Title

Reservoir water quality simulation with data mining models

Journal

Environmental Monitoring and Assessment, 192(7)

ISSN

0167-6369

Authors

Arefinia, Ali

Bozorg-Haddad, Omid

Oliazadeh, Arman

et al.

Publication Date

2020-07-01

DOI

10.1007/s10661-020-08454-4

Peer reviewed

Environ Monit Assess (2020) 192: 482
https://doi.org/10.1007/s10661-020-08454-4

Reservoir water quality simulation with data mining models

Ali Arefinia · Omid Bozorg-Haddad · Arman Oliazadeh · Hugo A. Loáiciga

Received: 6 February 2020 / Accepted: 23 June 2020 / Published online: 2 July 2020

Abstract Water pollution is a concern in the management of water resources. This paper presents a statistical approach for data mining of patterns of water pollution in reservoirs. Genetic programming (GP), artificial neural network (ANN), and support vector machine (SVM) are applied to reservoir quality modeling. Input data for GP, ANN, and SVM were derived with the CE-QUAL-W2 numerical water quality simulation model. A case study was carried out using measured reservoir inflow and outflow, temperature, and nitrate concentration to the Amirkabir reservoir, Iran. Data mining models were evaluated with the *MAE*, *NSE*, *RMSE*, and R^2 goodness-of-fit criteria. The results indicated that using the SVM model for determining nitrate pollution is time saving and more accurate in comparison with GP, ANN, and particularly CE-QUAL-W2. The SVM model reduces the runtime of nitrate concentration simulation by 581, 276, and 146 s compared with CE-QUAL-W2, GP, and ANN, respectively. The goodness-of-fit results showed that the highest values (R^2 = 0.97, *NSE* = 0.92) and the lowest values (*MAE* = 0.034 and *RMSE* = 0.007) corresponded to SVM predictions, indicating higher model accuracy. This study demonstrates the potential for application of data mining tools to solute concentration simulation in reservoirs.

Keywords Reservoir operation · Support vector machine · Genetic programming · Artificial neural network · CE-Qual-W2 · Amirkabir reservoir

A. Arefinia · O. Bozorg-Haddad (✉) · A. Oliazadeh
Department of Irrigation & Reclamation Engineering, Faculty of Agricultural Engineering & Technology, College of Agriculture & Natural Resources, University of Tehran, Karaj, Tehran, Iran
e-mail: OBHaddad@ut.ac.ir

A. Arefinia
e-mail: Ali.Arefinia@ut.ac.ir

A. Oliazadeh
e-mail: Arman.Oliazadeh@ut.ac.ir

H. A. Loáiciga
Department of Geography, University of California, Santa Barbara, CA 93016-4060, USA
e-mail: Hugo.Loaiciga@ucsb.edu

Introduction

Multiple sources of point and non-spatial, urban, industrial, agricultural, sudden, and non-sudden water pollution events have degraded the quality of many water bodies (Duda 1993; Jahandideh-Tehrani et al. 2015). Water quality simulation models have been developed to assess the spatial and temporal variations of the physical, chemical, and biological characteristics of water bodies. The first version of the well-known CE-QUAL-W2 water quality simulation model appeared in 1990. More than 100 other water quality models have been introduced (Wang et al. 2013). The CE-QUAL-W2 model was applied to simulate water temperature in the USA (Gelda et al. 1998, and Adams et al. 1993). CE-QUAL-W2 was applied to model the temperature in a river-reservoir system to explore management solutions to meet quantitative and qualitative fisheries requirements

(Annear and Wells 2002). A comparative study to evaluating the performance of three water quality models in the USA was reported by Bowen and Heironymuas (2003). There are numerous publications reporting applications of CE-QUAL-W2 to simulate water quality in reservoirs (Noori et al. 2015; Shourian et al. 2016; Aalami et al. 2018; Lindenschmidt et al. 2019; YoosefDoost et al. 2020; Kovač et al. 2020). However, the CE-QUAL-W2 model is computationally burdensome and requires multiple calibration data, which renders water quality simulation time consuming. It is in the context of expediting water quality simulations that data mining methods are gaining acceptance.

Data mining has found application in many fields of water resources management. Genetic programming (GP), artificial neural network (ANN), and support vector machine (SVM) are the most used methods among the comprehensive suite of data mining methods in water management. The assessment of water quality has been one of the fields of application of the latter methods.

Previous works have reported assessments of reservoir pollution with GP (Amirkhani et al. 2016, Nikoo et al. 2017, Ling et al. 2018, Soleimani et al. 2019, Saadatpour 2020). A water quality model was built combining a water quality physical model and ANN (Chaves et al. 2004; Kuo et al. 2006; Saadatpour et al. 2017; Shaw et al. 2017; Afshar et al. 2018; Hasanzadeh et al. 2020). The SVM model has been applied to simulate water quality indexes in reservoirs (Soleimani et al. 2016).

This work applies and evaluates the performance of common data mining algorithms and the CE-QUAL-W2 model in water quality simulations. Methods

The methodology includes three sub-sections: (1) simulate daily nitrate concentration in reservoir outflow with the CE-QUAL-W2, (2) implement data mining algorithms (i.e., GP, ANN, SVM) to extract a function to explains any hidden functional association between input and output variables, and (3) compare the accuracy and run-time of the developed data mining models and the CE-QUAL-W2 model. A flowchart of the employed methodology is displayed in Fig. 1.

The CE-QUAL-W2 model

The initial version of the CE-QUAL-W2 model was called the LARM model, developed by Edinger and Buchak (1975). Version 1.0 of this model was presented by the Water Quality Modeling Team of US Army Corps of Engineers at the Waterways Experimental Station (WES) in 1986. User-friendliness and the simulation capabilities of the CE-QUAL-W2 model have advanced over time leading to the current version 4.1 used in this work.

Model capabilities:

• Hydrodynamics: The ability to predict the surface water level, water velocity, and water temperature. Modeling of water density is based on water temperature and the chemical characteristics of water.

• Water quality: The CE-QUAL-W2 model is capable of simulating a wide range of water quality components and parameters. It simulates the water temperature and any combination of constituents such as (1) nonreactive compounds, (2) suspended solids, (3) phytoplankton, (4) epiphyton, (5) carbonaceous biochemical oxygen demand (CBOD), (6) ammonium, (7) nitrite and nitrate, (8) available biologically acceptable phosphorus, (9) soluble and degradable organic matter, (10) nondegradable organic solvents, (11) total carbon monoxide, (12) alkalinity, (13) total iron, and (14) dissolved oxygen.

• In addition to the above, the CE-QUAL-W2 model simulates several secondary parameters such as pH, total organic carbon, soluble organic carbon, nitrogen, and organic soluble and insoluble phosphorus. Moreover, the CE-QUAL-W2 model has the potential for long-term simulation of a water system with multiple branches, aquifers with irregular dimensions, ice cover effects, water diversions, and dynamic boundary conditions.

Model limitations:

The CE-QUAL-W2 model has limitations. The key ones are:

• Hydrodynamic flow: Assuming complete transverse (lateral) mixing. This assumption is not met in broad rivers. Also, the hydrostatic pressure assumption is not met in all fluid motion cases.

• Water quality: Water quality reactions in the water system descriptions are simplified in the specification of flow parameters such as water velocity, decay coefficient, and dispersion coefficient.

• The CE-QUAL-W2 model solves five laterally averaged hydrodynamic equations expressed in terms of five field variables plus laterally averaged equations of advection/diffusion for chemical constituents. The hydrodynamic equations are the longitudinal momentum equation (along the x coordinate), the vertical momentum equation (along the z coordinate), the

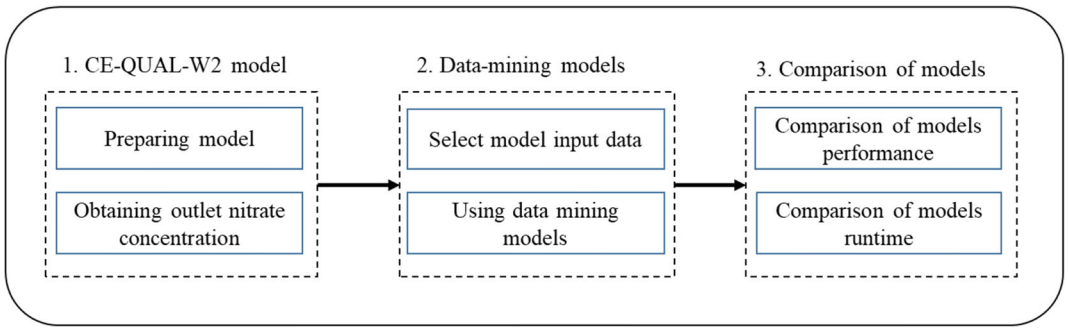

Fig. 1 The methodology's flowchart

continuity equation, the state equation, and hydrostatic pressure or free surface equation. The values of the longitudinal velocity, vertical velocity, density, depth, and pressure are solved for at each simulation location. The concentrations of water constituents are calculated by adding the corresponding transport equations and solving them jointly with the hydrodynamic equations.

The laterally averaged equations used in the CE-QUAL-W2 model are given by Eqs. (1)–(6) (Cole and Wells 2018):

Horizontal momentum equation:

$$\frac{\partial U_x.B}{\partial t}+\frac{\partial U_x.U_x.B}{\partial x}+\frac{\partial U_x.U_z.B}{\partial z}=\mathrm{g}\sin\alpha.B+\mathrm{g}\cos\alpha.B\frac{\partial \eta}{\partial x}-\frac{g\cos\alpha.B}{\rho}\int_{\eta}^{z}\frac{\partial \rho}{\partial x}dz+\frac{1}{\rho}\frac{\partial B\tau_{xx}}{\partial x}+\frac{1}{\rho}\frac{\partial B\tau_{xz}}{\partial z}+qBU_x \quad (1)$$

Vertical momentum equation:

$$0=\mathrm{g}\cos\alpha-\frac{1}{\rho}\frac{\partial P}{\partial Z} \quad (2)$$

Continuity equation:

$$\frac{\partial U_x.B}{\partial x}+\frac{\partial U_z.B}{\partial z}=qB \quad (3)$$

Equation of state for water density:

$$\rho=f(\mathrm{Tw},C_{TDS},\mathrm{C}_{ss}) \quad (4)$$

Hydrostatic or free-surface equation:

$$B\frac{\partial \eta}{\partial t}=\frac{\partial}{\partial x}\int_{\eta}^{h}U_x.Bdz-\int_{\eta}^{h}\boldsymbol{qBdz} \quad (5)$$

Constituent transport equation:

$$\frac{\partial B.C(x,t)}{\partial t}+\frac{\partial B.C(x,t)}{\partial x}+\frac{\partial U_x.B.C(x,t)}{\partial z}-\frac{\partial\left[\mathrm{B.D}_x\frac{\partial C(x,t)}{\partial x}\right]}{\partial x}-\frac{\partial\left[\mathrm{B.D}_z\frac{\partial C(x,t)}{\partial z}\right]}{\partial z}=q_{\Phi}B+S_{\Phi}B \quad (6)$$

in which η = free water level (m); U_z = mean vertical velocity (m/s); B = width of the water control volume (m); g = gravitational acceleration (m/s^2); C = laterally averaged constituent concentration (g/m^3); τ_{xx} = turbulent shear stress on the control volume in the direction x (kg/m/s^2);τ_{xz} = the turbulent shear stress applied to the control volume in the direction z(kg/m/s^2); α = the slope of the river; q = lateral inflow per unit volume of cell or control volume (time^{-1}); Tw = the water temperature (°C); C_{TDS} = the concentration of the soluble solids (g/m^3); C_{ss} = the concentration of suspended solids (g/m^3); D_x = longitudinal dispersion coefficient; D_z = vertical dispersion coefficient (m^2/s); q_{Φ} = input or output pollutant flux through boundary (g/(m^3/s)); U_x = longitudinal water velocity (m/s); U_z = vertical water velocity (m/s); ρ = water density (kg/m^3); and S_{Φ} = laterally averaged source/sink term (g/(m^3/s)). The CE-QUAL-W2 model solves Eqs. (1)–(6) in conjunction with initial and boundary conditions with the finite-difference method (Cole and Wells 2018).

Data mining models

Data mining is a powerful tool whose use is rising in water resources management and organization of high volume information. In fact, data mining is a collection of techniques that move beyond ordinary data and detect hidden, extractable, information. The work implements the data mining methods GP, ANN, and SVM, which are gaining acceptance in the hydrology and water

resources fields (Soleimani et al. 2016; Sarzaeim et al. 2017). These methods are described in the following sections.

Genetic programming

Evolutionary programming is a relatively new method (Koza 1992, 1994; Banzhaf et al. 1998; Khu et al. 2001). GP is inspired by Darwin's theory of evolution and is one of the GA derivative algorithms. An initial population of programs is generated at random. Each program is translated, compiled, and executed and assessed as to how well it performs with respect to task solving. This enables the calculation of a fitness value for each of the programs, and the best one is chosen for reproduction. Programs are combined or mutated into offspring, which are added to the next generation of programs. This process repeats until a termination condition is met. GP employs sets of numbers, operators, and functions as decision variables (Fallah-Mehdipour et al. 2014). The sets are known as terminal sets (Tset) and the functions (Fset). For example:

$$\boldsymbol{Tset} = \{x, \boldsymbol{1}, \boldsymbol{2}, \boldsymbol{-1}, \boldsymbol{-2}, \boldsymbol{\ldots}\} \tag{7}$$

$$\boldsymbol{Fset} = \{\div, \times, +, -, exp, sin, cos, log, \boldsymbol{\ldots}\} \tag{8}$$

GP creates possible solutions (chromosomes) by selecting a random primitive set of connection sets and functions. Figure 2 shows a sample of two chromosomes in GP. The objective function corresponding to each chromosome is calculated. Genetic operators are then applied, as exemplified in Fig. 3. Figure 4 depicts the creation of a new population of solutions by creating a cut and performing coupling and mutation on the parent chromosomes, and thus a new generation of chromosomes (improved possible solutions) is generated. The mutation operator is applied to the Tset and Fset sets. After a number of user-specified iterations, the objective function can no longer be improved, at which time an optimal or near-optimal solution has been obtained for the optimization problem at hand. This GP model was programmed with the MATLAB software.

Artificial neural network

ANN is a database model that detects relationships between outputs and inputs relying on a learning process. These relationships can be complex and non-linear. ANN introduces new inputs with which to predict the corresponding output according to a mathematical model after educating and understanding the relationships between inputs and outputs by ANN. Neural networks are generally non-linear learning mathematical systems.

Problem inputs are entered into the ANN model and outputs are calculated with the aim of minimizing an error criterion. The inputs are passed through a non-linear transfer function and the outputs of the ANN model are calculated. There are several algorithms for network training (i.e., calibration), followed by model testing calculations. The most widely used is the Lewenberg-Markow (LM) (Marquardt 1963) algorithm, an optimization algorithm that minimizes the sum of square errors. Figure 5 shows a three-layer neural network in which W is the weight of neurons, b shows the bias, and f represents the activation function.

A neural network does not require exact mathematical models, and, akin to human beings, it can learn through a number of specific examples, as opposed to digital computers, which require strictly explicit commands. Each neural network goes through the stages of training, testing, and validation. In fact, neural networks can be used to solve problems that do not have exact mathematical relations between inputs and outputs.

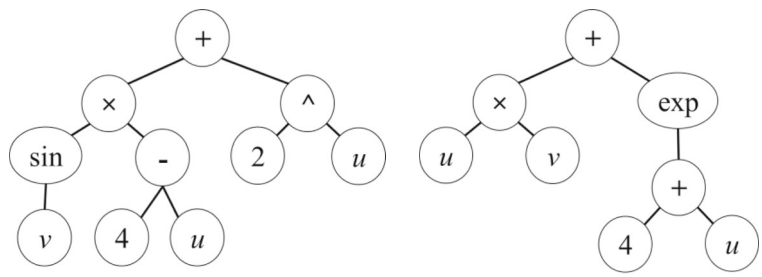

Fig. 2 Tree representation of mathematical relations in GP

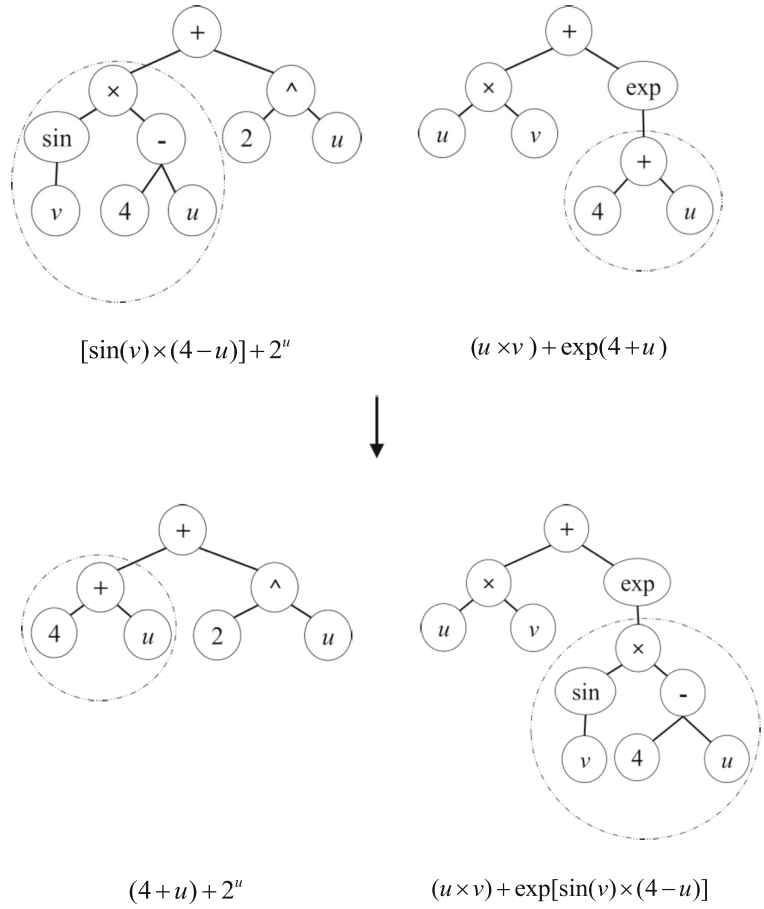

Fig. 3 The coupling genetic operator

ANN training is actually nothing more than adjusting the communication weights of the neurons so that the output of the network converges to the desired output. This paper predicts nitrate concentration by means of an ANN model that has three layers, whose numbers of neurons are calculated by trial and error based on best results. The numbers of neurons in the first, second, and third layer are 10, 5, and 1, respectively. The ANN's structure was achieved by sensitivity analysis and programmed in MATLAB software.

Support vector machine

SVM is a data mining method introduced by Vapnik (1995). SVM is widely used for categorization and regression. The regression form of SVM is called SVR. SVM, similar to ANN, is a data mining technique but its significant difference with ANN is that it is not trapped in local suboptimal solutions. SVR is defined by two functions. The first function calculates the error of the values

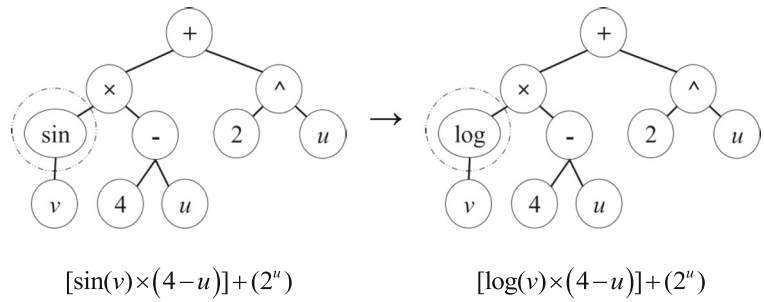

Fig. 4 The mutation genetic operator

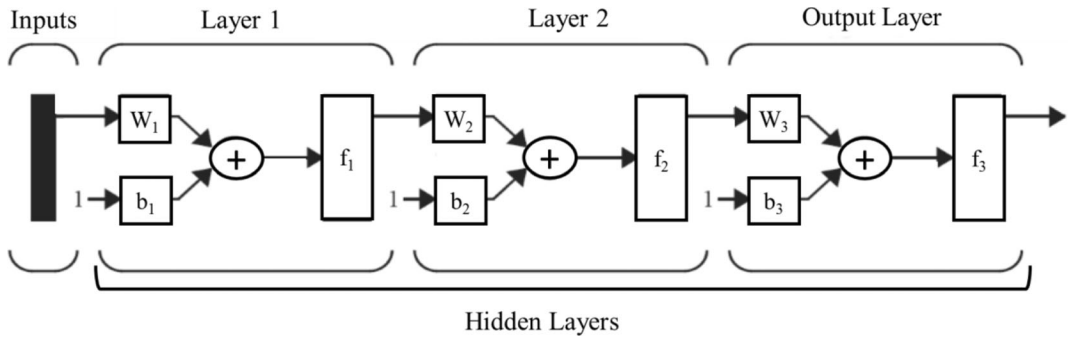

Fig. 5 A three-layer neural network

calculated with SVR (Eq. (9)); the second function calculates the values of outputs (Eq. (10)),

$$|y-f(x)| = \begin{cases} 0 \quad \textit{if} \quad |y-f(x)| \le \varepsilon | \\ |y-f(x)|-\varepsilon = \xi \quad \textit{otherwise} \end{cases} \tag{9}$$

$$f(x) = w^T.x + b \tag{10}$$

where y is the value of the output, $f(x)$ is the output value calculated by SVR, ε = the sensitivity of the function, ξ = the magnitude of the penalty, w = the weight of the variable x, b = the magnitude of the deviation x from its actual values, and T is the transpose sign.

Figure 6 shows the first function does not consider any penalty for values in which the difference between the real value and the value calculated by the model is in the range (− ε, + ε). The amount of the penalty is ξ for

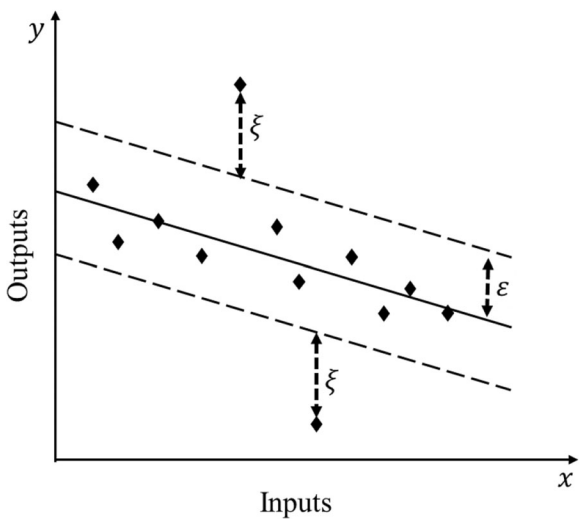

Fig. 6 The representation of the error value function of the SVR

predictions whose deviation with the actual value is outside this range.

The SVR model is an optimization problem that reduces prediction errors. It takes the form of Eq. (11):

$$min\frac{1}{2}\|w\|^2 + c\sum_{i=1}^{m}\left(\xi_i^- + \xi_i^+\right) \tag{11}$$

Subject to:

$(w^T. + b) - y_i < \varepsilon + \xi_i^+ \; i = 1, 2, 3, \ldots, m$ (12)

$y_i - (w^T.x + b) \le \varepsilon + \xi_i^- \; i = 1, 2, 3, \ldots, m$ (13)

where C = the coefficient of penalties, m = the number of training data, ξ_i^+ dξ_i^- the magnitude of penalties for the points above and below the range (− ε, + ε) respectively, y_i the actual data values. The values of w and b are found by solving problem (10)–(12), and these are used in prediction with Eq. (9).

SVM can simulate non-linear input-output relations. In such cases, transfer functions (or kernel functions) are used to convert the non-linear relations of the data to linear ones. Among the applications of SVM in the field of water resources prediction of the long-term water level in lakes (Khan and Coulibaly 2006), static micro-accounting using SVM is used in tropical regions of India (Tripathi et al. 2006). Nitrate concentration in reservoir outflow with the SVM data mining model was performed with the Tanagra software, which includes statistical training techniques. The kernel function used in this study is the radial basis function (RBF). The SVM parameters are achieved by trial and error to find the optimal training.

Nitrate concentration simulation

The simulation model CE-QUAL-W2 was applied to generate nitrate concentration at the outlet of the Amirkabir reservoir. There are numerous parameters

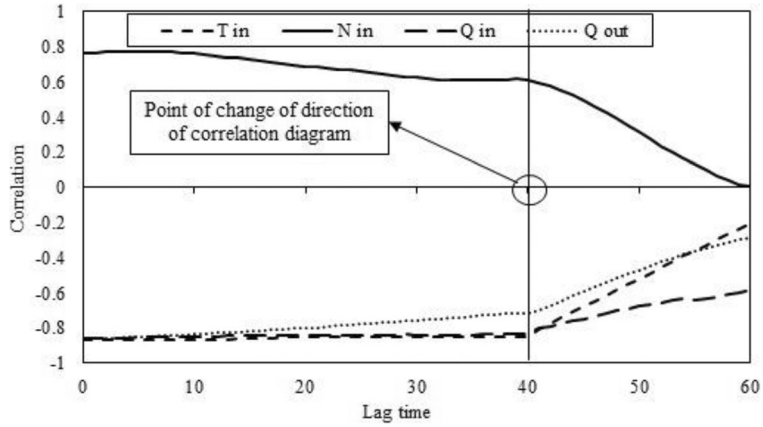

Fig. 7 Trend of correlation charts between the input and output parameters of data mining models at different lag times

such as bathymetry data, meteorological data, discharge rate, indices of water quality, inflow, wind protection, and shadow area in the reservoir which are crucial for CE-QUAL-W2 to create a water quality simulation for estimating the nitrate concentration in the reservoir outflow. Daily data in the year of 2015 were applied for calibration and validation of the CE-QUAL-W2 model. The model calibration step estimated optimal parameters for the CE-QUAL-W2 model, as discussed below.

Model performance was measured with four criteria, namely, the mean absolute error (*MAE*) and the Nash-Sutcliffe efficiency (*NSE*) indices, the correlation coefficient (R^2) and mean square error (*RMSE*) given by Eqs. (14) to (17).

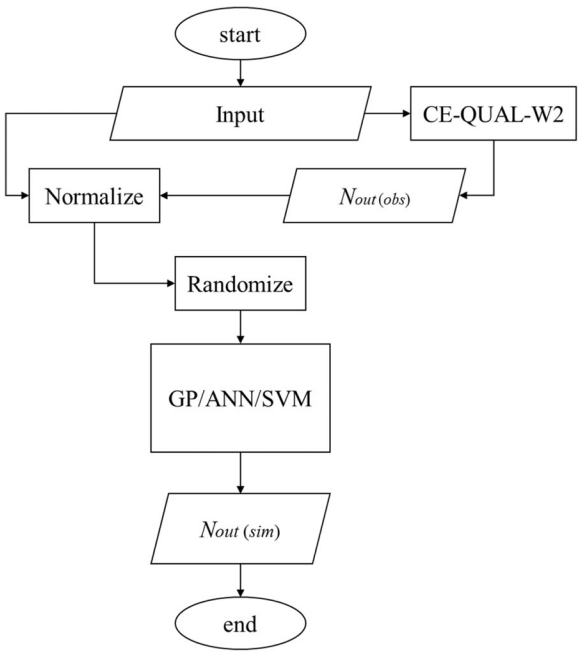

Fig. 8 Flowchart of this paper's method for nitrate concentration simulation

$$MAE = \frac{\sum_{i=1}^{n} |(N_{obs} - N_{sim})|}{n} \tag{14}$$

$$NSE = 1 - \frac{\sum_{t=1}^{n} (N_{obs} - N_{sim})^2}{\sum_{t=1}^{n} (N_{obs} - N_{avg})^2} \tag{15}$$

$$\boldsymbol{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (N_{obs} - N_{sim})^2}{n}} \tag{16}$$

$$R^2 = \frac{\sum_{i=1}^{n} \left(N_{sim} - \overline{N}_{sim}\right)^2 * \left(N_{obs} - \overline{N}_{obs}\right)^2}{\sum_{i=1}^{n} \left(N_{sim} - \overline{N}_{sim}\right)^2 * \sum_{i=1}^{n} \left(N_{obs} - \overline{N}_{obs}\right)^2} \tag{17}$$

where n = number of observational data, N_{obs} = observation nitrate concentration in reservoir outflow, N_{sim} = calculated nitrate concentration in reservoir outflow, $\overline{N}_{sim}$ = mean value of nitrate concentration by model simulation and, $\overline{N}_{obs}$ = average of nitrate concentration in observed flow.

The root mean square error (*RMSE*) and the correlation coefficient (R^2) were employed as performance indicators for each model. The closer the R^2 value is to 1 and the *RMSE* value to zero, the better the prediction skill of a prediction model.

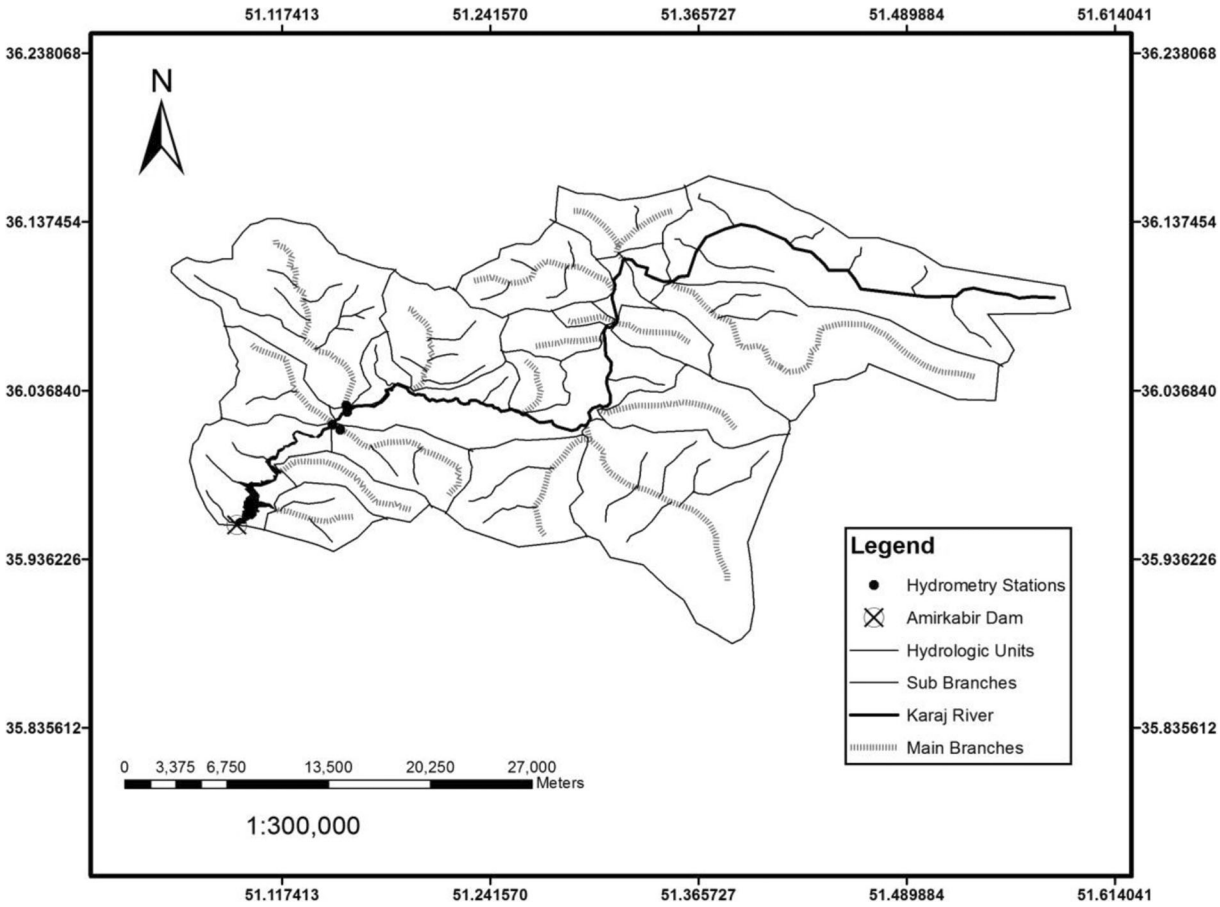

Fig. 9 Amirkabir dam and tributary area in the Karaj basin, Iran

To shorten the simulation time and increase the accuracy of results, data mining models are applied to simulate outflow nitrate concentration. Selecting the appropriate inputs to create the model structure is essential (Kashif Gill et al. 2007; Sarzaeim et al. 2017). The correlation criterion was adopted as a statistical criterion in selecting the input parameters of the data mining models. Accordingly, the correlation between the nitrate concentrations in the outflow and in the inflow at time t, and the discharge flow and the temperature in the reservoir at various lag times were investigated, and several variables were identified. Figure 7 displays the trend of correlation charts between the input and output parameters of the data mining models at different lag times. It is seen in Fig. 7 that the correlation between input and output parameters of the data mining models deteriorates after the 40th lag time which results in longer runtimes with negligible gain in accuracy.

Equation (18) represents a generic function relating the output parameter (i.e., the nitrate concentration in

Table 1 Performance indices for the CE-QUAL-W2 model

	Calibration				Validation			
	MAE	*NSE*	*RMSE*	R^2	*MAE*	*NSE*	*RMSE*	R^2
Water level	0.13	0.87	0.16	0.93	0.20	0.75	0.30	0.86
Temperature	0.12	0.85	0.23	0.94	0.15	0.84	0.27	0.92
Nitrate	0.32	0.71	0.42	0.83	0.38	0.69	0.51	0.82

reservoir outflow) at time t and the input parameters at times t through $t-40$.

$$(N_{out})_t = f\left[(N_{in}, T_{in}, Q_{in}, Q_{out})_t, \ldots, (N_{in}, T_{in}, Q_{in}, Q_{out})_{t-40}\right] \quad (18)$$

where $(N_{out})_t$ = the nitrate concentration in reservoir outflow at time t, $(Q_{in})_t$ =the reservoir inflow at time t, $(Q_{out})_t$ = the discharge outlet at time t, $(T_{in})_t$ = the water temperature of reservoir inflow t, $(N_{out})_{t-40}$ = the nitrate concentration in reservoir outflow at time $t-40$, $(Q_{in})_{t-40}$ = the reservoir inflow at time $t-40$, $(Q_{out})_{t-40}$ reservoir outflow at time $t-40$, $(T_{in})_{t-40}$ = water temperature of reservoir inflow at time $t-40$, f = function that converts inputs to output, and t = index of day.

All the input and output variables were normalized according to Eq. (19):

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (19)$$

where x_{norm} = normalized parameter, x = real value of the parameter before normalization, x_{min} and x_{max} denote the minimum and maximum values of the parameter before normalization, respectively. Upon normalization of the database the GP, ANN, and SVM were trained using input data and the amount of output nitrate calculated with the CE-QUAL-W2 model. Seventy percent of the input data were randomly selected for training purposes, and the rest of the data were used to test the GP, ANN, and SVM.

Furthermore, *MAE*, *NSE*, R^2, and *RMSE* were employed as performance indicators for all structure data mining models to choose the best one based on Eqs. (14) to (17). Figure 8 displays the flowchart for implementing the GP, ANN, and SVM models.

Case study

The Amirkabir dam and reservoir were inaugurated in 1961. The reservoir's functions are flood control, water supply to the city of Tehran, and water supply for Karaj agriculture and hydroelectric power generation. Nitrate pollution control in this reservoir is crucial to protect the health of the population of about 9 million people and to produce safe agricultural products. The Amirkabir dam is a two-arched concrete structure with a maximum height of 180 m from the foundation, width of 30 m at bottom and 9 m at crest. The Amirkabir dam is located 63 km northwest of Tehran. The reservoir has an average area equal to 764 km^2 and an average runoff of 472 million cubic meters. The reservoir is on the Karaj River, which originates in the Alborz Mountains and discharges to a salt lake, near the city of Qom. Figure 9 shows the location of the Amirkabir dam and its tributary basin.

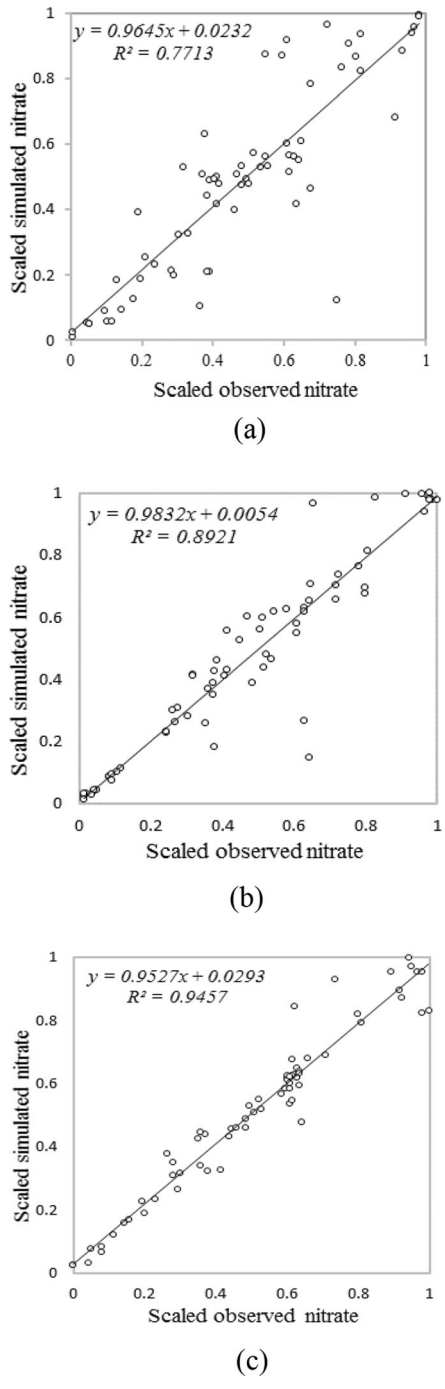

Fig. 10 Nitrate prediction diagrams obtained with (a) GP, (b) ANN, and (c) SVM

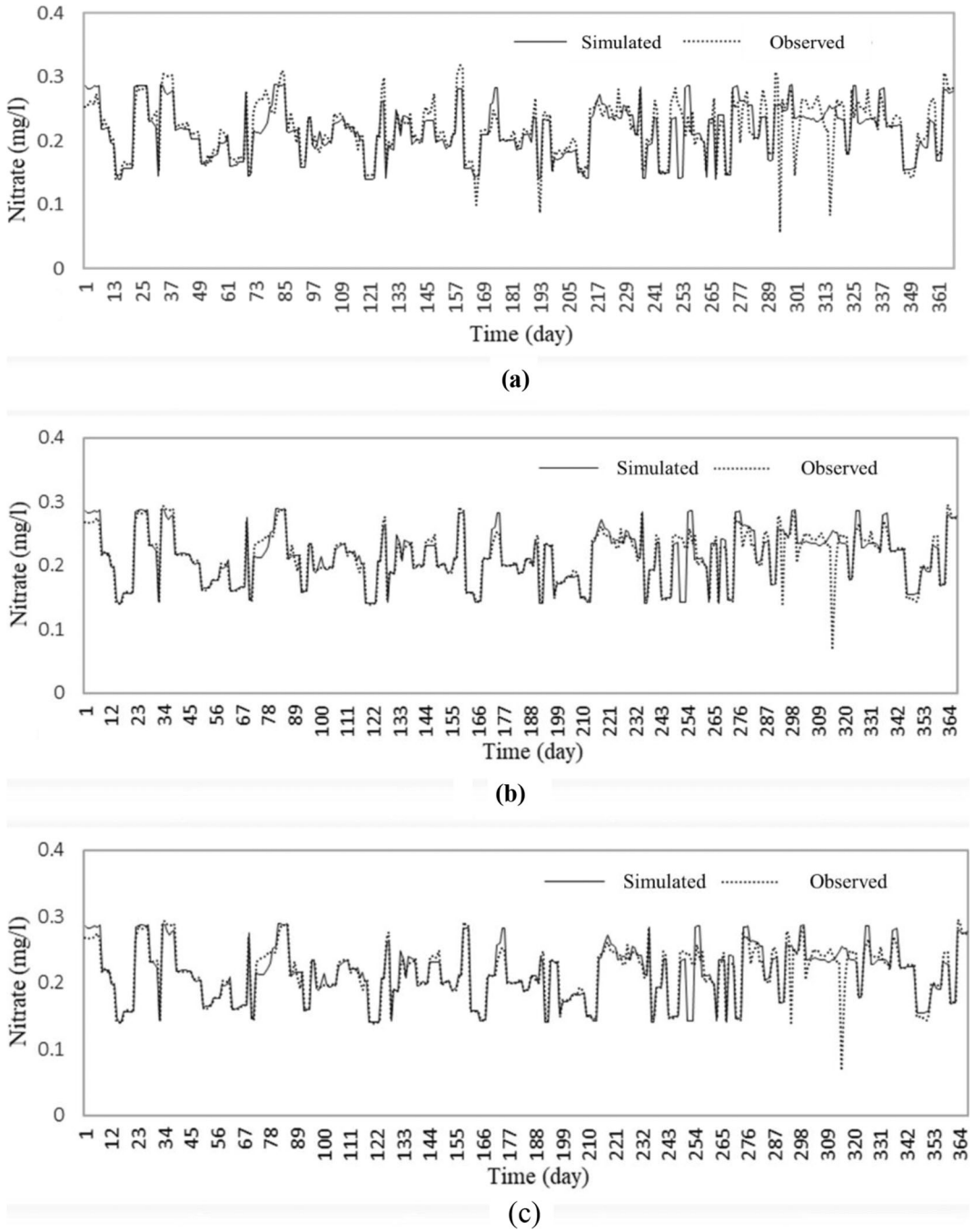

Fig. 11 Observed and simulated nitrate concentration with (a) GP, (b) ANN, and (c) SVM

Results and discussion

The calculated values of the efficiency indices for the CE-QUAL-W2 model for calibration and validation intervals are listed in Table 1.

Based on the results listed in Table 1, it is evident that the data obtained from CE-QUAL-W2 and the observed data are in proper agreement.

It is seen in Fig. 10 that the calculated R^2 values for nitrate concentration simulation stage equal 0.77, 0.89,

Table 2 Results for nitrate concentration prediction obtained with data mining models

Model	*MAE*		*NSE*		*RMSE*		R^2		Average Runtime (s)	Improvedruntime (s)
	Training	Testing	Training	Testing	Training	Testing	Training	Testing		
GP	0.065	0.073	0.846	0.834	0.091	0.116	0.856	0.771	324	305
ANN	0.049	0.066	0.896	0.842	0.062	0.097	0.928	0.892	194	435
SVM	0.034	0.056	0.921	0.875	0.007	0.013	0.970	0.945	48	581

and 0.94 with respect to GP, ANN, and SVM respectively. R^2 values close to 1 mean higher accuracy in predictions.

The results of the simulations of nitrate concentration in reservoir outflow calculated with GP, ANN, and SVM are presented in Fig. 11. It can be seen the plot depicted using GP predictions reveals results that were not as accurate as those obtained with ANN or SVM. It is also evident that the values predicted using SVM had higher accuracy than those of ANN. ANN-predicted nitrate values presented in Fig. 11 (c) display a relatively good correlation between observed nitrate concentrations.

Table 2 lists a summary of GP, ANN, and SVM performance results in terms of the *MAE*, *NSE*, *RMSE*, R^2 and average run time obtained in 10 runs of each model. It is clear the data mining tool of the SVM model exhibits better performance than the ANN and GP models. When comparing ANN to GP, the former performed better than the latter, though not as well as SVM. As shown in Table 2, the average run time of SVM is lower than those of ANN and GP models by 75% and 85%, respectively. When compared with CE-QUAL-W2, GP, and ANN, the SVM model reduced the runtime of nitrate concentration simulation by 305, 435, and 581 s, respectively. The SVM model achieved a more accurate and efficient performance in terms of error reduction and runtime than GP and ANN.

Concluding remarks

This work applied GP, ANN, and SVM data mining tools to predict nitrate concentrations in reservoir outflow using as input (predictors) the reservoir inflow, inflow water temperature, and the nitrate concentration in reservoir inflow.

The goodness-of-fit results listed in Table 2 indicate the highest values (R^2 = 0.97, *NSE* = 0.92) and the lowest values of (*MAE* = 0.034 and *RMSE* = 0.007), both corresponding to SVM predictions. Similarly, in the testing phase, the SVM model yield the highest values (R^2 = 0.94, *NSE* = 0.87) and the lowest (*MAE* = 0.056, *RMSE* = 0.01).

It is concluded that the SVM data mining tool outperforms GPs and ANNs with the particular data sets and water constituent (nitrate) selected in this study in terms of the goodness-of-fit or performance criteria (i.e., *MAE*, *NSE*, *RMSE*, and R^2). The superior performance of SVM was realized in the training phase and in the testing phase.

The complexity of the water quality prediction problem and a large number of time lags in the training phase poses a substantial computational burden. The run times of the implemented models associated with 10 model runs established the superiority of SVM, whose run times were 75% and 85% of those associated with GP and ANN, respectively. The selected data mining models were superior to CE-QUAL-W2.

Funding information The authors thank Iran's National Science Foundation (INSF) for its financial support of this research.

Compliance with ethical standards

Conflict of interests The authors declare that they have no conflict of interest.

References

Aalami, M. T., Abbasi, H., & Nourani, V. (2018). Sustainable management of reservoir water quality and quantity through reservoir operational strategy and watershed control strategies. *International Journal of Environmental Research, 12*(6), 773–788.

Adams, W., Thackston, E., Speece, R., Wilson, D., and Cardozo, R. (1993). Effect of Nashville's combined sewer overflows on the water quality of Cumberland River. . *Technical Rep*, 42

Afshar, A., Masoumi, F., & Sandoval Solis, S. (2018). Developing a reliability-based waste load allocation strategy for river-reservoir systems. *Journal of Water Resources Planning and Management, 144*(9), 04018052.

Amirkhani, M., Bozorg-Haddad, O., Fallah-Mehdipour, E., & Loáiciga, H. A. (2016). Multiobjective reservoir operation for water quality optimization. *Journal of Irrigation and Drainage Engineering, 142*(12), 04016065.

Annear, R. L., and Wells, S. A. (2002). "The Bull Run River–reservoir system model."

Banzhaf, W., Nordin, P., Keller, R. E., & Francone, F. D. (1998). *Genetic programming. An introduction Morgan and Kaufmann Publishers. California*: *Inc. San Francisco*.

Bowen, J. D., & Heironymuas, J. W. (2003). A CE-QUAL-W2 model of neuse estuary for total maximum daily load development. *Water Resources Planning and Management, ASCE, 129*(4), 283–294.

Chaves, P., Tsukatani, T., & Kojiri, T. (2004). Operation of storage reservoir for water quality by using optimization and artificial intelligence techniques. *Mathematics and Computers in Simulation, 67*(4-5), 419–432.

Cole, T. M., & Wells, S. A. (2018). *CE-QUAL-W2: A two-dimensional, laterally averaged, hydrodynamic and water quality model, version 4.1, user manual*. Department of Civil and Environmental Engineering: Portland State University, Portland, Oregon.

Duda, A. M. (1993). Addressing nonpoint sources of water pollution must become an international priority. *Water Science and Technology, 28*(3-5), 1–11.

Edinger, J., & Buchak, E. (1975). A hydrodynamic, two-dimensional reservoir model: the computational basis. In *US Army Engineer Division. Ohio River.*: *Cincinnati, OH*.

Fallah-Mehdipour, E., Bozorg-Haddad, O., & Mariño, M. A. (2014). Genetic programming in groundwater modeling. *Journal of Hydrologic Engineering, 19*(12), 04014031. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000987.

Gelda, R. K., Owens, E. M., & Effler, S. W. (1998). Calibration, verification, and an application of a two-dimensional hydrothermal model [CE-QUAL-W2 (t)] for Cannonsville Reservoir. *Lake and Reservoir Management, 14*(2-3), 186–196.

Hasanzadeh, S. K., Saadatpour, M., & Afshar, A. (2020). A fuzzy equilibrium strategy for sustainable water quality management in river-reservoir system. *Journal of Hydrology, 124892*.

Jahandideh-Tehrani, M., Bozorg-Haddad, O., & Loáiciga, H. A. (2015). Hydropower reservoir management under climate change: the Karoon reservoir system. *Water Resources Management, 29*(3), 749–770. https://doi.org/10.1007/s11269-014-0840-7.

Kashif Gill, M., Asefa, T., Kaheil, Y., & Mckee, M. (2007). Effects of missing data on performance of learning algorithms for hydrologic predictions: Implications to an imputation technique. *Journal of Water Resources Research, 43*(7), W07416.

Khan, M. S., & Coulibaly, P. (2006). Application of support vector machine in lake water level prediction. *Journal of Hydrologic Engineering, 11*(3), 199–205.

Khu, S. T., Liong, S. Y., Babovic, V., Madsen, H., & Muttil, N. (2001). Genetic programming and application in real-time runoff forecasting 1. *JAWRA Journal of the American Water Resources Association, 37*(2), 439–451.

Kovač, I., Šrajbek, M., Kranjčević, L., & Novotni-Horčička, N. (2020). Nonlinear models of the dependence of nitrate concentrations on the pumping rate of a water supply system. *Geosciences Journal*, 1–11.

Koza, J. R. (1992). *Genetic programming II, automatic discovery of reusable subprograms*. Cambridge, MA: MIT Press.

Koza, J. R. (1994). Genetic programming as a means for programming computers by natural selection. *Statistics and computing, 4*(2), 87–112.

Kuo, J.-T., Wang, Y.-Y., & Lung, W.-S. (2006). A hybrid neural–genetic algorithm for reservoir water quality management. *Water research, 40*(7), 1367–1376.

Lindenschmidt, K. E., Carr, M. K., Sadeghian, A., & Morales-Marin, L. (2019). CE-QUAL-W2 model of dam outflow elevation impact on temperature, dissolved oxygen and nutrients in a reservoir. *Scientific Data, 6*(1), 1–7.

Ling, Y., Wang, M., Chen, Q., & Mynett, A. (2018). Modelling spatial-temporal dynamics of cyanobacteria abundance in lakes by integrating cellular automata and genetic programming. *EPiC Series in Engineering, 3*, 1214–1223.

Marquardt, D. W. (1963). An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics, 11*(2), 431–441.

Nikoo, M. R., Pourshahabi, S., Rezazadeh, N., & Shafiee, M. E. (2017). Stakeholder engagement in multi-objective optimization of water quality monitoring network, case study: Karkheh Dam reservoir. *Water Science and Technology: Water Supply, 17*(4), 966–974.

Noori, R., Yeh, H. D., Ashrafi, K., Rezazadeh, N., Bateni, S. M., Karbassi, A., Kachoosangi, F. T., & Moazami, S. (2015). A reduced-order based CE-QUAL-W2 model for simulation of nitrate concentration in dam reservoirs. *Journal of Hydrology, 530*, 645–656.

Saadatpour, M. (2020). An adaptive surrogate assisted CE-QUAL-W2 model embedded in hybrid NSGA-II_ AMOSA algorithm for reservoir water quality and quantity management. *Water Resources Management*, 1–15.

Saadatpour, M., Afshar, A., & Edinger, J. E. (2017). Meta-model assisted 2D hydrodynamic and thermal simulation model (CE-QUAL-W2) in deriving optimal reservoir operational strategy in selective withdrawal scheme. *Water Resources Management, 31*(9), 2729–2744.

Sarzaeim, P., Bozorg-Haddad, O., Bozorgi, A., & Loáiciga, H. A. (2017). Runoff projection under climate change conditions with data-mining methods. *Journal of Irrigation and Drainage Engineering, 143*(8), 04017026.

Shaw, A. R., Smith Sawyer, H., LeBoeuf, E. J., McDonald, M. P., & Hadjerioua, B. (2017). Hydropower optimization using artificial neural network surrogate models of a high-fidelity hydrodynamics and water quality model. *Water Resources Research, 53*(11), 9444–9461.

Shourian, M., Moridi, A., & Kaveh, M. (2016). Modeling of eutrophication and strategies for improvement of water quality in reservoirs. *Water Science and Technology, 74*(6), 1376–1385.

Soleimani, S., Bozorg-Haddad, O., Saadatpour, M., & Loáiciga, H. A. (2016). Optimal selective withdrawal rules using a coupled data mining model and genetic algorithm. *Journal*

of Water Resources Planning and Management, 142(12), 04016064.

Soleimani, S., Bozorg-Haddad, O., Saadatpour, M., & Loáiciga, H. A. (2019). Simulating thermal stratification and modeling outlet water temperature in reservoirs with a data-mining method. *Journal of Water Supply: Research and Technology-Aqua, 68*(1), 7–19.

Tripathi, S., Srinivas, V., & Nanjundiah, R. S. (2006). Downscaling of precipitation for climate change scenarios: a support vector machine approach. *Journal of hydrology, 330*(3-4), 621–640.

Vapnik, V. (1995). *The nature of statistical learning theory*. New York: Springer.

Wang, Q., Li, S., Jia, P., Qi, C., & Ding, F. (2013). A review of surface water quality models. *The Scientific World Journal, 2013*.

YoosefDoost, A., Karrabi, M., Rezazadeh, N., & Mirabi, M. (2020). Development of the delta-normal stress combining CE-QUAL-W2 as a novel method for spatio-temporal monitoring of water quality in Karkheh Dam Reservoir. *Environmental Monitoring and Assessment, 192*, 1–13.