A Novel LSSVM Model Integrated with GBO Algorithm to Assessment of Water Quality Parameters

Mojtaba Kadkhodazadeh
Semnan University
Saeed Farzin (saeed.farzin@semnan.ac.ir)
Semnan University https://orcid.org/0000-0003-4209-9558

Research Article

Keywords: Novel hybrid model, LSSVM, GBO, Water quality parameters, Benchmark datasets, Karun river.

Posted Date: June 8th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-465707/v1

A novel LSSVM model integrated with GBO algorithm to assessment of water quality parameters

Mojtaba Kadkhodazadeh[1], Saeed Farzin[1]✉

ABSTRACT

In this study, a novel least square support vector machine (LSSVM) model integrated with gradient-based optimizer (GBO) algorithm is introduced for assessment of water quality parameters. For this purpose, three stations including Ahvaz, Armand, and Gotvand in the Karun river basin have been selected to model electrical conductivity (EC), and total dissolved solids (TDS). First, to prove the superiority of the LSSVM-GBO algorithm, the performance is evaluated with three benchmark datasets (Housing, LVST, Servo). Then, the results of the new hybrid algorithm were compared with those of artificial neural network (ANN), adaptive neuro-fuzzy interface system (ANFIS), and LSSVM algorithms. Input combination for assessment of water quality parameters EC and TDS consists of Ca^{+2}, Cl^{-1}, Mg^{+2}, Na^{+1}, SO_4, HCO_3, sodium absorption ratio (SAR), sum cation (Sum.C), sum anion (Sum.A), PH, and Q. The modelling results based on evaluation criteria showed the significant performance of LSSVM-GBO among all benchmark datasets and algorithms. Other results showed that in Ahvaz station, Sum.C, Sum.A, and Na^{+1} parameters, and in Gotvand and Armand stations, Sum.C, Sum.A, and Cl^{-1} parameters have the greatest impact on modelling EC and TDS parameters. In the next step, EC and TDS modelling was performed based on the best input combination and the best algorithm in

Mojtaba Kadkhodazadeh
mkadkhodazadeh@semnan.ac.ir

✉Saeed Farzin
saeed.farzin@semnan.ac.ir

[1] Department of Water Engineering and Hydraulic Structures, Faculty of Civil Engineering, Semnan University, Semnan 35131-19111, Iran

different time delays. Based on the results, the highest accuracy of modelling EC and TDS parameters in Gotvand station was [0] month time delays.

Keywords: Novel hybrid model, LSSVM, GBO, Water quality parameters, Benchmark datasets, Karun river.

1 Introduction

Surface water quality in a region depends on the nature and extent of human, industrial, and other human activities on the site. Rivers carrying water and nutrients are necessary for various regions of the earth and provide important resources for drinking, industrial, aquatic, recreational and agricultural consumption. Therefore, they require at least an acceptable level of water quality. In recent years due to the rapid population growth and urban extension, the increasing irregular water withdrawals, agriculture, economic development, and increasing industrial production, pollution has increased in rivers; therefore, the qualitative study of water resources is one of the most important challenges in most regions of the world (Ehteshami et al. 2014).

One of the most important ways to study the problems of water pollution is modelling and analysis of water quality using modern methods such as artificial intelligence. In recent years, many studies have been conducted about EC and TDS modelling in different regions using data mining methods as these methods have a lot of accuracies and, like physical and mathematical models, they do not need to specify a large number of parameters and reduce the cost of research work. Modelling TDS and EC concentration and predicting it is essential for pollution control and water resource management (Azad et al. 2019).

Naddafi et al. (2007) reported that the logarithmic and the exponential models describe the concentration-time relationships for Gotvand and Khorramshahr station stations in Karun river

better. Also, Mojahedi and Attari (2009) used two water quality indices for Karun river. Results showed that application of these indices was satisfactory. Faruk (2010) used a hybrid neural network and ARIMA models for water quality time series prediction. Their results showed that the hybrid model provides much better accuracy by itself and ensures a better method to include special parameters into water quality index due to superior capabilities of fuzzy logic in dealing with different systems (Semiromi et al. 2011). Asadollahfardi et al. (2011) applied two ANN networks models to predict TDS. The results of this study showed that the Elman network has higher accuracy. Emamgholizadeh et al. (2014) applied multi-layer perceptron (MLP), radial basis network, and ANFIS models to predict biochemical oxygen demand (BOD), dissolved oxygen (DO), and chemical oxygen demand (COD).The results showed that MLP was better than other models in predicting water quality variables. In another study, the accuracy of ANN, ANFIS, wavelet-ANN, and wavelet-ANFIS in predicting monthly water salinity levels was assessed. Their results showed that the ANFIS provides much better accuracy than the ANN (Barzegar et al. 2016). Salami et al. (2016) presented two mathematical and ANN methods to estimate the forecast river water quality. An acceptable precision was achieved, as shown in model verification results. Azad et al. (2018) reported that ANFIS-DE (differential evolution (DE)), ANFIS-GA (genetic algorithm (GA)), and ant colony optimization for continuous domains (ACO_R) models performed well in modelling EC, SAR, and Total Hardness (TH). Khosravi et al. (2018) reported that the hybrid models performed better than individual models. Haghiabi et al. (2018) used ANN, SVM, and group method of data handling (GMDH) to estimate EC and TDS. The evaluation of the accuracy of the applied models according to the error indexes declared that SVM was the most accurate model. Kisi et al. (2019) showed that compact genetic algorithm (CGA), ACO_R, DE, particle swarm optimization (PSO) improved the ANFIS performance in the modelling of EC and TH. Aryafar et al. (2019) used ANN, ANFIS, and Genetic programming (GP) to estimate the EC,

TDS and TH. Satisfactory performances were also produced by the ANN and ANFIS methods for the estimation of the intended groundwater quality parameters. Najafzadeh et al. (2019) used gene express programming (GEP), model tree (MT) and evolutionary polynomial regression (EPR) to estimate three indices including BOD, DO, and COD. The performance of the models indicated the relative superiority of the EPR approach to the GEP and MT models. Deterministic and numerical models have been applied extensively to model water quality; the counting convolutional neural network (CCNN) model has high accuracy for estimating the DO concentration (Zounemat-Kermani et al. 2019). Barzegari Banadkooki et al. (2020) used ANN, ANFIS, SVM to predict the TDS. Also, the moth flam optimization (MFO), cat swarm optimization (CSO), PSO, shark algorithm (SA), gray wolf optimization (GWO), and the gravitational search algorithm (GSA) were used to train the ANFIS, SVM, and ANN model. In this study, the ANFIS-MFO and ANFIS-CSO models showed superior performance to the other models. M.Melesse et al.(2020) applied two individual M5 Prime (M5P) and random forest (RF) and eight novel hybrid algorithms to predict EC. Results indicate that, in most cases, hybrid algorithms enhance predictive powers.

In the present study, for the first time, a new model is proposed to assess water quality parameters. For this purpose, the hybrid of LSSVM and GBO (LSSVM-GBO) is introduced for modelling electrical conductivity (EC), and total dissolved solids (TDS) at three hydrometric stations with different kinds of climate and water quality in Karun river area (as a case study). To prove the superiority of the LSSVM-GBO algorithm, the performance is evaluated with benchmark datasets. Then, the results of LSSVM-GBO are compared with the ANN, ANFIS, and LSSVM algorithms to demonstrate the ability and accuracy of the proposed algorithm. Finally, EC and TDS modelling is performed based on the best input combination and the best algorithm in different time delays.

2 Materials and Methods

2.1 Benchmark Datasets

In this article, the performance of the proposed LSSVM-GBO algorithm is compared with that of other algorithms on 3 real-world regression problems. Benchmarks that are real world issues, are a good criterion to determine algorithms working. In recent years, many studies have been conducted to compare algorithms with real-world regression problems (Breiman 2001; Zhang and Yang 2015; A.Henríquez and A.Ruz 2017). Specifications of benchmark datasets are shown in Table 1. Also, Fig. 1 shows the process of changing target data in benchmark datasets.

Table 1 Specification of real-world regression benchmark datasets

Fig. 1 The process of changing target data. a) Housing, b)LVST, c) Servo

2.2.1 Artificial Neural Network (ANN)

ANN is an information processing system that has certain performance characteristics resembling biological neural networks of the human brain (Tun Lee et al. 2008). ANN is based on a collection of connected units or nodes called artificial neurons (Anaraki et al. 2021). Each connection can transmit a signal to other neurons (Acharya et al. 2019). An artificial neuron receives a signal and then processes it. The connections between neurons are called edges. Neurons and edges typically have a weight.

2.2.2 Multi-Layer Perceptron (MLP)

MLP has three layers: input layer, output layer, and hidden layer. The first layer receives a set of data (x_1, x_2, …, x_n). The second layer is known as hidden layer. The number of hidden layers displays the intricacy of the MLP because a greater number of hidden layers increase the number of connections in the ANN. The number of nodes in each layer and the number of the hidden layer is evaluated by trial and error. The third layer is the output layer. The multi-layer perceptron

produces an expected result y in the output layer. The MLP is trained with a training set of input and known output data. So far, different methods have been put forward to determine these weights, the most famous of which is Levenberg-Marquart method (Hadi and Tombul 2018).

$$Z = f\left(\sum_{i=1}^{n} w_i x_i + b\right) \quad (1)$$

where, f: activation function, b: bias parameter, w_i: the weight of connection, Z: the output of ith neuron, and x_i: the received input from the ith neuron. Fig. 2 shows the structure of the ANN model.

Fig. 2 The schematic structure of ANN model

2.3 Adaptive Neuro-Fuzzy Interface System (ANFIS)

ANFIS model method is a well-known artificial intelligence method that has been used currently in water quality parameters, predicting rainfall and hydrological variables. ANFIS modelling is a reach where the combination of neural networks and fuzzy argument find their strengths (C.S.Bisht et al. 2011).

This model combines the advantage of both neural networks and fuzzy logic and can benefit from that at the same time (Kumar et al. 2019). ANFIS techniques can learn a system performance from enough large data sets and automatically procreate fuzzy sets to a pre-specified correctness level. The ANFIS model consists of five layers; input layers, rule layers, average layers, consequent layers, and total output layer. The main duty of the ANFIS is to the optimize values of the equivalent fuzzy such that the error between the target and the actual output is minimized. Two fuzzy "if-then" rules are used as follows (Yaseen et al. 2018):

$$\text{IF x is } A_1 \text{ and y is } B_1 \text{ THEN } f_1 = p_1x+q_1y+r_1 \quad (2)$$

$$\text{IF x is } A_2 \text{ and y is } B_2 \text{ THEN } f_2 = p_2x+q_2y+r_2 \quad (3)$$

where, x and y are input variables, A_i and B_i are the linguistic labels characterized by convenient membership functions (i= 1 or 2) and p_i, q_i, r_i: the output function parameters (i= 1or 2).

Many studies have been done using ANFIS algorithm. Refer to other research for more information on the ANFIS algorithm (Jang 1993; Ghordoyee Milan et al. 2021). Fig. 3 shows the structure of the ANFIS model.

Fig. 3 The schematic structure of ANFIS model

2.4 Least Square Support Vector Machine (LSSVM)

LSSVM is an implementation of support vector machine for the problem of classification and pattern identification, regression analysis, and the problem of learning a ranking function. The advantages of LSSVM include high precision, mathematical tractability, and direct geometric commentary. The algorithm converts the nonlinear relationship between inputs and outputs to a linear relationship (Keshtegar et al. 2019). The LSSVM uses the following equation to show the relationship between input and outputs.

$$M = \sum_{i=1}^{n} k(x,x_i)\alpha_i + b \tag{4}$$

where, M: the output value, α_i:weighting coefficient of input data, b: bias, k(x): the nonlinear mapping function. The LSSVM tries to minimize the difference between measured data and estimated data. The parameters α_i and b are computed as follows (Farzin and Valikhan-Anaraki 2021):

$$\begin{bmatrix} \text{kernel} & \bar{1}^T \\ \bar{1} & \text{kernel} + C^{-1}I \end{bmatrix} \begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ M \end{bmatrix} \tag{5}$$

C: regulation parameter; the parameters $\alpha, M, I, \bar{1}$ are computed as follows:

$$\alpha = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix}, \bar{1} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}, M = \begin{bmatrix} M_1 \\ \vdots \\ M_n \end{bmatrix}, I = \text{diag}(1,1,\ldots.1) \tag{6}$$

The radial basis function is used as a kernel function, as follows (Ghosh 2010):

$$k(x, x_i) = \exp\left(\frac{-\|x - x_i\|^2}{2\sigma^2}\right) \quad (7)$$

Fig. 4 shows the structure of the LSSVM model.

Fig. 4 The schematic structure of LSSVM model

2.5 Optimization Algorithms

2.5.1 GBO Algorithm

In this article, a novel meta-heuristic optimization algorithm, GBO, is used. The GBO uses two main operators: gradient search rule (GSR) and local escaping operator (LEO) and a set of vectors to explore the search space. GSR uses the slope-based method to reach better positions in the search space. In these algorithms, the search engine implements two stages of exploration and exploitation (Olorunda and P. Engelbrecht 2008). Simultaneous operation of two stages causes the optimal performance of this algorithm. Therefore, creating a suitable balance between these two processes is crucial (Patel and Savsani 2015). In recent years, many studies have been conducted to make performance of basic algorithms better by creating a suitable balance between two stages of exploration and exploitation or hybridization of optimization algorithms (Draa et al. 2015).

The GBO was a proposed as a meta-heuristic optimization algorithm by Ahmadianfar et al. (2020). They evaluated the performance of the GBO algorithm using 28 mathematical functions. Also, they observed that the GBO provided more optimal results than the other algorithms and was able to optimize the real-world problems with challenging and unknown search domains. They showed that this algorithm has a more promising operation capability than other optimization algorithms. The GBO algorithm uses a search engine based on the Newton method to find the optimal answer in the following. The GBO algorithm has a special feature in that it results from the combination of GSR and LEO (Hassan et al. 2021).

2.5.2 Gradient Search Rule (GSR)

GSR is the core of the GBO algorithm. The duty of GSR is to find better opportunities and increase convergence rate acceleration (DM) in the search space. Therefore, the equation to update the current vector position is (x_n^m) :

$$X1_n^m = x_n^m - randn \times \rho_1 \times \frac{2\Delta x \times x_n^m}{(yp_n^m - yq_n^m + \varepsilon)} + rand \times \rho_2 \times (x_{best} - x_n^m) \quad (8)$$

Where, randn is a normally distributed random number, ε: a small number within the range of [0, 0.10], rand is a random number in [0, 1], and x_{best} is the best solution. ρ_1 is an important parameter in the GBO to balance between two stages of exploration and exploitation, Δx is determined based on the difference between the best solution (x_{best}) and a randomly selected position (x_{r1}^m) and ρ_2 is a random parameter, which causes the vector to have a different step size.

By replacing the position of the best vector (x_{best}) with the current vector (x_n^m) in Eq. (8), the new vector ($X2_n^m$) ia obtained as follows:

$$X2_n^m = x_{best} - randn \times \rho_1 \times \frac{2\Delta x \times x_n^m}{(yp_n^m - yq_n^m + \varepsilon)} + rand \times \rho_2 \times (x_{r1}^m - x_{r2}^m) \quad (9)$$

Due to the positions $X1_n^m$, $X2_n^m$, X_n^m, the new solution at the next iteration (x_n^{m+1}) can be defined as:

$$x_n^{m+1} = r_a \times (r_b \times X1_n^m + (1 - r_b) \times X2_n^m) + (1 - r_a) \times X3_n^m \quad (10)$$

$$X3_n^m = X_n^m - \rho_1 \times (X2_n^m - X1_n^m) \quad (11)$$

r_a and r_b are two random numbers in [0, 1]. Fig. 5 shows the structure of the GBO model.

Fig. 5 The schematic structure of GBO algorithm

2.5.3 Local Escaping Operator (LEO)

The LEO effectively avoids being trapped in local optima and improves the convergence speed of the GBO algorithm. The LEO is capable of solving complex problems in the GBO algorithm. By using several solutions (the solutions $X1_n^m$ and $X2_n^m$, the best position (x_{best}) and the solutions $X1_n^m$ and $X2_n^m$), the LEO generates a solution with a superior performance (X_{LEO}^m). The solution X_{LEO}^m is produced as follows:

1: if rand $<$ pr (12)

if rand < 0.5

2: $X_{LEO}^m = X_n^{m+1} + f_1 \times (u_1 \times x_{best} - u_2 \times x_k^m) + f_2 \times \rho_1 \times (u_3 \times (X2_n^m - X1_n^m) + u_2 \times (x_{r1}^m - x_{r2}^m))/2$

3: $X_n^{m+1} = X_{LEO}^m$

4: Else

5: $X_{LEO}^m = x_{best} + f_1 \times (u_1 \times x_{best} - u_2 \times x_k^m) + f_2 \times \rho_1 \times (u_3 \times (X2_n^m - X1_n^m) + u_2 \times (x_{r1}^m - x_{r2}^m))/2$

6: $X_n^{m+1} = X_{LEO}^m$

7: End

8: End

where, f_1: uniform random number in the range of [-1,1], f_2: a random number from a normal distribution with mean of 0 and standard deviation of 1, pr: the probability, while u_1, u_2, and u_3: are three random numbers.

For more details, see Ahmadianfar et al. (2020).

2.6 Hybrid of LSSVM and GBO

The proposed LSSVM-GBO aims to reduce the probability to be trapped into local optima as well as accelerate the solution process. The parameters C and σ have a significant effect on the accuracy of the LSSVM model. In this study, the optimization algorithm GBO was used to find the optimal value of the LSSVM parameters. In the hybrid algorithm of LSSVM and GBO, the values of the LSSVM parameters are considered as decision variables. Also, the pseudo code of LSSVM-GBO is illustrated in Fig. 6. The steps of the LSSVM-GBO algorithm are described as follows:

1- Test and training data are randomly selected from the available data.

2- The initial parameters of the optimization algorithms GBO (The number of iterations and the population size) are randomly determined.

3- The LSSVM parameters (initial population) are initialized, and the GBO algorithm finds the optimal solution (values of parameters C and σ) in the search space.

4- After obtaining the optimal answer of the LSSVM parameters, training data and test data are used to obtain the LSSVM optimization model and to evaluate the predictive ability of the LSSSVM optimization model.

Fig. 6 Pseudo code of the proposed LSSVM-GBO algorithm

2.7 Data Collected

In this research, eleven input combinations including Ca^{+2}, Cl^{-1}, Mg^{+2}, Na^{+1}, SO_4, HCO_3, SAR, Sum.C, Sum.A, PH, Q were used to model the quality parameters of EC and TDS. Table 2 shows statistical specifications of input and output data.

Table 2 Statistical specifications of input and output data

2.8 Evaluation Criteria

To evaluate the accuracy of the model performances, four statistical criteria, mean absolute error (MAE), relative root mean square error (RRMSE), and correlation coefficient (R) and R^2 were calculated in their testing phases. Expressions for these measures are given as follows (Ehteram et al. 2018) :

$$\mathrm{MAE} = \frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} |\mathrm{O_i} - \mathrm{P_i}| \tag{13}$$

$$\mathrm{RRMSE} = \left(\sqrt{\frac{\sum_{\mathrm{i}=1}^{\mathrm{N}} (\mathrm{O_i} - \mathrm{P_i})^2}{\mathrm{N}}} \right) / \mathrm{SD(P)} \tag{14}$$

$$\mathrm{R} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{N}} (\mathrm{P_i} - \overline{\mathrm{P}})(\mathrm{O_i} - \overline{\mathrm{O}})}{\sqrt{\sum_{\mathrm{i}=1}^{\mathrm{N}} (\mathrm{P_i} - \overline{\mathrm{P}}) \sum_{\mathrm{i}=1}^{\mathrm{N}} (\mathrm{O_i} - \overline{\mathrm{O}})}} \tag{15}$$

$$\mathrm{R^2} = \left[\frac{\sum_{\mathrm{i}=1}^{\mathrm{N}} (\mathrm{P_i} - \overline{\mathrm{P}})(\mathrm{O_i} - \overline{\mathrm{O}})}{\sqrt{\sum_{\mathrm{i}=1}^{\mathrm{N}} (\mathrm{P_i} - \overline{\mathrm{P}}) \sum_{\mathrm{i}=1}^{\mathrm{N}} (\mathrm{O_i} - \overline{\mathrm{O}})}} \right]^2 \tag{16}$$

where, N: the number of data, O: observed values, P: predicted values. Other information on evaluation criteria is shown in Table 3 (Valikhan-Anaraki et al. 2019):

Table 3 Information on evaluation criteria

2.9 Study Area

The Karun river is situated in the south west of Iran and with a basin area of 67257 km^2 and with a length of 95 miles is the longest and most important river of Iran which collects the runoff of extensive areas and conveys it to the Persian Gulf. Karun river originates from Zagros mountain ranges which are stretched from northwest to southeast. Ahvaz metropolis due to having the largest and increasing population enters pollution into the river more than other cities in that

almost half of the incoming pollution is from Ahvaz metropolis, including domestic, urban, and hospital sewage. Karun river is the only navigable river in Iran. The average annual precipitation in Karun is 620 mm. The climate of the Karun domain is hot with dry summers and mild winters. In the present study, three stations including Ahvaz, Armand, and Gotvand in the Karun river have been selected to model water quality. Ahvaz and Gotvand stations are located in Khuzestan province, and Armand station is located in Chaharmahal Bakhtiari province. The choice of stations has been such that all climates are examined Gotvand has an arid and semi-arid climate, Ahvaz has a dry and extremely dry climate, and Armand has a humid and Mediterranean climate. Fig. 7 shows the location of the investigated stations in the Karun basin.

Fig. 7 Study area and hydrometric station

2.10 Assessment of Water Quality Parameters

Fig. 8 depicts the general framework for the assessment of water quality parameters. The steps of water quality modelling are described as follows:

1- Comparison of the performance of algorithms with 3 bench marks.

2- 70 % of the data is considered for the training data and 30 % for the test data.

3- Water quality is modelled based on the training data and the test data with the mentioned algorithms.

4-According to the evaluation criteria, the best algorithm and the best input combination are determined.

5- Creating time delays in the best algorithm and best input combination.

6-Calculating evaluation criteria in time delays and select the best time delay.

7-Time series calculation of EC and TDS.

Fig. 8 Flowchart for modelling water quality parameter

3 Results and Discussion

Tables 4, 5, and 6 show the correlation between water quality parameters in three stations of Ahvaz, Armand, and Gotvand. Based on the results of the correlation matrix, the highest amount of correlation of EC and TDS parameters with input parameters related to Sum.A and Sum.C at three stations. The correlation between EC and inputs is greater. The higher the correlation between inputs and outputs, the higher the modelling accuracy.

Table 4 Correlation matrix water quality parameters in Ahvaz station

Table 5 Correlation matrix water quality parameters in Armand station

Table 6 Correlation matrix water quality parameters in Gotvand station

According to Tables 7, the performance of the LSSVR-GBO algorithm is compared with that of other algorithms on 3 real-world regression problems. Results showed that the hybrid model provides much better accuracy than the ANN, ANFIS, and LSSVM model. Values of MAE, RRMSE, R in Housing dataset were 5.30, 0.91, 0.44, respectively. Also, in LVST dataset they were 99.94, 0.20, 0.98, respectively, and in Servo dataset were 0.46, 0.41, 0.91, respectively. In Fig. 9, the modelling accuracy for the LSSVM-GBO algorithm and benchmark dataset have been shown.

Table 7 Evaluation criteria for measuring precision in benchmarks modelling

Fig. 9 Comparison of scatter plots by LSSVM-GBO algorithm. a) Housing-train, b) Housing-test , c) LVST-train, d) LVST-test, e) Servo-train, f) Servo-test

In Table 8 the results of the algorithms are listed to the EC modelling of the Ahvaz station. Based on the results of Table 8, the LSSVM-GBO algorithm has the highest accuracy. The modelling results showed that Sum.C parameter has the most impact in modelling EC. In the optimal hybrid model, values of MAE, RRMSE, and R were 74.30, 0.14, 0.99, respectively. In Table 9, the

evaluation criteria of the listed to the EC modelling of the Armand station. Sum.C parameter has the most impact in modelling EC in this station. Also, the value of MAE is equal to 19.67, RRMSE equal to 0.22, R equal to 0.98. Also, the LSSVM-GBO algorithm has the highest accuracy. In Table 10, the results of the EC calculated by different algorithms and the combination of different inputs in the Gotvand station are compared. The modelling results showed that the LSSVM-GBO algorithm has the highest accuracy. Also, Sum.C parameter has the greatest impact in modelling EC in Gotvand station. Values of MAE, RRMSE, R were 57.12, 0.16, 0.99, respectively.

Table 8 Results of the algorithms to the EC modelling of the Ahvaz station

Table 9 Results of the algorithms to the EC modelling of the Armand station

Table 10 Results of the algorithms to the EC modelling of the Gotvand station

The results of TDS modelling in Ahvaz station by ANN, ANFIS, LSSVM, LSSVM-GBO algorithms are shown in Table 11. Sum.C parameter has the most significant impact in modelling TDS and the LSSVM-GBO algorithm has the highest accuracy. Values of MAE, RRMSE, R were 74.89, 0.43, 0.90, respectively. The results of TDS modelling in the Armand station are shown in Table 12. Based on the results of Table 12 the LSSVM-GBO algorithm has the highest accuracy. The modelling results showed that the Sum.C parameter has the most significant impact on modelling TDS. Values of MAE, RRMSE, R were 16.99, 0.26, 0.97, respectively. In Table 13, the results of the TDS calculated by different algorithms and the combination of different inputs in the Gotvand station are compared. Sum.C parameter has the most significant impact on modelling TDS in Gotvand station. Values of MAE, RRMSE, R were 37.18, 0.17, 0.99, respectively. The modelling results showed that the LSSVM-GBO algorithm has the highest accuracy.

Table 11 Results of the algorithms to the TDS modelling of the Ahvaz station

Table 12 Results of the algorithms to the TDS modelling of the Armand station

Table 13 Results of the algorithms to the TDS modelling of the Gotvand station

After detecting the best algorithm (LSSVM-GBO) and best input combination (in Ahvaz station, Sum.C, Sum.A, Na^{+1}, and Q parameters, and in Gotvand and Armand stations, Sum.C, Sum.A, Cl^{-1}, and Q parameters), in Tables 14-19, the effect of time delay on EC and TDS modelling results is investigated. According to Tables 14-16, the results of the EC in Ahvaz station, Armand stations, and Gotvand station the effect of different time delays are shown. The best results are related to the time delay of [0] months.

Table 14 Time delays in the modelling of the EC in the Ahvaz station by the best algorithm

Table 15 Time delays in the modelling of the EC in the Armand station by the best algorithm

Table16 Time delays in the modelling of the EC in the Gotvand station by the best algorithm

In Fig. 10, the modelling accuracy for the LSSVM-GBO algorithm and the EC parameter in Ahvaz, Armand, and Gotvand stations has been shown. According to Fig. 10, in the test and training period, the highest and lowest accuracy is related to Gotvand (MAE=49.86, RRMSE=0.14, R=0.99), and Armand stations (MAE=26.26, RRMSE=0.31, R=0.95), respectively.

Fig. 10 Comparison of scatter plots by LSSVM-GBO algorithm. a)Ahvaz-train, b)Ahvaz-test, c)Armand-train, d)Armand-test, e)Gotvand-train, f)Gotvand-test

Also, according to Tables 17, 18, and 19, for the TDS parameter in Ahvaz station, Armand stations, and Gotvand station, the best results are related to the time delay of [0] months.

Table 17 Time delays in the modelling of the TDS in the Ahvaz station by the best algorithm

Table 18 Time delays in the modelling of the TDS in the Armand station by the best algorithm

Table 19 Time delays in the modelling of the TDS in the Gotvand station by the best algorithm

In Fig. 11, the modelling accuracy for the LSSVM-GBO algorithm and the TDS parameter in three stations has been shown. According to Fig. 11, in the training period, the highest and lowest accuracy is related to Gotvand (MAE=35.41, RRMSE=0.26, R=0.96) and Armand stations (MAE=17.91, RRMSE=0.38, R=0.92), respectively, and in the test period, the highest and lowest accuracy is related to Gotvand (MAE=33.86, RRMSE=0.16, R=0.99) and Ahvaz stations (MAE=74.81, RRMSE=0.43, R=0.90), respectively.

Fig. 11 Comparison of scatter plots by LSSVM-GBO algorithm. a)Ahvaz-train, b)Ahvaz-test, c)Armand-train, d)Armand-test, e)Gotvand-train, f)Gotvand-test

In Fig. 12, the results of the EC and TDS time series model based on best input combination, best time delay, and best algorithm (LSSVM-GBO) are compared in three stations. According to these figures, the amount of EC and TDS fluctuations is well modelled by LSSVM-GBO algorithm, which indicates the high accuracy of this algorithm.

Fig. 12 The results of the EC and TDS time series model. a)Ahvaz-EC, b)Armand-EC, c)Gotvand-EC, d)Ahvaz-TDS, e)Armand-TDS, f)Gotvand-TDS

4 Conclusion

River water pollution is increasing due to various activities. Therefore, it is necessary to know the water quality of rivers. Machine learning algorithms are a good and efficient approach for the prediction of river qualitative parameters. In this research, a novel LSSVM model integrated with GBO algorithm was used to estimate EC and TDS values in the Karun river in three hydrometric stations of Gotvand, Ahvaz, and Armand. In the first step, the performance of the proposed LSSVM-GBO algorithm is compared with that of other algorithms on 3 real-world regression

problems(Housing, LVST, Servo). The modelling results showed that LSSVM-GBO indicate the highest accuracy on 3 regression problems. Values of MAE, RRMSE, R in Housing dataset were 5.30, 0.91, 0.44, respectively. Also, in LVST dataset were 99.94, 0.20, 0.98, respectively, and in Servo dataset were 0.46, 0.41, 0.91, respectively. Eleven input combinations including Ca^{+2}, Cl^{-1}, Mg^{+2}, Na^{+1}, SO_4, HCO_3, SAR, Sum.C, Sum.A, PH, Q was used to model the quality parameters of EC and TDS. The modelling results based on evaluation criteria showed the most significant performance of LSSVM-GBO among all algorithms. The modelling results showed that Sum.C, Sum.A, Na^{+1}, and Cl^{-1} parameters have the most noticeable impact on modelling EC and TDS parameters. Based on the results, the highest accuracy of modelling EC and TDS parameter in Gotvand station was [0] month time delays. Values of MAE, RRMSE, R in modelling EC were 49.86, 0.14, 0.99, respectively, and in modelling TDS were 33.86, 0.16, 0.99, respectively. Examination of EC and TDS time series modeled by LSSVM-GBO and observational time series showed high correlation between modelling and observational results. Armand station with an average TDS of 345 Mg/lit has a high water quality, and at Ahvaz station, the average TDS is equal to 1034 Mg/lit, which indicates that polluting effluents enter the river at this station. The LSSVM-GBO algorithm has various advantages, such as high estimation accuracy, balance between exploration and exploitation, and fast convergence, ability to find a global solution, and easy implementation. The GBO does not fall into the trap of local optima due to the use of a local escaping operator. Also, the GBO uses the direction of movement term to move towards the solution. According to the results of this study, and due to the many advantages of the proposed algorithm, the LSSVM-GBO is a good candidate to analyze other engineering problems.

Funding The research has not been supported through any funds.

Data Availability All data generated or used during the study are applicable if requested.

Compliance with Ethical Standards

Competing Interests The authors declare that they have no conflict of interest.

Ethical Approval Not applicable.

Consent to Participate The authors have made a significant contribution to this manuscript, have seen and approved the final manuscript.

Consent to Publish The authors have agreed to publish the study in water resource management.

References

Acharya R, Pal J, Das D, Chaudhuri S (2019) Long-range forecast of Indian summer monsoon rainfallusing an artificial neural network model. Meteorol. Appl. 26:347-361. https://doi.org/10.1002/met.1766

A.Henríquez P, A.Ruz G (2017) Extreme learning machine with a deterministic assignment of hidden weights in two parallel layers. Neurocomputing 226:109-116. https://doi.org/10.1016/j.neucom.2016.11.040

Ahmadianfar I, Bozorg-Haddad O, Chu X (2020) Gradient-based optimizer: a new metaheuristic optimization algorithm. Information Sciences 540:131-159. https://doi.org/10.1016/j.ins.2020.06.037

Aryafar A, Khosravi V, Zarepourfard H, Rooki R (2019) Evolving genetic programming and other AI-based models for estimating groundwater quality parameters of the Khezri plain, Eastern Iran. Environmental Earth Sciences 78:69. https://doi.org/10.1007/s12665-019-8092-8

Asadollahfardi G, Taklify A, Ghanbari A (2011) Application of artificial neural network to predict TDS in Talkheh Rud river. Journal of Irrigation and Drainage Engineering 138:363-370. https://doi.org/10.1061/(ASCE)IR.1943-4774.0000402

Azad A, Karami H, Farzin S, Saeedian A, Kashi H, Sayyahi F (2018) Prediction of water quality parameters using ANFIS optimized by intelligence algorithms (case study: Gorganrood river). KSCE Journal of Civil Engineering 22:2206-2213. https://doi.org/10.1007/s12205-017-1703-6

Azad A, Karami H, Farzin S, Mousavi SF, Kisi O (2019) Modelling river water quality parameters using modified adaptive neuro fuzzy inference system. Water Science and Engineering 12:45-54. https://doi.org/10.1016/j.wse.2018.11.001

Barzegar R, Adamowski J, Asghari-Moghaddam A (2016) Application of wavelet-artificial intelligence hybrid models for water quality prediction: a case study in Aji-Chay river, Iran. Stochastic Environmental Research and Risk Assessment 30:1797-1819. https://doi.org/10.1007/s00477-016-1213-y

Barzegari-Banadkooki F, Ehteram M, Panahi F, Sammen S, BintiOthman F, EL-Shafie A (2020) Estimation of total dissolved solids (TDS) using new hybrid machine learning models. J Hydrol 587:124989. https://doi.org/10.1016/j.jhydrol.2020.124989

Breiman L (2001) Random forests. Machine Learning 45:5–32. https://doi.org/10.1023/A:1010933404324

C.S.Bisht D, Jangid A (2011) Discharge modelling using Adaptive neuro-fuzzy inference system. International Journal of Advanced Science and Technology 31:99-114.

Draa A, Bouzoubia S, Boukhalfa I (2015) A sinusoidal differential evolution algorithm for numerical optimisation. Applied Soft Computing 27:99-126. https://doi.org/10.1016/j.asoc.2014.11.003

Ehteram M, Karami H, Farzin S (2018) Reducing irrigation deficiencies based optimizing model for multi-reservoir systems utilizing spider monkey algorithm. Water Resources Management 32:2315-2334. https://doi.org/10.1007/s11269-018-1931-7

Ehteshami M, Biglarijoo N, Salari M (2014) Assessment and quality classification of water in Karun, Dez and Karkheh rivers. Journal of River Engineering 2(8).

Emamgholizadeh S, Kashi H, Marofpoor I, Zalaghi E (2014) Prediction of water quality parameters of Karoon river (Iran) by artificial intelligence-based models. International Journal of Environmental Science and Technology 11:645-656. https://doi.org/10.1007/s13762-013-0378-x

Faruk DO (2010) A hybrid neural network and ARIMA model for water quality time series prediction. Engineering Applications of Artificial Intelligence 23:586-594. https://doi.org/10.1016/j.engappai.2009.09.015

Farzin S, Valikhan-Anaraki M (2021) Modelling and predicting suspended sediment load under climate change conditions: a new hybridization strategy. Journal of Water and Climate Change 2021317. https://doi.org/10.2166/wcc.2021.317

Ghordoyee-Milan S, Roozbahani A, Arya Azar N, Javadi S (2021) Development of adaptive neuro fuzzy system-evolutionary algorithms hybrid models (ANFIS-EA) for prediction of optimal groundwater exploitation. J Hydrol 598:126258. https://doi.org/10.1016/j.jhydrol.2021.126258

Ghosh S (2010) SVM-PGSL coupled approach for statistical downscaling to predict rainfall from GCM output. J Geophys Res 115:D22102. https://doi.org/10.1029/2009JD013548

Hadi SJ, Tombul M (2018) Streamflow forecasting using four wavelet transformation combinations approaches with data-driven models: a comparative study. Water Resources Management 32:4661- 4679. https://doi.org/10.1007/s11269-018-2077-3

Haghiabi AH, Nasrolahi AH, Parsaie A (2018) Water quality prediction using machine learning methods. Water Quality Research Journal 533:3-13. https://doi.org/10.2166/wqrj.2018.025

H.Hassan M, Kamel S, El-Dabah MA, Rezk H (2021) A novel solution methodology based on a modified gradient-based optimizer for parameter estimation of photovoltaic models. Journal Electronics 10:472. https://doi.org/10.3390/electronics10040472

Jang J-S.R (1993) ANFIS: adaptive-network-based fuzzy inference system. IEEE Transactions on Systems, Man, and Cybernetics 23:665-685.

Keshtegar B, Heddam S, Kisi O, Zhu Sh-P (2019) Modelling total dissolved gas (TDG) concentration at Columbia river basin dams: high-order response surface method (H-RSM) vs. M5Tree, LSSVM, and MARS. Arabian Journal of Geosciences 12:544. https://doi.org/10.1007/s12517-019-4687-3

Khosravi Kh, Mao L, Kisi O, Yaseen ZM, Shahid Sh (2018) Quantifying hourly suspended sediment load using data mining models: case study of a glacierized Andean catchment in Chile. J Hydrol 567:165-179. https://doi.org/10.1016/j.jhydrol.2018.10.015

Kisi O, Azad A, Kashi H, Saeedian A, Hashemi SAA, Ghorbani S (2019) Modelling groundwater quality parameters using hybrid neuro-fuzzy methods. Water Resources Management 33:847-861. https://doi.org/10.1007/s11269-018-2147-6

Kumar A, Kumar P, Kumar Singh V (2019) Evaluating different machine learning models for runoff and suspended sediment simulation. Water Resources Management 33:1217-1231. https://doi.org/10.1007/s11269-018-2178-z

M.Melesse A, Khosravi kh, P.Tiefenbacher J, Heddam S, Kim S, Mosavi A, Thai Pham B (2020) River water salinity prediction using hybrid machine learning models. Water 12: 2951. https://doi.org/10.3390/w12102951

Mojahedi SA, Attari J (2009) River a comparative study of water quality indices for Karun river. World Environmental and Water Resources Congress 2444-2452. https://doi.org/10.1061/41036(342)246

Naddafi K, Honari H, Ahmadi M (2007) Water quality trend analysis for the Karoon river in Iran. Environmental Monitoring and Assessment 134:305-312. https://doi.org/10.1007/s10661-007-9621-6

Najafzadeh M, Ghaemi A, Emamgholizadeh S (2019) Prediction of water quality parameters using evolutionary computing-based formulations. Int. J. Environ. Sci.Technol 16:6377-6396. https://doi.org/10.1007/s13762-018-2049-4

Olorunda O, P.Engelbrecht A (2008) Measuring exploration/exploitation in particle swarms using swarm diversity. IEEE Congress on Evolutionary Computation (IEEE World Congress on Computational Intelligence) 10221910:1128-1134.

Patel VK, Savsani VJ (2015) Heat transfer search (HTS): a novel optimization algorithm. Information Sciences 324:217-246. https://doi.org/10.1016/j.ins.2015.06.044

Salami ES, Salari M, Ehteshami M, Bidokhti NT, Ghadimi H (2016) Application of artificial neural networks and mathematical modelling for the prediction of water quality variables (case

study: southwest of Iran). Desalination and Water Treatment 57:27073-27084. https://doi.org/10.1080/19443994.2016.1167624

Semiromi FB, Hassani AH, Torabian A, Karbassi AR, Hosseinzadeh-Lotfi F (2011) Water quality index development using fuzzy logic: A case study of the Karoon river of Iran. African Journal of Biotechnology 10:10125-10133. 10.5897/AJB11.1608

Tun Lee K, Hung W-Ch, Meng Ch-Ch (2008) Deterministic insight into ANN model performance for storm runoff simulation. Water Resources Management 22:67-82. https://doi.org/10.1007/s11269-006-9144-x

Valikhan-Anaraki M, Farzin S, Mousavi SF, Karami H (2021) Uncertainty analysis of climate change impacts on flood frequency by using hybrid machine learning methods. Water Resources Management 35:199-223. https://doi.org/10.1007/s11269-020-02719-w

Valikhan-Anaraki M, Mousavi SF, Farzin S, Karami H, ... , El-Shafie A (2019) Development of a novel hybrid optimization algorithm for minimizing irrigation deficiencies. Sustainability 11:2337. https://doi.org/10.3390/su11082337

Yaseen Z.M, Ghareb M.I, Ebtehaj I, Bonakdari H, Siddique R, Heddam S, A.Yusif A, Deo R (2018) Rainfall pattern forecasting using novel hybrid intelligent model based ANFIS-FFA. Water Resources Management 32:105-122. https://doi.org/10.1007/s11269-017-1797-0

Zhang P, Yang Zh (2015) A robust AdaBoost.RT based ensemble extreme learning machine. Mathematical Problems in Engineering 260970. https://doi.org/10.1155/2015/260970

Zounemat-Kermani M, Seo Y, Kim S, Ghorbani M.A, Samadianfard S, Naghshara S, Kim N.W, Singh V.P (2019) Can decomposition approaches always enhance soft computing models?

Predicting the dissolved oxygen concentration in the St. Johns river, Florida. Appl. Sci. 9:2534. https://doi.org/10.3390/app9122534

Figures

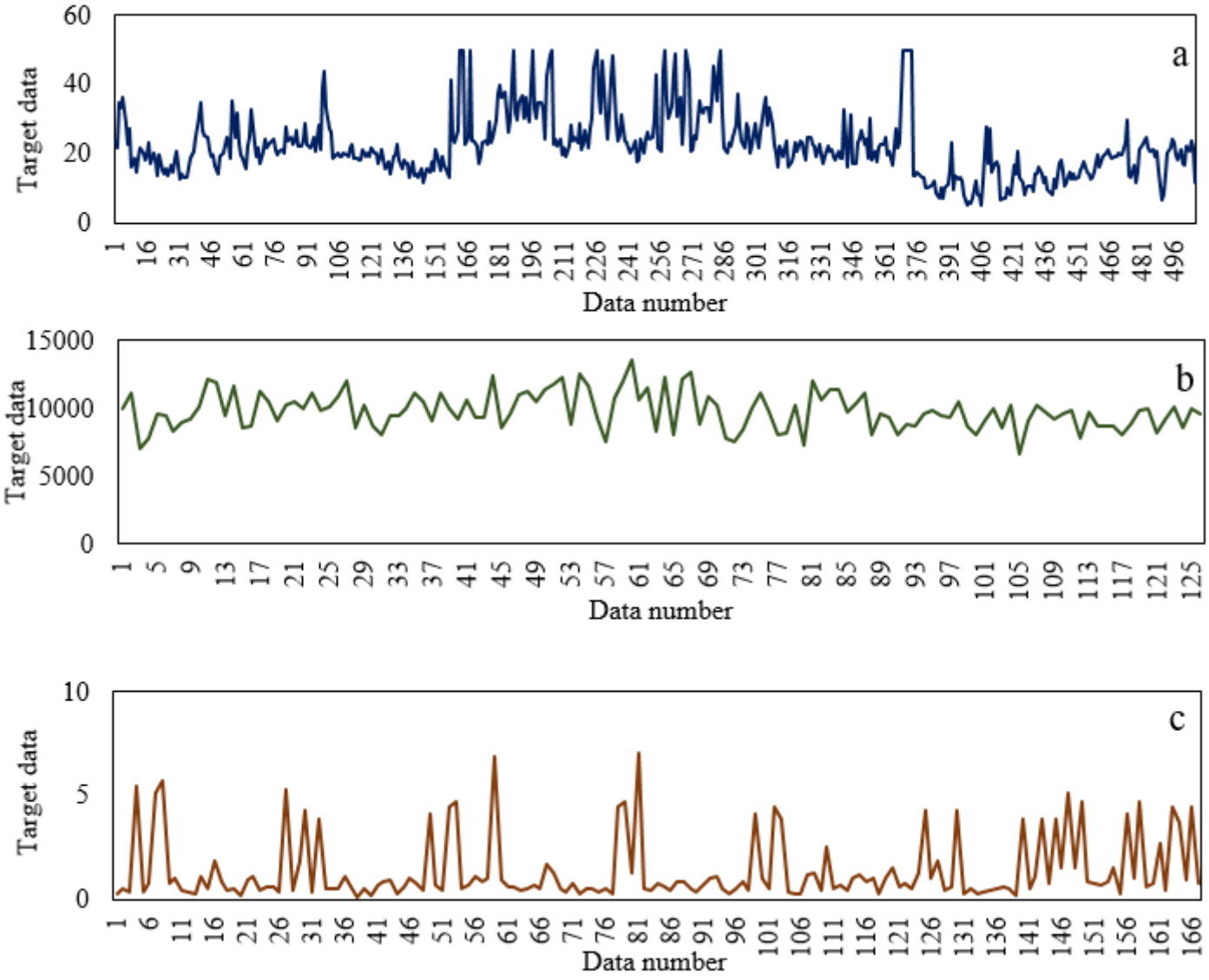

Figure 1

The process of changing target data. a) Housing, b)LVST, c) Servo

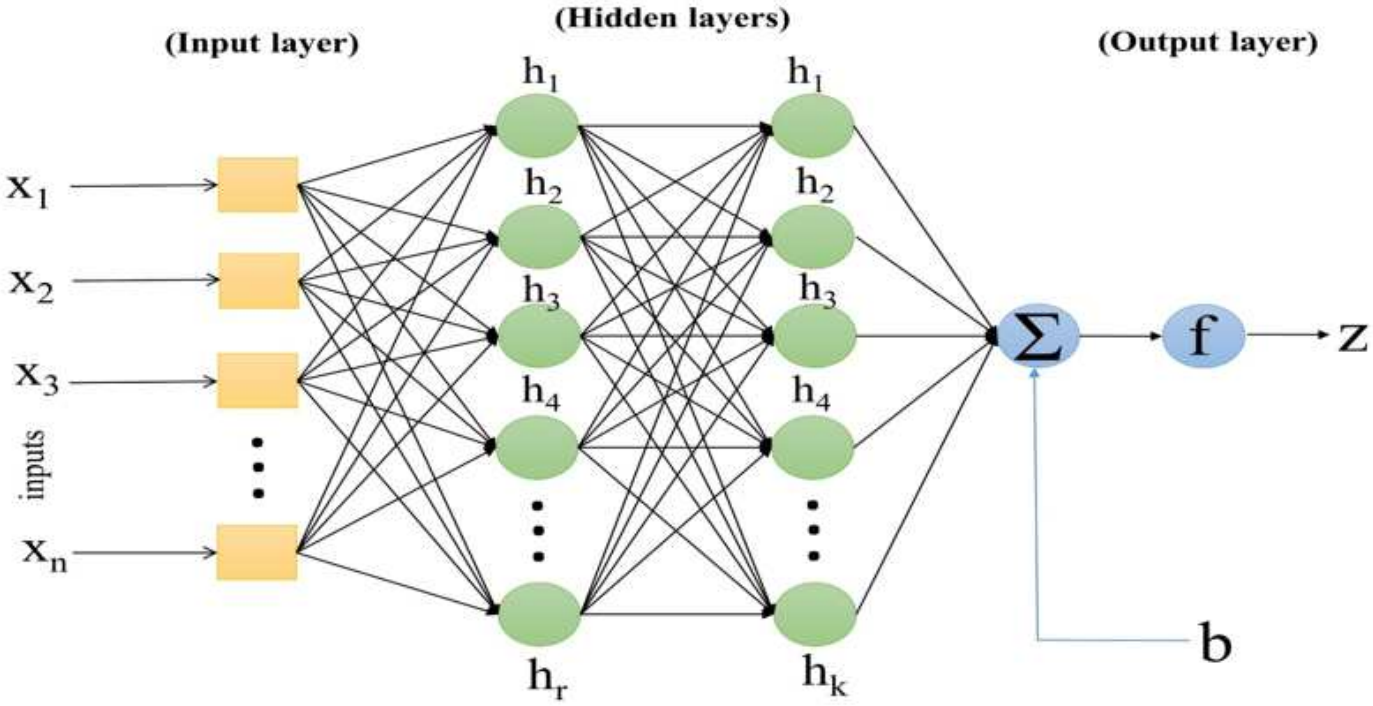

Figure 2

The schematic structure of ANN model

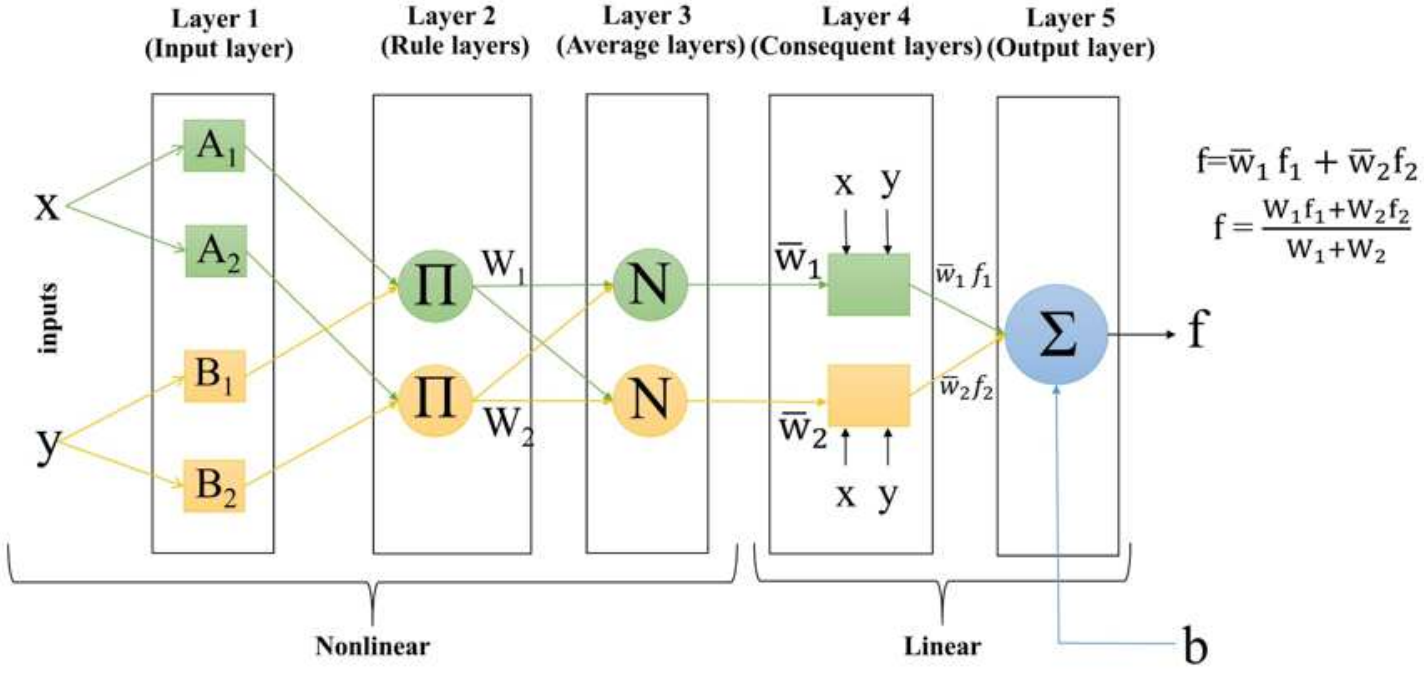

Figure 3

The schematic structure of ANFIS model

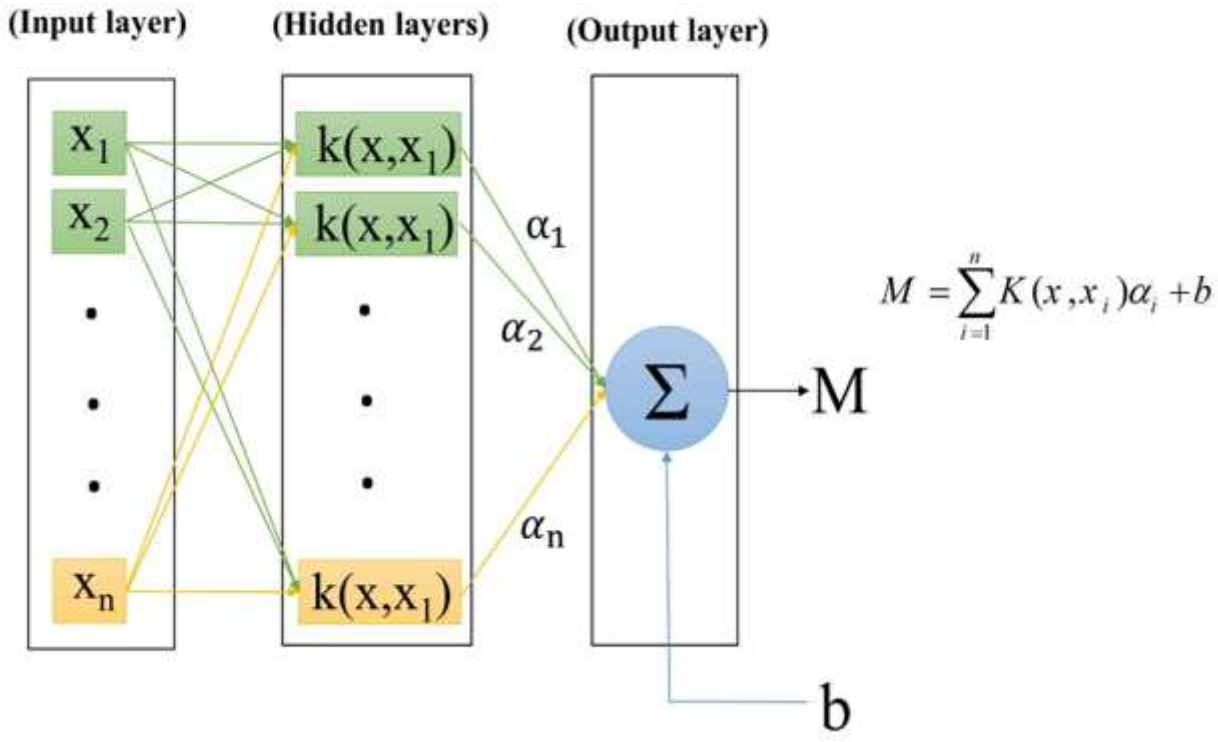

Figure 4

The schematic structure of LSSVM model

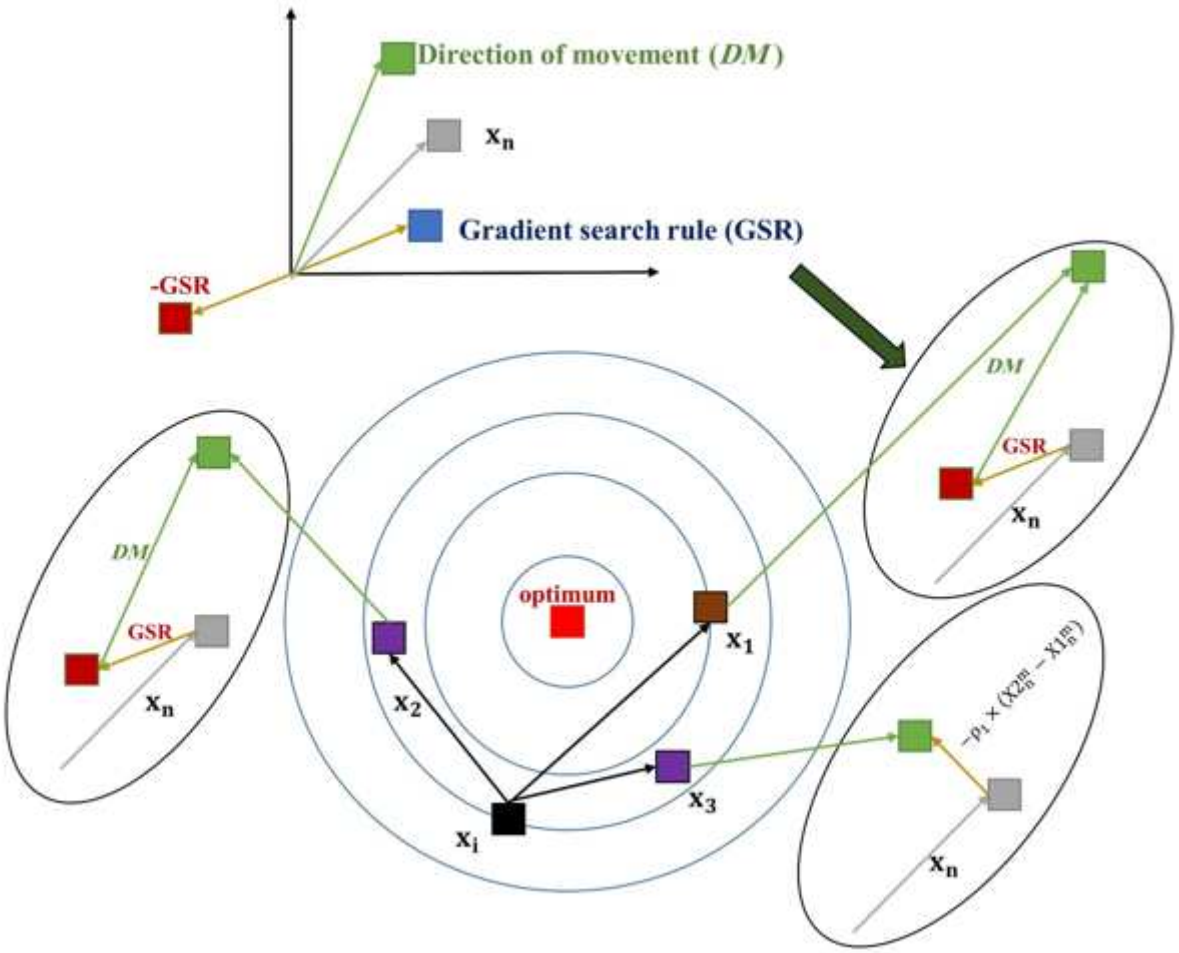

Figure 5

The schematic structure of GBO algorithm

1: Define initial GBO parameters: population size, the maximum number of iteration (M) and pr
2: Generate initial population of GBO $X_{0,d} = [X_{0,1}, X_{0,2}, \ldots, X_{0,D}]$
3: Execute LSSVM for each search agent GBO and calculate objective function for each GBO $f_{(X0)}$, $n = 1, \ldots, N$
4: Select best and worst solution GBO according to objective function of search agent (x^{m}_{best}, x^{m}_{worst})
5: While iteration (m) < M
6: For n=1: population size (N)
7: For i=1: D
8: Select randomly x^{m}_{r1} , x^{m}_{r2}, x^{m}_{r3}, x^{m}_{r4} : different integers randomly chosen from [1, N]
9: Calculate the positions x^{m+1}_{n}(Eq. 10)
10: End for
11: If rand < pr
12: Calculate the positions x^{m}_{LEO} (Eq. 12)
13: $x^{m+1}_{n} = x^{m}_{LEO}$
14: End if
15: Update the position x^{m}_{best}, x^{m}_{worst}
16: Run LSSVM by new values of parameters C and σ
17: Calculate the evaluation criteria
18: End for
19: m = m+1
20: End while
21: Return x^{m}_{best} (C and σ) whit the evaluation criteria

Figure 6

Pseudo code of the proposed LSSVM-GBO algorithm

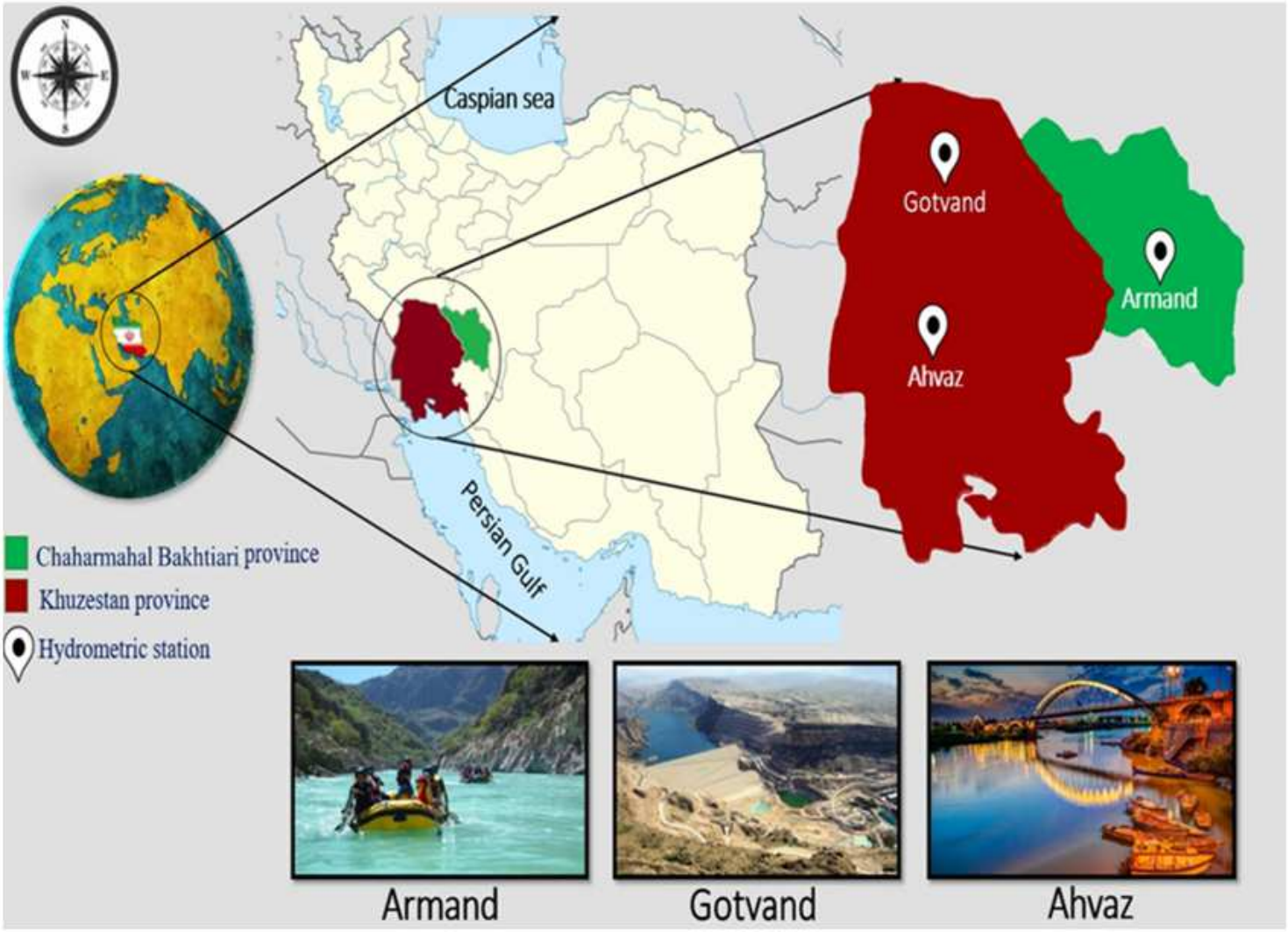

Figure 7

Study area and hydrometric station

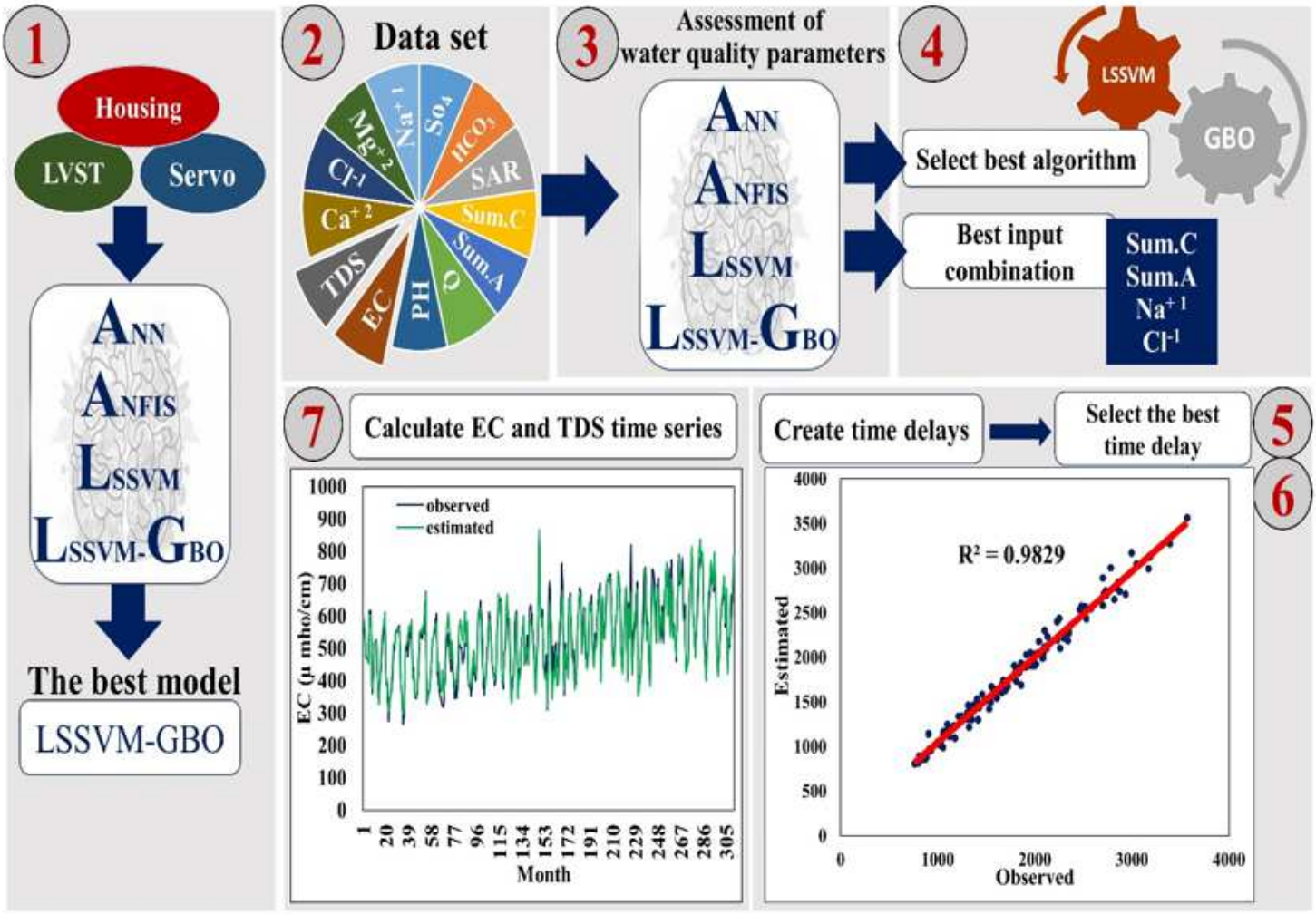

Figure 8

Flowchart for modelling water quality parameter

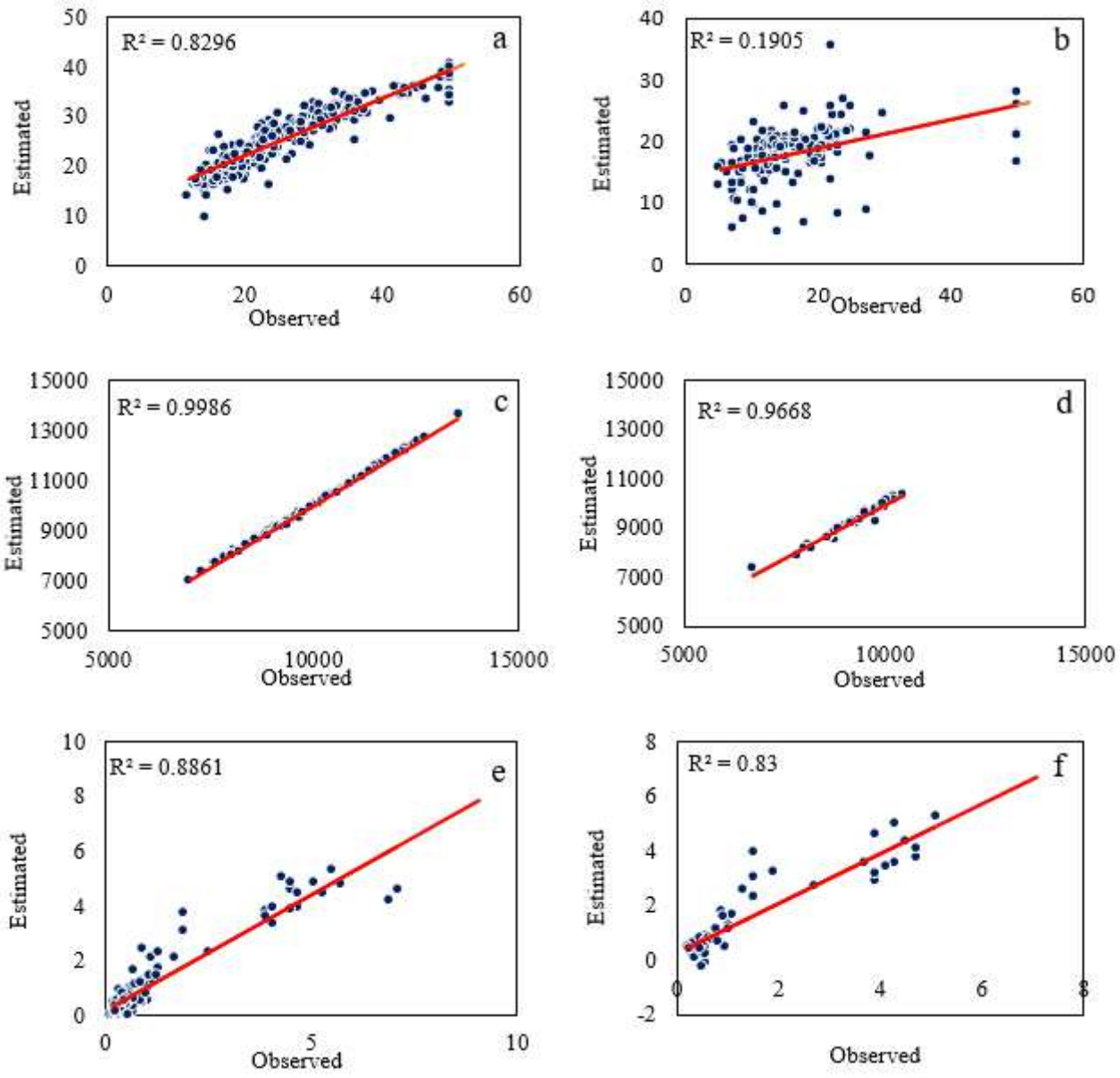

Figure 9

Comparison of scatter plots by LSSVM-GBO algorithm. a) Housing-train, b) Housing-test , c) LVST-train, d) LVST-test, e) Servo-train, f) Servo-test

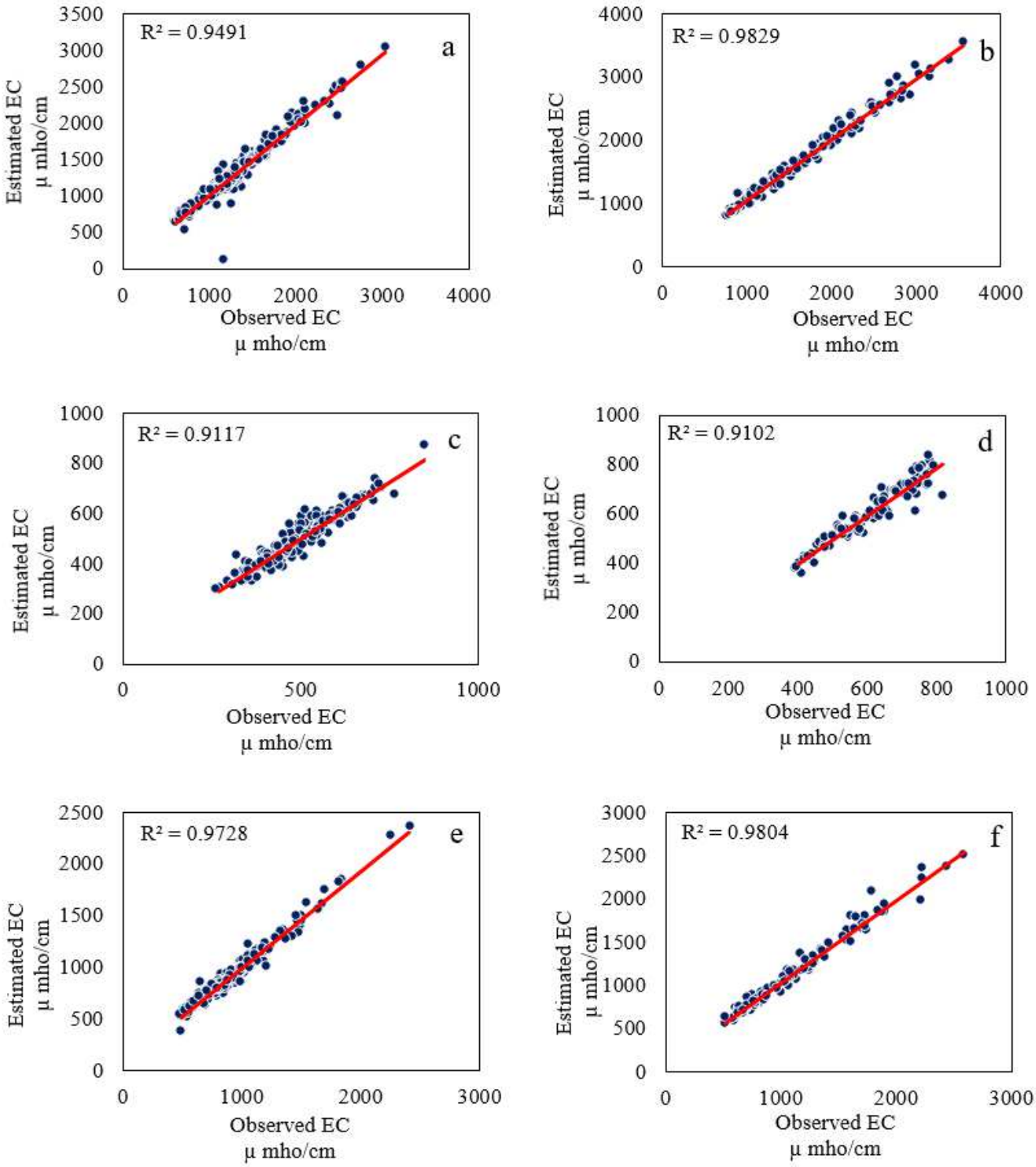

Figure 10

Comparison of scatter plots by LSSVM-GBO algorithm. a)Ahvaz-train, b)Ahvaz-test, c)Armand-train, d)Armand-test, e)Gotvand-train, f)Gotvand-test

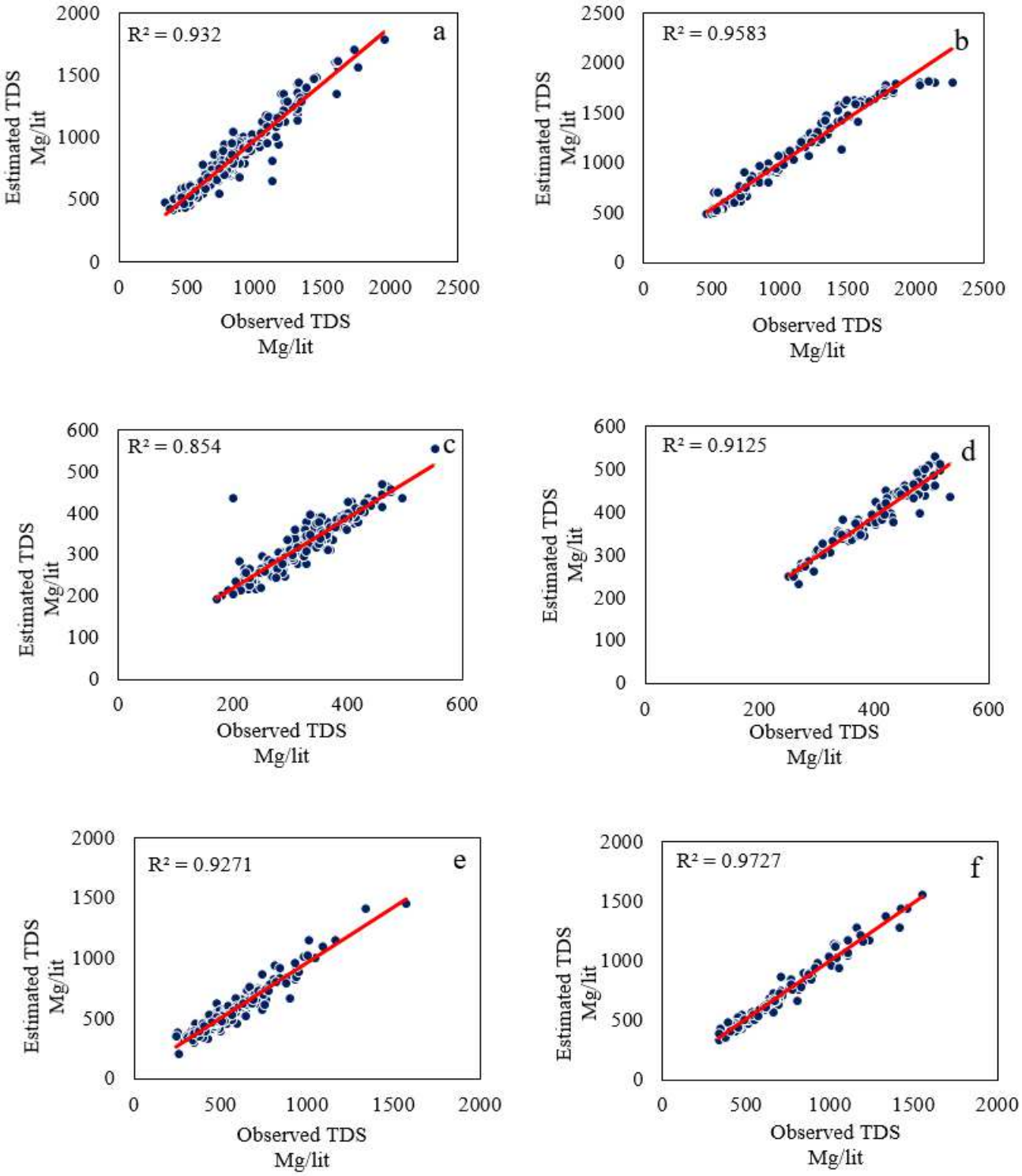

Figure 11

Comparison of scatter plots by LSSVM-GBO algorithm. a)Ahvaz-train, b)Ahvaz-test, c)Armand-train, d)Armand-test, e)Gotvand-train, f)Gotvand-test

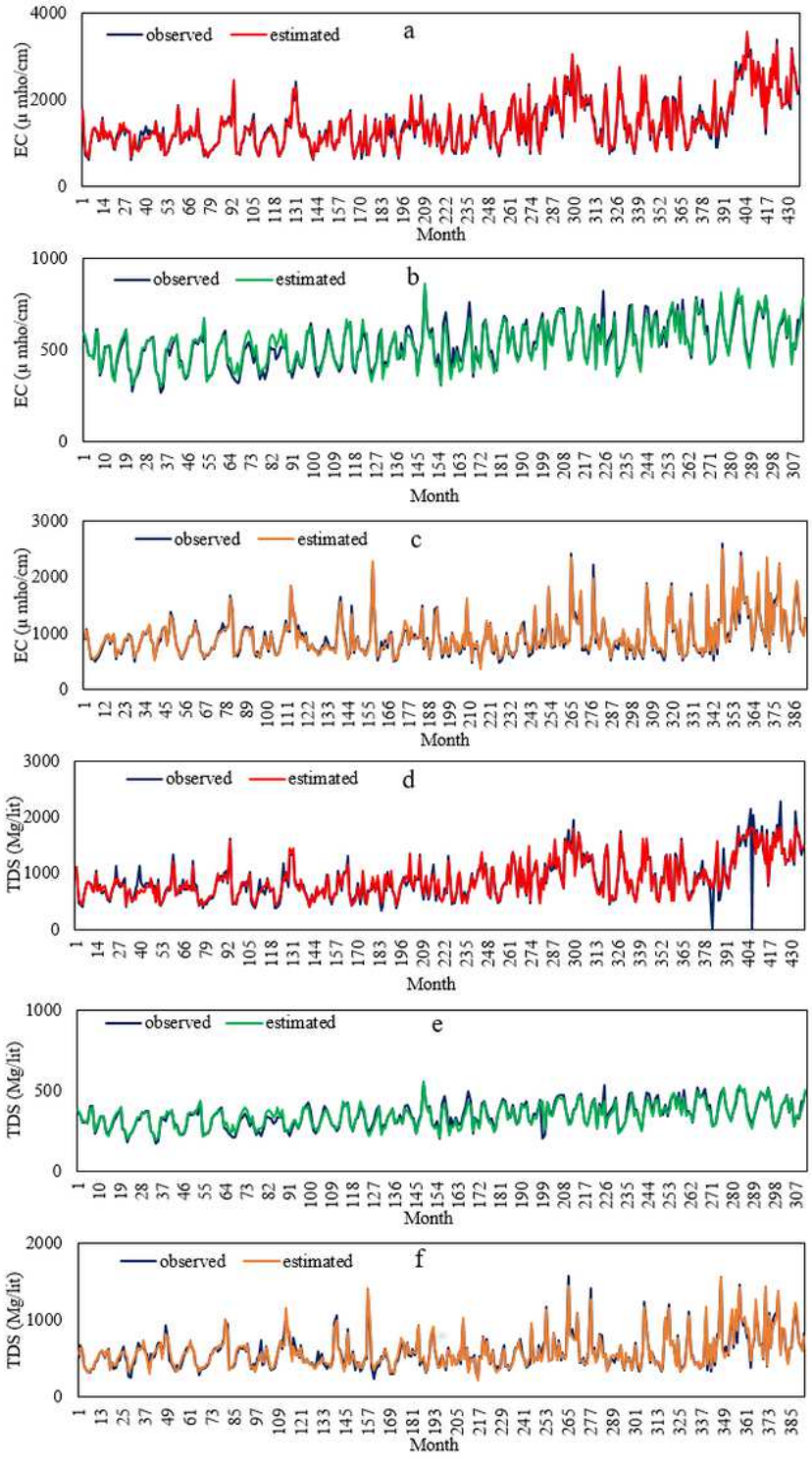

Figure 12

The results of the EC and TDS time series model. a)Ahvaz-EC, b)Armand-EC, c)Gotvand-EC, d)Ahvaz-TDS, e)Armand-TDS, f)Gotvand-TDS

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- graphicalabstract.png