Comparison of Data-Driven Methods for Estimating Deuterium and Oxygen-18 Isotopes of Groundwater

Bilal CEMEK
Ondokuz Mayis University Faculty of Agriculture: Ondokuz Mayis Universitesi Ziraat Fakultesi
Hakan ARSLAN
Ondokuz Mayis Universitesi
erdem küçüktopcu (erdem.kucuktopcu@omu.edu.tr)
Ondokuz Mayis University: Ondokuz Mayis Universitesi

Research Article

Keywords: Artificial intelligence, Isotope, Deuterium, Oxygen-18, Groundwater

Posted Date: May 6th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-316255/v1

Comparison of data-driven methods for estimating deuterium and oxygen-18 isotopes of groundwater

Bilal Cemek[1,*], Hakan Arslan[1], Erdem Küçüktopcu[1]

[1]Ondokuz Mayıs University, Agriculture Faculty, Agricultural Structures and Irrigation Department, Samsun, Turkey

bcemek@omu.edu.tr (Corresponding author)

hakan.arslan@omu.edu.tr

erdem.kucuktopcu@omu.edu.tr

Abstract

Isotope techniques are the most commonly used in cases where hydro-chemical analysis is insufficient to identify groundwater's origin and quality and reveal seawater intrusion into groundwater along coastlines. In this study, the potential of the Multilayer Perceptron (MLP), Radial Basis Neural Networks (RBNN), Generalized Regression Neural Networks (GRNN), Adaptive Neuro-Fuzzy Inference System (ANFIS), Support Vector Machines (SVM), Gaussian Process Regression (GPR), Classification and Regression Tree (CART), and Multiple Linear Regression Analysis (MLR) were compared using known hydro-chemical properties of waters for estimating deuterium (δD) and oxygen-18 ($\delta^{18}O$) isotopes in groundwater of the Bafra plain, Northern Turkey. A total of 61 water samples collected from the plain were chemically analyzed. All data were divided into training (70%) and test (30%) sets. Cluster analysis was performed to reduce the number of input variables, and electrical conductivity (EC), chloride (Cl), magnesium (Mg) and, sulphate (SO_4) were introduced into the models as input variables, after examining different combinations of these variables in the studied models. Three statistical indices were used to evaluate models' performances: determination coefficient (R^2), root mean square error (RMSE) and mean absolute error (MAE). Moreover, a visualization technique (Taylor diagram) was used to assess the similarities between the measured and estimated δD and $\delta^{18}O$ values. The comparison revealed that the performance accuracy of MLP was the best

among the applied models in δD and $\delta^{18}O$ estimations. Overall, the study suggests using data-driven methods, especially MLP, when lacking of appropriate laboratories for isotope analysis and facing with high cost.

Keywords Artificial intelligence. Isotope. Deuterium. Oxygen-18. Groundwater

Introduction

Groundwater, which is one of the most valuable natural resources, has a dual character. On the one hand, it is a resource that moves in the depths of the earth and abstracted from it; on the other hand, it is a part or total water resource. The dominant role of groundwater resources is clear; therefore, their use and protection are essential for human life and economic activity (Zektser and Everett 2004). The most beneficial use of groundwater resources is possible by determining the amount of recharge/discharge of groundwater. So, the studies of the identification origin of groundwater are of great importance in hydrology and hydrogeology (Nair et al. 2015; Seddique et al. 2019)

With the ever-increasing industrial development and increasing population, it becomes necessary to use advanced techniques alongside classical investigations to develop and use water resources efficiently and sustainably. Isotope hydrology is at the top of these techniques, and stable isotopes of oxygen-18 ($\delta^{18}O$) and hydrogen ($\delta^{2}H$) are frequently used to identify the origins of groundwater (Jayathunga et al. 2020; Maurya et al. 2019). The $\delta^{18}O$ and $\delta^{2}H$, among the environmental isotopes, are good tracers as they are natural water components. These isotopes are not affected by most of the hydro-geochemical processes that develop in the aquifer. However, they give clues about the physical and chemical processes by which water is affected. Investigating the distribution of the stable isotopes in natural waters, the origin of groundwater and mixing in aquifers can be determined.

The studies on water quality assessment revealed that the high volume of groundwater abstraction, excessive pumping, and less recharge in coastal wells lead to seawater intrusion,

thus increasing groundwater salinization (Klassen and Allen 2017; Mohanty and Rao 2019). The hydro-chemical characteristics of groundwater and stable isotopes ($\delta^{18}O$ and $\delta^{2}H$) have been used to identify groundwater origin and dynamics. Many studies revealed that there are close relations between the stable isotopes and other hydro-chemical properties of groundwater (Arslan et al. 2012; Bahir et al. 2018; Isawi et al. 2016; Jahnke et al. 2019; Lin et al. 2011; Mongelli et al. 2013).

Today, Artificial Intelligence (AI) techniques, such as the Artificial Neural Networks (ANN), the Adaptive Neuro-Fuzzy Inference System (ANFIS), and the Support Vector Machine (SVM), are being progressively used to solve a number of complicated problems in various research fields and are therefore becoming more and more popular (Shanmuganathan 2016). The most important advantage of AI models is that they can effectively tackle the nonlinearity and complexity of a system and overcome the drawbacks of using numerical models. In recent years, some studies have been employed AI techniques to predict water quality parameters, and successful results have been achieved. Barzegar and Moghaddam (2016) evaluated the performance of three different ANN models, Multilayer Perceptron (MLP), Generalized Regression Neural Network (GRNN), and Radial Basis Function Neural Network (RBNN) in the estimation of groundwater salinity of the Tabriz plain. Hameed et al. (2017) predicted water quality index (WQI) using a Backpropagation Neural Network (BPNN) and RBNN. Lal and Datta (2018) studied Genetic Programming (GP) and Gaussian Process Regression (GPR) models in modeling groundwater salinity. Juntakut et al. (2019) studied the Classification and Regression Tree (CART) to predict nitrate concentrations of groundwater in Nebraska. Jafari et al. (2019) estimated total dissolved solids of the groundwater aquifer in Tabriz plain using four soft computing approaches, namely, MLP, ANFIS, SVM, and Gene Expression Programming (GEP). Noori et al. (2020) developed a hybrid model by combining a process-based watershed model and ANN to improve the water quality predictions. Cahyadi

et al. (2020) applied BPNN model to estimate the hydraulic conductivity of fractured groundwater flow media.

These studies have contributed significantly to the knowledge base regarding the use of AI technology to estimate water quality parameters. However, no study has been reported to date on applying the AI to estimate deuterium (δD) and oxygen-18 ($\delta^{18}O$) isotopes in groundwater. Considering isotopes analysis is very expensive and very few laboratories equipped for to carry out this analysis, the objective of this research was to develop a simple, rapid, economical, and accurate model for estimating δD and $\delta^{18}O$ isotopes in groundwater by using different data-driven models, including MLP, ANFIS, RBNN, GRNN, SVM, GPR, CART, and MLR.

Material and Methods

Study area and data set

This study was conducted in Kızılırmak Delta of Bafra Plain, which is located in the middle black sea region of northern Turkey (41°30′-41°45′ North latitudes, 35°30′-36°15′ East longitudes) (Fig. 1). Soils in the study are composed of young alluvial deposits, brought by Kızılırmak River. While the predominant soils on the area's periphery are hydromorphic soils, coastal dunes are found at the seaside and colluvial-alluvial soils in the inland area. This plain is therefore regarded as one of having among the most fertile soils for agriculture in Turkey.

Fig. 1 is near here

The delta's climate is semi-humid, with temperatures ranging from 6.60 °C in January to 23.80 °C in July (average 14.40 °C). The average annual precipitation is between 536.4 mm and 783.5 mm. The elevation rises considerably from the coast to the inland and reaches 10 m within about 6 km. Because of the changes in elevation, there are great fluctuations in groundwater levels and drainage within the study area. Irrigation is generally performed by

using largely border and furrow irrigation methods. About 75% of the land is irrigated with surface water and the rest 25% with groundwater (Cemek et al. 2007).

For this analysis, a total of sixty-one water samples was taken from October 2007 to September 2008 in different location of the study area including fifty-six from the 28 different monitoring wells in Bafra plain, and five from Black sea. Samples filtered with a 0.45µm filter, enclosed in polyethylene bottles, and stored at 4ºC until processing. Electrical conductivity (EC) and pH of water samples were measured a handheld portable kit in situ. Major cation (K^+, Na^+, Ca^{+2}, Mg^{+2}) and anion (Cl^-, SO_4^{-2}) concentrations were analyzed in the laboratory using ion chromatography (761 Compact IC, Metrohm Schweiz AG, Switzerland). The HCO_3^- levels were determined by titration.

The isotopes (δD and $\delta^{18}O$) measurements were carried out at the General Directorate of Hydraulic Work (DSI), Department of Technical Research and Quality Control. As usual, the isotopic composition is expressed in δ per mil, i.e., deviation ‰ of the isotope ratios $^2H/^1H$ and $^{18}O/^{16}O$ from the reference Vienna Standard Mean Ocean Water (V-SMOW). The statistical summary of the analysis results is given in Table 1. As seen from Table 1, values for $\delta^{18}O$ range from -3.59 to -10.69‰ with an average of -8.16‰ and for δD range from -19.56 to -72.64‰ with an average of -50.42‰. The EC values vary between 1.86 and 28.80 dSm^{-1} with an average of 5.96 dSm^{-1}. The pH values vary from 7.28 to 8.16, having an average of 7.65. The concentrations (mgL^{-1}) of Ca^{+2}, Mg^{+2}, K^+, and Na^+ range from 37.00 to 305.00, 46.00 to 839.00, 4.00 to 200.00, and 247.70 to 6900.00, respectively. Average concentrations (mgL^{-1}) of these major cations are 102.44, 170.52, 35.03, and 1170.21, respectively. As regards the concentrations (mgL^{-1}) of Cl^-, SO_4^{-2}, and HCO_3^-, their values vary from 345.34 to 9800.00 (on average: 1551.02), 40.00 to 1950.00 (on average: 419.34), and 174.44 to 1617.00 (on average: 840.07), respectively.

Table 1. is near here

Data-Driven Techniques

Multilayer Perceptron (MLP)

The MLP is a widely used type of neural network. The structure of a simple MLP is shown in Fig. 2. In this network, the information moves from the input layer neurons, through hidden layer(s) neurons, and to the output layer neurons (Skansi 2018). Every layer except the output layer have a bias neuron and is connected to the next layer (Haykin 2001).

Fig. 2 is near here

Levenberg Marquardt (LM) learning algorithm was employed in this research (Cigizoglu and Kişi 2005; El-Bakry 2003). Different network topologies (single or double hidden layer) with a tangent sigmoid (tansig) and linear (purelin) transfer functions were used in hidden and output layers, respectively. The number of hidden nodes changed from 3 to 7 to achieve the optimal training network.

Radial Basis Neural Networks (RBNN)

Introduced into the neural network literature by Broomhead and Lowe (1988), the RBNN represent an efficient mechanism for local approximators complex nonlinear functions (Mai-Duy and Tran-Cong 2003), pattern classification and recognition (Umasankar and Kalaiarasi 2014), modeling and controlling dynamic systems (Zhao 2008) from the input-output data. RBNN are distinguished from other neural networks due to their universal approximation and shorter training phase. RBNN networks are a special class of feed-forward neural networks composed of three layers (Fig. 3). The first layer (input layer) represents the input data. The second (hidden layer) comprises a number of nodes that implement a nonlinear transformation to the input variables. The last corresponds to the final output of the network. This layer uses a linear activation function. Extensive detail on RBNN method can be obtained from Haykin (2001).

Fig. 3 is near here

153 **Generalized Regression Neural Networks (GRNN)**

154 The GRNN was initially proposed and developed by Specht (1991). As seen from Fig. 4, the
155 GRNN consists of input, pattern, summation, and output layers. The total number of variables
156 is equal to that of the first layer of input units. In pattern layer, each unit represents a training
157 pattern. In the summation layer, the S and D- summation neurons are connected to each layer
158 of the pattern. The S- summation neuron sums the product of the weights outputs of the pattern
159 layer, while the D- summation neuron calculates the sum of all weights. The last layer divides
160 S- summation neuron by D- summation neuron and produces the predicted output. The
161 outstanding aspect of GRNN is its use of a smoothing factor, which affects the extent to which
162 the network can be generalized. Low smoothing factors impair the ability of network to be
163 generalized, while high smoothing values, although good for generalization, often increase
164 prediction error. Detailed information for GRNN can be obtained from Haykin (2001).

165 *Fig. 4 is near here*

166 **Adaptive Neuro-Fuzzy Inference System (ANFIS)**

167 This system integrates the ANN's learning ability and relational structure with the decision-
168 making mechanism of the fuzzy inference system (FIS). ANFIS performs learning with samples
169 using a training dataset, as is done with ANN. In this way, the optimal ANN structure for
170 solving the associated problem is obtained. In order to identify its effect on samples that were
171 not previously observed, this structure is subjected to the test process. The lower error values
172 demonstrate the ANFIS model's suitability. One of the main drawbacks of ANN, though, is that
173 it cannot justify the weight values gained. This shortcoming is eliminated by the FIS, which is
174 found in the structure of the ANFIS. Because of this structure, it is used for solving many real-
175 world problems. Detailed information about ANFIS can be obtained from Daneshmand et al.
176 (2015).

177

178 **Support Vector Machines (SVM)**

179 SVM is a binary learning machine that can be used both for pattern recognition and nonlinear
180 regression problems. Similar to ANN, SVM can be interpreted as two-layer networks where the
181 weights are nonlinear in the first layer and linear in the second layer (Bray and Han 2004). The
182 SVM can efficiently handle a nonlinear classification/modeling using the kernel method.

183 Given a training data of n samples $\{z_i = (x_i, y_i),\ i = 1,2,3,\mathrm{K}\ ,n\}$, where $x_i \in R^n$ is the
184 training data and $y_i = R$ is the response for x_i. By solving the Kuhn-Tucker conditions of the
185 following quadratic optimization problem can be written as

$$\begin{aligned} &Min && \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \\ &Subject\ to && y_i - \omega.\varphi(x_i) - b \le \varepsilon + \xi_i,\quad \xi_i \ge 0 \\ & && \omega.\varphi(x_i) + b - y_i \le \varepsilon + \xi_i^*,\quad \xi_i^* \ge 0 \\ & && \forall i, i \in (1,2,\mathrm{K}\ ,n) \end{aligned} \tag{1}$$

187 where $1/2\,\omega^2$ is the regularization term, $\varphi(x_i)$ is nonlinear mapping function, C is the
188 error penalty factor to regulate the difference between the empirical error and the regularization
189 term, ξ_i and ξ_i^* are the positive slack variables, ε is the loss function, b is scalar, and ω represents
190 a normal vector.

191 The best decision function *f(x)* can be stated as;

$$f(x) = \sum_{i=1}^{N}(a_i - \alpha_i^*)k(x_i, x) + b \tag{2}$$

193 where, a_i α_i^* $(i = 1,2,3,\mathrm{K}\ ,n)$ are the Lagrange multipliers, and $k(x_i, x)$ is the kernel
194 function.

195 In general, different kernel functions (i.e., linear, polynomial, Gaussian, exponential)
196 were employed in SVM. This study used a polynomial kernel function (Achirul Nanda et al.
197 2018). Readers may refer to Vapnik (1995) and Vapnik et al. (1997) for more information on
198 SVM theoretical basis.

199 **Gaussian Process Regression (GPR)**

200 GPR is another popular nonparametric kernel-based machine learning method that clarifies the
201 response by presenting latent variables $f(x_i)$, $i = 1,2,3,\mathrm{K}\ ,n$, from a Gaussian process (GP)
202 and explicit basis functions. A GP is a finite number of random variables with a joint Gaussian
203 distribution.

204 A GP *f(x)* is specified by a mean *m(x)* function and covariance function which are
205 defined as follows (Williams and Rasmussen 1996):

206 $$m(x) = E(f(x)) \tag{3}$$

207 $$k(x_i, x_j) = E\left[\left\{f(x_i) - m(x_i)\right\}\left\{f(x_j) - m(x_j)\right\}\right] \tag{4}$$

208 when $f(x):\ GP(0, k(x_i, x_j))$ GPR model can be defined as:

209 $$P(f \mid X):\ N(f \mid 0, K(X, X)) \tag{5}$$

210 $$K(X, X) = \begin{pmatrix} k(x_1, y_1) & \mathrm{L} & k(x_1, y_1) \\ \mathrm{M} & \mathrm{O} & \mathrm{M} \\ k(x_1, y_1) & \mathrm{L} & k(x_1, y_1) \end{pmatrix} \tag{6}$$

211 The covariance function is described by different kernel functions, which can be
212 configured in kernel parameters in vector θ, therefore the covariance function is stated as
213 $k(x_i, x_j \mid \theta)$. In this study, four different kernel functions including rational quadratic,
214 exponential, squared exponential, and matern 5/2 were employed to estimate the δD and $\delta^{18}O$.
215 Detailed information for GPR were obtained by Rasmussen (2003).

216 **Classification and Regression Tree (CART)**

217 Breiman et al. (1984) developed the CART, a nonparametric modeling approach. This
218 technique is one of methods used most commonly in groundwater research management fields
219 (Choubin et al. 2019; Juntakut et al. 2019; Knoll et al. 2019; Naghibi et al. 2016; Zhao et al.
220 2016) as it is a model that is easy to understand and interpret (Genç et al. 2015).The process

begins with dividing the data into subgroups so that the samples in the child nodes become more homogeneous than those in the parent nodes. This process continues until the CART reaches a steady state in which the parent nodes do not improve the entire tree's capacity. In regression problems, CART can predict the targeted outcome effectively by applying the least-square deviation criteria. As it is not a black box model, in a CART model, the influence of each variable on the desired result can be visualized clearly in its corresponding tree structure. For a detailed information on CART model, readers are referred to Breiman et al. (1984) and Chipman et al. (1998).

Multiple Linear Regression Analysis (MLR)

In statistics, regression analysis is a statistical tool for predicting the nature of relation among different variables. The MLR technique uses a linear equation to match the observed data to the relations of two or more explanatory (independent) variables and a response (dependent) variable. The general equation for a MLR model is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \mathrm{L} + \beta_n x_n + \varepsilon \quad (7)$$

where y is the dependent variable; x_1, x_2 and x_n are the explanatory variables; β_1, β_2 and β_n are slope coefficients for each explanatory variable; β_0 is the constant term; ε is the model's error term.

Cluster analysis

Cluster analysis is a technique for grouping related observations into a set of clusters based on the observations of multiple variables for each individual. This technique has been applied to effectively classify water samples in many studies (Egbueri 2020; Kazakis et al. 2017; Rao and Chaudhary 2019; Yang et al.,2020). In this study, cluster analysis was applied to group groundwater samples for $\delta^{18}O$, δD, EC, pH, Ca, Mg, K, Na, Cl, SO_4, and HCO_3 content.

Data pre-processing

Before developing the model structure, the input and output data were standardized between 0 and 1 to certify the equal handling of all variables and enhance the efficiency of the training network. The following equation is used to normalize data:

$$Z_{norm} = \frac{\left(Z_i - Z_{\min}\right)}{\left(Z_{\max} - Z_{\min}\right)} \tag{8}$$

where, Z_{norm} is the standardized value, Z_i is the measured value, $Z_{\min}$ and $Z_{\max}$ are the minimum and maximum value, respectively.

In this research, we carried out a model selection in a training data set of 70%, and in a testing a data set of 30% by using the *k*-fold cross-validation technique. In this technique, the initial data is randomly portioned into *k* equally-sized subsets (*k*-folds). Of the *k* partitions, a single subset is designated as the validation data to evaluate the model's performance, and the remaining *k*-1 subsets are used as training data. This process is repeated *k* times, and the *k* results from the folds can be averaged then to produce a single estimation. In this research, *k* is set to 10. Detailed information on this procedure can be obtained from Cemek et al. (2020).

Performance criteria of model

In this study, the performances of models are evaluated by the use of the coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE). The equations were defined as below (Waller 2003).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(Z_i - Z_{i^*}\right)^2}{\sum_{i=1}^{n}\left(Z_i - \overline{Z_i}\right)^2} \tag{9}$$

$$RMSE = \sqrt{\frac{\sum\left(Z_{i^*} - Z_i\right)^2}{n}} \tag{10}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|Z_{i^*} - Z_i\right| \tag{11}$$

Where, Z_i is the measured value; Z_{i^*} is the predicted value; $\overline{Z}_i$ is the average value measured; n is the number of data.

In addition, Taylor diagrams were used to analyze the standard deviation (SD) and correlation coefficients (R) between the model-predicted and observed values.

Results and Discussions

In this paper, four different types of ANN models—MLP, GRNN, RBNN, and ANFIS— and MLR were used to estimate the δD and $\delta^{18}O$ isotopes in groundwater of the Bafra plain. Before developing the ANN and MLR models, hierarchical cluster analysis was applied to reduce the number of input variables, and similarities between the variables were determined. Three groups were generated from hierarchical cluster analysis (Fig. 5). Group A includes EC, Cl, Na, Mg, SO_4, δD, and $\delta^{18}O$. The variables of K, Ca, and pH composed group B, the HCO_3 was constituted mainly of the group C. According to this evaluation, it was seen that there was the similarity between the isotopes (δD and $\delta^{18}O$) and the variables of EC, Cl, Na, Mg, and SO_4. Therefore, δD and $\delta^{18}O$ isotopes were estimated by using these variables in the same group as isotopes.

Fig. 5 is near here

Different input combinations were tried in this study to assess the degree of effect of each variable on δD and $\delta^{18}O$ values. The input combinations test in the present article were (i) EC; (ii) EC and Cl; (iii) EC and SO_4; (iv) EC, Cl and SO_4; (v) EC, Cl, SO_4, and Mg.

In MLP models, different numbers of hidden nodes were tested, and the optimal one that generated the lowest RMSE in the testing phase was selected. In the ANFIS technique, different membership functions (MFs) such as triangle, trapezoidal, and Gaussian with varying MFs were tried to find the best outputs. In the RBNN model, the optimum spread and hidden node numbers were determined using a trial-error approach. In the GRNN application, optimal models were simply obtained using different spread values by trial and error method. In the

290 SVM technique, optimal parameters of SVM are selected using rule and the stopping criteria.
291 In the MLR analysis, the δD and $\delta^{18}O$ were used as dependent variables, whereas EC, Cl, SO_4,
292 and Mg were considered independent variables in determining the regression equations for the
293 training data set. The obtained regression equations were then used to determine the estimated
294 δD and $\delta^{18}O$ for the testing data set.

295 The training and testing results of the MLP, ANFIS, RBNN, GRNN, SVM, GPR,
296 CART, and MLR models for $\delta^{18}O$ estimates are given in Table 2. In the training period, the
297 range of RMSE values for MLP, ANFIS, RBNN, GRNN, SVM, GPR, CART, and MLR were
298 0.19-0.47, 0.34-0.47, 0.34-0.47, 0.50-0.72, 0.30-0.74, 0.21-0.35, 0.28-0.35, and 0.55-0.70‰,
299 respectively. The findings revealed that the minimum value of the RMSE (0.19‰) was obtained
300 for the MLP5(4,5,1) model whose inputs are the EC, Cl, SO_4, and Mg; however, the maximum
301 value was found as 0.72‰ for SVM2(Gaussian) model whose inputs are EC and SO_4. Similarly,
302 the lowest value of MAE was 0.10‰ for the MLP5(4,5,1) model, while the highest value was
303 0.56‰ for the MLR1 model. Also, R^2 values between the measured and estimated $\delta^{18}O$ were
304 between 0.91-0.98 for MLP models and 0.94-0.98 for GPR models, whereas this value varied
305 between 0.77 and 0.97 for the other models.

306 In the testing period, the range of RMSE values for MLP, ANFIS, RBNN, GRNN, SVM,
307 GPR, CART, and MLR were between 0.31-0.54, 0.35-0.70, 0.32-0.68, 0.44-0.74, 0.60-0.85,
308 0.33-0.44, 0.39-0.41, and 0.48-0.70‰, respectively. Also, the minimum MAE was attained for
309 the MLP5(4,5,1) model (0.20‰), whereas the maximum value was found as 0.64‰ for
310 SVM2(Gaussian) model. For MLP, ANFIS, RBNN, GRNN, SVM, GPR, CART, and MLR
311 models, the range of R^2 values were 0.93-0.98, 0.86-0.97, 0.87-0.97, 0.83-0.95, 0.85-0.94, 0.95-
312 0.98, 0.96-0.97, and 0.86-0.93‰, respectively.

313 *Table 2. is near here*

The training and testing results of the MLP, ANFIS, RBNN, GRNN, SVM, GPR, CART, and MLR models for δD estimates are given in Table 3. In the training period, the range of RMSE values for MLP, ANFIS, RBNN, GRNN, SVM, GPR, CART, and MLR were 2.05-2.64, 2.35-5.23, 3.46-9.03, 4.14-5.59, 2.36-5.96, 2.63-3.19, 2.97-3.24, and 4.56-5.69‰, respectively. Like the RMSE criterion, the minimum MAE value (1.36‰) was achieved using the SVM5(cubic) model, while the maximum MAE value (4.23‰) was found for RBNN3(2,0.6310) whose input variables, spread parameter value, and the number of hidden nodes were 2, 0.6, and 10, respectively. Similarly, R^2 values between the measured and estimated δD for MLP, ANFIS, RBNN, GRNN, SVM, GPR, CART, and MLR were 0.94-0.96, 0.86-0.95, 0.84-0.89, 0.71-0.85, 0.68-0.95, 0.90-0.94, 0.90-0.91, and 0.70-0.81, respectively.

In the testing period, the minimum and maximum RMSE values for MLP, ANFIS, RBNN, GRNN, SVM, GPR, CART, and MLR were 2.77-2.97, 3.02-4.76, 3.92-4.89, 4.04-5.29, 4.14-4.99, 3.17-3.57, 3.50-3.96, and 4.24-4.75‰, respectively. The lowest MAE value (1.89‰) was obtained for the MLP5(4,5,1) model, while the highest MAE value (4.03‰) was found for the RBNN2(2,0.5,10) model. The R^2 values ranged from 0.94 to 0.95 for MLP models, and 0.80 to 0.95 for other models.

Table 3. is near here

The measured and estimated $\delta^{18}O$ and δD values by the optimal models for MLP, ANFIS, RBNN, GRNN, SVM, GPR, CART, and MLR were plotted in Fig.6 and Fig.7, respectively.

Fig. 6. is near here

Fig. 7. is near here

From these figures, MLP5(4,5,1) models seem to have better results than the other studied models for $\delta^{18}O$ and δD estimation. Figure 8 shows the scatter plots of the MLP5(4,5,1) models for the measured and modelled $\delta^{18}O$ and δD values for testing period.

Fig. 8. is near here

Also, the $\delta^{18}O$ and δD estimation models were evaluated by using a Taylor diagram (Fig. 9). It is shown that the MLP5(4,5,1) models provided a lower SD and RMSE, and a higher correlation coefficient compared to the other studied models. Therefore, comparison of the findings of the models shows that the MLP5(4,5,1) models were the most accurate model in the prediction $\delta^{18}O$ and δD.

Fig. 9 is near here

Conclusions

In the studies related seawater intrusion, using traditional methods such as Piper diagram or Ca/Na, Cl/HCO_3, Ca/Cl, Mg/Cl, Mg/Ca, and SO_4/Cl molar ratios alone are not appropriate to determine the origin of waters. The study recommends using traditional methods and isotopic methods together when exploring the origins of the waters rather than conventional ones solely. In this study, different data-driven models called MLP, ANFIS, RBNN, GRNN, SVM, GPR, CART, and MLR were employed and their performances were assessed using known hydro-chemical properties of waters to estimate $\delta^{18}O$ and δD. The results of these techniques were statistically compared by using R^2, RMSE, and MAE. Taylor diagram was also employed to evaluate the studied models' performance.

Comparative analysis of the models' results indicated that MLP5 (4,5,1) generated the most suitable models for all estimations based on R^2, RMSE, and MAE for $\delta^{18}O$ estimation (0.98, 0.31‰, and 0.20‰. respectively), and δD estimation (0.95, 2.85‰, and 1.89‰ respectively) for the testing datasets. Also, in the case of scarce data, the MLP1 (1,5,1) model with only the EC generated satisfactory results in estimating of δD and $\delta^{18}O$.

Overall, the study suggests using data driven methods, especially MLP, when lacking of appropriate laboratories for isotope analysis and facing with high cost.

Declarations

Ethics approval and consent to participate Not applicable.

Consent for publication Not applicable.

Availability of data and materials All data is available in the paper.

Competing interests The authors declare that they have no competing interest.

Funding Not applicable.

Authors' contribution BC: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Supervision, Validation, Writing- original draft, Writing - review & editing. HA: Investigation, Resources, Writing - review & editing. EK: Formal analysis, Software, Validation, Visualization, Writing- original draft , Writing - review & editing.

References

Achirul Nanda M, Boro Seminar K, Nandika D, Maddu A (2018) A comparison study of kernel functions in the support vector machine and its application for termite detection. Information 9:5

Arslan H, Cemek B, Demir Y (2012) Determination of seawater intrusion via hydrochemicals and isotopes in Bafra Plain, Turkey. Water Resour Manag 26:3907–3922

Bahir M, Ouhamdouch S, Carreira PM (2018) Geochemical and isotopic approach to decrypt the groundwater salinization origin of coastal aquifers from semi-arid areas (Essaouira basin, Western Morocco). Environ Earth Sci 77:485

Barzegar R, Moghaddam AA (2016) Combining the advantages of neural networks using the concept of committee machine in the groundwater salinity prediction. Model Earth Syst Environ 2:26

Bray M, Han D (2004) Identification of support vector machines for runoff modelling. J Hydroinformatics 6:265–280

Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees (Wadsworth, Belmont, CA). ISBN-13 978–412048418

Broomhead DS, Lowe D (1988) Multivariable functional interpolation and adaptive networks, complex systems, vol. 2

Cahyadi TA, Syihab Z, Widodo LE, et al (2020) Analysis of hydraulic conductivity of fractured groundwater flow media using artificial neural network back propagation. Neural Comput Appl

Cemek B, GüLer M, Kiliç K, et al (2007) Assessment of spatial variability in some soil properties as related to soil salinity and alkalinity in Bafra plain in northern Turkey. Environ Monit Assess 124:223–234

Cemek B, Ünlükara A, Kurunç A, Küçüktopcu E (2020) Leaf area modeling of bell pepper (Capsicum annuum L.) grown under different stress conditions by soft computing approaches. Comput Electron Agric 174:105514

Chipman HA, George EI, McCulloch RE (1998) Bayesian CART model search. J Am Stat Assoc 93:935–948

Choubin B, Rahmati O, Soleimani F, et al (2019) Regional groundwater potential analysis using classification and regression trees. In: Spatial modeling in GIS and R for earth and environmental sciences. Elsevier, pp 485–498

Cigizoglu HK, Kişi Ö (2005) Flow prediction by three back propagation techniques using k-fold partitioning of neural network training data. Hydrol Res 36:49–64

Daneshmand H, Tavousi T, Khosravi M, Tavakoli S (2015) Modeling minimum temperature using adaptive neuro-fuzzy inference system based on spectral analysis of climate indices: A case study in Iran. J Saudi Soc Agric Sci 14:33–40

Egbueri JC (2020) Groundwater quality assessment using pollution index of groundwater (PIG), ecological risk index (ERI) and hierarchical cluster analysis (HCA): a case study. Groundw Sustain Dev 10:100292

El-Bakry MY (2003) Feed forward neural networks modeling for K–P interactions. Chaos, Solitons & Fractals 18:995–1000

Genç O, Gonen B, Ardıçlıoğlu M (2015) A comparative evaluation of shear stress modeling based on machine learning methods in small streams. J Hydroinformatics 17:805–816

Hameed M, Sharqi SS, Yaseen ZM, et al (2017) Application of artificial intelligence (AI) techniques in water quality index prediction: a case study in tropical region, Malaysia. Neural Comput Appl 28:893–905

Haykin SS (2001) Neural networks: a comprehensive foundation. Tsinghua University Press, Beijing

Isawi H, El-Sayed MH, Eissa M, et al (2016) Integrated geochemistry, isotopes, and geostatistical techniques to investigate groundwater sources and salinization origin in the Sharm EL-Shiekh Area, South Sinia, Egypt. Water, Air, Soil Pollut 227:151

Jafari R, Torabian A, Ghorbani MA, et al (2019) Prediction of groundwater quality parameter in the Tabriz plain, Iran using soft computing methods. J Water Supply Res Technol 68:573–584

Jahnke C, Wannous M, Troeger U, et al (2019) Impact of seawater intrusion and disposal of desalinization brines on groundwater quality in El Gouna, Egypt, Red Sea Area. Process analyses by means of chemical and isotopic signatures. Appl Geochemistry 100:64–76

Jayathunga K, Diyabalanage S, Frank AH, et al (2020) Influences of seawater intrusion and anthropogenic activities on shallow coastal aquifers in Sri Lanka: evidence from hydrogeochemical and stable isotope data. Environ Sci Pollut Res 1–13

Juntakut P, Snow DD, Haacker EMK, Ray C (2019) The long term effect of agricultural, vadose zone and climatic factors on nitrate contamination in Nebraska's groundwater system. J

Contam Hydrol 220:33–48

Kazakis N, Mattas C, Pavlou A, et al (2017) Multivariate statistical analysis for the assessment of groundwater quality under different hydrogeological regimes. Environ Earth Sci 76:349

Klassen J, Allen DM (2017) Assessing the risk of saltwater intrusion in coastal aquifers. J Hydrol 551:730–745

Knoll L, Breuer L, Bach M (2019) Large scale prediction of groundwater nitrate concentrations from spatial data using machine learning. Sci Total Environ 668:1317–1327

Lal A, Datta B (2018) Genetic Programming and Gaussian Process Regression Models for Groundwater Salinity Prediction: Machine Learning for Sustainable Water Resources Management. In: 2018 IEEE Conference on Technologies for Sustainability (SusTech). IEEE, pp 1–7

Lin I-T, Wang C-H, Lin S, Chen Y-G (2011) Groundwater–seawater interactions off the coast of southern Taiwan: evidence from environmental isotopes. J Asian Earth Sci 41:250–262

Mai-Duy N, Tran-Cong T (2003) Approximation of function and its derivatives using radial basis function networks. Appl Math Model 27:197–220

Maurya P, Kumari R, Mukherjee S (2019) Hydrochemistry in integration with stable isotopes (δ18O and δD) to assess seawater intrusion in coastal aquifers of Kachchh district, Gujarat, India. J Geochemical Explor 196:42–56

Mohanty AK, Rao VVSG (2019) Hydrogeochemical, seawater intrusion and oxygen isotope studies on a coastal region in the Puri District of Odisha, India. Catena 172:558–571

Mongelli G, Monni S, Oggiano G, et al (2013) Tracing groundwater salinization processes in coastal aquifers: a hydrogeochemical and isotopic approach in Na-Cl brackish waters of north-western Sardinia, Italy. Hydrol Earth Syst Sci Discuss 10:

Naghibi SA, Pourghasemi HR, Dixon B (2016) GIS-based groundwater potential mapping using boosted regression tree, classification and regression tree, and random forest machine learning models in Iran. Environ Monit Assess 188:44

Nair IS, Rajaveni SP, Schneider M, Elango L (2015) Geochemical and isotopic signatures for the identification of seawater intrusion in an alluvial aquifer. J Earth Syst Sci 124:1281–1291

Noori N, Kalin L, Isik S (2020) Water quality prediction using SWAT-ANN coupled approach. J Hydrol 590:125220

Rao NS, Chaudhary M (2019) Hydrogeochemical processes regulating the spatial distribution of groundwater contamination, using pollution index of groundwater (PIG) and hierarchical cluster analysis (HCA): a case study. Groundw Sustain Dev 9:100238

Rasmussen CE (2003) Gaussian processes in machine learning. In: Summer School on Machine Learning. Springer, pp 63–71

Seddique AA, Masuda H, Anma R, et al (2019) Hydrogeochemical and isotopic signatures for the identification of seawater intrusion in the paleobeach aquifer of Cox's Bazar city and its surrounding area, south-east Bangladesh. Groundw Sustain Dev 9:100215

Shanmuganathan S (2016) Artificial neural network modelling: An introduction. In: Artificial neural network modelling. Springer, pp 1–14

Skansi S (2018) Introduction to Deep Learning: from logical calculus to artificial intelligence. Springer, Cham, Switzerland

Specht DF (1991) A general regression neural network. IEEE Trans neural networks 2:568–576

Umasankar L, Kalaiarasi N (2014) Internal fault identification and classification of transformer with the aid of radial basis neural network (RBNN). Arab J Sci Eng 39:4865–4873

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

Vapnik V, Golowich SE, Smola AJ (1997) Support vector method for function approximation, regression estimation and signal processing. In: Advances in neural information processing systems. pp 281–287

Waller DL (2003) Operations management: a supply chain approach. Cengage Learning Business Press

Williams CKI, Rasmussen CE (1996) Gaussian processes for regression. In: Advances in neural information processing systems. pp 514–520

Yang J, Ye M, Tang Z, et al (2020) Using cluster analysis for understanding spatial and temporal patterns and controlling factors of groundwater geochemistry in a regional aquifer. J Hydrol 583:124594

Zektser IS, Everett LG (2004) Groundwater resources of the world and their use

Zhao T (2008) RBFN-based decentralized adaptive control of a class of large-scale non-affine nonlinear systems. Neural Comput Appl 17:357–364

Zhao Y, Li Y, Zhang L, Wang Q (2016) Groundwater level prediction of landslide based on classification and regression tree. Geod Geodyn 7:348–355

Figures

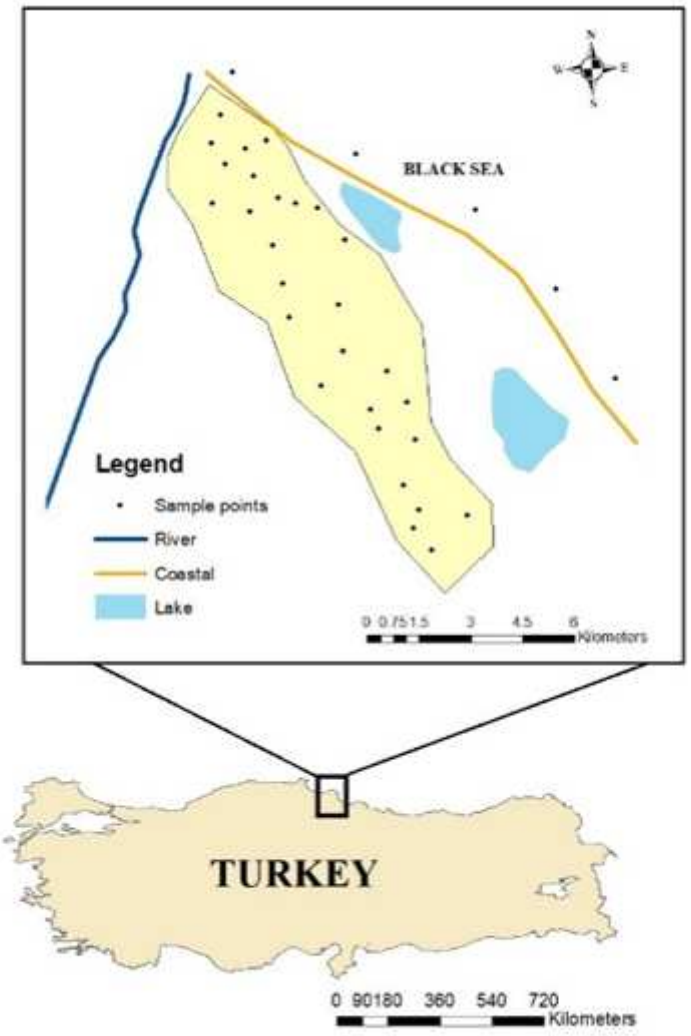

Figure 1

Study area and groundwater sampling sites Note: The designations employed and the presentation of the material on this map do not imply the expression of any opinion whatsoever on the part of Research Square concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. This map has been provided by the authors.

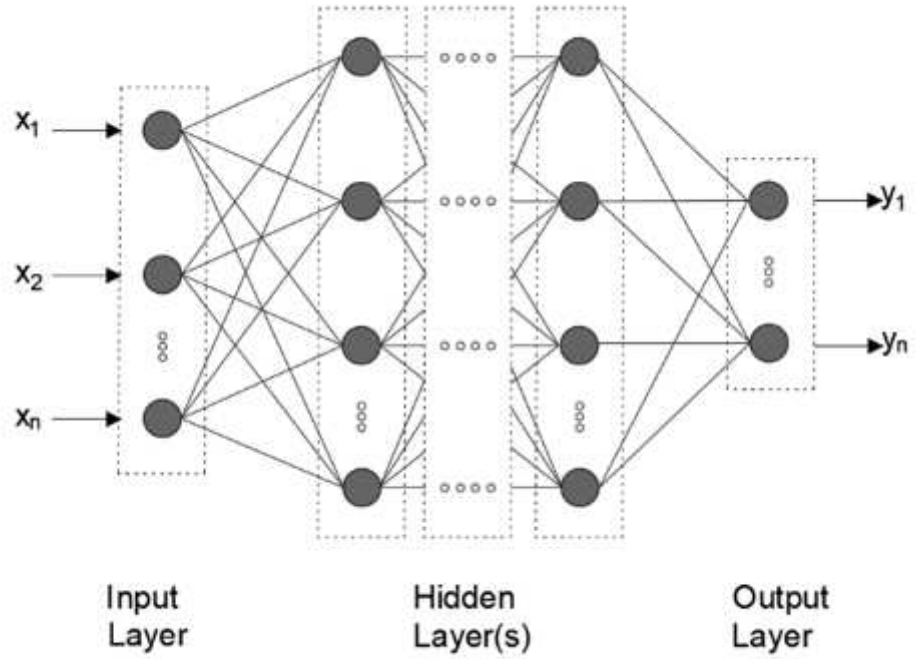

Figure 2

Schematic diagram of MLP architecture

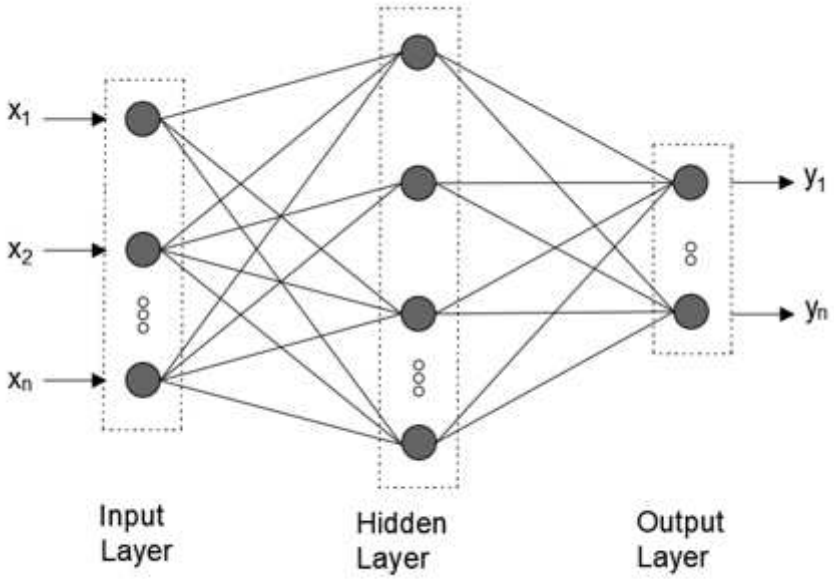

Figure 3

Schematic diagram of RBNN architecture

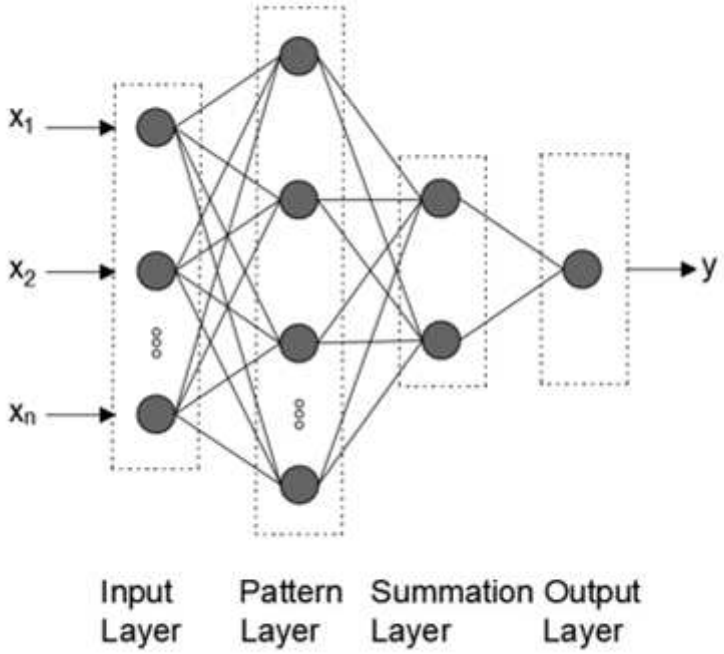

Figure 4

Schematic diagram of GRNN architecture

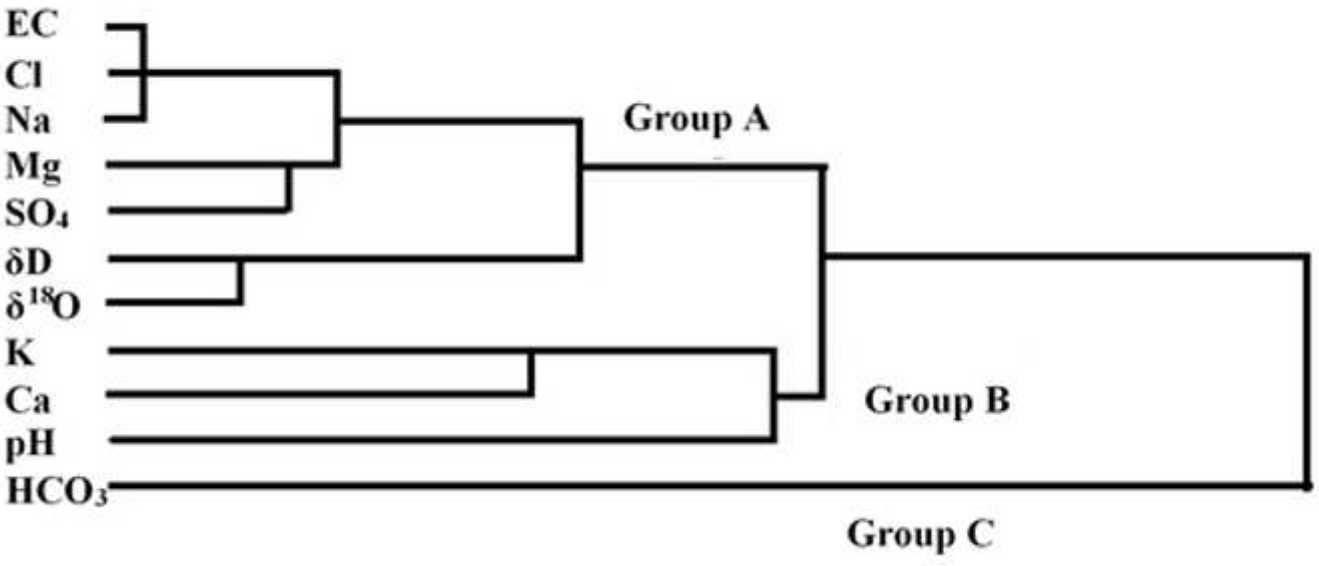

Figure 5

Dendrogram showing the clustering of some parameters of groundwater in Bafra plain

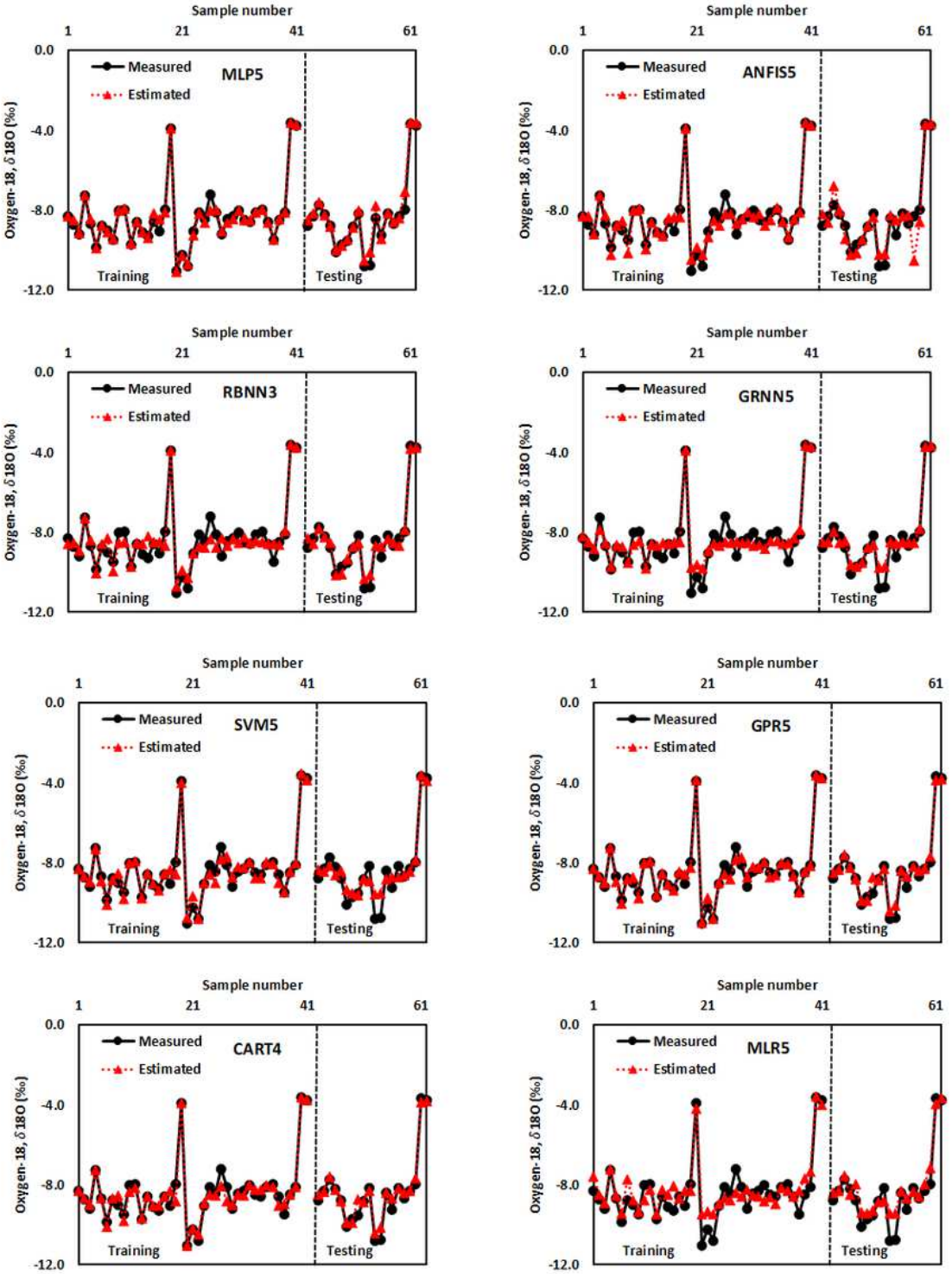

Figure 6

The measured and estimated δ18O values by the optimal models for MLP, ANFIS, RBNN, GRNN, SVM, GPR, CART, and MLR

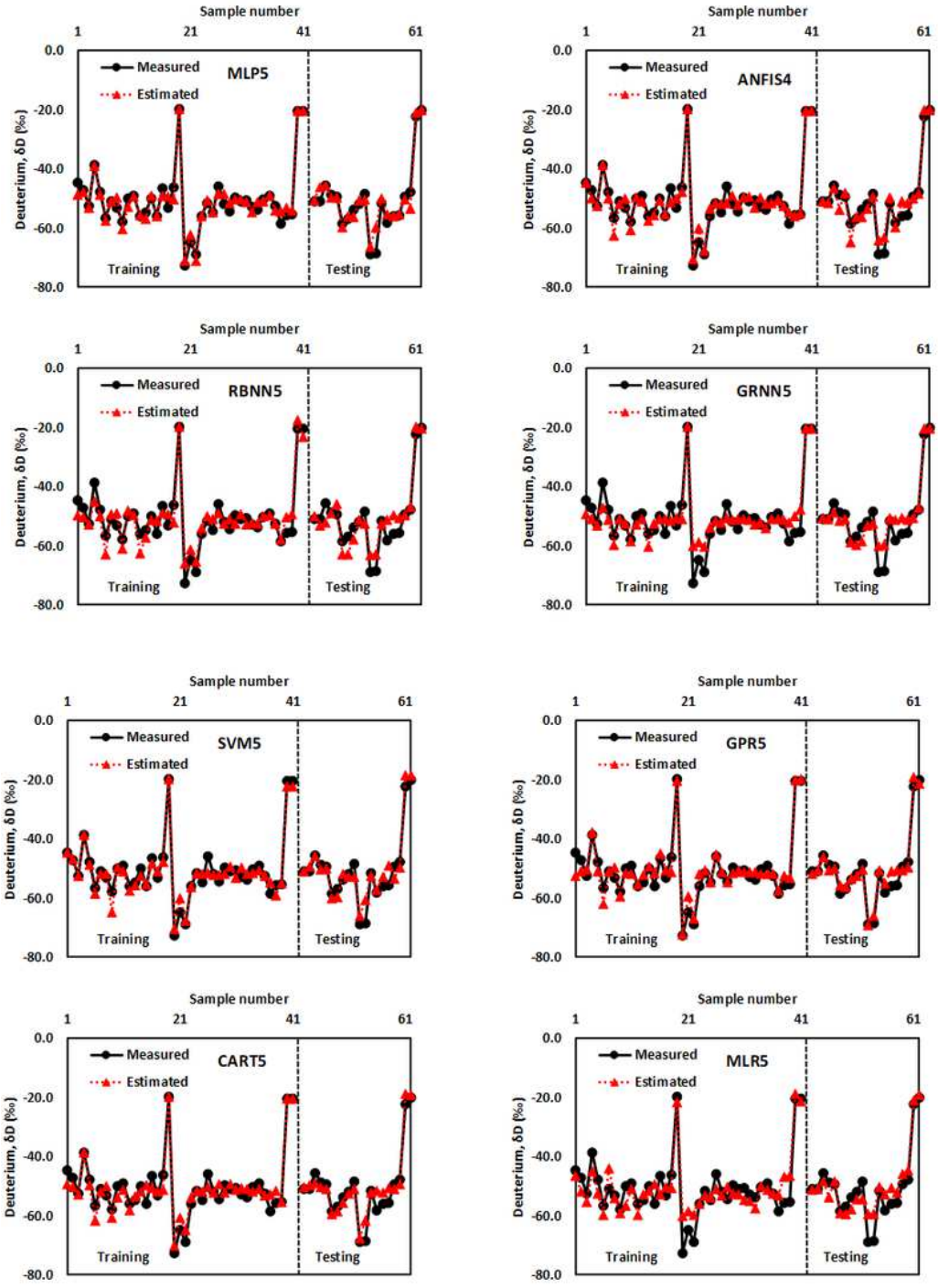

Figure 7

The measured and estimated δD values by the optimal models for MLP, ANFIS, RBNN, GRNN, SVM, GPR, CART, and MLR

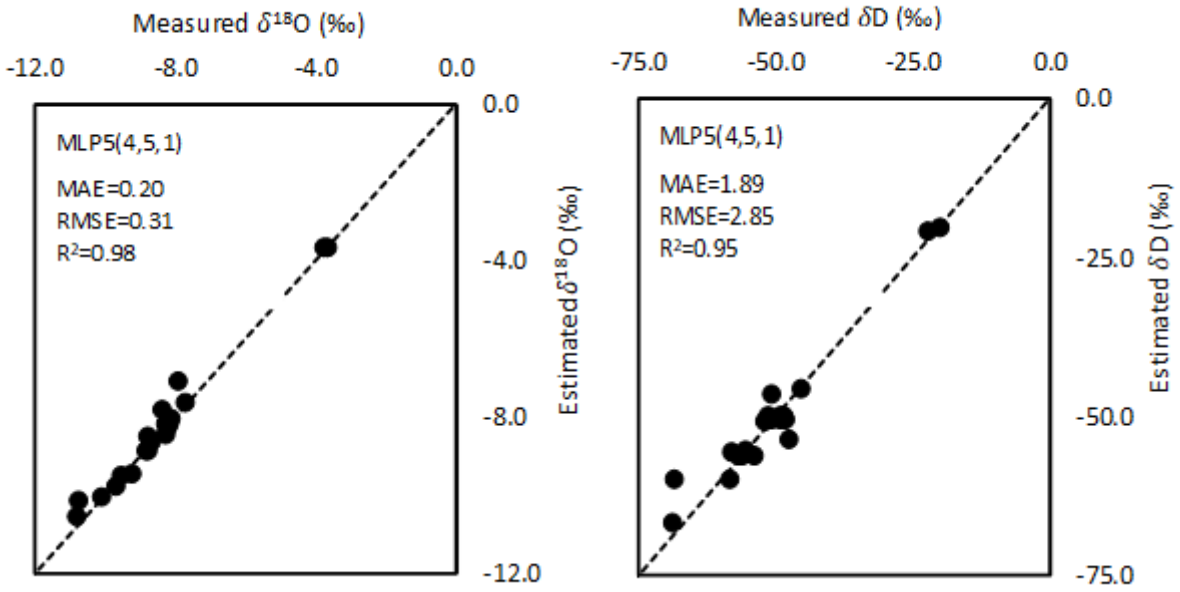

Figure 8

Measured and estimated δ18O and δD values by MLP5(4,5,1) models

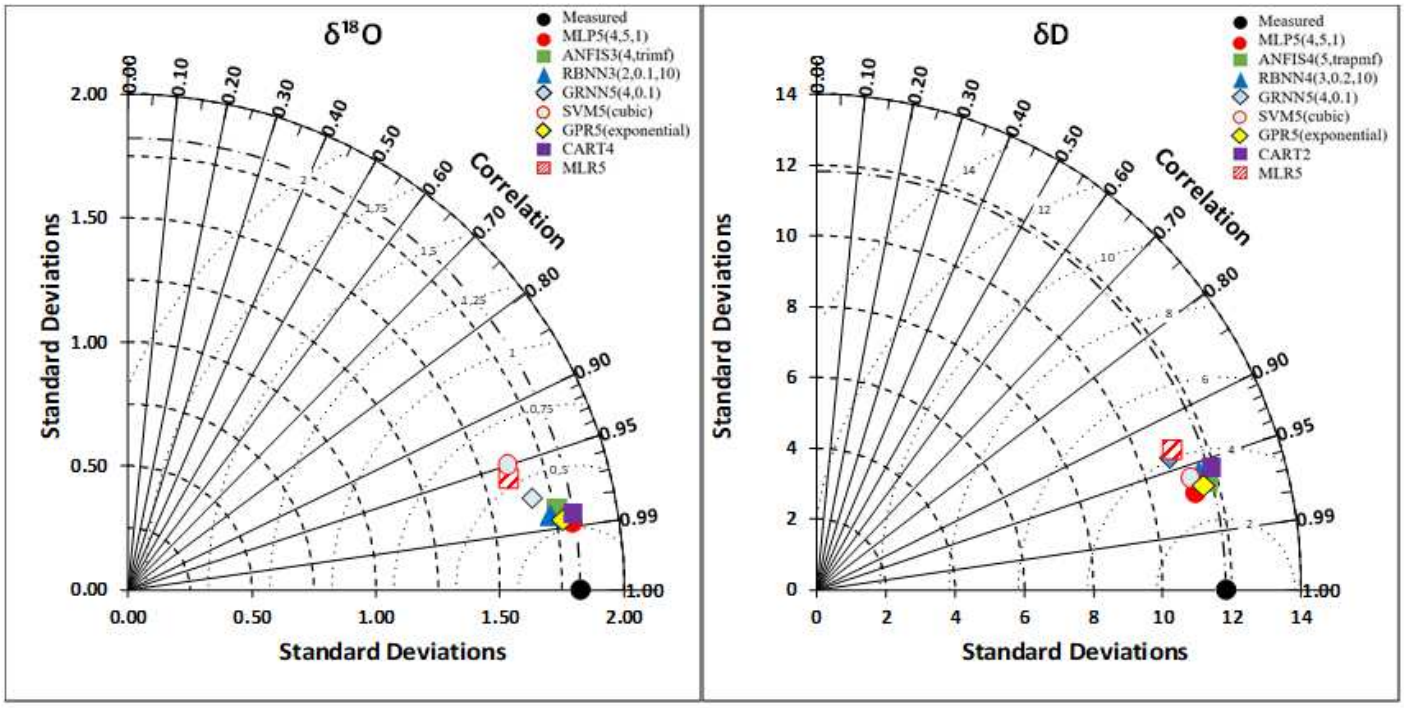

Figure 9

Taylor diagrams for evaluating the δ18O and δD estimation models