 doi:10.20944/preprints201905.0320.v1

Support vector regression integrated with fruit fly optimization algorithm for river flow forecasting in Lake Urmia basin

Saeed Samadianfard[1], Salar Jarhan[2], Ely Salwana [3] , Amir Mosavi [4,5,6] ,Shahaboddin Shamshirband [7,8*,] Shatirah Akib[9]

[1] Department of Water Engineering, Faculty of Agriculture, University of Tabriz, Tabriz, Iran
[2] Department of Hydrosciences, Technische Univeristät Dresden, Dresden, Germany
3 Institute of Visual Informatics, Universiti Kebangsaan Malaysia 43600 Bangi, Selangor, Malaysia.4 Centre for 4Accident Research Road Safety-Queensland, Queensland University of Technology, Brisbane QLD 4059, Australia;
5 School of the Built Environment, Oxford Brookes University, Oxford OX3 0BP, UK
6 Institute of Automation, Kando Kalman Faculty of Electrical Engineering, Obuda University, Budapest 1034, Hungary

7Department for Management of Science and Technology Development, Ton Duc Thang University, Ho Chi Minh City, Viet Nam
8Faculty of Information Technology, Ton Duc Thang University, Ho Chi Minh City, Viet Nam
9 School of Science and Engineering, Teesside University, Middlesbrough TS1 3BX, United Kingdom

Abstract

Adequate knowledge about the development and operation of the components of water systems is of high importance in order to optimize them. For this reason, forecasting of future events becomes greatly significant due to making the appropriate decision. Moreover, operational river management severely depends on accurate and reliable flow forecasts. In this regard, current study inspects the accuracy of support vector regression (SVR), and SVR regulated with fruit fly optimization algorithm (FOASVR) and M5 model tree (M5), in river flow forecasting. Monthly data of river flow in two stations of the Lake Urmia Basin (Vaniar and Babarud stations on the Aji Chay and the Barandouz Rivers) were utilized in the current research. Additionally, the influence of periodicity (π) on the forecasting enactment was examined. To assess the performance of mentioned models, different statistical meters were implemented, including root mean squared error (RMSE), mean absolute error (MAE), correlation coefficient (R), and Bayesian information criterion (BIC). Results showed that the FOASVR with RMSE (4.36 and 6.33 m3/s), MAE (2.40 and 3.71 m3/s) and R (0.82 and 0.81) values had the best performances in forecasting river flows in Babarud and Vaniar stations, respectively. Also, regarding BIC parameters, Q_{t-1} and π were selected as parsimonious inputs for predicting river flow one month ahead. Overall findings indicated that, although both FOASVR and M5 predicted the river flows in suitable accordance with observed river flows, the performance of FOASVR was moderately better than the M5 and periodicity noticeably increased the performances of the models; consequently, FOASVR can be suggested as the accurate method for forecasting river flows.

Keywords: River flow forecasting; Optimization; M5 model tree; Support vector regression; Fruit fly optimization algorithm

1. Introduction

Dependable approximation of discharge is imperative in water resources management (Onyari and Ilunga, 2010). Nowadays, stream flow predicting is a dynamic and active research zone which has been studied. The stream flow is censorious to numerous actions such as planning and designing flood protections for farming lands and urban areas, and evaluating the amount allowable extracted water from a river for irrigation (Ismail et al., 2010). By happening a dynamic climate change and continues altering of environmental situations, instantaneous approaches based on utilizing real past data rather than the hydrology of the catchments has become more applicable (Fernando et al., 2012).

Generally, data mining is an influential novel methodology based on the abstraction of concealed information from huge data. These tools forecast forthcoming movements of a special system using knowledge-driven decisions resulted from enormous input-output data. M5 model tree (M5), as a sub-technique of data mining, constructs tree based linear models for continues data. Lately, the implementation of M5, as decision tree based regression method, have been stated for hydrological and water-related studies (Bhattacharya and Solomatine, 2005, 2006; Khan and See, 2006; Siek and Solomatine, 2007; Stravs and Brilly, 2007; Samadianfard et al., 2014(a,b); Esmaeilzadeh et al. 2017). Londhe and Dixit (2011) implemented M5 to estimate river flow at two stations of India. The models were established by the preceding values of gauged river flow and rainfall for predicting river flow one day beforehand. Sattari et al., (2013) inspected the proficiencies of support vector machine (SVM) and M5 model tree in forecasting flows of Sohu River. They revealed that M5 provided precise predictions comparing with SVM.

SVM is a technique in which the strong points of traditional statistical methods, which are theory-oriented and analytically simple, are utilized. SVM approach has been frequently implemented in the areas of hydrology and forecasting time series. Liong and Sivapragasam (2002) applied the method for foreseeing floods. Yu et al. (2004) suggested a method for forecasting daily runoff through combining Chaos Theory and the SVM method. Recently, support vector regression (SVR) method has been developed based on SVM and shows superiority in the prediction of hydrologic processes. Kalteh (2013) with applying Artificial Neural Network (ANN) and SVR to monthly streamflow recorded in 2 different stations revealed that both models coupled with wavelet transformation produced more accurate outcomes than the regular models. Also, the results specified that SVR models had enhanced performances comparing ANN models. Wu et al. (2008) used a genetic algorithm to optimize the SVR model, and result exhibited that the suggested model could anticipate river flow precisely in comparison with other models. Londhe and Gavraskar (2015) utilized the SVR model to one-day ahead forecast river flow in two studied locations. The model results were favorable according to the low values of the evaluating metrics.

On the other hand, Cao and Wu (2016) coupled Fruit Fly Optimization algorithm (FOA) with SVR (named FOASVR) for optimizing the parameters of SVR. The obtained results exhibited that applying FOASVR had a significant role in increasing the prediction accuracy. Lijuan and Guohua (2016) used FOASVR which hybridizes the SVR model with FOA to estimate monthly incoming tourist flow. It was reported that the suggested FOASVR is a viable option for touristic applications.

The key objective of this research is exploiting the accuracy of M5, support vector regression, and optimized SVR with FOA for river flow forecasting in the Vaniar and Babarud stations on the Aji Chay and the Barandouz rivers, respectively, located in the Lake Urmia basin of Iran. Some evaluation parameters for error estimation are utilized for assessing the enactment of the considered models. To the best knowledge of the authors, FOA has not been integrated with SVR in river flow forecasting.

2. Techniques applied in modeling

2.1 M5 model tree

With a constant value at their leaves, model trees are based on regression trees (Witten and Frank, 2005). In this regard, M5, as one of the versions of model trees, has a high capability to forecast continuous numerical attributes (Quinlan, 1992). Moreover, two different steps are necessary to develop tree models. Firstly, a splitting principle should utilize for creating a decision tree. This criterion is constructed using the standard deviation (SD) of the class values which reach a node as a size of the error. So, the standard deviation reduction (SDR) is given by:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 May 2019 doi:10.20944/preprints201905.0320.v1

$$SDR = SD(T) - \sum \frac{|T_i|}{|T|} \times SD(T_i) \quad (1)$$

Where T is a set of data that reaches the node and T_i is the ith subset of data. After the first step, data in the secondary nodes have lower SD comparing with initial nodes, so, M5 selects the split which expects to maximize error reduction. Producing a large tree is the main drawback of this step which may cause overfitting problem. Pruning techniques should be employed in order to fix this problem and avoid overfitting. Therefore, the second step for developing M5 involves these techniques and substitution of subtrees with LR functions. As a result, by applying these two steps, M5 develops an LR model for each subspace.

2.2 Support vector regression (SVR)

SVM is recognized as a technique for classification and regression (Vapnik, 1995). Generally, regression-based SVM is called SVR. For solving complex problems effectively, SVR is constructed based on minimizing the structural risk. So, insensitive loss function ($\varepsilon -$) identified as the model tolerating errors up to ε in the training data. Thus, the SVR pursues a linear function as follows:

$$F(x) = w^T x + b \quad (2)$$

Where w and b represent the coefficients of the weight vector. This can be clarified as the following problem:

$$Min\ \frac{1}{2}||w||^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*) \qquad Subject\ to \begin{cases} F(x) - y_i \leq \varepsilon + \xi_i^* \\ y_i - F(x) \leq \varepsilon + \xi_i \\ \xi_i, \xi_i^* \geq 0,\ \ i = 1,2,\dots,N \end{cases} \quad (3)$$

Where C>0 is a penalty parameter which has to be selected earlier. The constant C can grade the experimental error. Moreover, $\xi_i\ and\ \xi_i^*$ which are known as slack variables indicating distathe nce between real values and the corresponding boundary values of $\varepsilon -$tube. Hence, in order to minimize Eq. (2) subject to Eq. (3), non-Lthe R function is given by (Gunn, 1998; Cimen, 2008):

$$f(x) = \sum_{i=1}^{N}(\alpha_i^* - \alpha_i)K(x, x_i) + B \quad (4)$$

Where $K(x,x_i)$ is the Kernel function, $a_i, a_i^* \geq 0$ are the Lagrange multipliers and B is a bias term. Kernel trick is an approach which is used to solve this problem by SVR (Smola and Scholkopf, 2004). In this study, the widely implemented kernel named radial basis function (RBF) is utilized for building optimum SVR model.

Converging fast, working well in high dimensional spaces and being simple are some advantages of the selected kernel (Wu et al., 2009).

$$k(x, x_i) = \exp(-\gamma \left\| x - x_i \right\|^2) \tag{5}$$

Where γ is the bandwidth of kernel function. C, γ and ε are three predefined parameters. In this research, 1, 0.01, and 0.001 were selected as default amounts which are used in WEKA software, respectively. Fig. 1 indicates scthe hematic configuration of SVthe R model.

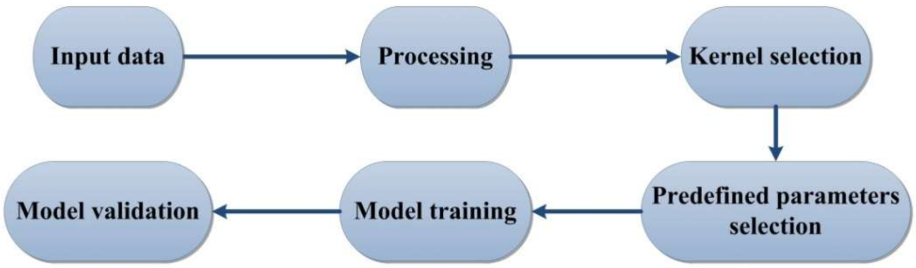


Fig. 1 schematic configuration of SVR model.

2.3. Fruit fly Optimization Algorithm (FOA)

FOA which was introduced by Pan (2012) is a swarm intelligent optimization algorithm that imitates the activities of fruit flies for searching the global optimum. Fruit flies can identify the smell from even 40 kilometers and fly on the way to it. Fig. 2 displays the food searching progression utilized by fruit fly iteratively. According to Pan (2012), the following equations are exploited to acquire the initial swarm location of a fruit fly:

$$X_0 = rand(LR)$$
$$Y_0 = rand(LR) \tag{6}$$

Where LR is the location range of the accidental, initial fruit fly swarm. Subsequently, unexpected search direction and distance for foraging of the fruit flies are given by:

$$X_i = X_0 + rand(FR)$$
$$Y_i = Y_0 + rand(FR) \tag{7}$$

where FR is the random flight range, so, smell concentration judgment value (S) can be computed by:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 May 2019 doi:10.20944/preprints201905.0320.v1

$$S_i = 1/\sqrt{x_i^2 + y_i^2} \tag{8}$$

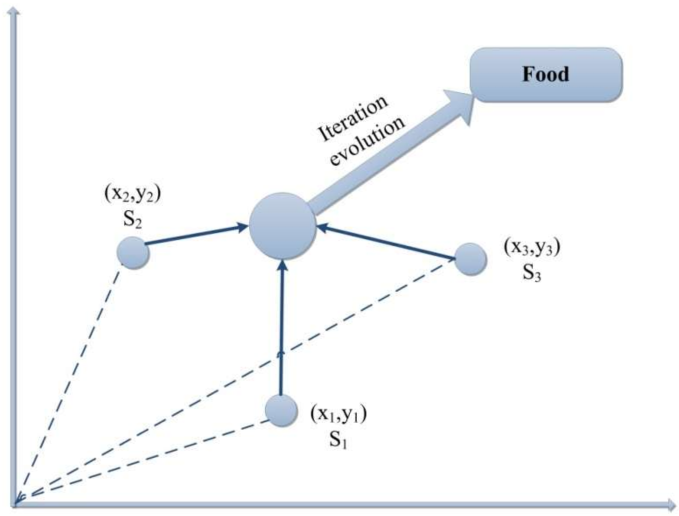


Fig. 2 food searching process utilized by fruit fly iteratively.

For improving the performance of river flow forecasting, FOA was implemented for choosing optimized values of three SVR parameters including C, ε, γ, which are connected to $(S_{ci}, S_{\varepsilon i}, S_{\gamma i})$ (i.e., $C = S_{ci}$, $\gamma = S_{\gamma i}$, and $\varepsilon = S_{\varepsilon i}$). The flowchart of the mentioned procedure (FOASVR) is displayed in Fig. 3. Moreover, the differences among the predicted and the actual values were evaluated by mean squared error (MSE), as presented in equation belthe ow:

$$MSE = \frac{\sum_{i=1}^{n}(p_i - o_i)^2}{n} \tag{9}$$

where p_i and o_i $a$re the i^{th} predicted and observed values and n is the entire number of data. The fruit fly saves the finest smell concentration value and the corresponding coordinate among the swarms, then flies towards the next place. When the new result is not superior to the previous iteration or the iteration number reaches its maximum, or the error of the prediction reaches the predefined value, this process will stop. Therefore, optimal values are acquired, and the model has the best performance with these values.

 doi:10.20944/preprints201905.0320.v1

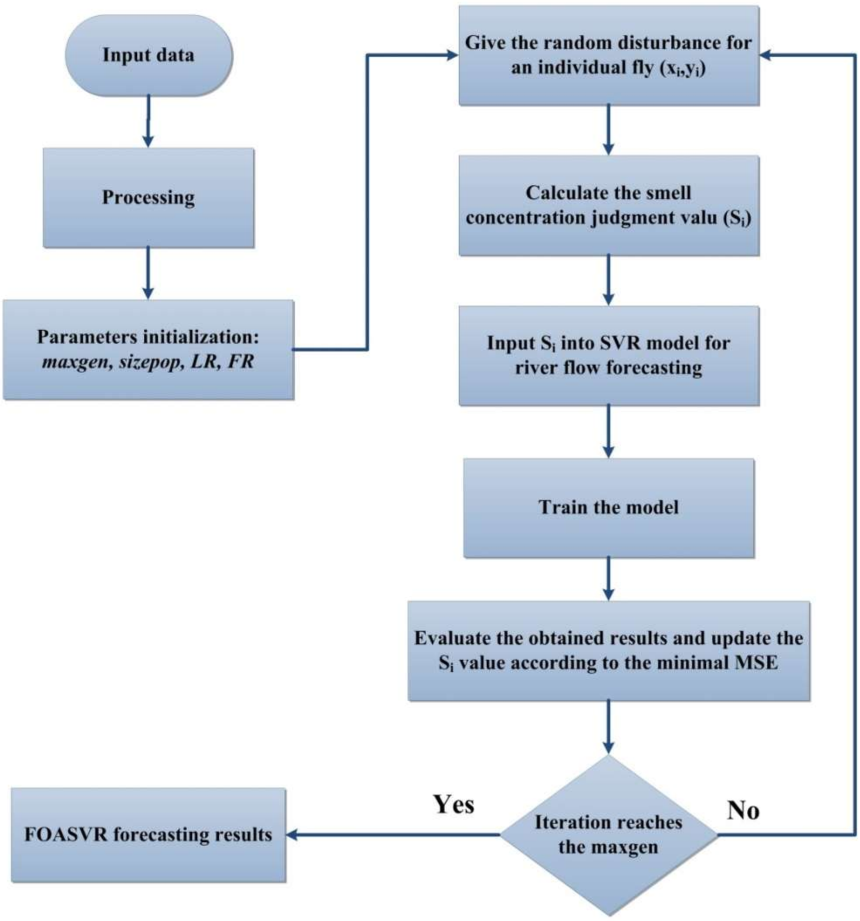


Fig. 3 The FOASVR flowchart

In this research, data were normalized to be between 0 and one because it helps to increase the accuracy of the model and to predict performance (Chang and Lin, 2001). Besides, LR and FR were chosen to be included [0, 10] and [-1, 1], respectively; Also, the maximum iteration number (maxgen) was equal to 100, and the population size (size pop) was selected to be 20 in order to have reasonable efficiency. Moreover, Libsvm toolbox was used to run SVR in this article.

3. Study area

In the current study, the monthly river flow was used for the Vaniar station on the Aji Chay Stream and the Babarud station on the Barandouz River, both located at Lake Urmia basin of Iran (Fig. 4). The observed data include 780 monthly river flows (65 years from 1952 to 2017) for Babarud station and 744 monthly records (62 years from 1952 to 2014) for Vaniar station. Moreover, there is no basic and technical way of separating training and testing data. For example, the study of Kurup and Dudani (2002) used a total of their 63% of data for model development whereas Pal (2006) used 69%, and Samadianfard et al. (2013, 2014) used 67% of total data, and Deo et al. (2018) and Samadianfard et al. (2018) used 70% of total data to develop their models.

 doi:10.20944/preprints201905.0320.v1

Thus, for developing the studied models, the data are divided into training (70%) and testing (30%). Additionally, Table 1 displays the statistics of implemented data for both stations. The observed data confirms high positive values of skewness (C_{sx} = 2.13 and 3.19). Furthermore, the low auto-correlations demonstrated the low persistence for both mentioned stations. According to the crisis of Lake Urmia, the amounts of precipitation and consequently river flow have decreased for the recent years; therefore, this may cause some difficulties in forecasting river flows.

Fig. 4 Babarud and Vaniar stations, located at Lake Urmia basin <URL1>

Table 1 Statistical parameters of the implemented data (X_{mean}, X_{max}, X_{min}, Sx, Csx, a_1, a_2, a_3 denote the overall mean, maximum, minimum, standard deviation, skewness, lag -1, lag -2, lag -3 auto-correlation coefficients, respectively)

Station	Data set	X_{mean} (m^3/s)	X_{max} (m^3/s)	X_{min} (m^3/s)	Sx (m^3/s)	Csx (m^3/s)	r_1	r_2	r_3
Babarud	Training data	8.75	66.50	0.00	9.63	2.05	0.70	0.25	-0.07
	Testing data	4.71	43.27	0.00	7.37	2.54	0.59	0.14	-0.12
	Entire data	7.74	66.50	0.00	9.28	2.13	0.69	0.25	-0.05
Vaniar	Training data	14.28	178.29	0.00	21.35	2.94	0.62	0.15	-0.11
	Testing data	5.66	65.30	0.00	10.50	3.02	0.50	0.11	-0.05
	Entire data	12.13	178.29	0.00	19.58	3.19	0.63	0.18	-0.07

4. Evaluation parameters

In this study, different evaluation parameters were considered for scrutinizing the precision of the mentioned models for river flow forecasting.

As one of the widely-used statistical parameters, root mean squared error (RMSE) measures the average amount of error (the difference between predicted and observed flows) appropriately, and it can be determined as follows:

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Q_p(i)-Q_o(i)\right)^2} \tag{10}$$

where $Q_p(i)$, $Q_o(i)$, and n represent the predicted river flow, the observed river flow, and the number of observations, respectively.

The bias in the predicted river flow is calculated by the mean absolute error (MAE) which measures the closeness of the predictions to the actual flows. Lower MAE values represent more precise predictions of river flow either equal or close to the observed values. It is calculated as follows:

$$MAE=\frac{1}{n}\sum_{i=1}^{n}\left|Q_p(i)-Q_o(i)\right| \tag{11}$$

The correlation coefficient (R), which describes the amount of linearity among simulated and observed values of river flows, ranges from -1 to 1 and is described as follows:

$$R=\frac{\left(\sum_{i=1}^{n}Q_o(i)Q_p(i)-\frac{1}{n}\sum_{i=1}^{n}Q_o(i)\sum_{i=1}^{n}Q_p(i)\right)}{\left(\sum_{i=1}^{n}Q_o(i)^2-\frac{1}{n}\left(\sum_{i=1}^{n}Q_o(i)\right)^2\right)\left(\sum_{i=1}^{n}Q_p(i)^2-\frac{1}{n}\left(\sum_{i=1}^{n}Q_p(i)\right)^2\right)} \tag{12}$$

Also, the Bayesian information criterion (BIC) was utilized to specify the best model parsimoniously which means that the model with fewer input parameters could have better performance in comparison to others. BIC measures models relative to each other; in fact, the model with the best performance has the smallest quantity of the BIC (Burnham and Anderson, 2002). It is given as follows:

$$BIC=n*\ln\left(\frac{RSS}{n}\right)+K*\ln(n) \tag{13}$$

where K indicates the number of input parameters and RSS can be determined as follows:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 May 2019 doi:10.20944/preprints201905.0320.v1

$$RSS = \sum_{i=1}^{n} \left(Q_p(i) - Q_o(i) \right)^2 \tag{14}$$

Furthermore, Taylor diagram (TD) which is a graphical illustration of the observed and forecasted data, was applied for inspecting the precision of models (Taylor, 2001). The TD has this capability to encapsulate some characteristics of the predicted and observed flows at the same time. This diagram can illustrate RMSE, R, and SD between the forecasted and actual data, simultaneously. In TD, the azimuth angle, the radial distance from the origin, and radial distance from the observed data point denote the R-value, the ratio of the normalized SD and the RMSE value of the prediction, respectively.

5. Results and discussion

For evaluating the effects of previous monthly flows, three input combinations were established. Moreover, the periodicity effect was inspected by appending a component π (1 to 12 for each month).

Table 2 Input parameters of the established models.

Model	Input parameters	Output parameters
1	Q_{t-1}	Q_t
2	Q_{t-1}, π	Q_t
3	Q_{t-1}, Q_{t-2}	Q_t
4	Q_{t-1}, Q_{t-2}, π	Q_t
5	Q_{t-1}, Q_{t-2}, Q_{t-3}	Q_t
6	Q_{t-1}, Q_{t-2}, Q_{t-3}, π	Q_t

The results of statistical parameters for studied techniques in the test phase for the Babarud station are given in Table 3. As mentioned before, π was appended to the input combinations 1,3, and 5 to examine the effect of periodicity. From the table, it is clear that the periodicity considerably increased each model's accuracy. For the FOASVR model, R increased from 0.63 (for input combination (1)) to 0.82 (for input combination (2)) and similarly, RMSE and MAE indices decreased from 5.74 to 4.36 and from 3.29 to 2.40, respectively. Regarding two previous cases, by adding periodicity component, R increased from 0.70 to 0.80 and RMSE, and MAE decreased from 5.33 to 4.50 and from 2.90 to 2.67, respectively. Finally, in the case of three previous discharges inputs, R increased from 0.67 to 0.79 and RMSE, and MAE decreased from 5.69 to 4.58 and from 3.20 to 2.67, respectively. Comparison of FOASVR, M5, and SVR models indicated that the FOASVR-2 model whose inputs are Q_{t-1}, and π had better accuracy than the M5 and SVR models. M5 also performed better than the SVR model. Overall, FOASVR performed better than SVR and M5s. Also, FOA increased the accuracy of SVR by approximately 27% for RMSE and 38% for MAE in the second scenario which performed roughly (4% RMSE and 14%MAE) better than M5. Without periodicity, FOASVR-3 indicated 6% better performance than M5-3, and they performed better than the SVR-5 model. The relative RMSE and MAE differences between the optimal FOASVR-3 model without periodicity and FOASVR-3 model with periodicity input were 18.2% and 17.2%, respectively. From the BIC point of view FOASVR-2, M5-2, and SVR-4 with the values of 597.55, 581.85, and 701.18 had better performance in comparison with other models which means that these scenarios had parsimonious inputs (accurate result with fewer input parameters), respectively. So, for this station input combination (2) was a reasonable choice. Time variation of observed and predicted river flows by the optimal periodic and non-periodic FOASVR, M5 and SVR models are illustrated in Fig. 5 and 6. It can be comprehended from the figures that all three periodic and non-periodic models considerably underestimate some peak flows. It seems that preciseness of these models decreases with increasing flow rate. However, the superior accuracy of FOASVR and M5 to the SVR model can be comprehended from these figures. Comparison of Fig. 5 and 6 visibly indicate that the periodic models better approximates the observed river flows than the non-periodic models. Fig. 9 displayes the scattered diagrams of the observed and predicted monthly river flows by each method. It is noticeably evident from the graphs that the SVR model performs worse than the other two methods especially in the prediction of peak river flows. Comparison of two figures reveals that the estimates of periodic models are more accurate than non-periodic models. Also, this figure indicates that all models (periodic and non-periodic) overestimate some low flows.

The test statistics of the FOASVR, M5 and SVR models for the Vaniar station are also provided in Table 3. Similarly, the encouraging influence of periodicity component on models' precision is clearly seen for this station. For the FOASVR model, R increased from 0.57 (for input combination (1)) to 0.79 (for input combination (2)) and similarly, RMSE and MAE values decreased from 8.78 to 6.58 and from 4.77 to 3.86, respectively. In the case of two previous discharges inputs, by adding periodicity component, R increased from 0.55 to 0.80 and RMSE, and MAE decreased from 8.88 to 6.48 and from 4.97 to 3.75, respectively. Finally, in the three previous discharges inputs case, R increased from 0.55 to 0.81 and RMSE, and MAE values decreased from 8.99 to 6.33 and from 5.53 to 3.71, respectively. Comparison of three models revealed that the optimal FOASVR-6 model whose inputs are Q_{t-1}, Q_{t-2}, Q_{t-3}, and π performed better than optimal M5-2 comprising Q_{t-1}

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 May 2019 doi:10.20944/preprints201905.0320.v1

and π inputs and both performed better than the optimal SVR-6 model whose inputs are same as FOASVR-6. Generally, FOASVR performed better than SVR and M5 models, moreover, accuracy of SVR was increased by 29.7% and 30.4% related to RMSE and MAE in the optimal scenario (FOASVR-6) by applying FOA, respectively; also, FOASVR showed 16.8% and 19.7% better performances than M5 in terms of RMSE and MAE for this scenario, respectively. Without the periodicity component, the optimal FOASVR-1 model performed better than the optimal M5-1 and SVR-3 model. The relative RMSE and MAE differences between the optimal FOASVR-1 model without periodicity and FOASVR-1 model with periodicity input were 25.1% and 19.1%, respectively. The best values for BIC in this station were related to FOASVR-6 with 703.64, M5-2 with 740.34, and SVR-2 with 825.05. According to the fact that FOASVR-6 was closely followed by FOASVR-4 with the value of 707.09 and FOASVR-2 with the value of 707.53, it is better to choose a combination with fewer input parameters. Thus, the input parameters of Q_{t-1} and π were selected as a parsimonious scenario for this station similar to the previous station. Fig. 7 and 8 demonstrate the time variation of observed and predicted river flows by the optimal periodic and non-periodic FOASVR, M5 and SVR models. As found for the Vaniar station, here also the three periodic and non-periodic models underestimate some peak flows. Comparison of Fig. 7 and 8 confirm that appending the periodicity component as the input increases the estimation capacity of the models. The scatterplots of the observed and predicted monthly river flows by each method are shown in Fig. 9. Alike to the previous station, the FOASVR and M5 perform better than the SVR model especially in the prediction of peak river flows. This figure indicates that the estimates of periodic models are more accurate. According to Fig. 9, same as Babarud station, the models overestimate low flows in the Vaniar station, so, forecasting shifts from overestimation to underestimation with increasing flow rate.

Table 3 The evaluation parameters of studied models in the test period

Model input	Model	Babarud Station				Vaniar Station			
		RMSE	MAE	R	BIC	RMSE	MAE	R	BIC
Q_{t-1}	SVR-1	6.10	4.10	0.59	706.88	9.33	5.60	0.50	831.52
	M5-1	5.94	3.62	0.61	696.57	9.57	5.44	0.54	840.91
	FOASVR-1	5.74	3.29	0.63	683.28	8.78	4.77	0.57	809.04
Q_{t-1}, π	SVR-2	5.97	3.88	0.61	703.79	9.04	5.31	0.52	825.05
	M5-2	4.54	2.73	0.80	597.55	7.19	4.46	0.77	740.34
	FOASVR-2	4.36	2.40	0.82	581.85	6.58	3.86	0.79	707.53
Q_{t-1}, Q_{t-2}	SVR -3	5.98	4.04	0.62	704.44	9.21	5.57	0.53	831.95
	M5-3	5.79	3.49	0.68	691.92	9.80	5.46	0.59	854.92

 doi:10.20944/preprints201905.0320.v1

	FOASVR-3	5.33	2.90	0.70	659.80	8.88	4.97	0.55	818.45
Q_{t-1}, Q_{t-2}, π	SVR -4	5.85	3.83	0.64	701.18	8.96	5.33	0.54	826.99
	M5-4	4.55	2.83	0.80	603.67	7.58	4.64	0.76	765.10
	FOASVR-4	4.50	2.67	0.80	599.39	6.48	3.75	0.80	707.09
Q_{t-1}, Q_{t-2},Q_{t-3}	SVR -5	5.91	3.90	0.62	705.14	9.22	5.73	0.52	837.57
	M5-5	5.79	3.50	0.68	697.18	9.79	5.55	0.60	859.76
	FOASVR-5	5.69	3.20	0.67	690.42	8.99	5.53	0.55	828.22
Q_{t-1}, Q_{t-2},Q_{t-3}, π	SVR -6	5.82	3.77	0.64	704.46	9.01	5.53	0.53	834.27
	M5-6	4.54	2.84	0.80	608.09	7.61	4.62	0.75	771.78
	FOASVR-6	4.58	2.67	0.79	611.49	6.33	3.71	0.81	703.64

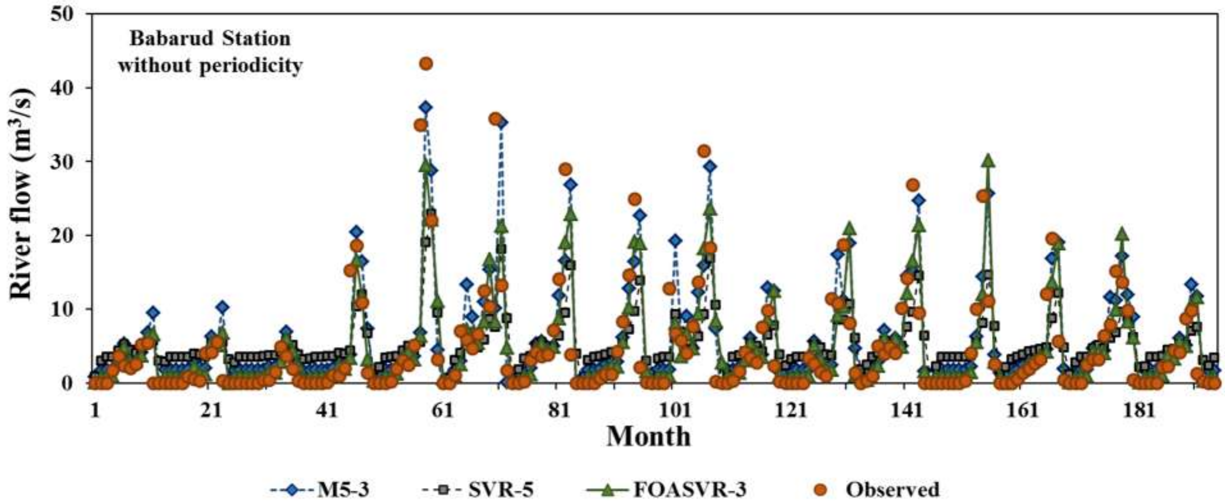


Fig. 5 The observed and forecasted monthly river flows without periodicity for Babarud station

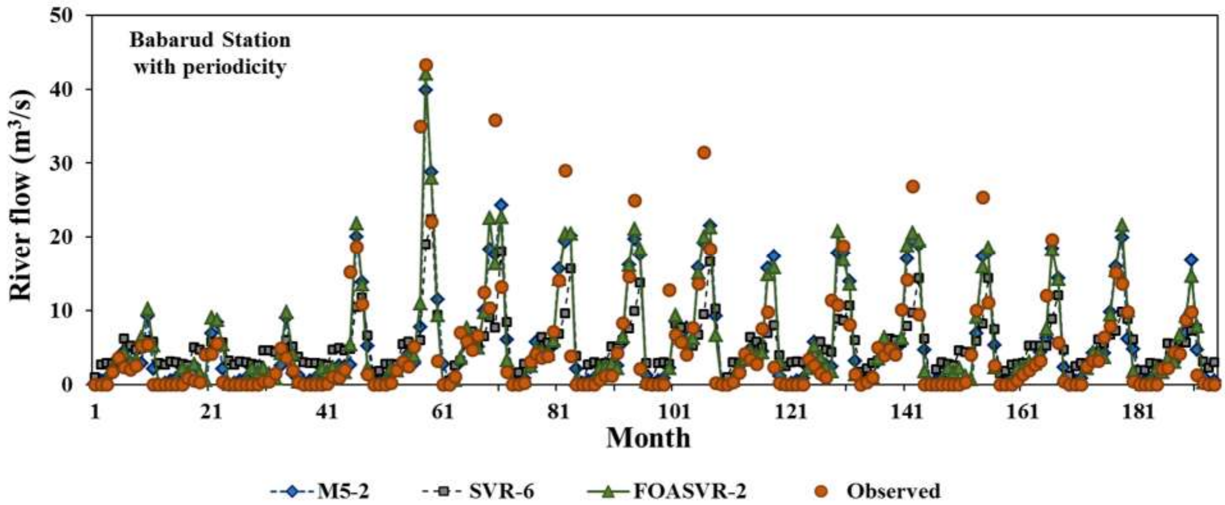


Fig. 6 The observed and forecasted monthly river flows with periodicity for Babarud station

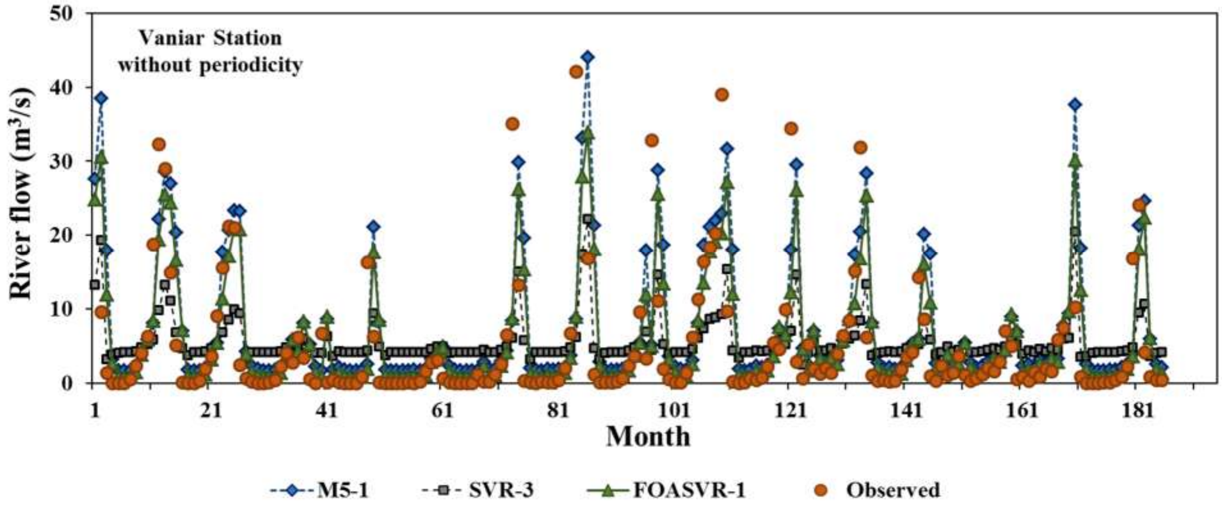


Fig. 7 The observed and forecasted monthly river flows without periodicity for Vaniar station

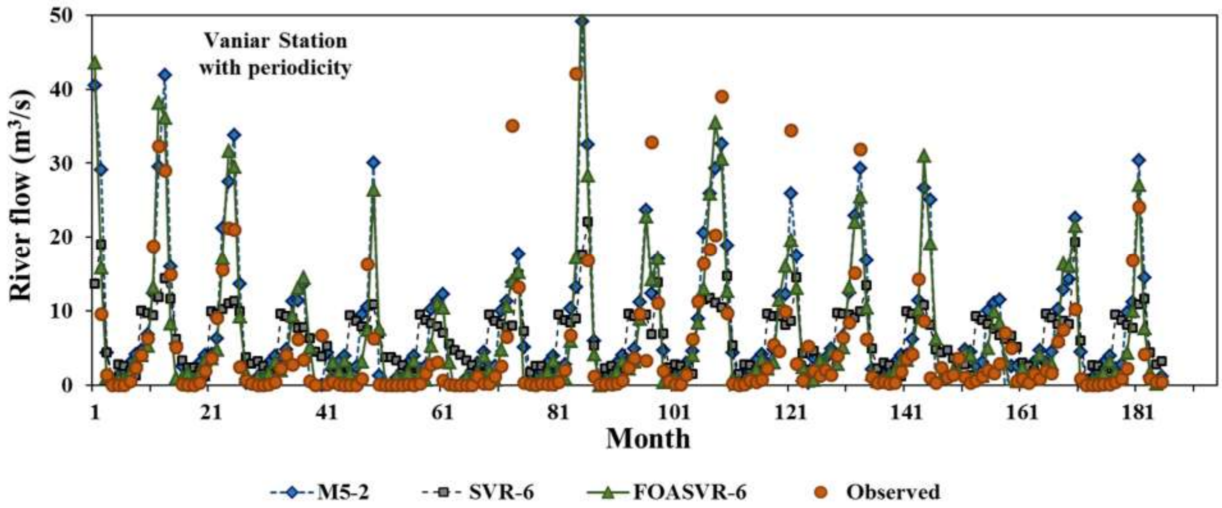
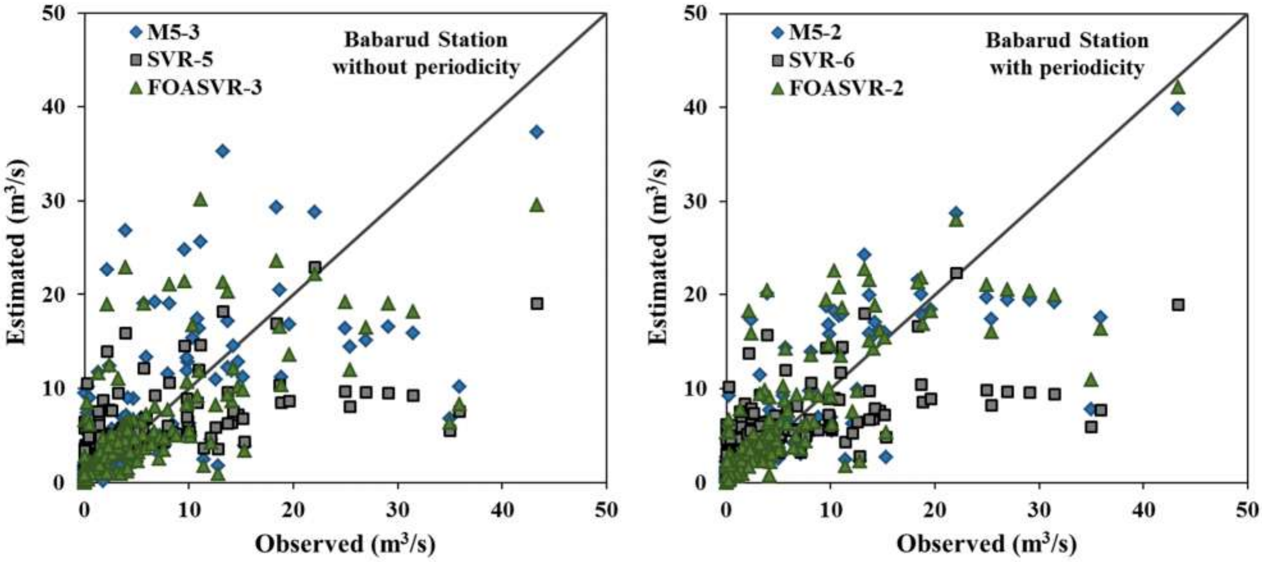


Fig. 8 The observed and forecasted monthly river flows with periodicity for Vaniar station



 doi:10.20944/preprints201905.0320.v1

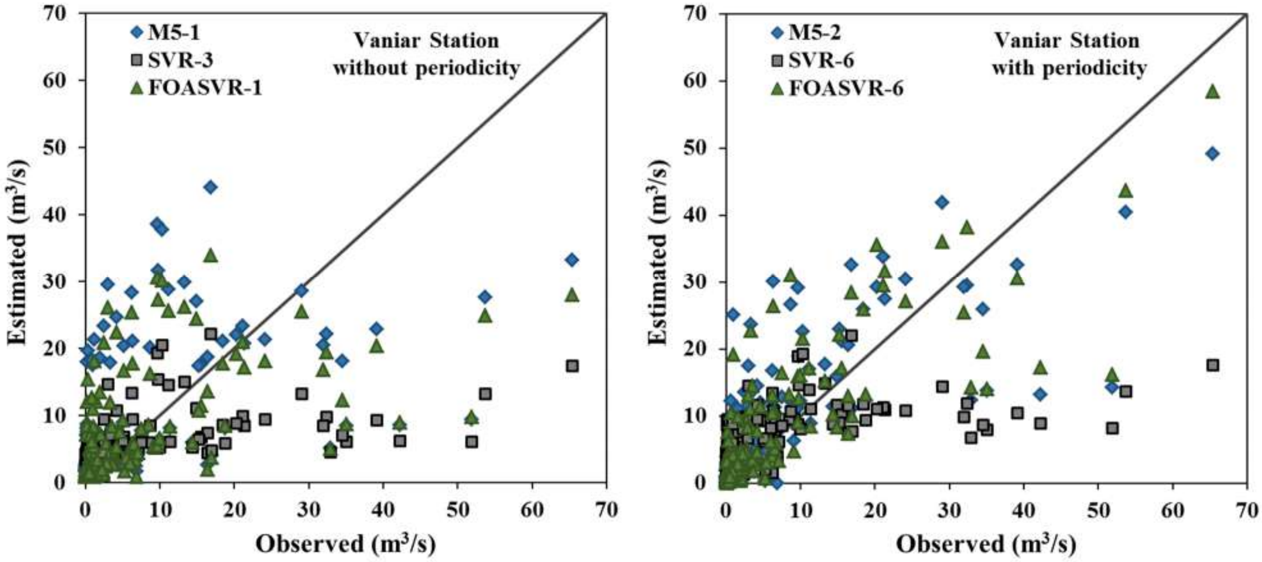


Fig. 9 The scatterplots of observed and forecasted monthly river flows with and without periodicity for both stations

Furthermore, TDs were utilized for examining SD and R values for the FOASVR, M5 and SVR models. Fig. 10 exhibits TDs for all models, where the space from the reference green point is an amount of the centered RMSE. So, it can be comprehended from Fig. 10 that FOASVR (a point with yellow color) provided relatively precise predictions of river flow in both stations.

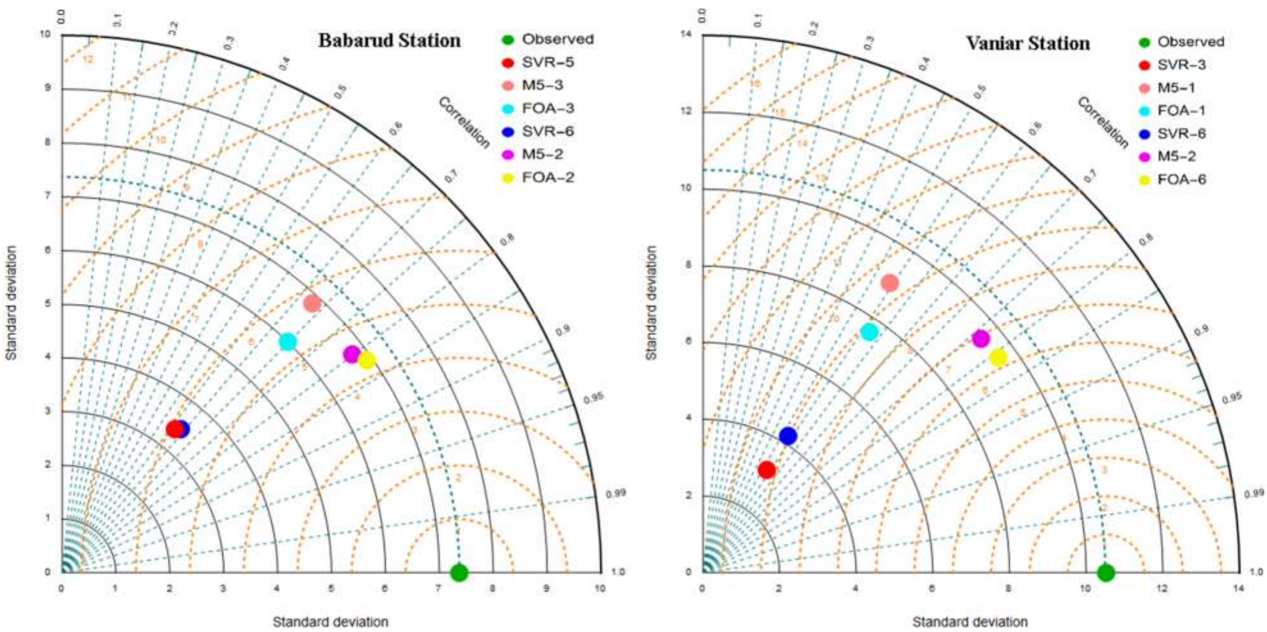


Fig. 10. TDs of the monthly predicted river flow

 doi:10.20944/preprints201905.0320.v1

6. Conclusion

In the current study, three different data-driven techniques, FOASVR, M5, and SVR were compared in one month ahead river flow forecasting in two stations, located in the Lake Urmia Basin of Iran. Comparison of three periodic models specified that the periodic FOASVR model had better accuracy than the periodic M5 and SVR models. M5 was also found to achieve more suitable results than the SVR model. Similar to periodic models, comparison of non-periodic models showed that the optimal FOASVR also had a better performance than M5 and SVR models. It was proved that appending periodicity component significantly increases models' accuracy in forecasting monthly river flows for both stations. For the Babarud station, the relative RMSE and MAE differences between the optimal periodic and non-periodic FOASVR models were found to be 18.2% and 17.2%, respectively. For the Vaniar station, the periodicity component decreased the RMSE and MAE values of the optimal FOASVR models by 27.9 and 22.2%, respectively. According to BIC, the second input combination (Q_{t-1} and π) opted as parsimonious inputs for FOASVR with values of 581.85 and 707.53 for Babarud and Vaniar stations, respectively. Generally, the performance of FOASVR models was better than the other two methods in forecasting monthly river flows. However, all methods indicated some difficulties in forecasting river flow peaks while the FOASVR models provided a better forecast in high river flows.

References

Bhattacharya, B., Solomatine, D.P., 2005. Neural networks and M5 model trees in modeling water level–discharge relationship. Neurocomputing. 63, 381–396.

Bhattacharya, B., Solomatine, D.P., 2006. Machine learning in sedimentation modeling. Neural Networks. 19, 208–214.

Burnham, K.P., Anderson, D.R., 2002. Model Selection and Inference: a Practical Information-theoretic approach, second edition. Springer-Verlag, New York.

Cao, G., Wu, L., 2016. Support vector regression with fruit fly optimization algorithm for seasonal electricity consumption forecasting. Energy 115, 734–745. doi:10.1016/j.energy.2016.09.065

Chang, C.-C., Lin, C.-J., 2001. Training v-support vector classifiers: theory and algorithms.Neural Comput. 13(9), 2119–47.

Cimen, M., 2008. Estimation of daily suspended sediments using support vector machines. Hydrol. Sci. J. 53 (3), 656–666.

Esmaeilzadeh, B., Sattari, M.T., Samadianfard, S., 2017. Performance evaluation of ANNs and an M5 model tree in Sattarkhan Reservoir inflow prediction. ISH Journal of Hydraulic Engineering. 1-10.

Fernando, A.K., Shamseldin, A.Y., Abrahart, B.J., 2012. River Flow Forecasting Using Gene Expression Programming Models. 10th International Conference on Hydroinformatics HIC 2012. Hamburg, Germany.

Gunn, S.R., 1998. Support Vector Machines for Classification and Regression, Technical Report. University of Southampton, England.

Ismail S., Samsudin, R., Shabri, A., 2010. River Flow Forecasting: a Hybrid Model of Self Organizing Maps and Least Square Support Vector Machine. Hydrol. Earth Syst. Sci. Discuss. 7, 8179–8212.

Kalteh, A.M., 2013. Monthly river flow forecasting using artificial neural network and support vector regression models coupled with wavelet transform. Comput. Geosci. 54, 1–8. doi:10.1016/j.cageo.2012.11.015

Khan, A.S., See, L., 2006. Rainfall-Runoff modeling using data-driven and statistical methods. Institute of Electrical and Electronics Engineers (IEEE).

Lijuan, W., Guohua, C., 2016. Seasonal SVR with FOA algorithm for single-step and multi-step ahead forecasting in monthly inbound tourist flow. Knowledge-Based Syst. 110, 157–166. doi:10.1016/j.knosys.2016.07.023

Liong, S.Y., Sivapragasam, C., 2002. Flood stage forecasting with support vector machines. J. AWRA. 38 (1), 173–186.

Londhe, S.N., Dixit, P.R., 2011. Forecasting Stream Flow Using Model Trees. International Journal of Earth Sciences and Engineering. 4(6), 282-285.

Londhe, S., Gavraskar, S.S., 2015. Forecasting One Day Ahead Stream Flow Using Support Vector Regression. Aquat. Procedia 4, 900–907. doi:10.1016/j.aqpro.2015.02.113

Pan, W.-T., 2012. a new Fruit Fly Optimization Algorithm: Taking the financial distress model as an example. Knowledge-Based Syst. 26, 69–74. doi:10.1016/j.knosys.2011.07.001

Onyari, E., Ilunga, F., 2010. Application of MLP neural network and M5P model tree in predicting streamflow: A case study of Luvuvhu catchment, South Africa. International Conference on Information and Multimedia Technology (ICMT). Hong Kong, China. V3, 156-160.

Quinlan, J.R., 1992. Learning with continuous classes. Proceedings Fifth Australian Joint Conf. on Artificial Intelligence (ed. by A. Adams & L. Sterling). World Scientific, Singapore. 343–348.

Samadianfard, S., Nazemi, A.H., Sadraddini, A.A., 2014a. M5 model tree and gene expression programming based modeling of sandy soil water movement under surface drip irrigation. Agriculture Science Developments. 3, 178-190.

Samadianfard, S., Sattari, M.T., Kisi, O., Kazemi, H., 2014b. Determining flow friction factor in irrigation pipes using data mining and artificial intelligence approaches. Applied Artificial Intelligence. 28, 793-813.

Sattari, M.T., Pal, M., Apaydin, H., Ozturk, F., 2013. M5 Model Tree Application in Daily River Flow Forecasting in Sohu Stream, Turkey. Water Resources. 40(3), 233-242.

Siek, M., Solomatine, D.P., 2007. Tree-like machine learning models in hydrologic forecasting: optimality and expert knowledge. Geophysical Research Abstracts. Vol. 9.

Smola, A.J., Scholkopf, B., 2004. A tutorial on support vector regression. Statistics and Computing. 14(3), 199-222.

Stravs, L., Brilly, M., 2007. Development of a low flow forecasting model using the M5 machine learning method. Hydrological Sciences. 52(3), 466–477.

Taylor, K.E., 2001. Summarizing multiple aspects of model performance in a single diagram. J. Geophys. Res. Atmos. 106, 7183–7192.

<URL1>https://earth.google.com/web/@32.205151,53.07029487,2852.42968574a,2667368.97567809d,35y,0.1175398 4h,16.72644158t,-0r

Vapnik, V., 1995. The Nature of Statistical Learning Theory. Springer Verlag, New York, USA.

Witten, I.H., Frank, E., 2005. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann: San Francisco.

Wu, C.L., Chau, K.W., Li, Y.S., 2008. River stage prediction based on a distributed support vector regression. J. Hydrol. 358, 96–111. doi:10.1016/j.jhydrol.2008.05.028

Wu, C.-H., Tzeng, G.-H., Lin, R.-H., 2009. A Novel hybrid genetic algorithm for kernel function and parameter optimization in support vector regression. Expert Syst. Appl. 36, 4725–4735. doi:10.1016/j.eswa.2008.06.046

Yu, X.Y., Liong, S.Y., Babovic, V., 2004. EC-SVM approach for realtime hydrologic forecasting. J. Hydroinf. 6 (3), 209–233.

Kurup, P.U., Dudani, N.K. Neural networks for profiling stress history of clays from PCPT data. Journal of Geotechnical and Geoenvironmental Engineering 2014; 128(7): 569-579.

Samadianfard, S, Delirhasannia, R, Kisi, O, Agirre-Basurko, E. Comparative analysis of ozone level prediction models using gene expression programming and multiple linear regression. GEOFIZIKA 2013; 30:43-74.

Samadianfard, S, Sattari, MT, Kisi, O, Kazemi, H. Determining flow friction factor in irrigation pipes using data mining and artificial intelligence approaches. Applied Artificial Intelligence 2014; 28: 793-813.

Samadianfard, S., Asadi, E., Jarhan, S., Kazemi, H., Kheshtgar, S., Kisi, O., Sajjadi, S., Abdul Manaf, A., 2018. Wavelet neural networks and gene expression programming models to predict short-term soil temperature at different depths, Soil and Tillage Research, 175: 37-50.

Deo, R.C., Ghorbani, M.A., Samadianfard, S., Maraseni, T., Bilgili, M., & Biazar, M., 2018. Multi-layer perceptron hybrid model integrated with the firefly optimizer algorithm for windspeed prediction of target site using a limited set of neighboring reference station data, Renewable Energy, 116: 309-323.