J. Earth Syst. Sci. (2019) 128:141
https://doi.org/10.1007/s12040-019-1170-1

Monthly prediction of streamflow using data-driven models

Behrouz Yaghoubi[1], Seyed Abbas Hosseini[2] and Sara Nazif[3,*]

[1] *Department of Water Engineering, Kermanshah Branch, Islamic Azad University, Kermanshah, Iran.*
[2] *Department of Civil Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran.*
[3] *School of Civil Engineering, College of Engineering, University of Tehran, Tehran, Iran.*
**Corresponding author. e-mail: snazif@ut.ac.ir*

MS received 17 September 2016; revised 14 September 2018; accepted 10 January 2019;
published online 31 May 2019

The estimation of river run-off is a complex process, but it is of vital importance to the proper operation of reservoirs, the design of hydraulic structures, flood control, drought management and the supply of water and electricity. The high uncertainty in rainfall–run-off modelling and lack of data has made the development of rainfall–run-off models with acceptable levels of accuracy and precision challenging. Furthermore, the rainfall–run-off models commonly do not provide an explicit relationship between run-off and other variables to be used for run-off-related investigations. To overcome the knowledge and information shortage in rainfall–run-off modelling, data-driven models have been used instead of conceptual models for the development of rainfall–run-off models. In this paper, three data-driven models, the genetic algorithm-support vector regression (GA-SVR), genetic algorithm-artificial neural network (GA-ANN) and the group method of data handling (GMDH) have been used to predict the monthly run-off of the Gavehroud basin. Their performances are compared with a conceptual hydrological model (HYMOD) whose parameters are calibrated using the GA. To this end, the monthly data on precipitation, temperature and run-off at the Gavehroud basin over 49 yr (1960–2009) were analysed. Evaluation of the results using performance evaluation indicators showed that the hybrid model of GA-SVR provided better accuracy in predicting the nonlinear behaviour of flow data than the GA-ANN, GMDH and HYMOD.

Keywords. Run-off prediction; GA-SVR; GA-ANN; GMDH; HYMOD.

1. Introduction

Stream flow prediction has been considered from different perspectives for the integrated planning and management of water resources and demands at the watershed scale. Rainfall–run-off models are used for this purpose. To use these models in a watershed scale, their parameters should be calibrated to watershed characteristics. This results in difficulties in model development, especially when limited data is available. As an alternative to conceptual rainfall–run-off models, data-driven models can be used for rainfall–run-off modelling, especially when there is limited data about different parameters that affect run-off production in the study area.

Data-driven methods are commonly based on regression and neural network techniques. Neural networks are used to model nonlinear hydrological processes such as the prediction of dam reservoir inflow and rainfall–run-off and soil moisture based on satellite data (Srivastava *et al.* 2013). Multi-layer perceptron (MLP) methods are artificial neural networks (ANNs) which normally use a back propagation algorithm for network training (Rumelhart *et al.* 1985). If the MLP coefficients are

determined based on smart optimisation methods such as genetic algorithms (GAs), the prediction results can improve (Sedki *et al.* 2009).

Even though ANNs offer good predictions for simple problems, they are not efficient at solving complex hydrological problems. The support vector machine (SVM) has been applied by numerous researchers in various fields to overcome limitations in the implementation of ANNs (Ishak *et al.* 2013; Sudheer *et al.* 2014). The SVM structure was developed by Vapnik *et al.* (1997) to include nonlinear cases. SVM models are based on probability training theory, a monitored learning method used for classification and regression analysis. Various researchers have implemented SVM to predict hydrological processes such as rainfall–run-off (Liong and Sivapragasam 2002; Khadam and Kaluarachchi 2004).

Some researchers have used SVM to determine the radial function structure in networks and for modelling the rainfall–run-off relationship (Choy and Chan 2003). Yu *et al.* (2004) presented a method for daily run-off prediction by combining SVM with the turbulence theory. Bray and Han (2004) used SVM to predict run-off, focusing on the selection of the proper model and identification of the model structure and the relevant parameters. Dibike *et al.* (2001) demonstrated the various features of SVM in hydrological predictions. The classification of remote sensing data and the model rainfall–run-off using this method provided a better performance than the ANN method.

Lin *et al.* (2009) used a SVM to predict effective reservoir flow. The results showed that SVM can be trained much faster than common ANNs. SVM also provided more accurate predictions than back propagationneural network methods. Asefa *et al.* (2006) used the SVM in hydrological models (HYMODs) to describe the relative uncertainty in data calibration. Kisi and Cimen (2011) used SVM and wave function to predict monthly stream flow rates. Mean square error (MSE) evaluation criteria indicated that the prediction results were more accurate.

Noori *et al.* (2011) used the SVM to select the inputs of SVM and ANN models to predict monthly flows. They introduced new evaluation criteria for smart prediction models based on the results. The new criteria showed the superiority of SVM for prediction. Sudheer *et al.* (2014) used a particle swarm optimisation (PSO) algorithm to modify the SVM modelling parameters. Su *et al.* (2015) used the SVM-GA hybrid model to predict the concentration of chlorophyll a in the Miyun reservoir in northern China. The results showed that the model could solve the nonlinear problem and complex system and it was appropriate for the simulation and prediction of chlorophyll a concentration in a reservoir. Cheng *et al.* (2015a) combined the ANN and SVM methods to predict the monthly flow in the Xinfengjiang reservoir in China and showed that the hybrid model was more efficient. By combining a quantum-behaved PSO (QPSO) algorithm with ANN to determine network weight, Cheng *et al.* (2015b) predicted the daily run-off of the Hongjiadu reservoir in China. The results showed that the hybrid QPSO-ANN predicted the run-off better than only ANN.

Wang *et al.* (2015) considered various neural network models for the daily run-off prediction. The results demonstrated that the singular spectrum analysis-ANN (SSA-ANN) model performs better than the nonlinear perturbation model (NLPM) based on ANN (NLPM-ANN). The SSA-ANN was evaluated using different inputs and showed that SSA-ANN for precipitation and run-off inputs and different time delays improved the results compared with those obtained only with the precipitation input. Ivakhnenko (1971) developed the extra-mental group method of data handling (GMDH), a data-based method that can be used for rainfall–run-off modelling, as a multivariate analysis method for identifying and modelling complex systems. GMDH can be used to model complex systems without specialised initial knowledge. The main idea of GMDH is to develop an analytic function based on a progressive network according to a binomial transfer function (Muller and Ivakhnenko 1996). Chang and Hwang (1999) and Samsudin *et al.* (2011) used GMDH for river flow forecasting. Samsudin *et al.* (2011) used GMDH and the least squares support vector machine (LSSVM) for this purpose. Badyalina and Shabri (2015) used GMDH on ungauged basins to predict flood quantiles and showed that its performance was much better than the traditional linear regression model.

HYMOD is a nonlinear conceptual model with applications in modelling rainfall–run-off and in flood warning systems. HYMOD was introduced and generalised using rainfall–run-off phase dispersion minimisation (Moore 1985; Boyle 2001).

The calibration of this model is an important issue and different optimisation algorithms are employed for this purpose. Singh and Bárdossy (2015) used the simple and effective optimisation

algorithm and sequential replacement of weak parameters (SRWP) to estimate HYMOD parameters and compared the results with those of other optimisation algorithms. Multi-objective particle swarm optimisation, non-dominated sorting GAs, multi-objective shuffled complex evolution metropolis algorithms and the multi-objective shuffled complex differential evolution algorithm have been used for the automatic calibration of HYMOD (Guo *et al.* 2013).

In the current study, the hybrid GA-support vector regression (GA-SVR), GMDH, the hybrid GA-GA-ANN and HYMOD were applied to a model Gavehroud river flow. This river flows in a mountainous watershed located in western Iran HYMOD is a conceptual rainfall–run-off model that includes five parameters. In this study, HYMOD was linked to GA to determine the optimal value of the model parameters. For the development of the data-driven methods (GA-SVR, A-ANN and GMDH), run-offs from the previous 1–6 months are used in addition to the common input variables of temperature and rainfall used in HYMOD as the model input variables. During the development of each data-driven model, the best set of its input variables and its optimal structure are determined using optimisation algorithms. Although the use of data-driven models for the prediction of a stream flow is not an innovation, in the present study, kernel function parameters are used as a chromosome in the GA models to optimise the SVR parameters which were not considered in previous studies. The GA-SVR model was used in a case study to compare the applicability of this new approach. The obtained results in this study through a comparison of different data-driven models with HYMOD will help in selecting the appropriate simulation approach.

2. Methodology

In this study, the GA-SVR, GA-ANN, GMDH and HYMOD methods were implemented to predict monthly flow rates in the Gavehroud watershed. The predictors used in the data-driven models are precipitation, temperature and river flow of up to three previous months. The inputs of HYMOD are precipitation and evapotranspiration based on the conceptual model. About 587 series of monthly data are used in the development of models. The representative temperature and rainfall of the study area are calculated as the average of the recorded temperature and rainfall in each time step.

2.1 *Genetic algorithm-support vector regression*

SVM is a supervised learning method that has different types and is used for classification and regression purposes. SVR is a type of SVM used to predict time series based on the regression structure. The prediction ability of SVR is fully dependent on its structure. Trial-and-error is generally used to determine the appropriate structure including kernel function type and SVR parameters, but this method is time-consuming and even may not lead to the best structure. Thus, optimisation methods such as GA were used to find the optimal kernel function type and SVR parameters. Figure 1 is a flowchart of the GA-SVR optimisation (Wu *et al.* 2009) method that is used in the current study for river flow prediction. A brief description of SVR is provided in the next section and further details can be found in the studies of Lin *et al.* (2009) and Zhu *et al.* (2016).

Different combinations of the inputs (predictors) are used in developing the GA-SVR model to find the combination with the best performance in river flow prediction. The most appropriate kernel function is determined through the optimisation process of GA-SVR. The kernel function parameters including the polynomial degree (d), the polynomial constant (b), the variance of the RBF function (σ) and C and ε are optimised through the application of GA-SVR to produce the best results for river flow prediction. The SVR model parameters and the type of kernel function and parameters were directly coded in the chromosome and the best one is determined by minimising the fitness function which is considered to be the MSE.

The parameters of GA, the crossover probability, the mutation function and the mutation rate have been considered to be 0.7, uniform and 0.04, respectively. Of the data, 80% was used for training and the rest was used for testing model performance. It should be noted that these values are determined based on initial investigations of model convergence time and the ability of the algorithm to find the global optimal solution.

2.1.1 *SVR structure*

SVR as a proper method for data classification and regression is briefly introduced. For the given training sample: Data $= \{X_i, d_i | X_i \in R^n, d_i \in R\}_{i=1}^{N}$, where X_i is the n-dimensional input vector and d_i

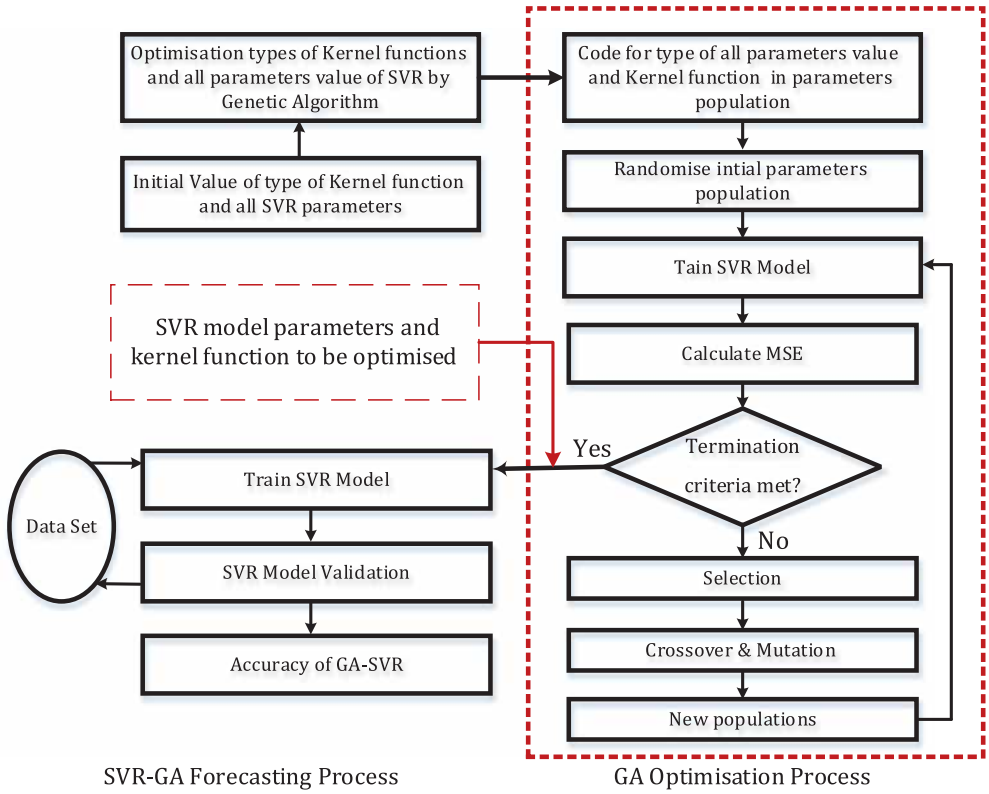

Figure 1. Flowchart of the GA-SVR optimisation process.

is the 1D desired output at the sample point i. The aim of SVR is to find the regression function in the form of equation (1) as

$$y_i = \omega\,\phi\,(X_i) + b \forall\, i = 1, 2, \ldots, N, \tag{1}$$

where $\phi\,(X_i)$ is the nonlinear mapping, ω and b are weights and bias of the regression function, respectively. The following penalty function is used in SVR (Gao and Jiang 2011):

$$\begin{cases} |d_i - y_i| \leq \varepsilon, \text{ not allocating a penalty,} \\ |d_i - y_i| > \varepsilon, \text{ allocating a penalty,} \end{cases} \tag{2}$$

where ε denotes the degree of tolerance to error. If the absolute simulation difference from the observation is less than ε, it is considered to be a perfect simulation. When the difference between the estimated and observed values is less than ε, the loss value will be zero. The parameter of this regression function can be acquired by minimising the flowing objective function as follows.

It can be demonstrated that $L_\varepsilon\,(y_i, f\,(x_i)) = \xi_i^+ + \xi_i^-$. To simplify the model, $1/2||\omega||^2$ is minimised as

$$\min\left[\frac{1}{2}||\omega||^2 + C\sum_{i=1}^{N} L^\varepsilon(y_i, d_i)\right] \tag{3}$$

$$L^\varepsilon\,(y_i, d_i) = \max(0, |d_i - y_i| - \varepsilon), \tag{4}$$

where parameter C is the positive regularisation constant that determines the trade-off between the generalisation ability and the accuracy in the training data. The importance of the simulation accuracy over its generalisation ability increases as the value of C increases. A dual problem can then be derived for the above model to minimise the following function using kernel function $k\,(x_i, x_j) = \Phi(x_i)\Phi(x_j)$:

$$\begin{aligned} &\underbrace{\text{minimise}}_{\alpha_i^+, \alpha_i^-} L_D \\ &= \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\alpha_i^+ - \alpha_i^-\right)\left(\alpha_j^+ - \alpha_j^-\right) \\ &\quad \times k\,(x_i, x_j) + \varepsilon\sum_{i=1}^{N}\left(\alpha_i^+ - \alpha_i^-\right) \\ &\quad - \sum_{i=1}^{N} y_i\left(\alpha_i^+ - \alpha_i^-\right) \end{aligned} \tag{5}$$

To simplify the dual equation for quadratic programming and to solve for $f(x)$, α^+ and α^- are determined and calculated in the following equations, and the nonlinear function becomes

$$\omega = \sum_{i=1}^{N} \left(\alpha_i^+ - \alpha_i^-\right) \times k\left(x_i, x\right), \tag{6}$$

$$b = \frac{1}{|S|}\left[\sum_{s\in S} y_s - \sum_{i=1}^{N}\left(\alpha_i^+ - \alpha_i^-\right) \times k\left(x_i, x\right) - \varepsilon \times \mathrm{Sign}\left(\alpha_i^+ - \alpha_i^-\right)\right], \tag{7}$$

where S = support vector $= \{i|0 < \alpha_i^+ + \alpha_i^- < C\}$.

The following types of kernel function are introduced:

(1) Polynomial kernel:

$$k\left(x_i, x_j\right) = \left(x_i^{\mathrm{T}}.x_j + b\right)^d. \tag{8}$$

(2) Radial basis function (RBF) kernel function:

$$k\left(x_i, x_j\right) = \exp\left(\frac{||x_i - x_j^2||}{2\sigma^2}\right). \tag{9}$$

(3) MLP kernel function:

$$k\left(x_i, x_j\right) = \tanh\left(\beta_0 + \beta_1 x_i^{\mathrm{T}}.x_j\right). \tag{10}$$

The parameter d denotes the degree of the polynomial kernel function, b denotes the constant polynomial kernel function and β_0 and β_1 are the constant values.

2.2 *GA-ANN model*

A hybrid of ANN and GA, GA-ANN, is also used for river flow prediction. In an MLP, the output is expressed as Out $= f\left(x|w,b\right) = f(w^{\mathrm{T}}x + b)$ where the vectors w and b represent weight and bias, respectively. Figure 2 shows the architectural graph of an ANN. MLP minimises the error in simulation of the outputs. The error is obtained as in equation (11) as the difference between the neural network output and the actual value of the inputs:

$$\mathrm{Error} = R_i^{\mathrm{Obs}} - R_i^{\mathrm{Sim}}, \tag{11}$$

where R_i^{Obs} and R_i^{Sim} are the observed (actual) and simulated run-offs calculated for the month i,

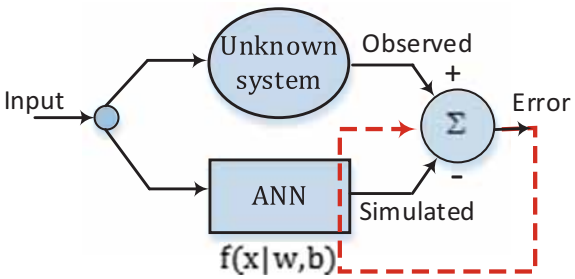

Figure 2. Architectural graph of an ANN.

respectively. To minimise the error, the MSE objective function is used:

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}\left(R_i^{\mathrm{Obs}} - R_i^{\mathrm{Sim}}\right)^2. \tag{12}$$

The weight and bias vectors are determined at the end of the calculation. In this study, GA is used to train neural networks similar to Sedki *et al.* (2009). In other words, w and b vectors are calculated using GA to minimise the fitness function of equation (12). Next, the fitness function is calculated using a classic minimisation scheme (figure 3). One of these methods is the Levenberg–Marquardt algorithm (trainlm function) where gradient descent, back propagation and adaptive learning techniques are implemented. In this process, different structures of the ANN model are also evaluated. For this purpose, the maximum number of hidden layers and the neurons in each hidden layer have been considered to be 2 and 20, respectively, with regard to the number of available data sets. Based on the initial investigations, the transition function of the hidden layers has been considered to be a hyperbolic tangent sigmoid (tansig) function. Furthermore, the crossover probability, the mutation function and the mutation rate have been considered to be 0.9, uniform and 0.08, respectively. About 80% of the data was used for training and 20%, for the testing of GA-ANN models.

2.3 *GMDH model*

The last data-driven model used for river flow prediction is GMDH. The GMDH is a self-organising unidirectional neural network which can be used to model a complex nonlinear system. This network has a multilayer structure with each layer playing the role of a nonlinear function of inputs (Samsudin *et al.* 2011). Each layer includes one or more units plus inputs and output arcs. Figure 4 illustrates a typical GMDH structure. Each unit corresponds to

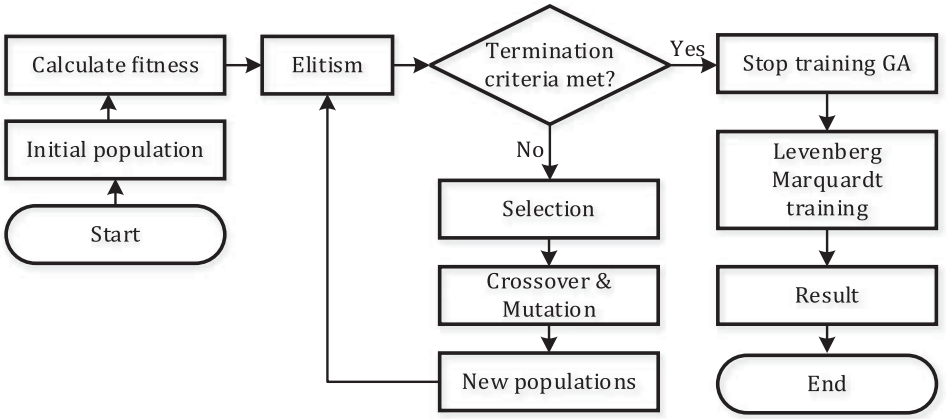

Figure 3. Flowchart of the optimisation process of the GA-ANN.

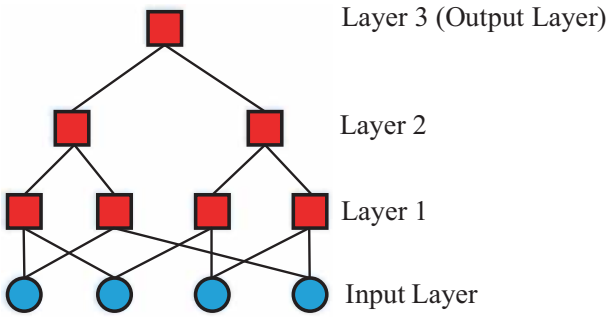

Figure 4. Evolved structure of the generalised GMDH-type neural network model with four inputs and seven units in three layers.

the Ivakhnenko polynomial form for a case having inputs x and y as follows (Ivakhnenko 1971):

$$z = a + bx + cy + dx^2 + exy + fy^2 \qquad (13)$$

or

$$z = a + bx + cy + dxy. \qquad (14)$$

Basic GMDH learning is a self-organisation method which includes the following steps:

(1) The learning data sample includes the dependent variable y and independent variables x_1, x_2, ..., x_m and is divided into a training and a test set.
(2) The input data of the m input variable is fed to the model and combined (m, 2) units from every two variable pairs at the first layer are generated.
(3) The weights of all units (parameters of each unit) are determined using the training set.
(4) The MSE between the model output and the data of each unit is determined using checking data.
(5) The units are sorted based on MSE and bad units are eliminated.
(6) The outputs of units in the first layer are used as inputs for the next layer and a multilayer structure is developed by applying steps 2 and 5.
(7) If the MSE is greater than that of the previous layer, the process of adding new layers is stopped and the minimum MSE unit in the highest layer is considered to be the final model output.

Steps 4 and 5 are the main and basic techniques of GMDH algorithm development, respectively. They are called regularity criteria and achieve the best structure at step 7. About 80% of available data series are used for GMDH model training while the remaining data are used to test their performance.

2.4 *Hydrological model*

To apply HYMOD, the watershed should be divided into infinitely small non-interacting areas. Each area (or point) has a specific storage capacity (C) and can be filled by storing rainfall. Rainfall and potential evapotranspiration during a specific period of time is a feature of these points. If the stored water at a point exceeds C, then the excess water flows from that point in the form of surface run-off.

The storage capacity is different for different points in accordance with the spatial distribution of watershed characteristics such as soil structure. The frequency distribution function for the

different storage capacities within the watershed can be expressed as

$$F(C) = 1 - \left(1 - \frac{C}{C_{\max}}\right)^{b_{\exp}}, \quad 0 < C < C_{\max}, \tag{15}$$

where F is the cumulative probability for storage C to occur at an arbitrary point in the watershed. $C_{\max}$ (mm) is the maximum possible storage capacity in the watershed. The exponent $b_{\exp}$ (with a value between 0.1 and 2) represents the spatial variability of soil moisture distribution that exists at different points in the watershed.

HYMOD is a relatively simple model for excess rain calculation and relates it to two series of reservoirs (three quick release reservoirs and one slow release reservoir). Figure 5 is a schematic of HYMOD (Vrugt *et al.* 2008). The inputs are potential evapotranspiration, rainfall and stream flow rate. The parameters and their allowable ranges are given in table 1 (Vrugt *et al.* 2008). The optimum values for these parameters were determined using a GA. The crossover probability was 0.7, the mutation function was uniform and mutation rate was 0.05. In this model, 80% of the data was used for training and 20% for testing.

2.5 *Optimisation*

In this study, the optimisation method is used for the calibration of HYMOD as a conceptual rainfall–run-off model and to determine the optimal structure of the data-driven models used for rainfall–run-off simulation. In using HYMOD in combination with GA, the decision variables are the five parameters of HYMOD to be calibrated. The objective function of the optimisation model is to minimise the MSE of the model outputs and observed run-off values. The constraints are the reasonable ranges for HYMOD parameter variation.

The objective of the combined GA and ANN is to find the optimal structure of the ANN model that results in the minimum simulation error as quantified by MSE. The decision variables are the number of ANN model hidden layers and neurons. In the SVR-GA model, GA is used to optimise the SVR parameters for minimum simulation error based on the MSE index. The decision variables are the type of kernel function and parameters d, ε, b, σ and C. In all applications of the optimisation tool (GA), optimisation stops when no further improvement is observed after 10 successive iterations.

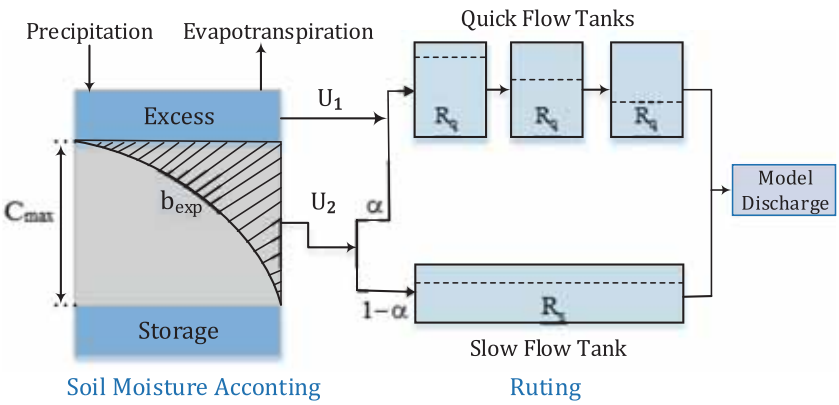

Figure 5. Schematic of the HYMOD.

Table 1. *Prior ranges and description of the HYMOD parameters.*

Parameter	Unit	Description	Minimum	Maximum
$C_{\max}$	mm	Maximum storage capacity in watershed	1	500
$b_{\exp}$	–	Spatial variability of soil moisture distribution	0.1	2
α	–	Distribution factor between two reservoirs	0.1	0.99
R_S	day^{-1}	Residence time of the slow release reservoir	0.001	0.1
R_q	day^{-1}	Residence time of the quick release reservoir	0.1	0.99

2.5.1 *Genetic algorithm*

GA is an optimisation method developed based on the principles of natural selection and genetics. GA encodes each decision variable into a gene. All decision variables gather in a finite-length string called a chromosome. The chromosomes are candidate solutions to the optimisation problem. To improve the solutions based on natural selection, an objective function is used to distinguish between good and bad solutions. To provide the next generation, the functions of crossover and mutation are used. In crossover, new children are generated based on the parent chromosomes. Mutation is used to provide diversity to the new generation (Holland 1975).

3. Evaluation criteria

RSR (RMSE-observation standard deviation ratio), NSE (Nash–Sutcliffe efficiency), EFF (error flood efficiency), CC (correlation coefficient) and error flood (EF) were selected to evaluate the performance of the model and compare the results. The EF index was used to evaluate rainfall–run-off performance in the reproduction of extreme run-off events with the probability of exceedance less than 30%. This index has a probability of exceedance of 40 cm. Smaller values for this index denote better performance of the model in the reproduction of extreme run-off values. The other considered performance evaluation indices consider the general performance of the model, but the EF index focuses on extreme values simulation. Because the rainfall–run-off model performance when dealing with maximum values is generally weak, this index helps to better compare the models:

$$\mathrm{RSR} = \frac{\mathrm{RMSE}}{\mathrm{STDEV}_{\mathrm{obs.}}} = \frac{\sqrt{\sum_{i=1}^{n}\left(R_i^{\mathrm{obs.}} - R_i^{\mathrm{sim.}}\right)^2}}{\sqrt{\sum_{i=1}^{n}\left(R_i^{\mathrm{obs.}} - R_{\mathrm{mean}}^{\mathrm{obs.}}\right)^2}}, \quad (16)$$

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{n}\left(R_i^{\mathrm{obs.}} - R_i^{\mathrm{sim.}}\right)^2}{\sum_{i=1}^{n}\left(R_i^{\mathrm{obs.}} - R_{\mathrm{mean}}^{\mathrm{obs.}}\right)^2}, \quad (17)$$

$$\mathrm{EFF} = \left(\frac{\sqrt{\sum_{i=1}^{n}\left(R_i^{\mathrm{sim.}} - R_{\mathrm{mean}}^{\mathrm{obs.}}\right)^2}}{\sqrt{\sum_{i=1}^{n}\left(R_i^{\mathrm{obs.}} - R_{\mathrm{mean}}^{\mathrm{obs.}}\right)^2}}\right)^2, \quad (18)$$

$$\mathrm{CC} = \frac{\sum_{i=1}^{n}\left[\left(R_i^{\mathrm{Sim.}} - R_{\mathrm{mean}}^{\mathrm{Sim.}}\right)\left(R_i^{\mathrm{Obs.}} - R_{\mathrm{mean}}^{\mathrm{obs.}}\right)\right]}{\sqrt{\sum_{i=1}^{n}\left(R_i^{\mathrm{Sim.}} - R_{\mathrm{mean}}^{\mathrm{Sim.}}\right)^2}\sqrt{\sum_{i=1}^{N}\left(R_i^{\mathrm{Obs.}} - R_{\mathrm{mean}}^{\mathrm{obs.}}\right)^2}}, \quad (19)$$

$$\mathrm{EF} = \frac{100}{m}\sum_{i=1}^{m}\frac{\left|R_i^{\mathrm{Obs.}} - R_i^{\mathrm{Sim.}}\right|}{R_i^{\mathrm{Obs.}}}. \quad (20)$$

In these equations, $R_i^{\mathrm{sim.}}$ and $R_i^{\mathrm{obs.}}$ are the simulated and observed flow rates in month i, respectively, and $R_{\mathrm{mean}}^{\mathrm{Sim.}}$ and $R_{\mathrm{mean}}^{\mathrm{obs.}}$ are the simulated and observed mean data, respectively. The value m represents the number of months with flow rates that exceed 40 cm. In hydrological simulation models, the considered assessment indices with the following values (NSE > 0.5; RSR < 0.7) usually produce satisfactory results (Moriasi *et al.* 2007). The closer values of CC to one and smaller values of EFF also indicate the better performance of the model.

4. Case study

As a case study, the monthly flow data at the outlet of the Gavehroud river watershed in western Iran was assessed. Gavehroud watershed is located in the Zagros mountain range in southern Kurdistan province and in northern Kermanshah province. The maximum altitude of this watershed is 1944 m. The regional climate is semi-dry (according to Domarten's classification) and the mean annual rainfall is 457 mm. About 44% of the annual precipitation occurs in winter. The mean temperature in the region is 14.2 °C. The mean flow in the river is 7.9 cm. The watershed area of the study region is 2081 km^2. A 49-yr (587 months) observational data set recorded between 1960 and 2009 was used. This data was collected by the Kermanshah Regional Water Company. The data corresponding to the first 39 yr (about 469 months; 80% of total data) was used for calibration and the data collected within the last 14 yr (168 months; 20% of total data) for testing (see figure 6).

5. Results and discussion

In this section the results of application of GA-SVR, GA-ANN and GMDH data-driven models as well as HYMOD conceptual model are presented and then results are compared to conclude which case would perform better in the study area.

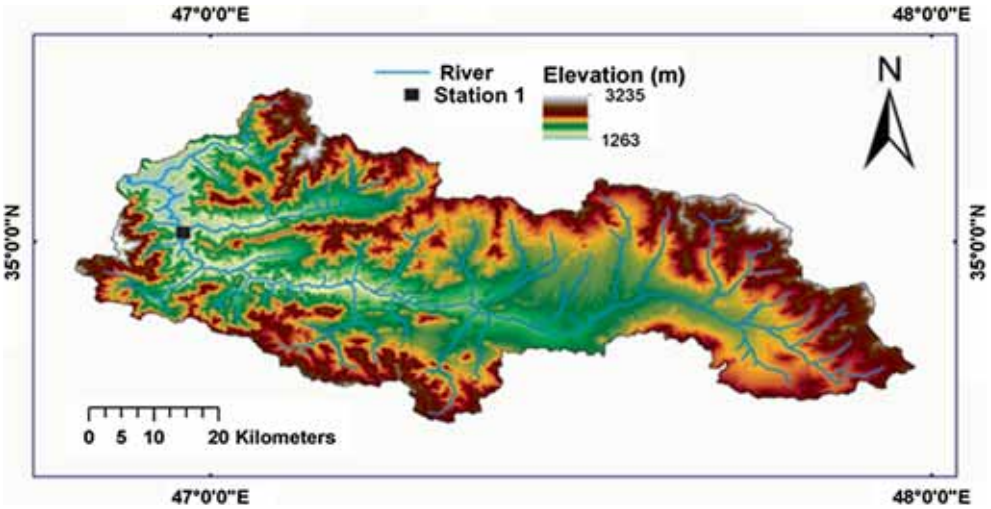

Figure 6. Geographical situation of the studied area and Station 1.

5.1 *Data preparation*

Before development of any data it is needed to prepare that and check its accuracy. At first, all considered data sets including rainfall, temperature and run-off are checked for any missing data. The few detected missing data are filled using the data of adjacent stations and considering their correlation with the used data in this study. The trend analysis of the run-off data using the Mann–Kendal method shows a slight descending trend even though it is not significant by a 90% confidence level. As expected, the seasonal behaviour is observed in the data on using the previous month's data for the prediction and it can be addressed during the development of the models.

5.2 *Genetic algorithm-support vector regression*

Based on the performance evaluation of different SVR models, the best GA-SVR model uses RBF kernel function with the C, σ and ε parameters equal to 7.13, 0.68 and 0.09, respectively. The outputs of the selected GA-SVR model are compared with the observed values in figure 7(a) for all data. Based on this figure, the model outputs well follow the observed value behaviours and fluctuations. The main weakness of model is in the simulation of peak values (flows more than 25 m^3/s which are commonly underestimated). When the river flow approaches zero, again the error in simulation is increased due to overestimations.

The values for RSR, NSE, EFF and CC for the training data were 0.24, 0.94, 0.87 and 0.97, respectively. These indices for the testing data were 0.41, 0.83, 0.76 and 0.92, respectively (see figure 8 and table 2). The model performance for the test data is weaker than that of the training data which can be due to overfitting of the training data. This is somehow expected because of the model's nature which is based on regression. The considered performance indices match well with the observed and simulated values based on CC and NSE. EFF shows that the simulated values have a more limited range in comparison with the observed values. This matches well with figure 7(a) and the model has underestimated peak values and overestimated low values. The model performance based on PSR shows that the RMSE does not exceed observed data variance and therefore is acceptable.

5.3 *Genetic algorithm-artificial neural network*

The selected ANN model structure includes two hidden layers with 14 and 7 neurons in the first and second hidden layers, respectively. Outputs of the selected GA-ANN model are compared with the observed values in figure 7(b) for all data. Based on this figure, the model outputs well follow the observed value behaviours, although more fluctuations especially in low values are observed. The model tends to underestimate the river flows especially in the simulation of peak values.

RSR, NSE, EFF, CC and EF indices results for selected GA-ANN models are presented in table 2 and figure 8. The values of the RSR, NSE, EFF and CC indices for the training data were 0.43, 0.81, 0.71, and 0.9, respectively. These indices for the test data were 0.43, 0.82, 0.75, and 0.9, respectively. The model performance for the test and

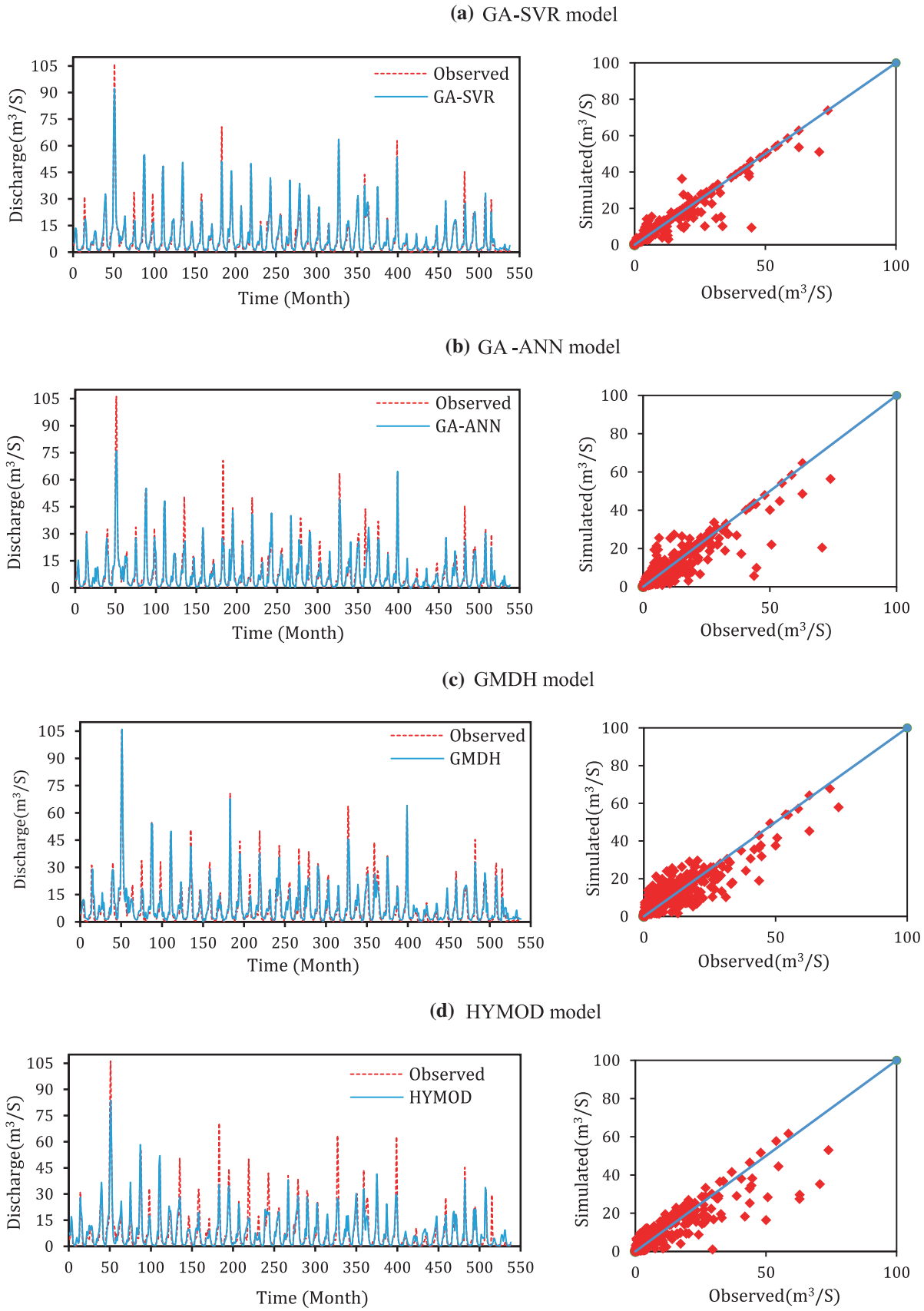

Figure 7. Observed and forecasted monthly streamflow for four models.

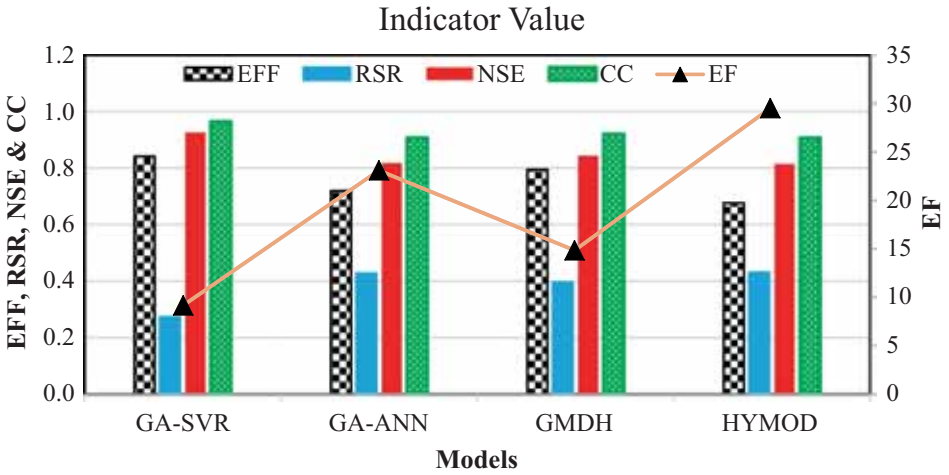

Figure 8. Results of the run-off simulation models for all data.

Table 2. *Results corresponding to the training and testing stages for GA-SVR, GA-ANN, GMDH and HYMOD.*

		Train				Test			
Model	EF	RSR	NSC	EFF	CC	RSR	NSC	EFF	CC
GA-SVR	9.1	0.24	0.94	0.87	0.97	0.41	0.83	0.76	0.92
ANN-GA	23.1	0.43	0.81	0.71	0.90	0.43	0.82	0.75	0.9
GMDH	15	0.41	0.83	0.79	0.91	0.4	0.83	0.8	0.91
HYMOD	29.5	0.42	0.83	0.68	0.91	0.52	0.73	0.66	0.85

train data is close which shows that overfitting does not happen in model development. The considered performance indices show a good match of observed and simulated values based on CC and NSE. EFF shows that the variation range of simulated values is about 75% of the observed values. This is due to the model weakness in the development of peak values that highly impact the data range. The model performance based on PSR shows that the RMSE is less than 50% of the observed values variance and therefore is acceptable.

5.4 *Group method of data handling*

The selected combination of input variables for the GMDH model comprised the precipitation of the previous month (P_{t-1}), temperature of the previous month (T_{t-1}), run-off in time step $t-12$ (R_{t-12}), run-off in time step $t-3$ (R_{t-3}), run-off in time step $t-2$ (R_{t-2}) and the run-off in time step $t-1$ (R_{t-1}). The GMDH model has six input vectors and when binary combinations of inputs are considered, the first layer of the GMDH model had 15 neurons. The results of the selected GMDH model (for all data) are given in figure 7(c). Similar to previous models, the peak values are underestimated while the minimum flow values are overestimated. It can be said that the flows of more than 20 m^3/s are underestimated and lesser than 20 m^3/s are overestimated.

RSR, NSE, EFF, CC and EF indices results are presented in table 2 and figure 8. The values of RSR, NSE, EFF and CC indices for the training data were 0.43, 0.83, 0.79 and 0.91, respectively, and for the test data were 0.4, 0.83, 0.8 and 0.91, respectively. The closeness of model performance in train and test data shows its reliability to be used for new data because there is no overfitting. The considered performance indices show a good match of observed and simulated values based on CC and NSE. EFF shows that the variation range of simulated values is about 80% of the observed values. This is due to the model weakness in developing peak and minimum values that highly impact the data range. The model performance based on PSR shows that the RMSE is about 40% of the observed values variance and therefore is acceptable.

5.5 *Hydrological model*

Table 3 shows the optimum values for HYMOD parameters determined by GA. The simulated river flows via observed values are given in figure 7(d). It is observed that some run-off variations are not well

Table 3. *Values of the HYMOD optimum parameters comparing the implemented models.*

C_{max}	b_{exp}	α	R_s	R_q
341	1.23	0.98	0.5	0.94

simulated by the model and the peak values are considerably underestimated. The minimum flow values are detected well by the model.

RSR, NSE, EFF, CC and EF indices results are presented in table 2 and figure 8. The RSR, NSE, EFF and CC indices for the training data were 0.42, 0.83, 0.68 and 0.91, respectively, and for the test data were 0.52, 0.73, 0.66 and 0.85, respectively. The simulation error is increased in the test data which may be due to model overfitting to training data or differences in statistics of the test and train data. The considered performance indices show a good match of observed and simulated values based on CC and NSE. EFF shows that the variation range of simulated values is about 67% of the observed values. This is due to the model weakness in the development of peak values higher than 20 m^3/s, which highly impacts the data range. The model performance based on PSR shows that the RMSE is about 47% of the observed value variance and, therefore, is acceptable.

5.6 *Model performance in the dry and wet periods*

To further investigate the considered model performance in river flow simulation, they were also developed for dry and wet periods separately. The dry period includes May–October and the remaining months are considered wet. In figure 9, the developed model results for the wet and dry periods are compared. As can be seen, the model performances in the wet period are completely different: some models have overestimated observed values such as GMDH while others have underestimated values such as HYMOD. All models have a weaker performance in comparison with the models developed with all data; however, based on the model performance indices that are given in table 4, there is no significant change in model performance. In the wet period, GA-SVR has the best performance and HYMOD shows the weakest results. The worse performance of data-driven models, in this case, can be due to less data used in the development of data-driven models which is of high importance in these kinds of model performance.

In the dry period, all models showed the same behaviour in simulation and underestimated the run-off peaks. Based on the PSR index, the GA-SVR model provides the best performance while the HYMOD is the second model. In other words, the performance of HYMOD is much better during the dry period in comparison with the wet period. This can be due to a variety of parameters that affect the run-off production and variability in the wet period that are not addressed in the HYMOD simulation structure, but in the dry periods, due to less uncertainty, the performance is improved. In the dry period, the worst performance is shown by GA-ANN.

5.7 *Model comparison*

Based on the given values of model performance indices in figures 8 and 9 and tables 2 and 4, it can be concluded that the GA-SVR model has performed better than the other models in all cases. This is because of the maximum CC and NSE and EFF which shows the model has the ability to reproduce the variation range of the observed values. Furthermore, this model has the minimum RMSE. This model also better simulated the peak values and the difference is less than the others. The only concern is about the considerable difference between the training and test results that questions the model reliability to be used for new data. This can be due to the regression-based nature of GA-SVR.

The GA-ANN and GMDH models have almost the same performance based on considered indices and the closeness of test and train results shows the model's trustable performance in dealing with new data. This can be due to the nonlinear structure of the relationships developed in these models as well as their robustness which is a very important issue in prediction models. A more detailed investigation of these model results shows that even though the performance indices show similar performance, the differences of observed and simulated peak values in the GA-ANN model are considerably more than that in the GMDH model. Therefore, it seems that the GMDH model is a better choice than the GA-ANN model for river flow prediction especially for flood warning systems.

A comparison of the rainfall and run-off data shows that the observed peak run-off values are not completely correlated with the rainfall data. In some cases, the very huge run-off peaks are a result of heavy rainfall, but in some cases, rainfall that is

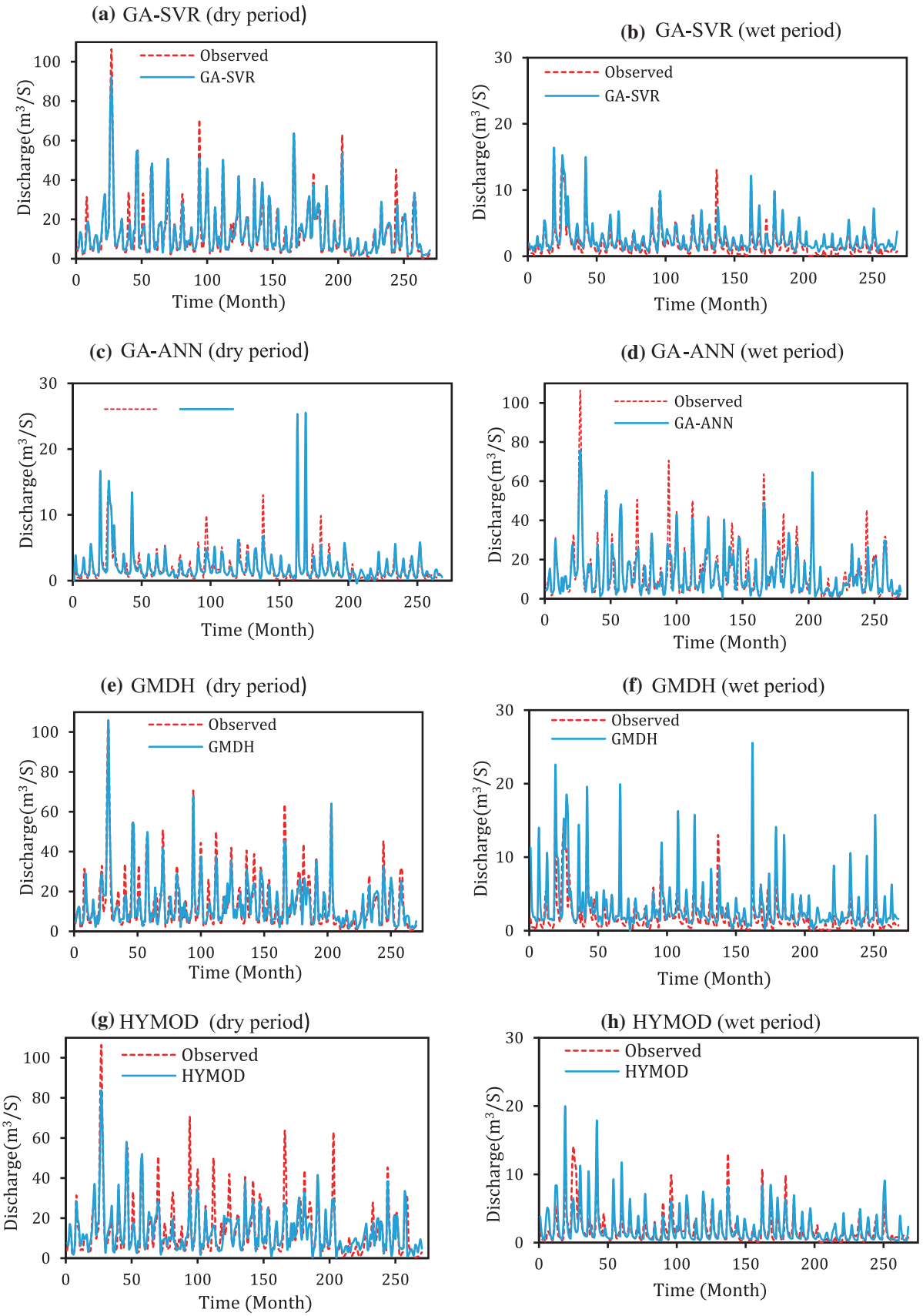

Figure 9. Observed and forecast monthly streamflow for the dry and wet periods in different models.

Table 4. *Results of run-off simulation models for wet and dry periods.*

	Wet period				Dry period			
Models	EF	RSR	NSC	CC	EF	RSR	NSC	CC
GA-SVR	9.1	0.31	0.91	0.95	–	0.50	0.75	0.91
GA-ANN	23.1	0.48	0.77	0.88	–	0.763	0.42	0.77
GMDH	15	0.41	0.83	0.91	–	1.27	−0.60	0.69
HYMOD	29.5	0.49	0.76	0.88	–	0.76	0.42	0.73

close to average has also resulted in a considerable run-off peak. In contrast, heavy rainfall does not produce a considerable run-off peak. The reason for this behaviour in the system can be the neglected snowmelt. As the study area is mountainous with considerable snow, snow melt can highly affect the run-off variability in the region.

6. Conclusion

In this study, three data-driven models of GA-SVR, GA-ANN and GMDH, besides the conceptual rainfall–run-off model of HYMOD, have been used for run-off simulation. In developing GA-SVR and GA-ANN, due to the parameters considered in SVR and the ANN structure that highly affect their performance, GA is used to find the optimal values of these parameters. It should also be noted that the optimal value of the HYMOD parameters are also determined using GA. The model performances were investigated in the three cases by using all data, wet data and dry data.

Different combinations of the considered predictors of rainfall, temperature and run-off of the previous months were checked and the results showed that the use of run-off data with a time lag beside the temperature and rainfall had a favorable impact on the results. In the case of using all data, GA-SVR shows the best performance while all models somehow show the same performance. The main issue is that in all cases, peaks are underestimated, and therefore, these models are not good choices for flood prediction. When models are used to simulate wet and dry data, separately, the performance of all models becomes worse. This can show a high sensitivity of model performance to data sets used for their development. The other issue in the case of separate simulation of dry and wet periods is that some peaks are underestimated while others are overestimated. In both wet and dry periods, the GA-SVR model shows the best performance.

The developed models in this study can be further improved by using the snow melt data as the study area is mountainous and therefore the snow melt can highly affect the run-off variations especially during the dry period. Furthermore, integrating different models can be helpful in providing more reliable results.

Acknowledgements

This study was supported by Islamic Azad University, Kermanshah Branch, Kermanshah, Iran.

References

Asefa T, Kemblowski M, McKee M and Khalil A 2006 Multi-time scale stream flow predictions: The support vector machines approach; *J. Hydrol.* **318**(**1**) 7–16.

Badyalina B and Shabri A 2015 Flood estimation at ungauged sites using group method of data handling in Peninsular Malaysia; *J. Teknologi.* **76**(**1**) 373–380.

Boyle D P 2001 *Multicriteria calibration of hydrologic models*; The University of Arizona Publisher, USA, chapter 4, pp. 95–122.

Bray M and Han D 2004 Identification of support vector machines for runoff modelling; *J. Hydroinform.* **6**(**4**) 265–280.

Chang F J and Hwang Y Y 1999 A self-organization algorithm for real-time flood forecast; *Hydrol. Process.* **13**(**2**) 123–138.

Cheng C T, Niu W J, Feng Z K, Shen J J and Chau K W 2015a Daily reservoir runoff forecasting method using artificial neural network based on quantum-behaved particle swarm optimization; *Water (Basel)* **7**(**8**) 4232–4246.

Cheng C T, Feng Z K, Niu W J and Liao S L 2015b Heuristic methods for reservoir monthly inflow forecasting: A case study of Xinfengjiang reservoir in pearl river, China; *Water (Basel)* **7**(**8**) 4477–4495.

Choy K and Chan C W 2003 Modelling of river discharges and rainfall using radial basis function networks based on support vector regression; *Int. J. Syst. Sci.* **34**(**14–15**) 763–773.

Dibike Y B, Velickov S, Solomatine D and Abbott M B 2001 Model induction with support vector machines: Introduction and applications; *J. Comput. Civil Eng.* **15**(**3**) 208–216.

Gao G Y and Jiang G P 2011 Zero-bit watermarking resisting geometric attacks based on composite-chaos optimized SVR model; *J. China Univ. Post. Telecomm.* **18**(**2**) 94–101.

Guo J, Zhou J, Zou Q, Liu Y and Song L 2013 A novel multi-objective shuffled complex differential evolution algorithm with application to hydrological model parameter optimization; *Water Resour. Manag.* **27**(**8**) 2923–2946.

Holland J H 1975 *Adaptation in natural and artificial systems*; University of Michigan Press, Michigan, USA.

Ishak A M, Remesan R, Srivastava P K, Islam T and Han D 2013 Error correction modelling of wind speed through hydro-meteorological parameters and mesoscale model: A hybrid approach; *Water Resour. Manag.* **27**(**1**) 1–23.

Ivakhnenko A 1971 Polynomial theory of complex systems; *IEEE Trans. Syst. Man Cybern.* **SMC-1** (**4**) 364–378.

Khadam I M and Kaluarachchi J J 2004 Use of soft information to describe the relative uncertainty of calibration data in hydrologic models; *Water Resour. Res.* **40**(**11**) W11505, https://doi.org/10.1029/2003WR002939.

Kisi O and Cimen M 2011 A wavelet-support vector machine conjunction model for monthly streamflow forecasting; *J. Hydrol.* **399**(**1**) 132–140.

Lin G F, Chen G R, Huang P Y and Chou Y C 2009 Support vector machine-based models for hourly reservoir inflow forecasting during typhoon-warning periods; *J. Hydrol.* **372**(**1**) 17–29.

Liong S Y and Sivapragasam C 2002 Flood stage forecasting with support vector machines; *J. Am. Water Resour. Assoc.* **38**(**1**) 173–186.

Moore R 1985 The probability-distributed principle and runoff production at point and basin scales; *Hydrol. Sci. J.* **30**(**2**) 273–297.

Moriasi D N, Arnold J G, Van Liew M W, Bingner R L, Harmel R D and Veith T L 2007 Model evaluation guidelines for systematic quantification of accuracy in watershed simulations; *Trans. ASABE* **50**(**3**) 885–900.

Muller J and Ivakhnenko A 1996 Self-organizing modelling in analysis and prediction of stock market; In: *Proceedings of the second international conference on application of fuzzy systems and soft computing–ICAFS′*96, Siegen, Germany, pp. 491–500.

Noori R, Karbassi A, Moghaddamnia A, Han D, Zokaei-Ashtiani M, Farokhnia A and Gousheh M G 2011 Assessment of input variables determination on the SVM model performance using PCA, gamma test, and forward selection techniques for monthly stream flow prediction; *J. Hydrol.* **401**(**3**) 177–189.

Rumelhart D E, Hinton G E and Williams R J 1985 Learning internal representations by error propagation; In: *Parallel distributed processing. Explorations in the microstructure of cognition, Vol 1, Foundations*, The MIT Press, Cambridge, Massachusetts, USA, pp. 318–362.

Samsudin R, Saad P and Shabri A 2011 A hybrid GMDH and least squares support vector machines in time series forecasting; *Neural Netw. World* **21**(**3**) 251.

Sedki A, Ouazar D and El Mazoudi E 2009 Evolving neural network using real coded genetic algorithm for daily rainfall–runoff forecasting; *Expert Syst. Appl.* **36**(**3**) 4523–4527.

Singh S K and Bárdossy A 2015 Hydrological model calibration by sequential replacement of weak parameter sets using depth function; *Hydrology* **2**(**2**) 69–92.

Srivastava P K, Han D, Ramirez M R and Islam T 2013 Machine learning techniques for downscaling SMOS satellite soil moisture using MODIS land surface temperature for hydrological application; *Water Resour. Manag.* **27**(**8**) 3127–3144.

Su J, Wang X, Zhao S, Chen B, Li C and Yang Z 2015 A structurally simplified hybrid model of genetic algorithm and support vector machine for prediction of chlorophyll a in reservoirs; *Water (Basel)* **7**(**4**) 1610–1627.

Sudheer C, Maheswaran R, Panigrahi B K and Mathur S 2014 A hybrid SVM-PSO model for forecasting monthly streamflow; *Neural Comput. Appl.* **24**(**6**) 1381–1389.

Vapnik V, Golowich S E and Smola A 1997 Support vector method for function approximation, regression estimation, and signal processing; In: *Advances in Neural Information Processing Systems 9*, (eds) Mozer M, Jordan M I and Petsche T, Cambridge, MA: MIT Press, 1996, pp. 281–287.

Vrugt J A, Ter Braak C J, Clark M P, Hyman J M and Robinson B A 2008 Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation; *Water Resour. Res.* **44**(**12**) W00B09, https://doi.org/10.1029/2007WR006720.

Wang Y, Guo S, Xiong L, Liu P and Liu D 2015 Daily runoff forecasting model based on ANN and data preprocessing techniques; *Water (Basel)* **7**(**8**) 4144–4160.

Wu C H, Tzeng G H and Lin R H 2009 A novel hybrid genetic algorithm for kernel function and parameter optimization in support vector regression; *Expert Syst. Appl.* **36**(**3**) 4725–4735.

Yu X, Liong S Y and Babovic V 2004 EC-SVM approach for real-time hydrologic forecasting; *J. Hydroinform.* **6**(**3**) 209–223.

Zhu S, Zhou J, Ye L and 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**(**6**) 1–12.

Corresponding editor: Prashant K Srivastava