Applied Water Science (2024) 14:110
https://doi.org/10.1007/s13201-024-02154-x

ORIGINAL ARTICLE

Enhancing spatial streamflow prediction through machine learning algorithms and advanced strategies

Sedigheh Darabi Cheghabaleki[1] · Seyed Ehsan Fatemi[1] · Maryam Hafezparast Mavadat[1]

Received: 20 January 2024 / Accepted: 7 March 2024

Abstract
Forecasting and extending streamflow is a critical aspect of hydrology, especially where the time series are locally unavailable for a variety of reasons. The necessity of preprocessing, model fine-tuning, feature selection, or sampling to enhance prediction outcomes for streamflow forecasting using ML techniques is evaluated in this study. In this regard, the monthly streamflow at Pol-Chehr station is analyzed using various monthly rainfall and streamflow time series data from different stations. The results of streamflow prediction in the k-folds cross-validator approach are generally better than those of the time series approach, except when raw data with no preprocessing or feature selection is used. Applying the simple SVR model to raw data leads to the weakest result, but using the GA-SVR model on raw data significantly increases the Nash coefficient by about 215% and 72%, decreases the NRMSE by about 48% and 36% in the k-fold and time series approaches, even with no feature selection. On the other hand, standardization produces highly accurate model predictions in both the k-fold and time series approaches, with a minimum Nash coefficient of 0.83 and 0.73 during the test period in the simple SVR model, respectively. Finally, using optimization algorithms like GA to fine-tune ML models and feature selection does not always yield improved prediction accuracy, but it depends on whether raw or preprocessed data is chosen. In conclusion, combining k-fold cross-validator and preprocessing typically yields highly accurate predictive results, with an *R* value exceeding 93.7% (Nash = 0.83, SI = 0.55, NRMSE = 0.09), without requiring any additional fine-tuning or optimization. Using feature selection is only significant when utilizing the TS approach as well.

Keywords Time series · SVR model · Preprocessing · K-fold cross-validator · Spatial prediction · GA-SVR

Introduction

Predicting streamflow is crucial for water resource planning and management. However, various relationships and complex patterns have been proposed for predicting river flow, including conceptual rainfall–runoff patterns, time series patterns, and hybrid patterns. Due to a lack of precise understanding and the complexity of factors affecting river flow, these relationships often fail to match observed values (Moeeni et al. 2017a, b). One of the most prevalent analytical techniques for forecasting data is statistical modeling and regression. However, they frequently yield errors due to their linear problem-solving approach, which fails to accurately model the time variations of the phenomenon in question. Hence, selecting a model capable of precisely estimating streamflow based on influential factors is crucial. Currently, machine learning techniques such as support vector machines and genetic programming models are extensively employed to predict nonlinear phenomena. In recent times, the use of intelligent models has garnered considerable interest from researchers aiming to predict river flows with the utmost accuracy (Moeeni et al. 2017a, b).

In recent years, changes in social conditions, climate change, population growth, and improper use of available water resources have been known to be the causes of the decline in available water resources (Pourkheirollah et al., 2023). Therefore, the need for integrated water resources management is clear. One of the most important parameters for the planning and sustainable management of water

✉ Seyed Ehsan Fatemi
se.fatemi@razi.ac.ir

Sedigheh Darabi Cheghabaleki
seddigheh.darabi73@gmail.com

Maryam Hafezparast Mavadat
maryam.hafezparast@gmail.com

[1] Water Engineering Department, Campus of Agriculture and Natural Resources, Razi University, Kermanshah, Iran

Published online: 03 May 2024

resources is streamflow estimation (Parvaz and Shahoei 2022). To manage water resources in the future, accurate and reliable river flow forecasts using intelligent data models are significantly provided (Yin et al. 2018). Modeling hydrological time series using historical records plays an essential role in predicting different hydrological processes (Sahoo et al. 2019). Data-driven models are relatively simple, but they are powerful methods for predicting river flow. The model-free hybrid methods are considered to use the strengths of each (Ebrahimi and Shourian 2022).

Forecasting models, especially support vector machine (SVM), provided outstanding performances in the prediction of various hydrological variables, such as groundwater level prediction (Fatemi and Parvini 2022; Ebrahimi and Rajaee 2017; Gorgani et al. 2017; Fatemi et al. 2018; Soltani et al. 2022 and Soltani et al. 2023), flood prediction (Sahoo et al. 2021; Wu et al. 2019), runoff prediction (Nourmohammadi et al., 2023; Samantaray et al. 2021; Zaini et al. 2018; Moeeni et al. 2017a, b; Bell et al. 2012; Okkan and Serbes 2012), sediment analysis (Samantaray et al. 2020) and rainfall forecasting (Ebtehaj et al. 2020). Recently, very important progress has been made in recognizing the capability of SVR in modeling the rainfall–runoff process. Wu et al. (2019) developed a new model, HGA-SVR, for kernel function type and kernel parameter value optimization in SVR. This model is fitted to search for the optimal kernel function types and their parameters to improve SVR accuracy. The results showed that the new HGA-SVR model performed better than the previous models. Particularly, the new model could successfully obtain the optimal kernel type for their parameters with the lowest prediction error values.

To obtain SVR models that can predict highly accurate set points, a new genetic algorithm method is applied (Sanz-Garcia et al. 2015). This proposal assigned feature selection, model tuning, and parsimonious model selection to accomplish robust SVR models. The results showed that GA-PARSIMONY, in comparison with classical GA, was able to produce more robust SVR models with fewer input features.

A set of 50 data-driven forecasting models such as SVR, Multivariate Adaptive Regression Line, MARS, and M5Tree for predicting river flow data in an ecologically important semi-arid mountainous region in the Pailugou watershed is applied in northwest China. To achieve stable and accurate prediction results, a random sampling of the entire river flow data is considered 80% for training and the rest for testing periods. They show that the M5Tree method can be successfully applied for short-term river flow forecasting in semi-arid mountainous regions, which may have useful implications in water resources management, ecological sustainability, and river systems assessment (Yin et al. 2018).

Baydaroğlu et al. (2018) investigated river flow forecasting using combined models of SVR with wavelet transform, singular spectrum analysis, and a chaos approach for the Kızılırmak River in Turkey. These three methods were successful in generating the input matrix for SVR, while the SVR-WT combination resulted in the highest determination coefficient and the lowest mean absolute error. Sahoo et al. (2019) analyzed the suitability of SVR to model the monthly low-flow time series for three stations in the Mahanadi River basin, India. The accuracy of the SVR model with two different framework models (ANN-ELM and GPR) is evaluated by different statistical criteria such as $r2$, RMSE, MAE, and Nash–Sutcliffe coefficient. To model monthly low flows in the Mahanadi River Basin, India, the results confirm that SVR can be suitably used. They suggest that to predict low flow (discharge Q75), the SVR model can be used as a new accurate data-intelligent approach based on past data on water resources and their dependent catchment.

A robust meta-model for river flow prediction, a feature-based adaptive combiner (FBAC), is introduced that uses features extracted from data. To build this model, some data-driven techniques are used, like Artificial Neural Networks (ANN), Random Forest Regression (RFR), SVR, and a modified Dynamic K-Nearest Neighbor (DKNN). FBAC is applied to two years of daily Azad reservoir inflow in western Iran. The results of FBAC in terms of reducing the RMSE value show a 31.8% and 29.5% improvement in accuracy compared to the best individual and combined models, respectively (Ebrahimi and Shourian 2022). Regardless of access to knowledge-based or data-driven models and various modeling techniques such as human activity and climate change, accurately predicting monthly runoff remains a challenging task. The application of a hybrid SVM-SSA model, support vector machine with Salp Swarming Algorithm, and conventional SVM and ANN models is investigated by Samantaray et al. (2022) for runoff forecasting in the Baitarani River Basin, Odisha, India. Test results indicated that SVM-SSA can be suggested for modeling the difficulty of relations between the rainfall–runoff process and forecasting runoff.

This study proposes a method for prediction of the spatial and temporal patterns of streamflow using the SVR and RF models, with optimized parameters through the GA model. The significance of this approach lies in its analysis of the impact of preprocessing, model fine-tuning, feature selection, and sampling on prediction results. Essentially, it identifies the essential training methods and those that are not necessary.

Materials and methods

Support vector machine and regrossor

The theory of the support vector machine (SVM) technique is comprehensively described by many researchers (Vapnik

1998; Chen and Yu 2007; Noori et al. 2011) which is briefly explained in this paper. SVR, a type of SVM, is a relatively new and improved data-driven model that is based on statistical learning theory (Vapnik, 1995). It is initially applied to solve pattern recognition and classification problems then it's generally used to solve regression problems.

In a regression SVM model, the functional dependence of the dependent variable *y* is estimated on a set of independent variables *x*. It assumes, like other regression problems that the relationship between the independent and dependent variables is given by a deterministic function *f*(*x*) plus the addition of some noise Eqs. (1, 2).

$$f(x) = W^T \cdot \phi(X) + b \tag{1}$$

$$y = f(x) + \text{noise} \tag{2}$$

The noise is also named error tolerance (ε). However, w and b are vectors of coefficients and constant, the regression function parameters, and ∅ the kernel function. Then finding a functional form for *f*(*x*) is a target. By training the SVR model on training data, it could be achieved. In this process, the sequential optimization of an error function is involved. An e-insensitivity loss function is presented for the convex optimization formula as follows:

$$\min \varphi(w;\xi) = \frac{1}{2}w^2 + C\left(\sum_{i=1}^{n} \xi_i\right) \tag{3}$$

Subject to:

$$y_i\left(w^T x_i + b\right) \geq 1 - \xi_i;\ \ \xi \geq 0;\ \ i = 1.\ldots.N \tag{4}$$

where ξ is a slack variable that penalizes training error by the loss function for the chosen error tolerance. *C* is a positive regularization parameter that shrinks the weight parameters while minimizing the empirical error in the optimization problem (see Fig. 1).

Genetic algorithm

The genetic algorithm (GA) is a powerful method for the heuristic development of large-scale combinatorial optimization problems. It encodes the problem as a set of strings that contain tiny particles; after that, they apply changes to the strings to stimulate the process of gradual evolution. It is a well-known metaheuristic algorithm that draws its inspiration from biological evolution. The GA apes the Darwinian notion of nature's survival of the fittest. J.H. Holland suggested GA in 1992. Chromosome representation, fitness selection, and biologically inspired operators make up the fundamental components of GA. Additionally, Holland developed a unique component known as Inversion, which is typically utilized in GA implementations (Holland 1975; Katoch et al. 2021). In comparison with traditional optimization algorithms, there are many advantages to genetic algorithms, like the ability to deal with complex problems and parallelism. It can also consider various types of objective functions, such as linear or nonlinear, stationary or nonstationary, continuous or discontinuous, or with random noise (Moeeni et al. 2017a, b). In this study, the GA application is used by the scikit-opt (Version 0.6.6) library in the Python programming language.

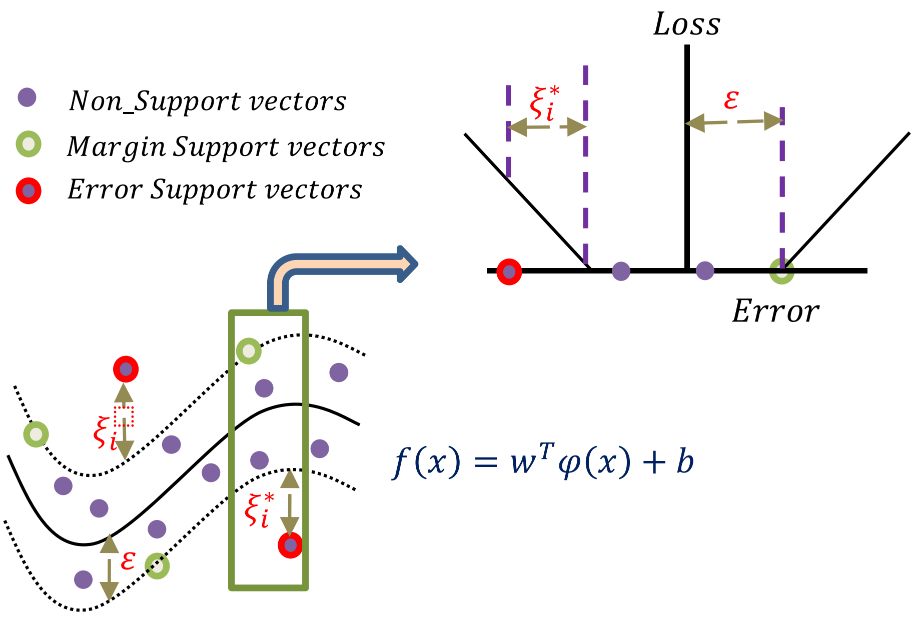


Fig. 1 Nonlinear support vector machine with Vapnik's e-insensitive loss function(Yaseen et. Al, 2016)

Feature selection

Feature selection is the process of reducing the number of variables. In this method, the variables that are most effective in terms of the desired feature are selected based on a series of specific criteria to predict the target variable in the prediction models. As the number of desired features increases, the model's prediction ability also increases. But it is only up to a certain level, so from this specific level onwards, the model would be faced with a problem called the curse of dimensionality. In this case, the performance of the model gets worse and worse, so only those features should be selected that can efficiently predict the desired variable. In this research, the Random Forest algorithm is used because it is a very efficient and simple method.

Random Forest

Random Forest is an ensemble learning method based on decision trees that are commonly used in classification and regression problems. It builds decision trees on different samples and takes their majority choice for classification and the average in the case of regression. It was first introduced by Breiman in 2001. Owing to its simple structure and high performance, it is widely used in many supervised learning applications.

k-fold cross-validator

In data resampling methods, one of the most commonly used algorithms is cross-validation to estimate the prediction model's error and to tune model parameters (Berrar 2019). It is usually used to avoid overfitting when using a supervised machine learning model to consider part of the available data as a test set. In this method, first the data sets are randomly mixed, and then they are divided into k-folds, dividing all the samples into k groups of samples. The prediction model is trained by $k-1$ folds, and the rest fold is used for the test period. In this study, the k-fold cross-validator scikit-learn library in the Python programming language is applied to find the best combination of a dataset in the train and test periods for the prediction model, disregarding the sequence of the data set.

Data preprocessing

Data normalization is one of the common methods in forecasting modeling that is used to better harmonize the data and increase the speed and accuracy of the models. In this method, the data is standardized using the formula (5). where Zi, t is the standardized data in month i of year T, xi, T is the initial data, $\overline{x_T}$ is the average of the data and ST is the standard deviation of the data. It should be noted that after the end of time series forecasting and extension, the model results should be inverted to the original data format. For this purpose, formula (6) is used; $y_{i.T}$ is the inverse function of Zi, t:

$$Z_{i.T} = \frac{\left(x_{i.T} - \overline{x_T}\right)}{S_T};\ y_{i.T} = \left(S_T \times Z_{i.T}\right) + \overline{x_T} \tag{5}$$

Study area

In this study, the time series of five different locations is used. The Pol-Chehr, Hydarabad, and Dooab hydrometric stations, as well as the Hydarabad and Aran climate stations, are located in the Gamasiab sub-basin in Kermanshah province, Iran (Fig. 2). The geographical characteristics of these stations are shown in Table 1.

In this research, to select the model features of Pol-Chehr discharge as dependent variable, the independent variables that have the greatest impact on the forecasting model are considered in Eq. (6):

$$Q_{\text{pol-chehr}} = f\left(Q_{\text{Hydarabad}} \cdot Q_{\text{Dooab}} \cdot P_{\text{Hydarabad}} \cdot P_{\text{Aran}}\right) \tag{6}$$

According to Eq. (7), the discharge flow of Pol-Chehr station is a function of two parameters of monthly rainfall at two stations, Hyderabad, and Doab, and also two monthly discharges at Hyderabad and Aran stations. The schematic of rivers and hydrometric stations located in this basin is shown in Fig. 3:

The discharge duration curve of each hydrometric station for 47 years as a monthly time series is presented in Fig. 4. According to the below diagram, it can be seen that the highest amount of discharge is recorded at Pol-Cheher station, which is almost equal to 260 CMS because this station is located at the end of the water basin. And also, the lowest flow rate recorded in it is equal to zero. Furthermore, in the mean flow of Pol-Cheher, Hyderabad, and Doab stations, the flow rate 50% of the time is equal to 11.52, 4.33, and 5.49 CMS, respectively. In Pol-Cheher station, in 5.5% of the time duration, river flow is equal to zero and the river is dry, but this value has happened less than 2 and completely zero percent of the time in the Hyderabad and Doab stations. For the monthly rainfall analysis in the Hyderabad and Aran stations, the rainfall duration curve is considered as shown in Fig. 5. According to this figure, the amount of rainfall in 50% of the time duration is observed at about 29.1 and 24 mm in

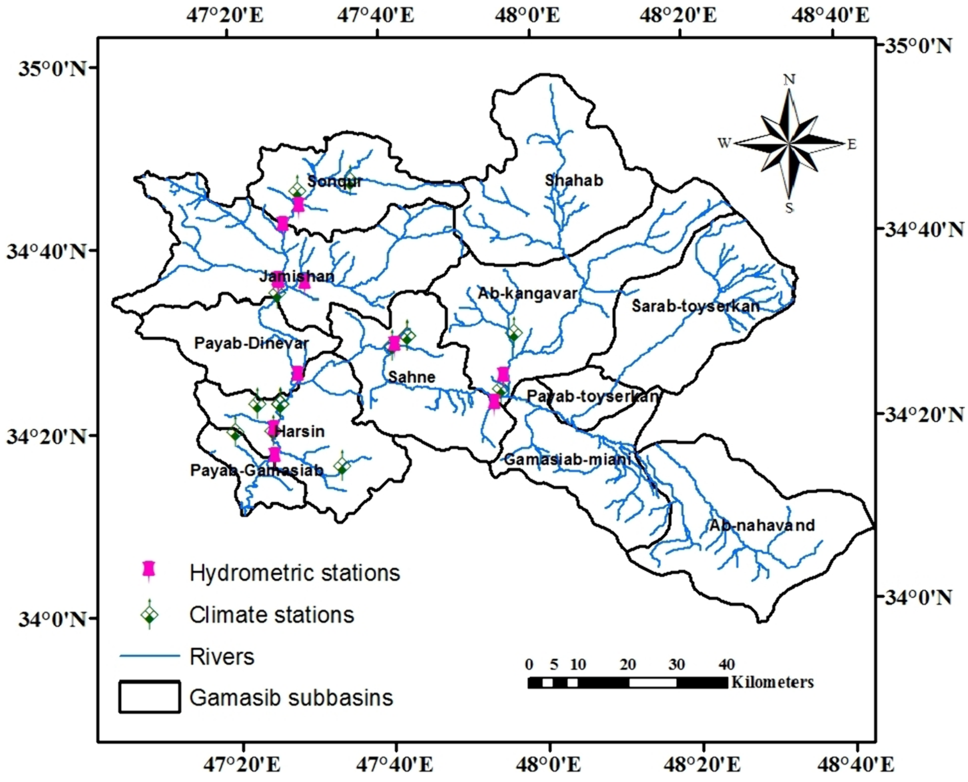


Fig. 2 Case studies location: gamasiab sub-basin

Table 1 Geographical characteristics of the stations

Latitude	Longitude	Site
Hydrometric S		
34° 20′ 11″	47° 25′ 36″	Pol-Chehr
34° 25′ 05″	47° 27′ 10″	Hydarabad
34° 22′ 13″	47° 54′ 04″	Dooab
Climate S		
34° 24′ 00”	47° 27′ 00″	Hydarabad
34° 24′ 52”	47° 55′ 15″	Aran

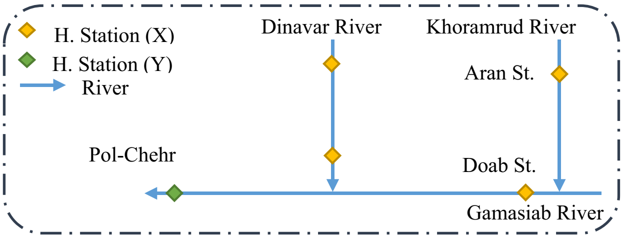


Fig. 3 The schematic of rivers and hydrometric stations

Hyderabad and Aran stations, respectively. Also, there is no rain in about 31 and 29 percent of the time duration for these two climate stations.

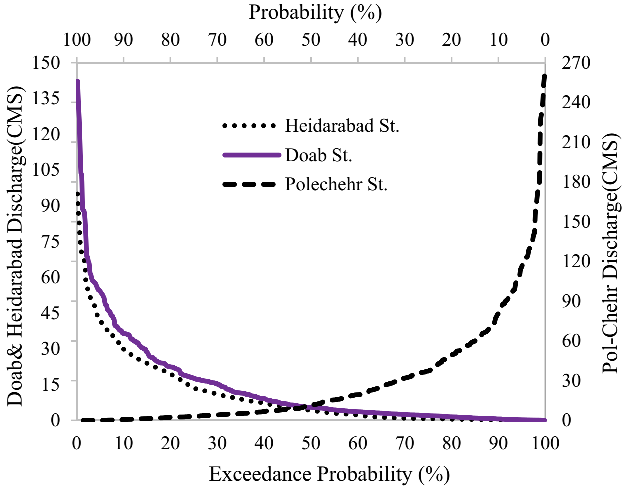
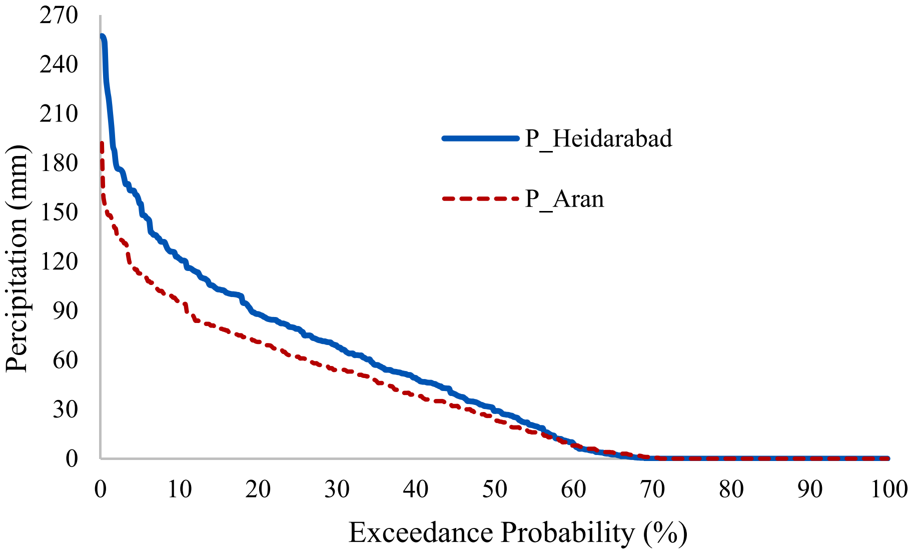


Fig. 4 The discharge duration curve of hydrometric stations

Introducing scenarios and cases

According to Fig. 3, different scenarios are defined as various parameters in different places used in the GA-SVR model inputs. Sixteen different types of input combinations are considered for the monthly discharge forecasting at Pol-Chehr station as follows:

Fig. 5 The rainfall duration curve of climate stations



1. Time series approach, raw data, all features, simple SVR-RBF; TS_raw_all_SVR
2. Time series approach, preprocessing, all features, simple SVR-RBF; TS_pre_all_SVR
3. Time series approach, raw data, feature selection, simple SVR-RBF; TS_raw_FS_SVR
4. Time series approach, preprocessing, feature selection, simple SVR-RBF; TS_pre_FS_SVR
5. k-fold approach, raw data, all features, simple SVR-RBF; K-fold_raw_all_SVR
6. k-fold approach, preprocessing, all features, simple SVR-RBF; K-fold _pre_all_SVR
7. k-fold approach, raw data, feature selection, simple SVR-RBF; K-fold _raw_FS_SVR
8. k-fold approach, preprocessing, feature selection, simple SVR-RBF; K-fold _pre_FS_SVR
9. Time series approach, raw data, all features, Genetic Algorithm + SVR-RBF; TS_raw_all_GA-SVR
10. Time series approach, preprocessing, all features, Genetic Algorithm + SVR-RBF; TS_pre_all_ GA-SVR
11. Time series approach, raw data, feature selection, Genetic Algorithm + SVR-RBF; TS_raw_FS_ GA-SVR
12. Time series approach, preprocessing, feature selection, Genetic Algorithm + SVR-RBF; TS_pre_FS_ GA-SVR
13. k-fold approach, raw data, all features, Genetic Algorithm + SVR-RBF; K-fold_raw_all_ GA-SVR
14. k-fold approach, preprocessing, all features, Genetic Algorithm + SVR-RBF; K-fold _pre_all_ GA-SVR
15. k-fold approach, raw data, feature selection, Genetic Algorithm + SVR-RBF; K-fold _raw_FS_ GA-SVR
16. k-fold approach, preprocessing, feature selection, Genetic Algorithm + SVR-RBF; K-fold _pre_FS_ GA-SVR

Evaluating model accuracy

For forecasting the monthly discharge at Pol-Chehr station, the performance of the SVR-GA model is evaluated by some statistical indices. Thus, the capability of the SVR-GA model in different scenarios is evaluated in terms of correlation coefficient (R), Nash–Sutcliffe (NSE), scatter index (SI), and normalized root mean square error (NRMSE), which are defined as follows:

$$R = \frac{\sum_{i=1}^{n} \left(xi - \bar{x}\right)\left(yi - \bar{y}\right)}{\sqrt{\sum_{i=1}^{n} \left(xi - \bar{x}\right)^2 \sum_{i=1}^{n} \left(yi - \bar{y}\right)^2}} \tag{7}$$

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{n} \left(x_i - y_i\right)^2}{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2} \tag{8}$$

$$\mathrm{SI} = \frac{\mathrm{RMSE}}{\bar{x}} \tag{9}$$

$$\mathrm{NRMSE} = \frac{\sqrt{\frac{\sum_{i=1}^{n} (xi-yi)^2}{n}}}{x_{\max} - x_{\min}} \tag{10}$$

Here xi and $\bar{x}$ represent the observed values and their mean values; yi and $\bar{y}$ are the predicted values and the mean of predicted values, respectively. For a better understanding of the overall workflow in this study, a block diagram and a flowchart on the algorithm is depicted in Figs. 6 and 7.

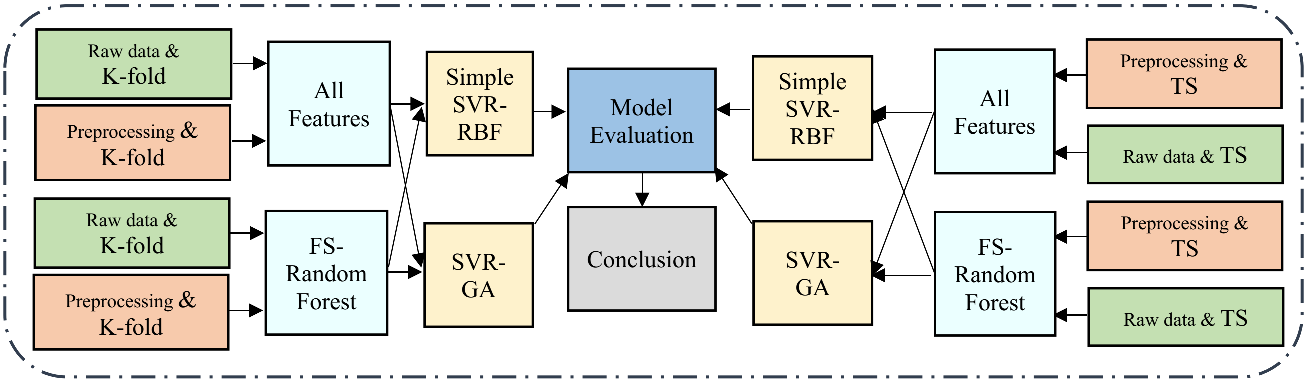


Fig. 6 A block diagram of the overall workflow

Results and discussion

In this research, to answer this question, do the AI models usually need preprocessing, model fine-tuning by an optimization algorithm, feature selection methods, or sampling for model training to improve the prediction results? In this regard, two scenarios of using time series or considering data with a k-fold cross-validation methodology are defined. Then, two approaches were considered for predicting monthly streamflow: using raw data or applying preprocessing to it. Furthermore, in each approach, two methods of model input selection have been investigated in the prediction model based on the feature selection method or including all inputs. Finally, in all cases, a simple RBF and an optimized SVR model with a genetic algorithm have been used for the forecasting of the monthly streamflow. In all cases, 80% and 20% of the data are considered for model training and testing, respectively. All the coding has been done on the Python programming language platform, especially by Numpy, Scipy, Matplotlib, Pandas, Sklearn, and Scikit-opt libraries. In addition, the simple SVM model is set to $C=1$, Epsilon = 0.01, and the RBF kernel. The max_depth in Random Forest is also considered equal to 30.

Table 2 shows the results of monthly discharge modeling for the combinations presented in cases 01–16. The comparison of indices from cases 01–08, considering the k-fold approach, demonstrates that the performance of the models for all cases is better than the time series approach, except for cases 03 and 04 in the test period. In these cases, if the raw data with no preprocessing and also feature selection are simultaneously considered, the results will lead to inferior, case03 ($R=0.77$, Nash = 0.44, SI = 0.99, NRMSE = 0.16) and case04 ($R=0.92$, Nash = 0.78, SI = 0.63, NRMSE = 0.10), rather than the same cases in the time series approach, cases 11($R=0.87$, Nash = 0.71, SI = 1.02, NRMSE = 0.05) and case 12 ($R=0.90$, Nash = 0.79, SI = 0.86, NRMSE = 0.04), regardless of which simple or SVR-GA models are applied in the test period.

If the k-fold approach is used on the raw data with no preprocessing or feature selection, the results of the simple SVR model would be the weakest among all cases ($R=0.76$, Nash = 0.25, SI = 1.14, NRMSE = 0.18) in the test period. By adding feature selection, the results have gradually improved, but the Nash coefficient and NRMSE are still below 0.5 and reduced by 11%, respectively ($R=0.77$, Nash = 0.44, SI = 0.99, NRMSE = 0.16). By applying the GA-SVR to the raw data, the indices would be significantly improved ($R=0.93$, Nash = 0.80, SI = 0.59, NRMSE = 0.09) so that the Nash coefficient increases more than three times even if feature selection is not used. In other words, unlike the simple SVR model, the use of feature selection in the GA-SVR model ($R=0.92$, Nash = 0.78, SI = 0.63, NRMSE = 0.10) will not have a significant effect on the results of the model. When using preprocessing and standardization in the k-fold approach, the model prediction results are very accurate for all cases, regardless of whether a simple SVR or GA-SVR model, feature selection, or all features are used. So that the minimum R value is more than 0.93 (the Nash coefficient is obtained between 0.83 and 0.96) in the model test period. Fatemi and Parvini (2022) show that using preprocessing, in particular, standardization on time series with a sinusoidal form of the ACF diagram always leads to improved forecasting model accuracy, and this property is more effective than using the model tuning. The Pole-Chehr ACF diagram is in sinusoidal form and is depicted in Fig. 8.

In the time series approach, applying simple SVR to the raw data achieved the weakest performance in the test period ($R=0.71$, Nash = 0.45, SI = 1.41, NRMSE = 0.07) between all cases, like the k-fold approach, but the minimum Nash coefficient started at 0.45, about 80% more than the k-fold approach. By adding GA-SVR or feature selection in this case, the model performance is considerably improved by reducing NRMSE by 28% ($R=0.89$, Nash = 0.77, SI = 0.90, NRMSE = 0.05) or ($R=0.87$, Nash = 0.71, SI = 1.02,

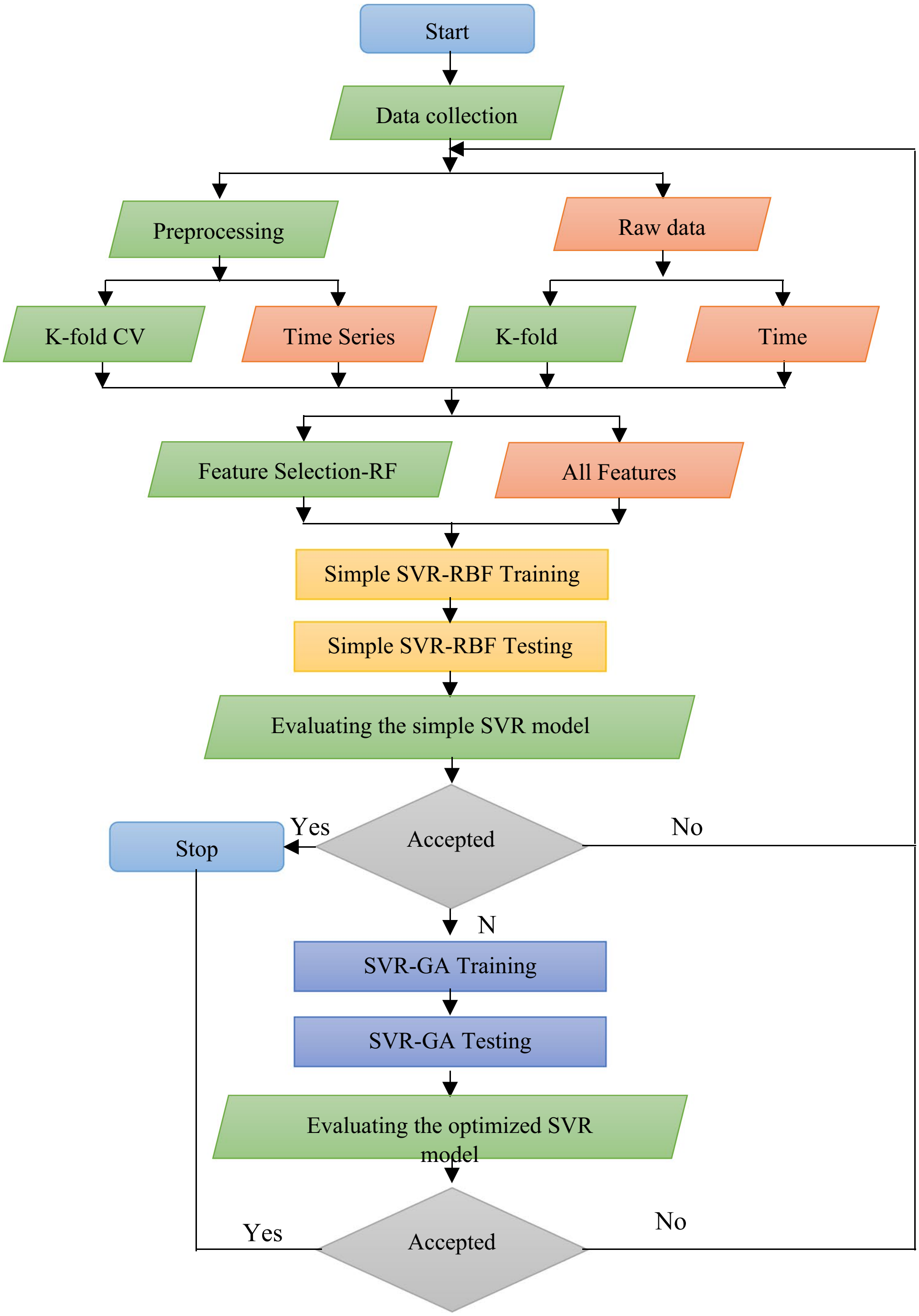


Fig. 7 A flowchart on the algorithm

NRMSE = 0.05). Applying feature selection and GA-SVR is simultaneously inefficient on the results, with less than 5% improvement in the Nash coefficient in this case.

Whenever preprocessing is added to the TS, like the k-fold approach, the results are significantly boosted for all cases, but the range of the Nash coefficient and NRMSE are changed from 0.73 to 0.79 and 0.045–0.05, respectively. It shows that applying the FS or GA-SVR is not necessary in this case. Finally, between the considered cases in the k-fold approach, the best results of the model are calculated in case 08, using preprocessing, adding FS, and applying GA-SVR ($R = 0.98$, Nash = 0.96, SI = 0.26, NRMSE = 0.04).

The Taylor diagram and the results of different models are depicted in Figs. 9 and 10, respectively, for the k-fold, cases 01–08, and the TS approach, cases 09–16, in the test period. As it can be seen from these Figs., the best cases in the k-fold and TS approaches are cases 08 and 12, respectively. For more analysis of the model prediction in low and peak flow, the box plot of all cases is depicted in Fig. 11. As the maximum discharge of Pol-Chehr is 259.9 cms, it is

Table 2 Results of statistical indicators for SVR models with different approaches

Case	Approach	Pre- process	Feature Sel	SVR model	Epsilon	*C*	Train				Test			
							R	Nash	SI	NRMSE	*R*	Nash	SI	NRMSE
01	k-fold	No	No	Simple	0.1	1	0.783	0.423	1.080	0.109	0.762	0.254	1.143	0.182
02	k-fold	No	No	SVR -GA	0.053	19.43	0.955	0.894	0.463	0.047	0.932	0.799	0.594	0.095
03	k-fold	No	Yes [1, 2]	Simple	0.1	1	0.830	0.610	0.888	0.090	0.770	0.441	0.990	0.158
04	k-fold	No	Yes [1, 2]	SVR -GA	0.018	17.37	0.953	0.897	0.457	0.046	0.915	0.776	0.627	0.100
05	k-fold	Yes	No	Simple	0.1	1	0.970	0.931	0.374	0.038	0.937	0.825	0.554	0.088
06	k-fold	Yes	No	SVR -GA	0.002	16.27	*0.996*	*0.991*	*0.135*	*0.014*	0.970	0.922	0.369	0.059
07	k-fold	Yes	Yes [1, 2]	Simple	0.1	1	0.977	0.949	0.321	0.033	0.962	0.884	0.451	0.072
08	k-fold	Yes	Yes [1, 2]	SVR -GA	0.013	3.99	0.992	0.985	0.176	0.018	*0.981*	*0.961*	*0.263*	*0.042*
09	TS	No	No	Simple	0.1	1	0.797	0.433	0.987	0.125	0.712	0.449	1.406	0.072
10	TS	No	No	SVR -GA	0.084	18.90	0.970	0.927	0.354	0.045	0.893	0.774	0.901	0.046
11	TS	No	Yes [1, 2]	Simple	0.1	1	0.830	0.612	0.817	0.103	0.869	0.708	1.023	0.053
12	TS	No	Yes [1, 2]	SVR -GA	0.047	19.38	0.980	0.953	0.284	0.036	0.897	**0.792**	**0.863**	**0.044**
13	TS	Yes	No	Simple	0.1	1	0.988	0.972	0.218	0.027	0.876	0.733	0.979	0.050
14	TS	Yes	No	SVR -GA	0.032	1.08	0.989	0.976	0.201	0.025	0.872	0.738	0.970	0.050
15	TS	Yes	Yes [1, 2]	Simple	0.1	1	0.992	0.981	0.179	0.023	**0.910**	0.788	0.871	0.045
16	TS	Yes	Yes [1, 2]	SVR -GA	0.052	4.10	**0.994**	**0.989**	**0.140**	**0.018**	0.897	0.783	0.883	0.045

Bold values indicate the best values of indices in the train and test for all cases

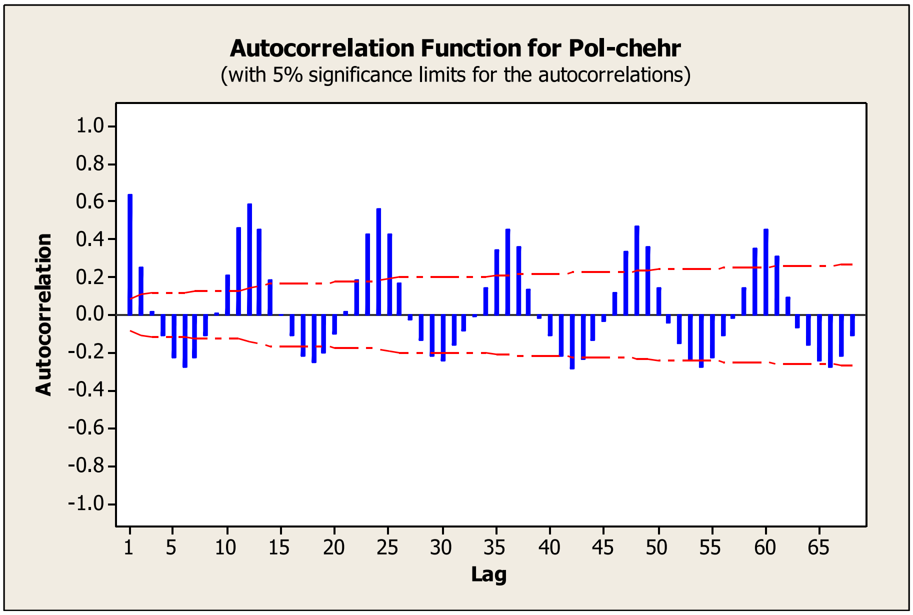
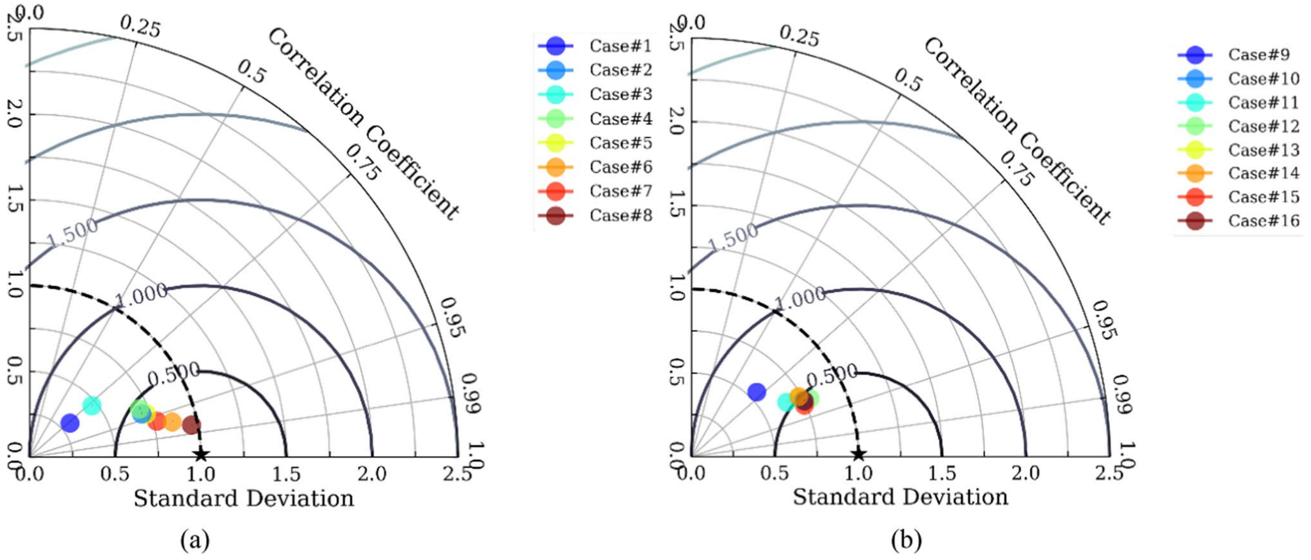


Fig. 8 The ACF diagram: pol-chehr discharge time series

calculated by a model to be more than 200 cms in cases 06 and 08, and especially in case 08, it is more than 250 cms for the k-fold approach, but this is about 100 cms for the best case. It mentions that the k-fold approach predicted the peak flows of Pol-Chehr discharge better than the TS approach, which is also defined in Figs. 10 and 11.

For a better low- and peak-flow analysis of different approaches in the test period, the lowest and highest values of five monthly discharges are considered. Q6, Q7, Q31, Q33, and Q45 in the range of 173–236 cms, and also Q84, Q85, Q96, Q107, and Q109 in the range of 47–174 cms, represent the peak flows in the k-fold and TS approaches, respectively. For the low flows, Q35, Q47, Q53, Q67, and Q86 are in the range of 0–0.42 cms, and also Q50, Q55, Q74, Q89, and Q100 are in the range of 0.01–0.43 cms. The model prediction for the low and peak flows in all cases of two different approaches is shown in Figs. 12 and 13. Based on Fig. 12, the maximum difference of low flows between the observed and model in case 08, the best-selected model for the k-fold approach, is calculated at 0.78 cms and



(a)

(b)

Fig. 9 Taylor diagram for different approaches in the test period: **a** k-fold, **b** TS

occurred in Q67. In case 12, this value of 2.47 cms happened in Q50 for the TS approach. In similar terms, the maximum difference is 91.8 and 93.2 cms for cases 08 and 12, which are in Q31 and Q96, respectively, in Fig. 13.

Conclusion

Streamflow prediction is a crucial parameter for planning and managing water resources sustainably, especially in climate change and improper use of water resources. The study focuses on whether ML models usually need preprocessing, model fine-tuning, feature selection, or sampling to enhance prediction results for streamflow forecasting. In this regard, the monthly streamflow in Pol-Chehr station is determined by the monthly rainfall time series in Hyderabad and Doab stations, as well as the monthly streamflow time series in Hyderabad and Aran stations. The results are shown that the k-fold cross-validator approach generally outperforms the time series approach, except in cases where raw data with no preprocessing (Nash = 0.25, NRMSE = 0.18) and feature selection (Nash = 0.44, NRMSE = 0.16) are used. In these cases, the time series approach yields better results regardless of which simple (Nash = 0.45, NRMSE = 0.07) or SVR-GA (Nash = 0.79, NRMSE = 0.04) model is applied in the test period.

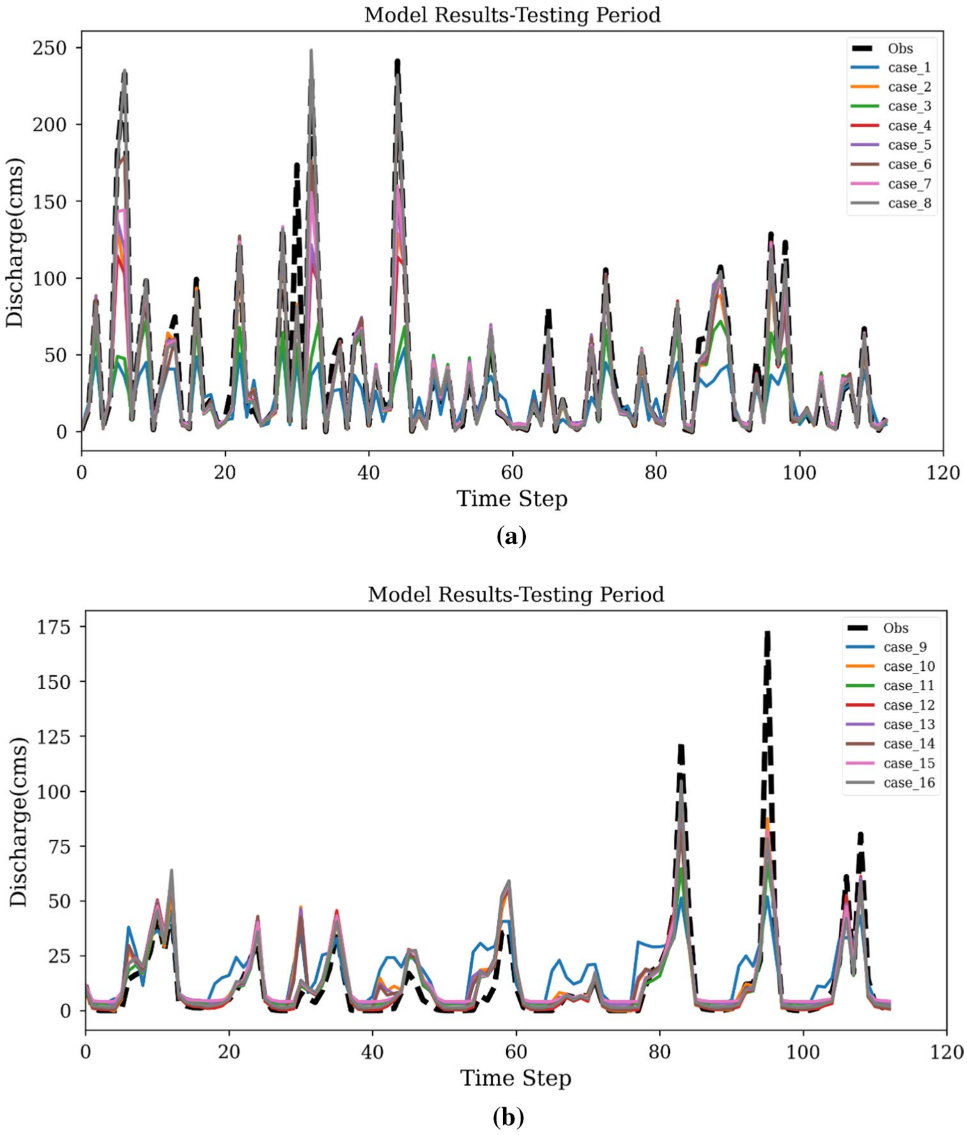
The k-fold scenario, when applied to raw data without preprocessing and feature selection, yields the weakest results for a simple SVR model. However, adding feature selection gradually improves the results although the Nash coefficient and SI remain below 0.5 and 1. On the other hand, using the GA-SVR model on raw data significantly improves the indices, including an increase of more than three times in the Nash coefficient, even without feature selection. Interestingly, feature selection does not have a significant effect on the results of the GA-SVR model.

Preprocessing, particularly standardization, produces highly accurate model predictions in the k-fold approach. This applies to both simple SVR and GA-SVR models, regardless of feature selection or the inclusion of all features. The minimum *R* value is above 0.93, with a Nash coefficient ranging from 0.83 to 0.96 during the model test period.

In the time series scenario, using simple SVR on raw data performed poorly in the test period, like the k-fold approach, with a minimum Nash coefficient of 0.45 and a maximum NRMSE of 0.072. However, adding feature selection significantly improved the model's performance in the Nash coefficient by increasing 58%. Replacing simple SVR with GA-SVR yields significant improvements in the Nash coefficient and NRMSE, the increase is 72% while the decrease is 36% in case.

The addition of preprocessing techniques to the time series scenario, like the k-fold scenario, greatly improves the results in all cases, even in simple SVR. However, it does alter the range of the Nash coefficient. This suggests that using feature selection (FS) or GA-SVR is not required in this particular scenario. The best results for all cases are obtained in case 08 of the k-fold approach, where preprocessing, FS, and GA-SVR are applied, resulting in high values for the model's Nash coefficient of 0.96.

In conclusion, preprocessing techniques significantly enhance the results in k-fold and time series scenarios; the best performance is achieved by combining preprocessing and GA-SVR, and using FS mostly improves the results



(a)

(b)

Fig. 10 a The prediction results of different models for the k-fold approach in the test period. b The prediction results of different models for the TS approach in the test period

in the time series scenario. The graphical conclusion is depicted in Fig. 14.

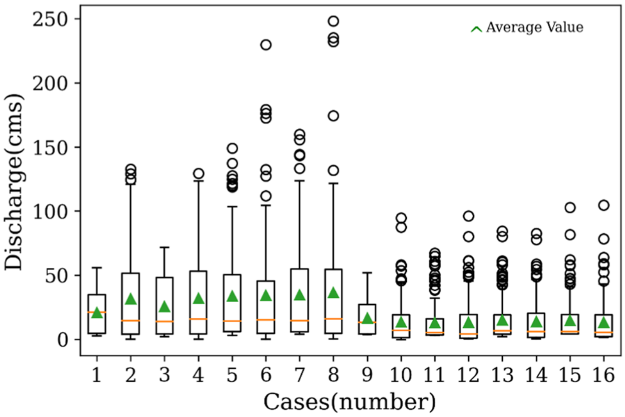


Fig.11 The box plot of different models in the test period

Fig. 12 **a** The model prediction in low flows for all cases of the k-fold approach. **b** The model prediction in low flows for all cases of the TS approach

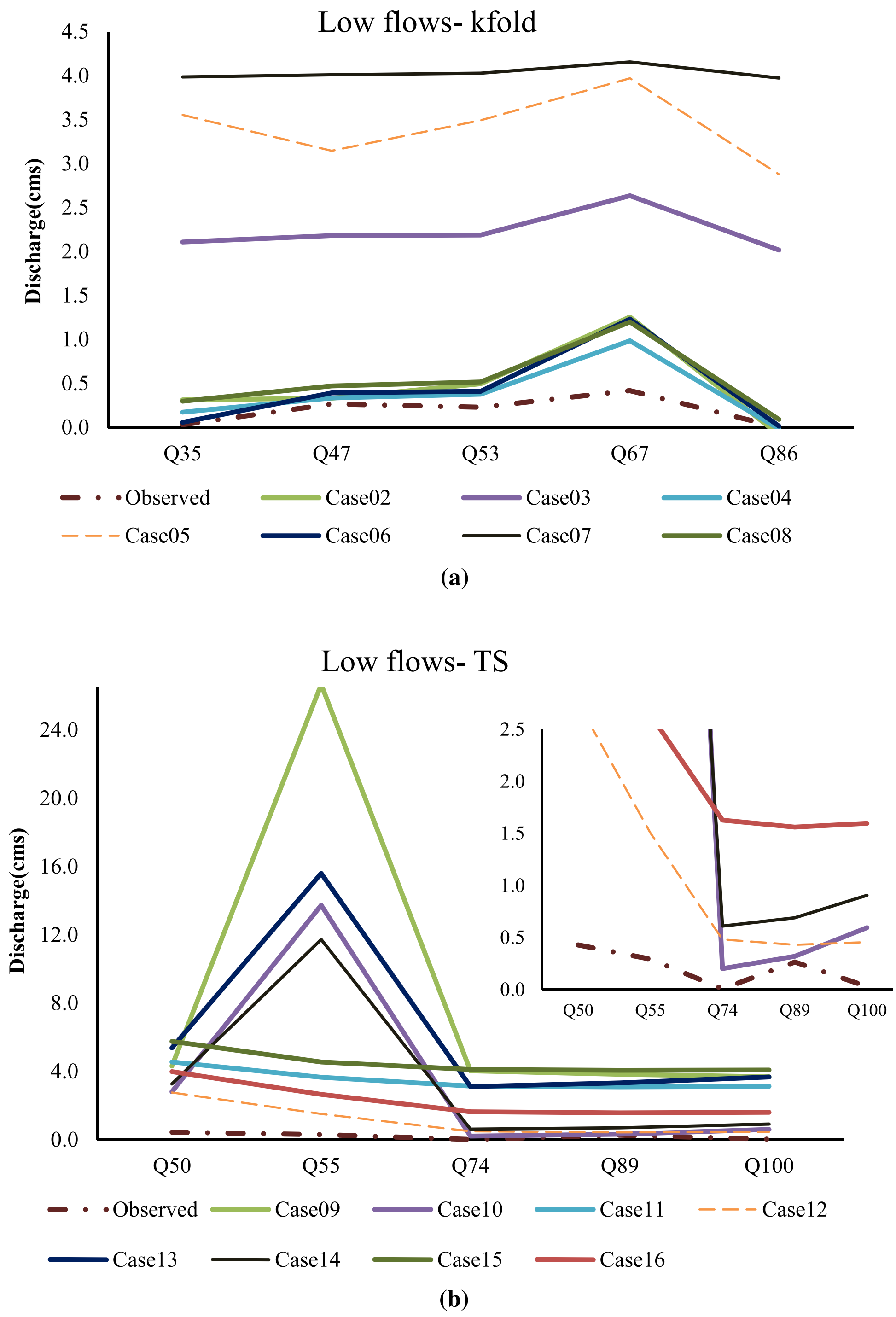
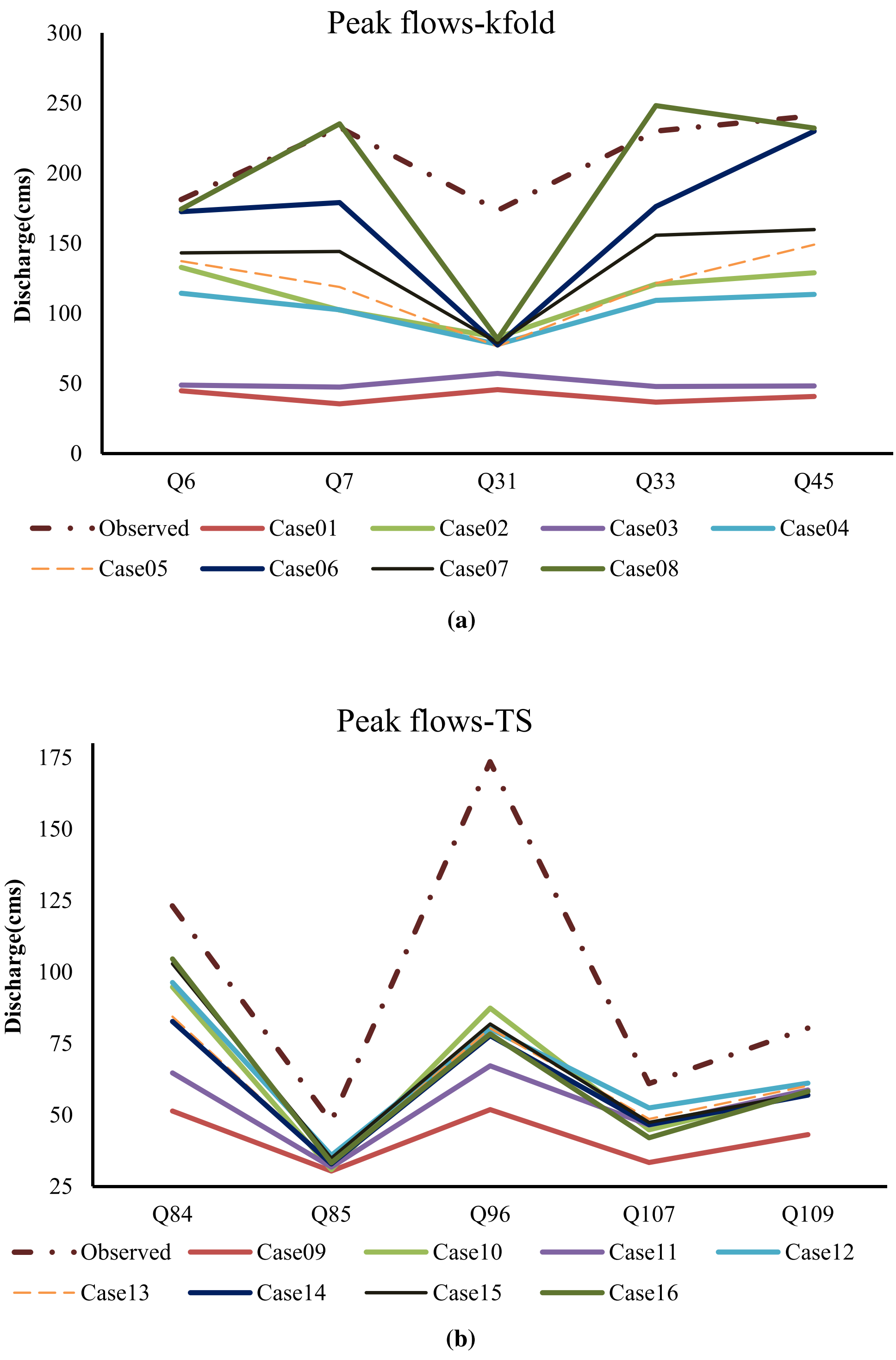


Fig. 13 **a** The model prediction in peak flows for all cases of the k-fold approach. **b** The model prediction in peak flows for all cases of the TS approach



(a)

(b)

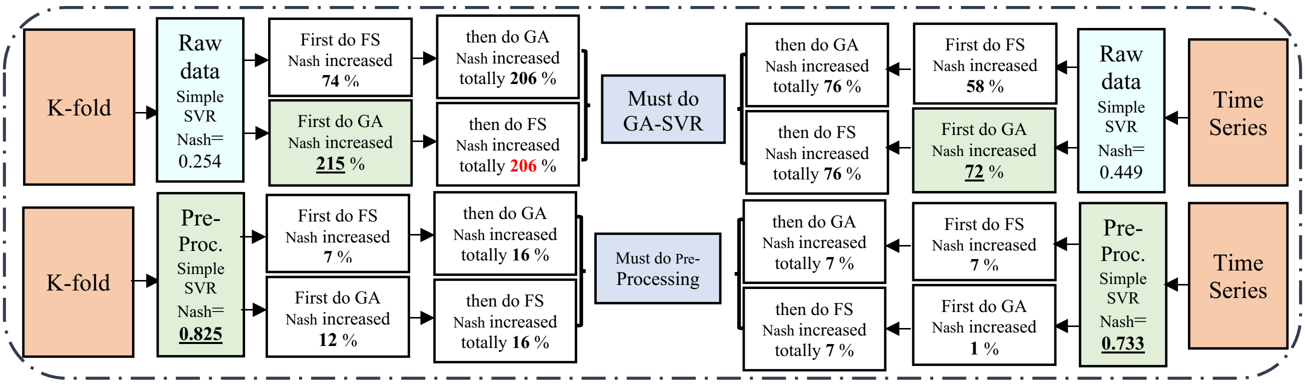


Fig. 14 The graphical conclusion

Funding The author(s) received no specific funding for this work.

Data availability The raw data are available from the corresponding author by request.

Declarations

Conflict of interest The authors have no conflicts of interest to declare that are relevant to the content of this article.

References

Baydaroğlu Ö, Koçak K, Duran K (2018) River flow prediction using hybrid models of support vector regression with the wavelet transform, singular spectrum analysis and chaotic approach. Meteorol Atmos Phys 130(3):349–359. https://doi.org/10.1007/s00703-017-0518-9

Bell B, Wallace B, Zhang D 2012 Forecasting river runoff through support vector machines. In: IEEE 11th Int. Conf. Cogn. Informatics Cogn. Comput., IEEE, pp 58–64

Berrar D (2019) Cross-validation encyclopedia of bioinformatics and computational biology. Academic Press, Oxford, pp 542–545

Chen ST, Yu PS (2007) Pruning of support vector networks on flood forecasting. J Hydrol 347:67–78. https://doi.org/10.1016/j.jhydrol.2007.08.029

Ebrahimi H, Rajaee T (2017) Simulation of groundwater level variations using wavelet combined with neural network, linear regression and support vector machine. Glob Planet Change 148:181–191. https://doi.org/10.1016/j.gloplacha.2016.11.014

Ebrahimi E, Shourian M (2022) A feature-based adaptive combiner for coupling meta-modelling techniques to increase accuracy of river flow prediction. Hydrol Sci J 67(14):2065–2081. https://doi.org/10.1080/02626667.2022.2130700

Ebtehaj I, Bonakdari H, Zeynoddin M, Gharabaghi B, Azari A (2020) Evaluation of preprocessing techniques for improving the accuracy of stochastic rainfall forecast models. Int J Environ Sci Technol 17:505–524. https://doi.org/10.1007/s13762-019-02361-z

Fatemi SE, Parvini H (2022) The impact assessments of the ACF shape on time series forecasting by the ANFIS model. Neur Comput Appl 34:12723–12736. https://doi.org/10.1007/s00521-022-07140-5

Fatemi SE, Ghobadian R, Pakbin M (2018) Forecasting groundwater depth using time series spectral analysis. Water and Soil Science 28(1):145–158

Gorgani S, Bafkar A, Fatemi SE (2017) Prediction of groundwater pollution potential using the DRASTIC index and annual time series analysis (case study: plain Mahidasht Kermanshah). Iran J Health Environ 10(3):317–328

Holland JH (1975) Adaptation in natural and artificial systems. University of Michigan Press, Ann Arbor

Katoch S, Chauhan SS, Kumar V (2021) A review on genetic algorithm: past, present, and future. Multimed Tools Appl 80(5):8091–8126. https://doi.org/10.1007/s11042-020-10139-6

Moeeni H, Bonakdari H, Fatemi SE (2017a) Stochastic model stationarization by eliminating the periodic term and its effect on time series prediction. J Hydrol 547:348–364. https://doi.org/10.1016/j.jhydrol.2017.02.012

Moeeni H, Bonakdari H, Fatemi SE, Zaji AH (2017b) Assessment of stochastic models and a hybrid artificial neural network-genetic algorithm method in forecasting monthly reservoir inflow. INAE Lett 2:13–23. https://doi.org/10.1007/s41403-017-0017-9

Noori R, Karbassi AR, Moghaddamnia A, Han D, Zokaei-Ashtiani MH, Farokhnia AM, GhafariGousheh (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(3/4):177–189. https://doi.org/10.1016/j.jhydrol.2011.02.021

Nourmohammadi Dehbalaei F, Azari A, Akhtari AA (2023) Development of a linear-nonlinear hybrid special model to predict monthly runoff in a catchment area and evaluate its performance with novel machine learning methods. Appl Water Sci 13(5):1–23. https://doi.org/10.1007/s13201-023-01917-2

Okkan U, Serbes ZA (2012) Rainfall–runoff modeling using least squares support vector machines. Environ Metrics 23(6):549–564. https://doi.org/10.1002/env.2154

Parvaz M, Shahoei SV (2022) Investigation using awbm model for monthly runoff simulation of urmia lake basin in Kurdistan

Province, sonnate station. J Environ Sci Stud 7(3):5347–5359. https://doi.org/10.22034/jess.2022.342020.1783

Pourkheirollah Z, Hafezparast Mavaddat M, Fatemi SE (2023) Nash bargaining optimization of released water from a reservoir dam under climate change conditions (case study: doiraj dam). J Agric Sci Technol 25(3):747–765. https://doi.org/10.22034/jast.25.3.747

Sahoo BB, Jha R, Singh A, Kumar D (2019) Application of support vector regression for modeling low flow time series. KSCE J Civ Eng 23:923–934. https://doi.org/10.1007/s12205-018-0128-1

Sahoo A, Singh UK, Kumar MH (2021) Estimation of flood in a river basin through neural networks: a case study. In: Satapathy SC, Bhateja V, Murty RM, Nhu NG, Kotti J (eds) Communication software and networks: proceedings of INDIA 2019. Springer, Singapore, pp 755–763. https://doi.org/10.1007/978-981-15-5397-4_77

Samantaray S, Sahoo A, Dillip KGh (2020) Assessment of sediment load concentration using SVM, SVM-FFA and PSR-SVM-FFA in arid watershed, India: a case study. KSCE J Civ Eng 24(6):1944–1957. https://doi.org/10.1007/s12205-020-1889-x

Samantaray S, Sahoo A, Mohanta NR, Biswal P, Das UK (2021) Runoff prediction using hybrid neural networks in semi-arid watershed. India A Case Study 134:729–736. https://doi.org/10.1007/978-981-15-5397-4_74

Samantaray S, Sawan Das S, Sahoo A, Satapathy DP (2022) Monthly runoff prediction at Baitarani river basin by support vector machine based on Slap swarm algorithm. Ain Shams Eng J 13(5):101732. https://doi.org/10.1016/j.asej.2022.101732

Sanz-Garcia J, Fernandez-Ceniceros F, Antonanzas-Torres AV, Pernia-Espinoza F-d-P (2015) GA-PARSIMONY: A GA-SVR approach with feature selection and parameter optimization to obtain parsimonious solutions for predicting temperature settings in a continuous annealing furnace. Appl Soft Comput 35:13–28. https://doi.org/10.1016/j.asoc.2015.06.012

Soltani K, Azari A (2022) Forecasting groundwater anomaly in the future using satellite information and machine learning. J Hydrol 612(2):128052. https://doi.org/10.1016/j.jhydrol.2022.128052

Soltani K, Azari A (2023) Terrestrial water storage anomaly estimating using machine learning techniques and satellite-based data (a case study of Lake Urmia Basin). Irrigat Drainage. https://doi.org/10.1002/ird.2863

Vapnik VN (1998) Statistical learning theory. Wiley, New York

Vapnik VN, Cortes C (1995) Support vector networks. Mach Learn 20:273–297

Wu J, Liu H, Wei G, Song T, Zhang C, Zhou H (2019) Flash flood forecasting using support vector regression model in a small mountainous catchment. Water 11(7):1327. https://doi.org/10.3390/w11071327

Yaseen ZM, Jaafar O, Deo RC, Kisi O, Quilty J, El-Shafie AA (2016) Stream-flow forecasting using extreme learning machines: a case study in a semi-arid region in Iraq. J Hydrol 542:603–614. https://doi.org/10.1016/j.jhydrol.2016.09.035

Yin Z, Feng Q, Wen X, Deo RC, Yang L, Si J, He Z (2018) Design and evaluation of SVR, MARS and M5Tree models for 1, 2 and 3-day lead time forecasting of river flow data in a semiarid mountainous catchment. Stoch Environ Res Risk Assess 32:2457–2476. https://doi.org/10.1007/s00477-018-1585-2

Zaini N, Malek MA, Yusoff M, Mardi NH, Norhisham S (2018) Daily river flow forecasting with hybrid support vector machine–particle swarm optimization. In: IOP Conf. Ser. Earth Environ. Sci., IOP Publishing Ltd., 140, pp 1315–755 https://doi.org/10.1088/1755-1315/140/1/012035

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