Predicting Water Quality Indicators from Conventional and Nonconventional Water Resources in Algeria Country: Adaptive Neuro-Fuzzy Inference Systems Versus Artificial Neural Networks

Salim Heddam, Ozgur Kisi, Abderrazek Sebbar, Larbi Houichi, and Lakhdar Djemili

Contents

S. Heddam (✉)
Faculty of Science, Agronomy Department, Hydraulics Division, Laboratory of Research in Biodiversity Interaction Ecosystem and Biotechnology, University 20 Août 1955, Skikda, Algeria
e-mail: heddamsalim@yahoo.fr

O. Kisi
School of Technology, Ilia State University, Tbilisi, Georgia
e-mail: ozgur.kisi@iliauni.edu.ge

A. Sebbar
Soil and Hydraulics Laboratory, Faculty of Engineering Sciences, Hydraulics Department, University Badji-Mokhtar Annaba, Annaba, Algeria
e-mail: rsebbar@yahoo.fr

L. Houichi
Department of Hydraulic, University of Batna 2, Batna, Algeria
e-mail: houichilarbi@yahoo.fr

L. Djemili
Research Laboratory of Natural Resources and Adjusting, Faculty of Engineering Sciences, Hydraulics Department, University Badji-Mokhtar Annaba, Annaba, Algeria
e-mail: lakhdardjemili@gmail.com

Abdelazim M. Negm, Abdelkader Bouderbala, Haroun Chenchouni, and Damià Barceló (eds.), *Water Resources in Algeria - Part II: Water Quality, Treatment, Protection and Development*,
Hdb Env Chem (2020) 98: 13–34, DOI 10.1007/698_2019_399,

Abstract Monitoring water quality is of great importance and mainly adopted for water pollution control of conventional and nonconventional water resources. Generally, water quality is evaluated using several indicators, including chemical oxygen demand (COD), biochemical oxygen demand (BOD), and dissolved oxygen concentration (DO). In the present investigation, two artificial intelligence techniques, namely, adaptive neuro-fuzzy inference system (ANFIS) and artificial neural networks (ANN), were applied for predicting two water quality indicators: (1) chemical oxygen demand (COD) at Sidi Marouane Wastewater Treatment Plant (WWTP), east of Algeria, and (2) dissolved oxygen concentration (DO) at the drinking water treatment plant of Boudouaou, Algeria. The models were developed and compared based on several water quality variables as inputs. Three ANFIS models, namely, (1) ANFIS with fuzzy c-mean clustering (FCM) algorithm called ANFIS_FC, (2) ANFIS with grid partition (GP) method called ANFIS_GP, and (3) ANFIS with subtractive clustering (SC) called ANFIS_SC, were developed. The ANFIS models were compared to standard multilayer perceptron neural network (MLPNN) and multiple linear regression model (MLR). Results obtained demonstrated that (1) for predicting COD, ANFIS_SC is the best model, and the coefficient of correlation (R), Wilmot's index (d), root-mean-square error (RMSE), and mean absolute error (MAE) were calculated as 0.805, 0.880, 6.742, and 4.944 mg/L for the validation dataset. The worst results were obtained using the MLR model with R, d, RMSE, and MAE equal to 0.750, 0.840, 0.7658, and 5.916 mg/L for the validation subset, and (2) for predicting DO concentration, the best results were obtained using ANFIS_SC with R, d, RMSE, and MAE equal to 0.856, 0.922, 1.528, and 1.123 mg/L for the validation subset, respectively.

Keywords ANFIS, Chemical oxygen demand, COD, Dissolved oxygen, DO, MLPNN, Modeling, Water quality indicators

1 Introduction

Over the year, the control of water pollution is becoming of great importance, and several regulations have been put in place [1]. Monitoring wastewater treatment plant (WWTP) using online sensors has become an essential and crucial task to handle rapid and seasonal variations that occur during all the months of years [2]. Consequently, real-time supervision of the process of WWTP is nowadays a challenge [3]. To deal with these challenges, WWTP must be highly efficient [4]. Evaluation of the WWTP performances is mainly based on the measure of water quality indicators (WQI), which are generally hard to measure regularly [5]. In the last few years, soft computing models have been largely employed for modeling and forecasting water quality indicators (WQI) in several water ecosystems. Chemical oxygen demand (COD), biochemical oxygen demand (BOD), and dissolved oxygen concentrations (DO) were the most important WQI that have received great importance, and modeling chemical oxygen demand in wastewater treatment plant (WWTP) is broadly discussed in the literature [6–13].

Ay and Kisi [6] compared several machine learning approaches in modeling daily COD measured at the upstream of a WWTP in Turkey, using discharge (Q) and three water quality variables as inputs: (1) suspended solid (SS), (2) temperature (T), and (3) pH. The proposed models included (1) multiple linear regression (MLR), (2) multilayer perceptron neural network (MLPNN), (3) radial basis function neural network (RBFNN), (4) generalized regression neural networks model (GRNN), (5) adaptive neuro-fuzzy inference system techniques with grid partitioning, (6) adaptive neuro-fuzzy inference system techniques with subtractive clustering, and (7) a new model called MLPNN embedded k-means clustering (K_MLP). The authors demonstrated that the K_MLPN using three input variables (SS, T, and pH) provided the best accuracy with a coefficient of determination (R^2) equal to 0.88 in the validation phase. Kisi and Parmar [7] applied three data-driven models, namely, (1) least square support vector machine (LSSVM), (2) multivariate adaptive regression splines (MARS), and (3) M5 model tree (M5Tree) for modeling monthly COD in India. According to the results obtained, the authors demonstrated that the MARS and LSSVM performed better than the M5Tree. Nadiri et al. [8] proposed a new model called supervised committee fuzzy logic (SCFL) for predicting COD in WWTP in Iran. The proposed model is a combination of the artificial neural network (ANN) paradigm and several individual fuzzy logic (FL) models: Takagi-Sugeno, Mamdani, and Larsen. According to the results obtained, the authors demonstrated that a linear combination of several FL models outperforms the individual FL model.

Moral et al. [9] applied the standard MLPNN for predicting effluent COD at the Iskenderun WWTP, Turkey. Yilmaz et al. [10] compared three data-driven models, GRNN, RBFNN, and MLPNN for predicting the effluent COD using influent COD, hydraulic retention time (HRT), and influent cyanide concentration (CN). MLPNN was found to be the best model compared to the two others, with an R^2 equal to 0.876 in the validation phase. Pai et al. [11] compared ANFIS and MLPNN for predicting effluent COD at WWTP in Taiwan. The authors selected four water quality variables as inputs, the influent SS, TE, and pH, in addition to the influent COD. According to the results obtained, the ANFIS model was found to be slightly better than the MLPNN. Perendeci et al. [12] proposed the use of the ANFIS model for predicting the effluent COD using the COD measured at previous 10 days and reported very encouraging results with an R^2 equal to 0.84. Singh et al. [13] compared several linear and nonlinear models for predicting weekly effluent COD at WWTP using four water quality variables as inputs measured at the influent of the WWTP. The proposed models were (1) partial least squares regression (PLSR),

(2) multivariate polynomial regression (MPR), and (3) MLPNN models, and it is observed that the MLPNN model has better performance than the other models with an R^2 equal to 0.84 in the test phase. Different modeling approaches can be found in the literature [14, 15]. Contrary to the COD, which has received great attention worldwide, modeling DO in DWTP is rarely reported in the literature. Hence, in the present study, we reported an application of the ANFIS, MLPNN, and MLR models for modeling DO in drinking water treatment plant (DWTP) and COD in WWTP.

2 Wastewater and Drinking Water Datasets

In the present study, effluent wastewater and drinking water data were obtained from two different stations (Fig. 1): (1) Sidi Marouane Wastewater Treatment Plant (WWTP) located at Sidi Marouane town, at about 12 km northeast of Mila Province, east Algeria.

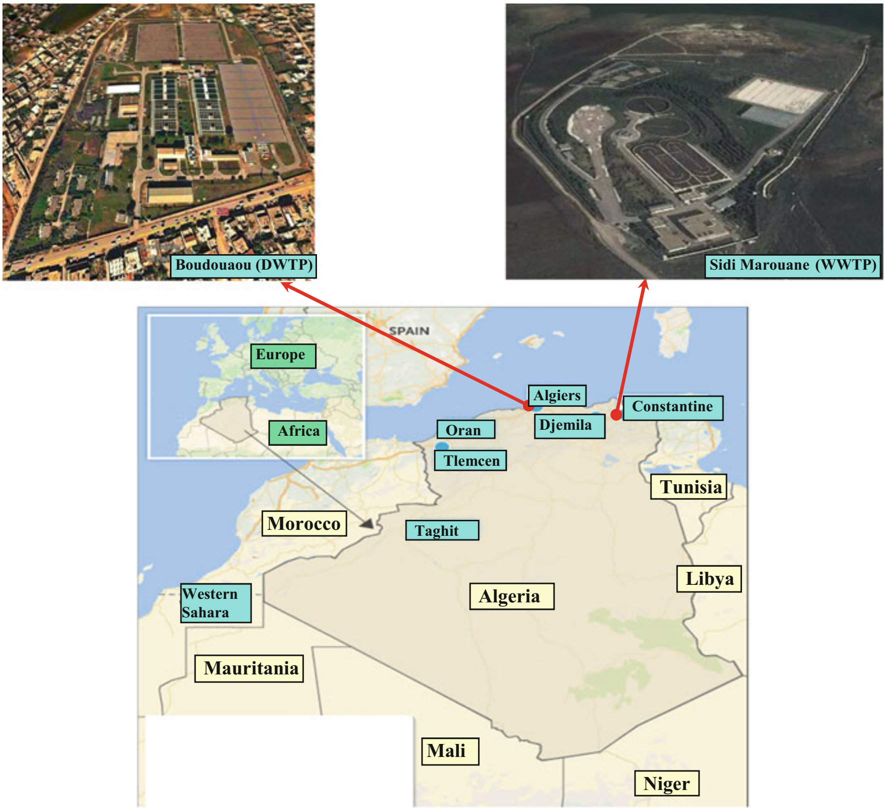

Fig. 1 Location of Sidi Marouane Wastewater Treatment Plant (WWTP) and Boudouaou Drinking Water Treatment Plant (DWTP) in Algeria country

The WWTP is located near the Beni Harroun Dam Reservoir [16], and (2) Boudouaou Drinking Water Treatment Plant (DWTP) is located at Boudouaou province and is the principal DWTP in Algeria [17]. The DWTP has a capacity of 540,000 m^3 of water per day and provides drinking water to more than four million inhabitants [18]. The treatment consists essentially of preliminary disinfection, coagulation-flocculation, settling, filtration, and final disinfection [18, 19] (Fig. 2). Regarding the WWTP, the treatment scheme is based on the conventional activated sludge plant and consists essentially of coarse and fine screens, grit and grease removal, primary sedimentation tanks, activated sludge aeration tanks, secondary sedimentation tanks, and final clarification and chlorination facilities [16] (Fig. 3). As explained above, two different datasets were used in the present study for modeling COD and DO, respectively. The first dataset was collected from the Sidi Marouane WWTP. It is composed of 364 patterns and includes four input variables:

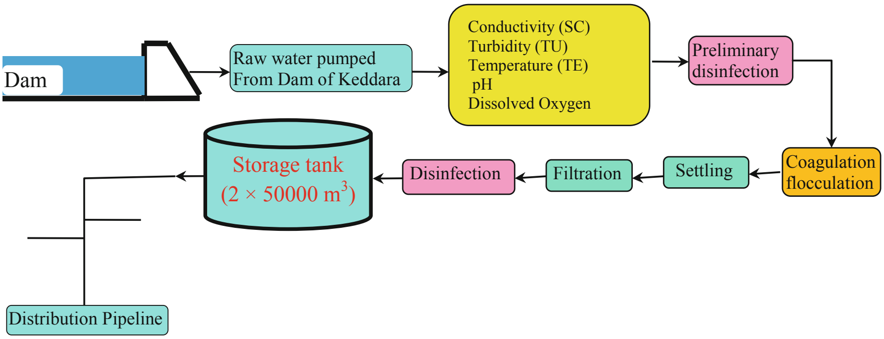

Fig. 2 Schematic diagram of Boudouaou Drinking Water Treatment Plant (scheme adopted from [12])

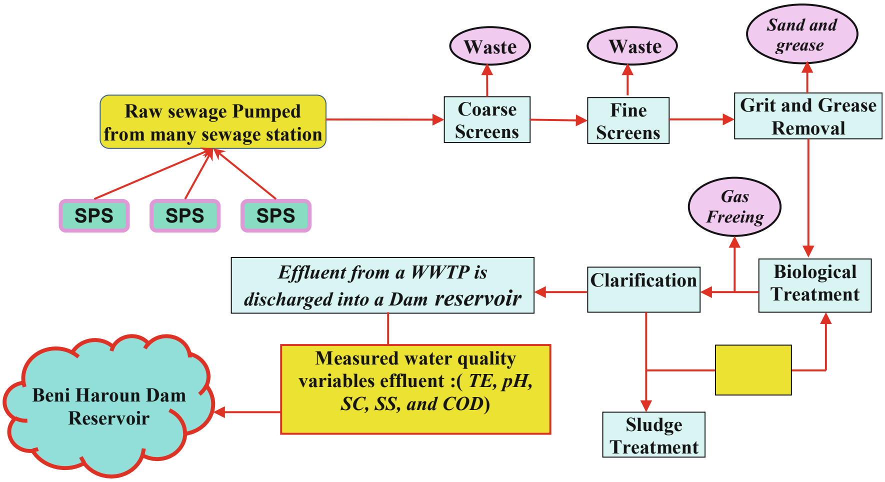
Raw sewage Pumped from many sewage station
SPS
SPS
SPS
Waste
Coarse Screens
Waste
Fine Screens
Sand and grease
Grit and Grease Removal
Gas Freeing
Biological Treatment
Clarification
Effluent from a WWTP is discharged into a Dam reservoir
Measured water quality variables effluent :(TE, pH, SC, SS, and COD)
Beni Haroun Dam Reservoir
Sludge Treatment

Fig. 3 Schematic diagram of Sidi Marouane Wastewater Treatment Plant (WWTP)

(1) effluent water temperature (TE), (2) effluent suspended solids (SS), (3) effluent specific conductance (SC), and (4) effluent pH. Consequently, the dependent variable (output) is the effluent chemical oxygen demand (COD), and the independent variables are TE, SS, SC, and pH. Of this 364 patterns, 255 (70%) were randomly selected as the model training subset, and 109 patterns (30%) were used as validation subset. The second dataset was collected from Boudouaou DWTP. It is composed of 902 patterns and includes four input variables: (1) raw water TE, raw water turbidity (TU), raw water SC, and raw water pH. Consequently, the dependent variable (output) is the dissolved oxygen (DO), and the independent variables are TE, TU, SC, and pH. Of these 902 patterns, 632 (70%) were randomly selected as the model training subset, and 270 patterns (30%) were used as validation subset.

In Table 1, we report the mean, maximum, minimum, standard deviation, coefficient of variation values, and the coefficient of correlation with COD and DO, i.e., X_{mean}, $X_{\max}$, $X_{\min}$, S_x, C_v, and R, respectively. Among the input variables for DO, TU has the highest variation, while for the COD, SS data have higher variation than the others (see the variation coefficients, C_v in the table). On the other hand, the training ranges of DO (5.084–11.16 mg/L) and COD (23.217–49.8 mg/L) do not cover the validation ranges (5.266–13.2 mg/L for DO and 22.289–55 mg/L for COD). This may cause some extrapolation difficulties for the applied models. In the present study, COD and DO and all the input variables were normalized using the Z-score method [20, 21]:

$$Z_n = \frac{x_n - x_m}{\sigma_x} \tag{1}$$

where Z_n is the normalized value of the observation n, x_{n} is the measured value of the observation n, and x_m and σ_x are the mean value and standard deviation of the variable x. This normalization was applied because it considerably improves the performances of the AI models [22, 23].

3 Methodology

In the present study, three kinds of models were developed and compared: multilayer perceptron neural network (MLPNN), adaptive neuro-fuzzy inference system (ANFIS), and multiple linear regression (MLR). The flow chart for training and validation of the MLPNN, ANFIS, and MLR is shown in Fig. 4.

3.1 Multilayer Perceptron Neural Network (MLPNN)

Artificial neural network (ANN) is a mathematical model that learns from examples similar to human brain, and the structure of the artificial neuron was inspired from the function of the biological neuron. ANN is structured in several layers,

Table 1 Daily statistical parameters of the dataset

Variables	Subset	Unit	X_{mean}	X_{max}	X_{min}	S_x	C_v	R
Boudouaou Drinking Water Treatment Plant (DWTP)								
TE	Training	°C	16.611	25.900	10.600	3.433	0.207	−0.255
	Validation		16.665	26.200	10.200	3.518	0.211	−0.297
	All data		16.627	26.200	10.200	3.457	0.208	−0.268
pH	Training	/	7.765	8.600	7.200	0.248	0.032	0.381
	Validation		7.764	8.510	7.290	0.239	0.031	0.350
	All data		7.765	8.600	7.200	0.245	0.032	0.372
SC	Training	μS/cm	1,043.954	1,610.000	668.000	147.247	0.141	0.317
	Validation		1,053.570	1,555.000	699.000	144.105	0.137	0.366
	All data		1,046.833	1,610.000	668.000	146.300	0.140	0.332
TU	Training	NTU	7.215	32.400	0.440	4.424	0.613	0.370
	Validation		7.160	27.000	0.500	4.306	0.601	0.389
	All data		7.199	32.400	0.440	4.387	0.609	0.375
DO	Training	mg/L	5.084	11.160	0.148	2.973	0.585	1.000
	Validation		5.266	13.200	0.143	2.939	0.558	1.000
	All data		5.139	13.200	0.143	2.962	0.576	1.000
Sidi Marouane Wastewater Treatment Plant (WWTP)								
SS	Training	Mg/L	6.736	23.000	0.200	5.052	0.750	0.498
	Validation		7.511	29.000	0.400	5.927	0.789	0.493
	All data		7.261	39.500	0.200	5.628	0.775	0.347
TE	Training	°C	19.845	28.000	12.500	4.546	0.229	−0.130
	Validation		19.871	27.300	7.300	4.680	0.236	−0.153
	All data		20.212	28.000	7.300	4.492	0.222	−0.113
pH	Training	/	7.673	8.500	7.000	0.406	0.053	0.528
	Validation		7.684	8.520	7.000	0.387	0.050	0.630
	All data		7.619	8.520	7.000	0.343	0.045	0.345

(continued)

Table 1 (continued)

Variables	Subset	Unit	X_{mean}	X_{max}	X_{min}	S_x	C_v	R
SC	Training	µS/cm	1,553.059	1,815.000	1,210.000	107.548	0.069	−0.365
	Validation		1,548.587	1,791.000	1,210.000	120.216	0.078	−0.547
	All data		1,602.226	1,970.000	1,210.000	116.764	0.073	−0.295
COD	Training	mg/L	23.217	49.800	5.900	11.564	0.498	1.000
	Validation		22.289	55.000	4.500	11.341	0.509	1.000
	All data		20.566	59.000	3.500	10.702	0.520	1.000

X_{mean} mean, X_{max} maximum, X_{min} minimum, S_x standard deviation, C_v coefficient of variation, *R* coefficient of correlation with DO/COD, *TE* water temperature, *SS* suspended solids, *SC* specific conductance, *TU* turbidity in Nephelometric Turbidity Unit (NTU), *COD* chemical oxygen demand, *DO* dissolved oxygen

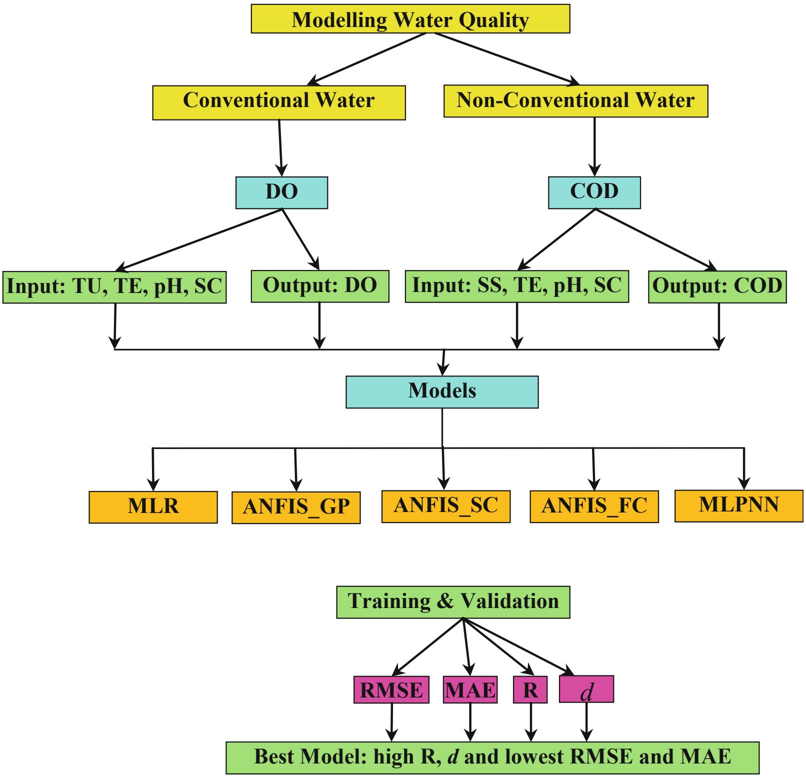

Fig. 4 Flow chart for the proposed MLPNN, ANFIS, and MLR models

and generally, there is an input, an output, and several hidden layers, and the information flows from the input to the output layer in which a series of processing operations is carried out, using a multiplication, summation, and transformation using a nonlinear activation (transfer) function. The available information represented by a matrix of input variables designed as x_i that represent the independent variable is stored in the input layer, while the response variable (y) is fixed into the output layer [24]. The hidden layers are the most important part of the ANN model, and its success and its ability to solve a highly complex problem are attributed to the role accomplished by the neurons arranged in the hidden layer that are characterized by the presence of a nonlinear function, generally the sigmoid function. The connection between different neurons, in different layers, is achieved using the weights and bias, sometimes called connection strengths. Similar to any other statistical models, weights and bias represent the parameters of the ANN model that must be optimized and adopted using a learning algorithm, generally the back-propagation, during a training process. Development of ANN models is mainly governed by the presence of dataset. The most well-known ANN model is certainly the MLPNN [25] that is frequently used for nonlinear mapping of input variables to an output variable based on function approximation. The goal of the training process is the minimization of an objective function. Generally the mean square error (MSE) is estimated between the measured value and the calculated value via the model [24]. In the present study, we used a MLPNN model having only one hidden layer with sigmoid activation function and a linear activation function also called identity function for the unique neuron in the output layer. MLPNN is a universal approximator [26, 27].

From the input layer to the output layer (Fig. 5), the mathematical formulation of the MLPNN can be split into the following equations:

$$I_j = \sum_{j=1}^{n} x_i w_{ij} + \delta_j \tag{2}$$

where x_i is the input variable, w_{ij} is the weight between the input i and the hidden neuron j, Ij is the net internal activity level of neuron j in the hidden layer, and δ_j is the bias of the hidden neuron j.

$$E_j = f_1(I_j) \tag{3}$$

E_j is the output from neuron j in the hidden layer, and f_1 is the activation sigmoid function, represented by Eq. (4).

$$f_1(x) = \frac{1}{1 + e^{-x}} \tag{4}$$

$$O = \sum_{j=1}^{n} E_j w_{jk} + \delta_0 \tag{5}$$

w_{jk} is the weight of connection of neuron j in the hidden layer to unique neuron k in the output layer; O is the input of the output neuron k, and δ_0 is the bias of the output neuron k. Finally the output of the neuron k in the output layer is calculated using a linear activation function f_0:

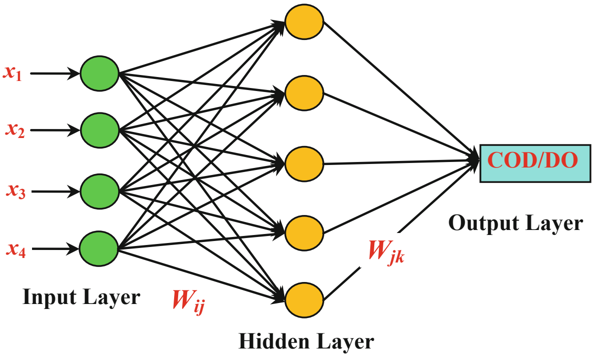

Fig. 5 Multilayer perceptron neural network (MLPNN) structure for modeling COD and DO concentrations

$$Y = f_2(O) \tag{6}$$

3.2 Adaptive Neuro-Fuzzy Inference System (ANFIS)

Adaptive neuro-fuzzy inference system designated as ANFIS is a data-driven model and belongs to the category of hybrid model, which combines two paradigms: the ANN and the fuzzy logic (FL) [28]. ANFIS is a nonlinear mathematical model that has a great capability of mapping any complex process characterized by a set of independent variables (inputs) and one dependent variable (the output). From the ANN approach, ANFIS is structured in several layers, and the information is circulated from the first to the last layer, while from the FL approach, ANFIS model uses the linguistic information and the concept of rules [28]. Among all the other artificial intelligence (AI) models, ANFIS needs a hybrid learning process to update the linear (consequent) and nonlinear (premises) parameters, composed of (1) back-propagation method for updating the nonlinear parameters found in the membership function and (2) the least squares (LS) for updating the linear parameters found in the IF-THEN rules base [28]. The hybrid algorithm is achieved in two steps: forward for updating the consequent parameters and backward pass for updating the premise parameters [28]. ANFIS general architecture is shown in Fig. 6. From Fig. 6, it is clear that the ANFIS model has five layers: two adaptive layers and three fixed layers. The first layer is used only for presenting the input variables. The fuzzy rule could be expressed as:

$$\text{Rule}\quad 1 = \text{If}\ (x\ \text{is}\ A_1)\quad \text{and}\quad (y\ \text{is}\ B_1)\quad \text{Then}\quad (f_1 = p_1x + q_1y + r_1) \tag{7}$$

$$\text{Rule}\quad 2 = \text{If}\ (x\ \text{is}\ A_2)\quad \text{and}\quad (y\ \text{is}\ B_2)\quad \text{Then}\quad (f_2 = p_2x + q_2y + r_2) \tag{8}$$

where x and y denote the inputs, A_i and B_i indicate the fuzzy sets, f_i are the outputs within the fuzzy region indicated by the fuzzy rule, and p_i, q_i, and r_i show the design parameters that are identified in the training phase.

Layer 1: the fuzzification layer with adaptive node

$$O_i^1 = \mu_{A_i}(x), \quad i = 1, 2, \tag{9}$$

$$O_i^1 = \mu_{B_{i-2}}(y), \quad i = 3, 4 \tag{10}$$

A_i (or B_{i-2}) is the linguistic label and $\mu_{A_i}\ (x)$, $\mu_{B_{i-2}}(y)$ fuzzy membership function.

For a Gaussian membership function, A_i can be computed as:

$$\mu_{A_i}(x) = \exp\left(-0.5 \times \{(x - c_i)/\sigma_i\}^2\right), \tag{11}$$

where σ_i, c_i are the premise parameters.

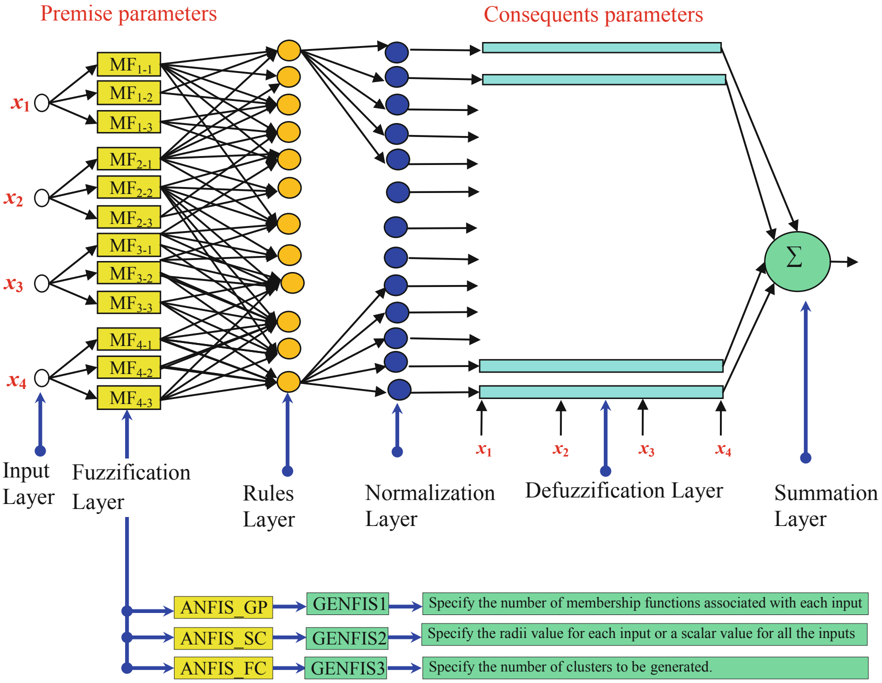

Fig. 6 ANFIS structure

Layer 2: the base rules layer

$$O_i^2 = w_i = \mu_{A_i}\mu_{Bi}, \quad i = 1, 2, \tag{12}$$

w_i is the firing strength of a rule. The node numbers in this layer equal the number of fuzzy rules.

Layer 3: the normalized firing strengths

$$O_i^3 = \overline{w}_i = (w_i/(w_1 + w_2)), \quad i = 1, 2, \tag{13}$$

Outputs of this layer are named as normalized firing strengths.

Layer 4: the defuzzification layer

$$O_i^4 = \overline{w}_i f_i = \overline{w}_i\,(p_i\,x + q_i\,y + r_i), \quad i = 1, 2 \tag{14}$$

where $\overline{w}_i$ the output of Layer 3 and p_i, q_i, and r_i are the consequent parameters.

Layer 5: the output of the ANFIS model

$$O_i^5 = \sum_{i=1} \overline{w}_i f_i = \left(\sum_{i=1} w_i f_i / (w_1 + w_2) \right). \tag{15}$$

ANFIS model can be built in three different forms: (1) ANFIS with grid partition method called ANFIS_GP, (2) ANFIS model with subtractive clustering called ANFIS _SC, and (3) ANFIS model with fuzzy *c*-means clustering *(*FCM) called ANFIS_FC. In the present study, ANFIS was developed using the software Matlab. For ANFIS_GP we used the GENFIS1 function; for ANFIS_SC and ANFIS_FC, we used GENFIS2 and GENFIS3 functions.

3.3 Multiple Linear Regression (MLR)

The multiple linear regression (MLR) is the well-known kind of linear models, and it represents an ideal relationship between a single variable called dependent (Y) and some explanatory variables (X_{i}) called independent variables. The relation between X_{i} and Y is given as:

$$\Psi = A_0 + A_1 \times x_1 + A_2 \times x_2 + A_3 \times x_3 + \ldots A_i \times x_i \tag{16}$$

where Ψ is the calculated or the predicted value of Y, A_0 is the intercept, and A_i are the partial regression coefficients associated with input variables.

3.4 Performance Assessment of the Models

In the present study, we used four performance indices to evaluate and compare the accuracy of the developed models: the coefficient of correlation (R), the Willmott index of agreement (d), the root-mean-squared error (RMSE), and the mean absolute error (MAE).

$$R = \left[\frac{\frac{1}{N} \sum (\mathrm{O}_i - \mathrm{O}_m)(\mathrm{P}_i - \mathrm{P}_m)}{\sqrt{\frac{1}{N} \sum_{i=1}^{n} (\mathrm{O}_i - \mathrm{O}_m)^2} \sqrt{\frac{1}{N} \sum_{i=1}^{n} (\mathrm{P}_i - \mathrm{P}_m)^2}} \right] \tag{17}$$

$$d = 1 - \frac{\sum_{i=1}^{N} (P_i - \mathrm{O}_i)^2}{\sum_{i=1}^{N} (|P_i - \mathrm{O}_m| + |\mathrm{O}_i - \mathrm{O}_m|)^2} \quad (18)$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (O_i - P_i)^2} \quad (19)$$

$$\mathrm{MAE} = \frac{1}{N} \sum_{i=1}^{N} |O_i - P_i| \quad (20)$$

where N is the number of data points, O_i is the measured value, P_i is the corresponding model prediction, and O_m and P_m are the average values of O_i and P_i [29–31].

4 Results

In the present work, we applied three types of models listed earlier: (1) the standard MLR, (2) the MLPNN, and (3) three types of ANFIS models, the ANFIS_GP, the ANFIS_SC, and the ANFIS_FC. The models were compared, and their performances were evaluated for modeling DO and COD concentrations. Several combinations of the water quality variables were selected, and in total five scenarios (Table 2) were compared. The performance of the models used in this study was computed using four performance criteria, including RMSE, MAE, *R*, and also the *d*. For the MLPNN models, the Levenberg-Marquardt (LM) algorithm was employed. For the hidden and output layers, the sigmoid and linear (identity) transfer functions were employed, respectively. The minimum number of the hidden neuron was one, and the maximum was 20, and a total of 100 epochs was adopted. By trial and error, we find that the optimal number of neurons in the hidden layer for both COD and DO was equal to 13 neurons. ANFIS is configured using three different identification

Table 2 The input combinations for different models

Models					Input combinations	
					Modeling DO	Modeling COD
MLR1	MLPNN1	ANFIS_SC1	ANFIS_GP1	ANFIS_FC1	TE, pH, SC, TU	SS,TE, pH, SC
MLR2	MLPNN2	ANFIS_SC2	ANFIS_GP2	ANFIS_FC2	TE, pH, SC	SS, TE, SC
MLR3	MLPNN3	ANFIS_SC3	ANFIS_GP3	ANFIS_FC3	TE, SC,TU	SS, pH, SC
MLR4	MLPNN4	ANFIS_SC4	ANFIS_GP4	ANFIS_FC4	pH, SC	SS, pH
MLR5	MLPNN5	ANFIS_SC5	ANFIS_GP5	ANFIS_FC5	SC,TU	SS, SC

methods: (1) grid partitioning using Genfis1 algorithm for ANFIS_GP, (2) subtractive clustering using Genfis2 algorithm for ANFIS_SC, and (3) fuzzy c-means clustering using Genfis3 algorithm for ANFIS_FC. For ANFIS_GP, the number of membership functions (MFs) for each input variable is two. The Gaussian curve membership function Gaussmf is used for input variables, and output membership function type is a linear type, and it is trained for 150 epochs. Using the grid partition method, the numbers of fuzzy rules exponentially increase with the increase of the number of MFs for each input variable, and it is hard to use this method if the number of input variables is rather than six because of computation time and/or memory limitations. Hence, the number of possible fuzzy rules is calculated as $(\text{MFs})^n$, where n is the number of input variables. For ANFIS_SC, contrary to the ANFIS_GP, the optimum number of MFs and consequently the number of fuzzy layers are determined using the subtractive clustering (SC) algorithm. The number of fuzzy rules generated using the SC algorithm is governed by one parameter: the radius value r_a, determined at the beginning of the training process. Large values of r_a generate fewer clusters and vice versa. Consequently, the number of fuzzy rules is equal to the number of clusters. In the present study, we determined the optimal value of r_a by trial and error, and the best values for DO and COD were 0.26 and 0.85, respectively. Finally, for the ANFIS_FC based on the fuzzy c-means clustering (FCM) algorithm, contrary to the SC algorithm, the number of clusters generated is known and fixed at the beginning of the training, and the number of fuzzy rules is equal to the number of clusters. In the present study, we determined the optimal number of the cluster by trial and error, and the best values for DO and COD were 20 and 3, respectively. Hereafter the results obtained are summarized and discussed.

4.1 *Modeling DO at Boudouaou Drinking Water Treatment Plant*

Table 3 shows the results obtained by the ANFIS, MLPNN, and MLR models applied and compared together. According to Table 3, acceptable accuracy between measured and calculated DO concentration was achieved by all the MLPNN and ANFIS models, while the MLR models perform worse with high (RMSE and MAE) and low (R and d) values. In the training phase as seen in Table 3, the best R and d across all compared models were achieved by the ANFIS_FC1 ($R = 0.939$, $d = 0.968$), followed by ANFIS_SC1 ($R = 0.909$, $d = 0.951$), the MLPNN1 ($R = 0.901$, $d = 0.947$) in the third place, and the ANFIS_GP1 ($R = 0.894$, $d = 0.942$) in the fourth place. The worst accuracy with low R and d was for the MLR1 model ($R = 0.649$, $d = 0.766$). As can be seen from Table 3, ANFIS_FC1 has the lowest RMSE and MAE values (RMSE = 1.021 mg/L, MAE = 0.710 mg/L), while MLR1 has the highest RMSE and MAE values (RMSE = 2.259 mg/L, MAE = 1.817 mg/L). Compared to the two other ANFIS models, it is clear that the ANFIS_FC1 model has smaller RMSE and MAE values and higher R and

Table 3 Performances of the developed models for modeling DO concentration

	Training				Validation			
	RMSE	MAE	*R*	*d*	RMSE	MAE	*R*	*d*
Models	(mg/L)	(mg/L)	/	/	(mg/L)	(mg/L)	/	/
MLR1	2.259	1.817	0.649	0.766	2.275	1.824	0.635	0.768
MLR2	2.379	1.938	0.599	0.725	2.352	1.915	0.599	0.732
MLR3	2.661	2.319	0.444	0.574	2.654	2.311	0.431	0.566
MLR4	2.614	2.206	0.475	0.611	2.490	2.117	0.531	0.648
MLR5	2.452	2.013	0.564	0.687	2.383	1.953	0.584	0.705
MLPNN1	1.286	0.942	0.901	0.947	1.844	1.246	0.796	0.891
MLPNN2	1.669	1.214	0.827	0.900	1.854	1.315	0.779	0.878
MLPNN3	1.725	1.284	0.814	0.892	2.400	1.782	0.617	0.786
MLPNN4	2.249	1.767	0.653	0.772	2.154	1.756	0.681	0.780
MLPNN5	2.042	1.581	0.726	0.827	2.161	1.634	0.685	0.811
ANFIS_SC1	1.235	0.888	0.909	0.951	1.528	1.123	0.856	0.922
ANFIS_SC2	1.658	1.198	0.830	0.901	1.874	1.334	0.775	0.874
ANFIS_SC3	1.899	1.431	0.769	0.861	2.171	1.694	0.681	0.813
ANFIS_SC4	2.317	1.879	0.626	0.747	2.389	1.951	0.594	0.737
ANFIS_SC5	2.160	1.697	0.686	0.797	2.248	1.723	0.646	0.775
ANFIS_GP1	1.331	0.973	0.894	0.942	1.662	1.191	0.831	0.908
ANFIS_GP2	1.848	1.363	0.783	0.869	1.865	1.417	0.772	0.863
ANFIS_GP3	2.101	1.662	0.707	0.815	2.193	1.749	0.667	0.793
ANFIS_GP4	2.443	2.036	0.569	0.694	2.399	2.046	0.576	0.695
ANFIS_GP5	2.266	1.782	0.647	0.767	2.103	1.655	0.700	0.795
ANFIS_FC1	1.021	0.710	0.939	0.968	1.662	1.152	0.836	0.914
ANFIS_FC2	1.407	0.999	0.881	0.934	1.804	1.256	0.801	0.893
ANFIS_FC3	1.725	1.284	0.814	0.892	2.400	1.782	0.617	0.786
ANFIS_FC4	2.060	1.602	0.720	0.824	2.240	1.701	0.652	0.787
ANFIS_FC5	1.738	1.311	0.811	0.889	2.105	1.524	0.713	0.838

d values than the ANFIS_GP1 and ANFIS_SC1 models. In the validation phase as seen in Table 3, the best accuracy was achieved using the ANFIS_SC1. It is clear that the ANFIS_SC1 model has smaller RMSE and MAE (RMSE = 1.528 mg/L, MAE = 1.123 mg/L) and higher R and d values ($R = 0.856$, $d = 0.922$) than the ANFIS_FC1, ANFIS_GP1, MLPNN, and MLR models. Also, the ANFIS_FC1 model has a lower RMSE and MAE, and slightly higher *R* and *d* than the ANFIS_GP1 model, and was again better than the MLPNN1 and MLR1 models. This indicates that in general, the ANFIS model is a good modeling tool for DO than MLPNN and MLR. It is also clear from Fig. 6 that the ANFIS models have less scattered estimates than the MLPNN and MLR models. This indicates that in general, the ANFIS model is a good modeling tool for DO compared to MLPNN and MLR. From Table 3, the second input combination provides better accuracy than the third input combination. Similarly, fifth input combination has better performance than the fourth one. Comparison of these combinations shows that the SU variable is more effective on DO than the pH variable.

4.2 *Modeling COD at Sidi Marouane Wastewater Treatment Plant*

Table 4 illustrates the obtained results from all five developed models for predicting COD. From these results, it can be observed that MLPNN, ANFIS_SC, ANFIS_FC, and ANFIS_GP have promising accuracy during the training and validation phases; for all the five combinations, the ANFIS_FC1 has the best accuracy in the training phase. However, during the training or validation phases, the models with two input variables (combination 4 and 5) give low efficiency, low (R and d) and high (RMSE and MAE) values. Also, the models with first and third input combinations give good results compared to the three other combinations; this is certainly due to the inclusion of the pH variable as input. Taking into account the four statistical indices, the ANFIS_FC1 model with four input variables (SS, TE, pH, and SC) gives the best estimation ($R = 0.771$, $d = 0.860$, RMSE $= 7.362$ mg/L, MAE $= 5.471$ mg/L)

Table 4 Performances of the developed models for modeling COD concentration

	Training				Validation			
	RMSE	MAE	*R*	*d*	RMSE	MAE	*R*	*d*
Models	(mg/L)	(mg/L)	/	/	(mg/L)	(mg/L)	/	/
MLR1	8.476	6.656	0.680	0.785	7.658	5.916	0.750	0.840
MLR2	9.431	7.546	0.579	0.693	8.810	7.280	0.651	0.760
MLR3	9.084	7.242	0.619	0.733	8.235	6.788	0.704	0.790
MLR4	9.192	7.344	0.607	0.721	8.645	7.040	0.663	0.757
MLR5	9.602	7.808	0.557	0.671	8.932	7.491	0.637	0.739
MLPNN1	7.824	5.896	0.736	0.837	6.971	5.069	0.790	0.870
MLPNN2	8.390	6.383	0.688	0.795	7.986	6.029	0.710	0.815
MLPNN3	7.454	5.690	0.765	0.854	7.883	5.768	0.726	0.837
MLPNN4	8.254	6.397	0.700	0.807	8.219	6.185	0.689	0.803
MLPNN5	8.647	6.768	0.664	0.773	8.574	6.596	0.655	0.764
ANFIS_SC1	7.922	6.054	0.729	0.828	6.742	4.944	0.805	0.880
ANFIS_SC2	8.565	6.691	0.672	0.781	8.052	6.000	0.704	0.805
ANFIS_SC3	8.235	6.399	0.702	0.807	7.462	5.525	0.755	0.837
ANFIS_SC4	8.678	6.899	0.661	0.772	8.066	6.167	0.705	0.793
ANFIS_SC5	8.859	7.013	0.643	0.757	8.421	6.411	0.671	0.767
ANFIS_GP1	7.689	5.873	0.747	0.841	7.128	5.119	0.779	0.867
ANFIS_GP2	8.263	6.308	0.700	0.805	7.684	5.731	0.736	0.829
ANFIS_GP3	8.083	6.224	0.715	0.816	7.328	5.502	0.765	0.845
ANFIS_GP4	8.657	6.881	0.663	0.774	8.186	6.277	0.693	0.789
ANFIS_GP5	8.823	6.941	0.646	0.761	8.174	6.273	0.696	0.782
ANFIS_FC1	7.362	5.471	0.771	0.860	7.299	5.356	0.767	0.864
ANFIS_FC2	7.987	6.116	0.723	0.824	8.615	6.435	0.665	0.796
ANFIS_FC3	7.579	5.712	0.755	0.850	7.432	5.344	0.757	0.854
ANFIS_FC4	8.262	6.384	0.700	0.807	7.966	5.983	0.712	0.815
ANFIS_FC5	8.681	6.720	0.661	0.774	8.268	6.443	0.685	0.781

among all the other models as shown in Table 4. According to the results in Table 4, it is clear that the MLR1 model has smaller R and d and higher RMSE and MAE in the training phase, compared to the MLPNN1 and the three ANFIS models, and provides the poorest accuracy. Some clear conclusions could be drawn from the Table 4 with respect to validation results. Firstly, for the entire five models developed, it can be concluded from the table that the poorest accuracy was obtained using the MLR1 model with the lowest R and d ($R = 0.771$, $d = 0.860$) and the highest RMSE and MAE (RMSE = 7.362 mg/L, MAE = 5.47 mg/L). Secondly, among the five input combinations (Table 2), the first (with all the input variables) is the best, and the fifth combination (SS, SC) is the worst. Thirdly, according to the results based on three ANFIS, the optimal models are ANFIS_SC1, ANFIS_FC1, and ANFIS_GP1, respectively. It is observed that the ANFIS_SC1 presented the lowest RMSE and MAE (RMSE = 6.742 mg/L, MAE = 4.944 mg/L) values and the highest R and d ($R = 0.805$, $d = 0.880$) values. Accordingly, ANFIS_SC1 is considered the optimal. It can simply be concluded that better performance results can be obtained with ANFIS_SC1. Fourthly and finally, the results of R, d, RMSE, and MAE suggest that the ANFIS_SC1 is slightly better compared to that of MLPNN1 and the MLPNN1 is slightly better compared to those of the ANFIS_FC1 and ANFIS_GP1. From Table 3, the third input combination has better accuracy than the second input combination. Moreover, the fourth input combination shows better performance than the fifth one. Comparison of these combinations indicates that the pH variable is more effective on COD than the TE and SC variables. It can be observed from Fig. 7 that the ANFIS_SC1 model has less scattered estimates than the other models. The slope and bias of its (ANFIS_SC1) fit line equation are, respectively, closer to the 1 and 0 with a lower R compared to MLPNN1, ANFIS_FC1, ANFIS_GP1, and MLR1 models (Fig. 8).

5 Discussion

In the present paper, we developed two artificial intelligence models, ANFIS and ANN, for predicting COD and DO in WWTP and drinking water treatment plant, respectively. Hereafter, we discussed the obtained results in comparison to the results reported in previous similar studies in the literature. Especially, we focused on the results related to the COD. Among all the developed models in our study, ANFIS represents the best accuracy compared to the ANN. Various researchers have attempted to develop models for COD using effluent and affluent water quality variables. The supervised committee fuzzy logic (SCFL) model proposed by Nadiri et al. [8] achieves an R^2 equal to 0.82, significantly superior to 0.648 obtained using ANFIS model in our study. Kisi and Parmar [7] utilized MARS, LSSVM, and M5Tree models for modeling COD and demonstrated that an $R^2 = 0.71$ is obtained using MARS model. In another study, Ay and Kisi [6] compared ANN, GRNN, and RBFNN models for COD estimation, and they obtained an R^2 equal 0.88. Similar to our approach, Pai et al. [11] compared ANFIS and ANN models for COD and demonstrated that ANFIS provided an R^2 equal 0.86 which is superior to the

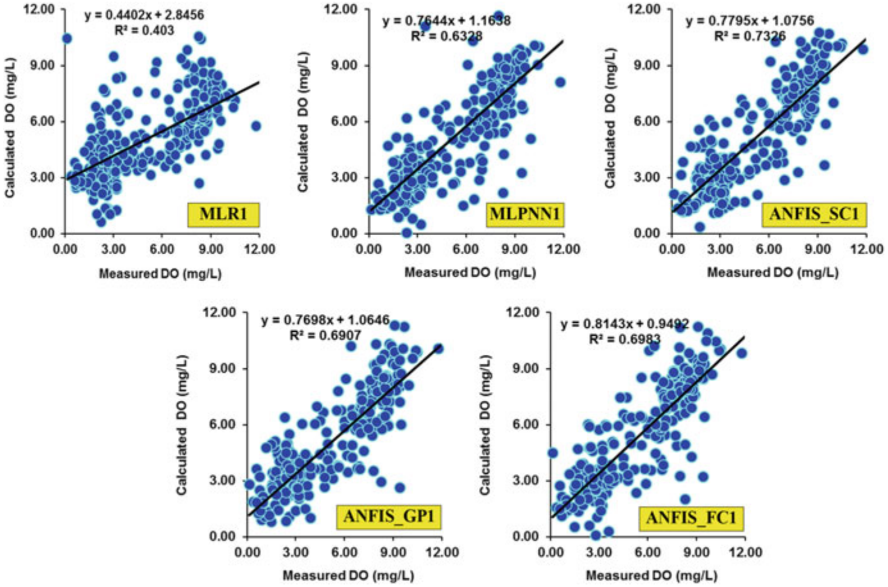

Fig. 7 Scatterplots of predicted versus measured values of dissolved oxygen concentration (DO) using MLPNN1, MLR1, ANFIS_SC1, ANFIS_FC1, and ANFIS_GP1 models in validation phase

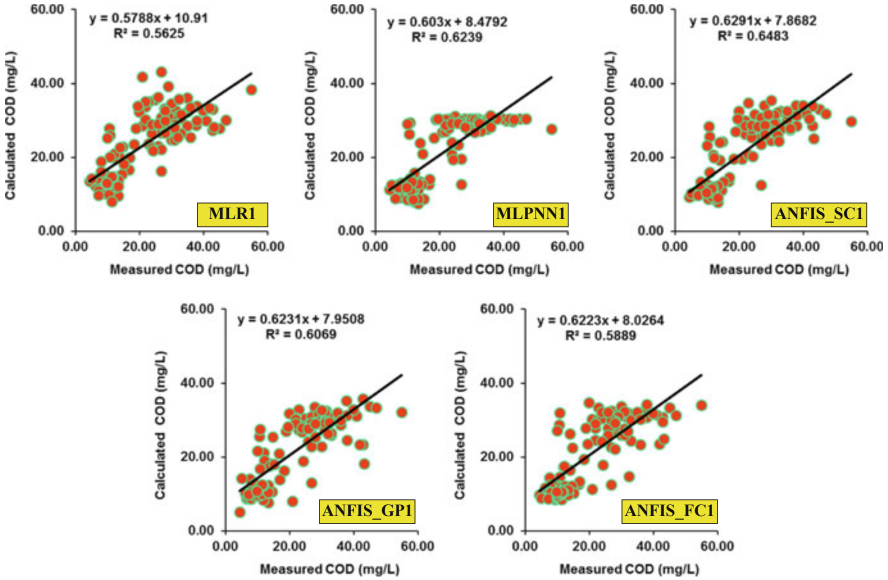

Fig. 8 Scatterplots of predicted versus measured values of chemical oxygen demand (COD) using, MLPNN1, MLR1, ANFIS_SC1, ANFIS_FC1, and ANFIS_GP1 models in the validation phase

R^2 (0.648) obtained using our model. Singh et al. [13] reported that ANN model had an R^2 of 0.84. In the study conducted by Yilmaz et al. [10], MLPNN model was more accurate than GRNN and RBFNN, with an R^2 equal 0.876. Finally, the ANFIS model proposed by Perendeci et al. [12] provided an R^2 equal 0.84. From the discussion reported above, it is clear that our models worked less accurate than the models proposed in the literature, and this is certainly related to the quality of the data used for developing the models.

6 Conclusions

In the present investigation, MLPNN, MLR, and three ANFIS models, namely, ANFIS_GP, ANFIS_SC, and ANFIS_FC, were developed to model two water quality indicators: (1) chemical oxygen demand (COD) and (2) dissolved oxygen concentration (DO). The models were developed using several water quality variables measured at daily time step at WWTP and DWTP, respectively. The input variables used for predicting COD are daily water temperature (TE), suspended solids (SS), specific conductance (SC), and pH, while the input variables used for modeling DO were turbidity (TU), TE, pH, and SC. From the results obtained in the present investigation, some conclusions can be drawn and are summarized as follows:

1. By comparing several combinations of the input variables for modeling DO concentration, the best results were obtained by the ANFIS_SC with TE, pH, SC, and TU inputs, followed by the ANFIS_FC in the second order, ANFIS_GP ranked third, MLPNN ranked fourth, and the MLR model in the last place.
2. In regard to modeling COD, the results showed that the ANFIS_SC with TE, pH, SC, and SS as inputs had the best results and it can be used to estimate COD with very acceptable accuracy, followed by the MLPNN, ANFIS_GP, ANFIS_FC, and MLR, respectively.

Another conclusion we can draw from the results obtained is that the accuracy of the proposed models is mainly dependent to the selection of the input variables, and to obtain good prediction accuracy, it is necessary that all the variables be included for the models.

7 Recommendations

Results obtained in the present study highlighted a number of points that need to be addressed in the future. Firstly, the quality of data must be improved, and the list of variables measured should be enlarged to other variables, notably to include chemical and physical variables that can be good predictors for COD. Secondly, the proposed models should be applied to other WWTP for further comparison of the models' performances.

References

1. Kisi O, Ay M (2014) Comparison of Mann-Kendall and innovative trend method for water quality parameters of the Kizilirmak River, Turkey. J Hydrol 513:362–375. https://doi.org/10.1016/j.jhydrol.2014.03.005
2. Cong Q, Yu W (2018) Integrated soft sensor with wavelet neural network and adaptive weighted fusion for water quality estimation in wastewater treatment process. Measurement 124:436–446. https://doi.org/10.1016/j.measurement.2018.01.001
3. Xiao H, Huang D, Pan Y, Liu Y, Song K (2017) Fault diagnosis and prognosis of wastewater processes with incomplete data by the auto-associative neural networks and ARMA model. Chemom Intell Lab Syst 161:96–107
4. Ruan J, Zhang C, Li Y, Li P, Yang Z, Chen X, Huang M, Zhang T (2017) Improving the efficiency of dissolved oxygen control using an on-line control system based on a genetic algorithm evolving FWNN software sensor. J Environ Manag 187:550–559. https://doi.org/10.1016/j.jenvman.2016.10.056
5. Fernandez de Canete J, Del Saz Orozco P, Baratti R, Mulas M, Ruano A, Garcia-Cerezo A (2016) Soft-sensing estimation of plant effluent concentrations in a biological wastewater treatment plant using an optimal neural network. Expert Syst Appl 63(8):19. https://doi.org/10.1016/j.eswa.2016.06.028
6. Ay M, Kisi O (2014) Modelling of chemical oxygen demand by using ANNs, ANFIS and k-means clustering techniques. J Hydrol 511:279–289
7. Kisi O, Parmar KS (2016) Application of least square support vector machine and multivariate adaptive regression spline models in long term prediction of river water pollution. J Hydrol 534:104–112
8. Nadiri AA, Shokri S, Tsai FTC, Moghaddam AA (2018) Prediction of effluent quality parameters of a wastewater treatment plant using a supervised committee fuzzy logic model. J Clean Prod 180:539–549. https://doi.org/10.1016/j.jclepro.2018.01.139
9. Moral H, Aksoy A, Gokcay CF (2008) Modeling of the activated sludge process by using artificial neural networks with automated architecture screening. Comput Chem Eng 32:2471–2478. https://doi.org/10.1016/j.compchemeng.2008.01.008
10. Yilmaz T, Seckin G, Yuceer A (2010) Modeling of effluent COD in UAF reactor treating cyanide containing wastewater using artificial neural network approaches. Adv Eng Softw 41:1005–1010. https://doi.org/10.1016/j.advengsoft.2010.04.002
11. Pai TY, Yang PY, Wang SC, Lo MH, Chiang CF, Kuo JL, Chu HH, Su HC, Yu LF, Hu HC, Chang YH (2011) Predicting effluent from the wastewater treatment plant of industrial park based on fuzzy network and influent quality. Appl Math Model 35:3674–3684. https://doi.org/10.1016/j.apm.2011.01.019
12. Perendeci A, Arslan S, Tanyolaç A, Celebi SS (2009) Effects of phase vector and history extension on prediction power of adaptive-network based fuzzy inference system (ANFIS) model for a real scale anaerobic wastewater treatment plant operating under unsteady state. Bioresour Technol 100:4579–4587
13. Singh KP, Basant N, Malik A, Jain G (2010) Modeling the performance of "up-flow anaerobic sludge blanket" reactor based wastewater treatment plant using linear and nonlinear approaches-a case study. Anal Chim Acta 658:1–11
14. Yang T, Zhang L, Wang A, Gao H (2013) Fuzzy modeling approach to predictions of chemical oxygen demand in activated sludge processes. Inf Sci 235:55–64
15. Erdirencelebi D, Yalpir S (2011) Adaptive network fuzzy inference system modeling for the input selection and prediction of anaerobic digestion effluent quality. Appl Math Model 35:3821–3832. https://doi.org/10.1016/j.apm.2011.02.015
16. Heddam S, Lamda H, Filali S (2016) Predicting effluent biochemical oxygen demand in a wastewater treatment plant using generalized regression neural network based approach: a comparative study. Environ Process 3:153–165

17. Heddam S, Bermad A, Dechemi N (2011) Applications of radial basis function and generalized regression neural networks for modelling of coagulant dosage in a drinking water treatment: a comparative study. J Environ Eng 137(12):1209–1214
18. Heddam S, Bermad A, Dechemi N (2012) ANFIS-based modelling for coagulant dosage in drinking water treatment plant: a case study. Environ Monit Assess 184:1953–1971. https://doi.org/10.1007/s10661-011-2091-x
19. Heddam S, Dechemi N (2015) A new approach based on the dynamic evolving neural-fuzzy inference system (DENFIS) for modelling coagulant dosage: case study of water treatment plant of Algeria country. Desalin Water Treat 53(4):1045–1053. https://doi.org/10.1080/19443994.2013.878669
20. Olden JD, Jackson DA (2002) Illuminating the "black box": understanding variable contributions in artificial neural networks. Ecol Model 154:135–150
21. Houichi L, Dechemi N, Heddam S, Achour B (2013) An evaluation of ANN methods for estimating the lengths of hydraulic jumps in U-shaped channel. J Hydroinf 15(1):147–154
22. Heddam S (2014) Modelling hourly dissolved oxygen concentration (DO) using two different adaptive neuro-fuzzy inference systems (ANFIS): a comparative study. Environ Monit Assess 186:597–619. https://doi.org/10.1007/s10661-013-3402-1
23. Keshtegar B, Heddam S (2018) Modeling daily dissolved oxygen concentration using modified response surface method and artificial neural network: a comparative study. Neural Comput Appl 30(10):2995–3006. https://doi.org/10.1007/s00521-017-2917-8
24. Haykin S (1999) Neural networks a comprehensive foundation. Prentice Hall, Upper Saddle River
25. Rumelhart DE, Hinton GE, Williams RJ (1986) Learning internal representations by error propagation. In: Rumelhart DE, McClelland PDP, Research Group (eds) Parallel distributed processing: explorations in the microstructure of cognition. Foundations, vol I. MIT Press, Cambridge, pp 318–362
26. Hornik K (1991) Approximation capabilities of multilayer feedforward networks. Neural Netw 4(2):251–257. https://doi.org/10.1016/0893-6080(91)90009-T
27. Hornik K, Stinchcombe M, White H (1989) Multilayer feedforward networks are universal Approximators. Neural Netw 2:359–366. https://doi.org/10.1016/0893-6080(89)90020-8
28. Jang JS (1993) ANFIS: adaptive-network-based fuzzy inference system. IEEE Trans Syst Man Cybern 23(3):665–685
29. Zhu S, Heddam S, Nyarko E, Hadzima-Nyarko M, Piccolroaz S, Wu S (2019) Modelling daily water temperature for rivers: adaptive neuro-fuzzy inference systems vs. artificial neural networks models. Environ Sci Pollut Res 26(1):402–420. https://doi.org/10.1007/s11356-018-3650-2
30. Zhu S, Heddam S, Wu S, Dai J, Jia B (2019) Extreme learning machine based prediction of daily water temperature for rivers. Environ Earth Sci 78:202
31. Zhu S, Nyarko E, Hadzima-Nyarko M, Heddam S, Wu S (2019) Assessing the performance of a suite of machine learning models for daily river water temperature prediction. PeerJ 7:e7065