Desalination and Water Treatment
www.deswater.com

doi: 10.5004/dwt.2019.24158

157 (2019) 79–89
July

Biochemical oxygen demand prediction in wastewater treatment plant by using different regression analysis models

Osman Tugrul Baki[a,*], Egemen Aras[b], Ummukulsum Ozel Akdemir[c], Banu Yilmaz[a]

[a]*Faculty of Technology, Department of Civil Engineering, Karadeniz Technical University, 61080 Trabzon, Turkey, Tel. +90 462 377 84 40; emails: tbaki@ktu.edu.tr (O.T. Baki), banuyilmaz@ktu.edu.tr (B. Yilmaz)*
[b]*Faculty of Engineering and Naturel Sciences, Department of Civil Engineering, Bursa Technical University, 16310 Bursa, Turkey, email: egemen.aras@btu.edu.tr*
[c]*Department of Civil Engineering, Giresun University, Faculty of Engineering, Giresun, Turkey, email: ummukulsum.akdemir@giresun.edu.tr*

Received 11 June 2018; Accepted 22 March 2019

ABSTRACT

The management and operation of the wastewater treatment plants (WWTP) have an important role in the controlling and monitoring of the plants' operations. Various performance data are taken into account in the controlling of the WWTP. The irregularities between operating parameters often lead to management problems that cannot be overcome. The aim of this study is to provide a simple and reliable prediction model to estimate the biochemical oxygen demand (BOD) with specific water quality parameters like wastewater temperature, pH, chemical oxygen demand, suspended sediment, total nitrogen, total phosphorus, electrical conductivity, and input discharge. The data records in this study were measured between June 2015 and May 2016 and obtained from the laboratory of Antalya Hurma WWTP. In the creation of the model, classical regression analysis, multivariate adaptive regression splines (MARS), artificial bee colony, and teaching-learning based optimization were used. The root mean square error and the mean absolute error were used to evaluate performance criteria for each model. When the results of the analyses were compared with each other, it was observed that the MARS method gave better estimation results than the other methods used in the study. As a result, it was evinced that the MARS method produces acceptable results in the BOD estimation.

Keywords: Biochemical oxygen demand; Wastewater treatment plant; Heuristic regression; Optimization algorithm

1. Introduction

Water is the basis of all biological and human activities. In the 21st century, the importance of water and water pollution is increasing day by day with the developments in technology and industry, irregular urbanization, increasing use of water resources together with increasing population. As a consequence of environmental pollution caused by humans, harmful substances are transported via water to regions hundreds of kilometers away. This polluted water cause pollution by interfering with other water and water sources in which in turn limits the living areas in aquatic environments.

Therefore, the necessity of collecting wastewater without harming human and environmental health and removing it from the receiving environment without causing any destruction has emerged. By treating wastewater, it is intended to prevent illnesses caused by wastewater and, pollution, and the damage to the environment into which the wastewater is discharged. The assessment of water quality parameters plays an important role in the management and performance evaluation of the wastewater treatment plants (WWTP).

* Corresponding author.

Biochemical oxygen demand (BOD) is an effective parameter in the planning and management of WWTPs. BOD is an approximate measure of the amount of biochemically degradable organic matter present in a water sample. Performing the BOD test involves the stages of preparation and analysis that require responsibility. This test takes approximately 5 d. The cost of the analysis increases if the experiments conducted as part of the test become too long and the measurements involve difficulties.

BOD is determined using laboratory tests. The advantage of laboratory tests is that the BOD can be determined accurately. Even though laboratory tests provide more accuracy, they are time intensive, demand commitment for preparation and analysis and a number of days are required in order to obtain and interpret the results of these tests. BOD varies depending on the characterization of the wastewater. Therefore, when the water quality changes quickly in unexpected and extreme cases, the results of the analysis may no longer be relevant for the current wastewater of the WWTP. This can result in a major failure in treatment of the water in the WWTP [1].

The control and safe operation of a WWTP can be achieved by developing a modeling tool for predicting the plant performance based on past observations of certain key product quality parameters. WWTPs involve several complex physical, biological and chemical processes. Often these processes exhibit nonlinear behaviors which are difficult to describe by linear mathematical methods [2].

2. Literature review

Recent studies have been directed toward developing a model that could quickly and reliably predict water quality parameters such as BOD. There are statistical and deterministic methods in modeling on water quality [3–5]. Multivariate statistical techniques such as factor analysis, principal component analysis, cluster analysis, and multiple regression analysis are widely used for water quality assessment [6–11]. In recent years, the development of environmental models has been provided through computer software and these models have started to be used universally in wastewater engineering [12–14].

In many studies conducted over the years, artificial neural networks (ANN), which are one of the soft computing methods, have shown very successful results in the context of wastewater prediction. Recent experiments have shown that ANN may be an alternative in estimating BOD [1,15–21]. According to Guclu, the ANN model trained the dynamic behavior of non-linear and complex WWTP processes satisfactorily [17].

In the same way, different intelligence techniques such as genetic algorithm [22], support vector machine [23], and adaptive neuro-fuzzy inference system [24–27] are used in the estimation of wastewater engineering applications. As a result of the literature review (Table 1), it was found that soft computing methods were commonly found on used for the BOD prediction. Studies using regression analysis were also found in the literature, however these studies were performed comparatively with ANN. In the literature review, regression analysis methods were not compared with each other.

In this study, the classical regression analysis (CRA), multivariate adaptive regression splines (MARS), artificial bee colony (ABC) and teaching-learning based optimization (TLBO) techniques were employed in the estimation of BOD. It was aimed to reveal the empirical relationship of BOD with plant operating parameters. The predictive capabilities of the obtained models were compared with each other. The main advantages of these methods, which are population based in order to give the optimum solution of the plant, can be summarized as; broad applicability, robust to dynamic changes, hybridization with other methods, ability to solve a problem that have no solution, high flexibility, being applicable to multidimensional optimization problems. In this study, it was aimed to estimate BOD by using water quality parameters and input flow which has a short measurement period.

3. Methods

In this part of the study, models were created to estimate BOD with different regression methods in the inlet pool located in Antalya Hurma WWTP. The input discharge (Q), wastewater temperature (t), pH, chemical oxygen demand (COD), suspended sediment (SS), total nitrogen (tN), total phosphorus (tP) and, electrical conductivity (EC) parameters were utilized as input parameters in BOD prediction. Four different methods which are; the CRA, ABC, TLBO, and MARS algorithms were used for estimation.

3.1. Classical regression analysis

Regression analysis is a statistical analysis technique which is used frequently to determine the relationship between two or more variables that have cause–effect relationship by using a mathematical function and to make estimation or prediction about the dependent variable. A mathematical model is used to explain the relationship between regression analysis and the dependent and independent variables and this model is called a regression model. This mathematical model can be univariate (simple) and multivariate (multiple). The model can be linear or curved.

Since the change of the BOD parameter is not related to a linear function (LF), three types of regression model were used in the study to estimate this parameter. In addition to the LF, the exponential function (EF) and power function (PF) were used in the study. Within the generated models, the regression coefficients were found by using the models giving the best result in the training set and the same steps were applied in the test set. The functions used to set the regression models are given below. The functions respectively expressed with,

$$y_{\mathrm{LF}} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_m x_m \tag{1}$$

$$y_{\mathrm{EF}} = b_0 + \exp\left(b_1 + b_2 x_1 + b_3 x_2 + \ldots + b_m x_{m-1}\right) \tag{2}$$

$$y_{\mathrm{PF}} = b_0 \times x_1^{b_1} \times x_2^{b_2} \times \ldots \times x_m^{b_m} \tag{3}$$

The functions used are expressed by the equations given above where; y shows the estimated value when the independent variables are x_1, x_2, …, x_m and regression coefficients are b_0, b_1, …, b_m [39].

Table 1
The input parameters used in previous studies

References	The name of parameters																		
	COD	Q	SS	TS	TN	TP	DO	t	pH	Od	Col	EC	MLSS	NH_3	NO_2	NO_3	O&G	BOD_i	Chl-a
Heddam et al. [1]	*		*					*	*			*							
Tomic et al. [5]	*		*			*	*	*	*			*				*			
Oliveira-Esquerre et al. [15]	*	*	*					*	*		*	*		*		*			
Doğan et al. [18]	*	*	*		*	*													
Baki and Aras [21]	*	*	*		*	*		*	*			*							
Hamed et al. [28]			*															*	
G. Onkal-Engin et al. [29]										*									
Mjalli et al. [30]	*		*															*	
Rene and Saidutta [31]	*		*		*							*		*				*	
Doğan et al. [32]	*	*					*	*						*	*	*			*
Lee et al. (2011) [33]	*	*	*		*								*					*	
Verma and Singh [34]			*	*			*	*	*								*		
Abyaneh [35]			*	*				*	*										
Li and Song [36]					*	*	*	*	*							*			*
Vijayan and Mohan [37]	*		*															*	
Ebrahimi et al. [38]		*	*		*	*	*		*				*					*	

*TS : Total suspended,
*Od: Odor
*Col: Color
*MLSS: Mixed liquor suspended solids
*O&G: Oil and grease
*BOD_i: Inlet BOD value
*Chl-a: Chlorophyll-a

3.2. Multivariate adaptive regression splines

The MARS is a form of nonparametric regression analysis, which was developed by Friedman in 1991. Non-parametric regression methods are used in most of the applied fields to represent events that have no linearity among variables. The main advantage of this model is that it can explain the complex and nonlinear relationships between the prediction variables and the dependent variables [40].

In this method, no assumption can be made between the variables. Base functions and coefficients related to these base functions are used in this method, which explains each region with a regression equation, separating the arguments by zeros. It is also predicted that the purpose of the change is both the explanatory variables and the contributions of the base functions that are shaped by the interactions between them. Fig. 1 shows a schematic view of MARS.

In Fig. 1, GCV, CM, and MSE represents generalized cross validation, penalty factor, and mean squared error, respectively.

3.3. Artificial bee colony algorithm

The ABC algorithm was developed by Dervis Karaboga in 2005. The algorithm was inspired by the food search behaviors of bees while looking for food. Behavioral food research, information sharing and memorization among individuals have been areas of research that have attracted considerable attention in recent years.

The most basic characterizations of the colony lifestyle of honey bees are the bees exist for the survival of the colony. When bees do work the must follow certain rules, thus a system that can survive successfully emerges. The prominence of this system is that the communication between each of the bees, who work independently, can be achieved successfully [42,43].

The basic flow of the algorithm is as follows;

- Initialize.
- Repeat.
- Employed bees are sent to explore food sources.
- Employed bees share their knowledge about the sources with other bees.
- Onlooker bees are sent to the food sources in the neighborhood which are selected according to the information shared by the employed bees.
- Memorize the best source ever.
- Send the scouts to the search field to find new food sources until the requirements are met [44].

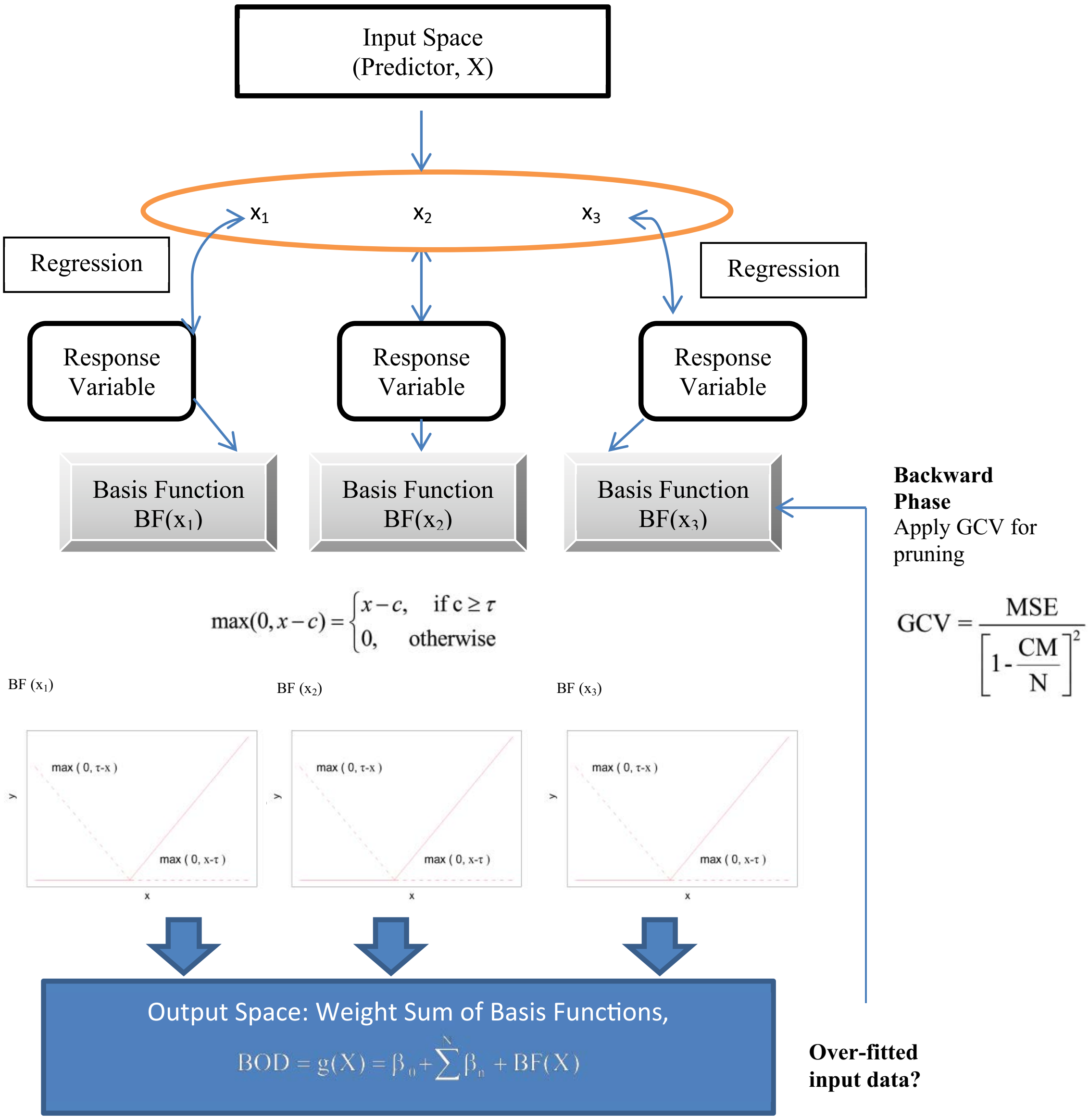
Input Space
(Predictor, X)
x1
x2
x3
Regression
Regression
Response Variable
Response Variable
Response Variable
Basis Function BF(x1)
Basis Function BF(x2)
Basis Function BF(x3)
Backward Phase
Apply GCV for pruning
$\max(0, x-c) = \begin{cases} x-c, & \text{if } c \geq \tau \\ 0, & \text{otherwise} \end{cases}$
$\mathrm{GCV} = \frac{\mathrm{MSE}}{\left[1 - \frac{\mathrm{CM}}{\mathrm{N}}\right]^2}$
BF (x1)
BF (x2)
BF (x3)
max (0, τ-x)
max (0, x-τ)
y
x
Output Space: Weight Sum of Basis Functions,
$\mathrm{BOD} = g(X) = \beta_0 + \sum^{N} \beta_n + \mathrm{BF}(X)$
Over-fitted input data?

Fig. 1. Structures of MARS model [41].

3.4. Teaching-learning based optimization algorithm

The TLBO algorithm is a new method, such as the intuitively known genetic algorithm, the ant colony algorithm and ABC. The TLBO is a social- optimization algorithm based on the interactions between students application of mechanical problems, was also used and teachers in a class [44]. The algorithm designed by Rao in 2011 for the in different engineering areas later years and compared with other nonlinear optimization techniques. At the same time, the algorithm has been used in data clustering and in the design of flat steel frames [45,46]. In the algorithm, which is based on teacher–student interaction in a class, the learning ability of the students is closely related to the capacity of the teacher. Successful students are selected at every stage of the algorithm and the best students are identified. The algorithm has three basic parameters: the number of students in the class, the number of classes, and the number of iterations. The algorithm consists of two phases; teacher phase and student phase.

The teacher's teaching process is done to improve the average knowledge of the students. The teaching process can be formulated as follows:

$$X_{\text{new}} = X_{\text{old}} + r\left(X_{\text{teacher}} - T_F \times X_{\text{mean}}\right) \tag{4}$$

Where X_{old} and X_{new} indicate the student status before and after the teaching process. X_{teacher} and X_{mean} are teacher status and class status respectively. r is a random coefficient ranging from 0 to 1. T_F is a random factor with a value ranging from 1 to 2, indicating a student's learning rate and is calculated by the following equation:

$$T_F = \text{round}\left[1 + \text{rand}\left(0,1\right)\right] \tag{5}$$

At the student status, each student randomly shares information with a student and teaches a student with a higher knowledge. This process is modeled as follows.

$$\begin{aligned} X_{j\text{new}} &= X_{j\text{old}} + r\left(X_i - X_j\right) \quad \text{if } f_i > f_j \\ X_{i\text{new}} &= X_{i\text{old}} + r\left(X_j - X_i\right) \quad \text{if } f_i < f_j \end{aligned} \tag{6}$$

Where, i and j are students' indices, r is a random coefficient ranging from 0 to 1, and f_i is the ith student's level of knowledge. The smaller objective function in the minimization problem expresses the higher knowledge of the student [44,47].

4. Study area and available data

The Antalya Hurma WWTP, from which data were obtained for this study, is a WWTP serving Antalya. It has a capacity to serve 1.4 million people. It is located at the 16th kilometer of Antalya–Kemer road and its construction was completed in 1999. The location of the plant shown in Fig. 2.

The data obtained for this study contained not only sufficient sample numbers but also fully represented possible conditions. In environmental processes, the source data set should cover data measured for at least 1 year, because temperature and the amount of precipitation can vary, by season etc. and affect the process. The data used in this study are the daily measurement results from the Antalya Hurma WWTP inlet pool. The data were recorded daily for 1 year between 2015 and 2016. Due to public holidays, weekends and technical maintenance days, measurements could not be taken everyday.

Abnormalities in the data were detected and corrected prior to the analysis phase. Compared to the other data, values outside the data set range were removed from the data set. These values are called outliers. Outliers can lead to the deviation of normal distribution and change in the analysis results. After the data were eliminated by these criteria, 232 data were used in the estimation of the BOD.

The statistical analysis results of the data set are given in Table 2. In the table, the mean values (x_{mean}), standard deviation (S_x), variance (C_v), skewness coefficient (C_{sx}), maximum value (x_{max}), minimum value (x_{min}) and maximum value to average ratio ($x_{\text{max}}/x_{\text{min}}$) are shown.

The statistical change interval was carried out for each parameter and examination was made to determine whether there were any measurement errors and values outside the logical boundary. The parameters with missing values were not included in the data set. In the applied model, the data were put in chronological order and divided into two. The first 80% of the data set was used in the training set and the remaining 20% was used in the test phase.

5. Discussion and results

The data used in all the models were divided into two groups. 80% of the data was used in the training phase and the remaining 20% was used in the testing phase of the models. After the training phase, it was determined that the model performed as well as using the remaining test data. In this study, 232 of the daily operating data of the Antalya Hurma WWTP were used, 186 of the data were used in the training and the remaining 46 were used for testing the model.

The determination of the input parameters is one of the most important factors in modeling. Table 3 shows the correlation coefficients between the input data of the inlet pool and the relation between the input data and the inlet pool. Before the modeling, these tables provide background

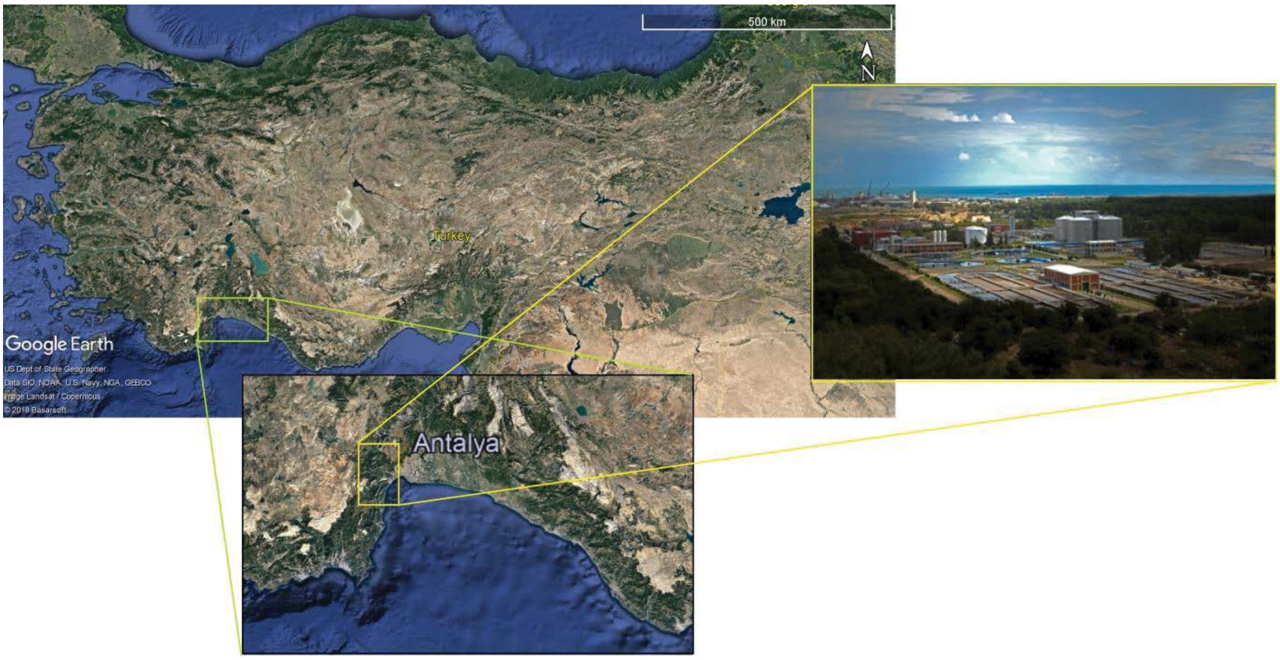
500 km
N
Turkey
Google Earth
US Dept of State Geographer
Data SIO, NOAA, U.S. Navy, NGA, GEBCO
Image Landsat / Copernicus
© 2018 Basarsoft
Antalya

Fig. 2. The location of Hurma WWTP.

Table 2
Basic statistics of Hurma WWTP analysis data

Name of Data	x_{mean}	S_x	C_{sx}	C_v (S_x/x_{mean})	x_{max}	x_{min}	$x_{\text{max}}/x_{\text{mean}}$
Q ($m^3 s^{-1}$)	164.24	16.8601	1.9360	0.10	259.38	115.73	1.5793
pH	7.86	0.2137	–0.3396	0.03	8.81	7.09	1.1202
t (°C)	19.65	4.6464	0.0209	0.24	30.60	10.10	1.5573
COD (mg L^{-1})	680.56	220.4367	0.4274	0.32	1,340.00	226.00	1.9689
BOD (mg L^{-1})	353.53	97.8359	–0.3104	0.28	500.00	120.00	1.4142
SS (mg L^{-1})	356.71	186.5927	1.6972	0.52	1,288.00	100.00	3.6107
tN (mg L^{-1})	39.48	8.4396	0.4524	0.21	65.80	17.50	1.6667
tP (mg L^{-1})	5.99	3.3728	4.9516	0.56	27.00	3.24	4.5141
EC (μs cm^{-1})	1,885.99	125.5650	–1.2239	0.07	2,080.00	1,392.00	1.1028

Table 3
Correlation coefficients of Hurma WWTP's analysis data

	r (Correlation coefficient)								
	BOD	Q	pH	*t*	COD	SS	tN	tP	EC
BOD	1.000								
Q	0.070	1.000							
pH	0.092	0.050	1.000						
T	0.097	0.190	0.070	1.000					
COD	0.569	0.048	0.122	0.025	1.000				
SS	0.472	0.094	0.060	0.008	0.525	1.000			
tN	0.124	0.292	0.083	0.073	0.158	0.084	1.000		
tP	0.135	0.054	0.074	0.131	0.201	0.091	0.237	1.000	
EC	0.098	0.239	0.153	0.099	0.163	0.026	0.191	0.221	1.000

information about which parameters change. These coefficients give understandable information about whether the parameters used before the modeling are logically related to each other. When the table is examined, it is observed that the COD and SS values affect the BOD parameter more than the other parameters. The model was created to observe the effect of the BOD on all parameters seen in the table. In this model, the input data were selected as the *Q*, pH, *t*, COD, SS, tN, tP, and EC.

Optimization of the coefficients is difficult because the magnitudes of the independent variables used in the analysis are at different intervals. In order to facilitate optimization, the values for the ABC and TLBO methods were normalized between 0.1 and 0.9 using Eq. (7). After the analyses were done, the normalized values obtained as the result of the equation were anormalized so that the equations obtained could be compared with the other methods and the raw data could be acquired more easily.

$$\text{Normalised Value} = \left[\frac{\begin{array}{c}\text{(Observed Value} -\\ \text{Minimum Value)}\end{array}}{\begin{array}{c}\text{(Maximum Value} -\\ \text{Minimum Value)}\end{array}}\right]\times\left(0.9-0.1\right)+0.1 \quad (7)$$

In the models used to estimate the BOD and to measure the performance of the model created in this study; the determination coefficient, root mean square error (RMSE), mean absolute error (MAE) values and prediction results were compared. The model that gave the best result was decided by selecting the best value among these results.

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\text{BOD}_{S_{\text{observed}}} - \text{BOD}_{S_{\text{predicted}}}\right)^2} \quad (8)$$

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|\text{BOD}_{S_{\text{observed}}} - \text{BOD}_{S_{\text{predicted}}}\right| \quad (9)$$

In the equations;
$\text{BOD}_{S_{\text{observed}}}$: represents the observed value of the BOD,
$\text{BOD}_{S_{\text{predicted}}}$: represents the modelled value of the BOD,
N: represents the number of data,

Studies on BOD prediction are usually carried out by using ANN method. Regression analysis methods were found to be based on BOD estimation studies. In the literature, the regression analysis model was used as a comparison criterion with the ANN method and there were no studies to compare different regression models with each other. In the present study, BOD was tried to be estimated by using different regression models (the CRA, ABC, TLBO, and MARS).

The ABC algorithm contains three control parameters. These are the number of food sources (SN), the limit value and the maximum number of cycles (MCN). It is very important to specify these parameters since any change in these parameters directly affects the performance of the algorithm [48]. The ABC algorithm parameters were set using different values, colony size (NP = 50, 100, and 200), the number of food sources (SN = (NP/2) = 25, 50, 100) and the limit value was 150, 300, 600 for all functions and 5,000 for the maximum number of cycles. On the other hand, population volume (NP = 50, 100, 200) and the number of maximum iteration (NMI) were chosen as 5,000 for the TLBO algorithm. In the ABC and TLBO algorithms, the weights of the parameters were distributed in the range of [–5, +5]. In the training process, the ABC and TLBO algorithms were used to obtain smaller error values by obtaining more acceptable parameters. The parameters were continuously updated until the convergence criterion was reached.

In the CRA method, the LP, EF, and QF functions were also used in the training of the ABC and TLBO algorithms. The functions obtained from all the methods and, the equation coefficients of the training set are summarized in Table 4. The ABC and TLBO algorithms were used to find more acceptable small error values in the training process. The parameters were continuously updated until the convergence criterion was reached.

In the MARS algorithm, 80% of the analysis data, as in the CRA, was used for training process and 20% for the test set. One of the important advantages of the MARS method is that the relative importance of each input variable can be determined on a scale of 0–100. Thus, it allows the determination of the contribution of different outputs on the model outputs [49]. The significant independent variables according to the MARS method and the relative importance of the changes of these variables are given in Table 5 proportionally. Table 6 shows the BF$_S$ of MARS method.

As seen in Table 6, the MARS method predicted the BOD with six BF$_S$ and the corresponding equation is given as Eq. (10).

$$\begin{aligned}\text{BOD} = 434.28 - 0.431546 \times BF_2 - 4.48034 \times \text{BF}_4 + 269.432 \times \text{BF}_7 \\ - 79.4639 \times \text{BF}_9 - 197.508 \times \text{BF}_{11} + 0.0917189 \times \text{BF}_{13}\end{aligned} \quad (10)$$

When the all models were evaluated according to performance evaluation criteria and reflecting BOD change, the MARS method was found to be more successful than other models. The scatter diagram and time series graphs of the

Table 5
Relatively importance of input parameters on the BOD for MARS model

Variable	Relatively importance (%)
COD	100.00
T	14.08

Table 6
Expressions of BFs for the MARS model

BF	Equation
BF2	max (0, 806-COD)
BF4	max (0, 19.6-COD)
BF7	max (0, t-23)
BF9	max (0, t-24.2)
BF11	max (0, t-22.5)
BF11	max (0, COD-609)

Table 4
Coefficients obtained from the CRA, ABC, and TLBO models

$$y_{\text{LF}} = b_0 + b_1x_1 + b_2x_2 + \ldots + b_mx_m$$

$$y_{\text{PF}} = b_0 \times x_1^{b_1} \times x_2^{b_2} \times \cdots \times x_8^{b_8}$$

$$y_{\text{EF}} = b_0 + \exp\left(b_1 + b_2x_1 + b_3x_2 + \ldots + b_9x_8\right)$$

Function	b_0	b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8	b_9
CRA										
LF	198.962	–0.089	0.397	–1.618	0.329	0.060	–0.788	–0.640	–0.010	
PF	15.806	–0.026	–0.048	–0.073	0.620	0.058	–0.063	–0.006	–0.078	
EF	–20,395.301	9.933	-4.13×10^{-6}	2.89×10^{-6}	-7.86×10^{-7}	1.58×10^{-5}	2.89×10^{-6}	-3.776×10^{-5}	-3.01×10^{-5}	4.48×10^{-7}
ABC										
LF	0.156	–0.105	0.171	–0.152	0.869	0.322	–0.210	0.013	0.118	
PF	1.014	0.048	–0.189	–0.020	0.662	0.066	–0.011	0.010	–0.006	
EF	–0.555	–0.047	–0.105	–0.001	–0.140	0.716	0.206	–0.063	0.007	–0.099
TLBO										
LF	0.227	–0.033	0.001	–0.087	0.964	0.197	–0.100	–0.040	–0.017	
PF	1.043	0.012	–0.010	–0.063	0.696	0.070	–0.043	–0.027	–0.0005	
EF	–4.999	1.655	–0.004	0.001	–0.017	0.169	0.035	–0.017	–0.005	–0.002

CRA, TLBO, ABC, and MARS are given in Fig. 3. As can be seen from Fig. 3, the MARS method closely predicts the BOD compared to other models, while the other methods are insufficient compared to the MARS method.

The results obtained from all methods are given in the Table 7. As it can be seen from Table 7, the CRA, ABC and TLBO were almost close to the accuracy of the training set. The most acceptable results for each criterion are marked

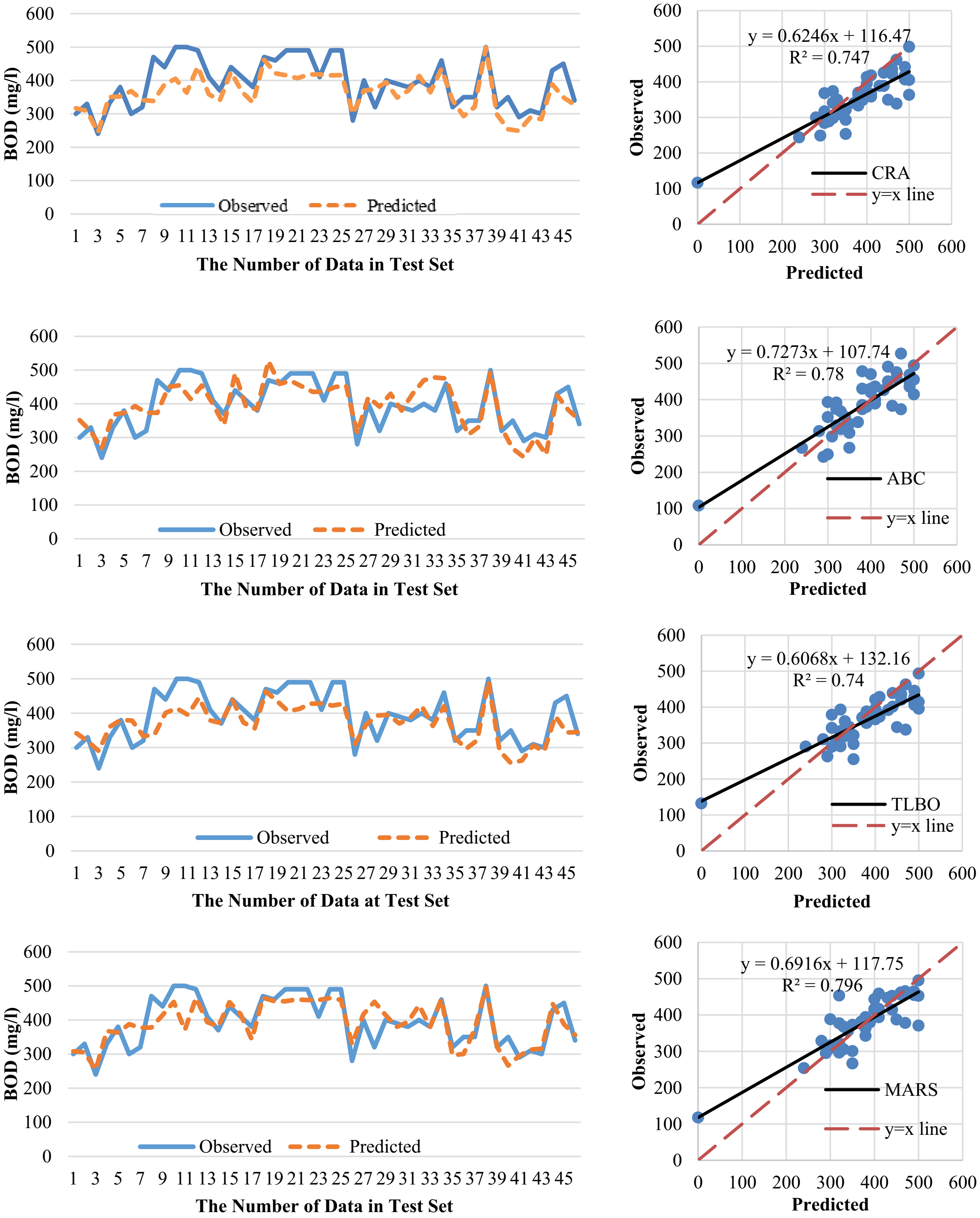

Fig. 3. The scatter diagrams and time series for the testing period using CRA, ABC, TLBO, and MARS methods.

Table 7
The test performances of the all models for the BOD estimation in Hurma WWTP

Models	Training set			Testing set		
	R^2	RMSE (mg L^{-1})	MAE (mg L^{-1})	R^2	RMSE (mg L^{-1})	MAE (mg L^{-1})
CRA						
LF	0.8491	54.0593	45.8947	0.7372	62.7499	59.9972
PF	*0.8606*	*52.1535*	*43.8336*	*0.7469*	*61.4826*	*58.8014*
EF	0.8522	68.8750	59.3735	0.7313	58.4138	54.3939
ABC						
LF	0.8363	56.0825	44.8623	0.7789	50.2081	38.8641
PF	*0.8560*	*52.9563*	*41.9584*	*0.7820*	*46.0179*	*38.2109*
EF	0.8156	59.2798	47.3908	0.7690	57.8442	45.6729
TLBO						
LF	0.8491	54.9096	43.5244	0.7789	72.8433	57.4252
PF	*0.8591*	*52.8924*	*41.3236*	*0.7442*	*72.7387*	*57.6410*
EF	0.8403	55.8052	44.3587	0.7690	71.3260	56.8979
MARS	*0.9039*	*43.6010*	*33.9190*	*0.7966*	*44.1090*	*31.4827*

Table 8.
The R^2 of previous studies

References	R^2
Oliveira-Esquerre et al. [15]	0.36–0.73
Rene and Saidutta [31]	0.55–0.6
Abyaneh [35]	0.53
Ebrahimi et al. [38]	0.82–0.83
Bhatt et al. [50]	0.92

in italic. Among the functions used in the training set for all models, the PF function had the best results. However, the MARS method is the most successful model for the RMSE 44.1090 mg L^{-1}, MAE 31.4827 mg L^{-1} of the test set of Antalya Hurma WWTP. As seen from Table 3, the MARS method gave a better prediction of the peak BOD than the CRA, ABC and TLBO methods.

The R^2 value derived from previous studies is shown in Table 8 for the estimation of BOD in WWTP with regression analysis models

6. Conclusion

In this study, the usability of the regression analysis methods without measuring the daily BOD value was investigated. The results of the CRA, ABC, TLBO, and MARS models were compared with each other. The RMSE and MAE values were used to compare the performances of the methods. These two indices were used to determine errors and similarities according to the observed values. The MARS method applied in the testing set achieved improvements between 4% and 39% compared to the other models. As a result of the comparisons, it is understood that the MARS method can be used for the prediction of BOD.

By means of the obtained model equation, BOD value can be reached with instantly measured parameters, without waiting for BOD test analysis results. Approximate BOD value can be obtained from this model without spending time and consuming material. It was observed that the established model of MARS method had estimates consistent with the measured values which showed the applicability of the model by giving close answers to the instantaneous changes. The MARS method also achieved reliable results in peak value changes. It was observed that it could be used in BOD modeling and analysis through any data set by MARS method.

In this study, only the daily operating data of a WWTP was used. Further studies can be carried out using different data sets in order to improve the results obtained in this study. In addition to the methods used in this study, different meta-heuristic methods can be used and compared with the MARS method. Because the analysis of BOD is difficult, more successful results can be obtained through different models.

Acknowledgments

We would like to express our sincere gratitude to Antalya Water and Sewerage Administration (ASAT) for providing the data used in the study.

References

[1] S. Heddam, H. Lamda, S. Filali, Predicting effluent biochemical oxygen demand in a wastewater treatment plant using generalized regression neural network based approach: a comparative study, Environ. Process., 16 (2016), 153–165.

[2] I. Plazl, G. Pipus, M. Drolka, T. Koloini, Parametric sensitivity and evaluation of a dynamic model for single-stage wastewater treatment plant, Acta Chim. Slov., 46 (1999) 289–300.

[3] K.P. Singh, A. Basant, A. Malik, G. Jain, Artificial neural network modeling of the river water quality a case study, Ecol. Model., 220 (2009) 888–895.

[4] X. Wen, J. Fang, M. Diao, C. Zhang, Artificial neural network modeling of dissolved oxygen in the Heihe River, Northwestern China, Environ. Monit. Assess., 185 (2013), 4361–4371.

[5] A.N.S. Tomić, D.Z. Antanasijević, M.Đ. Ristić, A.A. Perić-Grujić, V.V. Pocajt, Modeling the bod of Danube River in Serbia

using spatial, temporal, and input variables optimized artificial neural network models, Environ. Monit. Assess., 188 (2016).

[6] Q. Chen, A. Mynett, Modelling algal blooms in the Dutch coastal waters by integrated numerical and fuzzy cellular automata approaches, Ecol. Model., 199 (2006) 73–81.

[7] M.R. Kuppusamy, V.V. Giridhar, Factor analysis of water quality characteristics including trace metal speciation in the coastal environmental system of Chennai Ennore, Environ. Int., 32 (2006) 174–179.

[8] K.-W. Chau, N. Muttil, Data mining and multivariate statistical analysis for ecological system in coastal waters, J. Hydroinf., 9 (2007) 305–317.

[9] M.L. Wu, Y.S. Wang, Using Chemometeries to Evaluate Anthropogenic Effects in Daya Bay, China, Estuar, Coast. Shelf. Sci., 72 (2007) 732–742.

[10] A.F.M. Alkarkhi, A. Ahmad, A.M. Easa, Assessment of surface water quality of selected estuaries of Malaysia: multivariate statistical techniques, The Environmentalist, 29 (2009) 255–262.

[11] V. Kumar, A. Sharma, A. Chawla, R. Bhardwaj, K.T. Ashwani, Water quality assessment of river Beas, India, using multivariate and remote sensing techniques, Environ. Monit. Assess., 188 (2016) 137.

[12] B.K. McCabe, I. Hamawand, C. Baillie, Investigating wastewater modelling as a tool to predict anaerobic decomposition and biogas yield of abattoir effluent, J. Environ. Chem. Eng., 1 (2013) 1375- 1379.

[13] M.W. Lee, S.H. Hong, H. Choi, J.-H. Kim, D.S. Lee, J.M. Park, Real–time remote monitoring of small-scaled biological wastewater treatment plants by a multivariate statistical process control and neural network-based software sensors, Process Biochem., 43 (2008) 1107–1113.

[14] J. Tomperi, E. Koivuranta, A. Kuokkanen, K. Leiviskä, Modelling effluent quality based on a real-time optical monitoring of the wastewater treatment process, Environ. Technol., 38 (2017) 1-13,

[15] K.P. Oliveira-Esquerre, M. Mori, R.E. Bruns, Simulation of an industrial wastewater treatment plant using artificial neural networks and principal components analysis, Braz. J. Chem. Eng., 19 (2002), 365-370.

[16] S. Acikalin, R. Ileri, R. Keles, Estimation of Outflow Water Parameters and Yield Values of Adapazari Urban Wastewater Treatment Plant by Artificial Neural Networks, Üniversite Öğrencileri 2. Çevre Sorulari Kongresi, Istanbul, (In Turkish), (2007) 100–107.

[17] D. Guclu, Modeling of Full Scale Urban Wastewater Treatment Plants by Using Computer Program and Investigation of Treatment Performances, Phd Thesis, Selcuk University, Institute of Science and Technology, Konya (In Turkish), 2007.

[18] E. Dogan, R. Koklu, B. Sengorur, Modeling biological oxygen demand of the Melen River in Turkey using an artificial neural network technique, J. Environ. Manage., 90 (2009) 1229–1235.

[19] O.E. Denizci, Dynamic Simulation of Activated Sludge Systems: Investigation of Tuzla and Pasaköy Domestic Wastewater Treatment Plants in Istanbul, Master's Thesis, Yildiz Technic University, Institute of Science and Technology, İstanbul (In Turkish) 2009.

[20] Y-S.T. Hong, M.R. Rosen, R. Bhamidimarri, Analysis of a municipal wastewater treatment plant using a neural network-based pattern analysis, Water Res., 37 (2003) 1608–1618.

[21] O.T. Baki, E. Aras, Estimation of BOD in wastewater treatment plant by using different ANN algorithms, Membr. Water Treat., 9 (2018) 455-462.

[22] Y. Ma, M. Huang, J. Wan, K. Hu, Y. Wang, H. Zhang, Hybrid artificial neural network genetic algorithm technique for modeling chemical oxygen demand removal in anoxic/oxic process, J. Environ. Sci. Health A Tox. Hazard Subst. Environ. Eng., 46(2011) 574–580.

[23] H. Guo, K. Jeong, J. Lim, J. Jo, Y.M. Kim, J-P. Park, J.H. Kim, K.H. Cho, Prediction of effluent concentration in a wastewater treatment plant using machine learning models, J. Environ. Sci., 32 (2015) 90–101.

[24] Y.C. Huang, X.Z. Wang, Application of fuzzy causal networks to waste water treatment plants, Chem. Eng. Sci., 54 (1999) 2731-2738.

[25] G. Civelekoglu, Modeling of Treatment Processes with Artificial Intelligence and Multiple Statistical Methods, Ph.D Thesis, Suleyman Demirel University, Institute of Science and Technology, Isparta (In Turkish), 2006.

[26] G. Civelekoglu, N.O. Yigit, E. Diamadopoulos, M. Kitis, Modelling of COD removal in a biological wastewater treatment plant using adaptive neuro-fuzzy inference system and artificial neural network, Water Sci. Technol., 60 (2009) 1475–1487.

[27] T.-Y. Pai, S.C. Wang, C.F. Chiang, H.C. Su, L.F. Yu, P.J. Sung, C.Y. Lin, H.C. Hu,. Improving Neural Network Prediction of Effluent from Biological Wastewater Treatment Plant of Industrial Park Using Fuzzy Learning Approach, Bioprocess Biosyst. Eng., 32 (2009) 781–790.

[28] M.M. Hamed, M.G. Khalafallah, E.A. Hassanien, Prediction of wastewater treatment plant performance using artificial neural networks, Environ. Model. Softw., 19 (2004) 919–928.

[29] G. Onkal-Engin, I. Demir, S.N. Engin, Determination of the relationship between sewage odour and BOD by neural network, Environ. Model. Softw., 20 (2005) 843–850.

[30] F.S. Mjalli, S. Al-Asheh, H.E. Alfadala, Use of artificial neural network black-box modeling for the prediction of wastewater treatment plants performance, J. Environ. Manage., 83 (2007) 329–338.

[31] E.R. Rene, M.B. Saidutta, Prediction of Water Quality Indices by Regression Analysis and Artificial Neural Networks, Int. J. Environ. Res., 2 (2008) 183–188.

[32] E. Dogan, A. Ates, E.C. Yilmaz, B. Eren, Application of artificial neural networks to estimate wastewater treatment plant inlet biochemical oxygen demand, Environ. Prog. Banner, 27 (2008) 439–446.

[33] J.-W. Lee, C. Suh, Y.-S.T. Hong, H.-S. Shin, Sequential modelling of a full-scale wastewater treatment plant using an artificial neural network, Bioprocess Biosyst. Eng., 34 (2011) 963–973.

[34] A.K. Verma, T.N. Singh, Prediction of water quality from simple field parameters, Environ. Earth Sci., 69 (2013) 821–829.

[35] H.Z. Abyaneh, Evaluation of multivariate linear regression and artificial neural networks in prediction of water quality parameters, J. Environ. Health Sci. Eng., 12 (2014) 40.

[36] X. Li, J. Song, A New ANN-Markov Chain Methodology for Water Quality Prediction, 2015 Int. Joint Conf. on Neural Networks (IJCNN), Killarney, Ireland, 2015.

[37] A. Vijayan, G.S. Mohan, Prediction of effluent treatment plant performance in a diary industry using artificial neural network technique, J. Civil Environ. Eng., (2016) 6.

[38] M. Ebrahimi, E.L. Gerber, T.D. Rockaway, Temporal performance assessment of wastewater treatment plants by using multivariate statistical analysis. J. Environ. Manage., 193 (2017) 234–246.

[39] O.T. Baki, Modeling of Biochemical Oxygen Demand on Wastewater Treatment Plant by using Different Artificial Intelligence Methods: Antalya Hurma Wastewater Treatment Plant Example, Master Thesis, Karadeniz Technical University, Institute of Science and Technology, Trabzon, 2016 (In Turkish).

[40] O. Kisi, K.S. Parmar, Application of Least Square Support Vector Machine and Multivariate Adaptive Regression Spline Models in Long Term Prediction of River Water Pollution, J. Hydrol., 534 (2016) 104-112.

[41] R.C. Deo, O. Kisi, V.P. Singh, Drought forecasting in eastern Australia using multivariate adaptive regression spline, least square support vector machine and M5Tree model, Atmos. Res., 184 (2017) 149–175.

[42] D. Karaboga, An Idea on Honey Bee Swarm for Numerical Optimization, Technical Report-TR06, 2005.

[43] C. Ozkan, O. Kisi, B. Akay, Neural networks with artificial bee colony algorithm for modeling daily reference evapotranspiration, Irrig. Sci., 29 (2011) 431–441.

[44] R.V. Rao, V. Patel, An elitisit teaching-learning-based optimization algorithm for solving complex constrained optimization problems, Int. J. Ind. Eng. Comput., 3 (2012) 535–560.

[45] S.C. Satapathy, A. Naik, Data Clustering Based on Teaching Learning Based Optimization, SEMCCO 2011, Part II, LNCS 7077 (2011) 148–156.

[46] V. Togan, Design of Planar Steel Frames Using Teaching-Learning Based Optimization, Eng. Struct., 35 (2012) 225–232.

[47] E. Uzlu, M.I. Komurcu, M. Kankal, T. Dede, H.T. Ozturk, Prediction of berm geometry using a set of laboratory tests combined with teaching-learning-based optimization and artificial bee colony algorithms, Appl. Ocean Res., 48 (2014) 103–113.

[48] A. Bayram, E. Uzlu, M. Kankal, T. Dede, Modeling stream dissolved oxygen concentration using teaching–learning based optimization algorithm, Environ. Earth Sci., 73 (2015) 6565–6576

[49] V.N. Sharda, R.M. Patel, S.O. Prasher, P.R. Ojasvi, C. Prakash. Modeling runoff from middle Himalayan watersheds employing artificial intelligence techniques, Agric. Water Manage., 83 (2006) 233–242.

[50] A.H. Bhatt, R.V. Karanjekar, S. Altouqi, M.L. Sattler, M.D.S. Hossain, V.P. Chen, Estimating landfill leachate BOD and COD based on rainfall, ambient temperature, and waste composition: exploration of a MARS statistical approach, Environ. Technol. Innovation, 8 (2017) 1–16.