Prediction of groundwater quality parameter in the Tabriz plain, Iran using soft computing methods

Robabeh Jafari, Ali Torabian, Mohammad Ali Ghorbani, Seyed Ahmad Mirbagheri and Amir Hessam Hassani

ABSTRACT

Aquifers are one of the largest available freshwater resources. In this paper, total dissolved solids (TDS) of the groundwater aquifer in Tabriz plain is estimated by groundwater physicochemical parameters including Na, HCO_3, Ca, Mg, and SO_4 in the eastern region of Urmia Lake. For this purpose, four soft computing approaches, namely, multilayer perceptron (MLP), adaptive neuro-fuzzy inference system (ANFIS), support vector machine (SVM), and gene expression programming (GEP) were used to predict TDS for a period of 10 years (2002–2012). Data were collected from the East Azerbaijan Regional Water Organization, which totaled 1,742 samples. In the application, of the whole data set, 70% (1,220 samples) was used for training and 30% (522 samples) for testing. In the following, the correlation coefficient (*R*), root mean square error (*RMSE*), and mean absolute error (*MAE*) statistics were used for evaluating the accuracy of the models. According to the results, MLP, ANFIS, SVM, and GEP models could be employed successfully in estimating TDS alterations. A comparison was made between these soft computing approaches that corroborated the superiority of the GEP model over MLP, SVM, and ANFIS models with $RMSE = 58.93$, $R = 0.998$, and $MAE = 5.21$.

Key words | adaptive neuro-fuzzy inference system, gene expression programming, multilayer perceptron, support vector machine, total dissolved solids

Robabeh Jafari
Amir Hessam Hassani
Department of Environmental Engineering,
Islamic Azad University,
Science and Research Branch,
Tehran,
Iran

Ali Torabian (corresponding author)
Department of Civil Environmental Engineering,
University of Tehran,
Iran
E-mail: *atorabi@ut.ac.ir*

Mohammad Ali Ghorbani
Department of Water Engineering,
Tabriz University,
Tabriz,
Iran
and
Engineering Faculty, Near East University,
99138 Nicosia, North Cyprus, Mersin 10,
Turkey

Seyed Ahmad Mirbagheri
Department of Civil Environmental Engineering,
Khajeh Nasir Toosi University of Technology,
Iran

INTRODUCTION

Groundwater is one of the valuable resources of any country. These resources have been taken into account because of the lower pollution potential and the high storage capacity of surface water (Tabarmayeh & Vaezi Hir 2015; Qasemi *et al.* 2019). Globally 40% of the extracted groundwater is consumed for agricultural purposes (Qasemi *et al.* 2018a, 2018b; Wagh *et al.* 2018). In other words, groundwater is one of the main sources of agricultural water supply.

However, in Iran, due to the dry and semi-arid climate conditions, especially in the study area (Tabriz plain located in the eastern part of Lake Urmia), the importance of groundwater conditions becomes more intense. In Tabriz plain, the water needed for agricultural land is provided in addition to surface water resources from groundwater sources (Razaghmanesh *et al.* 2006). In general, changes in the patterns of land use, climate, urbanization, and population growth and the extensive use of chemical fertilizers and pesticides that penetrate underground in many areas are serious threats to the quality of groundwater, both quantitatively and qualitatively. The groundwater quality is getting worse gradually (Wagh *et al.* 2016, 2018).

The study of the GFZ German Research Centre for Geosciences, Potsdam (GFZ) states that the surface water

doi: 10.2166/aqua.2019.062

table in Iran over the past 15 years has dropped by an average of 1.5 feet (0.4572 m) a year. Groundwater quality is as important as its quantity, and is essential for its usability in various uses. Water quality analysis is an important part of groundwater studies, as the quality of these resources, such as surface water, is constantly changing. However, these changes are much slower than those of surface water (Mahdavi 1995). Modeling groundwater aquifers and consequently predicting the quality of water resources is important in terms of hydrological studies and agricultural management in order to achieve high quality groundwater (Todd & Mays 2005). Hence, the use of intelligent models for the evaluation of groundwater quality parameters has become widespread today and there are many studies in this field.

Nourani *et al.* (2008) studied temporal and spatial variations of groundwater level in Tabriz plain using the neural network. They used six types of architectures and algorithms to determine the best structure of the neural network to predict the groundwater level. The results of their study showed that the feedback artificial neural network along with the Levenberg–Marquardt (LM) algorithm provide the best forecasts among a variety of other networks.

Asadollahfardi *et al.* (2011) employed two types of artificial neural networks including recurrent neural network (RNN) and multilayer perceptron (MLP) to predict the total dissolved solids (TDS) in Talkheh Rood. The results showed that the performance of the RNN method was better than MLP in predicting TDS.

Talebizadeh & Moridnejad (2011) used adaptive neuro-fuzzy inference system (ANFIS) and artificial neural network (ANN) models to predict the level of Urmia Lake. The results showed that compared to ANN, the accuracy of the ANFIS model was higher in this regard. Karimi *et al.* (2012) used gene expression programming (GEP) and ANFIS models to estimate the water level of Urmia Lake. They also concluded that GEP was more efficient compared to the ANFIS model. Zaman Zad Ghvidel & Mntaseri (2014) evaluated the application of different data-based methods in predicting the TDS in Zarinehroud basin. They concluded that GEP, ANFIS with grid partition (ANFIS-GP), ANFIS with sub clustering (ANFIS-SC), and ANN models could be employed to predict the changes in TDS effectively. In addition, among the various methods of artificial intelligence, GEP was the superior method in the prediction of this parameter compared to the other models.

In a study by Kadam *et al.* (2019) in the Shivganga River in India, ANN and multiple linear regression (MLR) models were used to predict groundwater quality for drinking purposes. In this research, an MLR model was used to evaluate the ANN performance prediction and the results showed that ANN performance was satisfactory.

Taking the improved performances of different modeling techniques into account, this study compared the performance of four soft computing techniques, MLP, ANFIS, support vector machine (SVM), and GEP, for TDS prediction using physicochemical parameters of groundwater collected from the East Azarbaijan Regional Water Organization. It was compared because TDS is an important quality criterion for agricultural water.

MATERIALS AND METHODS

Groundwater quality data of the Tabriz plain were obtained from 1,742 samples studied during 2002–2012. The data were collected from Tabriz Regional Water Authority in East Azarbaijan Province. Data were collected twice a year for 10 years, in different piezometers in Tabriz plain. Generally, of the whole data set, 70% (1,220 samples) was used for training and 30% (522 samples) for testing. Excel software was used to prepare the statistical data. Then, four models including MLP, ANFIS, SVM, and GEP were used for modeling TDS. In this regard, the inputs of the model were selected in terms of the highest correlation with TDS (Na, HCO_3, Ca, Mg, and SO_4), respectively. The parameters *root mean square error (RMSE)*, *mean absolute error (MAE)*, and R were used to evaluate and select the best model.

Multilayer perceptron

ANN is a model based on the human brain, the idea for which was first proposed in 1940 in a School of Psychiatry. One of the most widely used types of ANN is MLP which is also employed in this study (Kouvhskzadeh & Bahmani 2006). Overall, the MLP neural network structure consists of three important layers, namely, input, hidden, and output layers and in each layer a number of neurons are considered in the network architecture. This model has been

successfully used in several studies (Deswal & Pal 2008; Singh *et al.* 2010; Khatibi *et al.* 2011; 2013; Ghorbani *et al.* 2013). In the MLP method, the number of neurons in the input and output layers is determined depending on the number of input and output variables of the system under study, respectively. In order to obtain the parameters of the MLP (number of hidden nodes, the learning rate, and the momentum value), the data set is divided into two parts: training and testing sets. Logistic sigmoid transfer function was applied in the hidden layer. One hundred trials of random initial weights were considered to prevent the local minima problem affecting the MLP method. The network was trained in 1,000 epochs using the LM algorithm with a learning rate of 0.001 and a momentum value of 0.7. Detailed theoretical information about MLP neural networks can be found in Haykin (1998). Figure 1 shows the structure of the multi-layer perceptron with five inputs. MLP models used in this study include an input layer with variable neurons including five input compounds (Na, HCO_3, Ca, Mg, and SO_4) and one hidden layer (with two neurons) and one output layer (TDS).

ANFIS model

Fuzzy system and fuzzy logic have an important role in modeling methods. Fuzzy system is a system based on logic 'If–Then' rules, that projects the space of input variables on the space of output variables by means of the concept of linguistic variables and fuzzy decision-making process. The combination of fuzzy systems which are based on logical rules with the artificial neural networks which have the ability to extract knowledge from numerical information, have led to the presentation of the neuro-fuzzy inference system (Zadeh 1965; Mohammadi *et al.* 2018; Takdastan *et al.* 2018). ANFIS has a five-layer structure of the forwarding neural network and these five layers are: layer 1 (fuzzification layer), layer 2 (rule layer), layer 3 (normalization layer), layer 4 (defuzzification layer), and layer 5 (a single summation neuron) or (output) layer, as shown in Figure 2. The training process in ANFIS is the propagation of error and in order to reach a rapid convergence the combined approach can be used, which is a combination of error propagation with the least squares method (Jang *et al.* 1997). The ANFIS network used in this study has a membership function of 'constant' in the output layer and a membership function of 'gaussmf' in the input layer. The membership function of 'constant' is constant for each rule. The ANFIS structure is shown in Figure 2. In this figure, (x) and (y) are inputs of the model, (A) and (B) are fuzzy sets, and (f) is the output of the model. In the first layer, all nodes are comparative and the output of layer 1 is the membership grade of the inputs. The output of layer 2 is the input signal coefficient, which is actually equivalent to the (if) rule. The output of layer 3 is normalized to the previous layer. In layer 4, the contribution of each rule is calculated for the final input and determination of the

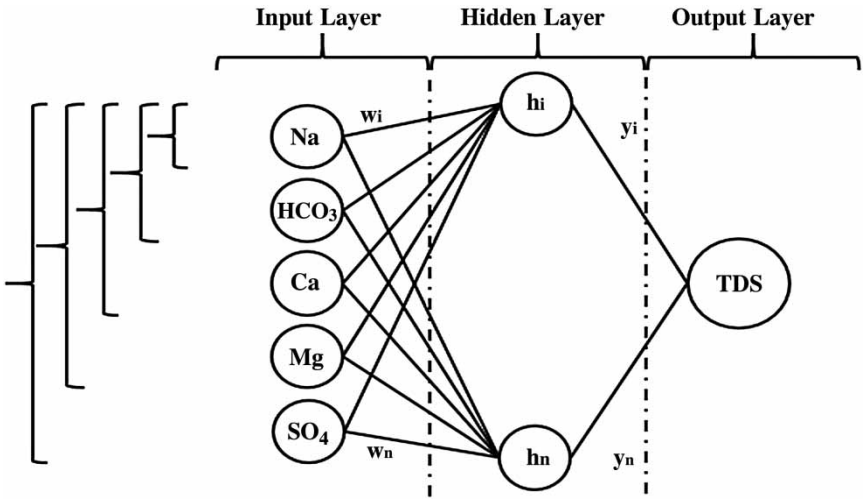

Figure 1 | Structure of multilayer perceptron with five inputs.

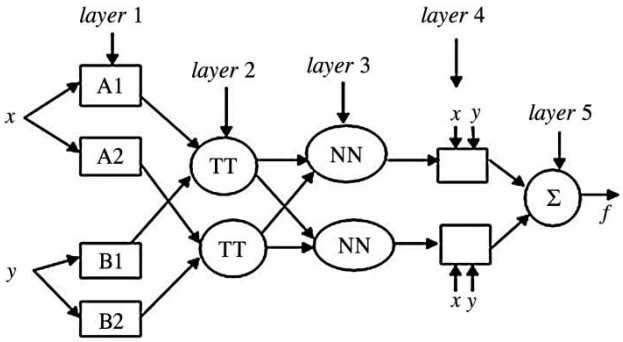

Figure 2 | The five layers of the ANFIS structure.

final function. Finally, the output of layer 5 is the total output of the system.

Support vector machine

Machine learning methods for the development of classification models are widely applied in numerous fields. The SVM was introduced as a regression tool for the first time by Vapnik (2013). The SVM method is a set of supervised learning algorithms used for classification and regression challenges (Cortes & Vapnik 1995). SVMs are kernel-based algorithms that transform data into a multi-dimensional space and form a hyperplane that enlarges the interval to the nearest data point of any of the input categories (Vapnik *et al.* 1997). For further information about the SVM method readers are referred to Gunn (1998). The SVM structure is shown in Figure 3. This figure contains input vectors, which are represented by x^1 to x^n, the x_1 to x_N parameters are represented as support vectors in the figure, k is the kernel function and the number of support vectors. Finally, the weight is applied to the output of the kernel function and the sum is displayed as output (y).

Gene expression programming

GEP, based on the principles of genetics, is a symbolic regression technique that automatically solves problems without pre-specified structure of the solution in advance (Babovic 2005). Five major preparatory steps for GEP are: (1) determining the set of terminals, (2) determining the set of functions, (3) determining the fitness measure, (4) determining the parameters for the run, and (5) determining the method for designating a result and the criterion for terminating a run (Koza 1994). The first step in the GEP algorithm is to generate the initial population of solutions. The chromosomes are then evaluated as a tree diagram (ET) with a fitness function to determine the suitability of a solution in the problem domain; if the satisfactory quality of a solution is found, evolution is stopped and the best solution is reported. On the other hand, if the conditions are not stopped, the best solution is kept from the current generation (this means elitist selection) and the rest of the solutions are left to the selective process (Ferreira 2001). For more information about the GEP method, readers are referred to Ghorbani *et al.* (2012).

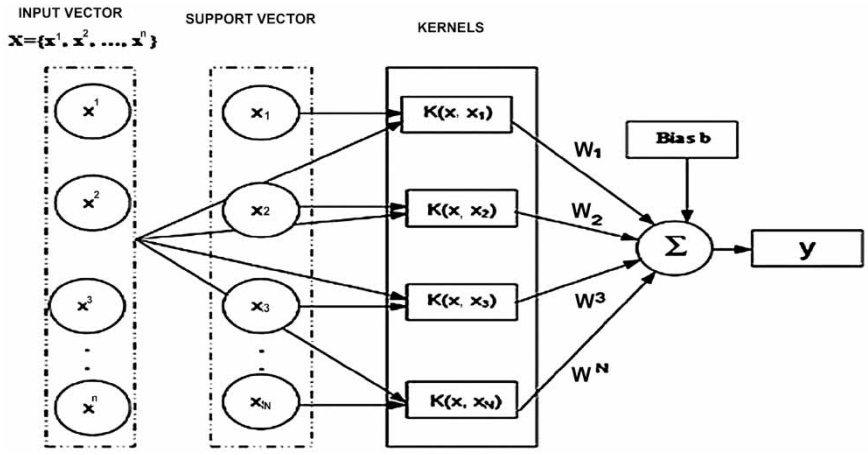

Figure 3 | Architecture of support vector machine method.

Study area and data used

The Tabriz plain is located at the eastern part of Urmia Lake in the north-west of Iran. This plain lies between latitude 37° 53′ to 38° 12′ N and longitude 45° 55′ to 46° 45′ E. The elevation of this plain ranges between 1,250 and 3,600 m above sea level. The climate of the area is semi-arid and average annual rainfall is about 230 mm. Figure 4 shows the Tabriz plain located in the eastern part of Lake Urmia in Iran. Figure 4 shows groundwater quality data of observation wells (piezometers) in the aquifer collected by the East Azarbayjan Regional Water Authority of Iran from 2002 to 2012, which were used in this research.

Table 1 provides the basic statistics of the water quality data used in this study. In this table, X_{Min}, X_{Max}, X_{Mean}, S_d, C_v, and C_{SX}, denote the overall minimum, maximum, mean, standard deviation, coefficient of variation, and skewness values, respectively. This information is provided for the two parts of training and testing.

Evaluation criteria for model performance

In this study, correlation coefficient (R), *RMSE*, and *MAE* were used to assess the performance of different models and defined as:

$$R = \left[\frac{\sum_{i=1}^{n}(TDS_o - \overline{TDS_o})(TDS_p - \overline{TDS_p})}{\sqrt{\sum_{i=1}^{n}(TDS_o - \overline{TDS_o})^2 \sum_{i=1}^{n}(TDS_p - \overline{TDS_p})^2}}\right] \quad (1)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(TDS_o - TDS_p)^2} \quad (2)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|TDS_o - TDS_p| \quad (3)$$

where TDS_o and TDS_p refer to observed and predicted TDS, $\overline{TDS_o}$ and $\overline{TDS_p}$ are the mean of the observed and predicted TDS, respectively, and n is the number of data sets.

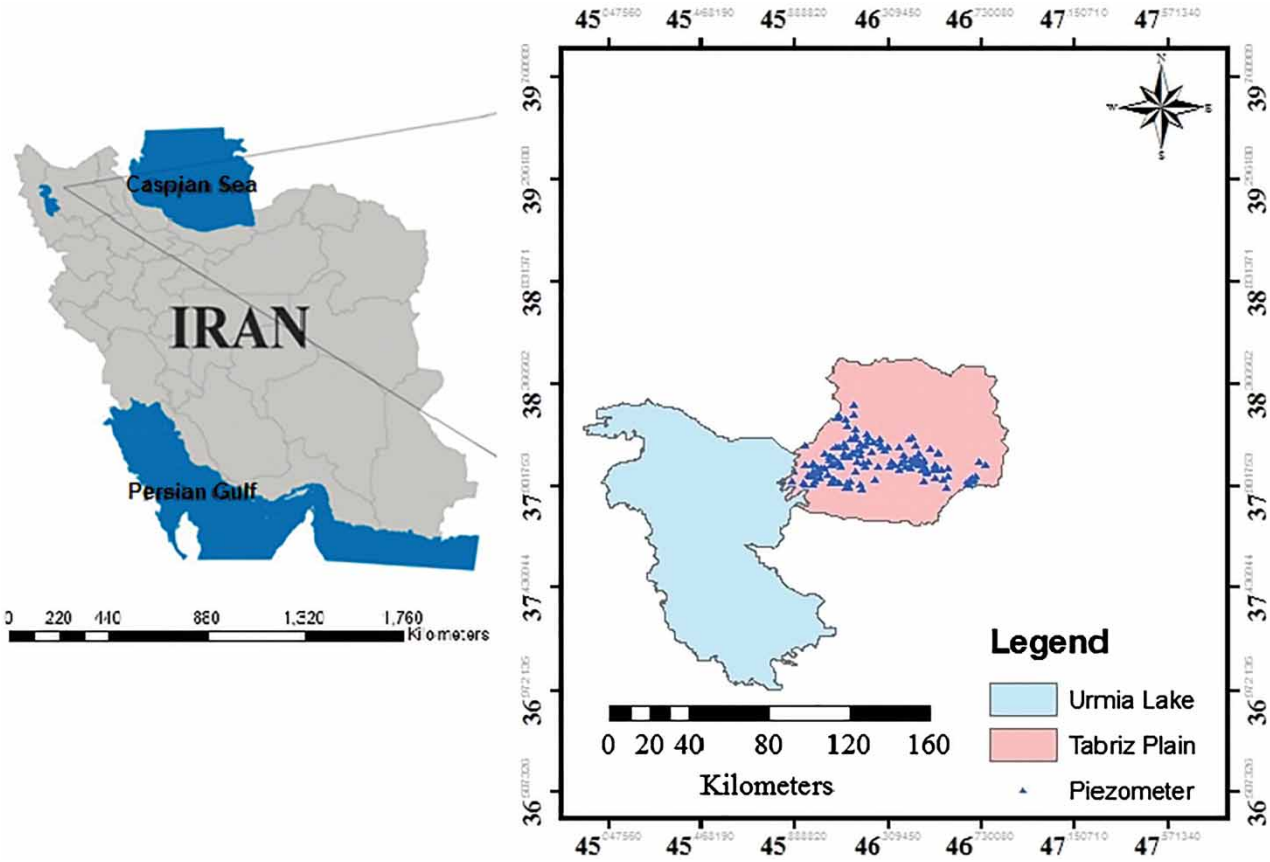

Figure 4 | Study area and piezometer distribution.

Table 1 | Statistical parameters of water quality variables over the Tabriz plain

		Statistical parameter					
	Variables	X_{Min}	X_{Max}	X_{Mean}	S_d	C_V	C_{SX}
Train	Na	0.38	112.35	13.22	16.55	1.25	2.18
	HCO_3	0.4	22.5	4.26	2.29	0.54	1.78
	Ca	0.4	46.4	6.85	7.61	1.11	2.19
	Mg	0.1	22.2	5.23	4.81	0.92	1.2
	SO_4	0.1	26	4.82	4.42	0.92	1.42
	TDS	130.2	8,775	1,646.72	1,645.33	0.99	1.48
Test	Na	0.4	91	7.39	13.29	1.8	3.28
	HCO_3	0.7	21.7	4.17	2.35	0.56	1.78
	Ca	0.64	50	4.57	5.77	1.26	4.27
	Mg	0.1	32	3.67	4.28	1.17	2.4
	SO_4	0	21	2.92	3.3	1.13	1.8
	TDS	96.24	9,763	1,013.86	1,382.95	1.36	2.79

The more the correlation coefficient (R) is closer to one, the more consistency there is between the estimated values and observational values. The *RMSE* value represents the root mean square of the error between the measured and estimated values, and the lower the value, the greater the accuracy of the model estimation. The *MAE* value represents an absolute computation error, which is close to zero, indicating the high accuracy of the model.

RESULTS AND DISCUSSION

MLP model

MLP models used in this study include an input layer with the neurons changing from 1 to 5 (five input combination) and a hidden layer with two neurons and an output layer. The performance of the MLP models with training and testing data sets are provided in Table 2. The input composition of each scenario was obtained using the correlation between the physicochemical parameters and the TDS. Accordingly Na, HCO_3, Ca, Mg, and SO_4 parameters had the highest correlation with TDS, respectively, which were used in the scenarios. The model is represented by one input as MLP1 and the model is represented by two inputs as MLP2. Finally, the model with five inputs is shown as MLP5.

According to the table, with the increasing number of inputs in the model, the results will be better (according to the evaluation criteria). This result is consistent with the results of Khashei *et al.*, who stated that the MLP model increases with increasing the number of input parameters, thus the accuracy of the model increases (Khashei *et al.* 2011; Khashei & Sarbazi 2015) so that the fourth and fifth scenarios are very close together. Therefore, the *MAE* value is better than the fifth scenario in both

Table 2 | Training and testing statistics of MLP models

		Training			Testing		
	Input combination	*R*	*RMSE*	*MAE*	*R*	*RMSE*	*MAE*
MLP1	Na	0.931	666.86	220.68	0.949	323.49	29.06
MLP2	Na, HCO_3	0.938	792.99	488.35	0.955	393.63	151.41
MLP3	Na, HCO_3, Ca	0.995	172.01	43.34	0.994	107.41	5.74
MLP4	Na, HCO_3, Ca, Mg	0.998	116.47	37.97	0.997	83.81	6.52
MLP5	**Na, HCO_3, Ca, Mg, SO_4**	**0.999**	**120.65**	**57.93**	**0.998**	**67.67**	**17.34**

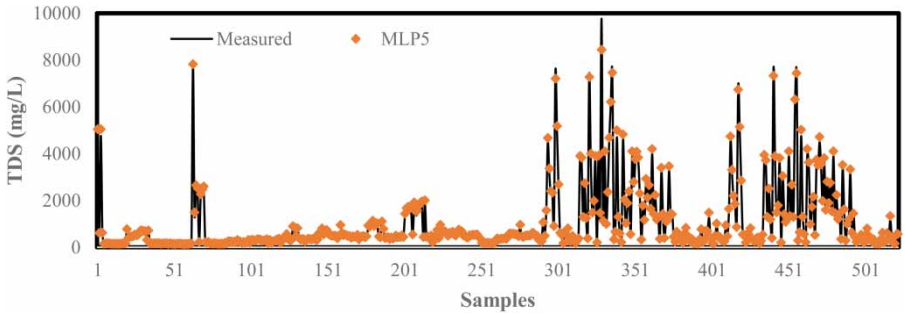

Figure 5 | Observed and estimated TDS data of the MLP model during the test period.

the test and training phases in the fourth scenario. However, considering the high R (in the test and training phase) and the lower $RMSE$ (in the testing phase) in the fifth scenario, the MLP model was accepted as the best model in the fourth scenario ($R = 0.998$, $RMSE = 67.67$, and $MAE = 17.34$). According to Figures 5 and 6, we see that there is a very good overlap between the data obtained from the model and observational data.

In Figure 6, $R = 0.9985$ indicates that these conditions are optimal. Thus, it can be stated that the fifth scenario (MLP5) predicts the high accuracy of TDS data.

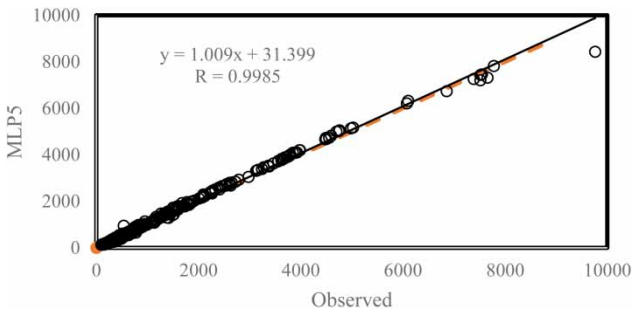

Figure 6 | Distribution of observed and estimated TDS data of the MLP model during the test period.

ANFIS model

The ANFIS method is implemented using the Fuzzy Logic Toolbox (MATLAB) for all short-term flow forecasting in this study. The number of membership functions are the same and equal to 4 for all inputs and membership function types are triangular-shaped, respectively. For performance of the ANFIS models, 70% of the data were for training and 30% for testing, as provided in Table 3. This table indicates that according to the selected evaluation criteria in this study, the results of the models with three inputs, four inputs, and five inputs are very similar. According to table information, the results of the model will be improved as the number of inputs increases. Therefore, the results for the ANFIS model with five inputs are better than other inputs; also the ANFIS model with five inputs, like the MLP5 model, has the best performance. According to the table, and at best, the values of the parameters are $R^{=}0.997$, $RMSE = 108.63$, and $MAE = 27.77$. The output of the data obtained with the model and observational data are shown in Figure 7. As shown in the figure, there is an overlap between estimated data and observational data. In Figure 8, the value of this overlap is shown as a correlation

Table 3 | Training and testing statistics of ANFIS models

		Training			Testing		
	Input combination	***R***	***RMSE***	***MAE***	***R***	***RMSE***	***MAE***
ANFIS1	Na	0.930	771.64	363.29	0.949	374.82	69.54
ANFIS2	Na, HCO_3	0.938	915.27	604.18	0.954	452.94	179.65
ANFIS3	Na, HCO_3, Ca	0.994	313.83	203.68	0.994	179.77	48.50
ANFIS4	Na, HCO_3, Ca, Mg	0.998	299.84	198.85	0.997	162.06	37.47
ANFIS5	**Na, HCO_3, Ca, Mg, SO_4**	**0.998**	**183.76**	**112.51**	**0.997**	**108.63**	**27.77**

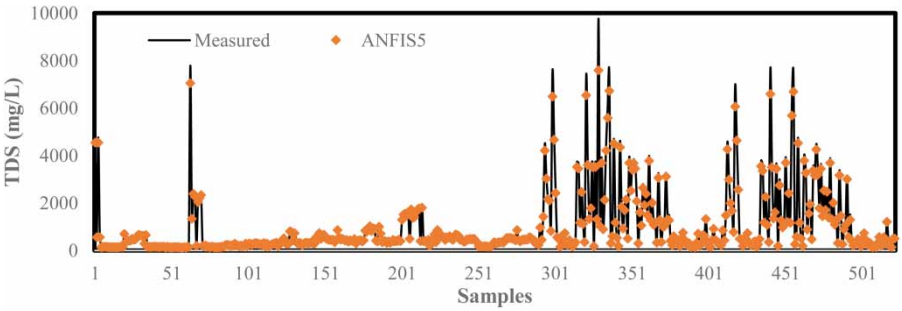

Figure 7 | Observed and estimated TDS data of the ANFIS model during the test period.

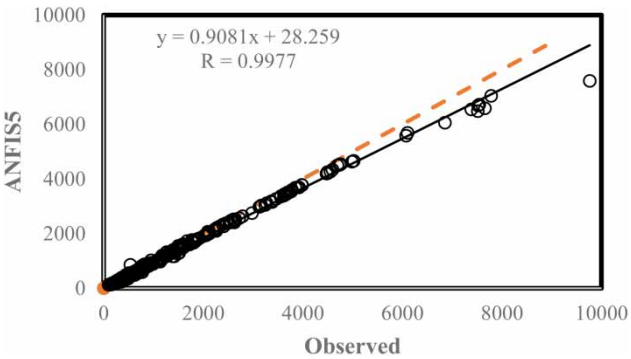

Figure 8 | Distribution of observed and estimated TDS data of the ANFIS model during the test period.

coefficient ($R = 0.9977$), which expresses the high accuracy of the model ANFIS5 for predicting computational TDS.

SVM model

Here, the SVM method was used for predicting the TDS of groundwater dominated by Urmia Lake salt water. The main benefit is to avoid attributes in greater numeric ranges dominating those in smaller numeric ranges. Another benefit is to avoid numerical difficulties during the computing. For the implementation of the SVM models, as with the previous models, 70% of the data were selected for training and 30% testing. The preparation of the model with different inputs (1 to 5) is shown in Table 4. This table indicates that the model results with two inputs (SVM2), four inputs (SVM4), and five inputs (SVM5) are very close together. However, in the end, according to the best assessment criteria in the fifth scenario, test value ($R = 0.953$, $RMSE = 274.56$, and $MAE = 13.83$) and training value ($R = 0.958$, $RMSE = 471.78$, and $MAE = 7.25$), the top model, SVM5 was selected. The results of the best model, SVM with five inputs (SVM5), are shown in Figures 9 and 10, illustrating that the predicted data are obtained with higher accuracy compared to the observational data, and clearly evident from the high ($R = 0.9533$) in Figure 10. Therefore, the SVM model, like the ANFIS and MLP models with five inputs (fifth scenario), has the best performance for predicting TDS.

GEP model

Various GEP models have been developed using input combinations similar to MLP, ANFIS, and SVM models. The performance of the GEP models with training and

Table 4 | Training and testing statistics of SVM models

		Training			Testing		
	Input combination	***R***	***RMSE***	***MAE***	***R***	***RMSE***	***MAE***
SVM1	Na	0.737	1111.8	9.82	0.779	576.82	41.75
SVM2	Na, HCO_3	0.937	598.37	120.7	0.956	288.85	4.45
SVM3	Na, HCO_3, Ca	0.882	776.57	29.65	0.818	542.64	100.33
SVM4	Na, HCO_3, Ca, Mg	0.95	511.45	13.02	0.933	339.24	83.81
SVM5	**Na, HCO_3, Ca, Mg, SO_4**	**0.958**	**471.78**	**7.25**	**0.953**	**274.56**	**13.83**

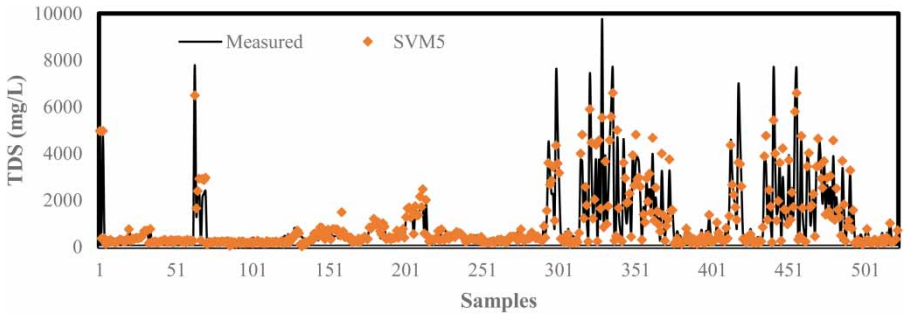

Figure 9 | Observed and estimated TDS data of the SVM model during the test period.

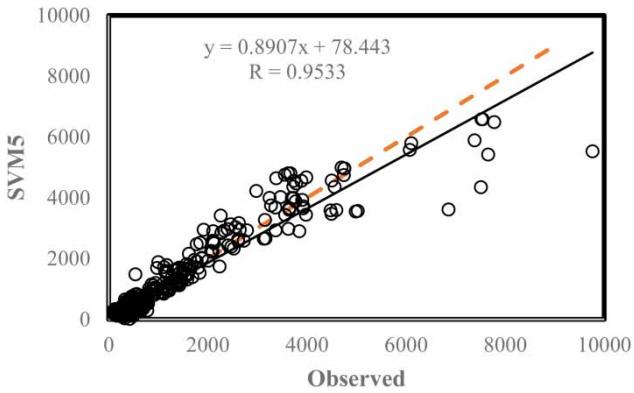

Figure 10 | Distribution of observed and estimated TDS data of the SVM model during the test period.

testing data sets are provided in Table 5. Unlike the MLP, ANFIS, and SVM models, where the fifth scenario was the best option, here, according to the evaluation criteria, the GEP model with four inputs has the best performance. This table shows that the GEP model with four inputs has a better performance than the other input combinations in terms of *R*, *RMSE*, and *MAE* values. Prediction accuracy of the GEP4 model is much better than other GEP models used with this data set in terms of *R* (0.998), *RMSE* (58.93), and *MAE* (5.21) values. Figure 11 shows the observational data and the data obtained from the output of the GEP4 model. It is clear that there is a high correlation between estimated data and observational data in Figure 12, with respect to the value of *R* (0.9992). This high correlation is proven. Therefore, according to the results obtained from the models, and considering the numbers and variables of the evaluation parameters of the models (*R, RMSE, MAE*), it was found that the GEP model was selected as the best model for predicting qualitative parameter (TDS) compared to other models of this study (MLP, ANFIS, SVM). This result is consistent with the studies of Zaman Zad Ghvidel & Mntaseri (2014) and Karimi *et al.* (2012). Given that the GEP model offers a mathematical equation, the equation of the optimal model in MATLAB format is shown in Figure 13. Also, Figure 14 shows the expression tree of the optimal GEP model (GEP4) for TDS estimation. The tree structure helps in each stage that the initial population be expressed in a simple linear structure, and all the changes are made to

Table 5 | Training and testing statistics of GEP models

		Training			Testing		
	Input combination	***R***	***RMSE***	***MAE***	***R***	***RMSE***	***MAE***
GEP1	Na	0.941	557.36	29.93	0.959	265.59	29.06
GEP2	Na, HCO_3	0.938	570.78	0.37	0.958	267.38	25.08
GEP3	Na, HCO_3, Ca	0.995	166.2	10.57	0.997	76.79	4.32
GEP4	**Na, HCO_3, Ca, Mg**	**0.998**	**100.15**	**2.34**	**0.998**	**58.93**	**5.21**
GEP5	Na, HCO_3, Ca, Mg, SO_4	0.993	205.4	53.06	0.995	89.77	4.26

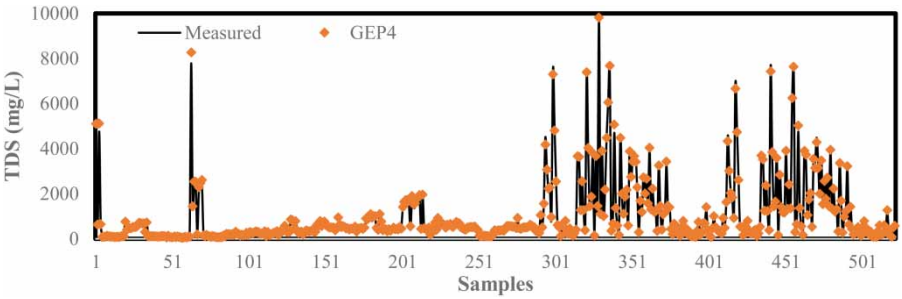

Figure 11 | Observed and estimated TDS data of the GEP model during the test period.

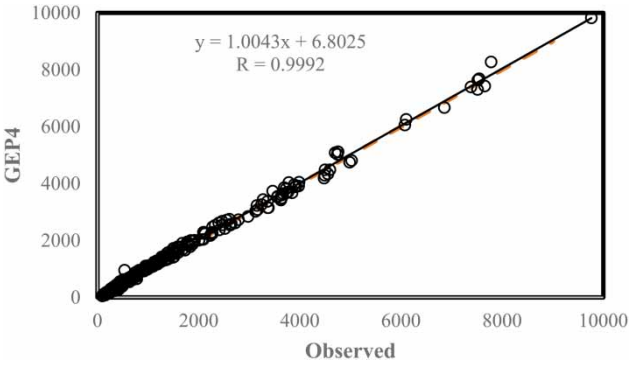

Figure 12 | Distribution of observed and estimated TDS data of the GEP model during the test period.

simple structures. As a result, there is no need for relatively complex structures to be expanded at each stage. In Figures 13 and 14, the values of C are related to the coefficients of gene expression planning and the values of d are related to the input parameter.

```
function result = gepModel(d)

G1C0 = 6.788452;
G1C1 = -5.448517;
G2C0 = 6.788452;
G2C1 = -5.448517;
G3C0 = -6.953949;
G3C1 = 4.18695;

varTemp = 0.0;

varTemp = (((log(d(4))+(d(4)+d(3)))*(G1C0^2))-sqrt(atan(d(1))));
varTemp = varTemp + (sqrt((d(4)+((d(4)^(1.0/3.0))+(G2C0*d(3)))))*d(4));
varTemp = varTemp + ((cos((d(1)+G3C0))+(G3C0^2))*(d(1)+d(3)));

result = varTemp;
```

Figure 13 | Matlab format equation for the GEP4 model.

CONCLUSIONS

Soft computing approaches can provide relatively accurate predictions for water quality parameters; however, these intelligent methods rely on data sets that take a long time in operation. In the current study, the applicability of MLP, ANFIS, SVM, and GEP models in predicting TDS values has been evaluated. Groundwater samples of Tabriz plain in northwest Iran were tested for physicochemical parameters using different models. Regarding the correlation coefficient of physicochemical parameters to TDS, the inputs of the model were selected according to the selected priority (Na, HCO_3, Ca, Mg, and SO_4) and in the scenarios 1 to 5. Finally, they were analyzed by the criteria of the evaluation (R, $RMSE$, MAE). The MLP model for scenarios 4 and 5, according to the $RMSE$, MAE, and R assessment criteria, had roughly similar results: MLP4 for test ($R = 0.997$,

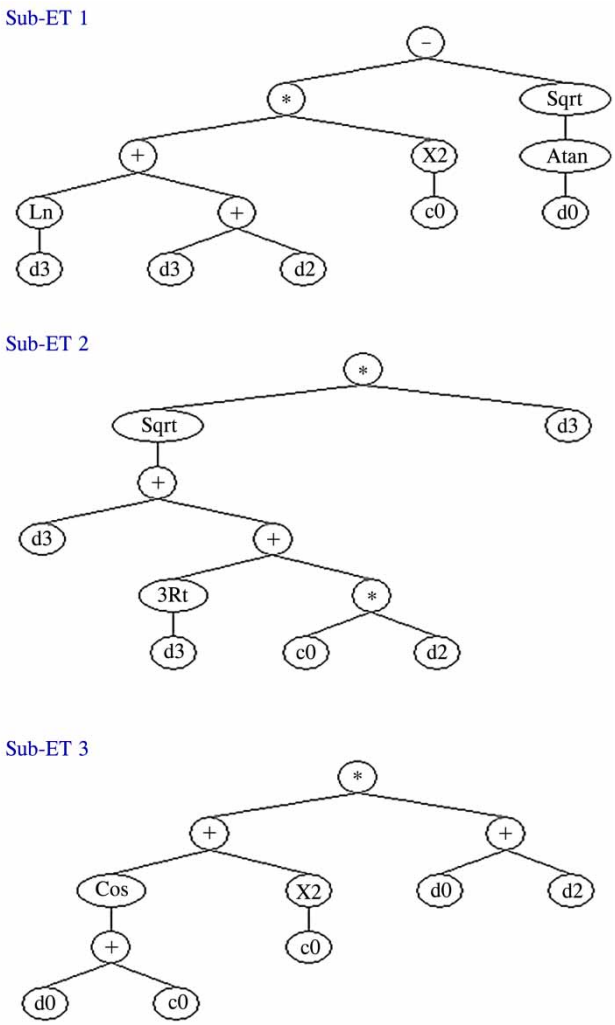

Figure 14 | Expression tree of the GEP4 model for TDS forecasting.

$RMSE = 83.81$, and $MAE = 6.52$) and MLP5 for test ($R = 0.998$, $RMSE = 67.67$, and $MAE = 17.34$). Therefore, considering the greater correlation coefficient (R) and lesser $RMSE$, the fifth scenario was chosen as a better scenario. The MLP, ANFIS, and SVM models had the best performance with five inputs, but the GEP model with four inputs was chosen as the best model. Also, according to the results, the MLP model was better than the ANFIS and SVM models according to ($R = 0.998$, $RMSE = 67.67$, and $MAE = 17.34$) evaluation criteria. On the other hand, the MLP model with five inputs (MLP5) and the GEP model with four inputs (GEP4) had similar results, but in the end, the GEP model was selected as the best model according to better conditions ($R = 0.998$, $RMSE = 58.93$, $MAE = 5.21$). Also, this method can provide mathematical modeling that is very effective for future predictions.

ACKNOWLEDGEMENTS

The authors want to thank the authorities of Islamic Azad University for their comprehensive support for this study. The authors of this article declare that they have no conflict of interests.

REFERENCES

Asadollahfardi, G., Taklify, A. & Ghanbari, A. 2011 Application of artificial neural network to predict TDS in Talkheh Rud River. *Journal of Irrigation and Drainage Engineering* **138**, 363–370.

Babovic, V. 2005 Data mining in hydrology. *Hydrological Processes* **19**, 1511–1515.

Cortes, C. & Vapnik, V. 1995 Support-vector networks. *Machine Learning* **20**, 273–297.

Deswal, S. & Pal, M. 2008 Artificial neural network based modeling of evaporation losses in reservoirs. *International Journal of Mathematical, Physical and Engineering Sciences* **2**, 177–181.

Ferreira, C. 2001 Gene expression programming in problem solving. Invited tutorial of the 6th online world conference on soft computing in industrial applications. *Origins of Functionalist Theory* **9**, 10–24.

Ghorbani, M. A., Khatibi, R., Asadi, H. & Yousefi, P. 2012 Intercomparison of an evolutionary programming model of suspended sediment time-series with other local models. *Computer and Information Science, Artificial Intelligence, Genetic Programming – New Approaches and Successful Applications* (S. Ventura, ed.). IntechOpen, London.

Ghorbani, M., Khatibi, R., Hosseini, B. & Bilgill, M. 2013 Relative importance of parameters affecting wind speed prediction using artificial neural networks. *Theoretical and Applied Climatology* **114**, 107–114.

Gunn, S. R. 1998 Support vector machines for classification and regression. *ISIS Technical Report* **14**, 5–16.

Haykin, S. 1998 *Neural Networks: A Comprehensive Foundation*, 2nd edn. Macmillan, New York.

Jang, J.-S. R., Sun, C.-T. & Mizutani, E. 1997 Neuro-fuzzy and soft computing-a computational approach to learning and machine intelligence. *IEEE Transactions on Automatic Control* **42**, 1482–1484.

Kadam, A. K., Wagh, V. M., Muley, A. A., Umrikar, B. N. & Sankhua, R. N. 2019 Prediction of water quality index using artificial neural network and multiple linear regression modelling approach in Shivganga River basin, India. *Modeling Earth Systems and Environment* **5**, 951–962.

Karimi, S., Shiri, J., Kisi, O. & Makarynskyy, O. 2012 Forecasting water level fluctuations of Urmieh Lake using gene expression programming and adaptive neuro-fuzzy inference system. *The International Journal of Ocean and Climate Systems* **3**, 109–125.

Khashei, S. A. & Sarbazi, M. 2015 Study of spatial distribution of groundwater quality using LS-SVM, MLP, and geostatistical models. *Water and Wastewater* **26**, 93–103.

Khashei, S. A., Kouchakzadeh, M. & Ghahraman, B. 2011 Predicting dryland wheat yield from meteorological data using expert system, Khorasan Province, Iran. *Journal of Agricultural Science and Technology* **13** (4), 627–640.

Khatibi, R., Ghorbani, M. A., Kashani, M. H. & Kisi, O. 2011 Comparison of three artificial intelligence techniques for discharge routing. *Journal of Hydrology* **403**, 201–212.

Khatibi, R., Naghipour, L., Ghorbani, M. A. & Aalami, M. T. 2013 Predictability of relative humidity by two artificial intelligence techniques using noisy data from two Californian gauging stations. *Neural Computing and Applications* **23**, 2241–2252.

Kouvhskzadeh, M. & Bahmani, A. 2006 Assessment of artificial neural networks revenue in reducing required parameters for estimation of reference evapotranspiration. *Journal of Agricultural Sciences* **11**, 87–97.

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

Mahdavi, M. 1995 Water management and artificial recharge in Djahrom city. *Environment* **17**, 16–23. (In Persian).

Mohammadi, A. A., Yousefi, M., Soltani, J., Ahangar, A. G. & Javan, S. 2018 Using the combined model of gamma test and neuro-fuzzy system for modeling and estimating lead bonds in reservoir sediments. *Environmental Science and Pollution Research* **25**, 30315–30324.

Nourani, V., Mogaddam, A. A. & Nadiri, A. O. 2008 An ANN-based model for spatiotemporal groundwater level forecasting. *Hydrological Processes* **22**, 5054–5066.

Qasemi, M., Afsharnia, M., Farhang, M., Bakhshizadeh, A., Allahdadi, M. & Zarei, A. 2018a Health risk assessment of nitrate exposure in groundwater of rural areas of Gonabad and Bajestan, Iran. *Environmental Earth Sciences* **77**, 551.

Qasemi, M., Farhang, M., Biglari, H., Afsharnia, M., Ojrati, A., Khani, F., Samiee, M. & Zarel, A. 2018b Health risk assessments due to nitrate levels in drinking water in villages of Azadshahr, northeastern Iran. *Environmental Earth Sciences* **77**, 782.

Qasemi, M., Shams, M., Sajjadi, S. A., Farhang, M., Erfanpoor, S., Yousefi, M., Zarei, A. & Afsharnia, M. 2019 Cadmium in groundwater consumed in the rural areas of Gonabad and Bajestan, Iran: occurrence and health risk assessment. *Biological Trace Element Research* 1–10. doi: https://doi.org/10.1007/s12011-019-1660-7

Razaghmanesh, M., Salami, T. & Saraj, M. 2006 Quantitative and qualitative study of groundwater in Tabriz plain. In: *First National Conference on Irrigation and Drainage Networks Management*. Chamran University, Ahvaz, Iran.

Singh, K. K., Pal, M. & Singh, V. 2010 Estimation of mean annual flood in Indian catchments using backpropagation neural network and M5 model tree. *Water Resources Management* **24**, 2007–2019.

Tabarmayeh, M. & Vaezi Hir, A. 2015 Investigation on vulnerability of Tabriz-plain unconfined aquifer. *Journal of Water and Soil* **28**, 1137–1151.

Takdastan, A., Mirzabeygi, M., Yousefi, M., Abbasnia, A., Khodadadia, R., Soleimani, H., Mahvi, A. H. & Naghan, D. J. 2018 Neuro-fuzzy inference system prediction of stability indices and sodium absorption ratio in Lordegan rural drinking water resources in west Iran. *Data in Brief* **18**, 255–261.

Talebizadeh, M. & Moridnejad, A. 2011 Uncertainty analysis for the forecast of lake level fluctuations using ensembles of ANN and ANFIS models. *Expert Systems with Applications* **38**, 4126–4135.

Todd, K. D. & Mays, L. W. 2005 *Groundwater Hydrology*. John Wiley & Sons, New York, USA, pp. 508.

Vapnik, V. 2013 *The Nature of Statistical Learning Theory*. Springer Science & Business Media, New York, USA.

Vapnik, V., Golowich, S. E. & Smola, A. J. 1997 Support vector method for function approximation, regression estimation and signal processing. *Advances in Neural Information Processing Systems* **1997**, 281–287.

Wagh, V. M., Panaskar, D. B., Muley, A. A., Mukate, S. V., Lolage, Y. P. & Aamalawar, M. L. 2016 Prediction of groundwater suitability for irrigation using artificial neural network model: a case study of Nanded tehsil, Maharashtra, India. *Modeling Earth Systems and Environment* **2**, 1–10.

Wagh, V. M., Panaskar, D. B., Muley, A. A., Mukate, S. V. & Gaikwad, S. 2018 Neural network modelling for nitrate concentration in groundwater of Kadava River basin, Nashik, Maharashtra, India. *Groundwater for Sustainable Development* **7**, 436–445.

Zadeh, L. A. 1965 Fuzzy sets. *Information Control* **8**, 338–353.

Zaman Zad Ghvidel, S. & Mntaseri, M. 2014 Application of different data-driven methods for the prediction of total dissolved solids in the Zarinehroud basin. *Stochastic Environmental Research and Risk Assessment* **28**, 2101–2118.

First received 30 April 2019; accepted in revised form 29 July 2019. Available online 29 August 2019