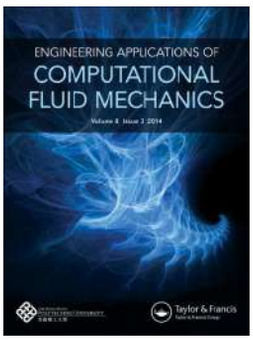

Engineering Applications of Computational Fluid Mechanics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tcfm20

An integrated machine learning, noise suppression, and population-based algorithm to improve total dissolved solids prediction

Kangjie Sun , Mohammad Rajabtabar , Seyedehzahra Samadi , Mohammad Rezaie-Balf , Alireza Ghaemi , Shahab S. Band & Amir Mosavi

To cite this article: Kangjie Sun , Mohammad Rajabtabar , Seyedehzahra Samadi , Mohammad Rezaie-Balf , Alireza Ghaemi , Shahab S. Band & Amir Mosavi (2021) An integrated machine learning, noise suppression, and population-based algorithm to improve total dissolved solids prediction, Engineering Applications of Computational Fluid Mechanics, 15:1, 251-271, DOI: 10.1080/19942060.2020.1861987

To link to this article: https://doi.org/10.1080/19942060.2020.1861987

Published online: 28 Jan 2021.

ENGINEERING APPLICATIONS OF COMPUTATIONAL FLUID MECHANICS
2021, VOL. 15, NO. 1, 251–271
https://doi.org/10.1080/19942060.2020.1861987

Taylor & Francis
Taylor & Francis Group

An integrated machine learning, noise suppression, and population-based algorithm to improve total dissolved solids prediction

Kangjie Sun[a], Mohammad Rajabtabar[b], Seyedehzahra Samadi[c], Mohammad Rezaie-Balf[d], Alireza Ghaemi[e], Shahab S. Band[f,g] and Amir Mosavi [h,i,j,k,l]

[a]School of Mechanical Engineering, Lanzhou Jiaotong University, Lanzhou 730070, Gansu, China; [b]Department of Computer Science and Software Engineering, Islamic Azad University Babol Branch, Babol, Iran; [c]Department of Agricultural Sciences, Clemson University, Clemson, SC, USA; [d]Department of Water Engineering, Graduate University of Advanced Technology, Kerman, Iran; [e]Department of Water Engineering, University of Sistan and Baluchestan, Zahedan, Iran; [f]Institute of Research and Development, Duy Tan University, Da Nang 550000, Vietnam; [g]Future Technology Research Center, College of Future, National Yunlin University of Science and Technology, Douliou, Yunlin 64002, Taiwan, ROC; [h]Faculty of Civil Engineering, Technische Universität Dresden, Dresden, Germany; [i]John von Neumann Faculty of Informatics, Obuda University, Budapest, Hungary; [j]School of Economics and Business, Norwegian University of Life Sciences, Ås, Norway; [k]School of the Built Environment, Oxford Brookes University, Oxford, UK; [l]Department of Informatics, J. Selye University, Komarno, Slovakia

ABSTRACT

Monitoring the water contaminants is of utmost importance in water resource management. Prediction of the total dissolved solid (TDS) is particularly essential for water quality management and planning in the areas exposed to a mixture of pollutants. TDS primarily includes inorganic minerals and organic matters, and various salts and increasing the concentration of TDS causes the esthetic problems. The reflection of the pollutant burden of the aquatic system can remarkably determined by TDS magnitudes. This study focuses on the prediction of TDS and several biochemical parameters such as Na, Ca, HCO3, and Mg in a river system. To overcome nonstationarity, randomness, and nonlinearity of the TDS data, a multi-step supervised machine learning evolutionary algorithm (MSMLEA) is proposed to improve the model's performance at two gaging stations, namely Rig-Cheshmeh and Soleyman-Tangeh, in the Tajan River, Iran. In addition, a hybrid model that recruits intrinsic time-scale decomposition (ITD) for frequency resolution of the input data as well as a multivariate adaptive regression spline (MARS) were adopted. A novel metaheuristic optimization algorithm, crow search algorithm (CSA), was also implemented to compute the optimal parameter values for the MARS model. To validate the proposed hybrid model, standalone MARS, empirical mode decomposition (EMD)-based models, and hybrid ITD-MARS as well as a MARS-CSA were considered as the benchmark models. Results suggest the ITD-MARS-CSA outperforms other models.

ARTICLE HISTORY
Received 7 July 2020
Accepted 29 November 2020

KEYWORDS
water pollution; multivariate adaptive regression splines; crow search algorithm; artificial intelligence; machine learning

Nomenclatures

Adaptive neuro-fuzzy inference system	ANFIS
Analysis of variance	ANOVA
Artificial neural network	ANN
Backpropagation neural networks	BPNN
Bicarbonate	HCO3
Basic function	BF
Calcium	Ca
Crow search algorithm	CSA
Complete ensemble empirical mode decomposition	CEEMD
Extreme learning machine	ELM
Generalized cross-validation	GCV
Gene expression programming	GEP
Intrinsic time-scale decomposition	ITD
Magnesium	Mg
Model tree	MT
Multiple linear regression	MLR
Multivariate adaptive regression splines	MARS
Multilayer perceptron	MLP
Multi-step supervised machine learning evolutionary algorithm	MSMLEA
Nash-Sutcliffe Efficiency	NSE
Percent Mean Absolute Relative Error	PMARE
Principal component regression	PCR
Proper rotation component	PRC
Ratio of RMSE to Standard Deviation	RSD
Root mean square error	RMSE
Total dissolved solids	TDS
Support vector machine	SVM
Sodium	Na
Vibrational mode decomposition	VMD
Wilmot's Index of agreement	WI

CONTACT Amir Mosavi amir.mosavi@mailbox.tu-dresden.de, amirhosein.mosavi@nmbu.no; Shahab S. Band shamshirbands@yuntech.edu.tw, shamshirbandshahaboddin@duytan.edu.vn
This article has been republished with minor changes. These changes do not impact the academic content of the article.

1. Introduction

The most common sources of river water are for irrigation, water supply, agriculture, etc. River systems are extremely susceptible to pollution as they are inherently dynamic and convenient environments for the disposal of waste material (Ahmed et al., 2019; Bui et al., 2020). For past decades, mismanagement of river systems caused widespread contamination that has hampered water bodies and rivers.

For the local water quality management, contamination is a significant issue. The amount of organic or inorganic matter (i.e. salts) dissolved in a water system is called TDS (total dissolved solids) and is usually measured as the amount/number of cations and anions contained in a sample. Inorganic and organic matter, minerals, and salts consist most of these dissolved solids (Miranda & Krishnakumar, 2015). Increasing the concentration of TDS may lead to adverse changes in esthetics with respect to precipitation, staining, or taste (Sibanda et al., 2014). TDS also leads to toxicity by increasing salinity and changing in the ionic composition of the water and toxicity of individual ions. Increases in salinity have acute or chronic influences on the biotic communities as well as specific life stages. The TDS concentration is one of the prominent water quality indexes (Jonnalagadda & Mhere, 2001; Weber-Scannell & Duffy, 2007). In this regard, it is crucial to have an accurate model to predict TDS that has significant social and practical values. As physical, biological, and chemical parameters for water quality parameters (WQPs) prediction are strongly nonlinear, non-mechanical computer training models were applied for the TDS prediction.

Since last decades, machine learning models like adaptive neuro-fuzzy inference system (ANFIS), artificial neural network (ANN), model tree (MT), gene expression programming (GEP), support vector machine (SVM), and extreme learning machine (ELM) have been widely extensively developed designed for solving various environmental engineering and water quality problems (Alizadeh et al., 2018; Anctil et al., 2008; Attar et al., 2018; Chen et al., 2020; Chen & Chau, 2019; Choubin et al., 2019; Hong et al., 2018; Kargar et al., 2020; Mouatadid et al., 2018; Najafzadeh et al., 2016; Najafzadeh et al., 2019; Noori & Kalin, 2016; Rezaie-Balf & Kisi, 2018; Shamshirband et al., 2019; Shiri et al., 2011; Solomatine & Xue, 2004; Taormina & Chau, 2015; Yassin et al., 2016; Zounemat-Kermani et al., 2018).

In terms of TDS estimation, a plethora of studies have been carried out that a couple of them can be mentioned here. Abudu et al. (2012) applied ANN, transfer function-noise, and Autoregressive Integrated Moving Average (ARIMA) techniques for the monthly prediction of TDS content in the Rio Grande in El Paso, Texas. Ghavidel and Montaseri (2014) employed ANN, GEP, and ANFIS with grid partition as well as ANFIS with subtractive clustering to predict TDS values of the Zarinehroud basin, Iran. In a sequence, Khaki et al. (2015) evaluated the ability of ANN and ANFIS for the TDS estimation in the Langat Basin, Malaysia. The performance of ANN in the estimation of TDS was further strengthened by Mustafa (2015) and Asadollahfardi et al. (2018) who applied multilayer perceptron (MLP) and Box–Jenkins time series approaches for the TDS prediction in the Zayande Rud River, Iran. Soon after, Pan et al. (2019) assessed the potential of hybrid principal component regression (PCR), dual-step multiple linear regression (MLR), and backpropagation neural networks (BPNN) to model the TDS for an aquifer system in Canada. Although there are strong approaches with high ability, achieving more accurate predictive methods remains a challenging task for the TDS assessment.

Owing to the forgoing hydorogical components, WQPs behavior is known by high non-stationarity, nonlinearity, and anthropogenic changes. In this sense, creating an accurate TDS prediction model due to the existing high complexity issue is highly challenging. MARS model is one of the reliable machine learning (ML) models, which has demonstrated its capability in solving engineering regression problems (Rezaie-Balf et al., 2019; Zhang & Goh, 2016). The construction of MARS model highly depended on three parameters, namely maximum basis function (*MaxFun*), penalty parameter (*d*), and interaction (I_{max}). In this regard, it is hard to select the optimum parameters simultaneously because of the variety choices. The DDMs can be modeled as optimization problems in continuous domains to identify the optimum value of parameters.

Considering the intelligent behavior of crows which are among the most intelligent birds, Askarzadeh (2016) proposed an original Crow Search Algorithm (CSA). The important advantages of CSA are the simple implementation and setting a few parameters. To overcome the difficulty of MARS model, a metaheuristic optimization technique, CSA, is employed in this study to optimize the three aforementioned parameters of the MARS model applied for the TDS prediction.

Due to seasonal data and non-linearity of time-series records, feeding the raw metadata directly to the model may not provide significant insights for water quality parameter estimation. More often, a data pre-processing technique is recommended to enhance model fidelity and performance. Various strategies have been proposed in order to extract embedded features in dynamical and non-stationary time series signal including streamflow (Rezaie-Balf & Kisi, 2018), evaporation (Ghaemi et al.,

2019; Yaseen et al., 2020), rainfall (Ouyang et al., 2016; Wu & Chau, 2013), solar ultraviolet index, groundwater level (Rezaie-Balf, Naganna, et al., 2017; Roshni et al., 2020), wind power (Niu & Wang, 2019), water quality parameters (Fijani et al., 2019), soil moisture (Prasad et al., 2019). More recently, intrinsic time-scale decomposition (ITD) as a new noise assisted data analysis technique is proposed to decompose input/output variables with a few proper rotation components (PRCs) that change non-stationary signals into stationary statues (Martis et al., 2013). Interestingly, ITD is fully data-dependent; thereby making the tool significantly robust for relevant feature extraction without any loss of information.

The scope of this study is to develop a multi-step supervised machine learning evolutionary algorithm (MSMLEA) for predicting TDS. The focus is on using various physicochemical parameters to predict TDS at the Rig-Cheshmeh and the Soleyman-Tangeh Rivers in Iran. The main contributions of the research are as follows:

(1) The CSA optimization technique is used to determine the optimum value of the hyper-parameters of the MARS model and avoiding trial and error procedure. This model is developed with automated workflow and settings without human intervention.
(2) Presenting an accurate and stable formula for TDS using physicochemical parameters and comparing it with Ghavidel and Montaseri's empirical equation at both aforementioned stations.
(3) To convert non-stationarity and non-linearity time series to stationary ones, ITD was recruited and MSMLEA is proposed to predict monthly TDS records.
(4) By evaluation metrics and several visual plots, experimental outcomes indicated that the proposed MSMLEA method can provide better prediction accuracy compared to several traditional equations and models. This study is the first attempt, known to the authors, that combine MARS, CSA, and ITD models for the TDS prediction. Thus this study has the potential to fill a significant research gap in TDS simulation based on intelligent techniques.

This paper is organized as follows. In Section 2, the case study and data are explained. In addition, the procedures, algorithms, and the functionality of proposed models are introduced and discussed in this section. Data screening and analysis is carried out in Section 4. Section 5 discusses the implementation and case studies. The conclusion is provided, in Section 6.

2. Material and methods

2.1. Case study and sampling locations

The case study in this research is Tajan River basin, located in Mazandaran, Iran. Tajan Basin (53° 56' – 36° 17' north latitude and 53° 7' – 53° 42' east longitude) passes through the urbanized region (the City of Sari), with roughly 4147.22 km^2 area (Ghanbarpour et al., 2013). The climate system of this catchment is dominantly humid; either cold and/or partially humid. The average area slope, river discharge, and annual rainfall are respectively 85%, 20 m^3/s (cubic feet per second), and 539 mm (Rezaie-Balf & Kisi, 2018). The lowest and the highest elevations of the Tajan basin are 26 and 3728 m, respectively. Brown soil covers about 90% of the forest surface. Alluvial soil, rendzina, colluvial soil, and ranker are the next widespread types (Talebi et al., 2014). The river is host to various agricultural, aquacultural/aquafarming, and industrial activities and operations such as damming and sand mining, as the average amount of measured TDS is directly depended to those processes. A big dam has been constructed in the past to separate up- and downstreams of the river (Shahid-Rajaie Dam). So, this parameter should be monitored twice a year; the first one in fall and winter since the rate of rainfall is relatively high and the second one when active season of agriculture is coming. Currently, there are nine active hydrometric gaging stations in the basin and this research used Rig-Cheshmeh and Soleyman Tange gages for addressing TDS modeling assessment (see Figure 1). The characteristics and climatic and physical parameters of the basin are presented in Tables 1 and 2. According to Table 2, among input and output variables, TDS showed the maximum amount of concentration at two proposed stations (Rig-Cheshmeh (1270) and Soleyman-Tangeh (650)). Moreover, standard deviation (Sx) value computed for this parameter indicated that the Sx values of TDS records were spread over a wider range of values compare to input variables. The WQPs data are obtained from the Meteorological Organization of Mazandaran Province (MOMP). It is undeniable that there are a large number of variables which have significant influences on the TDS estimation. For example, Ghavidel and Montaseri (2014) selected Bicarbonate (HCO3), Calcium (Ca), Sodium (Na), Magnesium (Mg), and river discharge as input variables to estimate TDS. Asadollahfardi et al. (2016) selected HCO3, pH, Na, Mg, carbonate (CO3), Ca, and chloride (Cl) as input variables for the TDS study. Barzegari-Banadkook et al. (2020) considered the Na, HCO3, Mg, Ca, Cl, and sulfate (SO4) for the TDS study. In this study, Bicarbonate (HCO3), Calcium (Ca), Sodium (Na), Magnesium (Mg) were considered

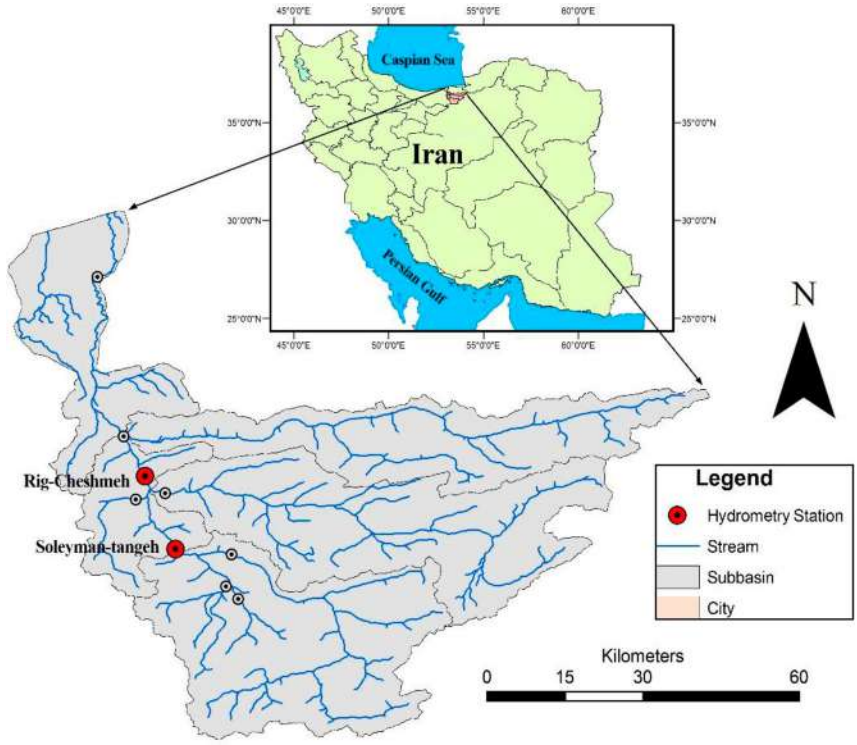

Figure 1. Location map of the study sites at the Tajan basin.

Table 1. The location and characteristics of selected gaging stations across the Tajan basin.

Station	Sub-basin	Latitude (N)	Longitude (E)	Altitude (m)	Number of samples
Rig-Cheshmeh	Tajan	36.22	53.10	240	505
Soleyman-Tangeh	Dodangeh	36.15	53.13	400	390

Table 2. Monthly values of statistical indices for the study sites located at the Tajan basin.

Station	Variable	Min	Mean	Max	S_x	C_v	C_{sx}
Rig-Cheshmeh	Hco_3 (mg/L)	1.60	3.88	12.2	0.89	0.79	2.05
	Ca (mg/L)	1.1	3.16	7.5	0.68	0.46	0.55
	Mg (mg/L)	0.1	2.17	6	0.69	0.48	0.39
	Na (mg/L)	0.2	1.54	6.50	0.75	0.57	1.86
	TDS (mg/L)	271	446.49	1270	78.7	6194.38	2.85
Soleyman-Tangeh	Hco_3 (mg/L)	1.2	3.84	7.70	0.91	0.83	0.55
	Ca (mg/L)	1.2	3.41	6.3	0.66	0.44	0.07
	Mg (mg/L)	0.5	2.07	4.5	0.68	0.46	0.29
	Na (mg/L)	0.08	0.87	2.94	0.42	0.18	1.75
	TDS (mg/L)	156	408.87	650	63.1	3981.8	0.46

Note: S_x, C_v, and C_{sx} denote the standard deviation, variation coefficient, and skewness coefficient, respectively.

as the input variables to predict TDS. Monthly time series of Bicarbonate (HCO3), Calcium (Ca), Sodium (Na), Magnesium (Mg), and TDS were obtained for March 1974–August 2016 and March 1984–August 2016 at Rig-Cheshmeh and Soleyman-Tangeh gauging stations, respectively. Approximately 75% and 25% of the datasets were used for training and testing periods, respectively.

2.2. Multivariate adaptive regression splines (MARS)

Friedman (1991) introduced one form of non-parametric regression analysis that is called a multivariate adaptive regression spline. In this technique, there is not any assumption of basic function (BF) regarding independent and dependent variables; thereby the segment's

endpoints (nodes) can estimate the endpoint of each region (Kim et al., 2019).

One of the abilities of this method is the splines, which cause increasing the performance model and considering linear function deviations (e.g. curvatures and thresholds). The adaptive algorithm is selected to determine the position of nodes. Suppose y as a deterministic output is a function of the input variable X ($X = (X_1, \ldots, X_p)$). Hence, y is provided as follows (Najafzadeh & Ghaemi, 2019; Yilmaz et al., 2018),

$$y = f(X_1, \ldots, X_p) + e = f(x) + e \tag{1}$$

where e is defined as error distribution. Basic functions containing piecewise-cubic and piecewise-linear functions, that help the model to calculate f function accurately. Piecewise-linear function is a kind of max $(0, x–t)$, where a node is suited at the value t. max (.) indicates the positive part of (.) is only used and otherwise, it is equal to zero (Zhang & Goh, 2016).

$$\max(0, x - t) \begin{cases} x - t & \text{if } x \geq t \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

MARS is a combination of linear BFs and their mutual relationships which is given by Equation (3) expressed by Rezaie-Balf et al. (2019).

$$f(x) = \beta_0 + \sum_{m=1}^{M} \beta_m \lambda_m(x) \tag{3}$$

Here $f(x)$ denotes the predicted response corresponding to the predictor variable x. Also, β_0 and β_m are the predicted constant coefficients (can be determined thorough least-squares technique) in order to attain the best data fit; BF is represented by $\lambda_m(x)$ and M is the number of basis functions. The MARS model is taken into account as a data-driven technique that is firstly conducted based on the calibration dataset. By cutting off the β_0 and basis pair, a model with a significant reduction of calibration error is built. The next pairs are then added to the model based on the M BFs (Zhang & Goh, 2016),

$$\hat{\beta}_{M+1}\lambda_1(X)\max(0,\ X_j - t) + \hat{\beta}_{M+2}\lambda_1(X)\max(0,\ \mathrm{t} - X_j) \tag{4}$$

where the estimation of β is performed by means of LS approach. Mutual interplay for the basic functions in that model is carefully selected when the new BF is added to the space of the model. Hence, BFs are added on the model for achieving the maximum special term numbers that cause an appropriate fitness model. Afterward, backward elimination discipline is recruited for reducing the term numbers. The major aim removing process is to find a closest to the optimum model thoroughly omitting the inessential variables. In the backward process, for selecting the proper sub-model, the lowermost effective BFs are eliminated. Therefore, the remaining BFs in the optimal model, is utilized in the initial step. More significantly, to compare model subsets generalized cross-validation (GCV) represented by Equation 5 is applied as a less computationally expensive function (Ghaemi et al., 2019; Sharda et al., 2008).

$$\mathrm{GCV} = \frac{\mathrm{MSE}}{\left[1 - \frac{N+\mathrm{dN}}{M}\right]^2} \tag{5}$$

where M and N are, respectively the number of observations and basic functions and d is the penalty of BFs.

2.3. Crow search algorithm (CSA)

Among the category of birds and animals, crows are the most intelligent birds with a wide brain compared to their body. They have significant ability to use tools, memorize faces, communicate in sophisticated ways, and hide and retrieve food during different seasons. Crows's features cause them to be able to find the hidden food places of other crows and steal them in their absence. If a crow recognizes that it is being chased by another one, it flies to another place to mislead the pursuer. According to this strategy, Askarzadeh (2016) suggested CSA as an evolutionary algorithm to solve a wide range of problems based on the following roles:

(1) The living of crow is as the flock form
(2) They have more ability in memorizing the hidden places of their food
(3) The pursuit each other for stealing their own hidden foods
(4) Crows used a probability to protect their hidden foods from robbery

The optimization process of the crow search algorithm begins with a dimensional environment including several crows. Each crow number N with its position i at each iteration the search space is provided using a vector $x^{i,iter} = [x_1^{i,iter}, x_2^{i,iter}, \ldots, x_d^{i,iter}]$, where $i = (1, 2, \ldots, N)$ and $iter = (1, 2, \ldots, iter_{max})$, which the maximum iteration is shown by $iter_{max}$. At iteration $iter$, each crow can memorize the location of its hidden location (i) and save it in its mind as the best place that the memory of that crow is shown by $m^{i,iter}$. At each iteration, two statuses can happen when crow j flies towards its hiding situation ($m^{j,iter}$), and crow i pursuit crow j for stealing the foods of crow j (Díaz et al., 2018; Gupta et al., 2018):

(1) If crow j cannot understand that it is being followed by crow i, crow i will find the hiding food place of crow j and it is defined as a new position of crow i as follows:

$$x^{i,\,iter+1} = x^{i,\,iter} + r_i \times fl^{i,iter} \times (m^{j,\,iter} - x^{i,\,iter}) \tag{6}$$

where r_i is a random number between 0 and 1, $fl^{i,iter}$ is the flight length for crow i at each iteration ($iter$).

(2) If crow j can understand that it is being followed by crow i, it flies to another place in the environment to protect its hiding food place.

In general, first and second conditions are summarized as below:

$$x^{i,\,iter+1} = \begin{cases} x^{i,\,iter+1} = x^{i,\,iter} + r_i \\ \times fl^{i,iter} \times (m^{j,\,iter} - x^{i,\,iter}) & r_j \geq AP^{i,\,iter} \\ \text{a random position} & \text{otherwise} \end{cases} \tag{7}$$

where $AP^{i,iter}$ is the amount of awareness probability for crow j at iteration $iter$. One of the main features of meta-heuristic algorithms is providing a permissible balance between diversification and intensification that this feature is performed by awareness probability (Askarzadeh, 2016; Mohammadi & Abdi, 2018).

2.4. Development of MARS using CSA

In computing science, choosing the best parameters is an important stage to attain well performance for machine learning techniques in modeling. Considering ANN as an example, the number of hidden layers and the number of hidden units (both discrete) or the weight and bias parameters can be prominent parameters in the ANN optimization. Various methods for finding appropriate parameters combine various experiences with a limited heuristic searching for possible optimal solutions which is time consuming for the users. In this regard, using a meta-heuristic algorithm (partial search algorithm) can ease the modeling processes (Rezaie-Balf et al., 2019) and may produce as the proper solution to an optimization problem, particularly with incomplete or imperfect information or limited computation capability.

Machine learning approach is highly dependent on maximum basis function ($M_{\max)}$, penalty parameter (d) and interaction (mi). But it is hard to select the optimum parameters in the MARS model simultaneously due to various choices, selecting the proper parameters can add to the MARS model fidelity. Focusing on various modeling procedures, we aim to integrate MARS with CSA (MARS-CSA) to make this complex problem easier to encounter (Figure 3). At the first stage, MARS addresses the basic function. New MARS-es are designed afterwards for every CSA-produced parameter values and the model quality is compared with the CSA's greedy selector regarding fitness function evaluation. Finally, the fitness function has been evaluated by the following objective function;

$$f = E_{\text{calibration}} + E_{\text{validation}} \tag{11}$$

where $E_{\text{calibration}}$ is the error at the calibration stage and $E_{\text{validation}}$ the error at the validation stage. Root means square error (RMSE) in the above equation is defined as the prediction error index. Moreover, the fitness function indicates the trade-off between model complexity and model generalization. It also appears that over-fitting in models arises from good training data fitness so that the error combination of calibration error and validation can build on a model that balances the minimum calibration error. Secondly, CSA begins searching for finding the most suitable parameter setting values, including $M_{\max}$, mi, and d. Once the convergence criterion is satisfied, the optimization process is terminated. This study used the generation number as the convergence criterion to reach certain iteration numbers. After performing the convergence criterion (and finishing calibration), the best predictive model containing the optimal setting of parameters with the best parameter settings is found and it is ready to apply for the validation dataset.

This research utilized 'ARESLab', which is an open-source code, developed by Jekabsons (2011), and CSA (Askarzadeh, 2016) for the evolutionary MARS-CSA model design. According to the MARS-CSA model, the best values of the three mentioned parameters of the model for both stations are described in Table 5. The maximum BFs number and maximum interaction level for the TDS prediction were 24 and 2, respectively, and the pairwise BFs products are permissible can be allowed (second-order interaction). At last, Finally, 16 piecewise-linear BFs at Rig-Cheshmeh and 8 piecewise-linear BFs at Soleyman-Tangeh stations were found, all with containing the intercept term, were found to achieve the best model, respectively, at Rig-Cheshmeh and Soleyman-Tangeh stations. The details of the BFs for both stations are presented in Table A1 in Appendix 1 shows the details of the BFs for both stations for prediction of monthly TDS. In addition, 10-fold cross-validation helped eliminate the possibility of performance bias10-fold cross-validation was applied to prevent model performance bias. Analysis of variance (ANOVA) decomposition has been employed in high-dimensional methods for training dataset in order to select important variables and interactions between them their interactions in high-dimensional methods. Consequently, the ANOVA

decomposition for the MARS-CSA model in TDS predicting has been carried out (see Table A2 in Appendix 1). According to Table A2, the GCV lists GCV value for the proposed model with all BFs (Table A2) for the specific ANOVA function removed to indicate the significance of the corresponding ANOVA function indicating the significance of the corresponding ANOVA function, by listing the GCV value for the proposed model with all BFs (Table A2) for the specific ANOVA function removed. Finally, the MARS model equations for both stations were computed as follows:

$$\begin{aligned}\mathrm{TDS}_{\text{Rig - Cheshmeh}} &= 647.6 - 54.921 \times \mathrm{BF1} + 73.774 \times \mathrm{BF2} \\ &\quad - 33.214 \times \mathrm{BF3} + 33.002 \times \mathrm{BF4} \\ &\quad - 9.1184 \times \mathrm{BF5} + 113.19 \times \mathrm{BF6} \\ &\quad - 58.14 \times \mathrm{BF7} - 12.408 \times \mathrm{BF8} \\ &\quad - 95.847 \times \mathrm{BF9} + 10.382 \times \mathrm{BF10} \\ &\quad + 39.116 \times \mathrm{BF11} + 6.1189 \times \mathrm{BF12} \\ &\quad + 10.192 \times \mathrm{BF13} - 16.932 \times \mathrm{BF14} \\ &\quad - 33.066 \times \mathrm{BF15} + 62.001 \times \mathrm{BF16}\end{aligned}$$

$$\begin{aligned}\mathrm{TDS}_{\text{Soleyman - Tangeh}} &= 283.82 + 50.373 \times \mathrm{BF1} - 71.797 \times \mathrm{BF2} + 64.492 \\ &\quad \times \mathrm{BF3} + 395.18 \times \mathrm{BF4} - 133.12 \times \mathrm{BF5} \\ &\quad - 265.7 \times \mathrm{BF6} + 183.62 \times \mathrm{BF7} - 125.56 \times \mathrm{BF8}\end{aligned}$$

2.5. Intrinsic time-scale decomposition

In 2007, Frei and Osorio introduced ITD as a time–frequency indicator for non-stationary, complicated time series assessment. To categorize the datasets, Proper Rotation Components (PRCs) functions are used. ITD as one of the decomposition-based methods is an EMD improvement, which effectively processes nonlinear and non-stationary signals with many successful applications in hydrological modeling (see Frei & Osorio, 2007; Guo et al., 2014; Martis et al., 2013).

ITD process technique includes four steps with an operator L that, from in input signal $x(t)$, generates the baseline signal. This causes a precise rotation and a lower frequency in residuals (Frei & Osorio, 2007). $Lx(t) = Lx(t)$ denotes the signal mean, expressed as $L(t)$. The PRCs are selected as $Hx(t) = (1 - L)x(t)$, presented as $H(t)$. The input signal $x(t)$ is then decomposed as (according to Martis et al., 2013):

$$x(t) = Hx(t) + L(t) = (1 - L)\, x(t) \tag{12}$$

The steps to develop an ITD algorithm proceed as follows:

(1) Determining τ_k, which is the corresponding occurrence time, and $x(t)$, which is the extreme points of the input signal, in which $k = 0, 1, 2, \cdots$ the first signal would be $\tau_0 = 0$.
(2) Supposing the input signal $x(t)$ in the interval of 0 and $\tau_k + 2$ and $L(t)$ and $H(t)$ as operators over the time interval $[0, \tau_k]$ that the baseline-providing operator L is considered as linear function on the interval $[\tau_k, \tau_k + 1]$. The baseline extraction operator is:

$$Lx(t) = L(t) = L_k + \left(\frac{L_{k+1} - L_k}{x_{k+1} - x_k}\right)(x(t) - x_k), \quad t \in (\tau_k, \tau_{k+1}), \tag{13}$$

and

$$L_{k+1} = \alpha\left[x_k + \frac{(\tau_{k+1} - \tau_k)}{\tau_{k+2} - \tau_k}(x_{k+1} - x_k)\right] + (1 - \alpha)x_{k+1} \tag{14}$$

where α *is* a constant value between 0 and 1 and taken as a fixed value ($\alpha = 1/2$).
(3) Applying an operator function to extract PRCs:

$$H(t) = Hx(t) = x(t) - L(t) = x(t) - L(t) \tag{15}$$

The principal purpose of ITD is to integrate the highest signals into some PRCs. It is clear in Equation 15 that PRCs can be achieved if the baseline is subtracted from the input signal. In general, ITD has various advantages which can be summarized in different concepts namely providing the transient smoothing, solving the smearing in time-scale space, and constant sifting, and time-saving of computation.

(1) Repeating the process for Equations 13 and 14 iteratively until the baseline $L(t)$ changes to a monotonous function that the single signal is divided into PRCs.

$$x(t) = \sum_{i=1}^{p} H^i(t) + L^p(t), \tag{16}$$

where p represents the number of obtained PRCs.

2.6. Description of the MSMLEA prediction model

Providing an estimation of the TDS by physicochemical input variables at a disparate natural stream is the preliminary goal of ITD-based MLMs. Figure 2 shows the workflow and procedure about how we developed and implemented the MSMLEA for TDS. Before beginning three

important steps of approaches based on decomposition, TDS and other physicochemical measurements during a one-month period were gathered and put into two distinct groups, training and validation periods and the appropriate model is determined independently from the training period. The randomness of the applied dataset and the parameter numbers play crucial roles in computing the number of data points (Fijani et al., 2019). In this study, random data variations indicated that a suitable technique with an adequate number of accessible observations could be predicted. Following a study by Rezaie-Balf et al. (2019), the accuracy of the suggested procedures was improved through the following 3 prominent steps:

Step 1: ITD method is applied to decompose both input and output datasets into some PRCs and a remaining component.

Step 2: The MSMLEA is proven as a robust TDS prediction tool for computing the decomposed PRC and calculating each component by means of the same sub-series

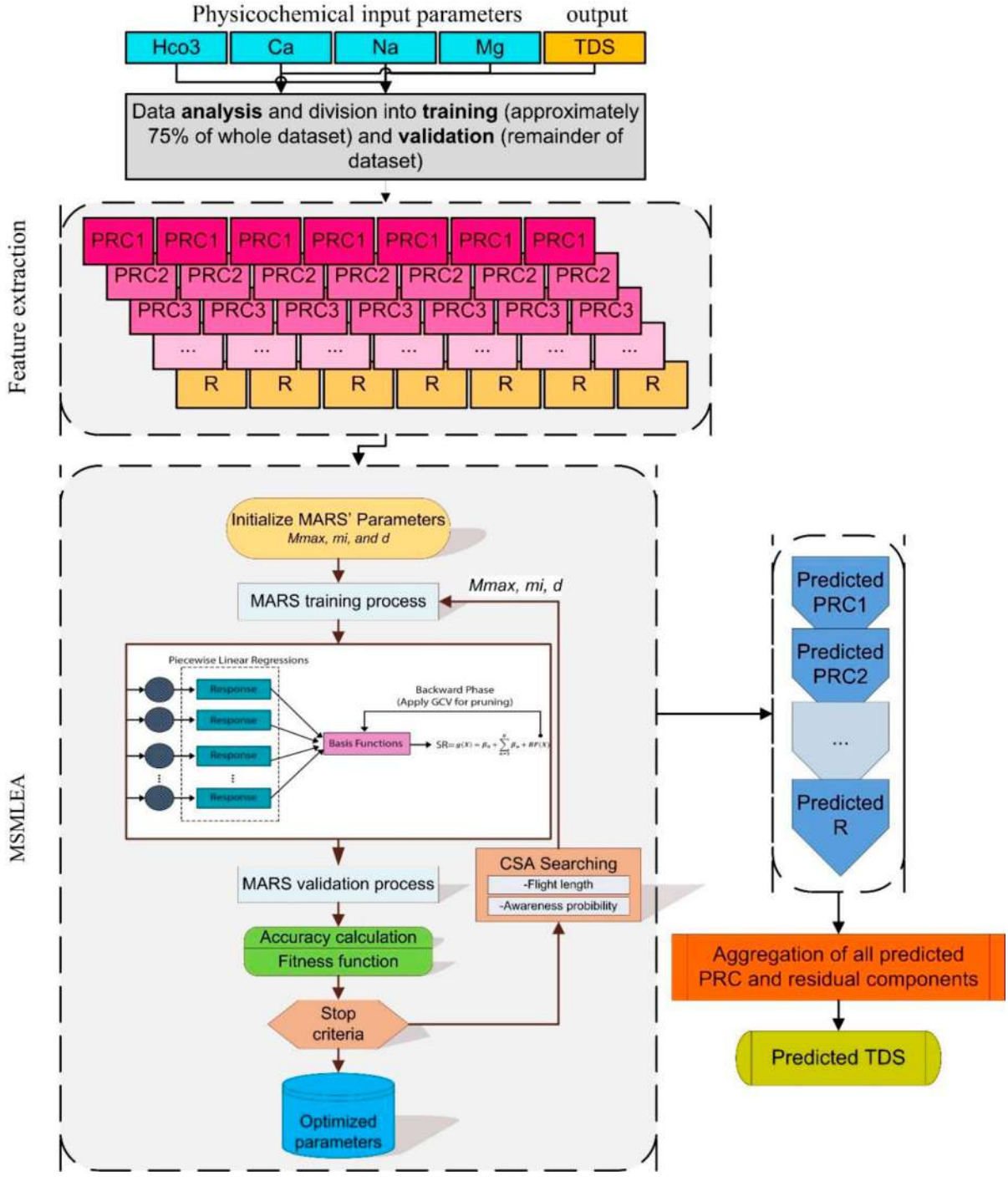

Figure 2. Workflow of the proposed MSMLEA.

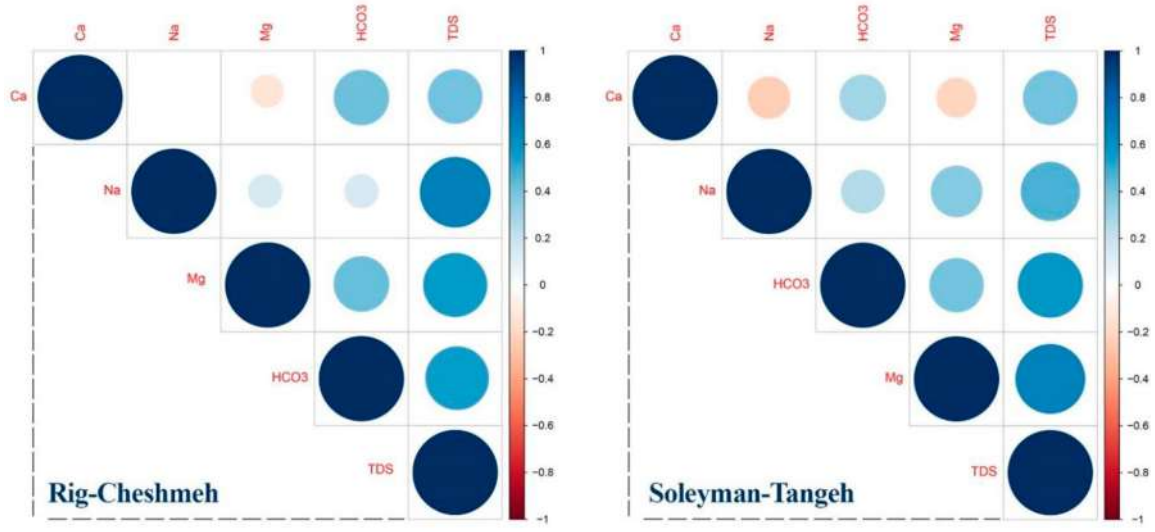

Figure 3. Linear relationship between each physiochemical parameter and the TDS using Pearson correlation matrix.

(PRC1) and the residue component of input variables, respectively.

Step 3: To produce the TDS value, a combination is made from the estimated values of every extracted PRC and residual components by MSMLEA.

ummarily, the MLMs which are based on ITD (ITD-MARS-CSA) recommended the idea of 'decomposition and ensemble'. The ensemble can produce a consensus formula to predict the original datasets; while the decomposition is a proper tool to make the estimating method easier.

2.7. Statistical analysis and performance assessment

2.7.1. Physiochemical–Covariate correlation

TDS co-variability with Hco3, Ca, Mg, and Na as physiochemical variables are investigated by the Pearson Coefficient that provides the dependency among several variables simultaneously. For evaluating the relationships among the datasets, the correlation factor which varies between -1 to $+1$ has been applied. In addition, the linear dependency between two variables for the Rig-Cheshmeh and the Soleyman-Tangeh stations is plotted as a graphical correlation matrix (Figure 3). As illustrated, the monthly TDS has a high correlation with monthly Na (0.68) for the Rig-Cheshmeh and Mg (0.65) for the Soleyman-Tangeh stations.

2.7.2. Statistical analysis of variance

Evaluating the dependent and independent variables is an important problem for data validation. One of these approaches is Analysis of variance (ANOVA) according to which modeler can use it to determine if there is any interaction among independent variables that may modulate the variability of the dependent variable (e.g. Lam et al., 2016). The GLM-ANOVA is one of the diagnostic tools that reduce the error variance overtime during the prediction period. In this study, the statistical significance of independent variables (Hco3, Ca, Na, and Mg) was set at 0.05.

The GLM-ANOVA was employed for each variable; the results are presented with a variable quantity, the sequential sum of squares, and the number of (in percents) independent variables given by the properties at two proposed stations.

The effect of the null hypothesis (i.e. the variances are equal) or significance test was defined to evaluate the effect of independent variables on the TDS variability at a probability level (p-value).

Table 3 shows that, by comparing the significance level factor (0.05), p-values provided the significance of independent variables.

The independent variables were all considered significant due to their p-value $\leq$ 0.05. Additionally, an evaluation was performed of the contribution of individual input variables for the above-mentioned stations. For the Rig-Cheshmeh, Na (84.16%) was the highest and Ca (71.35%) the lowest contributors. In contrast, at the Soleyman-Tangeh, Mg with 81.55% and Na with 73.85% has the highest and lowest contributions, respectively.

3. Results and discussion

To evaluate ML techniques, several performance metrics (see Appendix 2) are employed for evaluating the predictive performance criteria during calibration and validation periods.

Table 3. Analysis of variance (ANOVA) results.

		Statistical parameters					
Station	Source of Variation	DF	Seq. SS	Computed F	*P* value	Significance	Co. (%)
Rig-Cheshmeh	Na	125	64,158.4	6.03	0.00	Yes	84.16
	Ca	64	33,658.7	4.88	0.007	Yes	71.35
	HCO_3	68	140,513.1	5.23	0.0005	Yes	78.59
	Mg	59	64,152.9	9.35	0.00	Yes	81.56
	Error	188	88,913		-	-	-
Soleyman-Tangeh	Na	82	9063.16	3.41	0.0001	Yes	73.85
	Ca	38	13,035.48	4.33	0.004	Yes	69.42
	HCO_3	57	12,620.05	5.03	0.00	Yes	77.18
	Mg	38	40,672.63	9.98	0.00	Yes	81.55
	Error	175	1,552,902	–	–	–	–

Note: DF: degree of freedom; Seq. SS: Sequential sum of squares; Co.: Contribution.

3.1. Application and prediction outcomes

3.1.1. Rig-Cheshmeh station case study

This section discusses monthly TDS predictions for the Rig-Cheshmeh gauging station. The predictive ability of the standalone and integrated MARS models for the TDS prediction for both calibration and validation dataset presented concisely in Table 4.

Clearly, MSMLEA yielded better prediction (i.e. generally lowest RMSE, as well as the largest WI) compared to the rest of models. This indicates that intrinsic time-scale decomposition is a robust technique for performing non-stationary assessment and increasing the precision of the MARS-CSA model at this location during calibration and validation periods. For instance, the integrated ITD-MARS-CSA model provided the best performance compared to the rest of hybrid methods based on the performance criteria (NSE = 0.97 and WI = 0.992, lowest RMSE = 14.85, and RSD = 0.183). The ITD-MARS model stands next based on the same creations applied in this study.

The standalone MARS model and combined approaches such as MARS-CSA, ITD-MARS as well as MSMLEA used for the validation period (Table 4). According to this table, the evaluation metrics of the ITD-MARS-CSA model in terms of 95% uncertainty interval (95% confidence interval; 12.599), PMARE (2.51) and RSD (0.21) outperformed better while MARS-CSA with higher percentage error in case of RMSE (46.91%) and U95 (2.508) ranked second.

As mentioned above, among various models, i.e. Ghavidel and Montaseri (2014), ANN, GEP, ANFIS-GP, and ANFIS-SC, GEP (Eq.17) performed better for the TDS estimation in Zarinehroud basin compared to the rest of approaches.

$$\mathrm{TDS} = 91.2\mathrm{Na} + \mathrm{Na} - 14.5\mathrm{Ca}[\mathrm{HCO}_3 - (4.97 + \mathrm{HCO}_3)] + 2(\mathrm{HCO}_3)^2 + \mathrm{Mg} + (\mathrm{Mg})^{1/3} - \mathrm{Ca} \quad (17)$$

Based on Table 4, the GEP equation obtained by Ghavidel and Montaseri (2014) performed poorly with high error and uncertainty that made this model less capable of predicting the TDS records.

The goodness-of-fit and Pearson's Correlation Coefficients (R) values are presented as a scatterplot in Figure 4.

Table 4. Evaluation benchmarks of the proposed models for the calibration and validation periods at the Rig-Cheshmeh station.

	Statistical error indices				
Models	MARS	MARS-CSA	ITD-MARS	ITD-MARS-CSA	Ghavidel and Montaseri (2014)
Total available data in calibration period					
NSE	0.94	0.95	0.92	0.97	0.25
RMSE	19.49	17.17	24.4	14.85	70.01
RSD	0.24	0.21	0.301	0.183	0.86
U95	16.329	16.229	16.581	16.141	20.984
PMARE	2.84	2.49	4.6	2.62	13.02
WI	0.982	0.988	0.971	0.992	0.857
Total available data in validation period					
NSE	0.88	0.90	0.85	0.95	0.26
RMSE	21.26	19.76	23.88	13.45	53.79
RSD	0.338	0.314	0.37	0.21	0.85
U95	13.006	12.915	13.179	12.599	16.216
PMARE	3.038	2.98	4.75	2.51	10.09
WI	0.971	0.976	0.961	0.988	0.86

The scatterplots display the agreement between predicted and output variables and a least-squares regression (LSR) line and the determination coefficient (R^2) with a linear fit equation ($y = ax + b$) in each sub-panel. As presented, the gradient and b denotes intercept on the *y-axis* which is applied to outline the method's performance (Deo et al., 2016) based on the correlation coefficient (R^2).

As illustrated in Figure 4, ITD-MARS-CSA skillfully predicted the TDS values thereby it would be our recommended model for the Rig-Cheshmeh station. Figure 5 illustrates the time series of estimated and observed TDS for the entire calibration and validation records for the periods of March 1974 to August 2016 at the Rig-Cheshmeh station. Clearly, ITD-MARS-CSA (dotted orange line) proved to be the potential model to predict the TDS records whilst Ghavidel and Montaseri (2014) proposed model (the solid red line) underestimated the peak values indicating the poor performance of this model for this case study. Time required

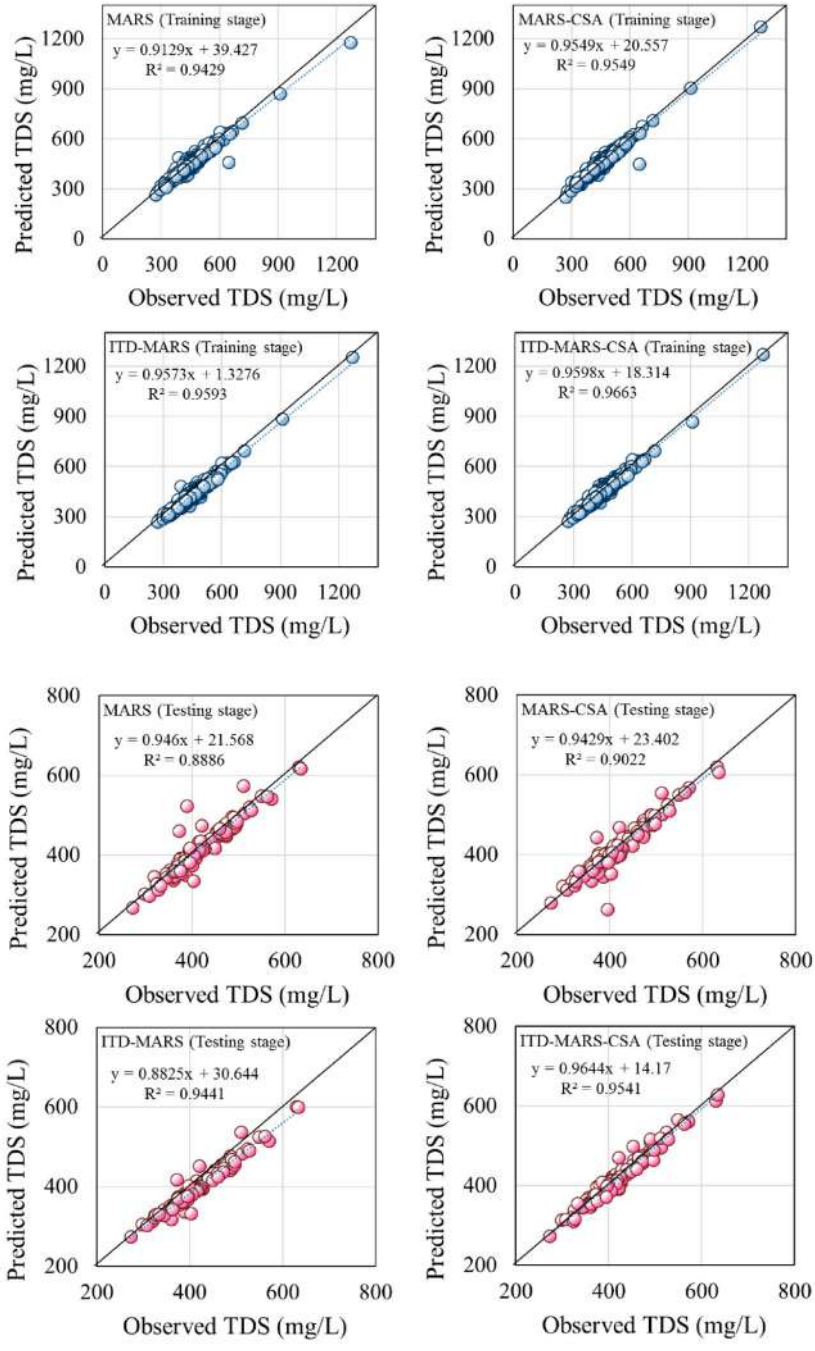

Figure 4. Scatter plots between observed and predicted TDSs at the Rig-Cheshmeh station in training and testing periods.

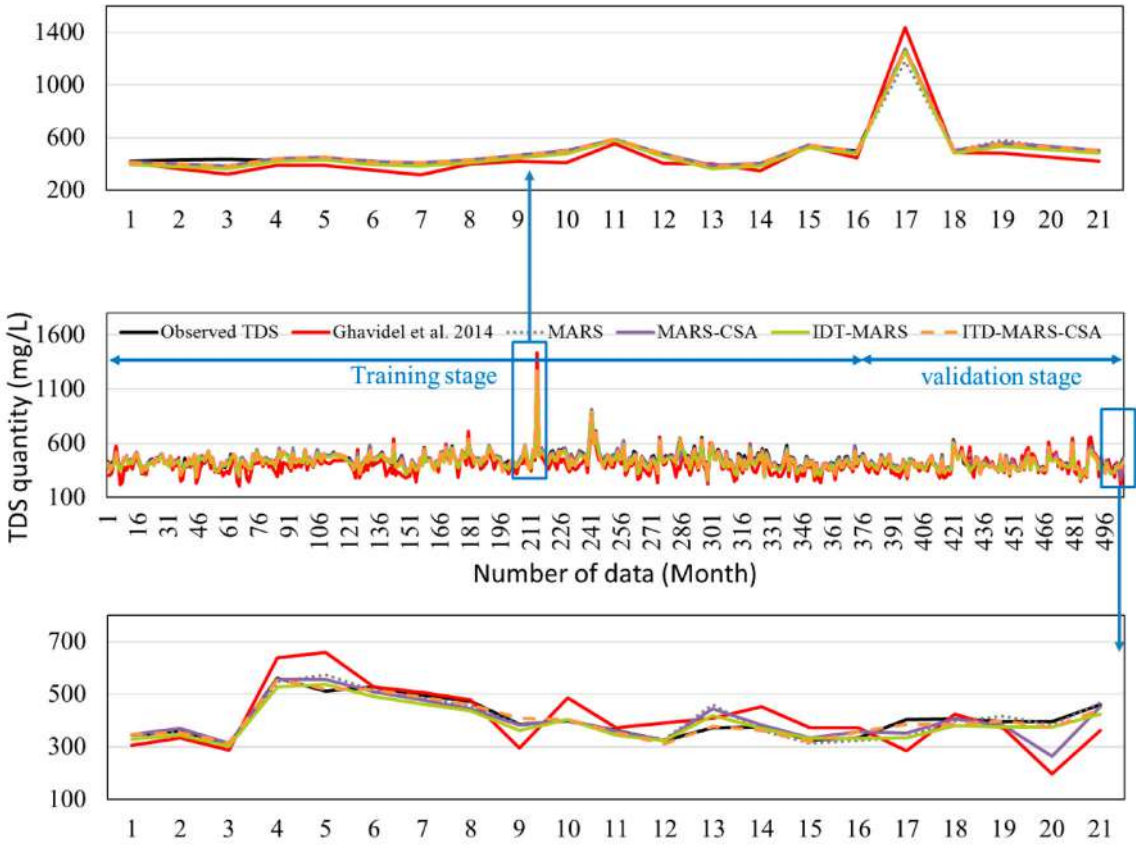

Figure 5. Monthly TDS predictions for training and validation periods at the Rig-Cheshmeh station.

to obtain the optimal solution for the proposed prediction problem with and without CSA optimization was computed for comparison purpose. For this station, the execution time using a laptop with 2.20 GHz Intel Core i7 4702MQ processor (8 GB RAM) was 0.58 and 0.64 s, respectively, for standalone MARS and MARS-CSA models.

3.1.2. Soleyman-Tangeh station case study

A similar evaluation was also performed for the Soleyman-Tangeh gauging station (Table 5). As perceived in Table 5, when we used the hybrid model (ITD-MARS-CSA) the performance improved significantly with respect to all metrics (NSE, RMSE, RSD, U95, PMARE, and WI).

Table 5. Evaluation metrics of the proposed models in the training and validation periods at the Soleyman-Tange gauging station.

	Statistical error indices				
Models	MARS	MARS-CSA	ITD-MARS	ITD-MARS-CSA	Ghavidel and Montaseri (2014)
Total available data in training stage					
NSE	0.899	0.91	0.902	0.92	0.36
RMSE (mg/l)	19.93	18.34	18.15	17.14	73.51
RSD	0.316	0.279	0.264	0.27	1.168
U95	12.939	12.725	12.551	12.784	18.968
PMARE	3.42	3.08	2.81	2.93	14.98
WI	0.972	0.975	0.978	0.98	0.72
Total available data in validation stage					
NSE	0.51	0.86	0.85	0.94	0.15
RMSE	29.81	15.45	16.02	9.72	45.48
RSD	0.71	0.36	0.37	0.22	1.071
U95	10.171	8.859	8.898	8.541	12.199
PMARE	6.76	3.43	3.48	2.28	8.91
WI	0.84	0.96	0.95	0.98	0.76

In the other hand, ITD-MARS-CSA relatively superior to the rest of predictive methods by achieving the lowest prediction error (RMSE = 17.14), and the highest predictive power (NSE = 0.92, and WI = 0.98) whereas *ITD-MARS* and MARS-CSA performed poorly with relatively high error and low predictivity. As expected, Ghavidel and Montaseri (2014) method performed poorly with significant differences in PMARE (14.98) and U95 (18.968). MARS model, on the other hand, showed unsatisfactory results for the WQI prediction during the calibration period.

In the validation period, among several predictive models the MARS-CSA and RSD proved to have the potential to predict TDS compare to the rest of the techniques. Comparing the performance of MARS-CSA and ITD-MARS-CSA models, the computed value of WI slightly increased from 0.96 to 0.98. Likewise, the magnitude of RMSE and RSD largely decreased by 9.72 (mg/l)

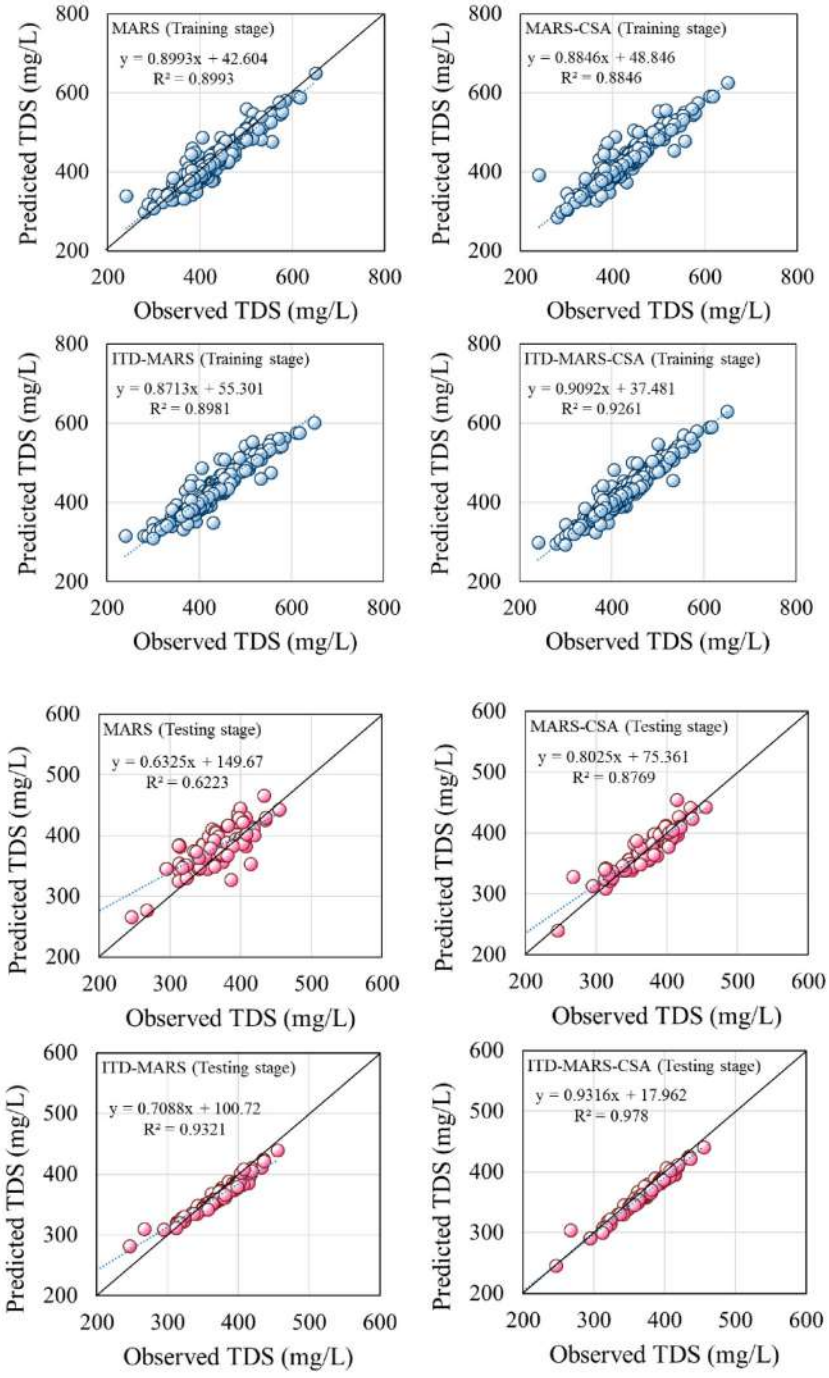

Figure 6. Scatter plots between the observed and predicted TDS at the Soleyman-Tangeh gauging station in training and validation periods.

and 0.22, respectively. Similar to the Rig-Cheshmeh gauging station, Ghavidel and Montaseri (2014) model provided unsatisfactory results with the highest error compared to the other ML methods. According to the above results, the ITD-MARS-CSA hybrid technique was the outstanding model because it combines the strengths and knowledge of time-scale decomposition, multivariate adaptive regression spline, and the CSA as a metaheuristic method.

In addition, scatter plots of predicted and observed TDS records for the Soleyman-Tangeh gauging station is presented in Figure 6. As illustrated, the slopes of the TDS values for the ITD-MARS-CSA method are closest to the best-fitting line although a number of TDS values are underestimated. In addition, MARS model was unable to estimate WQ parameter well compare to the other models which indicate the less capable of this model to the TDS prediction.

The prediction results of observed and predicted monthly TDS at the Soleyman-Tangeh station were similar to that of the Rig-Cheshmeh station. It should be noted that the proposed ITD-MARS-CSA exhibits the most accurate result than other approaches, in terms of both general tendency and estimating capacities of the TDS peak values (Figure 7). As shown previously, Equation (17) was poorly predicted the TDS values revealing that an empirical equation is less capable of predicting TDS variability. On the other hand, ML prediction can be more valuable and expressive than the outputs of empirical techniques.

3.2. Compression of the MSMLEA and the empirical equation

To compare the ITD-MARS-CSA with the empirical equation proposed by Ghavidel and Montaseri (2014), we analyzed the scatter plots and the validation results for the TDS prediction. As illustrated in Figure 8, TDS records predicted by the ITD-MARS-CSA showed more consistency with observations at both stations. Further, the trend of TDS records predicted by the ITD-MARS-CSA was remarkably similar to the observations and followed the same patterns. These results indicate that compared to empirical methods, a Multi-Step Supervised machine learning evolutionary algorithm can be used as a sophisticated and intelligent approach to deal with both nonstationary and trends in the data while putting more emphasis on predicting the TDS peak values well. The idea of ML algorithm is that a system can learn from data and can adapt to the patterns in the data. This could be integrated

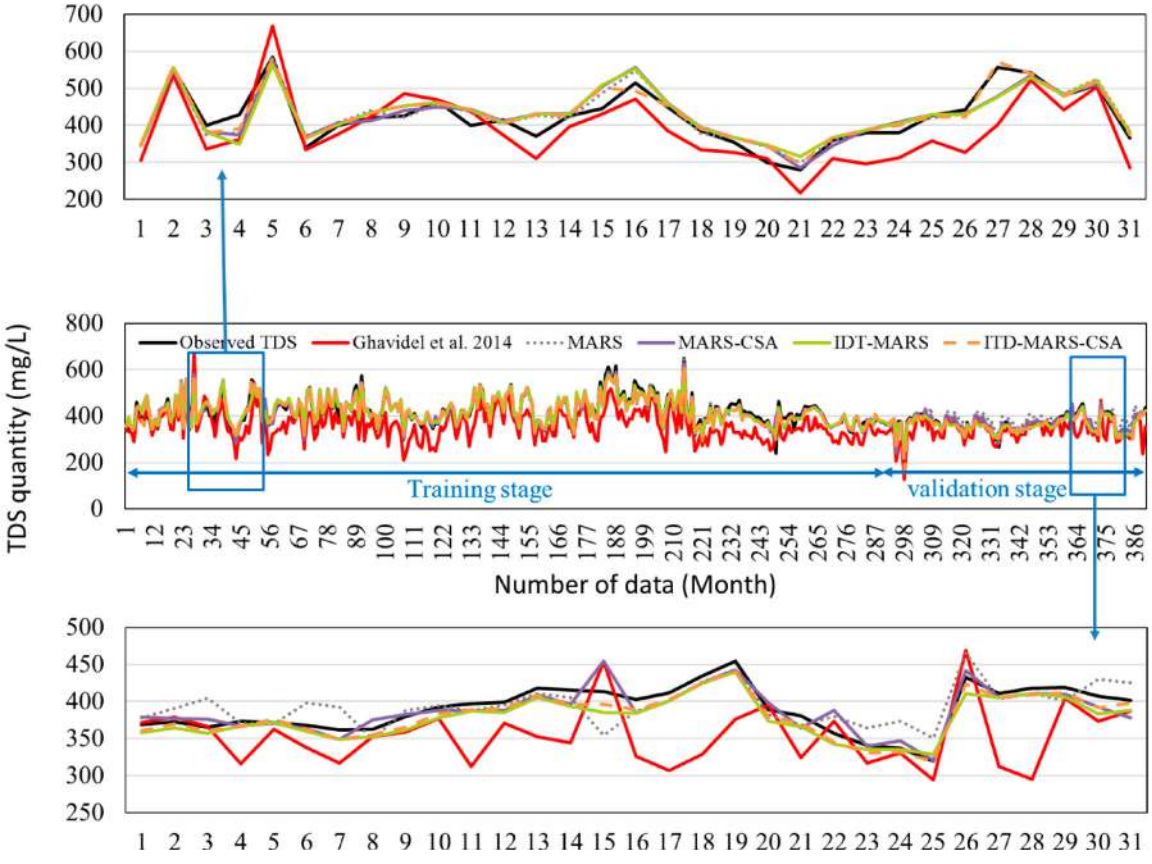

Figure 7. Monthly TDS predictions for training and validation periods at the Soleyman-Tangeh gauging station.

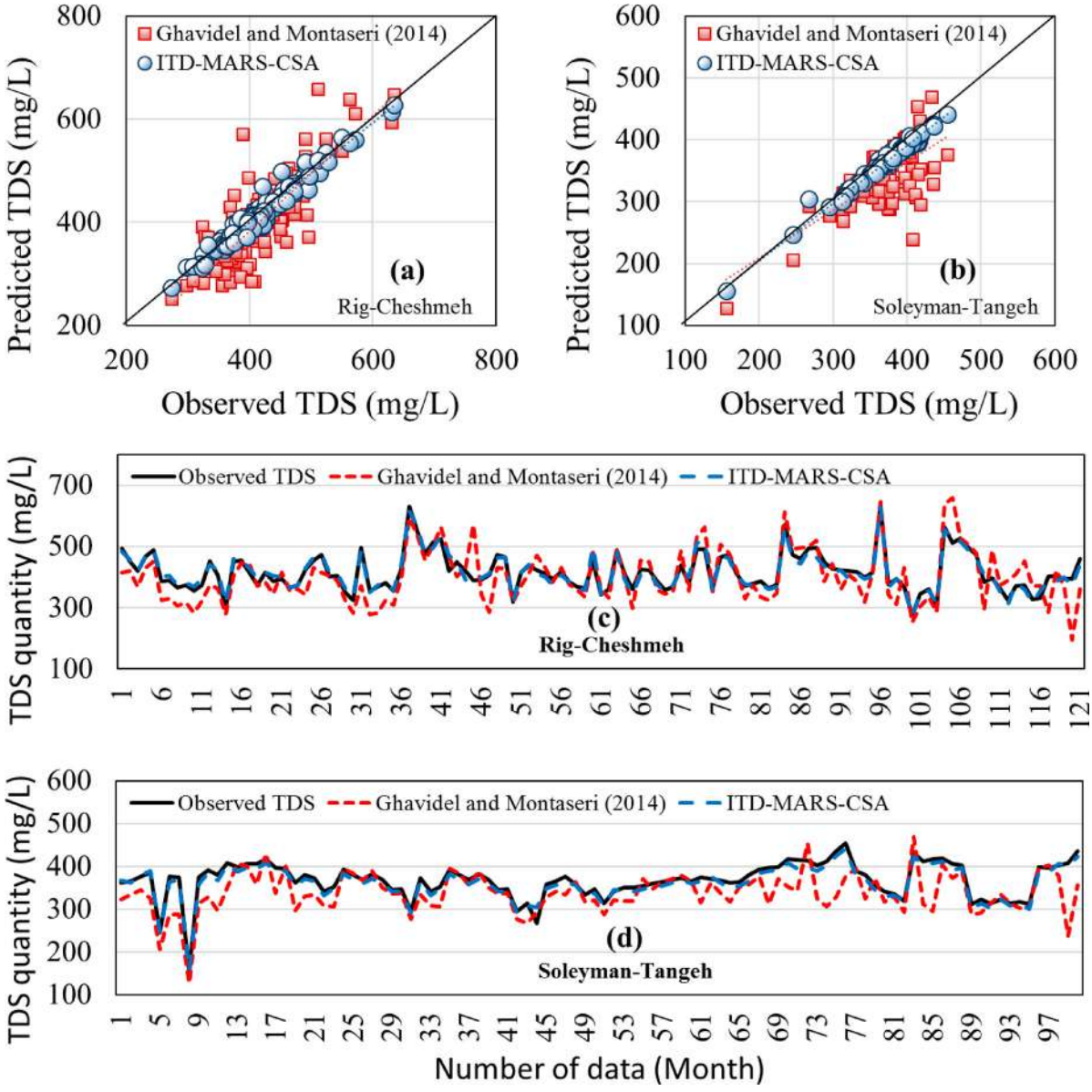

Figure 8. Comparison of ITD-MARS-CSA and Ghavidel and Montaseri (2014) for the TDS prediction using scatter plots (a, b) and hydrographs (c and d) for validation period.

with empirical or even physically-based models to make informed decisions for water resources systems.

In addition, error histograms of standalone (MARS) and hybrid models (MARS-CSA, ITD MARS, and ITD-MARS-CSA) as well as Ghavidel, and Montaseri equation were plotted for both stations in order to compute error concentrations (Figures 9 and 10). Interestingly, the density of errors is approximately scattered around zero for all models with the exception of the empirical approach. Among all models used in this study, the accuracy of the ITD-MARS-CSA approach is superior and it is a more robust algorithm compare to the rest of models. Times of execution for training and building the MARS and MARS-CSA models were almost similar, 0.76 s and 0.74 respectively. That is, in this research, models' runtime could not be compared for TDS prediction at both stations and it may occurred for a few number of independent (input) parameters.

As aforementioned above, for the standalone MARS model, it is often hard to fully reflect the information mechanisms of natural hydrological variables such as TDS accurately tied to a few resolution components that applied to establish the prediction models. This reveals that other resolution subcomponents in the original TDS time series cannot be separated effectively. To avoid this problem, decomposition methods can be proposed to select various resolution intervals and then the features of each subseries can be separated. As a

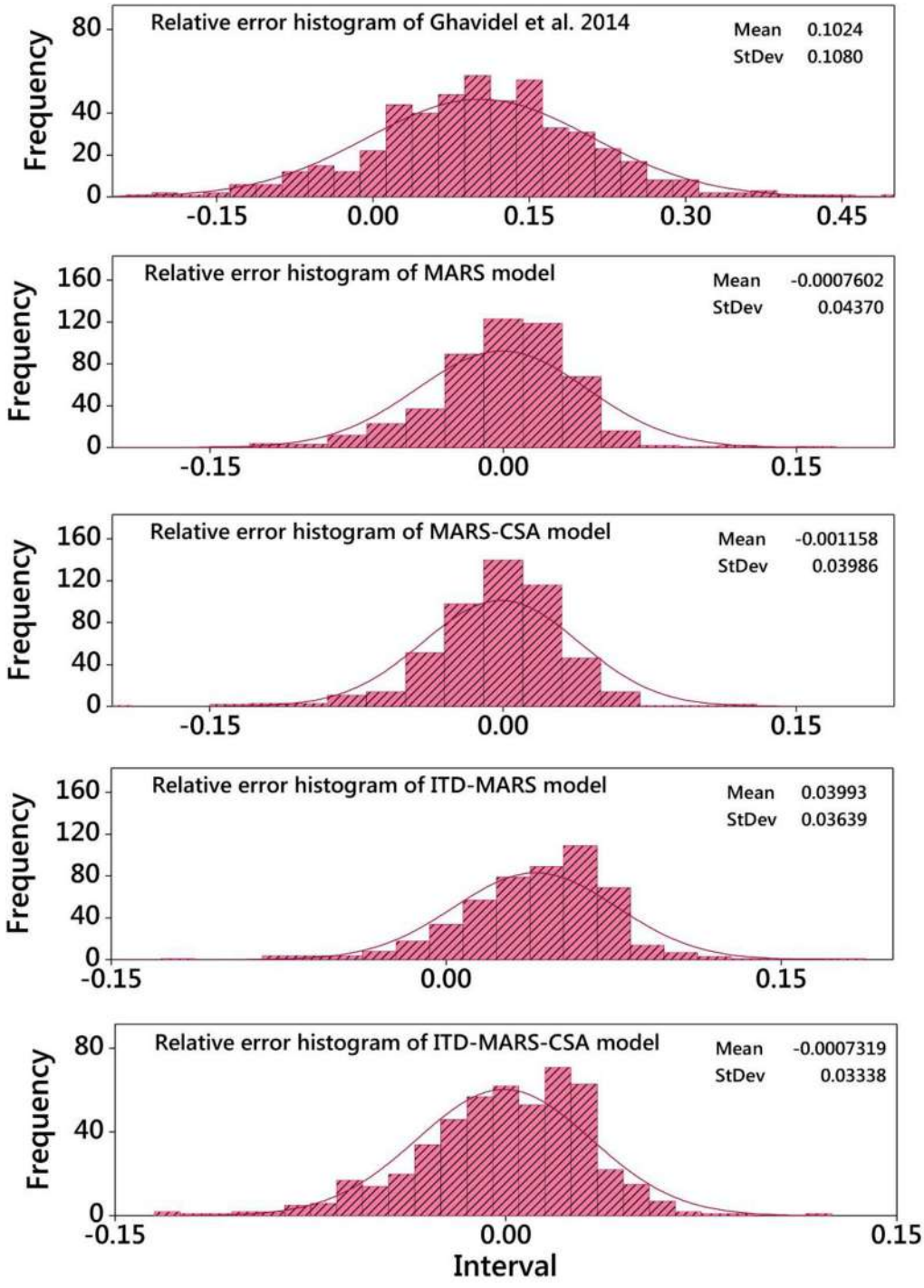

Figure 9. Relative error histograms for proposed standalone (MARS) and hybrid models (MARS-CSA, ITD-MARS, and ITD-MARS-CSA) as well as Ghavidel and Montaseri empirical equation at the Rig-Cheshmeh station.

result, the performances of the hybrid method (ITD-MARS) were outperformed to those of the standalone MARS model. However, MARS development is hugely dependent on maximum basis function (Mmax), penalty parameter (d), and interaction (mi). Although it was a challenge to select the optimum parameters simultaneously. Owing to the various choices, selecting the proper parameters may diminish the performance of MARS model. This may increase the error in simulation when the number of input variables increases that can be a prominent factor for decreasing the model accuracy (such as ITD-MARS). In this regards, developing a optimizing algorithm such as CSA could help find the best parameter setting values and improve the model

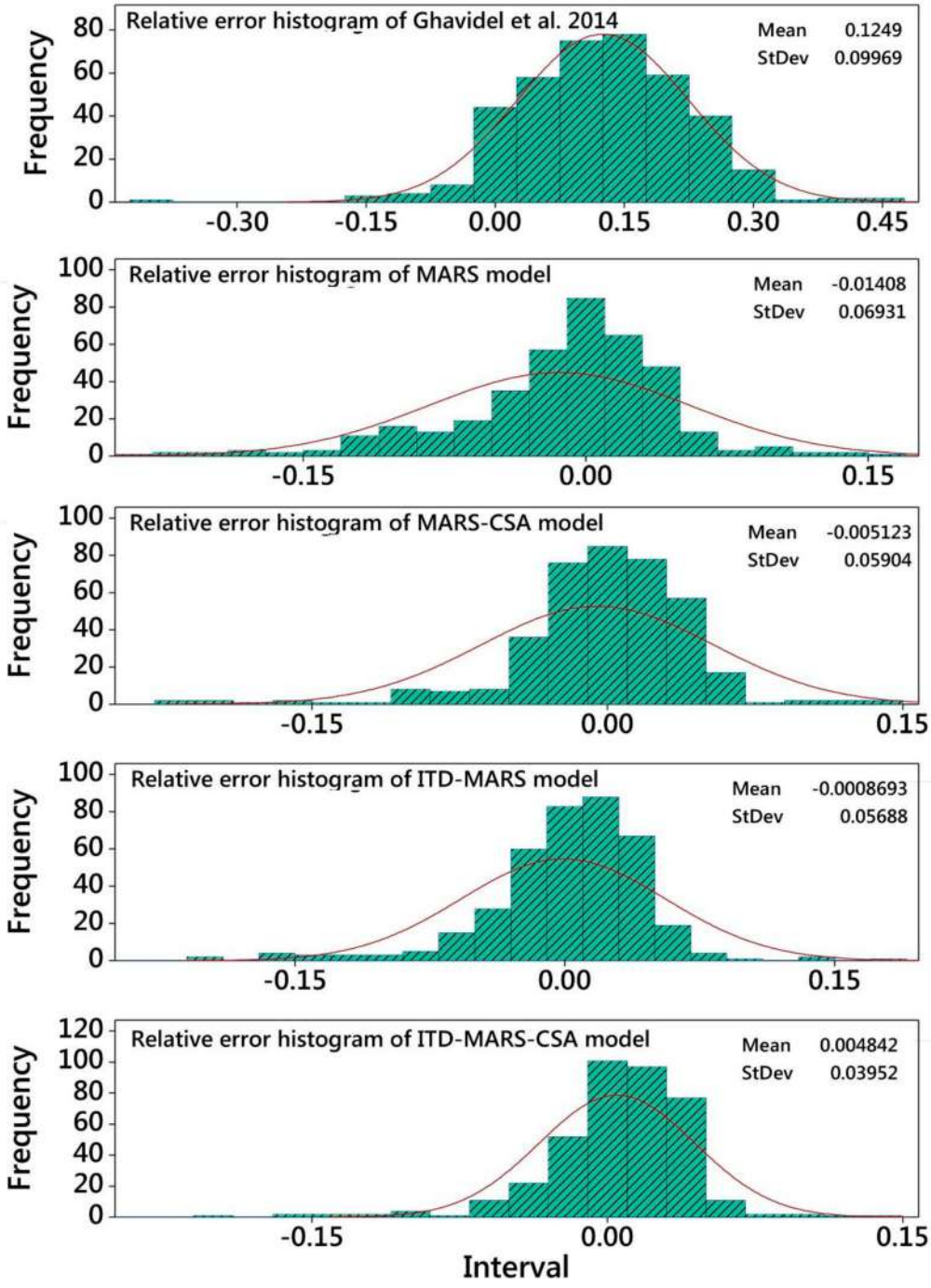

Figure 10. Relative error histograms for proposed standalone (MARS) and hybrid models (MARS-CSA, ITD-MARS, and ITD-MARS-CSA) as well as Ghavidel and Montaseri empirical equation at the Soleyman-Tangeh station.

accuracy. In this study, CSA with optimizing the parameter setting of ITD-MARS model could enhance the model accuracy.

The authors recommend the utilization of both decomposition and optimization-based methods for other TDS assessment with the same scale of input/output parameters as well as watershed physical characteristics in order to assess the generalization of the MSMLEA. Furthermore, nonlinear and/or dynamic ML programming based on simulation models could be used to find the contributors to the TDS in the river system, however, this type of assessment typically imposes a prohibitive computational burden, especially for large and complex river systems prediction.

It should be noted that the amount of data used to train a ML algorithm has a rather large impact on the accuracy of the prediction. There appears to be the expected improvement in prediction that as the size of the data

increased, the accuracy increased up to a significant level. This causes the model being more optimized and capable of predicting TDS variability over time. The outcomes of this research may assist in providing a range for how much data is needed to create an optimized model for water quality modeling system. The future of TDS modeling using machine learning algorithms seems to be very bright and remarkable with the continuous evolution of AI techniques that created (and likely will create) more intelligent and modern algorithms.

4. Conclusion

In this study, the capability of a multi-step supervised based machine learning approach incorporated with the evolutionary algorithm, MSMLEA, was evaluated for monthly TDS prediction at two stations, Rig-Cheshmeh and Soleyman-Tangeh, in Tajan River, Iran. This study focused on predicting the most influential water quality parameters such as Na, Ca, HCO3, and Mg. Analysis suggests that Na and Ca contributed, respectively 84.16% and 71.35% to the total dissolved solid for the Rig-Cheshmeh while Mg subsidized 81.55% to the Soleyman-Tangeh River. Comparing the results of the standalone and hybrid models revealed that ITD data-decomposition technique has a significant influence on models' accuracy. This approach can successfully decompose the dataset and solve the non-stationary associated with time series records. At the Rig-Cheshmeh and Soleyman-Tangeh gauging stations, the predicted TDS records were investigated in term of evaluation metrics. Comparing the performance of MARS-CSA and ITD-MARS-CSA, it is noted that the computed values of the WI increased. Whereas, the magnitude of RMSE and RSD also decreased significantly. On the other hand, the outcomes showed that MSMLEA was the accurate model with the help of data decomposing by the ITD algorithm. Besides, our proposed equation was compared with Ghavidel and Montaseri's empirical method. Results suggest that MARS's equation provided lower error than empirical method for the TDS prediction in terms of RMSE and PMARE at both stations.

Although the proposed model had an acceptable accuracy, it is possible to employed other evolution machine learning and modern algorithms and integrating them with pre-processing methods such as vibrational mode decomposition (VMD), complete ensemble empirical mode decomposition (CEEMD). This will create more accurate models in the WQPs prediction. With the aim of increasing the accuracy of TDS estimation, we recommend using large data samples with various input variables based on daily or hourly timescales. As the potential avenue for future research, the uncertainty associated with the input/output variables and models can be investigated to present more reliable predictive models. It can be considered how model input and parameter uncertainty may affect the TDS prediction results. As the final suggestion and limitation of the present research, other WQPs and hydrological parameters such as rainfall, temperature, and river discharge can be fed to the model as input layers to better compute the TDS variability and patterns, particularly during low and high flow events.

Acknowledgements

The authors acknowledge and appreciate the Regional Water Organization of Mazandaran Province and Meteorological Organization of Mazandaran Province (MOMP) of Iran for giving us access to their meteorological data. We acknowledge the 'Open Access Funding by the Publication Fund of the TU Dresden'.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Amir Mosavi http://orcid.org/0000-0003-4842-0613

References

Abudu, S., King, J. P., & Sheng, Z. (2012). Comparison of the performance of statistical models in forecasting monthly total dissolved solids in the Rio Grande 1. *JAWRA Journal of the American Water Resources Association*, *48*(1), 10–23. https://doi.org/10.1111/j.1752-1688.2011.00587.x

Ahmed, A. N., Othman, F. B., Afan, H. A., Ibrahim, R. K., Fai, C. M., Hossain, M. S., & Elshafie, A. (2019). Machine learning methods for better water quality prediction. *Journal of Hydrology*, *578*, Article 124084. https://doi.org/10.1016/j.jhydrol.2019.124084

Alizadeh, M. J., Kavianpour, M. R., Danesh, M., Adolf, J., Shamshirband, S., & Chau, K. W. (2018). Effect of river flow on the quality of estuarine and coastal waters using machine learning models. *Engineering Applications of Computational Fluid Mechanics*, *12*(1), 810–823. https://doi.org/10.1080/19942060.2018.1528480

Anctil, F., Lauzon, N., & Filion, M. (2008). Added gains of soil moisture content observations for streamflow predictions using neural networks. *Journal of Hydrology*, *359*(3–4), 225–234. https://doi.org/10.1016/j.jhydrol.2008.07.003

Asadollahfardi, G., Meshkat-Dini, A., Homayoun Aria, S., & Roohani, N. (2016). Application of artificial neural networks to predict total dissolved solids in the river Zayanderud, Iran. *Environmental Engineering Research*, *21*(4), 333–340. https://doi.org/10.4491/eer.2015.096

Asadollahfardi, G., Zangooi, H., Asadi, M., Tayebi Jebeli, M., Meshkat-Dini, M., & Roohani, N. (2018). Comparison of Box-Jenkins time series and ANN in predicting total dissolved solid at the Zāyandé-Rūd River, Iran. *Journal of Water Supply: Research and Technology-Aqua*, *67*(7), 673–684. https://doi.org/10.2166/aqua.2018.108

Askarzadeh, A. (2016). A novel metaheuristic method for solving constrained engineering optimization problems:

Crow search algorithm. *Computers & Structures*, *169*, 1–12. https://doi.org/10.1016/j.compstruc.2016.03.001

Attar, N. F., Khalili, K., Behmanesh, J., & Khanmohammadi, N. (2018). On the reliability of soft computing methods in the estimation of dew point temperature: The case of arid regions of Iran. *Computers and Electronics in Agriculture*, *153*, 334–346. https://doi.org/10.1016/j.compag.2018.08.029

Barzegari-Banadkook, F. B., Ehteram, M., Panahi, F., Sammen, S. S., Othman, F. B., & Ahmed, E. S. (2020). Estimation of total dissolved solids (TDS) using new hybrid machine learning models. *Journal of Hydrology*, Article 124989. https://doi.org/10.1016/j.jhydrol.2020.124989

Bui, D. T., Khosravi, K., Tiefenbacher, J., Nguyen, H., & Kazakis, N. (2020). Improving prediction of water quality indices using novel hybrid machine-learning algorithms. *Science of the Total Environment*, Article 137612. https://doi.org/10.1016/j.scitotenv.2020.137612

Chen, X. Y., & Chau, K. W. (2019). Uncertainty analysis on hybrid double feedforward neural network model for sediment load estimation with LUBE method. *Water Resources Management*, *33*(10), 3563–3577. https://doi.org/10.1007/s11269-019-02318-4

Chen, H., Xu, L., Ai, W., Lin, B., Feng, Q., & Cai, K. (2020). Kernel functions embedded in support vector machine learning models for rapid water pollution assessment via near-infrared spectroscopy. *Science of The Total Environment*, *714*, Article 136765. https://doi.org/10.1016/j.scitotenv.2020.136765

Choubin, B., Moradi, E., Golshan, M., Adamowski, J., Sajedi-Hosseini, F., & Mosavi, A. (2019). An ensemble prediction of flood susceptibility using multivariate discriminant analysis, classification and regression trees, and support vector machines. *Science of the Total Environment*, *651*, 2087–2096. https://doi.org/10.1016/j.scitotenv.2018.10.064

Deo, R. C., Samui, P., & Kim, D. (2016). Estimation of monthly evaporative loss using relevance vector machine, extreme learning machine and multivariate adaptive regression spline models. *Stochastic Environmental Research and Risk Assessment*, *30*(6), 1769–1784. https://doi.org/10.1007/s00477-015-1153-y

Díaz, P., Pérez-Cisneros, M., Cuevas, E., Avalos, O., Gálvez, J., Hinojosa, S., & Zaldivar, D. (2018). An improved crow search algorithm applied to energy problems. *Energies*, *11*(3), 571. https://doi.org/10.3390/en11030571

Fijani, E., Barzegar, R., Deo, R., Tziritis, E., & Skordas, K. (2019). Design and implementation of a hybrid model based on two-layer decomposition method coupled with extreme learning machines to support real-time environmental monitoring of water quality parameters. *Science of the Total Environment*, *648*, 839–853. https://doi.org/10.1016/j.scitotenv.2018.08.221

Frei, M. G., & Osorio, I. (2007). Intrinsic time-scale decomposition: Time–frequency–energy analysis and real-time filtering of non-stationary signals. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *463*(2078), 321–342. https://doi.org/10.1098/rspa.2006.1761

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics*, 1–67.

Ghaemi, A., Rezaie-Balf, M., Adamowski, J., Kisi, O., & Quilty, J. (2019). On the applicability of maximum overlap discrete wavelet transform integrated with MARS and M5 model tree for monthly pan evaporation prediction. *Agricultural and Forest Meteorology*, *278*, Article 107647. https://doi.org/10.1016/j.agrformet.2019.107647

Ghanbarpour, M. R., Goorzadi, M., & Vahabzade, G. (2013). Spatial variability of heavy metals in surficial sediments: Tajan river watershed, Iran. *Sustainability of Water Quality and Ecology*, *1*, 48–58. https://doi.org/10.1016/j.swaqe.2014.04.002

Ghavidel, S. Z. Z., & Montaseri, M. (2014). Application of different data-driven methods for the prediction of total dissolved solids in the Zarinehroud basin. *Stochastic Environmental Research and Risk Assessment*, *28*(8), 2101–2118. https://doi.org/10.1007/s00477-014-0899-y

Guo, Z., Xie, L., Ye, T., & Horch, A. (2014). Online detection of time-variant oscillations based on improved ITD. *Control Engineering Practice*, *32*, 64–72. https://doi.org/10.1016/j.conengprac.2014.07.002

Gupta, D., Rodrigues, J. J., Sundaram, S., Khanna, A., Korotaev, V., & de Albuquerque, V. H. C. (2018). Usability feature extraction using modified crow search algorithm: A novel approach. *Neural Computing and Applications*, 1–11. https://doi.org/10.1007/s00521-018-3688-6

Hong, H., Panahi, M., Shirzadi, A., Ma, T., Liu, J., Zhu, A. X., & Kazakis, N. (2018). Flood susceptibility assessment in Hengfeng area coupling adaptive neuro-fuzzy inference system with genetic algorithm and differential evolution. *Science of The Total Environment*, *621*, 1124–1141. https://doi.org/10.1016/j.scitotenv.2017.10.114

Jekabsons, G. (2011). ARESLab: Adaptive regression splines toolbox for Matlab/Octave. http://www.cs.rtu.lv/jekabsons.

Jonnalagadda, S. B., & Mhere, G. (2001). Water quality of the Odzi River in the eastern highlands of Zimbabwe. *Water Research*, *35*(10), 2371–2376. https://doi.org/10.1016/S0043-1354(00)00533-9

Kargar, K., Samadianfard, S., Parsa, J., Nabipour, N., Shamshirband, S., Mosavi, A., & Chau, K. W. (2020). Estimating longitudinal dispersion coefficient in natural streams using empirical models and machine learning algorithms. *Engineering Applications of Computational Fluid Mechanics*, *14*(1), 311–322. https://doi.org/10.1080/19942060.2020.1712260

Khaki, M., Yusoff, I., & Islami, N. (2015). Application of the artificial neural network and neuro-fuzzy system for assessment of groundwater quality. *CLEAN–Soil, Air, Water*, *43*(4), 551–560. https://doi.org/10.1002/clen.201400267

Kim, S., Seo, Y., Rezaie-Balf, M., Kisi, O., Ghorbani, M. A., & Singh, V. P. (2019). Evaluation of daily solar radiation flux using soft computing approaches based on different meteorological information: Peninsula vs continent. *Theoretical and Applied Climatology*, *137*(1–2), 693–712. https://doi.org/10.1007/s00704-018-2627-x

Lam, T. C., Ge, H., & Fazio, P. (2016). Energy positive curtain wall configurations for a cold climate using the analysis of variance (ANOVA) approach. *Building Simulation*, *9*, 297–310. https://doi.org/10.1007/s12273-016-0275-6

Martis, R. J., Acharya, U. R., Tan, J. H., Petznick, A., Tong, L., Chua, C. K., & Ng, E. Y. K. (2013). Application of intrinsic time-scale decomposition (ITD) to EEG signals for automated seizure prediction. *International Journal of Neural Systems*, *23*(5), Article 1350023. https://doi.org/10.1142/S0129065713500238

Miranda, J., & Krishnakumar, G. (2015). Microalgal diversity in relation to the physicochemical parameters of some industrial sites in Mangalore, South India. *Environmental Monitoring and Assessment*, *187*(11), 664. https://doi.org/10.1007/s10661-015-4871-1

Mohammadi, F., & Abdi, H. (2018). A modified crow search algorithm (MCSA) for solving economic load dispatch problem. *Applied Soft Computing*, *71*, 51–65. https://doi.org/10.1016/j.asoc.2018.06.040

Mouatadid, S., Raj, N., Deo, R. C., & Adamowski, J. F. (2018). Input selection and data-driven model performance optimization to predict the Standardized precipitation and evaporation index in a drought-prone region. *Atmospheric Research*, *212*, 130–149. https://doi.org/10.1016/j.atmosres.2018.05.012

Mustafa, A. S. (2015). Artificial neural networks modeling of Total dissolved solid in the selected Locations on Tigris River, Iraq. *Journal of Engineering*, *21*(6), 162–179.

Najafzadeh, M., & Ghaemi, A. (2019). Prediction of the five-day biochemical oxygen demand and chemical oxygen demand in natural streams using machine learning methods. *Environmental Monitoring and Assessment*, *191*(6), 380. https://doi.org/10.1007/s10661-019-7446-8

Najafzadeh, M., Ghaemi, A., & Emamgholizadeh, S. (2019). Prediction of water quality parameters using evolutionary computing-based formulations. *International Journal of Environmental Science and Technology*, *16*(10), 6377–6396. https://doi.org/10.1007/s13762-018-2049-4

Najafzadeh, M., Rezaie Balf, M., & Rashedi, E. (2016). Prediction of maximum scour depth around piers with debris accumulation using EPR, MT, and GEP models. *Journal of Hydroinformatics*, *18*(5), 867–884. https://doi.org/10.2166/hydro.2016.212

Niu, X., & Wang, J. (2019). A combined model based on data preprocessing strategy and multi-objective optimization algorithm for short-term wind speed forecasting. *Applied Energy*, *241*, 519–539. https://doi.org/10.1016/j.apenergy.2019.03.097

Noori, N., & Kalin, L. (2016). Coupling SWAT and ANN models for enhanced daily streamflow prediction. *Journal of Hydrology*, *533*, 141–151. https://doi.org/10.1016/j.jhydrol.2015.11.050

Ouyang, Q., Lu, W., Xin, X., Zhang, Y., Cheng, W., & Yu, T. (2016). Monthly rainfall forecasting using EEMD-SVR based on phase-space reconstruction. *Water Resources Management*, *30*(7), 2311–2325. https://doi.org/10.1007/s11269-016-1288-8

Pan, C., Ng, K. T. W., Fallah, B., & Richter, A. (2019). Evaluation of the bias and precision of regression techniques and machine learning approaches in total dissolved solids modeling of an urban aquifer. *Environmental Science and Pollution Research*, *26*(2), 1821–1833. https://doi.org/10.1007/s11356-018-3751-y

Prasad, R., Deo, R. C., Li, Y., & Maraseni, T. (2019). Weekly soil moisture forecasting with multivariate sequential, ensemble empirical mode decomposition and Boruta-random forest hybridizer algorithm approach. *Catena*, *177*, 149–166. https://doi.org/10.1016/j.catena.2019.02.012

Rezaie-Balf, M., & Kisi, O. (2018). New formulation for forecasting streamflow: Evolutionary polynomial regression vs. extreme learning machine. *Hydrology Research*, *49*(3), 939–953. https://doi.org/10.2166/nh.2017.283

Rezaie-Balf, M., Maleki, N., Kim, S., Ashrafian, A., Babaie-Miri, F., Kim, N. W., & Alaghmand, S. (2019). Forecasting daily solar radiation using CEEMDAN decomposition-based MARS model trained by crow search algorithm. *Energies*, *12*(8), 1416. https://doi.org/10.3390/en12081416

Rezaie-Balf, M., Naganna, S. R., Ghaemi, A., & Deka, P. C. (2017). Wavelet coupled MARS and M5 model tree approaches for groundwater level forecasting. *Journal of Hydrology*, *553*, 356–373. https://doi.org/10.1016/j.jhydrol.2017.08.006

Roshni, T., Jha, M. K., & Drisya, J. (2020). Neural network modeling for groundwater-level forecasting in coastal aquifers. *Neural Computing and Applications*, 1–18. https://doi.org/10.1007/s00521-020-04722-z

Shamshirband, S., Jafari Nodoushan, E., Adolf, J. E., Abdul Manaf, A., Mosavi, A., & Chau, K. W. (2019). Ensemble models with uncertainty analysis for multi-day ahead forecasting of chlorophyll a concentration in coastal waters. *Engineering Applications of Computational Fluid Mechanics*, *13*(1), 91–101. https://doi.org/10.1080/19942060.2018.1553742

Sharda, V. N., Prasher, S. O., Patel, R. M., Ojasvi, P. R., & Prakash, C. (2008). Performance of multivariate adaptive regression splines (MARS) in predicting runoff in mid-Himalayan micro-watersheds with limited data. *Hydrological Sciences Journal*, *53*(6), 1165–1175. https://doi.org/10.1623/hysj.53.6.1165

Shiri, J., Makarynskyy, O., Kisi, O., Dierickx, W., & Fard, A. F. (2011). Prediction of short-term operational water levels using an adaptive neuro-fuzzy inference system. *Journal of Waterway, Port, Coastal, and Ocean Engineering*, *137*(6), 344–354. https://doi.org/10.1061/(ASCE)WW.1943-5460.0000097

Sibanda, T., Chigor, V. N., Koba, S., Obi, C. L., & Okoh, A. I. (2014). Characterisation of the physicochemical qualities of a typical rural-based river: Ecological and public health implications. *International Journal of Environmental Science and Technology*, *11*(6), 1771–1780. https://doi.org/10.1007/s13762-013-0376-z

Solomatine, D. P., & Xue, Y. (2004). M5 model trees and neural networks: Application to flood forecasting in the upper reach of the Huai River in China. *Journal of Hydrologic Engineering*, *9*(6), 491–501. https://doi.org/10.1061/(ASCE)1084-0699(2004)9:6(491)

Talebi, K. S., Sajedi, T., & Pourhashemi, M. (2014). Forests of Iran. In *A treasure from the past, a hope for the future* (Vol. 10). Springer.

Taormina, R., & Chau, K. W. (2015). ANN-based interval forecasting of streamflow discharges using the LUBE method and MOFIPS. *Engineering Applications of Artificial Intelligence*, *45*, 429–440. https://doi.org/10.1016/j.engappai.2015.07.019

Weber-Scannell, P. K., & Duffy, L. K. (2007). Effects of total dissolved solids on aquatic organism: A review of literature and recommendation for salmonid species. *American Journal of Environmental Sciences*. https://doi.org/10.3844/ajessp.2007.1.6

Wu, C. L., & Chau, K. W. (2013). Prediction of rainfall time series using modular soft computing methods. *Engineering Applications of Artificial Intelligence*, *26*(3), 997–1007. https://doi.org/10.1016/j.engappai.2012.05.023

Yaseen, Z. M., Al-Juboori, A. M., Beyaztas, U., Al-Ansari, N., Chau, K. W., Qi, C., & Shahid, S. (2020). Prediction of evaporation in arid and semi-arid regions: A comparative study using different machine learning models. *Engineering Applications of Computational Fluid Mechanics*, *14*(1), 70–89. https://doi.org/10.1080/19942060.2019.1680576

Yassin, M. A., Alazba, A. A., & Mattar, M. A. (2016). Artificial neural networks versus gene expression programming

for estimating reference evapotranspiration in arid climate. *Agricultural Water Management*, *163*, 110–124. https://doi.org/10.1016/j.agwat.2015.09.009

Yilmaz, B., Aras, E., Nacar, S., & Kankal, M. (2018). Estimating suspended sediment load with multivariate adaptive regression spline, teaching-learning based optimization, and artificial bee colony models. *Science of the Total Environment*, *639*, 826–840. https://doi.org/10.1016/j.scitotenv.2018.05.153

Zhang, W., & Goh, A. T. (2016). Multivariate adaptive regression splines and neural network models for prediction of pile drivability. *Geoscience Frontiers*, *7*(1), 45–52. https://doi.org/10.1016/j.gsf.2014.10.003

Zounemat-Kermani, M., Ramezani-Charmahineh, A., Adamowski, J., & Kisi, O. (2018). Investigating the management performance of disinfection analysis of water distribution networks using data mining approaches. *Environmental Monitoring and Assessment*, *190*(7), 397. https://doi.org/10.1007/s10661-018-6769-1

Appendices

Appendix 1

Table A1. BFs and corresponding equations of evolutionary MARS-CSA for TDS prediction at Rig-Cheshmeh and Soleyman-Tangeh stations.

	Equation	
BF	Rig-Cheshmeh	Soleyman-Tangeh
BF1	max(0,3.4 – Na)	max(0, Na – 1.17)
BF2	max(0, Ca – 4)	max(0,1.17 – Na)
BF3	max(0,4 – Ca)	max(0, Mg – 1.4)
BF4	BF3 ∗ max(0, Na – 3.4)	max(0,4.4 – HCO_3)
BF5	BF3 ∗ max(0,3.4 – Na)	max(0,4.2 – HCO_3)
BF6	max(0, Mg – 3.2)	max(0,4.5 – HCO_3)
BF7	max(0,3.2 – Mg)	max(0, Ca – 2)
BF8	BF5 ∗ max(0, Mg – 3)	max(0, Ca – 2.2)
BF9	BF6 ∗ max(0,3.5 – HCO_3)	
BF10	max(0,3.7 – Mg) ∗ max(0, Ca – 2.7) ∗ max(0, Na – 1.8)	
BF11	BF1 ∗ max(0,1.9 – Ca)	
BF12	max(0, HCO_3 – 4.2)	
BF13	max(0,4.2 – HCO_3)	
BF14	BF13 ∗ max(0, Ca – 2.7)	
BF15	BF13 ∗ max(0,2.7 – Ca)	
BF16	BF14 ∗ max(0, Mg – 2.4)	

Appendix 2

Model's evaluation metrics

1. Nash-Sutcliffe Efficiency (NSE)

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{N} (\mathrm{TDS}_{\mathrm{pre}} - \mathrm{TDS}_{\mathrm{obs}})^2}{\sum_{i=1}^{N} (\mathrm{TDS}_{\mathrm{obs}} - \overline{\mathrm{TDS}_{\mathrm{obs}}})^2}$$

2. Root Mean Square Error (RMSE)

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (\mathrm{TDS}_{\mathrm{pre}} - \mathrm{TDS}_{\mathrm{obs}})^2}$$

3. The ratio of RMSE to Standard Deviation (RSD):

$$\mathrm{RSD} = \frac{\mathrm{RMSE}}{\mathrm{STDE}V_{\mathrm{obs}}} = \frac{\left[\sqrt{\sum_{i=1}^{N} (\mathrm{TDS}_{\mathrm{obs}} - \mathrm{TDS}_{\mathrm{pre}})^2}\right]}{\left[\sqrt{\sum_{i=1}^{N} (\mathrm{TDS}_{\mathrm{obs}} - \overline{\mathrm{TDS}_{\mathrm{obs}}})}\right]}$$

4. Uncertainty at 95% (U95):

$$U_{95} = 1.96\sqrt{(\mathrm{STDE}V^2 + \mathrm{RMS}E^2)}$$

5. Percent Mean Absolute Relative Error (PMARE)

$$\mathrm{PMARE} = \frac{100}{N}\sum_{i=1}^{N} (\mathrm{TDS}_{\mathrm{pre}} - \mathrm{TDS}_{\mathrm{obs}})^2$$

6. Wilmot's Index of agreement (WI)

$$\mathrm{WI} = 1 - \frac{\sum_{i=1}^{N} (\mathrm{TDS}_{\mathrm{obs}} - \mathrm{TDS}_{\mathrm{pre}})^2}{\sum_{i=1}^{N} (|\mathrm{TDS}_{\mathrm{pre}} - \overline{\mathrm{TDS}_{\mathrm{obs}}}| + |\mathrm{TDS}_{\mathrm{obs}} - \overline{\mathrm{TDS}_{\mathrm{obs}}}|)^2}$$

where $\mathrm{TDS}_{\mathrm{obs}}$ and $\mathrm{TDS}_{\mathrm{pre}}$ are the observed and predicted values of the TDS, and $\overline{\mathrm{TDS}_{\mathrm{obs}}}$ is the mean value of the $\mathrm{TDS}_{\mathrm{obs}}$. In addition, N is the number of sample.

Table A2. ANOVA decomposition of MARS-CSA technique.

	Rig-Cheshmeh				Soleyman-Tangeh			
Function	GCV	STD	No. of BFs	Variable(s)	GCV	STD	No. of BFs	Variable(s)
1	358.22	4.85	2	HCO_3	527.65	4.96	3	HCO_3
2	426.21	26.21	2	Ca	1881.5	40.17	2	Ca
3	1070.81	42.38	2	Mg	1769.3	36.12	1	Mg
4	522.53	33.43	1	Na	1208.32	27.03	2	Na
5	374.53	8.47	2	HCO_3, Ca				
6	370.49	5.37	1	HCO_3, Mg				
7	386.45	12.18	3	Ca, Na				
8	375.86	6.87	1	HCO_3, Ca, Mg				
9	361.21	6.94	2	Ca, Mg, Na				