0.0360	0.0036	0.9296	0.0414	0.0034
M5	0.9057	0.0085	0.00024	0.9217	0.0336	0.0102	0.9432	0.0345	0.0031	0.9408	0.0382	0.0027
M6	0.9145	0.0078	0.00019	**0.9335**	**0.0246**	**0.0019**	0.9418	0.0360	0.0032	**0.9467**	**0.0327**	**0.0021**

Table 4. Forecast results by MARS

	Indus River MARS						Chenab River MARS					
	Training Data Set			Testing Data Set			Training Data Set			Testing Data Set		
	CE	MAE	MSE	CE	MAE	MSE	CE	MAE	MSE	CE	MAE	MSE
M1	0.8715	0.0112	0.0006	0.7991	0.0250	0.0080	0.8840	0.0863	0.0129	0.7869	0.1180	0.0271
M2	0.9061	0.0094	0.0003	0.8561	0.0266	0.0087	0.8976	0.0354	0.0029	0.8174	0.1152	0.0297
M3	0.9034	0.0097	0.0003	**0.8968**	**0.0267**	**0.0085**	0.9069	0.0313	0.0025	0.7869	0.1213	0.0352
M4	0.9054	0.0096	0.0003	0.8958	0.0265	0.0083	0.8895	0.0300	0.0023	0.7684	0.1243	0.0396
M5	0.9042	0.0098	0.0003	0.8825	0.0275	0.0089	0.9004	0.0281	0.0021	0.7336	0.1352	0.0458
M6	0.9056	0.0096	0.0003	0.8774	0.0274	0.0089	0.9397	0.0344	0.0031	**0.8471**	**0.1112**	**0.0278**

Table 5. Comparison of forecast results

Dataset	Model	CE	MAE	RMSE
Indus River	**GMDH**	**0.9635**	**0.0247**	**0.0011**
	LSSVM	0.9335	0.0246	0.0019
	MARS	0.8968	0.0267	0.0085
Chenab River	**GMDH**	**0.9547**	**0.0290**	**0.0015**
	LSSVM	0.9467	0.0327	0.0021
	MARS	0.8471	0.1112	0.0278

indeed an authentic evidence for its robustness; hence, GMDH is appropriate for estimating the forecasting the real value.

6. Conclusion

Three different time-series forecasting models have been compared with computed statistical estimations. The findings showed that the GMDH model was indeed more

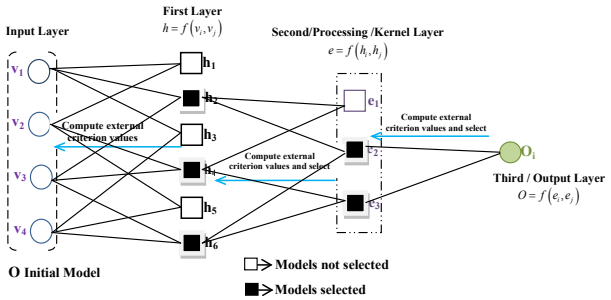

Figure 1. Architecture of GMDH model.

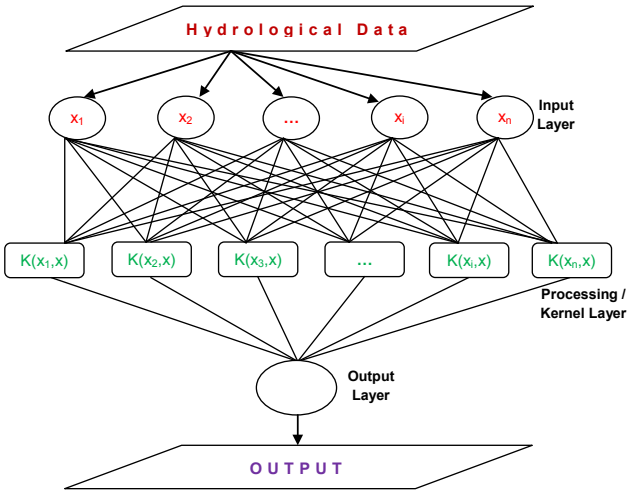

Figure 2. Architecture of LSSVM model.

useful and more accurate compared to the other two models, LSSVM and MARS, in estimating the optimal forecasting real value on the monthly river streamflow datasets for both Indus and Chenab Rivers of Pakistan. Based on the numerical investigations, it is concluded that the GMDH model's forecasting performance is more stable and robust than LSSVM and MARS models.

References

1. Water management, impacts and conflicts: case of Indus water distribution in Sindh, Pakistan. [cited 2012 Oct]. https://www.researchgate.net/publication/233933586_Water_Management_Impacts_and_Conflicts_Case_of_Indus_water_distribution_in_Sindh_Pakistan.
2. Ahmed A, Iftikhar H, Chaudhry G. Water resources and conservation strategy of Pakistan. In: The Pakistan development review; 2007. P.997–1009.
3. Li Q, Tian Y, Zhang G. GMDH modeling based on polynomial spline estimation and its applications. Int J Math Comput Phys Electr Comput Eng. 2013;7(3):458–62.
4. Uddin SM, Rahman A, Ansari EU. Comparison of some statistical forecasting techniques with GMDH predictor: a case study. Mech Engineering. 2017(47):16–21.
5. Fereydooni M, Rahnemaei M, Babazadeh H, Sedghi H, Reza Elhami M. Comparison of artificial neural networks and stochastic models in river discharge forecasting, (Case study: Ghara-Aghaj River, Fars Province, Iran). Afr J Agric Res. 2012;7(40):5446–58.
6. Samsudin R, Saad P, Shabri A. River flow time series using least squares support vector machines. Hydrol Earth Syst Sci. 2011;15(6):1835–52.
7. Kisi O. Modeling discharge-suspended sediment relationship using least square support vector machine. J Hydrol. 2012;456–457:110–20.
8. Rate AIFE. Comparison between MEMD-LSSVM and MEMD-ARIMA in forecasting exchange rate. J Theor Appl Inf Technol. 2017;95(2):328–39.
9. Ebtehaja HBo, Khoshbina F, Joo Bongc CH, Ab Ghanid A. Development of group method of data handling based on genetic algorithm to predict incipient motion in rigid rectangular storm water channel. SCI IRAN TRANS A: Civil Eng. 2017;24(3):1000–9.
10. Pandhiani SM, Shabri A. Time series forecasting by using hybrid models for monthly streamflow data. Appl Math Sci. 2015;9(57):2809–29.
11. Ismail S, Shabri A, Samsudin R. A hybrid model of self-organizing maps and least square support vector machine for river flow forecasting. Hydrol Earth Syst. 2012;16:4417–33.
12. Andres JD, Lorca P, de Cos Juez FJ, Sánchez-Lasheras F. Bankruptcy forecasting: a hybrid approach using fuzzy c-means clustering and multivariate adaptive regression splines (MARS). Expert Syst Appl. 2010;38:1866–75.
13. Adamowski J, Chan HF, Prasher SO, Sharda VN. Comparison of multivariate adaptive regression splines with coupled wavelet transform artificial neural networks for runoff forecasting in Himalayan micro-watersheds with limited data. J Hydroinform. 2011;14(3):731–44.
14. Sharda V, Patel R, Prasher SO, Ojasvi P, Prakash C. Modeling runoff from middle Himalayan watersheds employing artificial intelligence techniques. Agric Water Manag. 2006;83(3):233–42.
15. Fathian F, Mehdizadeh S, Kozekalani Sales A, Safari MJS. Hybrid models to improve the monthly river flow prediction: integrating artificial intelligence and non-linear time series models. J Hydrol. 2019;575:1200–13.