International Journal of Environmental Science and Technology (2022) 19:11397–11408
https://doi.org/10.1007/s13762-022-04356-9

ORIGINAL PAPER

Groundwater quality prediction based on LSTM RNN: An Iranian experience

D. Valadkhan[1] · R. Moghaddasi[1] · A. Mohammadinejad[1]

Received: 27 September 2021 / Revised: 4 June 2022 / Accepted: 15 June 2022 / Published online: 5 July 2022

Abstract
Groundwater quality prediction has practical significance for the prevention of water pollution. Based on the exogenous variables which are effective on water quality indicators, this paper proposes a new method with new effective parameters based on LSTM RNN for groundwater quality index prediction. The effective parameters on the groundwater quality index include rainfall rate, temperature, and humidity, and groundwater abstraction was collected. Monthly time series data selection was done from five different locations in the Damavand region in Iran from 2009 to 2021. Neural network architecture is tested by "f-score" tested to obtain the best neural network performance. A comparison of the real value and the result of the prediction show that the water quality index prediction has been done sensibly and quite properly in most cases.

Keywords Groundwater · Water quality · Prediction · Neural network · LSTM

Introduction

Developments of the world's population in the present century cause growing food demand, expansion of agricultural activity, increasing urbanization, and development of industries will make humans destroy the groundwater resources, more than ever (Madani 2014). Population growth, development irrigated farming, and raise up industrial formation are the main factors for water demand increases in the world (Wada et al. 2013). In arid regions with limitation of water resource, groundwater is almost indicated as the major water resource. If groundwater usage surpasses groundwater renewal, overutilization or obstinate groundwater depletion may happen (Gleeson et al. 2010). Decreasing levels of groundwater may have destructive effects on the natural hydrologic cycle and related ecosystems (Wada et al. 2010). The greatest difficulty of the world is the quality reduction in groundwater supplies because it is known as the major drinkable stock water resource for people expenditures and irrigation purposes (Kumar 2012). For the first time, the Ministry of the Environment of proposed the reaction of Water Quality Index (wqi) to evaluate the degree of water pollution (Hameed et al. 2016). Groundwater structures are dynamic that move smoothly from areas of recharge to areas of discharge. The aquifers are primarily unprotected and endangered resource stock, although in some cases they can be extracted each year and recharge by penetration of precipitation into the ground (Koundouri 2004). Because of these dynamical properties and limitation of groundwater resources, it is necessary to manage and predict the future situation that plays the main role to get the optimal management of resources. Water quality prediction is so similar to air quality prediction which has been studied by Chen et al. (Chen et al. 2020). We must use a flexible time-series prediction model according to the time series modality data of effective groundwater quality parameters. There are many algorithms for sequence modeling, which have been proposed recently. One of the earliest popular algorithms is the Hidden Markov Model (HMM). The arithmetical intricacy of HMM is too large, and if it is employed the dynamic programming, it will be an incomputable problem when there are too many situations. The Recurrent Neural Network (RNN) idea was introduced in the late 1980s. The RNNs are primarily used for the learning of time series data. It has end-to-end direction, a specific framework, and formalize methods such as a dropout mechanism, and a floating

Editorial responsibility: Binbin Huang.

✉ R. Moghaddasi
r.moghaddasi@srbiau.ac.ir

[1] Department of Agricultural Economics, Extension and Education, Science and Research Branch, Islamic Azad University, Tehran, Iran

degree of freedom to improve the over-fitting appearances. But the conventional RNNs have a Long-Term Dependencies problem. Hochreiter and Schmidhuber (1997) proposed an upgraded form of RNN and recently developed and promoted by Alex Graves. The new type of RNN resolves the problem of gradient disappearance or gradient burst and long-term memory deficiency of RNN. This effective model is called the LSTM RNN which is described below (Fang et al. 2020).

Hydrological cycle and effective parameters

Hydrologic sequence variables such as rainfall, surface flows, temperature, evapotranspiration, intrusion, and break into the soil moisture and aquifers are parts of nature's active ecosystem. The hydrologic feature of groundwater resources also is the main impressive attribute of groundwater quality analysis. Ground topology and sea level height of aquifer are effective indirect groundwater quality parameters. Because of ocean warming and loss of mass from glaciers and ice sheets, sea level rise (SLR) happened in the recent century. The sea level height also has a direct effect on rainfall, which consider an important variable. Because of SLR, the coastal areas were mainly destroyed by saltwater intrusion (Sreekesh et al. 2018). Groundwater is usually getting in subsurface generation known as aquifers, which may be an important hydrological component. Because aquifers may be connected, the accessibility and quality of the groundwater within them just be local issues, defined by both surface and subsurface surveys. The condition and characteristics of a given aquifer and its quality are determined by the hydrologic features. Because of the dynamic intricate nature of the groundwater feature, it becomes very hard to simulate groundwater quality and level responses. Numerical physically based models of groundwater circulation can simulate potential for detailed groundwater levels and reproduce complex hydrogeological settings, but require a massive amount of special information to understand the basic physical process and determine the hydrologic system, much of which is unknown and difficult to catch (Sahoo et al. 2015). Rainfall intensity in climate situations is one of the main hydrological parameters on groundwater resources in arid and semiarid areas. Infiltrated water that reaches the flows relatively penetrates streams as surface runoff. Surface flows usually permeate into the deep ground, finally join the springs and outside flows, or storing in the aquifers. The ground saves the penetrated water and expands the land moisture, then it moves to the ground aquifers if the soil is fully soaked. This process will happen smoothly as the belowground flows join the aqua circulation during the dry period. If no water provide frequently from penetration sources of recharge, the groundwater quality index would progressively fall down by decreasing the level of ground resource (Jan et al. 2007).

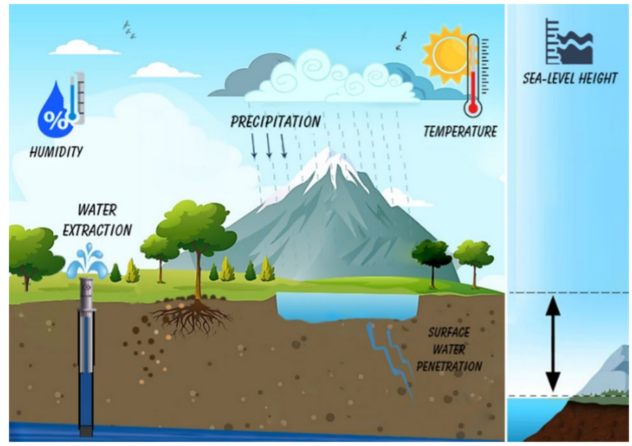


Fig. 1 Hydrological cycle and effective parameters on groundwater quality

Due to the affectivity of rainfall in the middle term, it can be known as an effective parameter to predict groundwater quality index in the research progress. Because most of the ground aquifers are regenerated by precipitation and runoff penetration, global warming on rainfall and external flows evaporation deplete ground aquifers in the long run. It is so clear that we are in a period of global warming which occurred by rising greenhouse gases and the persistent increase in carbon dioxide levels since the last century. The continuation of this event (global warming) may seriously modify worldwide and local climate specifications, especially increasing temperature. Warming up the weather also affects the hydrological elements by the vaporization of external flows and plant heat exhaustion. While climate variation influences external water resources through main variables such as heat, rainfall, and evaporation, the correlation between temperature enhancement and groundwater quality is so complicated to survey. The unpredictability of climate change causes concern about the supply and management of groundwater resources. And increasing of temperature plays an important role in groundwater re-management (Kumar, 2012). Groundwater resource is mostly used in the arid and semiarid area and agricultural section. In agricultural operations, many kinds of reputed and natural resources that are not easily renewable are used, depleted, and damaged, as well as fertile lands, soil structure, and water resources (Bacquart et al. 2015). Some effective parameters on groundwater quality in the hydrological cycle are shown in Fig. 1.

Groundwater quality prediction methods

GWQI[1] prediction is an impressive way to control and preserve groundwater storage that cautioned about future

[1] Groundwater Quality Index.

Table 1 Groundwater quality index classification

No	GWQI range	Groundwater quality classification
1	GWQI > 85	Excellent water
2	70.1 < GWQI < 85	Good water
3	45 < GWQI < 70	Normal water
4	30 < GWQI < 44.9	Poor water
5	15 < GWQI < 29.9	Very poor water
6	15 > GWQI	Unsuitable water

water quality state (Jiang et al. 2019). Groundwater quality prediction is one of the main research contents of water environmental problems and the basic work of groundwater assets operation and control. The groundwater quality index is impressed by different elements, which are not fully structured. Also, the systematic and unsystematic features specify the complication of groundwater quality prediction (Ting et al. 2018). Groundwater is mostly used in irrigation in the agricultural sector. So we used the irrigation water quality index classification developed by Meireles (Meireles et al. 2010). The water quality index grading is shown in Table 1.

Many researchers have studied water quality index prediction to assist an established managerial structure for managing groundwater pollution. The performance of neural networks in GWQI prediction has been acceptable. The effective parameters of GWQI prediction model are interrelated similar to a compound sequence with time lags. RNN presents the concept of "memory" as a neural network to cope with sequence problems, because it looks to generate sequences as the instrument of dealing with problems. The RNN can manage long-term dependence hitches. But, actually, as time goes on, RNN will recurrently lose the ability to learn the time-lagged information such as rainfall and temperature. To resolve the long-term dependence problem, the LSTM network presents some memory blocks to replace the implicit nodes in the cyclic. A typical LSTM network has a specific unit called "Forgotten Gate" which enables the network to learn to "forget" and bypass the overload. (Fang et al. 2020).

Using long short-term memory prediction model

Long short-term memory, which is called LSTM, is widely used in the prediction job. Fang et al. 2019 proposed a prediction model by using radar image data based on Long Short-Term Memory (LSTM) and Deep Convolution Generative Adversarial Networks (DCGAN) (Chen et al. 2020). Yan et al. (2020) introduced an improved method for fitting and predicting the number of COVID-19 confirmed cases based on LSTM (Yan et al. 2020). Wang et al. (2017) introduced a model based on LSTM NN for water quality prediction (Wang et al. 2017). Also, many same studies have been done on prediction jobs based on LSTM neural network. Due to the traditional water quality prediction models that cannot comprehensively reflect the influence of hydrologic parameters, meteorology, and hydraulics factors, scientists generally pay attention to improving the relevance and reliability of water quality prediction models. They introduced a variety of new technologies, such as stochastic models, fuzzy mathematics, 3S technology, artificial neural networks, and other models based on ANN such as LSTM RNN, to improve water quality prediction (Ping et al. 2019). In Damavand (the study area), Ministry of Agriculture and Groundwater Resources Management Organization collect the groundwater quality data, for every particular period of time by sampling. Their reports give the water quality variations at regular intermissions at the observing points. Also, the other effective data mentioned above have a time series modality, such as monthly temperature, rainfall, and humidity. Thus, water quality parameters are in the form of time-series data. To classify the database, they input it to the model in a multivariate string array format. With the appropriate performance of long short-term memory (LSTM) models in time-series forecasts, the request for LSTM in environmental research has become broader. LSTM achieves good forecasting performance because of its capacity to handle time-dependent datasets (Hossain et al. 2019). In the case of predicting the groundwater quality index, we faced natural time-series parameters such as temperature and rainfall, which are exogenous variables. Additionally, groundwater resources are used in agriculture and cultivation, so their kind of usage can have an important effect on groundwater quality. Applying LSTM RNN to predict the drinkable water quality is a bit different, but it can use to predict the drinking-water quality index, by customizing the time series effective parameters. Some studies have applied the LSTM models to the prediction of drinkable water quality index. For predicting environmental parameters, two major features highly recommend application of the neural network algorithm. First is data absoluteness, implying when network methods are performed for clarification, it will be much easier if the data is task-related, organized, and rich. The second feature is its effectiveness and appropriateness in describing complex connections between features and the target variable (Ye et al. 2020). To improve the performance of LSTM RNNs, we have to do more training on a model to increase the accuracy of results. The ranking of accuracy in the neural network is measured by "f-score" testing. It tests by trial and error model on the sample data, to identify the best neural architecture. So, apply LSTM with the "f-score" testing helps us to use this type of neural networks to predict the water quality index (Ping et al. 2019). The typical LSTM network is developed and applied on many time

Table 2 Sample bore wells information

No	Bore well location	Elevation above sea level (m)	Depth (m)	Most Utilization
1	Absard	1936	55	Garden irrigation
2	Seyed Abad	2400	30	Garden irrigation & Swimming pool
3	Garmabsard	1962	65	Garden irrigation
4	Sarbandan	2240	40	Garden irrigation
5	Aieenevarzan	2146	86	Garden irrigation & Swimming pool

series datasets such as the predicting number of infections and deaths of infectious diseases, operating image feature extraction and image restoration, forecasting weather, realizing the measurement of content and literature and predicting the citation time of biomedical literature, etc. (Yan et al. 2020).

Materials and methods

Study area

Damavand is located in the northeastern region of Tehran province. Its total area is 1932 km^2 within approximately 2000 m elevation above sea level. This region has arid and semiarid weather and a cold climate with an average of 400 mm of annual rainfall. Groundwater resources are the main source of water demand in the region, while there are more than 4,500 unauthorized wells in the region. There are many same areas in the world, with mentioned features, such as sea level height and semiarid climate, and groundwater quality challenges. This is the main reason for selecting the Damavand region for research analysis.

Data collection

A representative groundwater sample was obtained from five wells selected with an average depth of 30 to 90 m in several locations in the region. Table 2 shows the sample information.

Water quality data and underground resource usage information of mentioned resource were gotten from Damavand Water Department. Weather information was collected from Mosha weather station for a period from 2009 to 2021 (13 years).

Rainfall

Monthly rainfall rates were collected from 2009 to 2021 (13 years) in millimeters.

Temperature

The monthly temperature in centigrade was calculated as the average of maximum and minimum temperature in a month.

Air Humidity

There is a close relation between rainfall and humidity at the same time. When the rain is falling, it will increase the relative humidity because of the evaporation. The air where the rain is falling may not be full of water vapor. However, the longer rains cause more humidity because of the air constantly drawing the water Figs. 2, 3, 4 and 5.

Groundwater abstraction

Groundwater usage is dependent on some factors such as plant water requirement, air temperature, and age of plants in case of agricultural irrigation. It is dynamic on the growing season, fruiting period, and plant hibernation. Groundwater abstraction is the most important factor on reduced groundwater levels and decrease in resource quality. In this research, due to plant growth and change in irrigation patterns, and the other usage of water programs such as swimming pools, dynamic water extraction was collected. Figure 6 with five charts shows the groundwater abstraction in different locations in the Damavand region.

It noticed that parameters of rainfall, temperature, and air humidity are general and not limited to the specific location around the region, while the water extraction and index of quality of desired water well are related to mentioned location. The function model shows in Eq. 1.

$$\mathrm{GWQI} = f\left(R_i, T_i, \mathrm{Hu}_i, \mathrm{WE}_i\right) \tag{1}$$

where is GWQI is the groundwater quality index, R_i is the rainfall rate, T_i is the average temperature, Hu_i is air humidity and WE_i is the amount of groundwater abstraction.

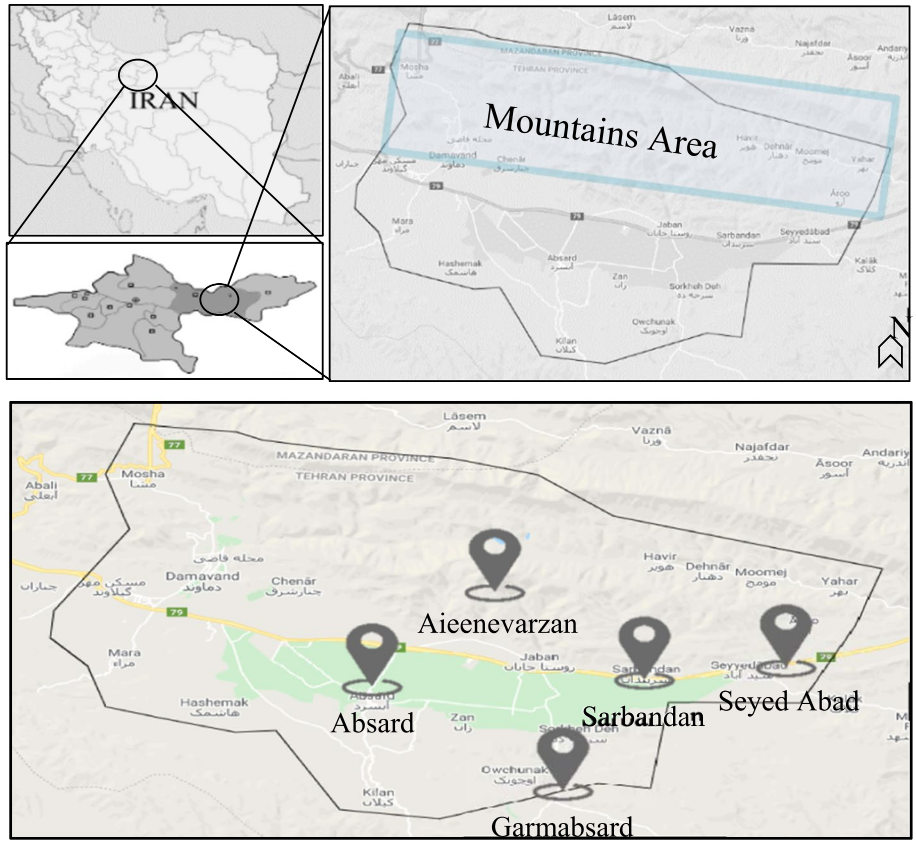


Fig. 2 Location of area and sample data

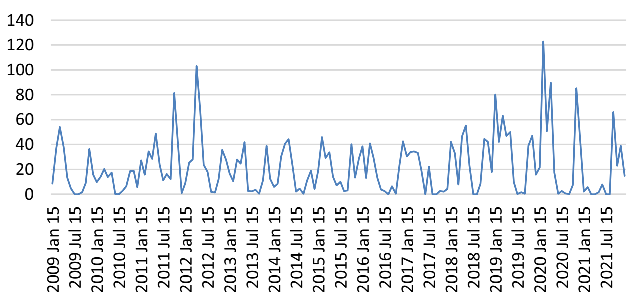


Fig. 3 Rainfall trend (millimeter)

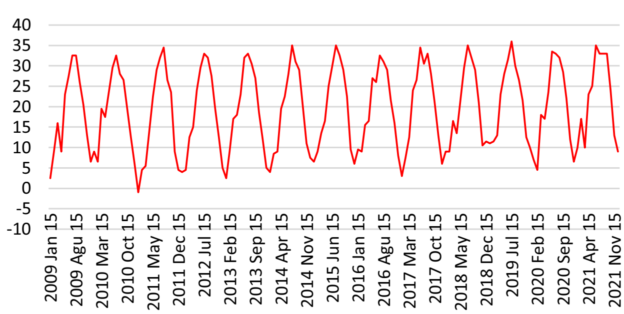


Fig. 4 Average temperature (degrees centigrade)

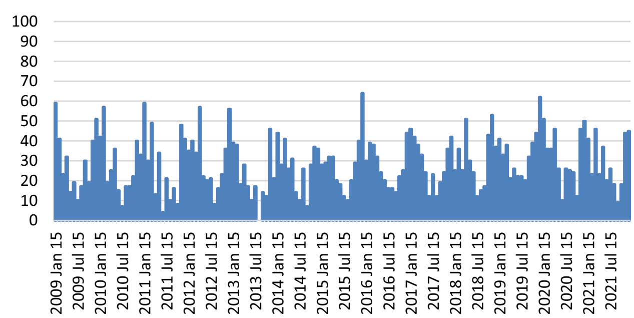


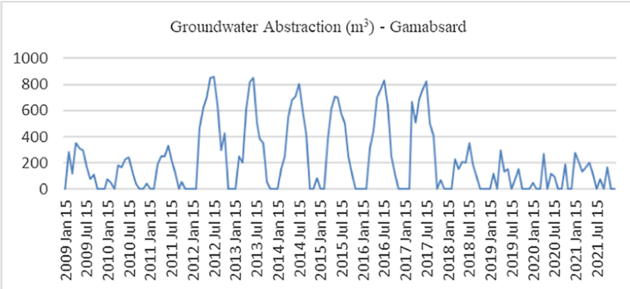
Fig. 5 Air humidity (%)

Data processing

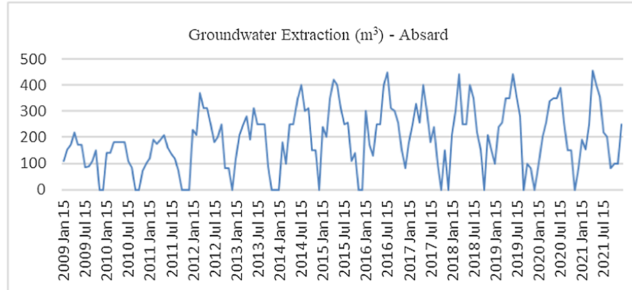
Normalization test

Normalization technique is resizing the dataset to a particular spacing so that the out-of-range data are detached and changed into a net void value so that can be compared and weighed. It is defined as Eq. 2.

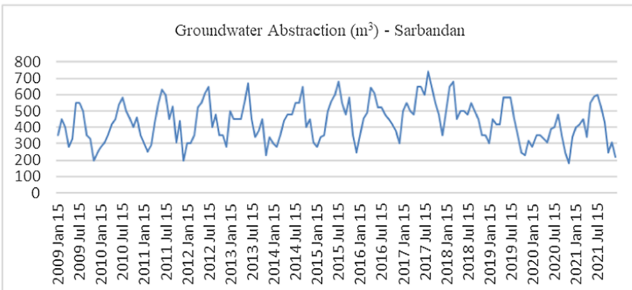
$$x_{\text{norm},i} = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \tag{2}$$



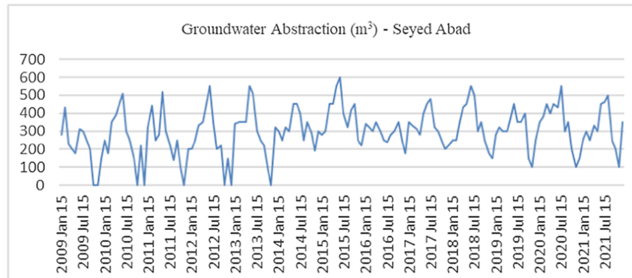
a. Garmabsard Groundwater Abstraction (m^3)



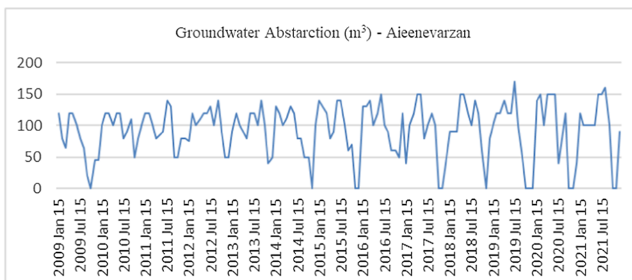
b. Absard Groundwater Abstraction (m^3)



c. Garmabsard Groundwater Abstraction (m^3)



d. Seyed Abad Groundwater Abstraction (m^3)



e. Avieenevarzan Groundwater Abstraction (m^3)

Fig. 6 Groundwater abstraction

where $x_{norm,i}$ is the normalized value, x_i, x_{max}, x_{min} are the sample, maximum, and the minimum value in samples. Normalization test is done by Eviews v.9.

Cross-correlation test

Parsing the relationships between variables is a crucial component of input data preprocessing. It is necessary to describe data correlations with a unit-free statistic called the "correlation coefficient" which varies from − 1 to + 1.

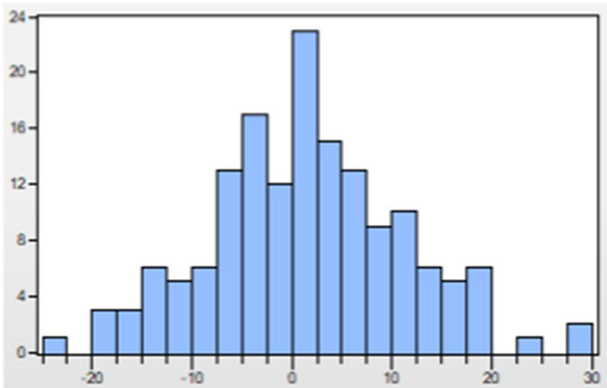
Most of the correlation coefficients between the water qualities index and rainfall rate indicate show a relatively strong positive correlation (Table 3). A negative value of the GWQI-WE correlation coefficient in most cases shows that an increase in water abstraction causes a decrease in the water quality index. The high elevation of sea level in the case of the Aieenevarzan area and the difference in ground topology probably cause a non-relation between water abstraction and groundwater quality. Other coefficients value means there is no correlation between GWQI and fertilizer usage, percentage of humidity, and temperature.

Prediction model

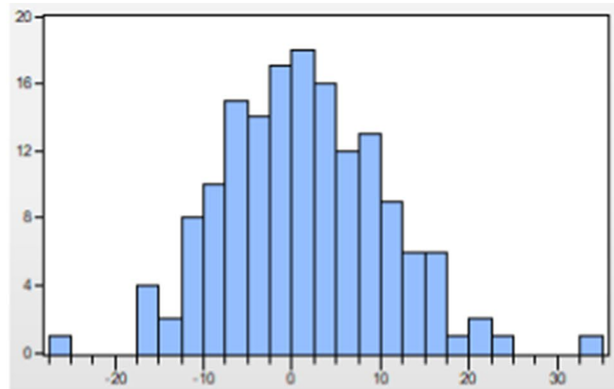
A recurrent neural network has a grouping of sequence networks. Every network unit takes input and data from

Table 3 Correlation coefficients of variables

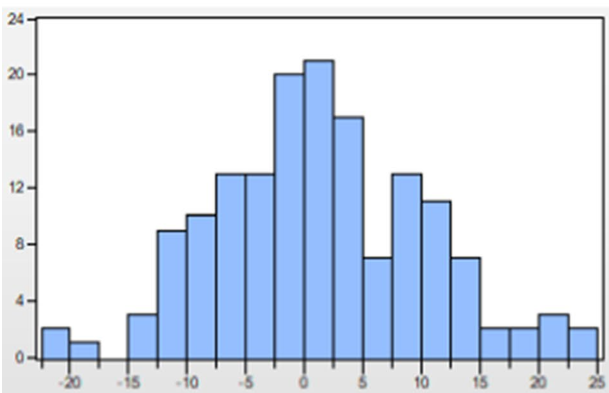
Collected Bore-Wells	R	WE	Hu	T
GWQI [Garmabsard]	0.10	− 0.2	0.06	0.1
GWQI [Abesard]	0.12	− 0.10	0.06	− 0.12
GWQI [Sarbandan]	0.20	− 0.07	0.03	0.02
GWQI [Seyed Abad]	0.07	− 0.03	0.07	0.00
GWQI [Aieenevarzan]	0.02	0.01	0.07	0.01



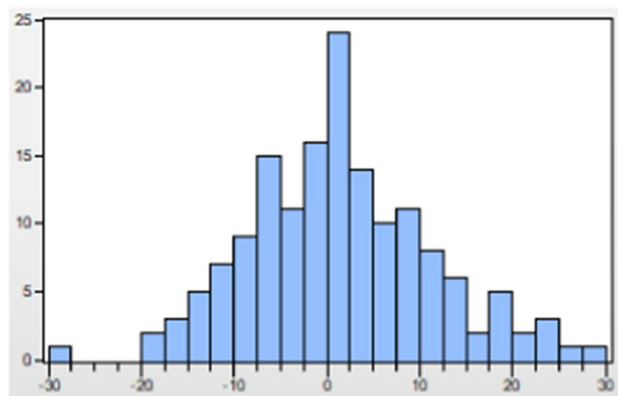
a. Absard data normalization test
Jarque-Bera: 0.90 | Probability: 0.63



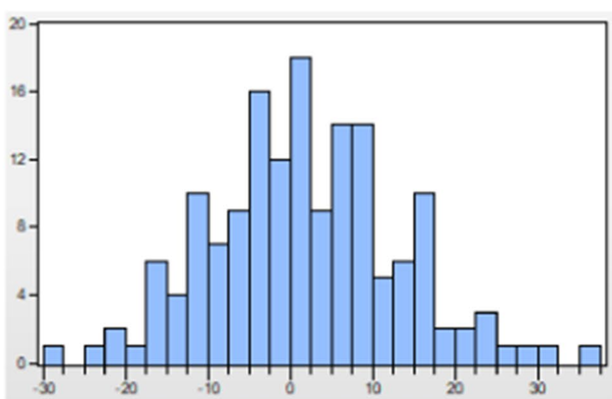
b. Aieenevarzan data normalization test
Jarque-Bera: 2.93 | Probability: 0.23



c. Garmabsard data normalization test
Jarque-Bera: 0.76 | Probability: 0.68



d. Sarbandan data normalization test
Jarque-Bera: 1.76 | Probability: 0.41



e. Seyed Abad data normalization test
Jarque-Bera: 1.01 | Probability: 0.60

Fig. 7 Normalization test results

the past unit, carries out the identified process, and produces results along with temporary information to the next section. Recently, with the progress of AI, researchers have proposed upgrade RNNs system to prevail over the restriction of the basic RNN systems. The hidden layer components keep the past production of hidden neurons, which are attached.

Gradients errors in progress caused trouble in standard RNNs (Jiang et al. 2019). To fix this trouble, the long short-term memory model parametrizes the typical RNN

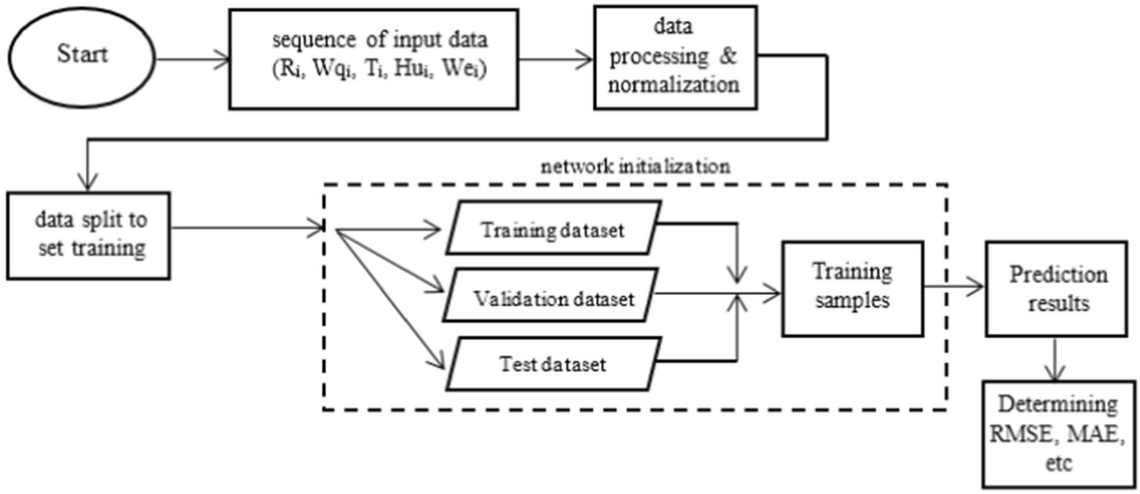


Fig. 8 Model framework

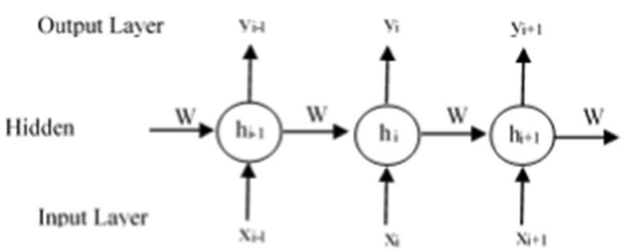


Fig. 9 Unfolded structure of typical recurrent neural networks

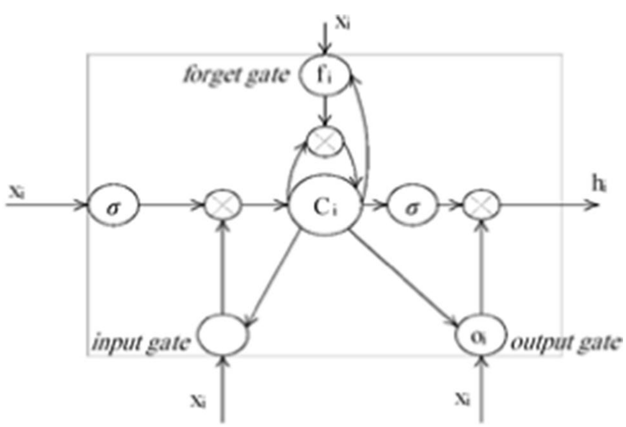


Fig. 10 Structure of memory segment

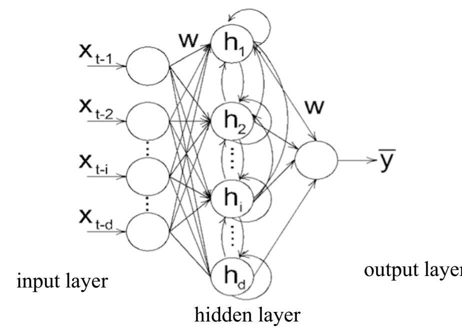


Fig. 11 Groundwater quality prediction based on LSTM RNN

(Greff et al. 2017). The typical RNN sequence units are converted into LSTM sequence units. In the next, the input ports in the model can import data from the preceding moment into the temporary memory one. This type of recurrent neural network (LSTM) is absolutely mapped out to overcome dependence in sequence issues Figs. 7 and 8.

Figure 9 shows the total framework of the model. Figure 10 shows the typical RNN structure include input layer, hidden layer and output layer. Figure 11 shows the internal structure of memory segment. Its activation is same to simple RNNs. In Fig. 11, the explanation of x_i, y_i, and h_i is the same on Fig. 10. C_i is the state of memory segment at ith time, which controls the renew and productivity of the networks. To keep C_i, three gates are active around the memory segment, input gate, forget gate and output gate. Although the RNN can draw the input established by the neuron node at the current time during training, polynomial multiplication can cause a grave gradient disappearance problem when the parameter is updated by the sequence instruction (Fang et al. 2020).

Input gate controls which particulars should be save in memory. 'i_i' is the input gate value, which is defined as follows:

$$i_i = \sigma\left(W_{ix}X_i\right) + W_{hh}h_{i-1} + b_i \quad (3)$$

W_{ix} is the weight of 'input gate—input layer' and W_{hh} is the weight of 'last time hidden layer—current time hidden layer'. Also, '

σ' is the sigmoid function and 'b_i' is the bias angel. In recurrent neural networks such as the LSTM, the sigmoid and hyperbolic tangent functions are regularly used as activation functions in the memory segment units. Forget gate and output gates are similar to input gate, follow as 'Eq. 4' and 'Eq. 5':

$$f_i = \sigma\left(W_{fx}X_i\right) + W_{fh}h_{i-1} + b_f \tag{4}$$

where W_{fx} is the weight of 'forget gate—input layer,' W_{fh} is the weight of 'forget gate—hidden layer,' and b_f is the bias angel.

$$o_i = \sigma\left(W_{ox}X_i + W_{hh}h_{i-1} + W_{oc}C_{i-1} + b_o\right) \tag{5}$$

where W_{ox} is the weight of 'output gate–input layer' and W_{oc} is the weight of 'output gate –memory segment state at ith time'. Also, the b_o is the bias angle. At 'Eq. 4' and 'Eq. 5', f_i is the value of forget gate and o_i is the value of output gate. The estimate of C_i defined as follows:

$$C_i = f_i \times C_{i-1} + i_i\sigma\left(W_{cx}X_i + W_{hh}h_{t-1} + W_{cc}C_{i-1} + b_c\right) \tag{6}$$

W_{cx} is the weight of 'input layer –memory segment state at ith time' and W_{cc} is the weight of 'memory segment state at ith time–memory segment state at next time.' 'b_c' is the bias angel. Eventually, we presented a prediction model with fully connected structure which includes input layer, hidden layer and output layer.

As shown in Fig. 12, the input $X = (x_{t-1}, x_{t-2}, \ldots, x_{t-d})$, $x_{t\text{-}d}$ is the sequential values of effective parameters in 'd' months before tth month. The number of neurons in the input layer is detected by time step (d). $H = (h_1, h_2, \ldots, h_i, \ldots, h_n)$ is the value of hidden layer. The number of neurons in hidden layer is defined by replication. '

$\overline{y} = \overline{x}_t$' is the prediction for tth month as the model output. The number of neurons in output layer is one, because the single step prediction model has been set up. 'σ' is sigmoid function in each neurons. The model total estimation is as follows:

$$h_i = H\left(W_{hx}X + W_{hh}h_{t-1} + b_n\right) \tag{7}$$

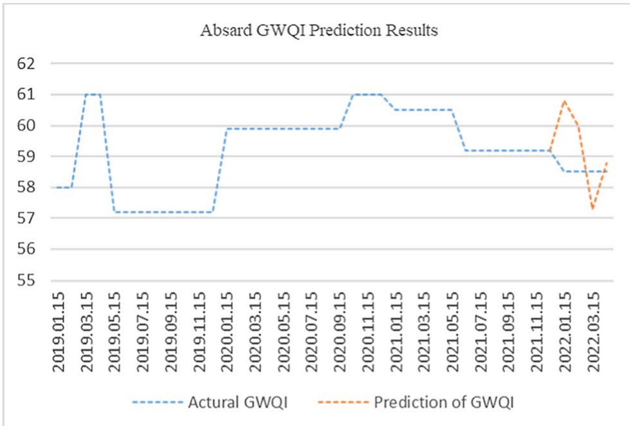
$$\overline{y} = W_{hx}X + b_y \tag{8}$$

W_{hx} is the weight of 'hidden layer–input layer,' W_{hh} is the weight of 'last time hidden layer–current time hidden layer' and b_y is the bias angel.

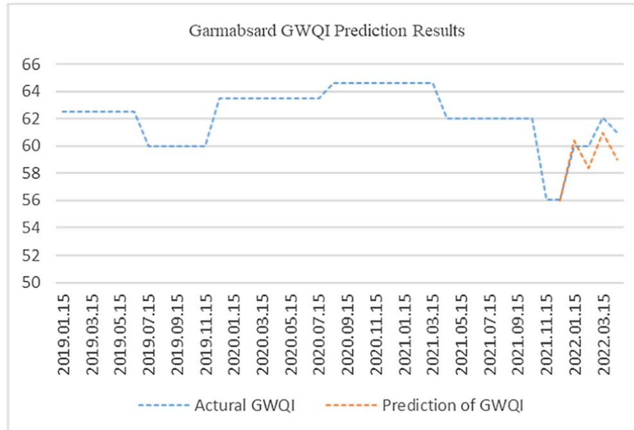
Results and discussion

In this research, we have multi-sequences of inputs that vary in length and may require the model to learn the long-term framework or dependencies between parameters in the input sequence. Sequence classification is a predictive modeling problem where some sequences of inputs over space or time and the task is to predict a category for the sequence. We used an LSTM model with rectified linear activation function which allows the neural network to learn nonlinear dependencies. According to Aish et al. (Aish et al. 2015), neural networks with too many layers rarely show good performance in the prediction process. It calls the "over-fitting" problem. In this research, various replication was run to obtain the optimal numbers of layers and neurons and the best performance and possible framework were used to obtain the desired result. To obtain the best neural network application, we used "f-score" testing. F-score test is a configurable single-score metric for evaluating a binary classification model based on the predictions made for the optimistic course. The f-score test result is about 0.74% in our prediction model. Neural network architecture is set up by 10 hidden layers which is optimal value between input and output layer, and the optimal length of time step is set to '4.' The neural network has been trained 250 times (epoch = 250). We performed data as 80% of training data and 20% of the test data. Figure 12 shows the output result of LSTM RNN prediction in five different sample cases. Data processing application is done by Python coding in the Google Colab platform.

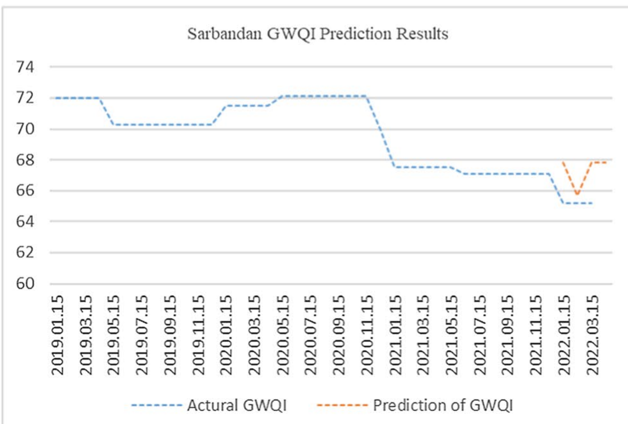
Figure 12 shows the five prediction results on the select bores. It can be seen that 4 out of 5 predictions are approximately true. The water quality index prediction in Garmabsard, Sarbandan, and Aieenevarzan has been done more sensibly and quite properly. Low accuracy of prediction results in Seyed Abad and one step wrong prediction in the Absard groundwater quality analysis may show that some effective parameters have been dropped out or the lack of an existing an exhaustive database. In most charts, the lowest error in the first time steps prediction shows that near-future prediction is more accurate than the long run. The aim of neural networks structure is to detect the optimum weight coefficients. F-Score testing has been employed to find which times of training network create the best learning architecture. Networks frequently training is to decline errors manufactured between the entry data and the final generation of the networks. The root-mean-square error (RMSE) identifies the predictive value distance from the true value. It shows the prediction efficiency. Minimum of RMSE presents the higher predictive accuracy, which is shown as follows:



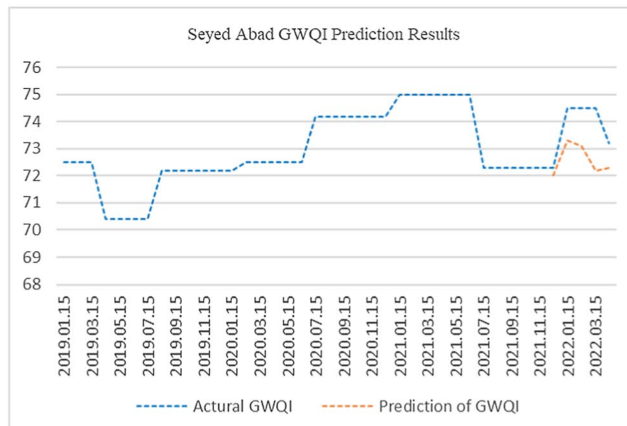
a. Absard GWQI Prediction Results



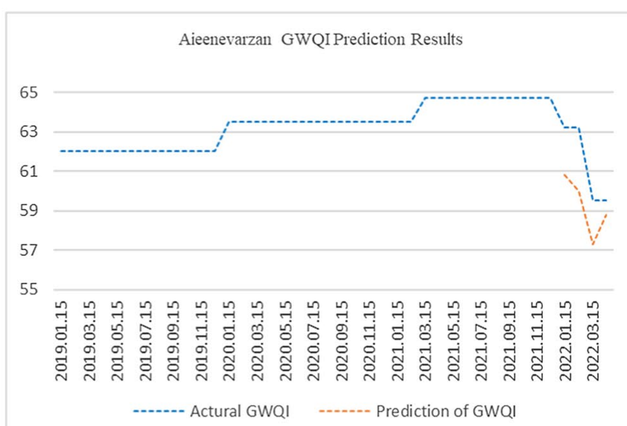
b. Garmabsard GWQI Prediction Results



c. Sarbandan GWQI Prediction Results



d. Seyed Abad GWQI Prediction Results



e. Aieenevarzan GWQI Prediction Results

Fig. 12 Prediction results

Table 4 Result of statistical accuracy indicators

Sample location	Model	Hidden layer	Length of look back	Epochs	MAE	MSE	R^2
Absard	LSTM RNN	10	4	250	0.15	0.28	0.99
Garmabsard	LSTM RNN	10	4	250	0.1	0.12	0.99
Sarbandan	LSTM RNN	10	4	250	0.17	0.43	0.99
Seyed Abad	LSTM RNN	10	4	250	0.2	0.41	0.99
Aieenevarzan	LSTM RNN	10	4	250	0.26	0.66	0.99

$$\text{RMSE} = \sqrt{\frac{\sum \left(y_0 - y_t\right)^2}{n}} \tag{9}$$

Also, coefficient of determination (R^2), Mean Squared Error (MSE) and Mean Absolute Error (MAE) are given as the following 'Eq. 10', 'Eq. 11' and 'Eq. 12':

$$R^2 = 1 - \frac{\sum \left(y_0 - y_t\right)^2}{\sum (y_t)^2} \tag{10}$$

$$\text{MSE} = \frac{\sum (y_0 - y_t)^2}{n} \tag{11}$$

$$\text{MAE} = \frac{\sum \left|y_0 - y_t\right|}{n} \tag{12}$$

where is 'y_0' is predicted value, 'y_t' is true value at the tth month. Table 4 shows the comparison and result of statistical indicators.

Conclusion

This research is an attempt to review the exogenous natural and humanity effective variables, which influence the water quality index. It is set up a new plan with new effective parameters based on LSTM RNN for groundwater quality index prediction. LSTM RNN method has been widely used by the various scientists (Tyagi et al. 2013). The model is trained by the historical data of rainfall, temperature, humidity, water abstraction amount, and water quality indicators in five different sample locations in Damavand. These variables define various human intervention and biological parameters. Due to increase in errors in subsequent courses, it is clear that prediction in short term of future is more accurate. Also, an expansion in effective parameters as input sequence variables may cause the swerve of water quality prediction. As the vast extent of the study area and inhomogeneous distribution of soil structure and aquifers, we can't expand our results to another part of the area, because of differences in soil features and ground topology. To uphold and develop of the present research, we suggested that some operation must be done in future: (1) Seeking more effective parameter optimization such as air pressure, surface evapotranspiration, groundwater aquifers features, etc. (2) Don't abstract water more than irrigating needed because of long-run renewable processing.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Aish AM, Zaqoot HA, Abdeljawad SM (2015) Artificial neural network approach for predicting reverse osmosis desalination plants performance in the Gaza Strip. Desalination 367:240–247

Bacquart T, Frisbie S, Mitchell E, Grigg L, Cole C, Small C, Sarkar B (2015) Multiple inorganic toxic substances contaminating the groundwater of Myingyan Township, Myanmar: Arsenic, manganese, fluoride, iron, and uranium. Sci Total Environ 517:232–245. https://doi.org/10.1016/j.scitotenv.2015.02.038

Chen B, Zhu G, Ji M, Yu Y, Zhao J, Liu W (2020) Air quality prediction based on Kohonen Clustering and ReliefF feature selection. Comput Mater Cont CMC 64(2):1039–1049

Fang W, Zhang F, Ding Y, Sheng J (2020) A new sequential image prediction method based on LSTM and DCGAN. Comput Mater Cont 64(1):217–231

Gleeson T, VanderSteen J, Sophocleous A, Taniguchi M, Alley WM, Allen DM, Zhou Y (2010) Groundwater sustainability strategies. Nat Geosci 3(6):378–379

Greff K, Srivastava RK, Koutnik J, Steunebrink, BR, Schmidhuber, J (2017) LSTM: a search space odyssey, 28(10): 2222–2232

Hameed M, Sharqi SS, Yaseen ZM, Afan H, Hussain A, Elshafie A (2016) Application of artificial intelligence (AI) techniques in water quality index prediction: a case study in tropical region, Malaysia. Neural Comput Appl 28(1):893–905

Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9:1735–1780

Hossain B, Morooka T, Okuno M, Nii M, Yoshiya Sh, Kobashi S (2019) Surgical outcome prediction in total knee arthroplasty using machine learning. Intell Autom Soft Comput 25(1):105–115

Jan CD, Chen TH, Lo WC (2007) Effect of rainfall intensity and distribution on groundwater level fluctuations. J Hydrol 332(3–4):348–360

Jiang Q, Chen L,Xu R, Ao X, Yang M (2019) A challenge dataset and effective models for aspect-based sentiment analysis. In:Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint-conference on natural language processing (EMNLP-IJCNLP).

Association for Computational Linguistics, Hong Kong, China, pp 6280–6285
Khan MN, Mobin M, Abbas ZK, Alamri SA (2018) Fertilizers and their contaminants in soils, surface and groundwater. Encycl Anthr. https://doi.org/10.1016/B978-0-12-809665-9.09888-8
Koundouri P (2004) Potential for groundwater management: Gisser-Sanchez effect reconsidered. Water Resour Res. https://doi.org/10.1029/2003WR002164
Kumar CP (2012) Climate change and its impact on groundwater resources. Int J Eng Sci 1(5):43–60
Li L, Jiang P, Xu H, Lin G, Guo D, Wu H (2019) Water quality prediction based on recurrent neural network and improved evidence theory: a case study of Qiantang River, China. Environ Sci Pollut Res 26(4):19879–19896
Madani K (2014) Water management in Iran: what is causing the looming crisis? J Environ Stud Sci 4(4):314–325
Meireles ACM, de Maia Andrade E, Guerreiro Chaves LC, Frischkorn H, Crisostomo AC (2010) A new proposal of the classification of irrigation water. Agric Eng 41(3):41–57
Ning CC, Gao PD, Wang BQ, Lin WP, Jiang NH, Cai KZ (2016) Impacts of chemical fertilizer reduction and organic amendments supplementation on soil nutrient, enzyme activity and heavy metal content. J Integr Agric 16(8):1819–1831
Norse D (2005) Non-point pollution from crop production: global, regional and national issues. Pedosphere 15(4):1–10
Ping L, Jin W, Sangaiah A, Xie Y, Yin X (2019) Analysis and prediction of water quality using LSTM deep neural network in IoT environment. Sustainability 11(7):2058–2074
Sahoo S, Jha MK (2015) On the statistical forecasting of groundwater levels in unconfined aquifer systems. Environ Earth Sci 73(7):3119–3136
Scholz RW, Geissler B (2018) Feebates for dealing with trade-offs on fertilizer subsidies: a conceptual framework for environmental management. J Clean Prod 189:898–909
Sreekesh S, Sreerama SR, Naik S, Seema R (2018) Effect of sea level changes on the groundwater quality along the coast of Renakulam District, Kerala. J Clim Ch 4(2):51–65
Ting Xu, Dengming Y, Baisha W, Wuxia B, Pierre D, Fang L, Ying W, Jun M (2018) The effect evaluation of comprehensive treatment for groundwater overdraft in Quzhou County, China. Water 10(7):874–892
Tyagi S, Sharma B, Singh P, Dobhal R (2013) Water quality assessment in terms of water quality index. Am J Water Resour 1(3):34–38
Wada Y, van Beek LPH, van Kempen CM, Reckman JWTM, Vasak S, Bierkens MFP (2010) Global depletion of groundwater resources. Geophys Res Lett. https://doi.org/10.1029/2010GL044571
Wada Y, Wisser D, Eisner S, Flörke M, Gerten D, Haddeland I, Hanasaki N, Masaki Y, Portmann FT, Stacke T, Tessler Z, Schewe J (2013) Multimodel projections and uncertainties of irrigation water demand under climate change. Geophys Res Lett 40(17):4626–4632
Wang Y, Zhou J, Chen K, Wang Y, Liu L (2017) Water quality prediction method based on LSTM neural network. In: International Conference on Intelligent Systems and Knowledge Engineering, 12th
Yan B, Wang J, Zhang Z, Tang X, Zhou Y, Zheng G, Zou Q, Lu Y, Liu B, Tu W, Xiong N (2020) An improved method for the fitting and prediction of number of COVID-19 confirmed cases based on LSTM. Comput Mater Cont CMC 64(3):1473–1490
Ye Y, Liu F, Zhao S, Hu W, Liang Z (2020) Ensemble learning based on GBDT and CNN for adoptability prediction. Comput Mater Cont CMC 65(2):1361–1372