Large-scale association analysis of climate drought and decline in groundwater quantity using Gaussian process classification (case study: 609 study area of Iran)

Saeed Azimi¹ · Mehdi Azhdary Moghaddam¹ · Seyed Arman Hashemi Monfared¹

Received: 25 March 2018 / Accepted: 9 April 2018 / Published online: 19 May 2018
© Springer International Publishing AG 2018

Abstract

Background The level of groundwater resources is changing rapidly and this requires the discovery of newer groundwater resources. Drought is one of the most significant natural phenomena affecting different aspects of human life and environment. During the last decades, the application of artificial intelligent techniques has been recognized as effective approaches to forecast an annual precipitation rate.

Method In this study, the association analysis of climate drought and a decline in groundwater level is addressed using Gaussian process classification (GPC) and backpropagation (BP) artificial neural network (ANN). This methodology is proposed to create a framework for decision making and reduce uncertainty in water resource management calculations, and in particular to optimize the management of groundwater drinking water sources.

Results Underground water levels in 609 study plains in Iran were used to predict drought over the test period, extending from 2017 to 2021. The artificial intelligence methods were implemented in the Python programming environment to achieve an annual precipitation rate. A statistical summary of the Rasterized Cells of the zoning maps was used to validate the prediction results. Considering the relationship between water quality reductions and drought in Iranian aquifers due to the occurrence of groundwater drought periods, the results were validated by analysis of the effect of climate drought using the Standardized Precipitation Index (SPI) on the occurrence of observed droughts with the Groundwater Resources Index (GRI). The results are well-illustrated by the observation of the predicted digits in the third dimension of the Gaussian distribution.

Conclusion According to the SPI indicator, the southern regions of the country, and especially the central parts of the plain, can be considered the most affected areas by the most severe future droughts. The prediction results indicate a decrease in drought severity as part of a two-year sequence involving a recurrence of drought exacerbation and relative decline, as well as a failed state after the critical condition of aquifers.

Keywords Gaussian process classification · Groundwater quality · Drought · Artificial neural network

Introduction

Having enough information about the status of water resources in a region plays a decisive role in water planning and agriculture. Therefore, awareness of the future status of water resources, especially in standard drinking water classes, is very important for urban and rural communities [1]. The level of groundwater resources is changing rapidly and this requires the discovery of newer groundwater resources [2]. The water supply reservoirs are shared with multiple uses of underground aquifers and vary in short-term cycles from year to year. Thus, predicting possible changes in the concentration of quality parameters is considered at least annually for the adoption of management measures [3].

✉ Mehdi Azhdary Moghaddam
mazhdary@eng.usb.ac.ir

Saeed Azimi
saed.azimi@pgs.usb.ac.ir

Seyed Arman Hashemi Monfared
hashemi@eng.usb.ac.ir

¹ Civil Engineering Department, Faculty of Engineering, University of Sistan and Baluchestan, P.O. Box 9816745563-161, Zahedan, Iran

One of the basic needs for sustainable water resources planning is the prediction of the amount of water for agricultural, industrial and urban loading. Rainfall plays a key role in the decline in the quality of water in the reservoirs in dry and semi-arid areas where permanent currents in the supply of groundwater resources are not significant. [4]. In times of droughts, changing the amount of available water, or deficiencies in water resources exploits is clearly a source of tension in a part of the community, which is generally part of the underground water resources. Therefore, it is necessary to predict the water power of each area at different times for efficient planning through appropriate and reliable methods.

One of the best available forecasting techniques is the artificial neural network, which is a new tool in nonlinear and indeterminate systems, as a prerequisite for accessing the database for forecasting data on water resources [5]. The artificial neural network is capable of predicting of changes in the water level of aquifer in dry and semi-arid areas. An artificial neural network is a simplified model of the natural nervous system and has the ability to learn, by processing on experimental data. The advantage of a neural network is the direct learning of data without the need to estimate their statistical characteristics. A neural network, regardless of any initial hypothesis and previous knowledge of the associations between the parameters studied, is able to find the relationship between the set of inputs and outputs to predict each output corresponding to the desired input.

To reduce the uncertainty caused by the prediction of the water level in aquifers and the total annual precipitation by statistical methods and in order to create the macro decision-making management platform, it is only possible to predict the accuracy of the results. One of the ways to accurately analyze the results is to use data generation methods, as well as to use statistical distribution methods to find trends. A non-random process in a natural phenomenon can occur due to different causes, which in any case leads to errors in the output of the prediction model and thus leads to the adoption of managerial decisions. As a solution in deductive discovery, the best way of guessing the future values of a hydrological phenomenon is to use the Gaussian processing classifier. In the theory of probability and statistics, a Gaussian process is a statistical model in which observations occur on a continuous domain, for example, time or space. In a Gaussian process, every point of the input space is a random variable with a normal statistical distribution. In addition, each finite set of random variables has a multivariate Gaussian distribution. The distribution of the Gaussian process is a common distribution of all these random variables (limited and unlimited). The Gaussian process is useful for statistical modeling since this process utilizes the inherent advantages of the normal distribution.

Several artificial intelligent and statistical methods have been used for forecasting the groundwater quality [6].

Among them, several studies have been carried out on the use of neural networks. Daliakopoulos, Coulibaly [7] used various structures of artificial neural networks to predict the groundwater level. The results showed that the Levenberg-Marquardt model has a higher accuracy in prediction. Lallahem, Mania [8] used an artificial neural network to determine the time needed to estimate the groundwater level of a piezometer. The results showed that the multilayer perceptron neural network performed the most appropriate simulation with minimal latency.

Chandramouli, Lingireddy [9] provided a benchmark for determining the number of repetitions for the training of post-propagation neural networks. This study showed that training much or less than the amount could lead to passing the appropriate amount or failing to get the optimal response to determine the relationship between input and output data. Mishra and Singh [10] proposed a model for climate impact on severity-area-frequency (SAF) for annual droughts in the Kansabati River, India. Chronological droughts have been compared with the historical curves of SAF based on predicted rainfall using the general circulation model under uncertainty. The Downscaling method, based on the Bayesian Neural Network (BNN), was used to obtain results. Standardized rainfall indicators are used as drought indicators for the construction of the SAF curve for two periods (2000–2050 and 2051–2100). The results show that the probability of severe droughts in the period from 2001 to 2050 is more historically significant.

Sreekanth, Geethanjali [11] showed that the use of a neural network with a standard grid model, and the training of the Levenberg-Marquardt (LM) algorithm for predicting groundwater level, is a suitable model. Mohanty, Jha [12] used an artificial neural network model for predicting groundwater levels, and it was found that although groundwater surface prediction accuracy decreases with increasing the examination time period, the underground water level prediction was more acceptable for long period of time. Banerjee suggested an artificial neural skeleton model as an alternative to predict the salinity of groundwater. Hosseini-Moghari, Ebrahimi [13] used the fuzzy water quality index (FWQI) method to assess groundwater quality according to the water quality index.

Supreetha, Nayak [14] examined the ability of a hybrid ANN model and genetic algorithm (GA) to predict groundwater levels of several observation wells. The groundwater level for a ten-year period and rainfall data were used during the same period for model training. The groundwater surface prediction model was developed using an artificial neural network. GA was also used to determine the optimized weight of the artificial neural network. The study showed that the ANN-GA model could successfully be used to predict groundwater levels in observation areas. Additionally, the comparative study showed that the combined ANN-GA

model was more effective than the traditional ANN back-propagation approach.

Sakizadeh [15] have used ANNs for predicting the Water Quality Index (WQI) over the years 2006 to 2013. Yoon, Hyun [16] proposed a weighted error function to improve the performance of ANN and support vector machine (SVM) recursive prediction models for predicting long-term groundwater abundance. The results showed that weight-dependent errors could improve the stability and accuracy of recursive prediction models, especially for the ANN model. Shamsuddin, Kusin [17] used an ANN model to predict groundwater levels in two vertical wells in the aquifer area near the Langat River. According to the results of this study, accurate forecasts can be made with time series 1 day ahead of the predictive levels of groundwater and interactions between the river and the aquifer.

Nourani [18] implemented the Emotional ANN (EANN) as a new generation of artificial intelligence-based models for modeling the daily rainfall-runoff (R-R) prediction of the basin. It has been mentioned that inspired by the brain neurophysiology, in addition to normal weight, an EANN contains simulated parameters to improve the network learning process. A general comparison of the results of the EANN against Feed Forward Neural Networks (FFNN) showed that the training and productivity measures were superior up to 13 and 34%, respectively. The superiority of EANN over classical ANN was based on the ability to detect and identify dry (rainless days) and wet (rainy days) study locations using artificial hormonal parameters of the system. Ahmad, Rahim [19] designed a dynamic prediction method for the estimation of water quality by excluding the biological and chemical oxygen demands. They employed feed-forward ANN to predict the water quality indicators in Perak River basin Malaysia. The empirical results indicated that the proposed single feed-forward ANN can predict water quality with R^2 and mean squared error (MSE) of 90 and 17%, respectively. The outcomes show that the combination of multiple neural networks (forward selection and backward elimination) can further enhance the performance and accuracy of the prediction model. The generated result has the higher accuracy in terms of R^2 and MSE of 93 and 11.5%, respectively. Qaderi and Babanejad [20] developed an ANN-based forecasting model to study the groundwater quality which results in the accurate prediction of the costs of drinking water treatment. The historical data of the groundwater quality was collected during a 13-year period in Ilam, Iran. The ANN method trained by the Levenberg-Marquardt algorithm. The outcomes were compared against the RMSE through ROSA until the most effective model was designated. According to the obtained results, the ANN structured with three hidden layers and 18 neurons was the best-performing model which results in cost reduction by about 150,000 dollars daily. Charulatha, Srinivasalu [21] employed different regression and ANN

models for quality assessment of groundwater pollutants. The study was conducted to detect of nitrite ion concentration for the potential pollutants in the groundwater. Using Nash–Sutcliffe efficiency test, the generated result indicated that the proposed prediction model has superior performance in the validation step. Sahoo, Patra [22] proposed Bayesian and entropy methods for water quality assessment. The Aggregative Index Evaluation method was implemented to calculate water quality index in the Brahmani River. The trends of water quality indicators were predicted using Bayes' rule. The outcomes indicated that the WQI was enhanced during dry periods than during wet periods as a result of the reduction of pollutants.

He, Liang [23] proposed a spatial-temporal analysis model to investigate the hydrological drought characteristics of drainage basins in a region in southern China. In this paper, features such as the severity and frequency of hydrological droughts and patterns of spatial evolution in the drainage basins are studied. According to the results of the research, hydrological droughts have risen from the 1970s to 2010. This incremental trend was mainly due to the gradual reduction of hydrological droughts from the 1970s to the 1990s and its gradual escalation from 2000 to 2010. Deng, Chen [24] studied the spatial and temporal features including rainfall and drought effects based on temperature and precipitation data in 48 weather stations from 1959 to 2012 in the Pearl River Basin in China. The probable effect of events indicates that annual and seasonal rainfall has decreased slightly in most areas and annual and seasonal precipitation has decreased in some areas; monthly rainfall has irregular distribution, but no significant trends have been identified.

The literature review suggests that various techniques of artificial neural networks are widely used in predicting groundwater level variations in different water sources, despite different results. Due to the existence of nonlinear and indefinite relations, the use of artificial neural network model is recommended as against for Markov chain models. With the help of this model, it is possible to predict and describe the relationships between components and system parameters that are not well described but need to be analyzed and simulated.

It should be noted that the non-random process in a natural phenomenon can occur for various reasons, which in any case leads to errors in the output of the prediction model and, as a result, make difficulties to take managerial decisions. As a solution in deductive discovery, the best way to estimate the future values of a hydrological phenomenon is to use the Gaussian classification method. Other reasons for using this methodology are summarized below:

- To reduce the uncertainty caused by the extraction (prediction) of the water surface in aquifers and the total annual precipitation by statistical methods
- Establishing macro decision-making platforms by validating the results of that prediction

The application of artificial neural network models has been further enhanced to increase accuracy; however, the application of other methods such as time series and hybrid models has also led to improved output and error reduction in some cases.

The analysis provided in the present study presents an original contribution to the literature in two respects: First, although there have been previous studies that address climate drought analysis (e.g., [25]), no focused and concrete analysis dealt with large-scale association analysis of climate drought and a decline in groundwater quality. Recognizing that limited attention has been devoted to [drought and regional hydrologic variation](#) in uncertain environment, this research aims to obtain patterns for changes in the water level of aquifers in dry and semi-arid areas. The effects of water quality decline in Iranian aquifers on the occurrence of observed droughts have been investigated using SPI and GRI. Second, this paper proposes a Gaussian Process Classification approach that provides a solution for the climate prediction model. As presented in methodology Section, rather than adopting the simple probability distributions which are typical and custom view in previous studies, this paper accounts for the machine learning techniques such as Back propagation (BP) to reduce the uncertainty associated meteorological observations (water surface in aquifers and the total annual precipitation). To the best of our knowledge, this study is among the few articles that addressed the large-scale association analysis of climate drought and a decline in groundwater quality using Gaussian process classification methods in different time scales.

Research objectives

In this study, the objective is to conduct an association analysis of climate drought and decline in groundwater quality using GPC and back propagation (BP) artificial neural network. The methodology provides a decision support framework for managing uncertainty in the area of water resource planning. The calculations here are based on the data obtained from the previous calculations of rainfall data and water levels in the aquifer of the country and in the period 2004 to 2015 for the maximum of synoptic stations and observation wells in Iran. This model enables to analyze spatial-temporal drought characteristics using data of underground water levels recorded from 2017 to 2021 in 609 study plains in Iran. The artificial intelligence techniques have been programmed in the Python language designed to predict an annual precipitation rate. The probabilistic effects of water quality decreases in Iranian aquifers on the occurrence of observed droughts have been investigated using SPI and GRI. The outcomes of the prediction

model have been validated using a statistical summary of the rasterized cells of zoning maps. At the end of this study, it is expected that higher risk aquifers, as well as certain areas of Iran that are exposed to severe drought stresses, will be detected with the lowest overall error.

Materials and methods

The study area is the range of Iran (25 to 40 degrees north latitude and 44 to 64 degrees east longitude) and the zoning is based on peeling 609 plains in the country. The situation includes climatic variation and climate change in warm and dry areas in the middle, to wet and overcast areas in the northern regions; In addition, the unique geological structure and type of each aquifer are enclosed to half free and fully open. Due to the existence of these environmental conditions, the results of simultaneous studies of all aquifers are associated with various difficulties. In areas with higher population density, like the western regions of Iran, underground aquifers are directly affected by human factors. Quantitative changes followed by the quality of aquifer feeding waters, while in the eastern and central regions, generally, the river's water levels have little effect on groundwater levels. In addition, it impossible to simultaneous studies of all aquifers in the aquifer section, because of the unique geological form and structure of each aquifer (enclosed to semi-free and free-form). Figures. 1 and 2 show the position of Iran's plains and simultaneously the position of synoptic stations and observation wells in the country.

Methodology

Back propagation (BP) is a technique that is processed in artificial neural networks to calculate the contribution of each neuron error after a batch of data. This is a special case of the older and more general method named automatic differentiation. The list of symbols and notations used in this paper is provided in Table 1. In the learning context, BP algorithm is typically adjusted using the gradient calculation of the loss function. This method also sometimes has the back propagation of errors because the error in the output is computed and distributed through network layers [26].

The purpose of any supervised learning algorithm is to find a function that maps the set of inputs to its correct output. The main purpose of the back propagation is to calculate partial derivatives or gradients, $\partial E / \partial w$ of a function based on the loss of E according to any weights in the network [27]. The artificial neural network BP has the following definitions:

Fig. 2 Observational wells in the country's plains

Table 1 The basic notations used in this paper

Notations	Description
E	Error functions
x	Input vector
y	Output vector
w	Weight vector
g	The sigmoid function
$a_0^{(2)}$	
$a_1^{(2)}$	
$a_2^{(2)}$	
$\Theta_{10}^{(2)}$	
$h_{\Theta}(x)$	

$$a_0^{(2)} = g\left(\Theta_{00}^{(1)}x_0 + \Theta_{01}^{(1)}x_1 + \Theta_{02}^{(1)}x_2\right) = g(\Theta_0^T x) = g\left(z_0^{(2)}\right) \quad (1)$$

$$a_1^{(2)} = g\left(\Theta_{10}^{(1)}x_0 + \Theta_{11}^{(1)}x_1 + \Theta_{12}^{(1)}x_2\right) = g(\Theta_1^T x) = g\left(z_1^{(2)}\right) \quad (2)$$

$$a_2^{(2)} = g\left(\Theta_{20}^{(1)}x_0 + \Theta_{21}^{(1)}x_1 + \Theta_{22}^{(1)}x_2\right) = g(\Theta_2^T x) = g\left(z_2^{(2)}\right) \quad (3)$$

$$h_{\Theta}(x) = a_1^{(3)} = g\left(\Theta_{10}^{(2)}a_0^{(2)} + \Theta_{11}^{(2)}a_1^{(2)} + \Theta_{12}^{(2)}a_2^{(2)}\right) \quad (4)$$

In Eq. (1), g is a sigmoid function that refers to a particular state of the logistic function and is defined as Eq. (2).

$$g(z) = \frac{1}{1 + e^{-z}} \quad (5)$$

The learning algorithm can be divided into two stages: 1) release and 2) update of weights [28]. The update process includes the removal of a gradient of weight. This percentage affects the speed and quality of learning, which is also called learning rate. Stages 1 and 2 are repeated until the network performance is satisfactory. For the output unit ($L = 3$), if an error from node j in layer l is denoted by $\delta_j^{(l)}$, the actual value activation is equal to:

$$\delta_j^{(3)} = a_j^{(3)} - y_j = h_{\Theta}(x) - y_j \quad (6)$$

If the vector format is used, then:

$$\delta^{(3)} = a^{(3)} - y \quad (7)$$

$$\delta^{(2)} = \left(\Theta^{(2)}\right)^T \delta^{(3)} \cdot g'\left(z^{(2)}\right) \quad (8)$$

It should be noted that the condition $\delta^{(1)}$ does not exist because the input layer is considered as observational values

and is used as a training set. Therefore, there are no errors with input [27]. Correspondingly, the derivative of the cost function can be expressed as Eq. (6).

$$\frac{\delta}{\delta \Theta_{ij}^{(l)}} J(\Theta) = a_j^{(l)} \delta_i^{(l+1)} \quad (9)$$

This amount is used to update the weight and also the training rate can be multiplied by the weight adjustment.

Gaussian process classification (GPC)

In principle, a Gaussian process is the statistical distribution of X_t with $t \in T$, which for any finite number of linear combinations of samples has a Gaussian common distribution. More precisely, any linear function applied to X_t results in a Gaussian distribution. It can be written as $X \sim GP(m, K)$, which means that the random function X has a Gaussian process distribution with the mean function m and the covariance function K [29].

As another definition, it is a continuous process in Gaussian time if and only if for each finite set of indices $t_1, \dots, t_k$ exist in the index set T of Eq. (7).

$$X_{t_1, \dots, t_k} = (X_{t_1}, \dots, X_{t_k}) \quad (10)$$

The marginal distribution $P(y_i)$ of one component y_i is Gaussian. In addition, the joint marginal distribution of any subset of the components is multivariate-Gaussian [30]. Using the characteristic function of random variables, the Gaussian character can be expressed as follows:

Set $\{X_t; t \in T\}$ is Gaussian if and only if for each finite set of real-value indices $t_1, \dots, t_k$ that exists μ_1, σ_{1j} ($\sigma_{1j} > 0$) so that Eq. (8) holds for all $\{s_1, s_2, \dots, s_k \in \mathbb{R}\}$:

$$E\left(\exp\left(i \sum_{l=1}^k s_l X_{t_l}\right)\right) = \exp\left(-\frac{1}{2} \sum_{lj} \sigma_{lj} s_l s_j + i \sum_l \mu_l s_l\right) \quad (11)$$

In fact, Gaussian processes (GPs) are a commonly used learning approach designed to solve regression and probabilistic issues [29]. The Gaussian process can be used as a probabilistic distribution of functions on the Bayesian inference [31]. Assuming a set of N points in the domain of the desired functions, a covariance matrix of the multivariate Gaussian parameter that is identified by the Gram matrix of the N points with some desired kernel functions is an example of Gaussian processing.

Gaussian process classification (GPC)

In the classification methods, some of the popular kernel classifications can be called as a support vector machine (SVM), a

Bayes Point Machines (BPM), and a Gaussian Process Classification (GPC). In contrast to these two methods of classification, the GPCs approach, as mentioned, is a Cornell classification, which was developed by prior Gaussian's processing functions that were originally for regression [32]. The term Gaussian Process Classifier implements Gaussian processes (GPs) for classification purposes, and more specifically for probabilistic categorization, in which test predictions are categorized as probabilities. In principle, the classification of the Gaussian process in its mechanism employs a Gaussian process in the hidden function f , which then extends through a resulting binding function to obtain a possible classification of processing. The function of hidden function f is a so-called extra function, whose value is not observations, and is not related to this function itself. The purpose is to allow the development of a suitable formula in the model and to remove this function (integration process) when predicting probabilistic results.

Contrary to the regression setting, since the former Gaussian probability for the class label is inappropriate, the formerly hidden function f does not exist even for a Gaussian structure GP. Instead, a non-Gaussian probability related to the logit function is used. The Gaussian process classification provides the non-Gaussian approximation with a Laplace-based Gaussian method [31].

In GPCs, the likelihood of belonging to a class in a particular input location is uniformly related to the value of the hidden functions in that location. Starting from a previous value on this hidden function, the data are used for deduction in the hidden function and the determination of the value of the sub-parameters of the function in different aspects. GPCs can be viewed graphically with hidden variables.

It is assumed that the GP average is zero from the beginning. Primary covariance is determined by passing a component of the kernel function. The kernel function metadata is optimized for connections of the Gaussian process classification method by optimizing the marginal log-likelihood (LML) based on the specified optimizer. As the LML may have some local optimization, the optimizer can be repeated over and over again by specifying a re-optimizer. In the first run, which always begins with the initial value of the metadata, the kernel function is executed; then the metadata whose values are randomly selected will be calculated from a range of allowable values. It should be noted that if the initial metadata are kept constant, then they cannot be translated as optimizers [32].

Assuming that a data set (D) of x_i data points with binary class labels is as follows:

$$\begin{aligned} y_i &\in \{-1, 1\} : D = \{(x_i, y_i) | i = 1, 2, 3, \dots, n\} \\ X &= \{x_i | i = 1, 2, 3, \dots, n\} \\ Y &= \{y_i | i = 1, 2, 3, \dots, n\} \end{aligned} \quad (12)$$

Given the values of this dataset, the goal is to find the correct class label for new values of $\tilde{x}$ data points. This is done by calculating the probability of the class $P(\tilde{y}|\tilde{x}, D)$.

Graphical representation of GPCs with n data and 1 test data is illustrated in Fig. 3. x_i and y_i are observable data, $\tilde{x}$ is given, $\tilde{y}$ is predicted, f_i and $\tilde{f}$ are the hidden function and the Gaussian connection, respectively. It is assumed that the class label is obtained by transferring some values of the hidden variables $\tilde{f}$. The values of some hidden functions $f(\cdot)$ are evaluated in $\tilde{x}$ [33]. The former maximum Gaussian processing is placed on this function, which means how many evaluated points of the function will have a multivariate Gaussian density. It is assumed that this Gaussian former processing is parameterized by Θ which is called meta-parameters. The probability Θ can be written as Eq. (10).

$$P(\tilde{y}|\tilde{x}, D, \Theta) = \int P(\tilde{y}|\tilde{f}, \Theta) d\tilde{f} \quad (13)$$

The second part of the equation is obtained by an additional integral over $f = [f_1 \ f_2 \dots f_n]$. The value of the hidden function at the data points is obtained from Eqs. (11) to (13).

$$P(\tilde{f}|D, \tilde{x}, \Theta) = \int P(\tilde{y}|\tilde{f}, \Theta) P(\tilde{f}|D, \tilde{x}, \Theta) d\tilde{f} \quad (14)$$

$$P(\tilde{f}|\tilde{x}, f, \Theta) = \frac{P(\tilde{f}, f|\tilde{x}, \{x_i\}, \Theta)}{P(f|\{x_i\}, \Theta)} \quad (15)$$

$$\begin{aligned} P(f|D, \Theta) &\propto P(Y|f, X, \Theta) P(f|X, \Theta) \\ &= \left\{ \prod_{i=1}^n P(y_i|f_i, \Theta) P(f_i|X, \Theta) \right\} \end{aligned} \quad (16)$$

The first condition for each observation class is the specified value of the hidden function, and the second condition is the previous Gaussian processing on the evaluation functions

Fig. 3 Graphical representation of GPCs

in the data $\tilde{x}$. An approach for the condition $P(y_i | f_i, \Theta)$, which correlates $f(x_i)$ uniformly to the probability $y_i = +1$, is equal to:

$$P(y_i | f_i, \Theta) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y_i f(x_i)} \exp\left(-\frac{z^2}{2}\right) dz = \text{erf}(y_i f(x_i)) \quad (17)$$

By rewriting the dependence of f on the virtual x , the prior Gaussian processing functions can be written as Eq. (15):

$$P(f | \Theta) = \frac{1}{(2\pi)^{\frac{n}{2}} |C_\Theta|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} (f - \pi)^T C_\Theta^{-1} (f - \pi)\right) \quad (18)$$

In which the mean μ is generally assumed to be equal to the zero vector, and any condition of a covariance matrix C_{ij} is a function of x_i in x_j . Generally, the probability of a class by integrating on the meta-parameters is obtained by Eq. (16) [34].

$$P(\tilde{y} | \tilde{x}, D, \Theta) = \int P(\tilde{y} | \tilde{x}, D, \Theta) P(\Theta | D) d\Theta \quad (19)$$

The artificial neural network model was used in order to analyze the statistical distribution of drought and groundwater level [35]. The probability distribution of SPI and GRI are obtained in the period from 2017 to 2021 for all the country's plains. The Gaussian Process classification method was implemented in the Python coding environment. The code was executed based on the hidden function of Eq. (17):

$$z = 5 - y - \left(\frac{1}{2} (x^2)\right) \quad (20)$$

Figure 4 shows the 3D structure of the hidden function of the Gaussian process in a probabilistic classification for space $z > 0$. The intersection point of this function with the z -axis of magnitude 5 and its collision point with the x -axis is approximately 3 times the standard deviation of the normal distribution. Considering that the final results of the climate-drought and underground water indicators are standardized, using this function, we can analyze how far each third

dimension of the prediction is different from the Gaussian distribution as the probability of occurrence. In the case of assigning each regular pair to the positive class, it can be confirmed that the occurrence is probable by confirming the trend of the effect of climate drought on the reduction of groundwater level.

Results and discussion

Artificial neural network

Python code and the BP artificial neural network is employed to predict the annual rainfall and groundwater levels in the country's aquifers from years 2017 to 2021. According to the aforementioned descriptions in section three, it is required to define the format of input-layer digits. Each time series in 362 synoptic stations and 11,383 observational wells of the country are separated into n rows and m columns of the matrix, in order to provide the maximum statistical relation in the input layer column. The order of producing random variables is used in each implementation of the Python code of the artificial neural network post-marketing. The results of the code for the mean of data are normalized between 0 and 1. For all 362 synoptic stations and 11,383 observation wells in the implementation of the forecast period, the approximate random velocities of simulation training are shown in Table 2.

Finally, by converting the normal output figures to the base values, these numbers are zoned in the GIS using the optimal method for both the individual predictions of the annual accumulated rainfall and the groundwater levels of the country's aquifers.

The spatial variations obtained in the output of zoning of the prediction interval for SPI and GRI are illustrated in Figs. 5 and 6. The results show the occurrence of drought with a constant trend in a specific region. Considering the four-year

Fig. 4 Three-dimensional hidden function view in Gaussian Process classification

Table 2 Bounds weights for artificial neural network

	Weight number	2016	2017	2018	2019	2020	2021
SPI	1	58.390	58.510	53.652	53.770	53.150	52.713
	2	50.147	47.553	47.585	52.279	51.806	51.902
	3	52.050	47.524	48.201	48.023	47.750	48.067
	4	40.192	40.615	40.426	42.644	42.779	42.796
	5	46.308	46.392	47.122	47.266	48.041	48.082
	6	37.613	37.757	37.664	37.717	37.754	37.820
	7	41.856	43.107	38.491	37.696	38.092	38.127
	8	39.247	39.256	39.295	39.294	39.245	39.248
	9	44.229	45.054	45.979	45.968	45.970	46.181
GRI	1	97.998	99.222	98.740	98.744	98.478	98.478
	2	88.949	89.331	89.734	89.780	89.692	88.858
	3	73.387	74.639	74.703	74.719	74.732	74.728
	4	67.532	69.057	70.311	70.427	73.131	71.118
	5	79.979	80.338	80.421	76.228	75.235	75.305
	6	77.040	79.963	80.762	81.043	78.122	78.131
	7	93.084	93.315	93.346	93.411	93.402	93.307
	8	84.121	84.130	72.771	74.652	74.689	74.904
	9	92.260	92.385	92.501	93.004	93.418	93.524

training interval used in BP artificial neural network, the prediction period can have the least uncertainty. According to the obtained results, the least drought was observed in three areas with the northeastern-southwest structure: 1) in the northwest of the country in the Azerbaijan region, 2) in the northern region of the country, and 3) in the south of Khorasan to the south region of the country. Other parts of the areas belong to the drought class that is commonly known. This situation indicates that with the change in the southern parts of the country, the main drought will return to the same areas in 2016. In the year 2017, just as occurred in 2016, the only difference is the decline of the northern regions of Iran, which occurs in the least SPI drought class, and thus a wider area of the country will suffer from drought in this year. In the three consecutive years, 2019, 2020, and 2021, the largest droughts will continue to occur in the southern regions of Iran. Despite the relative decline of drought in the northeast of the country in 2019 and extending it to vast areas of northern Iran, the country will continue to face drought stress in 2021. In total, from the point of view of spatial distribution in the forecast period, the southern regions of Iran and especially the mid-sectional areas of the study plains are exposed to the most severe climate droughts more than other areas.

According to GRI drought prediction maps, the least drought is expected in the central parts of the country. However, changes in the drought class are clearly observed in all regions without any correlation with a particular location. The wetness class in 2016 and in the five years of its forecast

was limited to the central and west areas. In 2017, the occurrence of groundwater droughts is predicted in many parts of the country. The observed trend over the next four years is also very similar to the initial two years. Indeed, the trend of increasing and decreasing relative droughts of water levels in aquifers is predicted in the form of a sinusoidal function during each two-year period. However, in all periods, the two central parts of the country and the middle regions of the eastern coast, in the south of Khorasan province, will still be protected from drought. This can be attributed to the absence of specific aquifers in these areas, as well as to the difference in the nature and structure of the aquifer in relation to mountainous regions. In the western regions of Iran, the rate of exploitation of underground water is much higher than the eastern regions and especially the central parts of the country due to the development of the agricultural sector and the high population density.

In short, it can be concluded that despite the unchanged conditions of exploitation and according to the prediction maps of the GRI, the severity of drought is reduced compared to the observation period. The reason for reducing the occurrence of droughts in the forecast period is the repetition of a randomized process of exploitation of the permanent potential of aquifers during the 22-year period. In other words, from 2017 to 2021, a large portion of groundwater aquifers is expected to be unable to meet the water needs. Thus in practice, with the change in water harvesting resources for agriculture, this phenomenon will intensify and relative droughts will also decrease.

Fig. 5 GRI calculated in theyear 2021

Fig. 6 SPIcalculated in theyear 2021

Fig. 7 Two-dimensional map of hidden function classification by Gaussian process method

Validation by Gaussian process classification

In the Gaussian process method, the hidden function is defined in Eq. (17). This is a function of the integer classes in which the input is the pair of a statistical average of the value of each SPI and GRI for each plain. Accordingly, the average value of zoning indicators in each plain area and each period are obtained in the GIS environment. The average SPI for each year is selected as the x variable and the mean of the GRI for each year is regarded as the y variable. The third dimension variable is calculated by Eq. (17) as mentioned before. The z dimension divides the class of each regular pair (x, y) into two

distinct parts smaller than 0. In this learning model, the training points are in pairs (x_i, y_i) in which the samples are given with their identifiers; and i is the index of each sample in a set of training points. The goal of this learning is to obtain the function f , which can return the appropriate class value for the other input samples ($f(x) = y$). The Eq. (17) is of a quadratic type with a Gaussian structure. The importance of classes in the conjecture of each pair of data is the Gaussian statistical distribution. With respect to the normal basis of the Gaussian function, if the value of the third variable is negative, then each pair of data is assigned a red color. However, the examination of the validity of each paired membership belonging to

Fig. 8 Continues membership probability obtained by GPC method in 2016

Fig. 9 Continues membership probability obtained by GPC method in 2019

each class is conditional probability of the membership calculated for that point position.

Figure 7 shows the two-dimensional model of the hidden function of classes with the Gaussian distribution. The limits of L here are about 3 times of the standard deviation of the normal distribution. Each pair of data is classified into two positive and negative probability classes. The possibility of the orientation of the probability lines here depends on the Gaussian learning during the extraction of the kernel function. For the period 2016 to 2019, the lines with the same probability with negative z -data are depicted in Figs. 8 and 9 and are obtained at three sizes 0.334, 0.500 and 0.666 (Figs. 10 and 11).

The continuous value of the non-membership class (red points) is equivalent to the estimate $P[G(x) < 0]$. These probability values are obtained in the form of raster cells of 50×50 units and with an incremental value. In this case, each regular pair in each cell has a certain probability value of its membership to the negative class. By examining the results, it was found for all points the bounds are intended to be around 0%. Correspondingly, as shown in Figs. 12 and 13, all points in the positive class are obtained. Based on the continuous probabilistic structure and the average output of the indices, it can be concluded that the prediction model has been able to accurately predict the allocation of significant points of the points to the positive class. In addition to validating the

Fig. 10 The third dimension's distribution in a positive class of kernel function (2020)

Fig. 11 The third dimension's distribution in a positive class of kernel function (2021)

predicted variables for the two indicators in the future, we can substantially confirm the effect of climate drought on groundwater in a specific and nonlinear relationship. It should be noted that this relationship is not in linear conditions, and each aquifer independent of the other aquifer has its own unique complexity, such as geological structure, the way and the nature of the operation, the time and resources of water extraction. However, due to the dry climate of Iran, rain is the main source of aquifer nutrition. As a result of negative changes, the average amount of rainfall is an important factor

in reducing the groundwater level, although it has been associated with seasonal delays. Table shows the main kernel functions generated in the probabilistic classifying step by the Gaussian process. The parameters of the kernel functions have been optimized during fitting of Gaussian Process by maximizing the log-marginal-likelihood (LML). Gaussian Process uses these kernel functions to define the covariance of a prior distribution over the target functions and uses the experiential training data to term a probability function (Table 3).

Fig. 12 The average of continues membership probability obtained by GPC method

Fig. 13 The average of continues membership probability obtained by GPC method

Conclusion

In this study, the relationship between climatic drought effects on groundwater resources as the most important factor in water quality changes in aquifers was investigated. The association analysis of climate drought and a decline in groundwater quality was performed using Gaussian process classification and back propagation artificial neural network. Underground water levels in 609 study plains in Iran were used to predict drought over the test period extending from 2017 to 2021. The artificial intelligence methods were implemented in the Python coding environment to achieve an annual precipitation rate. The Gaussian Process modeling was used for validating the results and confirming the effect of climate drought with the SPI on the occurrence of observed droughts with GRI

Table 3 The main kernel functions generated in the probabilistic classifying step by Gaussian process

Year	Educational core kernel function
2016	Learned kernel: $41.8 \times 2 * \text{DotProduct}(\sigma_0 = 3.98) ** 2$
2017	Learned kernel: $43 \times 2 * \text{DotProduct}(\sigma_0 = 4.58) ** 2$
2018	Learned kernel: $16.2 \times 2 * \text{DotProduct}(\sigma_0 = 4.46) ** 2$
2019	Learned kernel: $31.2 \times 2 * \text{DotProduct}(\sigma_0 = 4.93) ** 2$
2020	Not having at least two classes
2021	Not having at least two classes
Average	Learned kernel: $4.28 \times 2 * \text{DotProduct}(\sigma_0 = 4.46) ** 2$

calculations. In addition, a statistical summary of the Rasterized Cells of zoning maps was used in order to validate the prediction results. Considering the relationship between water quality reductions in Iranian aquifers due to the occurrence of groundwater drought periods, the results were validated by analysis of the effect of climate drought using SPI on the occurrence of observed droughts with GRI. The results are well illustrated by the observation of the predicted digits in the third dimension of the Gaussian distribution.

Artificial Neural Network prediction zoning for the SPI drought index indicates that from the perspective of spatial dispersion over the prediction period, the southern regions of Iran, and especially the middle parts of the plain, are exposed to the most severe future droughts more than other areas. In short, it can be concluded that despite the unchanged conditions of exploitation and according to the prediction maps of the GRI, the severity of drought is reduced compared to the observation period. The reason for reducing the occurrence of droughts in the forecast period is the repetition of a randomized process of exploitation of the permanent potential of aquifers during the 22 years period. In other words, from 2017 to 2021, a large portion of groundwater aquifers is expected to be unable to meet the water needs. Thus in practice, with the change in water harvesting resources for agriculture, this phenomenon will intensify and relative droughts will also decrease.

According to the obtained results, the effect of climate drought on groundwater drought was confirmed. As a result, the causes of intensification of harvesting from underground

water sources, in addition to factors such as population growth, may also be the reduction of surface water reserves. In many aquifers, an increase in harvest could be due to an increase in the area under cultivation in the nearest aquifer area. Therefore, in addition to lack of nutritional resources and increased harvest, irrigation water returns to the aquifer has led to a reduction in the quality standard of drinking. According to the obtained results, it is possible to create a decision-making model for the proper management of aquifer, as well as to reduce the cost of decision making due to the uncertainties of the problem.

Compliance with ethical standards

Disclosure statement No potential conflict of interests was reported by the authors.

References

- Eberhard F, Hamawand I. Selective electrodialysis for copper removal from brackish water and coal seam gas water. *International Journal of Environmental Research*. 2017;11(1):1–11.
- Bajc AF, et al. *Evaluating the groundwater resource potential of the Dundas buried bedrock valley, southwestern Ontario: an integrated geological and hydrogeological case study*. *Can J Earth Sci*. 2017;999:1–18.
- Mosaferi M, Pourakbar M, Shakerkhatibi M, Fatehifar E, Belvasi M. Quality modeling of drinking groundwater using GIS in rural communities, northwest of Iran. *J Environ Health Sci Eng*. 2014;12(1):99.
- Nasrabadi T, Bidabadi NS. Evaluating the spatial distribution of quantitative risk and hazard level of arsenic exposure in groundwater, case study of Qorveh County, Kurdistan Iran. *Iranian Journal of Environmental Health Science and Engineering*. 2013;10(1):30.
- Nourani V, Ejlali RG, Alami MT. Spatiotemporal groundwater level forecasting in coastal aquifers by hybrid artificial neural network-geostatistics model: a case study. *Environ Eng Sci*. 2011;28(3):217–28.
- Zaltsberg E. Application of statistical methods to forecasting of natural groundwater tables. *Can J Earth Sci*. 1982;19(7):1486–91.
- Daliakopoulos IN, Coulibaly P, Tsanis IK. Groundwater level forecasting using artificial neural networks. *J Hydrol*. 2005;309(1):229–40.
- Lallahem S, Mania J, Hani A, Najjar Y. On the use of neural networks to evaluate groundwater levels in fractured media. *J Hydrol*. 2005;307(1):92–111.
- Chandramouli V, Lingireddy S, Brion G. Robust training termination criterion for back-propagation ANNs applicable to small data sets. *J Comput Civ Eng*. 2007;21(1):39–46.
- Mishra A, Singh VP. *Analysis of drought severity-area-frequency curves using a general circulation model and scenario uncertainty*. *J Geophys Res Atmos*. 2009;114(D6).
- Sreekanth P, et al. Forecasting groundwater level using artificial neural networks. *Curr Sci*. 2009;933–9.
- Mohanty S, Jha MK, Kumar A, Sudheer KP. Artificial neural network modeling for groundwater level forecasting in a river island of eastern India. *Water Resour Manag*. 2010;24(9):1845–65.
- Hosseini-Moghari S-M, Ebrahimi K, Azarnivand A. Groundwater quality assessment with respect to fuzzy water quality index (FWQI): an application of expert systems in environmental monitoring. *Environ Earth Sci*. 2015;74(10):7229–38.
- Supreetha B, Nayak PK, Shenoy NK. Groundwater level prediction using hybrid artificial neural network with genetic algorithm. *International Journal of Earth Sciences and Engineering*. 2015;8(6):2609–15.
- Sakizadeh M. Artificial intelligence for the prediction of water quality index in groundwater systems. *Modeling Earth Systems and Environment*. 2016;2(1):8.
- Yoon H, Hyun Y, Ha K, Lee KK, Kim GB. A method to improve the stability and accuracy of ANN-and SVM-based time series models for long-term groundwater level predictions. *Comput Geosci*. 2016;90:144–55.
- Shamsuddin, M.K.N., et al. *Forecasting of Groundwater Level using Artificial Neural Network by incorporating river recharge and river bank infiltration*. in *MATEC Web of Conferences*. 2017. EDP Sciences.
- Nourani V. An emotional ANN (EANN) approach to modeling rainfall-runoff process. *J Hydrol*. 2017;544:267–77.
- Ahmad Z, Rahim NA, Bahadori A, Zhang J. Improving water quality index prediction in Perak River basin Malaysia through a combination of multiple neural networks. *International Journal of River Basin Management*. 2017;15(1):79–87.
- Qaderi F, Babanejad E. Prediction of the groundwater remediation costs for drinking use based on quality of water resource, using artificial neural network. *J Clean Prod*. 2017;161:840–9.
- Charulatha G, et al. Evaluation of ground water quality contaminants using linear regression and artificial neural network models. *Arab J Geosci*. 2017;10(6):1–9.
- Sahoo MM, Patra KC, Swain JB, Khatua KK. Evaluation of water quality with application of Bayes' rule and entropy weight method. *European Journal of Environmental and Civil Engineering*. 2017;21(6):730–52.
- He Z, Liang H, Yang C, Huang F, Zeng X. Temporal-spatial evolution of the hydrologic drought characteristics of the karst drainage basins in South China. *Int J Appl Earth Obs Geoinf*. 2018;64:22–30.
- Deng S, et al. Spatial and temporal distribution of rainfall and drought characteristics across the Pearl River basin. *Sci Total Environ*. 2018;619:28–41.
- Valipour M. Optimization of neural networks for precipitation analysis in a humid region to detect drought and wet year alarms. *Meteorol Appl*. 2016;23(1):91–100.
- Zhang G, Patuwo BE, Hu MY. *Forecasting with artificial neural networks: the state of the art*. *Int J Forecast*. 1998;14(1):35–62.
- Schmidhuber J. Deep learning in neural networks: an overview. *Neural Netw*. 2015;61:85–117.
- Li, Y., et al. *The improved training algorithm of back propagation neural network with self-adaptive learning rate*. In *Computational Intelligence and Natural Computing*. 2009. CINC'09. International Conference on. 2009. IEEE.
- Kang F, Han S, Salgado R, Li J. System probabilistic stability analysis of soil slopes using Gaussian process regression with Latin hypercube sampling. *Comput Geotech*. 2015;63:13–25.
- MacKay, D.J., *Information theory, inference and learning algorithms*. 2003: Cambridge university press.
- Rasmussen CE, Williams CK. Gaussian processes in machine learning. *Lect Notes Comput Sci*. 2004;3176:63–71.
- Pal M, Deswal S. Modelling pile capacity using Gaussian process regression. *Comput Geotech*. 2010;37(7):942–7.
- Pasolli L, Melgani F, Blanzieri E. Gaussian process regression for estimating chlorophyll concentration in subsurface waters from remote sensing data. *IEEE Geosci Remote Sens Lett*. 2010;7(3):464–8.
- Rasmussen CE, Nickisch H. Gaussian processes for machine learning (GPML) toolbox. *J Mach Learn Res*. 2010;11(Nov):3011–5.
- Banerjee P, Singh VS, Chattopadhyay K, Chandra PC, Singh B. Artificial neural network model as a potential alternative for groundwater salinity forecasting. *J Hydrol*. 2011;398(3):212–20.