Arabian Journal of Geosciences (2019) 12: 544
https://doi.org/10.1007/s12517-019-4687-3

ORIGINAL PAPER

Modeling total dissolved gas (TDG) concentration at Columbia river basin dams: high-order response surface method (H-RSM) vs. M5Tree, LSSVM, and MARS

Behrooz Keshtegar[1,2] · Salim Heddam[3] · Ozgur Kisi[4] · Shun-Peng Zhu[5]

Received: 7 March 2019 / Accepted: 31 July 2019 / Published online: 22 August 2019

Abstract

The accuracy of ordinary response surface method (RSM) is improved using the high-nonlinear polynomial basis functions for modeling total dissolved gas (TDG). The third-order (3O), fourth-order (4O), and fifth-order (5O) polynomial functions are applied as the mathematical relations of TDG. The accuracy of third-, fourth-, and fifth-order polynomial basis function based on high-order RSM (H-RSM) is compared with least squares support vector machine (LSSVM), M5 model tree (M5Tree), and multivariate adaptive regression spline (MARS) models. The H-RSM, LSSVM, MARS, and M5Tree models were developed and compared using four input combinations and evaluated using several statistical indices namely coefficient of correlation (R), Willmott index of agreement (d), Nash-Sutcliffe coefficient of efficiency (NSE), RMSE, and MAE. The models were developed using data collected from four USGS stations at Columbia River, USA. According to the obtained results, it was demonstrated that the models worked with high level of satisfactory accuracy with respect to the five statistical indices. Overall, the 5H-RSM1 with four input variables provided the best accuracy at the four stations with R, NSE, d, RMSE, and MAE ranging from 0.911 to 0.965, 0.829 to 0.931, 0.952 to 0.982, 1.456 to 2.263, and 1.022 to 1.751, respectively.

Keywords Modeling · Total dissolved gas · TDG · H-RSM · M5Tree · LSSVM · MARS

Introduction

Over the last few decades, with the great importance attributed to the management and control of dam's reservoirs, the linkage between total dissolved gas (TDG) supersaturation and some environmental problems has been widely recognized (Parker et al. 1984; Boyd et al. 1994; Tanner et al. 2012). In particular, and without limitation to the foregoing, the "gas bubble trauma GBT" in fish also designated by "gas bubble disease GBD" is the most important and serious problem associated with the elevation of TDG (Beeman et al. 2003; Skov et al. 2013; Wang et al., 2018a), and the interest to demonstrate that TDG supersaturation may be having an adverse effect on the aquatic life of fish is not new and dates back to

Editorial handling: Santanu Banerjee

✉ Salim Heddam
heddamsalim@yahoo.fr

Behrooz Keshtegar
beh.keshtegar@tdtu.edu.vn

Ozgur Kisi
ozgur.kisi@iliauni.edu.ge

Shun-Peng Zhu
zspeng2007@uestc.edu.cn

[1] Division of Computational Mathematics and Engineering, Institute for Computational Science, Ton Duc Thang University, Ho Chi Minh City, Vietnam

[2] Faculty of Civil Engineering, Ton Duc Thang University, Ho Chi Minh City, Vietnam

[3] Faculty of Science, Agronomy Department, Hydraulics Division, Laboratory of Research in Biodiversity Interaction Ecosystem and Biotechnology, University 20 Août 1955, Route El Hadaik, 26 Skikda, BP, Algeria

[4] School of Technology, Ilia State University, 0162 Tbilisi, Georgia

[5] Center for System Reliability & Safety, University of Electronic Science and Technology of China, Chengdu 611731, China

the beginning of the last century (Gorham, 1898; Marsh and Gorham 1904; Alikhuni et al., 1951; Parker et al. 1984; Boyd et al. 1994; Colt 1986). In general, TDG supersaturation is quantified in percent (%) and calculated as the difference between TDG pressure and atmospheric pressure, for which TDG pressure is calculated for any temperature, as the sum of the partial pressures of all TDG plus the vapor pressure of water, and generally limited to 110% (Skov et al. 2013; Politano et al. 2016). Nitrogen and oxygen form the essential dissolved gas and hence play an important role in the determination of the final TDG concentration. However, the limited level of gas concentration varies depending on the nature of gas dissolved in water, and generally, nitrogen and dissolved oxygen have been considered unsafe above 110% and 300%, respectively (Parker et al. 1984). TDG supersaturation is produced in a major part by the "spill" at "hydroelectric projects" (Weitkamp et al. 2003). Water is generally spilled whether through drum gates or also through a series of outlet work conduits (Bragg and Johnston 2016). According to Beeman and Maule (2006), the rate of TDG supersaturation increases if the pressure of TDG falls below the hydrostatic pressure in dam tailraces.

The objectives related to the monitoring and control of TDG supersaturation in water ranged, on the one hand, from providing the relevant physicochemical phenomena which are causing TDG supersaturation and, on the other hand, understanding the anthropogenic and natural causes (CCME 1999). Consequently, an effective understanding of the TDG supersaturation represents a high priority, especially for the regions with high network of dam reservoirs (e.g., Columbia River, USA), and in this case, the database for all measured variables plays a crucial role and can help in the development of models for estimating TDG. Previous studies have shown that TDG can be analyzed using numerical models (NM), using laboratory and field experiment. However, NM possesses an important disadvantage that generalization is inappropriate for the majority of empirical equation derived. Therefore, the results cannot be generalized statistically to the outside of used data during the calibration phase (Wang et al., 2018b). Review of literature clearly demonstrated that various empirical and physical models were proposed for predicting TDG supersaturation (Roesner and Norton 1971; Hibbs and Gulliver 1997; Shaw 1998; Geldert et al. 1998; Orlins and Gulliver 2000; Hadjerioua et al. 2012; Feng et al. 2013; Picket et al. 2004; Tawfik and Diez 2014; Wang et al., 2018a; Yuan et al. 2018).

Roesner and Norton (1971) were the first authors in the literature that have attributed particular importance to develop a predictive model for TDG downstream of spillway. The proposed model was applied at Columbia and Snake rivers, USA. As a result of the model, it is reported that an effective depth to reduce the rate of TDG supersaturation is directly related to a specific discharge. Tawfik and Diez (2014) proposed a model for predicting TDG supersaturation up to the onset of bubble nucleation at the electrodes (heterogeneous nucleation). In order to demonstrate the effect of solid media in water, Yuan et al. (2018) conducted an investigation and proposed a model linking the density of vegetation with the TDG dissipation process. The model was able to significantly increase the performances of predictive models of TDG "transport" and "dissipation" in downstream high dam spilling. From the obtained results, the authors demonstrated that the dissipation rate of TDG supersaturation increased with the increase in vegetation density. A first-order kinetics reaction model is proposed by Picket et al. (2004). Feng et al. (2013) used the CE-QUAL-W2 developed by the US Army Corps of Engineers, for developing a two-dimensional laterally averaged hydrodynamic model for TDG transportation and dissipation in the deep Dachaoshan reservoir. The authors demonstrated that the percent saturation increases with the water depth at the layers near the water surface. A model based on volume of fluid method was proposed by Wang et al. (2018b) and applied for predicting TDG downstream of spillways. TDG was quantified using a mathematical formula taking into account the masse transfers between bubbles and water.

Orlins and Gulliver (2000) used a combined model having two components: physical and numerical for total dissolved gas concentration. The physical model takes into account the hydraulic information related to the modification of spillway, while the numerical model estimates the concentration of TDG using mass transport relation between air and water. Shaw (1998) implemented a new equation for TDG production from spill using the Columbia River salmon passage model (CRiSP.1). The proposed equation links the discharge to the TDG % exiting the tailrace of a dam. Hibbs and Gulliver (1997) proposed a numerical model for determining an effective bubble depth (EPD) in spillway taking into account three variables: spillways, flow parameters, and the tailwater depth. The determined EPD is incorporated into the models used for predicting TDG in downstream of spillways. In another study, Geldert et al. (1998) used data collected from three spillways on the Columbia and Snake rivers for developing a physical predictive model for TDG caused by the stilling basin in downstream of the dam. The proposed model takes into account several ideas, namely transfer across the air water interface, and the hypothesis that TDG% is mainly related to the stilling basin depth and the downstream river depth. Hadjerioua et al. (2012) proposed a generalized model for predicting TDG level. The proposed model included a large number of parameters including structural, operational, and environmental parameters, among them are the following: (*i*) the depth of the stilling basin, (*ii*) total head, (*iii*) volume spill, (*iv*) powerhouse flow, (*v*) TDG pressure, (*vi*) water temperature, (*vii*) barometric pressure, (*viii*) geometry of the spillway, (*ix*) spillway flow deflectors, (*x*) training walls, and (*xi*) and baffle blocks. Several other models can be found in the literature (Politano et al. 2007, 2009, 2012, 2017; Stewart et al. 2015; Witt et al.2017a, b).

Paradoxically, despite the importance attached to the study and modeling TDG supersaturation and to the application of DD models in many areas of scientific researches, none of the above reported investigations have applied or used data-driven (DD) models for predicting TDG. Recently, Heddam (2017) proposed for the first time the generalized regression neural network (GRNN) for predicting TDG at Columbia River dams. In the present study, we applied four DD models for predicting TDG (%), namely high-order response surface method (H-RSM), least squares support vector machine (LSSVM), M5 model tree (M5Tree), and multivariate adaptive regression splines (MARS) using data collected at four dam reservoirs at Columbia rivers, USA.

Methods

Study area and data used

The data used for developing the models were obtained from the United States Geological Survey (USGS) database (https://waterdata.usgs.gov). The study includes four stations, namely USGS453439122223900 at Columbia River, right bank, at Washougal, Clark County, WA (latitude, 45° 34′ 39″; longitude, 122° 22′ 39″); USGS453630122021400 at Columbia River, left bank, near Dodson, Multnomah County, OR (latitude, 45° 36′ 30″; longitude, 122° 02′ 14″); USGS453845121562000 at Columbia River at Bonneville Dam Forebay, Skamania County, WA (latitude, 45° 38′ 45″; longitude, 121° 56′ 20″); and USGS453845121564001 at Columbia River at Cascade Island, Skamania County, WA (latitude, 45° 38′ 45″; longitude, 121° 56′ 40″). The statistical parameters of the selected data used in our investigation are summarized in Table 1. Four variables measured at daily time step were selected as input variables for developing the models. These variables were respectively the following: (*i*) daily water temperature (TE), (*ii*) daily barometric pressure (BP), (*iii*) daily spill from dam (SFD), and (*iv*) daily discharge (DIS). In addition, the total dissolved gas measured in % of saturation is used as the output of the models. Note from Table 1 that the SFD and DIS have high correlations with the TDG in all stations. Data used in the present investigation are based on large period of measure and cover a period from 1 January 1998 to 31 December 2017 for three stations: USGS 453439122223900, USGS 453630122021400, and USGS 453845121562000, respectively, while for the fourth station (USGS 453845121564001), the data cover a period from 1 January 2004 to 31 December 2017. However, during the period of record, the stations have some incomplete data from year to year. The data set was randomly divided into two sub-data sets in this study: (*i*) training subset (70%) for model calibration and (*ii*) validation subset (30%) for model testing. The validation subset, sometimes called test subset, was used to assess the performances of the proposed models (Moghaddasi and Noorian-Bidgoli 2018; Li et al. 2019).

High-order response surface method

The response surface method (RSM) is an efficient and simple approximating experiment-based the quadratic polynomial set function corresponding to several data bases (Keshtegar and Kisi 2017). Generally, the second-order polynomial function is utilized for predicting TDG in the RSM by the following function (Keshtegar and Heddam 2018).

$$T\hat{D}G = a_0 + \sum_{i=1}^{n} a_i x_i + \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j \tag{1}$$

where, $T\hat{D}G$ is the approximated TDG using the input data sets x including TE, BP, SFD, and DIS. n is the number of input variables, and a_0, a_i, and a_{ij} are the unknown coefficients. The total number of coefficients is NC = $(n + 1)(n + 2)/2$. Least squared estimator is applied to approximate the unknown coefficients of the response surface function in Eq. (1). Thus, the predicted data basis RSM is given as follows (Keshtegar and Seghier 2018):

$$TDG_p = P(x_i)\left[P(x)^T P(x)\right]^{-1}\left[P(x)^T\right]\boldsymbol{TDG} \tag{2}$$

where, TDG$_P$ is the predicted TDG using ith input variables (x_i ϵ [TE$_i$, BP$_i$, SFD$_i$, and DIS$_i$]). Here, x and TDG are respectively the input data set and the observed TDG in the training phase. $P(.)$ is the polynomial basis function. Second-order functions for 4-input data with 15 coefficients can be expressed as:

$$P(x) = \left[1, x_1, x_2, x_3, x_4, x_1^2, x_1, x_2, x_1, x_3, x_1, x_4, x_2^2, x_2, x_3, x_2, x_4 x_3^2, x_3, x_4, x_4^2\right] \tag{3}$$

As seen from Eq. (3), the $P(x_i)$ is computed based on the input data of (x_i ϵ [TE$_i$, BP$_i$, SFD$_i$, and DIS$_i$]) at ith data; thus, we have a matrix with $N \times NC$ dimensions for $P(x)$ where N is the number of the training data and NC is the number of the unknown coefficients and the polynomial basis functions $P(x)$ is developed based on the linear cross-correlations of the input variables (i.e., $x_i\, x_j\; i \neq j$). Consequently, highly nonlinear cross-correlation between variables x_i and x_j is not considered in the RSM. The recent studies showed that the original RSM may provide inaccurate results compared with the modified versions of the response surface method which transfers the input

Table 1 Statistical parameters of the used data sets for all stations

Station	Data set	Unit	X_{mean}	X_{max}	X_{min}	S_{x}	C_{v}	R
USGS 453439122223900	TE	°C	16.58	23.7	6.40	4.527	0.274	− 0.194
	BP	mmHg	763.8	779	748	3.607	0.005	− 0.167
	DIS	kcfs	214.6	506	71.2	91.65	0.427	0.735
	SFD	feet	86.75	303	0.00	53.51	0.617	0.888
	TDG	% sat.	111.6	129	98.0	5.194	0.047	1.000
USGS 453630122021400	TE	°C	15.67	23.3	3.40	5.010	0.320	− 0.037
	BP	mmHg	763.5	781	747	3.798	0.005	− 0.091
	DIS	kcfs	211.3	506	72.1	91.85	0.435	0.700
	SFD	feet	77.74	303	0.00	58.37	0.751	0.929
	TDG	% sat.	112.3	131	98.0	6.307	0.056	1.000
USGS 453845121562000	TE	°C	16.39	23.5	6.40	4.498	0.274	− 0.426
	BP	mmHg	761.4	779	747	3.523	0.005	− 0.137
	DIS	kcfs	213.1	506	71.2	91.58	0.430	0.838
	SFD	feet	85.14	303	0.00	54.22	0.637	0.799
	TDG	% sat.	108.9	126	97.0	5.498	0.050	1.000
USGS 453845121564001	TE	°C	16.40	23.5	6.40	4.625	0.282	− 0.325
	BP	mm Hg	761.9	774	747	3.344	0.004	− 0.011
	DIS	kcfs	212.5	455	71.2	87.12	0.410	0.827
	SFD	feet	89.33	280	0.00	45.91	0.514	0.872
	TDG	% sat.	116.4	128	104	3.952	0.034	1.000

TDG, total dissolved gas; *TE*, water temperature; *BP*, barometric pressure; *SFD*, spill from dam; *DIS*, discharge; X_{mean}, mean; X_{max}, maximum; X_{min}, minimum; S_x, standard deviation; C_v, coefficient of variation; *R*, correlation coefficient with TDG; *kcfs*, thousands of cubic feet per second; *mmHg*, millimeter of mercury

database by power or exponential forms (Keshtegar and Heddam 2018; Keshtegar and Kisi; 2017; Keshtegar and Seghier 2018). Consequently, the highly nonlinear cross-correlations of the input data may improve the accuracy of the RSM as well as the modified RSM. In this current study, the high-nonlinear basis polynomial functions are investigated for the prediction of TDG using nonlinear forms of the basic RSM.

As seen from Eq. (1), the RSM is a simple and efficient modeling approach for predicting TDG, but this explicit mathematical form basis polynomial function with cross-linear terms may provide inaccurate results for complex processes. Therefore, the second-order polynomial basis set functions are needed to improve RSM in solving complex real engineering problems. The flexibility of RSM in Eq. (1) may be enhanced for predictions of TDG using the high-order polynomial basis function with second, third or fourth cross terms, which is presented by the following polynomial set functions:

$$T\hat{D}G = \underbrace{\underbrace{\underbrace{a_0 + \sum_{i\;1}^{n} a_i x_i + \sum_{i=1}^{n}\sum_{j=i}^{n} a_{ij} x_i\, x_j + \sum_{i\;1}^{n}\sum_{j\;1}^{n} b_{ij} x_i\, x_j^2}_{3-order} + \sum_{i\;1}^{n}\sum_{j\;1}^{n} d_{ij} x_i\, x_j^3}_{4-order} + \sum_{i=1}^{n}\sum_{j=1}^{n} k_{ij} x_i\, x_j^4}_{5-order} \qquad (4)$$

in which TĎG is the predicted TDG with high-order polynomial function (PF) with *n* input data set. a_0, a_i, a_{ij}, b_{ij}, d_{ij}, and k_{ij} are the unknown coefficients with a total number of coefficients:

$$NC = (n+1)(n+2)/2 + (Or{-}2)n^2 \qquad (5)$$

where (O_r) is the order of PF and 3 to 5 orders were used in this study. The least square estimator is generally applied to solve Eq. (4) using training data. Consequently, the predicted data using high-order PF is computed based on Eq. (2) where *P*(.) and $P(x_i)$ are determined by using the high-order polynomial basis function. For example, the high-order polynomial basis functions for three input data of x_1, x_2, x_3, and PF with 3-order is given as below:

$$P(x) = \begin{bmatrix} 1, x_1, x_2, x_3, x_1^2, x_1, x_2, x_1, x_3, x_2^2, x_2, x_3, x_3^2, \\ x_1^3, x_1, x_2^2, x_1, x_3^2, x_2, x_1^2, x_2^3, x_2, x_3^2, x_3, x_1^2 x_3, x_2^2, x_3^3 \end{bmatrix} \qquad (6)$$

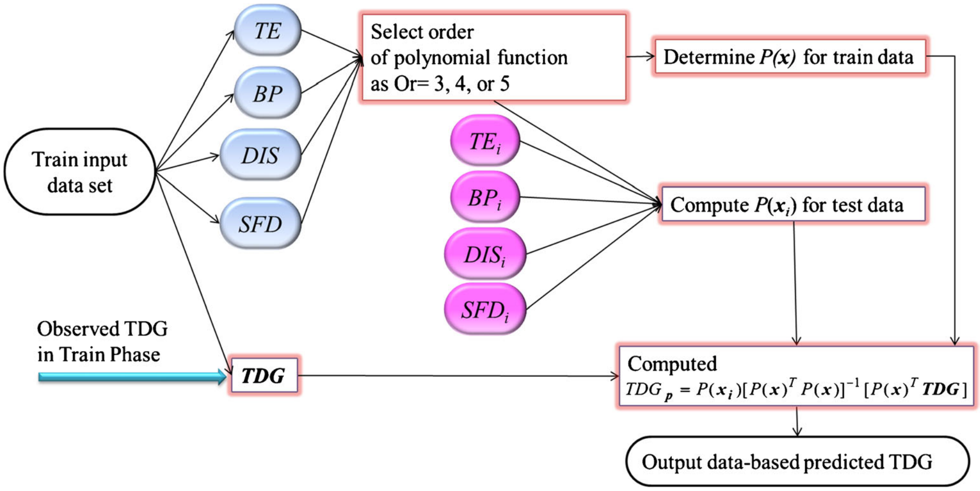
TE
BP
DIS
SFD
Train input data set
Select order of polynomial function as Or= 3, 4, or 5
Determine $P(x)$ for train data
TE_i
BP_i
DIS_i
SFD_i
Compute $P(\mathbf{x}_i)$ for test data
Observed TDG in Train Phase
TDG
Computed $TDG_p = P(x_i)[P(x)^T P(x)]^{-1}[P(x)^T \mathbf{TDG}]$
Output data-based predicted TDG

Fig. 1 The schematic view of the H-RSM for predictions of TDG

As seen from Eqs. (3) and (4), the high-order PF is structured by the high cross-correlation nonlinear of the input data in the H-RSM modeling approach. More accurate results may be obtained using the high-order PF compared with the second-order PF for complex processes. The steps of the predicted TDG using H-RSM are presented as below:

Step 1: Give the training databases including input data set as (x_i ∊ [TE, BP, SFD, and DIS]) and output TDG.

Step 2: Set the order of PF and compute $P(x)$ in terms of high-order polynomial basis function.

Step 3: Give the input data in test phase (x_i ∊ [TE$_i$, BP$_i$, SFD$_i$, and DIS$_i$]) and determine the high-order PF P (x_i).

Table 2 Performances of different models in modeling TDG at USGS 453439122223900 station

Models	Training					Validation				
	R	NSE	d	RMSE	MAE	R	NSE	d	RMSE	MAE
M5TRee1	0.983	0.966	0.991	0.970	0.617	0.897	0.796	0.947	2.272	1.751
M5TRee2	0.976	0.953	0.988	1.135	0.706	0.890	0.784	0.943	2.339	1.802
M5TRee3	0.969	0.939	0.984	1.293	0.851	0.892	0.788	0.944	2.316	1.809
M5TRee4	0.974	0.949	0.987	1.187	0.776	0.880	0.761	0.938	2.462	1.809
LSSVM1	0.926	0.857	0.960	1.979	1.543	0.922	0.849	0.958	1.955	1.508
LSSVM2	0.917	0.842	0.955	2.085	1.626	0.914	0.836	0.954	2.037	1.564
LSSVM3	0.918	0.843	0.956	2.079	1.627	0.915	0.837	0.954	2.027	1.572
LSSVM4	0.918	0.843	0.956	2.079	1.615	0.915	0.838	0.954	2.023	1.562
MARS1	0.929	0.862	0.962	1.944	1.517	0.923	0.838	0.959	2.024	1.564
MARS2	0.924	0.854	0.959	2.002	1.575	0.913	0.832	0.953	2.064	1.600
MARS3	0.926	0.858	0.961	1.972	1.544	0.905	0.797	0.949	2.270	1.690
MARS4	0.912	0.832	0.953	2.147	1.662	0.913	0.832	0.953	2.060	1.580
3H-RSM1	0.923	0.853	0.959	2.012	1.572	0.921	0.849	0.958	1.957	1.520
3H-RSM2	0.916	0.839	0.955	2.102	1.645	0.913	0.833	0.954	2.057	1.591
3H-RSM3	0.920	0.847	0.957	2.048	1.595	0.916	0.839	0.955	2.020	1.549
3H-RSM4	0.920	0.847	0.957	2.048	1.581	0.918	0.843	0.956	1.995	1.540
4H-RSM1	0.930	0.864	0.963	1.930	1.496	0.922	0.850	0.959	1.950	1.498
4H-RSM2	0.923	0.851	0.959	2.021	1.578	0.915	0.837	0.955	2.032	1.569
4H-RSM3	0.922	0.850	0.958	2.032	1.578	0.916	0.839	0.955	2.019	1.551
4H-RSM4	0.921	0.849	0.958	2.035	1.570	0.918	0.842	0.956	1.998	1.543
5H-RSM1	0.931	0.867	0.963	1.914	1.488	0.923	0.851	0.959	1.943	1.496
5H-RSM2	0.924	0.854	0.959	2.002	1.573	0.917	0.840	0.956	2.014	1.560
5H-RSM3	0.922	0.851	0.958	2.026	1.575	0.917	0.841	0.955	2.010	1.550
5H-RSM4	0.923	0.852	0.959	2.019	1.562	0.919	0.845	0.957	1.983	1.542

Step 4: Predict the TDG (TDG_p) using the training datasets of x, TDG, and test input data of x_i as:

$$TDG_p = P(x_i)\left[P(x)^T P(x)\right]^{-1}\left[P(x)^T \boldsymbol{TDG}\right] \tag{7}$$

The framework of H-RSM for prediction of the TDG is presented in Fig. 1. As seen from Fig. 1, the H-RSM is a simple modeling approach as well as the RSM. For applying the above steps to predict TDG, a program code was developed by MATLAB software.

Least squares support vector machine

Least squares support vector machine (LSSVM) proposed by Suykens and Vandewalle (1999) is a supervised machine learning model introduced as an improved version of the original SVM. The LSSVM possesses the general form of any regression model that works on linking a set of inputs variables (x_i) to one target output variable (y) by using the linear least squares (LLS) criteria rather than the convex quadratic programming (CQP) used for the SVM as follow (Suykens and Vandewalle 1999):

$$y = f(x) = w^t \phi(x) + b\left(w, x\in \mathrm{R}^{\mathrm{d}}\right) \tag{8}$$

for which w and b are the weights and biases, x is the matrix of input variables, and y is the target variable. Using the LSSVM, the objective function (OF) is calculated using the principle of structural risk minimization (SRM) (Zhu et al. 2018) as follow:

$$\mathrm{Min}_{w,\gamma}\,\mathrm{J}(w,\gamma) = \frac{1}{2}w^T w + \frac{1}{2}\gamma \sum_{k=1}^{n} e_k^2 \tag{9}$$

Consequently, Eq. (8) is written as follow:

$$y = w.\varphi(x) + b + e_k \tag{10}$$

where γ denotes regularization parameter and e_k denotes the random error between observed and calculated value, and $\varphi(x)$ is the kernel function Mercer conditions as follow (Zhu et al. 2018; Yang 2018):

$$k\left(x_i, x_j\right) = \varphi(x_i)^T \varphi\left(x_j\right) \tag{11}$$

Table 3 Performances of different models in modeling TDG at USGS 453630122021400 station

Models	Training					Validation				
	R	NSE	*d*	RMSE	MAE	*R*	NSE	*d*	RMSE	MAE
M5TRee1	0.993	0.986	0.996	0.749	0.447	0.951	0.902	0.975	1.998	1.453
M5TRee2	0.990	0.980	0.995	0.896	0.556	0.952	0.905	0.976	1.973	1.487
M5TRee3	0.987	0.974	0.993	1.015	0.657	0.943	0.887	0.971	2.147	1.576
M5TRee4	0.988	0.977	0.994	0.951	0.607	0.952	0.905	0.976	1.975	1.431
LSSVM1	0.964	0.930	0.981	1.660	1.285	0.962	0.925	0.980	1.750	1.354
LSSVM2	0.960	0.921	0.979	1.757	1.370	0.959	0.919	0.979	1.822	1.422
LSSVM3	0.957	0.916	0.977	1.820	1.394	0.958	0.918	0.978	1.832	1.429
LSSVM4	0.960	0.921	0.979	1.762	1.368	0.960	0.922	0.979	1.783	1.396
MARS1	0.966	0.934	0.983	1.610	1.241	0.961	0.924	0.980	1.762	1.388
MARS2	0.964	0.930	0.982	1.661	1.288	0.957	0.916	0.978	1.859	1.464
MARS3	0.962	0.925	0.980	1.714	1.328	0.960	0.922	0.979	1.790	1.389
MARS4	0.957	0.916	0.978	1.814	1.409	0.960	0.921	0.979	1.794	1.402
3H-RSM1	0.965	0.932	0.982	1.637	1.273	0.964	0.929	0.982	1.705	1.331
3H-RSM2	0.961	0.924	0.980	1.731	1.356	0.959	0.920	0.979	1.812	1.421
3H-RSM3	0.959	0.919	0.979	1.784	1.372	0.960	0.922	0.979	1.789	1.405
3H-RSM4	0.960	0.922	0.979	1.755	1.365	0.960	0.922	0.980	1.784	1.396
4H-RSM1	0.966	0.934	0.983	1.610	1.242	0.965	0.931	0.982	1.685	1.313
4H-RSM2	0.962	0.926	0.980	1.707	1.328	0.961	0.922	0.980	1.781	1.393
4H-RSM3	0.960	0.922	0.979	1.753	1.350	0.960	0.922	0.980	1.781	1.403
4H-RSM4	0.962	0.925	0.980	1.717	1.329	0.963	0.927	0.981	1.730	1.360
5H-RSM1	0.968	0.936	0.983	1.581	1.219	0.965	0.931	0.982	1.677	1.300
5H-RSM2	0.963	0.928	0.981	1.679	1.311	0.961	0.923	0.980	1.776	1.385
5H-RSM3	0.961	0.923	0.980	1.744	1.342	0.962	0.925	0.980	1.755	1.381
5H-RSM4	0.963	0.928	0.981	1.685	1.298	0.964	0.929	0.981	1.703	1.332

Table 4 Performances of different models in modeling TDG at USGS 453845121562000 station

Models	Training					Validation				
	R	NSE	*d*	RMSE	MAE	*R*	NSE	*d*	RMSE	MAE
M5TRee1	0.981	0.962	0.990	1.076	0.660	0.814	0.600	0.898	3.457	2.102
M5TRee2	0.971	0.943	0.985	1.309	0.850	0.877	0.756	0.935	2.702	2.056
M5TRee3	0.963	0.927	0.981	1.486	0.993	0.855	0.718	0.923	2.902	2.191
M5TRee4	0.967	0.934	0.983	1.409	0.929	0.858	0.720	0.925	2.893	2.150
LSSVM1	0.922	0.850	0.958	2.130	1.661	0.906	0.821	0.949	2.317	1.797
LSSVM2	0.901	0.811	0.945	2.387	1.874	0.896	0.801	0.943	2.439	1.940
LSSVM3	0.904	0.818	0.947	2.346	1.855	0.893	0.798	0.942	2.459	1.918
LSSVM4	0.900	0.809	0.945	2.400	1.881	0.881	0.776	0.934	2.589	1.989
MARS1	0.924	0.854	0.959	2.097	1.632	0.910	0.828	0.951	2.269	1.742
MARS2	0.908	0.825	0.950	2.299	1.798	0.892	0.794	0.942	2.482	1.923
MARS3	0.907	0.822	0.949	2.320	1.832	0.899	0.808	0.944	2.398	1.865
MARS4	0.889	0.791	0.939	2.515	1.976	0.877	0.769	0.931	2.629	2.028
3H-RSM1	0.920	0.845	0.957	2.161	1.686	0.909	0.826	0.951	2.280	1.772
3H-RSM2	0.900	0.809	0.945	2.400	1.883	0.895	0.801	0.943	2.442	1.944
3H-RSM3	0.903	0.815	0.947	2.364	1.875	0.896	0.802	0.943	2.431	1.908
3H-RSM4	0.896	0.803	0.943	2.443	1.920	0.880	0.775	0.934	2.597	1.992
4H-RSM1	0.922	0.850	0.958	2.131	1.658	0.907	0.823	0.950	2.301	1.782
4H-RSM2	0.903	0.816	0.947	2.358	1.845	0.895	0.799	0.943	2.451	1.939
4H-RSM3	0.904	0.816	0.947	2.355	1.868	0.894	0.800	0.942	2.446	1.910
4H-RSM4	0.901	0.811	0.945	2.387	1.867	0.885	0.783	0.937	2.548	1.950
5H-RSM1	0.924	0.853	0.959	2.105	1.641	0.911	0.829	0.952	2.263	1.751
5H-RSM2	0.906	0.820	0.949	2.331	1.826	0.899	0.807	0.946	2.402	1.912
5H-RSM3	0.904	0.818	0.948	2.344	1.854	0.897	0.805	0.944	2.417	1.892
5H-RSM4	0.903	0.816	0.947	2.357	1.845	0.887	0.786	0.939	2.530	1.924

Finally, the regression equation of the LSSVM model is obtained by a group of single linear equations as follow (Xiong et al. 2018; Wang et al. 2018):

$$f(x) = \sum_{i=1}^{n} \alpha_i k\left(x_i, x_j\right) + b \tag{12}$$

In the present study, LSSVM model is developed using the LS-SVMlab software (http: //www.esat.kuleuven.be/sista/lssvmlab/.).

Multivariate adaptive regression splines

One of the most and well-known adaptive, nonlinear and non-parametric regression models is certainly the multivariate adaptive regression spline (MARS) model introduced by Friedman (1991). The MARS is mainly used for mapping a set of predictors to a dependent variable using high-dimensional arguments (Nalcaci et al. 2018), with respect to the principal of "divide" and "conquers" method (Zhang et al. 2018). The MARS approach divides the input space into several subspaces and then for each subspace, a basic function (BF) is developed. Each BF takes into account the information provided by one or several predictors, and the BFs are fixed between two limits: "primary" and "end" points called "knote" (Arabameri et al. 2018). The BF relates the regressors to the dependent variable in the form of Max (0, *X*−c) or Max (0, *c*−*X*) (Friedman 1991), for which *c* is the threshold and *X* is the input variable. MARS model can be expressed as follow:

$$Y = f(x_i) = \psi_0 + j = \sum_{j=1}^{P} \sum_{b=1}^{B} \left[\psi_{jb}(+) Max\left(0, x_i - \mathrm{K}_{bj}\right) + \psi_{jb}(-) Max\left(0, \mathrm{K}_{bj} - x_j\right)\right] \tag{13}$$

where *x* is one of the predictors, *Y* is the dependent variable; *P* and *B* are the number of the predictors and the number of the generated BF, respectively, ψ_0 is constant or intercept, ψ_{jb} is the coefficient of the *j*th BF, and the *K* values are called knots (Arabameri et al. 2018). In the present study, MARS model is developed using the Matlab toolbox ARESLab (Jekabsons, 2016b).

M5Tree model

Inspired from the original regression trees (RT), the M5Tree model (M5Tree) was developed by Quinlan (1992) as a new

Table 5 Performances of different models in modeling TDG at USGS 453845121564001 station

Models	Training					Validation				
	R	NSE	*d*	RMSE	MAE	*R*	NSE	*d*	RMSE	MAE
M5TRee1	0.989	0.978	0.994	0.580	0.293	0.912	0.829	0.954	1.674	1.036
M5TRee2	0.984	0.969	0.992	0.687	0.35	0.917	0.840	0.957	1.621	1.002
M5TRee3	0.972	0.944	0.985	0.926	0.465	0.921	0.847	0.959	1.585	1.031
M5TRee4	0.978	0.957	0.989	0.813	0.454	0.887	0.782	0.94	1.889	1.265
LSSVM1	0.931	0.867	0.963	1.427	0.960	0.923	0.851	0.958	1.564	1.080
LSSVM2	0.926	0.857	0.960	1.478	0.991	0.922	0.850	0.955	1.566	1.090
LSSVM3	0.927	0.859	0.961	1.469	0.978	0.922	0.850	0.958	1.568	1.083
LSSVM4	0.915	0.837	0.953	1.579	1.148	0.905	0.819	0.948	1.720	1.287
MARS1	0.941	0.886	0.969	1.318	0.919	0.926	0.856	0.959	1.535	1.062
MARS2	0.940	0.884	0.968	1.331	0.926	0.920	0.846	0.956	1.589	1.112
MARS3	0.933	0.870	0.964	1.410	0.962	0.912	0.828	0.954	1.678	1.150
MARS4	0.909	0.826	0.950	1.631	1.181	0.907	0.822	0.950	1.707	1.283
3H-RSM1	0.925	0.856	0.960	1.481	0.999	0.925	0.856	0.960	1.534	1.065
3H-RSM2	0.922	0.849	0.958	1.518	1.038	0.923	0.852	0.960	1.556	1.101
3H-RSM3	0.922	0.850	0.958	1.513	1.024	0.922	0.850	0.959	1.568	1.105
3H-RSM4	0.909	0.826	0.951	1.630	1.197	0.905	0.818	0.949	1.726	1.312
4H-RSM1	0.933	0.871	0.965	1.404	0.951	0.932	0.868	0.964	1.470	1.039
4H-RSM2	0.930	0.865	0.963	1.435	0.985	0.929	0.864	0.962	1.495	1.058
4H-RSM3	0.927	0.859	0.961	1.465	0.983	0.926	0.858	0.961	1.524	1.064
4H-RSM4	0.917	0.841	0.956	1.557	1.117	0.917	0.841	0.955	1.613	1.207
5H-RSM1	0.936	0.876	0.966	1.375	0.939	0.933	0.871	0.964	1.456	1.022
5H-RSM2	0.933	0.869	0.964	1.413	0.974	0.932	0.868	0.964	1.472	1.031
5H-RSM3	0.928	0.860	0.962	1.461	0.981	0.927	0.859	0.961	1.519	1.061
5H-RSM4	0.920	0.845	0.957	1.537	1.102	0.916	0.839	0.955	1.622	1.203

machine learning technique. The M5Tree model builds recursively a RT model (Pulkknen et al. 2018), by dividing the space of the training data into several subsets, and a multivariate linear regression is formulated for each one, which relates the input variables to the dependent variable. The partitioning is achieved by recursive splits that minimize intra-subset variation (Ajmera and Goyal 2012). The M5Tree model contains three steps: splitting, creating, and extracting the knowledge from the tree (Fatehnia et al. 2016). During the first stage, a linear regression model is developed based on the division of the space of the data, and at the end of this step, the required information is provided that form the tree and the nodes are constructed. Generally, M5Tree model employs splitting criterion using the standard deviation reduction (SDR), calculated as follow (Pal and Deswal 2009; Fatehnia et al. 2016; Sattari et al. 2018):

$$SDR = sd(T) - \sum \frac{|T_i|}{T} sd(T_i) \qquad (14)$$

where T is the set of data points that reach the node, T_i denotes the subset of cases that have the ith outcome of the potential test, and sd represents the standard deviation of the observed values. In the present study, the M5Tree is implemented using the Matlab toolbox M5PrimeLab (Jekabsons, 2016a).

Performance assessment of the models

To evaluate and compare the accuracy of the developed models, we used five performance indices. These five indices are the following: the coefficient of correlation (R), the Nash-Sutcliffe efficiency (NSE), the Willmott index of agreement (d), the root mean squared error (RMSE), and the mean absolute error (MAE).

$$R = \left[\frac{\frac{1}{\mathrm{N}} \sum (\mathrm{O}_i - \mathrm{O}_m)(\mathrm{P}_i - \mathrm{P}_m)}{\sqrt{\frac{1}{\mathrm{N}} \sum_{i=1}^{n} (\mathrm{O}_i - \mathrm{O}_m)^2} \sqrt{\frac{1}{\mathrm{N}} \sum_{i-1}^{n} (\mathrm{P}_i - \mathrm{P}_m)^2}} \right] \qquad (15)$$

$$NSE = 1 - \frac{\sum_{i=1}^{N} [\mathrm{O}_i - P_i]^2}{\sum_{i=1}^{N} [\mathrm{O}_i - \mathrm{O}_m]^2} \qquad (16)$$

$$d = 1 - \frac{\sum_{i=1}^{N} (\mathrm{P}_i - \mathrm{O}_i)^2}{\sum_{i=1}^{N} (|\mathrm{P}_i - \mathrm{O}_m|| + ||\mathrm{O}_i - \mathrm{O}_m|)^2} \qquad (17)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(O_i - P_i\right)^2} \quad (18)$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|O_i - P_i\right| \quad (19)$$

where N is the data number, O_i is the measured TDG value, and P_i is the predicted TDG. O_m and P_m indicate the average of O_i and P_i.

Results and discussion

In this study, TDG measured at four dam reservoirs at Columbia River, USA, was predicted using four data-driven models as reported above. The developed models were LSSVM, MARS, M5Tree, and high-order RSM (H-RSM) with three different orders, from 3 to 5. The models were developed using four input variables, namely TE, BP, SFD, and DIS, served as predictors. The models were developed and evaluated using five statistical indices: R, NSE, d, RMSE, and MAE. Several combinations of the input variables were utilized and four scenarios were evaluated: (*i*) SFD, DIS, BP, and TE; (*ii*) SFD, DIS, and TE; (*iii*) SFD, DIS, and BP; (*iv*) SFD, BP, and TE. Consequently, LSSVM1, MARS1, M5TRee1, 3H-RSM1, 4H-RSM1, and 5H-RSM1 correspond to the first combination; LSSVM2, MARS2, M5Tree2, 3H-RSM2, 4H-RSM2, and 5H-RSM2 correspond to the second and so on, until the fourth combination. The performances of the simulated models compared with measured TDG are shown in Tables 2, 3, 4, and 5, illustrating the calculated statistical indices.

According to the obtained results, the following conclusions can be derived. First, comparison of models indicated that model 5H-RSM1 that includes the four input variables (SFD, DIS, BP, and TE) yields the best performance among the considered models with respect to R, NSE, d, RMSE, and MAE criteria, at four stations. Using 5H-RSM1, a good agreement is observed in all stations with R, NSE, and *d* ranging from 0.911 to 0.965, 0.829 to 0.931, and 0.952 to 0.982, respectively. The highest accuracy using the 5H-RSM1 was obtained at USGS453630122021400, while the lowest accuracy has been achieved at USGS 453845121562000. Similarly, the RMSE ranges from 1.456 to 2.263% with an average of 1.835%. In addition, the MAE ranges from 1.022 to 1.751% with an average of 1.392%. Second, it is clear from the results reported in Tables 2, 3, 4, and 5 that the M5Tree approach produced a model with low accuracy at the four

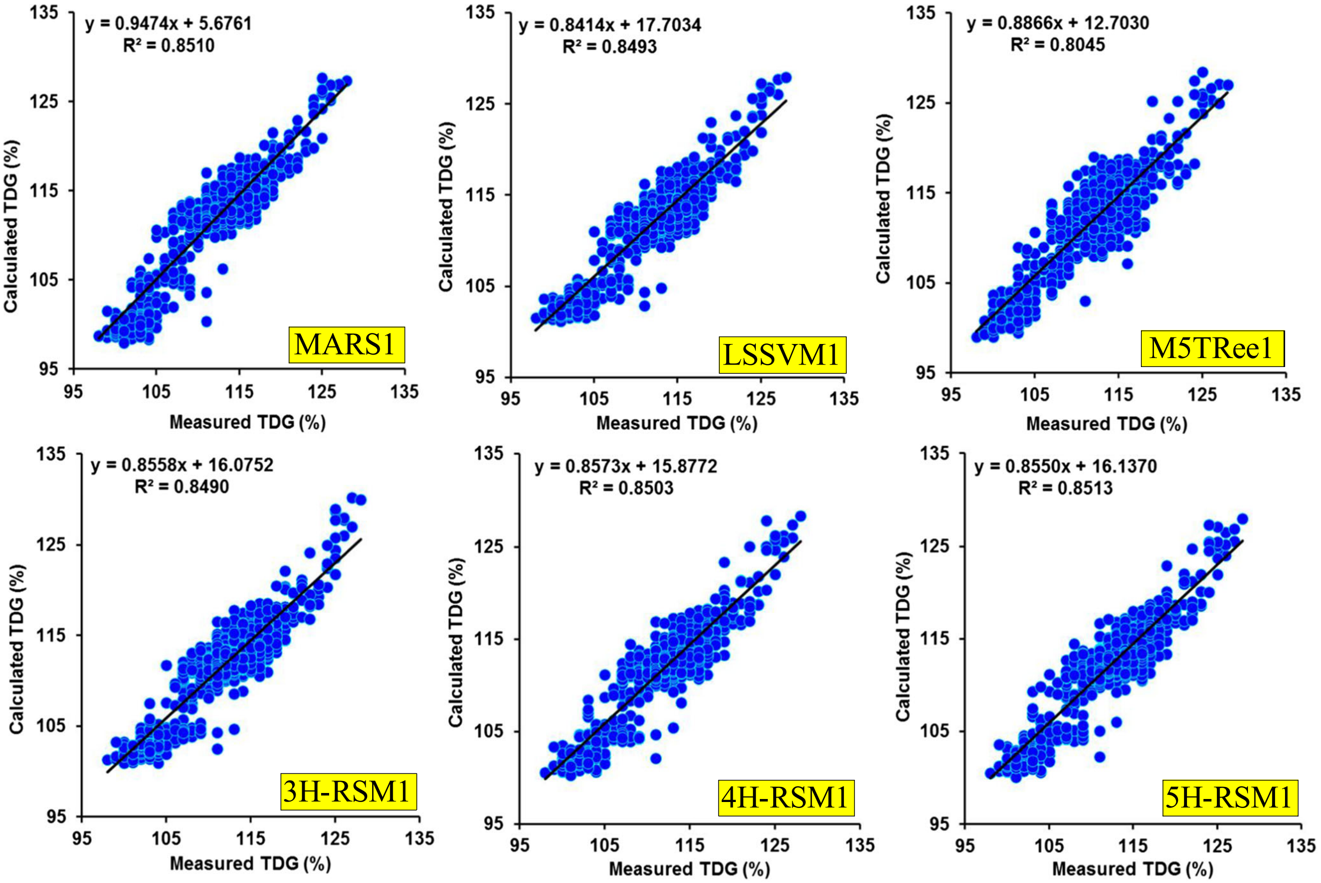

Fig. 2 Scatterplot of calculated versus measured TDG (%) for the optimum developed models during the validation phase: USGS 453439122223900

stations, with high RMSE and MAE, and low R, NSE, and d. Finally, although the 5H-RSM1 provided the high accuracy, it is evident that the 5H-RSM1 slightly improved the accuracy of the other three models which use only three input variables, and the improvement is marginal especially between 5H-RSM1 and 5H-RSM2, for which the MAE observed from validation data between the two models in the four stations tended to be of a similar magnitude. Similarly, RMSE values were fairly consistent across stations between the two models.

Results obtained at USGS453439122223900 station are reported in Table 2. Hereafter, we compared the performance (e.g., RMSE, MAE, R, NSEN, and d) of the H-RSM methods against the other three methods: LSSVM, MARS, and M5TRee. An accuracy assessment was conducted to evaluate the results. Table 2 showed that the best accuracy was obtained using the 5H-RSM1 slightly better than 4H-RSM1, 3H-RSM1, LSSVM1, and MARS1, and considerably higher than the M5TRee1. 5H-RSM1 yielded a result with an RMSE of 1.943%, an MAE of 1.496%, an R of 0.923, an NSE of 0.851, and an $d = 0.959$. For numerical comparison, using the 5H-RSM1, the RMSE of the M5TRee1 was decreased by 14.48%, the MAE is reduced by 14.56%, the R was lifted by 2.6%, the NSE was promoted by 5.5%, and finally the d is increased by 1.2%. For further analysis, when looking at the models with only three inputs, it is obvious that the best performance was obtained using combination number 4 with SFD, BP, and TE as input variables, with the exception of the M5Tree models so that the best accuracy with three input variables was found for the M5TRee3 with SFD, DIS, and BP as input variables. However, it is worth noting that the most significant improvement between four-input and three-input models was achieved using MARS and LSSVM. Four-input LSSVM1 decreased the RMSE and MAE of the three-input LSSVM4 model by 3.36% and 3.45%, respectively. In addition, MARS1 decreased the RMSE and MAE of the MARS4 by 1.748% and 1.013%, respectively. For the other approaches, the difference between the models with four and three input variables is completely negligible. Finally, according to Table 2, the four M5Tree models provided relatively similar accuracy with very marginal difference. Calculated TDG (%) using the best models are plotted against corresponding measurements values in Fig. 2.

For USGS453630122021400 station, the R, NSE, d, RMSE, and MAE statistics are provided in Table 3, for all models, with all input combinations. Among the all proposed methods, model 5H-RSM1 contains the smaller RMSE and MAE, with values equal to 1.677 and 1.300, respectively. Compared with the worst method, the 5H-RSM1 decreased

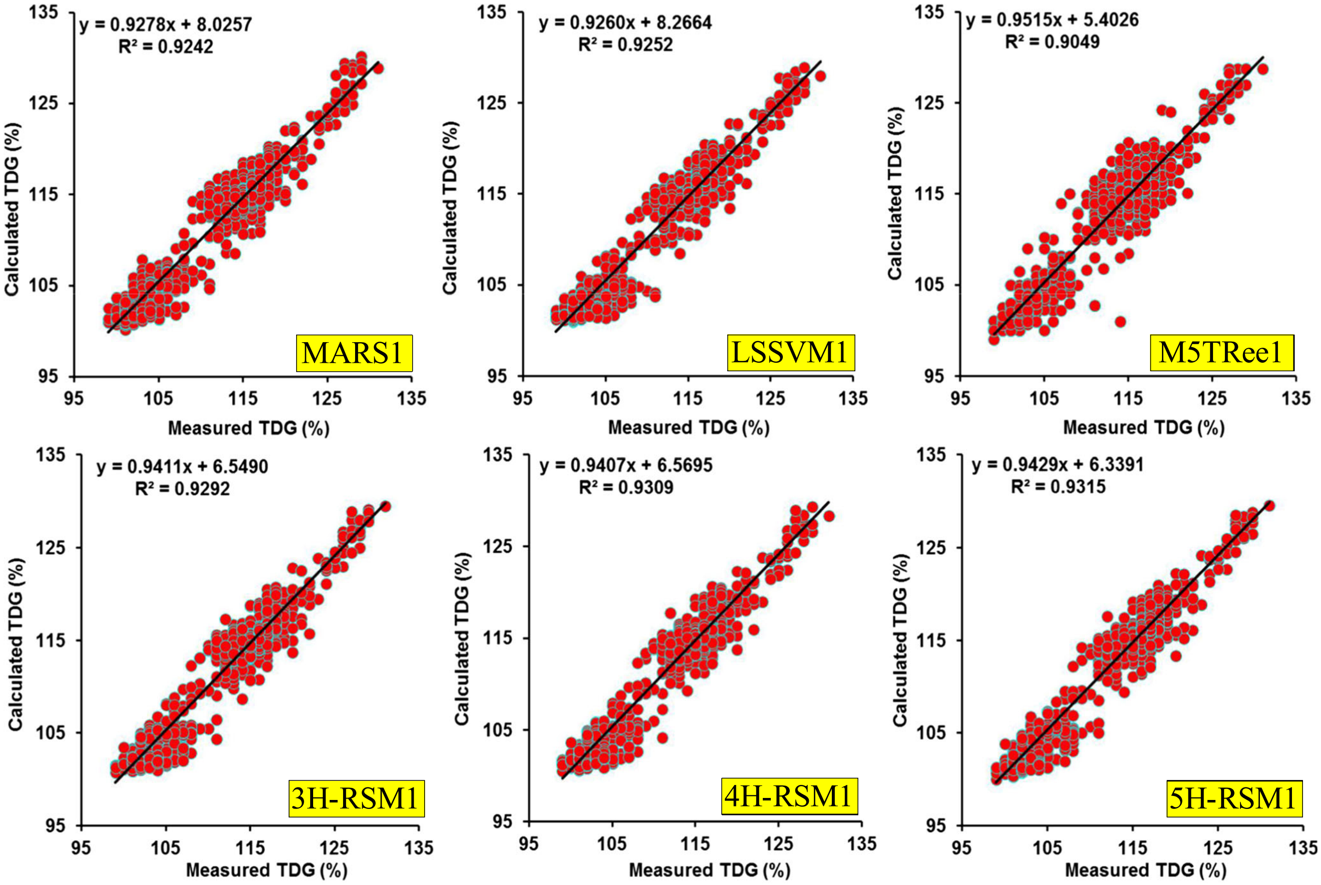

Fig. 3 Scatterplot of calculated versus measured TDG (%) for the optimum developed models during the validation phase: USGS 453630122021400

the RMSE and MAE of the M5Tree1 by 16.066% and 10.53%, respectively. According to Table 3, the results obtained demonstrated that combination of three input variables compared with the best model with four input variables slightly decreases the performance. Specifically, 5H-RSM4 with SFD, BP, and TE was slightly less than the 5H-RSM1 with negligible difference in the RMSE and MAE values. Based on the validation results in Table 3, the 5H-RSM, 4H-RSM, and 3H-RSM are promising for TDG modeling while the LSSVM and MARS are comparable to the H-RSM with relatively similar accuracy when using only three input variables. The statistical indices showed that there was no considerable difference between LSSVM and MARS predictions, but they were both considerably different from the M5Tree1 for TDG (%) modeling. LSSVM1 decreased the RMSE and MAE of the M5Tree1 by 12.41% and 6.81%, respectively. In addition, the MARS1 decreased the RMSE and MAE of the M5Tree1 by 11.81% and 4.47%, respectively. The 4H-RSM1 and 3H-RSM1 achieved a comparable result with similar R and d values ($R = 0.965$, $d = 0.982$) and slightly different values of NSE: 0.929 for 3H-RSM1 and 0.931 for 4H-RSM1. Figure 3 shows the linear regressions (scatterplot) between the calculated and the measured TDG (%) values using the validation dataset. As clearly observed from the scatter graphs, the 5H-RSM1 has less scattered estimates ($R^2 = 0.9315$) and its fit line is closer to the exact line (slope and bias of the fit line equation is closer to 1 and 0, respectively) compared with other models.

Table 4 reports the results of the proposed models for TDG prediction at USGS 453845121562000 station. The overall accuracy of the best models compared with the measured TDG was discussed hereafter. It can be seen from Table 4 that the 5H-RSM1 model has both the highest R, NSE, and d values ($R = 0.911$, NSE = 0.829, $d = 0.952$) and the lowest error measures (RMSE = 2.263 and MAE = 1.751). The 5H-RSM1 model is more accurate compared with the other models, is considerably higher than the M5TRee1, and it slightly improved the accuracy of 4H-RSM1, 3H-RSM1, LSSVM1, and MARS1. The 5H-RSM1 model reduced the MAE and RMSE measurements of the M5TRee1 with a percentage reduction of 34.54% and 16.70%, respectively. In addition, the values of R, NSE, and d of the M5TRee1 were promoted by 9.7%, 22.9%, and 5.4%, respectively, while for the 3H-RSM1 and 4H-RSM1, the values of RMSE, MAE, R, NSE, and d did not show considerable differences. MARS1 model predictions are slightly more accurate than those from LSSVM1 in terms of the all five statistical indices. Utilizing the MARS1 model resulted in a reduction of RMSE and MAE of about 2% and 3%, respectively, compared with the

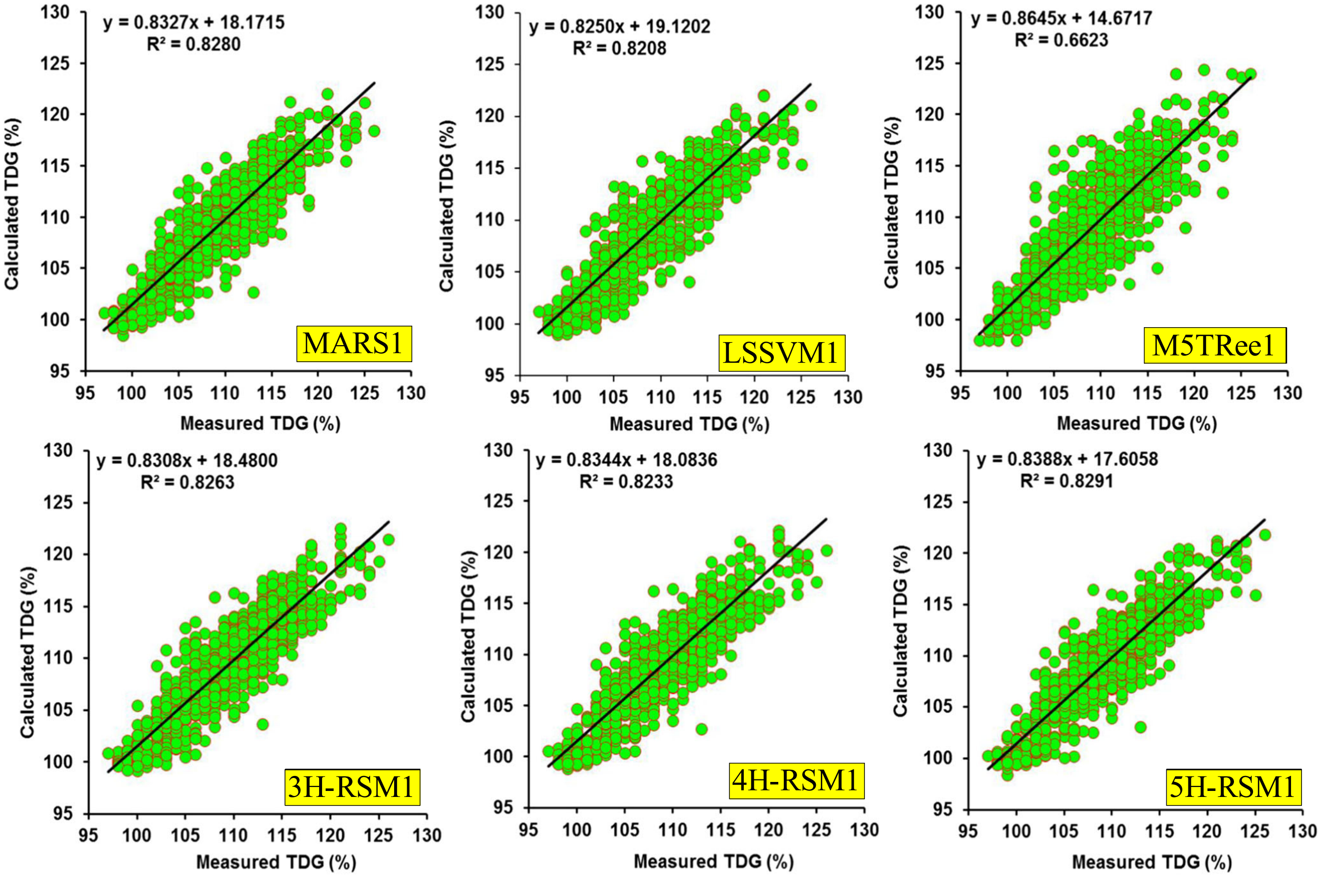

Fig. 4 Scatterplot of calculated versus measured TDG (%) for the optimum developed models during the validation phase: USGS 453845121562000

LSSVM1. Table 4 indicates that, on the average, the results from the two input combinations (combinations 2 and 3) differ by only a few. Table 4 highlights a better performance of the MARS3 having input variables SFD, DIS, and BP, with respect to all statistical indices. For instance, MARS3 has RMSE, MAE, *R*, NSE, and *d* of 2.398, 1.865, 0.899, 0.808, and 0.944, respectively, slightly superior to the values provided by the LSSVM3, 5H-RSM3, 4H-RSM3, and 3H-RSM3 and significantly better than the M5TRee3. Contrary to the two previous stations, the lowest accuracy was obtained by the models using combination 4, with SFD, BP, and TE. Overall, except the models using all the four input variables, for which the best accuracy was obtained, the models with three input variables provided relatively similar accuracy with low differences. Figure 4 shows the scatterplot of measured and predicted TDG, for the six best models. Here also the less scattered estimates of the 5H-RSM1 model can be clearly seen especially for the peak TDG values compared with other models. This confirms the lower RMSE values (see Table 4) of this model than the alternative models.

Table 5 summarizes the results of applied models in terms of statistical indices at USGS453845121564001. Clearly, the results suggest that the 5H-RSM1 model consistently outperforms the other models and produces high accuracies in terms of *R*, NSE, *d*, RMSE, and MAE; however, the results also show that the 5H-RSM1 and 4H-RSM1 give similar accuracy regarding all the statistical indices. In addition, the difference among the 3H-RSM1, LSSVM1, and MARS1 models are marginal. MARS1 had the second best accuracy, and LSSVM1 performed less than the MARS1 and better than the M5TRee1. Similarly, poor accuracy was obtained using M5TRee1 compared with that obtained using all developed models. The M5TRee1 model yielded the substantially lowest *R*, NSE, and *d* and the highest RMSE and MAE. Using three input variables, the best accuracy was obtained using 5H-RSM2 with SFD, DIS, and TE as input variables. However, the RMSE and MAE differences between the proposed models are small when the SFD, DIS, and BP are included as predictor variables suggesting that the relationships between TDG and those predictor variables can be captured equally by all models, and none of them was capable to provide an accuracy improvement. Comparison of first and second input combinations reveals that including BP variable in inputs considerably increases the models' accuracies in estimating TDG even though there is a very low correlation between BP and TDG as seen from Table 1. This is also valid for the other three stations. This indicates that there might be nonlinear relationship between the BP and TDG variables.

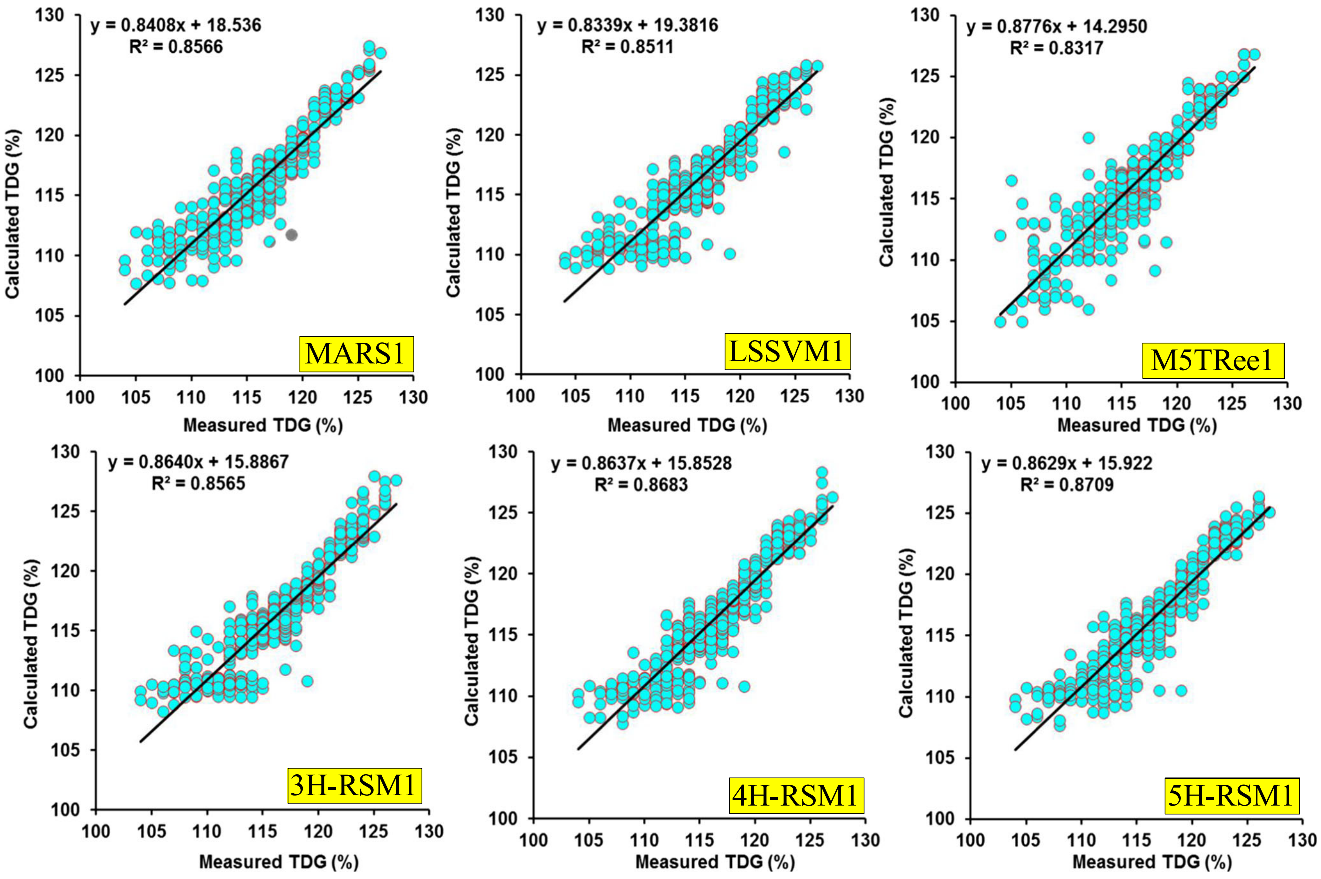

Fig. 5 Scatterplot of calculated versus measured TDG (%) for the optimum developed models during the validation phase: USGS 453845121564001

Figure 5 shows the scatterplot of measured and calculated TDG using the best models. The figure clearly shows the superiority of the 5H-RSM1.

Conclusions

The investigation presented in this paper has demonstrated the potential of data-driven models to accurately predict an important variable at dam's reservoirs: total dissolved gas (TDG) concentration. The estimation of TDG was made possible through the exploitation of large data base freely available and contains several easily measured variables. Providing such kind of models can be of great interest, compared with the physical and numerical models which require a large number of variables to be calibrated. Four models were developed and compared, three well known and widely reported in the literature as a powerful tool for solving several environmental problems (LSSVM, MARS, and M5Tree) and one new model (RSM) presented for the first time as a powerful tool. While the demonstration of the usefulness and robustness of the proposed approaches is based on data collected at four USGS stations, the generalization of the methods is relatively straightforward, in the light of all data-driven models. Several conclusions can be drawn at the closing of this investigation. Firstly, among the proposed four models, M5Tree models had the lowest accuracy at the four stations, and this leads us to conclude that M5Tree was unable to provide high and precise accuracy for modeling TDG concentration. Secondly, LSSVM and MARS models provided relatively similar accuracy with slightly and marginal difference. Finally, high-order response surface method (5H-RSM) using all the four input variables successfully built the relationship between TDG concentration and the selected predictors. The high-order RSM provides the high correlation terms based on the polynomial functions while the coefficient vector is increased by increasing the polynomial term basis high-order consideration. According to these reasons, applying the high-order correlation of basis input data may improve the accuracy prediction of nonlinear problems compared with original RSM. However, the H-RSM needs more data point for training a nonlinear model due to the increasing number of coefficients, which could provide a useful method with high flexibility for both accuracy and efficiency. However, it can be applied for modeling the problems with smaller input variables with larger training database. Furthermore, it was demonstrated that the performance of the 5H-RSM was only slightly decreased when the model included fewer input variables.

Acknowledgments We would like to thank all scientists from USGS for allowing permission for using the data that made this study possible. Once again, we would like to thank anonymous reviewers and the editor of Arabian Journal of Geosciences (AJGS) for their invaluable comments and suggestions on the contents of the manuscript which significantly improved the quality of the paper.

Compliance with ethical standards

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

References

Arabameri A, Pradhan B, Pourghasemi H, Rezaei K, Kerle N (2018) Spatial modelling of gully erosion using gis and r programing: a comparison among three data mining algorithms. Appl Sci 8(8): 1369. https://doi.org/10.3390/app8081369

Ajmera TK, Goyal MK (2012) Development of stage-discharge rating curve using model tree and neural networks: an application to Peachtree Creek in Atlanta. Expert Syst Appl 39(5):5702–5710. https://doi.org/10.1016/j.eswa.2011.11.101

Alikhuni KH, Ramachandran V, Chaudhri H (1951) Mortality of carp fry under supersaturation of dissolved oxygen in water. Proc Natl Inst Sci India 17:261–264

Beeman JW, Venditti DA, Morris RG, Gadomski DM, Adams BJ, Vanderkooi SJ, Robinson TC, Maule AG (2003) Gas bubble disease in resident fish below Grand Coulee Dam: final report of research. US Bureau of Reclamation https://pubs.er.usgs.gov/publication/70179865

Beeman JW, Maule AG (2006) Migration depths of juvenile Chinook salmon and steelhead relative to total dissolved gas supersaturation in a Columbia River reservoir. Trans Am Fish Soc 135:584–594. https://doi.org/10.1577/T05-193.1

Bragg HM, Johnston MW (2016). Total dissolved gas and water temperature in the lower Columbia River, Oregon and Washington, water year 2015: U.S. Geological Survey Open-File Report 2015-1212, p 26. 10.3133/ofr20151212.

Boyd CE, Watten B, Goubier V, Wu R (1994) Gas supersaturation in surface waters of aquaculture ponds. Aquac Eng 13(1):31–39. https://doi.org/10.1016/0144-8609(94)90023-X

CCME (1999) Canadian Council of Ministers of the Environment, Canadian water quality guidelines for the protection of aquatic of aquatic life: dissolved gas supersaturation. Canadian Environmental Quality Guidelines http://ceqg-rcqe.ccme.ca/download/en/176/

Colt J (1986) Gas Supersaturation-impact on the design and operation of aquatic systems. Aquac Eng 5:49–85

Fatehnia M, Tawfiq K, Ye M (2016) Estimation of saturated hydraulic conductivity from double-ring infiltrometer measurements. Eur J Soil Sci 67(2):135–147. https://doi.org/10.1111/ejss.12322

Feng JJ, Li R, Yang HX, Li J (2013) A laterally averaged two-dimensional simulation of unsteady supersaturated total dissolved gas in deep reservoir. J Hydrodyn 25(3):396–403. https://doi.org/10.1016/S1001-6058(11)60378-9

Friedman JH (1991) Multivariate adaptive regression splines. Ann Stat 19(1):1–67. https://doi.org/10.1214/aos/1176347963

Geldert DA, Gulliver JS, Wilhelms SC (1998) Modeling dissolved gas supersaturation below spillway plunge pools. ACSE J Hyd Eng 124(5):513–521. https://doi.org/10.1061/(ASCE)0733-9429(1998)124:5(513)

Gorham FP (1898) Some physiological effects of reduced pressure on fishes. J Boston Sot Med Sci 3:50 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2121825/pdf/jbsms00023-0019.pdf.

Hibbs DE, Gulliver JS (1997) Prediction of effective saturation concentration at spillway plunge pools. ACSE J Hyd Eng 123(11):940–949. https://doi.org/10.1061/(ASCE)0733-9429(1997)123:11(940)

Hadjerioua B, Pasha MD, Stewart KM, Bender M, Schneider ML (2012). Prediction of total dissolved gas exchange at hydropower dams (No. ORNL/TM-2011/340). Oak Ridge National Laboratory (ORNL). https://info.ornl.gov/sites/publications/Files/Pub32242.pdf.

Heddam S (2017) Generalized regression neural network based approach as a new tool for predicting total dissolved gas (TDG) downstream of spillways of dams: a case study of Columbia River Basin Dams, USA. Environ Process 4:235–253. https://doi.org/10.1007/s40710-016-0196-5

Jekabsons G (2016b) ARESLab Adaptive regression splines toolbox for Matlab/Octave ver. 1.13.0. Institute of Applied Computer Systems Riga Technical University, Latvia Available: http://www.cs.rtu.lv/jekabsons/Files/ARESLab.pdf

Jekabsons G (2016a) M5PrimeLab: M5' regression tree and model tree ensemble toolbox for Matlab/Octave ver. 1.7.0. Institute of Applied Computer Systems Riga Technical University, Latvia Available: http://www.cs.rtu.lv/jekabsons/Files/M5PrimeLab.pdf

Keshtegar B, Kisi O (2017) Modified response-surface method: new approach for modelling pan evaporation. J Hydrol Eng 22(10): 04017045. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001541

Keshtegar B, Seghier MEAB (2018) Modified response surface method basis harmony search to predict the burst pressure of corroded pipelines. Eng Fail Anal 89:177–199. https://doi.org/10.1016/j.engfailanal.2018.02.016

Keshtegar B, Heddam S (2018) Modeling daily dissolved oxygen concentration using modified response surface method and artificial neural network: a comparative study. Neural Comput & Applic 30(10):2995–3006. https://doi.org/10.1007/s00521-017-2917-8

Li S, Kazemi H, Rockaway TD (2019) Performance assessment of stormwater GI practices using artificial neural networks. Sci Total Environ 651:2811–2819. https://doi.org/10.1016/j.scitotenv.2018.10.155

Moghaddasi MR, Noorian-Bidgoli M (2018) ICA-ANN, ANN and multiple regression models for prediction of surface settlement caused by tunneling. Tunn Undergr Space Technol 79:197–209. https://doi.org/10.1016/j.tust.2018.04.016

Marsh H.C., Gorham F.P. (1904). The gas disease in fishes. Rep US Bur Fish, pp. 343-376.

Nalcaci G, Özmen A, Weber GW (2018) Long-term load forecasting: models based on MARS, ANN and LR methods. Central Eur J Oper Res:1–17. https://doi.org/10.1007/s10100-018-0531-1

Orlins JJ, Gulliver JS (2000) Dissolved gas supersaturation downstream of a spillway II: computational model. J Hydraul Res 38(2):151–159. https://doi.org/10.1080/00221680009498350

Pal M, Deswal S (2009) M5 model tree based modelling of reference evapotranspiration. Hydrol Process 23(10):1437–1443. https://doi.org/10.1002/hyp.7266

Parker NC, Suttle MA, Fitzmayer K (1984) Total gas pressure and oxygen and nitrogen saturation in warmwater ponds aerated with airlift pumps. Aquac Eng 3(2):91–102. https://doi.org/10.1016/0144-8609(84)90001-3

Picket J., Rueda H., Herold M. (2004). Total maximum daily load for total dissolved gas in the Mid-Columbia River and Lake Roosevelt. Submittal Report. No. 04-03-002, Washington State Department of Ecology, Olympia, WA. http://www.ecy.wa.gov/biblio/0403002.html.

Politano M, Carrica PM, Turan C, Weber L (2007) A multidimensional two phase flow model for the total dissolved gas downstream of spillways. J Hydraul Res 45(2):165–177. https://doi.org/10.1080/00221686.2007.9521757

Politano M, Carrica P, Weber L (2009) A multiphase model for the hydrodynamics and total dissolved gas in tailraces. Int J Multiphase Flow 35:1036–1050. https://doi.org/10.1016/j.ijmultiphaseflow.2009.06.009

Politano M, Arenas Amado A, Bickford S, Murauskas J, Hay D (2012) Evaluation of operational strategies to minimize gas supersaturation downstream of a dam. Comput Fluids 68:168–185. https://doi.org/10.1016/j.compfluid.2012.08.003

Politano M, Lyons T, Anderson K, Parkinson S, Weber L (2016). Spillway deflector design using physical and numerical models. 6th International Symposium on Hydraulic Structures Portland, Oregon, USA, 27-30 June 2016. Hydraulic Structures and Water System Management. ISBN 978-1-884575-75-4. 10.15142/T3470628160853.

Politano M, Castro A, Hadjerioua B (2017) Modeling total dissolved gas for optimal operation of multireservoir systems. J Hydraul Eng. https://doi.org/10.1061/(ASCE)HY.1943-7900.0001287

Pulkknen M, Ginzler C, Traub B, Lanz A (2018) Stereo-imagery-based post-stratification by regression-tree modelling in Swiss National Forest Inventory. Remote Sens Environ 213:182–194. https://doi.org/10.1016/j.rse.2018.04.052

Quinlan J.R. (1992). Learning with continuous classes. In: Proceedings of the Fifth Australian Joint Conference on Artificial Intelligence, Hobart, Australia, 16-18 November. World Scientific, Singapore, pp. 343-348.

Roesner LA, Norton WR (1971). Nitrogen Gas (N2) Model for the Lower Columbia River. Water Resources Engineers. Report N° 1-350, water resources Engineers, Inc., Walnut Creek, Calif

Sattari MT, Mirabbasi R, Sushab RS, Abraham J (2018) Prediction of groundwater level in Ardebil plain using support Vector regression and M5 tree model. Groundwater 56(4):636–646. https://doi.org/10.1111/gwat.12620

Stewart KM, Witt A, Hadjerioua B (2015) Total dissolved gas prediction and optimization in riverware. Prepared for US Department of Energy Wind and Water Program by Oakridge National Laboratory, Oak Ridge https://info.ornl.gov/sites/publications/Files/Pub59285.pdf

Suykens JAK, Vandewalle J (1999) Least square support vector machine classifiers. Neural Processing Letters 9 (3), 293-300 https://doi.org/10.1023/A:1018628609742

Shaw P (1998) In: University of Washington (ed) Gas generation equations for CRiSP 1.6, Seattle, Washington www.cbr.washington.edu/d_gas/tdg_manual.pdf

Skov PV, Pedersen LF, Pedersen PB (2013) Nutrient digestibility and growth in rainbow trout (Oncorhynchus mykiss) are impaired by short term exposure to moderate supersaturation in total gas pressure. Aquaculture 416:179–184. https://doi.org/10.1016/j.aquaculture.2013.09.007

Tanner DQ, Bragg HM, Johnston MW (2012). Total dissolved gas and water temperature in the lower Columbia River, Oregon and Washington, water year 2011: quality-assurance data and comparison to water-quality standards: U.S. Geological Survey Open-File Report 2011-1300, p 28. http://pubs.usgs.gov/of/2011/1300

Tawfik ME, Diez FJ (2014) On the relation between onset of bubble nucleation and gas supersaturation concentration. Electrochim Acta 146:792–797. https://doi.org/10.1016/j.electacta.2014.08.147

Weitkamp DE, Sullivan RD, Swant T, DosSantos J (2003) Gas bubble disease in resident fish of the lower Clark Fork River. Trans Am Fish Soc 132(5):865–876. https://doi.org/10.1577/T02-026

Wang Y, Politano M, Weber L (2018a) Spillway jet regime and total dissolved gas prediction with a multiphase flow model. J Hydraul Res 57:26–38. 1-13. https://doi.org/10.1080/00221686.2018.1428231

Witt A, Stewart K, Hadjerioua B (2017a) Predicting total dissolved gas travel time in hydropower reservoirs. J Environ Eng 143(12): 06017011. https://doi.org/10.1061/(ASCE)EE.1943-7870.0001281

Wang P, Liu C, Li Y (2018b) Estimation method for ET0 with PSO-LSSVM based on the HHT in cold and arid data-sparse area. Clust Comput. https://doi.org/10.1007/s10586-018-1726-x

Witt A, Magee T, Stewart K, Hadjerioua B, Neumann D, Zagona E, Politano M (2017b) Development and implementation of an optimization model for hydropower and total dissolved gas in the mid-Columbia River System. J Water Resour Plan Manag 143(10): 04017063. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000827

Xiong J, Wang T, Li R (2018).Research on a hybrid LSSVM intelligent algorithm in short term load forecasting. https://doi.org/10.1007/s10586-018-1740-z.

Yuan Y, Feng J, Li R, Huang Y, Huang J, Wang Z (2018) Modelling the promotion effect of vegetation on the dissipation of supersaturated total dissolved gas. Ecol Model 386:89–97. https://doi.org/10.1016/j.ecolmodel.2018.08.016

Yang J. (2018). A novel short-term multi-input-multi-output prediction model of wind speed and wind power with LSSVM based on improved ant colony algorithm optimization. Clust Comput, 1-8. https://doi.org/10.1007/s10586-018-2107-1.

Zhu X, Ma SQ, Xu Q (2018) A WD-GA-LSSVM model for rainfall-triggered landslide displacement prediction. J Mt Sci 15(1):156–166. https://doi.org/10.1007/s11629-016-4245-3

Zhang W, Zhang R, Goh AT (2018) Multivariate adaptive regression splines approach to estimate lateral wall deflection profiles caused by braced excavations in clays. Geotech Geol Eng 36(2):1349–1363. 1-15. https://doi.org/10.1007/s10706-017-0397-3