Multivariate modeling of groundwater quality using hybrid evolutionary soft-computing methods in various climatic condition areas of Iran

Alireza Emadi, Sarvin Zamanzad-Ghavidel, Reza Sobhani and Ali Rashid-Niaghi

ABSTRACT

In the current study, several soft-computing methods including artificial neural networks (ANNs), adaptive neuro-fuzzy inference system (ANFIS), gene expression programming (GEP), and hybrid wavelet theory-GEP (WGEP) are used for modeling the groundwater's electrical conductivity (EC) variable. Hence, the groundwater samples from three sources (deep well, semi-deep well, and aqueducts), located in six basins of Iran (Urmia Lake (UL), Sefid-rud (SR), Karkheh (K), Kavir-Markazi (KM), Gavkhouni (G), and Hamun-e Jaz Murian (HJM)) with various climate conditions, were collected during 2004–2018. The results of the WGEP model with data de-noising showed the best performance in estimating the EC variable, considering all types of groundwater resources with various climatic conditions. The Root Mean Squared Error (RMSE) values of the WGEP model were varied from 162.068 to 348.911, 73.802 to 171.376, 29.465 to 351.489, 118.149 to 311.798, 217.667 to 430.730, and 76.253 to 162.992 μScm^{-1} in the areas of UL, SR, K, KM, G, and HJM basins. The WGEP model's performance (R-values) for deep wells, semi-deep wells, and aqueducts of the areas of the KM basin associated with the arid steppe cold (Bsk) dominant climate classification was the best. Also, the WGEP's extracted mathematical equations could be used for EC estimating in other basins.

Key words | climatic condition, groundwater quality, modeling, soft-computing, wavelet theory

HIGHLIGHTS

- Iran's groundwater resources face a critical situation.
- The Electrical Conductivity (EC) variable of various groundwater resources was estimated using single and hybrid wavelet theory methods.
- The impact of various climatic categories on the EC estimation was evaluated.
- The data de-noising by wavelet tools can improve the performance of EC estimation models.

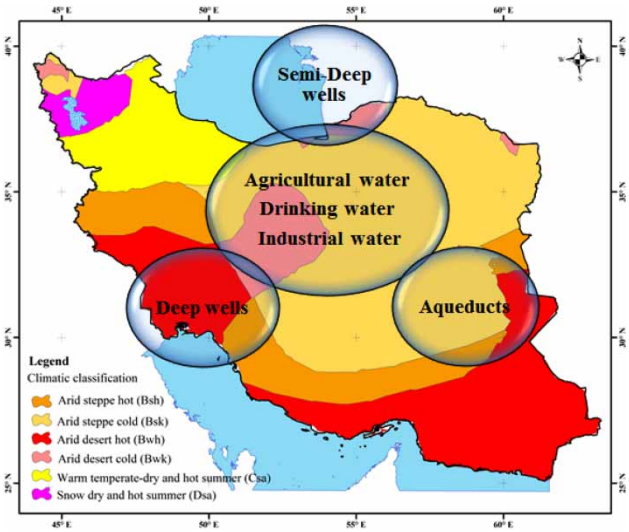
Alireza Emadi (corresponding author)
Reza Sobhani
Department of Water Engineering,
Sari Agricultural Science and Natural Resources University,
Sari,
Iran
E-mail: *emadia355@yahoo.com*

Sarvin Zamanzad-Ghavidel
Department of Irrigation & Reclamation Engineering, Faculty of Agriculture Engineering & Technology, College of Agriculture & Natural Resources,
University of Tehran-INSF,
Karaj, Alborz,
Iran

Ali Rashid-Niaghi
College of Food, Agriculture, and Natural Resource Science (CFAN),
University of Minnesota,
Minneapolis,
USA

doi: 10.2166/aqua.2021.150

GRAPHICAL ABSTRACT



INTRODUCTION

In recent decades, the increasing growth of industry and agriculture, as well as climate changes and human development, have increased the demand for water resources which have caused a serious challenge of water quality deterioration (Yang *et al.* 2017; Kaur *et al.* 2020). Groundwater, as the world's second-largest source of fresh water, could be a safe and reliable source of drinking water for many communities and regions around the world (Hekmatnia *et al.* 2020). Groundwater pollution arising from the deleterious consequences of industrial, agricultural, and service development is caused by unauthorized disposal of a variety of industrial and municipal wastewater which contains various pollutants, regardless of environmental considerations, which can endanger human health (Nigam & Yadav 2019). Considering the importance of groundwater resources and the country's location in arid and semi-arid climates, it is important to study the quality of water resources and their contaminants as well as modeling via soft-computing methods (Sharghi *et al.* 2019). In recent decades, soft-computing methods such as multi-linear regression (MLR), artificial neural network (ANN), support vector machines (SVM), adaptive neuro-fuzzy inference system (ANFIS), multivariate adaptive regression splines (MARS), gene expression programming (GEP) methods, and their hybrids with wavelet theory (W), have been successfully employed in surface and groundwater quality modeling (Wang *et al.* 2020).

Khudair *et al.* (2018) applied the ANN method without hybrid methods to predict the Water Quality Index (WQI). Their results showed that the pH and chloride variables have a significant influence on WQI prediction with the R^2 value of 0.973 for the optimal model. Wagh *et al.* (2018) used the ANN method without hybrid methods for modeling the nitrate concentration in the groundwater resources of Kadava river basin. Their results reflected that the Levenberg–Marquardt (LM) back-propagation algorithm was the effective algorithm of ANN models for the prediction of water quality variables. Zaqoot *et al.* (2018) applied the multi-layer perceptron-ANN (MLP-ANN) model successfully to predict the nitrate concentration of 15 groundwater wells in the Khan Younis and Rafah areas (semi-arid climate) with R and RMSE values of 0.838 and 63.236. Azad *et al.* (2018) applied different hybrid evolutionary algorithms (EA) with ANFIS without data de-noising to estimate the Gorganrood river water quality. The results showed that the ANFIS-Genetic Algorithm (ANFIS-GA)

has a suitable performance for estimating Sodium Adsorption Ratio (SAR). Kisi *et al.* (2019) used various soft-computing methods to estimate electrical conductivity (EC), total hardness (TH), and SAR variables of groundwater resources in Isfahan-Borkhar, Iran. The results indicated that the hybrid of the continuous genetic algorithm (CGA) algorithm with the ANFIS method had the best performance for estimating the EC, SAR, and TH, respectively. Aryafar *et al.* (2019) applied GP, ANFIS, and ANN models without hybrids for estimating groundwater quality variables such as TH, total dissolved solids (TDS), and EC of 12 wells in the Khezri plain. According to the results, the GP can be considered as a promising method to estimate the quality variables of groundwater resources with various uses. Kadam *et al.* (2019) used ANN and MLR methods without hybrid models to estimate the WQI of Shivganga River basin, and 34 representative groundwater characteristics including pH, EC, TDS, TH, Ca, Mg, Na, K, Cl, HCO_3, SO_4, NO_3, and PO_4 variables. Their results indicated that the estimation of ANN models had an acceptable performance of RMSE values. Jafari *et al.* (2019) used MLP, ANFIS, SVM, and GEP methods to estimate the TDS of a groundwater aquifer in Tabriz plain. They found that the GEP model had superior performance to other methods with the R^2 and RMSE values of 0.998 and 58.930. Maroufpoor *et al.* (2019) applied ANN and ANFIS models without hybrids for estimating the spatial distribution of groundwater's EC variable in Keshit, Bam Normashir, and Rhmtabad plains. They illustrated that the ANN method had the best performance with the R^2 value of 0.992 and RMSE value of 142.462 compared to the ANFIS approach.

W-hybrid models are mostly developed in estimating surface water quality without reducing data noises. Montaseri *et al.* (2018) used single and W-hybrid soft-computing methods, including ANN, ANFIS, GEP, wavelet-ANN (WANN), wavelet-ANFIS (WANFIS), and wavelet-GEP (WGEP) without data de-noising, to estimate the amount of TDS in rivers of four basins with different climatic categories (e.g. snow, dry-hot summer (Dsa), arid steppe, cold (Bsk), arid desert, cold (Bwk), and arid steppe, hot (Bsh) Koppen Geiger climate categories), located in Iran. The results indicated the superior performance of W-hybrid methods compared to single methods. Chen *et al.* (2020) and Rajaee *et al.* (2020) investigated single and W-hybrid soft-computing methods to estimate the river water qualitative variables. Their evaluations emphasized the applicability of the soft-computing methods in the estimation of water quality indicators for rivers and the superiority of W-hybrid methods compared to the single soft-computing methods for various climatic conditions.

The current study with de-noising the groundwater quality variables corresponding to various climate categories and types of groundwater sources could fill the gaps of previous studies in EC estimating. Our selected basins are located in Iran and were selected based on Koppen Geiger classification that includes Urmia Lake (UL) with a snow, dry and hot summer (Dsa), Sefid-rud (SR) with warm temperate and dry and hot summer (Csa), Karkheh (K) with an arid steppe, hot (Bsh), Kavir-Markazi (KM) with an arid steppe, cold (Bsk), Gavkhouni (G) with an arid desert, cold (Bwk), Hamun-e Jaz Murian (HJM) with an arid desert, hot (Bwh) climate categories. Therefore, the objectives of the current study are to use single and W-hybrid soft-computing methods for the area of the six selected basins with various climate conditions to: (1) apply soft-computing methods (ANN, ANFIS, and GEP) with a novelty structure of W-hybrid model (WGEP) for data de-noising and groundwater's EC estimating; (2) investigate the impact of climate variability and different types of groundwater sources (deep well, semi-deep well, and aqueducts) on the EC estimating models; (3) explore mathematical relationships of groundwater quality variables at spatial and temporal scales for the various groundwater resource types and their validations.

METHODOLOGY

Research concepts and steps

The results of the current research lead to the extraction of the mathematical relation governing the qualitative variables of different types of groundwater resources that could be applied at different time and spatial scales. Figure 1 displays the steps of the current research. The Köppen and Geiger climate classification scheme separates the climates into five main groups (A, B, C, D, and E).

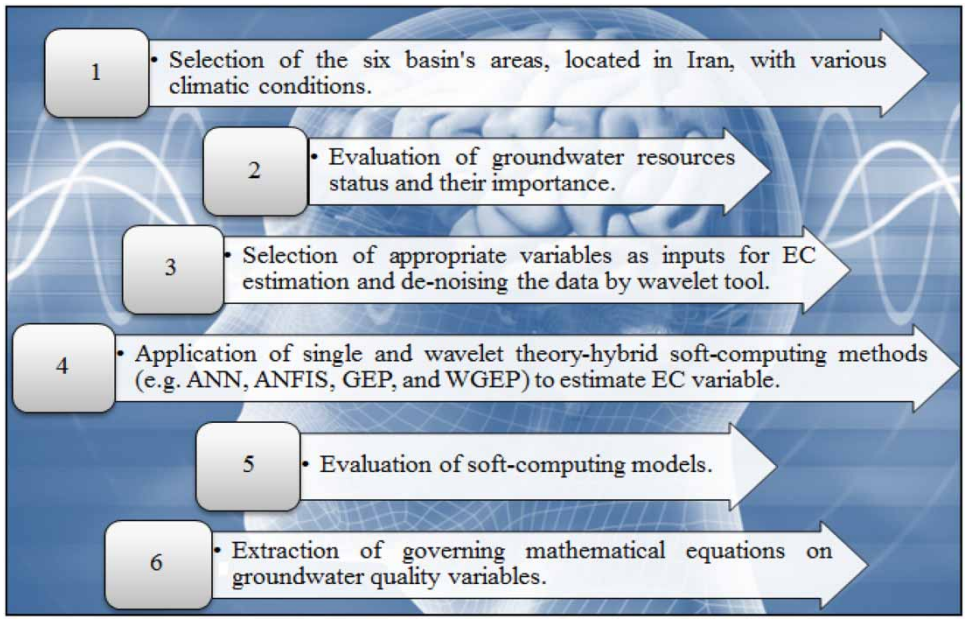


Figure 1 | Steps of the current research.

ANNs approaches

Various types of soft-computing models, such as ANNs, can be used for a variety of applications. The learning process of ANN models has a similarity with the human brain's performance (Haykin 1998; Ham & Kostanic 2001). Three basic characteristics of ANNs for determining the optimal model include: (1) applied learning algorithm, (2) activation function, and (3) neuron numbers. In the current study, the LM algorithm with three-layers was used for the training of the ANNs estimation models. The logsig, tansig, and purelin functions were applied in the hidden and output layers as activation functions. Also, MATLAB software (Montaseri *et al.* 2018) was used to develop the ANNs models and the trial-error methods were applied to determine the optimal number of neurons in the third layer of the ANN models (Barzegar *et al.* 2016).

ANFIS approaches

The relationship between fuzzy logic and neural networks has led to the development of various types of systems that are called 'adaptive neural-fuzzy inference systems (ANFIS)' (Jang *et al.* 1997). The first orders Sugeno fuzzy If/Then rules are represented below:

$$\text{Rule 1: If } x \text{ is } A_1 \text{ and } y \text{ is } B_1, \text{ then } f_1 = p_1x + q_1y + r_1 \quad (1)$$

$$\text{Rule 2: If } x \text{ is } A_2 \text{ and } y \text{ is } B_2, \text{ then } f_2 = p_2x + q_2y + r_2 \quad (2)$$

Membership functions (MFs) of inputs *Cl* and *K* variables (*x, y*), are defined by A_1, A_2, and B_1, B_2 (*LOW, LOW* and *HIGH, MEDIUM,* respectively). The importance of the cluster's number is in determining the efficient radius amount. The best radius values ranged from 0.200 to 0.600.

GEP approaches

The GEP is one of the newest methods of evolutionary algorithms that is more applicable because of its high accuracy (Ferreira 2006). Design and implementation steps of the GEP methods include: (1) defining the fitness function; (2) defining the terminals and functions; (3) determining the structure of chromosomes (number of generations, length, and number of genes); (4) determining the linking function of genes; (5) specifying the operators and finally, (6) executing the methods (Ferreira 2006). For the terminal set (T), SO_4, Cl, SAR, K, Mg, and Ca variables were selected based on their significant correlation coefficients with the EC variable.

Wavelet theory

Wavelet analysis is a set of mathematical functions that are used to continuously analyze the signal to its frequency components that are based on Fourier transform (Montaseri *et al.* 2018). It can be calculated as follows:

$$DWT(m,t) = \frac{1}{\sqrt{a^m}} \sum_n x(n) f\left(\frac{t - nba^m}{a^m}\right) \quad (3)$$

where a and b are the scaling and translation function of integer variable m; t is an integer variable that refers to a point of the input signal; n is the discrete-time index; $x(t)$ is a given signal and $f(t)$ is the mother wavelet. In the present study, one-dimensional Daubechies (db4) wavelet, based on its similarity in shape with data series, is used to decompose data into main and detailed sub-series. The structures of ANN, ANFIS, GEP, and WGEP are shown in Figure 2.

Case study and dataset

In this study, the groundwater quality data and the volume of agricultural, drinking, and industrial water uses were collected by the Iranian Water Resources Management Company (http://wrbs.wrm.ir/) for the study period 2004–2018. The groundwater quality data included three groundwater resources, namely deep wells, semi-deep well, and aqueducts ($z = 1$, 2, and 3), for six basins with various

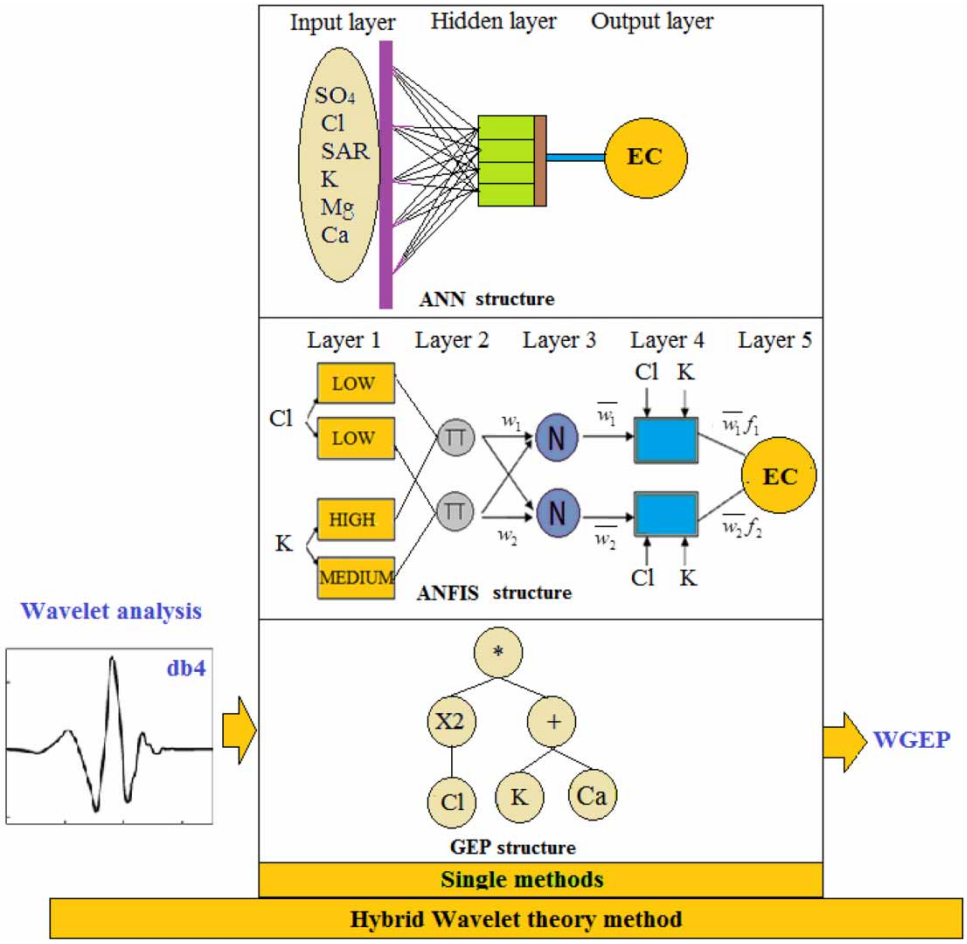


Figure 2 | ANN, ANFIS, GEP, and WGEP structure.

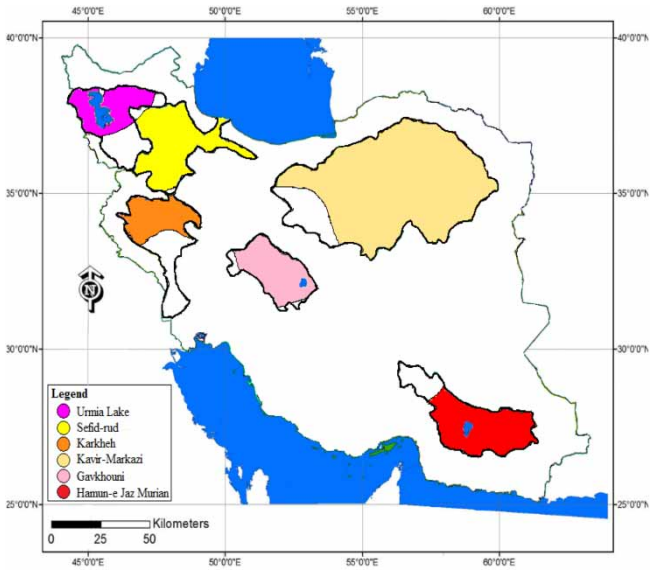


Figure 3 | Geographical location of the study basins.

climate conditions. Figure 3 displays the geographical location of the studied basins. The characteristics of the EC variable for the selected area's basins are listed in Table 1. The maximum variation coefficient of EC variable for deep and semi-deep wells, and aqueducts with the values of 1.339, 1.421, and 1.383, are related to the areas of UL, KM, and SR basins, respectively. On the other hand, the minimum variation coefficient with the value of 0.663 for deep wells was related to the HJM basin, and semi-deep wells and aqueducts with values of 0.740 and 0.259 were related to the K basin.

To create soft-computing models for EC estimation, 75 and 25% of collected data was used for training and testing. Also, SO_4, Cl, SAR, K, Mg, and Ca variables (meq.L^{-1}) were selected as input variables for EC estimation due to a significant Pearson correlation coefficient ($\alpha = 0.050$) with the EC variable.

Calculating the model's performance

To determine the optimal models for EC estimation, four evaluation criteria, namely the correlation coefficient (R), root mean squared error (RMSE), mean absolute error (MAE), and relative error (RE) were used as follows:

$$R = \left[\frac{\sum_{i=1}^{N} (EC_{io} - \overline{EC_o})(EC_{ie} - \overline{EC_e})}{\sqrt{\sum_{i=1}^{N} (EC_{io} - \overline{EC_o})^2 \sum_{i=1}^{N} (EC_{ie} - \overline{EC_e})^2}} \right] \quad (4)$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (EC_{io} - EC_{ie})^2} \quad (5)$$

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |EC_{io} - EC_{ie}| \quad (6)$$

Table 1 | The characteristics of EC variable (μScm^{-1}) data in the area of the selected basins

(z)	Min	Max	Mean	S[a]	CV[b]	Min	Max	Mean	S	CV
	Urmia Lake					**Sefid-rud**				
1	165.700	28,900.000	2,088.028	2,795.581	1.339	4.000	41,010.000	769.997	843.591	1.096
2	100.000	44,150.000	2,884.476	3,577.346	1.240	30.000	9,340.000	1,050.924	968.232	0.921
3	207.000	13,940.000	1,166.112	1,079.360	0.926	484.000	14,800.000	2,238.099	3,094.439	1.383
	Karkheh					**Kavir-Markazi**				
1	10.140	8,445.000	1,706.736	1,409.279	0.826	71.000	33,330.000	4,036.771	3,570.677	0.885
2	180.000	5,460.000	892.035	659.896	0.740	318.000	14,600.000	1,454.732	2,066.838	1.421
3	183.000	772.000	430.612	111.619	0.259	312.000	17,180.000	2,885.502	2,374.714	0.823
	Gavkhouni					**Hamun-e Jaz Murian**				
1	198.000	21,200.000	3,892.439	3,516.595	0.903	400.000	7,870.000	2,274.334	1,508.659	0.663
2	286.000	33,700.000	6,304.933	5,644.910	0.895	380.000	9,660.000	2,862.012	1,558.490	0.882
3	188.000	11,520.000	1,464.257	1,469.432	1.004	509.000	5,520.000	1,145.295	577.912	0.505

[a]Standard deviation.
[b]Coefficient of variation.

$$RE = \frac{|EC_{io} - EC_{ie}|}{EC_{io}} \quad (7)$$

where N shows the total number of groundwater quality samples, $\overline{EC}_o$ and $\overline{EC}_e$ are the average of observed and estimated EC, and EC_{io} and EC_{ie} are observed and estimated EC variables, respectively. A model with R and RE values closer to 1 and $RMSE$ and MAE values closer to 0 were chosen as optimal models (Montaseri *et al.* 2018).

RESULTS AND DISCUSSION

Status of the groundwater resources

During the study period, the total number of deep wells had an increasing trend in the areas of all studied basins. Deep wells located in the K and KM basins with Bsh and Bsk climate classes, semi-deep wells located in the UL, G, and HJM basins with Dsa, Bwk, and Bwh climate classes, and aqueducts located in the SR basin with Csa climate type, had the highest average amount of EC variable. Also, in the areas of K and HJM with dominant Bsh and Bwh climate types and minimum average of EC values in semi-deep wells and aqueducts, which are more affected by climate change, with the reduction of surface water resources, groundwater resources have begun to meet the needs. The results showed that the highest water abstraction for agricultural uses, with the highest average EC for deep wells and aqueducts, is related to the KM basin. Also, the lowest water abstraction for agricultural use, with the lowest average amount of EC for deep wells, is related to the SR basin. Therefore, in most cases, groundwater resources that had low quality (more salinity) are prioritized for agricultural uses. Areas with warmer and drier climates had lower quality of groundwater resources with higher agricultural uses.

EC estimation using ANN, ANFIS, and GEP

Three soft-computing approaches, namely ANN, ANFIS, and GEP, were applied to estimate groundwater quality variables of deep wells, semi-deep wells, and aqueducts in different areas of Iran with various climatic conditions. The results of optimal ANN, ANFIS, and GEP models by groundwater resource types in the areas of the studied basins are listed in Table 2. The numbers of neurons in the hidden layer of three-layer structure ANN models for deep wells, semi-deep wells, and aqueducts were (2, 5, 2), (3, 2, 4), (2, 2, 4), (4, 5, 2), (2, 4, 2) and (3, 2, and 2) for the areas of UL, SR, K, KM, G, and HJM basins, respectively. The activation functions of output nodes were obtained linear-purelin or tangent sigmoid-tansig for all the areas with various groundwater resource types. The activation functions of hidden nodes of ANN models were respectively (tansig, tansig and tansig), (logsig, tansig and tansig), (tansig, tansig and tansig), (tansig, tansig and tansig), (logsig, tansig and tansig) and (tansig, tansig and logsig) for the groundwater resource types ($z = 1$, 2, and 3) located in the areas of UL, SR, K, KM, G, and HJM basins. The radii values of ANFIS models for the deep wells, semi-deep wells, and aqueducts groundwater resources type were UL: 0.260, 0.320, 0.350; SR: 0.420, 0.370, 0.510; K: 0.350, 0.410, 0.280; KM: 0.550, 0.230, 0.480; G: 0.550, 0.220, 0.430 and HJM: 0.330, 0.270, 0.460, respectively. The R values for soft-computing models are close to 1, with the quality relations being: $R_{GEP} > R_{ANFIS} > R_{ANN}$ for all groundwater resource types and studied areas. The ANFIS model exceeded the ANN model's performance; also, the GEP models had a better performance for EC estimation than the ANFIS and ANN models for all areas (see Supplementary Data, Figure S1 for more details).

The values of a coefficient of fitted equation ($y = ax$) to the observed and residual values of GEP optimal models for deep wells, semi-deep wells, and aqueducts were UL: 0.051, 0.012, 0.004; SR: 0.090, 0.003, 0.027; K: 0.002, 0.055, 0.005; KM: 0.002, 0.005, 0.011; G: 0.002, 0.011, 0.027 and HJM: 0.016, 0.040, 0.049, respectively. The minimum value of a coefficient, close to 0, indicates the randomness and independence of the estimated values of the GEP optimal models. The minimum value of a coefficient with the value of 0.002 was obtained for deep wells of Bsh, Bsk, and Bwk climate classes with an arid climate group. These results indicated that the performance of the models is influenced by the climatic characteristics and groundwater resource types. The variation coefficients of a, by considering all types of groundwater resources for Bwh, Bsk, Bwk, Csa, Dsa, and Bsh climate categories,

Table 2 | The results of optimal ANN, ANFIS, and GEP models by the groundwater source types in the area of the studied basins

(z)	Models	R	RMSE (μSm^{-1})	MAE (μSm^{-1})	Models	R	RMSE (μSm^{-1})	MAE (μSm^{-1})
	Urmia Lake				**Sefid-rud**			
1	ANN1	0.911	461.134	300.768	ANN1	0.531	288.716	213.592
	ANFIS1	0.959	346.136	214.611	ANFIS1	0.817	221.153	133.547
	GEP1	0.978	331.825	185.716	GEP1	0.897	188.341	111.198
2	ANN2	0.936	753.345	537.098	ANN2	0.891	426.661	336.418
	ANFIS2	0.956	701.267	396.788	ANFIS2	0.962	252.955	176.031
	GEP2	0.972	516.127	298.854	GEP2	0.994	100.639	62.725
3	ANN3	0.922	397.020	251.201	ANN3	0.982	593.152	393.587
	ANFIS3	0.958	246.651	155.002	ANFIS3	0.992	433.703	238.563
	GEP3	0.966	220.015	136.854	GEP3	0.994	316.688	234.540
	Karkheh				**Kavir-Markazi**			
1	ANN1	0.909	610.482	353.319	ANN1	0.985	903.513	566.441
	ANFIS1	0.931	526.781	235.205	ANFIS1	0.994	509.732	303.528
	GEP1	0.949	425.025	226.207	GEP1	0.995	479.211	269.172
2	ANN2	0.914	439.747	276.407	ANN2	0.963	669.115	400.140
	ANFIS2	0.950	280.684	203.633	ANFIS2	0.986	420.022	255.860
	GEP2	0.974	204.892	112.064	GEP2	0.997	198.718	131.336
3	ANN3	0.861	50.151	37.443	ANN3	0.962	570.052	437.290
	ANFIS3	0.909	45.249	30.297	ANFIS3	0.986	325.840	229.170
	GEP3	0.943	35.540	26.254	GEP3	0.998	144.612	97.267
	Gavkhouni				**Hamun-e Jaz Morian**			
1	ANN1	0.929	1,408.750	682.036	ANN1	0.970	367.961	292.385
	ANFIS1	0.976	794.411	465.461	ANFIS1	0.977	347.390	191.247
	GEP1	0.992	402.501	293.063	GEP1	0.991	214.817	157.640
2	ANN2	0.989	1,026.550	733.365	ANN2	0.953	503.411	406.534
	ANFIS2	0.993	759.044	502.505	ANFIS2	0.966	481.098	384.778
	GEP2	0.997	493.792	324.234	GEP2	0.990	258.490	209.264
3	ANN3	0.914	708.084	383.206	ANN3	0.702	201.766	152.124
	ANFIS3	0.984	313.800	181.666	ANFIS3	0.841	202.112	160.730
	GEP3	0.988	272.315	153.081	GEP3	0.977	98.744	70.362

were 0.478, 0.718, 0.942, 1.133, 1.129, and 1.457, respectively. These results reflected that the Bwh and Bsh climate types with an arid climate group had the lowest and the highest effect on the amounts of WGEP model's performance to estimate the EC values for various groundwater resource types, respectively. Also, the variation coefficients of *a*, for aqueducts, semi-deep wells, and deep wells, were 0.835, 1.025, and 1.325. Therefore, aqueducts and deep wells located in six of the study climates had the lowest and the highest effect on the amounts of WGEP model's performance for EC estimating, respectively. The model's performances (R-values) for deep wells and aqueducts located in the areas of KM with Bsk and for semi-deep wells located in the areas of G with Bwk dominant type of Koppen climate classification were the best. The model's performances for deep wells, semi-deep wells, and aqueducts of the SR, UL, and K areas basins associated with the Csa, Dsa, and Bsh dominant type of Koppen climate classification were the poorest among the other climates, respectively. The Csa climate type is a climate where the coldest month is warmer than –3 °C but colder than +18 °C and summers are dry and hot. The Dsa is a climate where there is at least one month colder than –3 °C and summers are dry and hot. The Bsh is a climate which means annual temperature is greater than or equal to 18 °C and is too dry to support a forest, but not dry

enough to be a desert, usually consisting of grassland plains. The climatic conditions and type of groundwater resource only affect the performance amount of the models and not their priority. However, the results of the three methods could be acceptable for estimating the EC variable in groundwater resources with various climatic conditions. The *RE* values of EC estimated data via GEP models at different ranges ($25\%_{max}$, $50\%_{mid}$, $25\%_{min}$) are listed in Table 3. The maximum values of *RE* are related to the areas of SR and K basins with the values of 1.231 and 1.215, respectively at the range of $25\%_{min}$. Also, the *RE* values of the GEP models EC estimated data for all cases were acceptable. The GEP models percentage of performance improvement compared to the ANN and ANFIS models for semi-deep wells groundwater sources type equaled approximately 31, 76, 53, 70, 52, and 49% and 26, 60, 27, 53, 35, and 46% in the areas of UL, SR, K, KM, G, and HJM basins, respectively.

EC estimation using WGEP

For improving the performance of GEP models and constructing the hybrid GEP estimation models with wavelet tools (WGEP), the first point is to decompose the groundwater quality variables, into the subseries of mains and details (A and D sub-series) via a wavelet tool. To make and develop the WGEP models, D decomposed subseries are introduced as noise and removed from the models. After data de-noising

Table 3 | RE values of the estimated EC data based on the GEP model at different ranges

		Basins					
(z)	Range	Urmia Lake	Sefid-rud	Karkheh	Kavir-Markazi	Gavkhouni	Hamun-e Jaz Murian
1	$25\%_{max}$	0.176	0.145	0.105	0.047	0.061	0.052
	$50\%_{mid}$	0.235	0.117	0.122	0.089	0.105	0.084
	$25\%_{min}$	0.181	1.231	1.215	0.283	0.487	0.098
2	$25\%_{max}$	0.081	0.054	0.076	0.037	0.033	0.050
	$50\%_{mid}$	0.190	0.064	0.076	0.153	0.085	0.083
	$25\%_{min}$	0.803	0.079	0.380	0.135	0.148	0.109
3	$25\%_{max}$	0.092	0.148	0.055	0.035	0.074	0.081
	$50\%_{mid}$	0.135	0.138	0.068	0.047	0.084	0.050
	$25\%_{min}$	0.429	0.224	0.054	0.081	0.463	0.040

Table 4 | Results of optimal WGEP models by the groundwater source types in the areas of the study basins

(z)	Models	R	RMSE (μSm^{-1})	MAE (μSm^{-1})	Models	R	RMSE (μSm^{-1})	MAE (μSm^{-1})
	Urmia Lake				**Sefid-rud**			
1	WGEP1	0.978	227.536	142.813	WGEP1	0.910	152.741	110.455
2	WGEP2	0.974	348.911	226.128	WGEP2	0.994	73.802	28.516
3	WGEP3	0.967	162.068	99.114	WGEP3	0.997	171.376	137.144
	Karkheh				**Kavir-Markazi**			
1	WGEP1	0.963	351.489	218.147	WGEP1	0.995	311.798	126.870
2	WGEP2	0.979	161.708	111.094	WGEP2	0.998	161.596	127.654
3	WGEP3	0.949	29.465	19.475	WGEP3	0.998	118.149	35.509
	Gavkhouni				**Hamun-e Jaz Morian**			
1	WGEP1	0.993	316.300	195.660	WGEP1	0.992	162.992	91.992
2	WGEP2	0.998	430.730	313.070	WGEP2	0.991	159.008	115.160
3	WGEP3	0.992	217.667	124.331	WGEP3	0.980	76.253	64.243

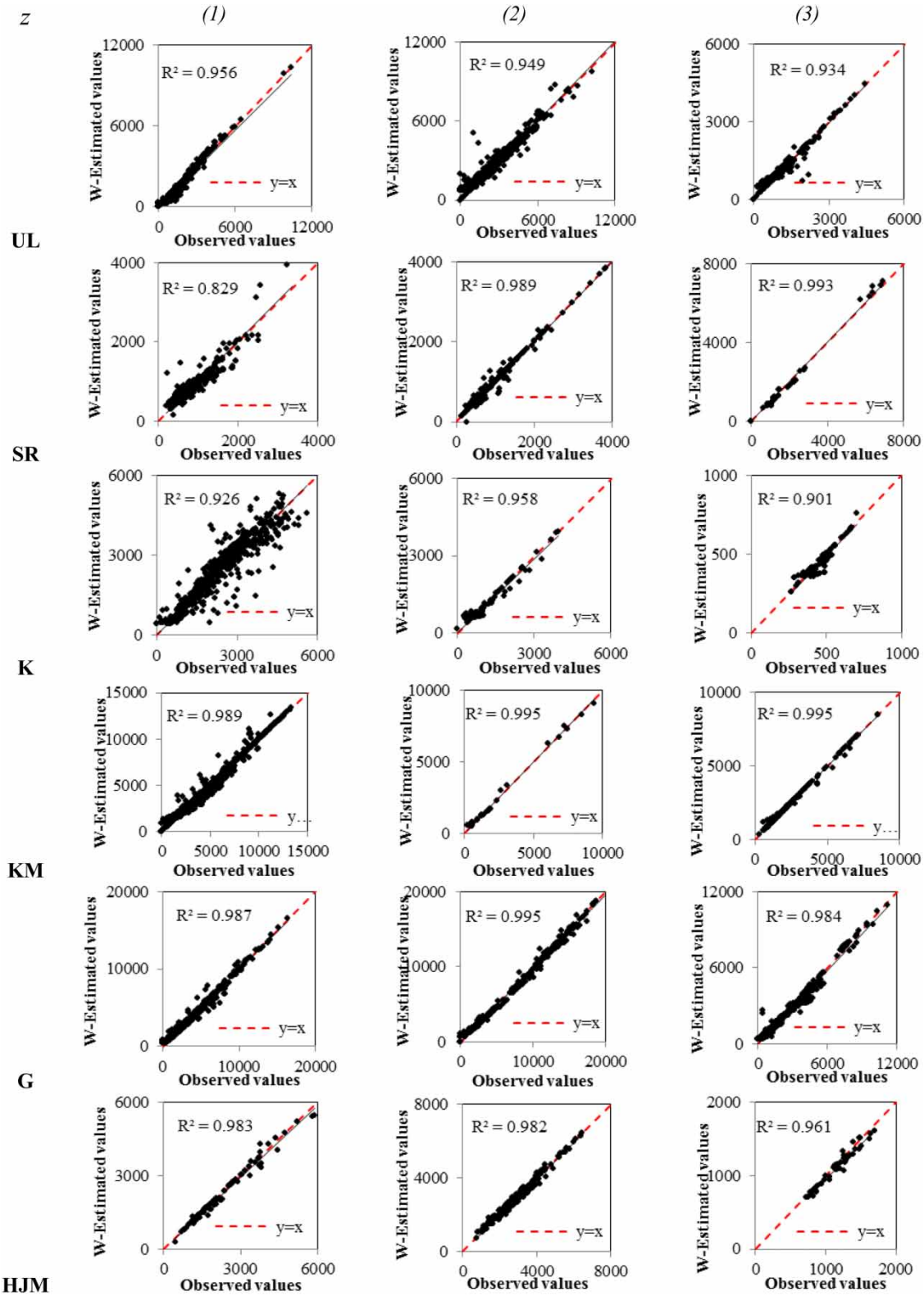


Figure 4 | The observed and estimated EC values for ($z = 1$, 2, and 3) with the WGEP models.

(see Tables S1 and S2 for more details), $A \geq 0$ decomposed subseries values are estimated separately with GEP models. The results of WGEP optimal models during the test period are listed in Table 4. The RMSE values of the WGEP models with db4 mother wavelet varied from 162.068 to 348.911, 73.802 to 171.376, 29.465 to 351.489, 118.149 to 311.798, 217.667 to 430.730, and 76.253 to 162.992 (μScm^{-1}) in the areas of UL, SR, K, KM, G, and HJM basins for various groundwater source types. Figure 4 displays the observed and estimated EC variable during the test period using the

Table 5 | Developed mathematical equations using the WGEP model

Area	(z)	Equations
UL	1	$EC_{st} = (-3.80)(3Ca_{st} - SAR_{st}) - 119.10(K_{st} - SO_{4st} - Cl_{st}) - 0.40((Cl_{st})(Ca_{st}) + SAR_{st}) - Cos(Ca_{st}/SO_{4st}) + 200.40$
	2	$EC_{st} = 82.92(SO_{4st} + Cl_{st}) + 2Mg_{st} + 17.24\,SAR_{st} + 10.62Cl_{st} + Ca_{st} - (SAR_{st})(K_{st}) + 717.21$
	3	$EC_{st} = SO_{4st}(SAR_{st} - Mg_{st}) + 9.92Ca_{st} + SO_{4st} + Mg_{st}\,(21.34 + 0.22Mg_{st}) + (Mg_{st} + SO_{4st} + Cl_{st} + 4.32)(76.39 - Sin(Mg_{st}))$
K	1	$EC_{st} = (-0.28)(SO_{4st}^2) + a\tan(SO_{4st}).(0.59K_{st} - 5.75SO_{4st}) + 92.48(5.75 + SO_{4st}) + 92.20(Cl_{st})$
	2	$EC_{st} = 6(Mg_{st} + SAR_{st})(36 + K_{st}) - 5.83(Mg_{st}).(SO_{4st}^{(1/3)}) + 78.65(Ca_{st} - Mg_{st})$
	3	$EC_{st} = Ln(Cl_{st}).(4.42K_{st} + 2.16)^3 + 83.10(SAR_{st} + 2.16)^{(1/3)}.[(SAR_{st})(SO_{4st}) + Mg_{st} + Ca_{st}] + [(SAR_{st})^{(1/2)} - 5.42SO_{4st} + SAR_{st} - 4.42)].(Ca_{st}) + 4.42$
G	1	$EC_{st} = 9.42(76.77 - 9.42K_{st}).(SAR_{st}^{(1/3)}) - SO_{4st}(6.55 - (\frac{2.88 - Cl_{st}}{6.55}) - SO_{4st}) + 43.59Cl_{st}(a\tan(K_{st} - 9.72))^2$
	2	$EC_{st} = (12.89 - SAR_{st})(Cl_{st} + Mg_{st} - Ca_{st} - 1.31K_{st}) + (Ln(SO_{4st}) + SAR_{st})(K_{st} - SO_{4st} - 3.67) + (86.12 + SAR_{st})(Cl_{st} + SO_{4st} + 3.40)$
	3	$EC_{st} = SO_{4st}(37.21 - 0.16Cl_{st} + SO_{4st} + 4.83) + 112.68(Cl_{st} + 4.34) + SAR_{st}(SAR_{st} - (Cl_{st} - 10.59))$
SR	1	$EC_{st} = Cl_{st} - 4.79(1 + (6.00 - (K_{st}).(Mg_{st})(Cl_{st})^{-1})) + Ca_{st}(0.50 + Mg_{st}) - (SAR_{st} + 0.79)(SAR_{st} + 1.13) - 12.69(6.32(6.32 - SO_{4st}) - 9.59Cl_{st})$
	2	$EC_{st} = (Mg_{st} + SAR_{st})(146.00 + Cl_{st} + K_{st}) - (13.84 + Ca_{st})(Ca_{st} - Cl_{st} - SO_{4st}) + 96.24Ca_{st}$
	3	$EC_{st} = 87.24(6.47 + SO_{4st}) + Cl_{st}^2 - ((SAR_{st} + K_{st})(Mg_{st} - 0.37)Ca_{st}^{(1/3)}) - (Cl_{st} - 46.06(3.83 + K_{st})) + 14.67$
KM	1	$EC_{st} = 50.18(Cl_{st} + Mg_{st} + 2SO_{4st}) + (20.43 - Ln(SO_{4st}))(K_{st} + SAR_{st} + 2Cl_{st}) + Cl_{st} + SAR_{st}$
	2	$EC_{st} = (Ca_{st} + SO_{4st} + 16.59)^2 + (Cl_{st} + SAR_{st})(K_{st}^{(1/2)} - K_{st} + 80.64) + Mg_{st}$
	3	$EC_{st} = 72.66(Cl_{st} + SAR_{st} + Ca_{st}) + (3.77Mg_{st} + 16.55)SO_{4st} - Cos(3.77Mg_{st})$
HJM	1	$EC_{st} = 3.35(Ca_{st} + 2SAR_{st})(Cl_{st})^{(-1)} + 5.10Ca_{.st}(K_{st} - Cl_{st}) - 26.01SO_{4st}.(Ca_{st})^{(1/2)} + (185.19 - Cl_{st}).(Ca_{st} + Cl_{st}) - 11.78(Cl_{st} - SAR_{st})$
	2	$EC_{st} = 7.18SO_{4st}(7.18 + Mg_{st}) + 52.56Mg_{st}^{(1/3)} + 39.94(Cl_{st} + SAR_{st}) + Ln(Mg_{st}) - 6.32Cl_{st} + 74.82(Cl_{st} - K_{st}) - 1.21SAR_{st}$
	3	$EC_{st} = Cl_{st}^{(1/2)}((Ln(SAR_{st}))^2 - (458.31 - SAR_{st})) + (15.84 + Cl_{st})((SAR_{st} + SO_{4st}) - 2.07(Ca_{st})^{(-1)} + (Cl_{st} + 8.77)(Mg_{st} - a\tan(SAR_{st}) + SAR_{st})$

st: denotes spatial and time for s and t, respectively.

WGEP method. The R values obtained were more than 0.910 for all areas with various groundwater source types in the WGEP optimal models. The results of R, RMSE, MAE, and graphical methods reflected the best performance of the WGEP models.

Wavelet analysis and developing WGEP models greatly enhances the performance of GEP models in all climate types and groundwater sources, by de-noising the data noises and the ability to establish very complex nonlinear relationships in its structure. The results indicated that the performance improvement of WGEP models compared to GEP varied from 17 to 35%, 13 to 32%, and 17 to 46% for deep wells, semi-deep wells, and aqueducts of study areas, respectively.

The WGEP model's performance (R-values) for deep wells, semi-deep wells, and aqueducts of the areas of the KM basins associated with the Bsk dominant type of Koppen climate classification was the best. The Bsk type of Koppen climate classification indicated the climate whose mean annual temperature is less than 18 °C and is too dry to support a forest, but not dry enough to be a desert, usually consisting of grassland plains. However, the main advantage of the GEP over other soft-computing methods (ANFIS and ANN) is in producing predictive equations. The equations obtained with the optimal WGEP models are listed in Table 5. The fitted equations can be applied at variable spatial and temporal scales. Due to the high performance of the KM basin for EC estimating in three types of groundwater resources, it is selected as a 'basic basin' for validating the extracted mathematical equations in other study basins. The validation results are listed in Table 6. Our results reflected the high ability of WGEP model's extracted mathematical equations for EC estimating of corresponding groundwater resource types in the areas of various basins. For instance, R values of deep well's extracted mathematical equations in validating were 0.984, 0.921, 0.968, 0.993, and 0.989 for UL, SR, K, G, and HJM basins, respectively. The highest R-value in the validating section is related to the basin with an arid climate, which is in the same climate categories as the basic basin. The result of the present study is in agreement with the findings by Khudair *et al.* (2018), Zaqoot *et al.* (2018), Aryafar *et al.* (2019), Maroufpoor *et al.* (2019), and Chen *et al.* (2020).

Table 6 | Validation of the basic basin's extracted mathematical equations

(z)	Basins	R	RMSE (μSm^{-1})	MAE (μSm^{-1})
1	UL	0.984	294.371	260.064
	SR	0.921	381.319	351.537
	K	0.968	590.083	457.742
	G	0.993	396.648	267.190
	HJM	0.989	430.465	372.695
2	UL	0.967	465.196	263.918
	SR	0.951	186.787	136.215
	K	0.970	298.122	187.384
	G	0.992	1,039.717	794.691
	HJM	0.958	448.666	320.393
3	UL	0.973	332.127	287.411
	SR	0.991	486.911	463.771
	K	0.796	190.740	181.727
	G	0.992	318.253	272.482
	HJM	0.975	228.150	218.876

CONCLUSION

Due to the reflected main results of our research, climate categories and type of groundwater resources had a major impact on the amount of model's performance in the groundwater resources quality estimating, but not on the priority of applied model's performance. The priority of the model's performance was: $R_{WGEP} > R_{GEP} > R_{ANFIS} > R_{ANN}$, without interfering with the climate classes and groundwater resource types. Our results strongly confirm the high ability of the EC estimating of WGEP model's improvements with a new structure of data de-noising models. These results reflected an important message including the data noise impact on the soft-computing model's performance in estimating EC values. The Bwh and Bsh climate types had the lowest and the highest effect, respectively on the amounts of WGEP model's performance to estimate the EC values for various groundwater resource types. On the other hand, aqueducts and deep wells located in six study climates had the lowest and the highest impact on the EC values of the WGEP model. The results reflected that the percentage improvement of WGEP models compared to GEP ranged from 13 to 46% for deep wells, semi-deep wells, and aqueducts of study areas. Also, the obtained R-values of WGEP optimal models (>0.910) for all areas with various groundwater source types are also in line with the suitable performance

of extended models. The *RE* values of WGEP models varied from 0.033 to 1.231 for three ranges of $25\%_{min,max}$ and $50\%_{mid}$, which could confirm the optimal estimation of the extended new structure model. As the existence of uncertainty in meteorological-hydrological variables is undeniable and introduced structure of soft-computing methods by eliminating data noise could improve the performance of models, selecting the meteorological-quantitative hydrological variables as model's input variables for EC estimation can be suggested for future research. The main practical point of our research indicated the high ability of WGEP model's extracted mathematical equations for EC estimating of corresponding groundwater resource types in the areas of other basins.

ACKNOWLEDGEMENTS

The authors would like to thank Sari Agricultural Science and Natural Resources University for financing this research [Code Number: 02-1399-08].

DATA AVAILABILITY STATEMENT

All relevant data are included in the paper or its Supplementary Information.

REFERENCES

Aryafar, A., Khosravi, V., Zarepourfard, H. & Rooki, R. 2019 Evolving genetic programming and other AI-based models for estimating groundwater quality parameters of the Khezri plain, Eastern Iran. *Environmental Earth Sciences* **78** (3), 69–82.

Azad, A., Karami, H., Farzin, S., Saeedian, A., Kashi, H. & Sayyahi, F. 2018 Prediction of water quality parameters using ANFIS optimized by intelligence algorithms (case study: Gorganrood River). *KSCE Journal of Civil Engineering* **22** (7), 2206–2213.

Barzegar, R., Adamowski, J. & Moghaddam, A. A. 2016 Application of wavelet-artificial intelligence hybrid models for water quality prediction: a case study in Aji-Chay River, Iran. *Stochastic Environmental Research and Risk Assessment* **30** (7), 1797–1819.

Chen, Y., Song, L., Liu, Y., Yang, L. & Li, D. 2020 A review of the artificial neural network models for water quality prediction. *Applied Sciences* **10** (17), 5776–5825.

Ferreira, C. 2006 *Gene Expression Programming: Mathematical Modeling by an Artificial Intelligence*. Vol. 21. Springer, Berlin.

Ham, F. M. & Kostanic, I. 2001 *Fundamental Neurocomputing Concepts. Principles of Neurocomputing for Science and Engineering*. Arnold Publishers, London, pp. 24–91.

Haykin, S. 1998 *Neural Networks A Comprehensive Foundation*, 2nd edn. Prentice-Hall, Upper Saddle River, USA, pp. 26–32.

Hekmatnia, H., Barzegari Banadkooki, F., Moosavi, V. & Zare Chahouki, A. 2020 Evaluation of groundwater suitability for drinking, irrigation, and industrial purposes (Case study: Yazd-Ardakan Aquifer, Yazd Province, Iran). *ECOPERSIA* **9** (1), 11–21.

Jafari, R., Torabian, A., Ghorbani, M. A., Mirbagheri, S. A. & Hassani, A. H. 2019 Prediction of groundwater quality parameter in the Tabriz plain, Iran using soft computing methods. *Journal of Water Supply: Research and Technology – AQUA* **68** (7), 573–584.

Jang, J. S. R., Sun, C. T. & Mizutani, E. 1997 Neuro-fuzzy and soft computing-a computational approach to learning and machine intelligence [Book review]. *IEEE Transactions on Automatic Control* **42** (10), 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.

Kaur, S., Singh, R., Vashisht, B. B., Gill, K. K. & Aggarwal, R. 2020 Modelling the response of paddy water balance on groundwater level fluctuations in Central Punjab. *Journal of Hydroinformatics* **22** (6), 1663–1671.

Khudair, B. H., Jasim, M. M. & Alsaqqar, A. S. 2018 Artificial neural network model for the prediction of groundwater quality. *Civil Engineering Journal* **4** (12), 2959–2970.

Kisi, O., Azad, A., Kashi, H., Saeedian, A., Hashemi, S. A. A. & Ghorbani, S. 2019 Modeling groundwater quality parameters using hybrid neuro-fuzzy methods. *Water Resources Management* **33** (2), 847–861.

Maroufpoor, S., Fakheri-Fard, A. & Shiri, J. 2019 Study of the spatial distribution of groundwater quality using soft computing and geostatistical models. *ISH Journal of Hydraulic Engineering* **25** (2), 232–238.

Montaseri, M., Ghavidel, S. Z. Z. & Sanikhani, H. 2018 Water quality variations in different climates of iran: toward modeling total dissolved solid using soft computing techniques. *Stochastic Environmental Research and Risk Assessment* **32** (8), 2253–2273.

Nigam, U. & Yadav, S. M. 2019 Development of computational assessment model of fuzzy rule based evaluation of groundwater quality index: comparison and analysis with conventional index. In *Proceedings of International Conference on Sustainable Computing in Science, Technology and Management (SUSCOM)*. Amity University Rajasthan, Jaipur, India.

Rajaee, T., Khani, S. & Ravansalar, M. 2020 Artificial intelligence-based single and hybrid models for prediction of water quality in rivers: a review. *Chemometrics and Intelligent Laboratory Systems* **200**, 103978–104042.

Sharghi, E., Nourani, V., Molajou, A. & Najafi, H. 2019 Conjunction of emotional ANN (EANN) and wavelet transform for rainfall-runoff modeling. *Journal of Hydroinformatics* **21** (1), 136–152.

Wagh, V., Panaskar, D., Muley, A., Mukate, S. & 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.

Wang, Y., Yuan, Y., Pan, Y. & Fan, Z. 2020 Modeling daily and monthly water quality indicators in a canal using a hybrid wavelet-based support vector regression structure. *Water* **12** (5), 1476–1497.

Yang, Q., Zhang, J., Hou, Z., Lei, X., Tai, W., Chen, W. & Chen, T. 2017 Shallow groundwater quality assessment: use of the improved Nemerow pollution index, wavelet transform and neural networks. *Journal of Hydroinformatics* **19** (5), 784–794.

Zaqoot, H. A., Hamada, M. & Miqdad, S. 2018 A comparative study of Ann for predicting nitrate concentration in groundwater wells in the southern area of Gaza Strip. *Applied Artificial Intelligence* **32** (7–8), 727–744.

First received 26 November 2020; accepted in revised form 30 January 2021. Available online 10 March 2021