

Modified DRASTIC model for groundwater vulnerability to nitrate contamination in the Dagujia River Basin, China

Aihua Wei, Pan Bi, Jie Guo, Shuai Lu and Duo Li

ABSTRACT

Due to rapid economic growth and over-exploitation of groundwater, nitrate pollution in groundwater has become very serious. The main objective of this study is to modify the DRASTIC model to identify groundwater vulnerability to nitrate pollution. The DRASTIC model was firstly used to analyze the intrinsic vulnerability. The DRASTIC model with the inclusion of a land-use factor (DRASTIC-LU) was put forward to map the specific vulnerability of groundwater. Furthermore, the Support Vector Machine (SVM) was introduced to avoid the drawback of the overlay and index methods, and the improved integrated models of DRASTIC + SVM and DRASTIC-LU + SVM were built. Moreover, 103 groundwater samples were collected for building and validating the models. The Root Mean Squared Error (RMSE) of DRASTIC, DRASTIC-LU, DRASTIC + SVM, and DRASTIC-LU + SVM was found to be 0.853, 0.755, 0.631, and 0.502, respectively. The model DRASTIC-LU was more precise than the original one. The results also showed that the integrated model using SVM exhibited better correlation between the vulnerability value and the nitrate pollution. The study indicated that the modified models including the land-use factor as well as SVM in the DRASTIC model were more suitable to assess the groundwater vulnerability to nitrate.

Key words | DRASTIC, integrated model, land-use, nitrate pollution, SVM, vulnerability

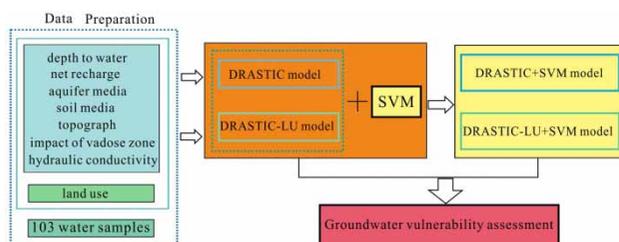
Aihua Wei
Pan Bi
Shuai Lu
Duo Li (corresponding author)
 School of Water Resources and Environment,
 Hebei GEO University,
 Shijiazhuang 050031,
 China;
 Hebei Province Key Laboratory of Sustained
 Utilization and Development of Water
 Resources,
 Shijiazhuang 050031,
 China
 and
 Hebei Province Collaborative Innovation Center for
 Sustainable Utilization of Water Resources and
 Optimization of Industrial Structure,
 Shijiazhuang 050031,
 China
 E-mail: liduo@hgu.edu.cn

Jie Guo
 Key Laboratory of Shale Gas and Geoengineering,
 Institute of Geology and Geophysics,
 Chinese Academy of Sciences,
 Beijing 100029,
 China

HIGHLIGHTS

- The improved DRASTIC-LU model was put forward to delineate groundwater specific vulnerability.
- Support vector machine was introduced to further build the novel integrated model to analyze the intrinsic and specific vulnerability of groundwater.

GRAPHICAL ABSTRACT



This is an Open Access article distributed under the terms of the Creative Commons Attribution Licence (CC BY 4.0), which permits copying, adaptation and redistribution, provided the original work is properly cited (<http://creativecommons.org/licenses/by/4.0/>).

doi: 10.2166/ws.2021.018

INTRODUCTION

Groundwater is a crucial resource for ensuring sustainable development in different regions of the world. Protecting groundwater is considered as a fundamental requirement on a global scale. However, groundwater pollution has become more serious recently. Nitrate contamination is one of the most widespread problems faced by many countries (Arabgol *et al.* 2016; Thapa *et al.* 2018). Many researchers have mentioned that nitrate contamination has severe impacts not only on human health, but also on the ecosystem (Li *et al.* 2017). Vulnerability assessment has proven to be an efficient and cost-effective tool for protecting groundwater resources from contamination (Machiwal *et al.* 2018).

The concept of groundwater vulnerability was first introduced by Margat in the late 1960s, and is defined as the tendency for contaminants to reach the groundwater system (Huan *et al.* 2012). Generally, this term consists of the intrinsic and specific vulnerabilities. The intrinsic vulnerability indicates a vulnerability to all contaminants, while the specific vulnerability is based on the intrinsic vulnerability combined with the specific contaminant characteristics. Groundwater vulnerability assessment projects, especially those about the intrinsic vulnerability, have successfully been used in many regions, and can be summarized into three categories: statistical models, process mathematical simulation methods and index-overlay methods (Kim & Hamm 1999; Pacheco *et al.* 2015). Statistical methods are used to easily obtain the relationships between the selected variables and the actual occurrence of pollutants in the groundwater. However, the assessment results are easily influenced by the selected data or methods. Although process-based models can obtain accurate results, this method needs a large input of data and is difficult to realize especially on the regional scale. By contrast, index-overlay is a commonly used method due to its simplicity and small data requirement (Sajedi-Hosseini *et al.* 2018). DRASTIC, COP, EPIK, SINTACS and GOD are some of the overlay and index methods, among which DRASTIC is one of the most widely used methods (Pacheco *et al.* 2015; Caprario *et al.* 2019; Rajput *et al.* 2020). DRASTIC has an outstanding advantage of permitting a simplicity and

flexibility criteria structure to realize the estimation. However, the weights and rates are originally given or dependent on the experiences of assessment experts, which is the major drawback of this method. In order to deal with this issue, some studies have proposed various techniques, such as changing the weights and/or rates of the structure, subtracting or adding additional factors, using sensitivity analyses and calibration approaches, and combining with the analytic hierarchy process (Secunda *et al.* 1998; Huan *et al.* 2012; Thapa *et al.* 2018).

Recently, machine learning and soft computing techniques have successfully been applied to the assessment of environmental issues, such as flood susceptibility, landslide susceptibility, groundwater level prediction, and groundwater pollution (Yoon *et al.* 2011; Tehrani *et al.* 2015; Chen *et al.* 2018; Sajedi-Hosseini *et al.* 2018; Deng *et al.* 2020). Groundwater vulnerability delineation using intelligent algorithms has attracted much research attention. Some scholars have just begun to study groundwater vulnerability using Artificial Intelligence (AI) (Jia *et al.* 2019; Rajput *et al.* 2020). It is necessary to explore novel approaches for assessing highly reliable groundwater vulnerability mapping.

The Support Vector Machine (SVM) is a relatively new AI technique in the field of data-driven predictions. Due to its strong theoretical statistical framework, SVM has proved to be significantly robust in several fields, especially for noise mixed data (Raghavendra & Deka 2014; Deng *et al.* 2019). SVM can effectively improve the objectivity and accuracy of the results. Many researchers have verified SVM as an efficient classifier of divisible and non-divisible categorical problems (Asefa *et al.* 2006; Pradhan 2013; Naghibi *et al.* 2017). Another obvious advantage of SVM is that it uses structural risk minimization instead of the empirical risk. Therefore, it can adequately deal with a high dimensionality or limited number of training samples (Cortes & Vapnik 1995).

The basin of the Dagujia river, known as the 'Mother River' by the local people, is located in Yantai city. Recently, groundwater exploitation has increased enormously, and the release of municipal and industrial wastes has seriously

threatened the local groundwater environment. Meanwhile, intense agricultural activities have resulted in the leaching of nutrient constituents. Generally, this basin is a typical area with groundwater nitrate pollution, however, little attention has been paid to groundwater nitrate contamination in this area.

In this paper, based upon the DRASTIC model, integrated models were put forward to assess groundwater vulnerability. The DRASTIC method was firstly used to assess the intrinsic vulnerability of groundwater. Then, the DRASTIC-LU method was put forward to study the specific vulnerability of the aquifer to nitrate pollution. Meanwhile, considering the deficiency of the index-overlay method, the SVM methods combined with DRASTIC and DRASTIC-LU were used to build the novel integrated models. Generally, the main objective of this research is to produce groundwater intrinsic and specific vulnerability maps by employing the ensembles of the DRASTIC (and DRASTIC-LU) method and intelligence technique, and evaluate their respective performances.

STUDY AREA

The Dagujia river basin is located in the north-central part of Yantai city, Shandong, China (Figure 1). The basin is between the longitude and latitude of $120^{\circ}44'00''$ – $121^{\circ}27'00''$ and $37^{\circ}02'00''$ – $37^{\circ}36'00''$, respectively, covers an area of about 2,308 km², and is bound by the Yellow Sea to the north. The climate is warm temperate monsoon with four distinctive seasons. The average temperature and the annual precipitation are 12.5 °C and 629 mm, respectively. The general trend of the topography is high in the south and low in the north. The maximum elevation has been observed to be 814 m in the western part, whereas the northern part is dominated by coastal regions. The main river, the Dagujia river, is drained by the inner Jia river in the west and the outer Jia river in the east. Meanwhile, there are three reservoirs in this watershed: the Menlou reservoir, the Anli reservoir and the Taoyuan reservoir.

In the study area, groundwater is used for domestic, industrial and agricultural applications. Types of groundwater include pore water in loose ground, fracture karst

water in carbonates, bedrock fracture water and clastic-rock-type pore–fissure water. The pore and karst water is the target aquifer of intensive exploitation in this area. Increase in the amount of groundwater withdrawal is the dominant factor, resulting in the lowering of groundwater level. The annual groundwater extraction has increased continually since 1976. Since the 1990s, the government has begun to control the exploitation of groundwater, and the problem of saltwater intrusion has tended to slow down. The groundwater contamination problem caused by seawater intrusion mainly occurs near the Dagujia river estuary. However, as a result of the rapid development of society in recent decades, nitrate pollution of groundwater has become more serious. The maximum concentration of nitrate in the groundwater has reached more than 200 mg/L. Therefore, there is an urgent need to initiate study of groundwater nitrate pollution in the river basin.

DATA AND METHODS

The classic DRASTIC model and its modified model, DRASTIC-LU, were introduced. Meanwhile, an SVM-based improved method was also used. The natural break classification scheme could determine the best arrangement of values into different classes (Yoon *et al.* 2011; Thapa *et al.* 2018). Therefore, in order to compare the results using different methods, the natural break classification method was always selected to divide the assessment area into the same four ranks: very low, low, medium, and high vulnerability zones.

Preparation of the nitrate concentration

In this study, the nitrate concentration was selected as the primary pollution parameter. To this end, 103 groundwater samples were collected from wells or piezometers during December 2014. The spatial distribution of water samples is shown in Figure 1. According to the World Health Organization (WHO), a concentration of nitrate in groundwater exceeding the permissible limit (50 mg/L) is considered as unsuitable for drinking purposes. Meanwhile, when the concentration reaches 30 mg/L, based upon the Chinese Groundwater Quality Standards (GB/T 14848–2017), the

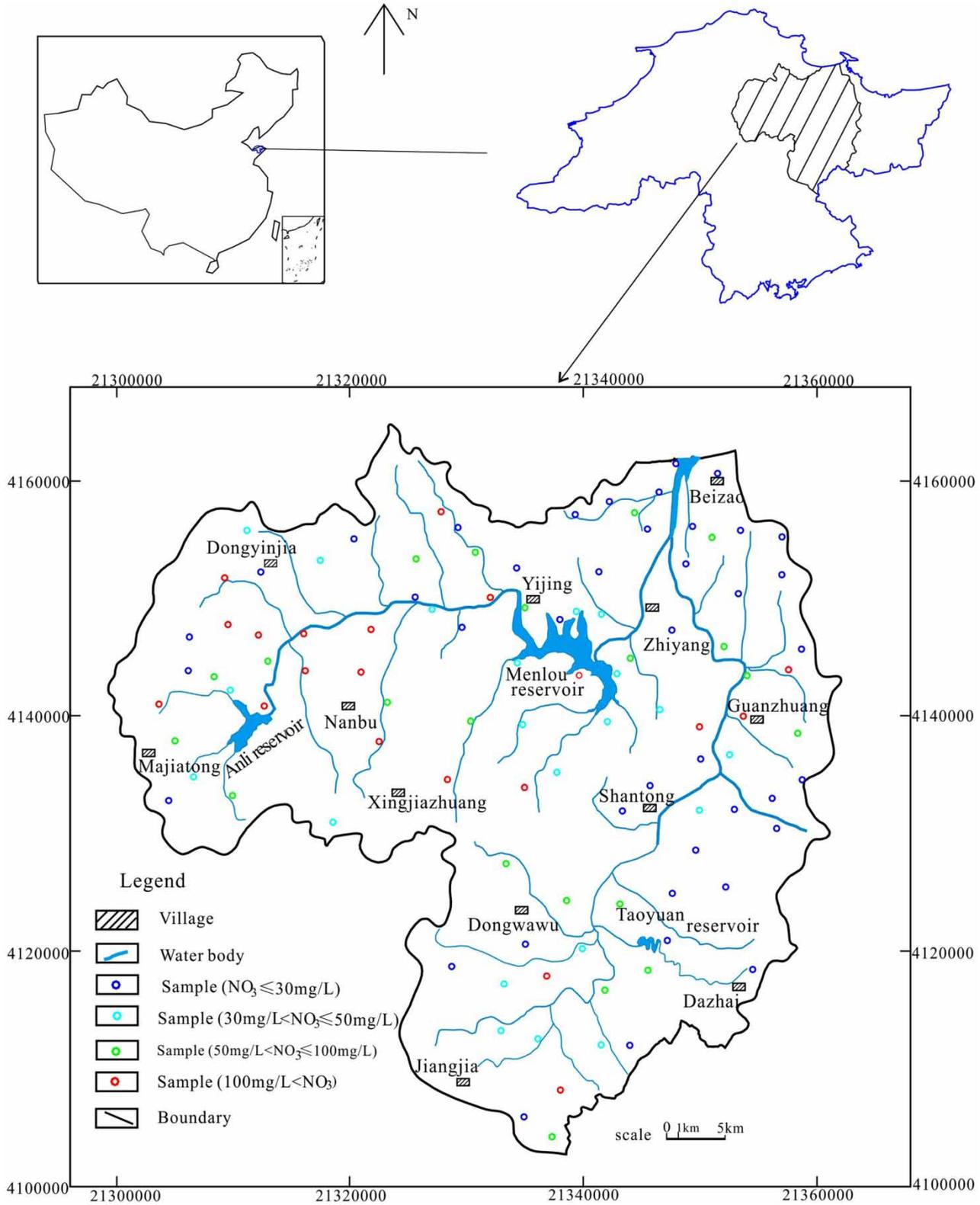


Figure 1 | The location of the study area and water samples.

groundwater is considered not suitable for drinking. Among the collected groundwater samples, nitrate concentration exceeding 30 mg/L and 50 mg/L accounted for 60.2% and 40.8% of the total samples, respectively. Areas with higher nitrate concentrations (more than 50 mg/L) were mainly distributed sporadically in the alluvial region, proluvial-alluvial region and the southwest part among Dongyinjia, Nanbu, and Xingjiazhuang villages.

For the validation of obtained groundwater vulnerability maps from different assessment methods, the nitrate concentration was incorporated into the validation process. The concentration could also serve as the basis for deriving target values for the supervised process using SVM. Considering that the concentration values are directly proportional to the vulnerability index, the collected 103 groundwater samples were divided into four classes (as noted above) based on the abovementioned threshold values: less than 30 mg/L (very low), 30–50 mg/L (low), 50–100 mg/L (medium), and greater than 100 mg/L (high), and these categories included 41, 21, 21, and 21 samples, respectively. The detailed distribution is shown in Figure 1.

DRASTIC method

DRASTIC is believed to be a very popular method for groundwater vulnerability assessment. The original DRASTIC model includes seven hydrogeological parameters: depth to water (D), net recharge (R), aquifer media (A), soil media (S), topography (T), impact of vadose zone (I), and hydraulic conductivity (C) (Singha *et al.* 2019). Both the rates and the weights of the seven parameters are based on local conditions. The given weights of the above seven factors range between 1 and 5 and represent the relative importance of those parameters. The Vulnerability Index (VI) is used to divide the region, which is calculated as the sum of the product of prescribed ratings (r) and weights (w) assigned to each of the above seven data layers (Pacheco *et al.* 2015). In order to obtain the VI value, Equation (1) was used. According to the related literature reviews and local data conditions, Table 1 presents the rating and weighting values in the DRASTIC model.

$$VI = D_r D_w + R_r R_w + A_r A_w + S_r S_w + T_r T_w + I_r I_w + C_r C_w \quad (1)$$

Table 1 | The rating and weight values of DRASTIC and DRASTIC-LU models

Parameters	Range	Rating	Weight
D (m)	0–2	10	5
	2–4	8	
	4–8	6	
	8–12	3	
	>12	1	
R (mm/y)	>200	9	4
	150–200	8	
	100–150	6	
	50–100	3	
	0–50	1	
A	Limestone with intercalation of shale and granulite	9	3
	Quaternary sediments	6	
	Metamorphic rocks	5	
	Clastic rocks, thin-layer metamorphic rocks	3	
	Granite, gneiss	2	
S	Sand loam	7	2
	Silt loam	3	
	Clay loam	1	
T (%)	0–2	10	1
	2–6	9	
	6–12	5	
	12–18	3	
	>18	1	
I	Gravel, coarse sand with gravel	9	5
	Sand	7	
	Sand with clay	4	
	Bedrock	3	
C (m/d)	0–5	1	3
	5–10	2	
	10–20	5	
	20–30	7	
	>30	10	
LU	Agricultural area	9	4
	Water body	8	
	Built-up area	6	
	Land for transportation	4	
	Tree-clad area	3	

Here D , R , A , S , T , I , and C are the seven parameters and the subscripts r and w indicate the corresponding ratings and weights, respectively.

DRASTIC-LU method

Different types of land uses cause separate sources of nitrogen pollution, resulting in variant input intensity of nitrate

contamination sources. Land cover can also affect the infiltration rate and surface runoff. Therefore, the factor of land use (LU) was also considered to build the DRASTIC-LU model for the assessment of specific vulnerability to nitrate contamination. The DRASTIC-LU Vulnerability Index (VIL) could be obtained by adding the land use factor with appropriate weights and ratings. On the basis of VI, Equation (2) is used to calculate the VIL value. In this study, the weight of the land use parameter was set to be 4. The detailed information on this model is presented in Table 1.

$$VIL = VI + LU_r LU_w \quad (2)$$

Here VI is the vulnerability index calculated using Equation (1), and LU_r and LU_w are the rate and weight values of the land use parameter, respectively.

Support vector machine (SVM)

The SVM, a relatively new machine learning method, is a supervised machine learning algorithm. SVM is one of the most cogent prediction methods, which is based on the structural risk minimization method. By contrast, most artificial intelligence models, such as Artificial Neural Networks, use empirical risk minimization techniques. Therefore, the SVM method can reduce the empirical error, model the complexity and overfit the probability (Raghavendra & Deka 2014; Chen et al. 2018).

A training data set represented by T was given, as can be seen from Equation (3):

$$T = \{x_i, y_i\}_{i=1}^m \quad (3)$$

Here m is the number of samples in the data set, x_i is the input vector of data sample i ($x_i \in R^N$), y_i is the corresponding output value ($y_i \in R$), and R^N and R are the N -dimensional and one-dimensional vector spaces, respectively.

The aim of SVM is to find the optimal separation hyperplane, which can specify the widest margin between different classes and minimize the distance of the same class. The separating hyperplane is obtained by solving Equation (4):

$$w \cdot x + b = 0 \quad (4)$$

Here w is the normal vector of the separating hyperplane ($w \in R^N$), and b is the bias value.

As to the linearly separable problem, the hyperplane can be formulated as a quadratic programming problem by solving the objective function given by Inequality (5):

$$\begin{aligned} \text{Minimize} \quad & \frac{1}{2} \|w\|^2 \\ \text{Subject to} \quad & y_i((w \cdot x_i) + b) \geq 1 \quad i = 1, \dots, m \end{aligned} \quad (5)$$

Meanwhile, the kernel function is employed to transform a nonlinear classification into a linear classification problem to find an optimum separating hyperplane. The slack variable (ξ) and the penalty factor (C) can modify the constraint condition and objective function, whereas the corresponding objective function (Inequality (5)) can be converted to Inequality (6):

$$\begin{aligned} \text{Minimize} \quad & \frac{1}{2} \|w\|^2 + C \sum_{i=1}^m \xi_i \\ \text{Subject to} \quad & y_i((w \cdot x_i) + b) \geq 1 - \xi_i \quad i = 1, \dots, m \\ & \xi_i \geq 0 \end{aligned} \quad (6)$$

Many studies using SVM demonstrate that the Radial Basis Function (RBF) has favorable performance over other kernels in groundwater and hydrologic predictions (Arabgol et al. 2016). Therefore, RBF with the Gaussian kernel was selected in this study and is given by Equation (7):

$$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\gamma}\right) \quad (7)$$

Here γ is the Gaussian parameter.

The parameters C and γ in Equations (6) and (7), respectively, have a significant effect on the accuracy of the SVM model, which is noted as the major drawback of SVM. Therefore, a ten-fold cross-validation was used to select the optimal kernel parameters in this process (Pradhan 2013; Naghibi et al. 2017).

RESULTS AND DISCUSSION

Data preparation

The maps of *D*, *R*, *A*, *S*, *T*, *I*, *C*, and *LU* were prepared in a raster format with a resolution of 30 m using geographic information system (GIS), and the inverse distance weighted interpolation technique was used to transform the statistical discrete data to a continuous surface.

D: At the lower groundwater level depth, contaminants are easier to get into the upper aquifer. The water level data of 30 boreholes and 60 piezometers in July 2014 were used to interpolate the depth to groundwater in the basin. This map was reclassified into five classes, and the subarea can be seen from Figure 2(a).

R: Net recharge is the amount of water that reaches the groundwater aquifer, which is greatly influenced by the surface cover and topography (Singha *et al.* 2019). The average annual rainfall ranged from 650 to 400 mm during the period from 2000 to 2014. The rainfall infiltration factor values were set in terms of soil characteristics. The net recharge in the study area was estimated using the rainfall infiltration method, which is shown in Figure 2(b).

A: The type and formation of aquifers are often used to identify the vulnerability to geo-environmental issues. In comparison to fine soil media, coarse soil media and limestone have a higher permeability and lower attenuation capacity. The aquifer media map was constructed using 30 boring logs and hydrogeological maps (Figure 2(c)).

S: The type and the grain size of soil cover influence the infiltration and downward movement of contaminants. The large size has a high potential to infiltrate contamination from the source. This layer was prepared from the soil association map and boring logs, as seen from Figure 2(d).

T: The topography condition affects the confluence and infiltration of surface water, especially the wastewater. The influence of slope gradient on the infiltration rate and direction could represent the topographical conditions in this basin. The map of the degree of slope was produced in a GIS environment using the digital elevation model, as seen from Figure 2(e).

I: The vadose zone is often defined as the ground portion between the water table and the soil cover. The material of the vadose zone influences the attenuation and

permeability characteristics of pollution. This map was extracted from the borehole and well logs (Figure 2(f)).

C: This parameter describes the ability of the aquifer to transmit water under a given hydraulic gradient. Except for the transport of pollution in the aquifer, the hydraulic conductivity also influences the subsurface flow rate and infiltration. This map was prepared using the transmissivity data and 20 pumping tests in this area (Figure 2(g)).

LU: This parameter is a specifically added parameter to the modified DRASTIC-LU model. The type of land use plays a vital role in determining the occurrence and transport of nitrate pollution, caused by anthropogenic activities (Jia *et al.* 2019). This additional factor map was prepared using the land use map of Yantai city in 2014 (Figure 2(h)).

Multicollinearity test

The independence of the selected parameters in the vulnerability assessment models is very important to ensure the accuracy of results. Correlation analysis has shown that a relationship among two or more of the input variables may cause deviations. In order to diagnose multi-collinearity among various factors, Tolerance (TOL) and Variance Inflation Factor (VIF) are the two common statistical parameters, using Equation (8). When the TOL value is <0.10 and VIF value is >5 , there is a high multicollinearity among the predictor variables (Choubin *et al.* 2019):

$$\begin{aligned} TOL &= 1 - R_j^2 \\ VIF &= 1/TOL \end{aligned} \quad (8)$$

Here R_j^2 is the *R*-squared value of regression using the regressing parameter *j* on all of the others.

The possibility of the presence of multicollinearity among the seven or eight conditioning factors was examined before the assessment process, and the corresponding results are presented in Tables 2 and 3. The results indicated that no high multicollinearity was observed among the selected parameters.

DRASTIC vulnerability map

The sums of the seven DRASTIC thematic parameters were used to estimate the value of VI, according to Equation (1).

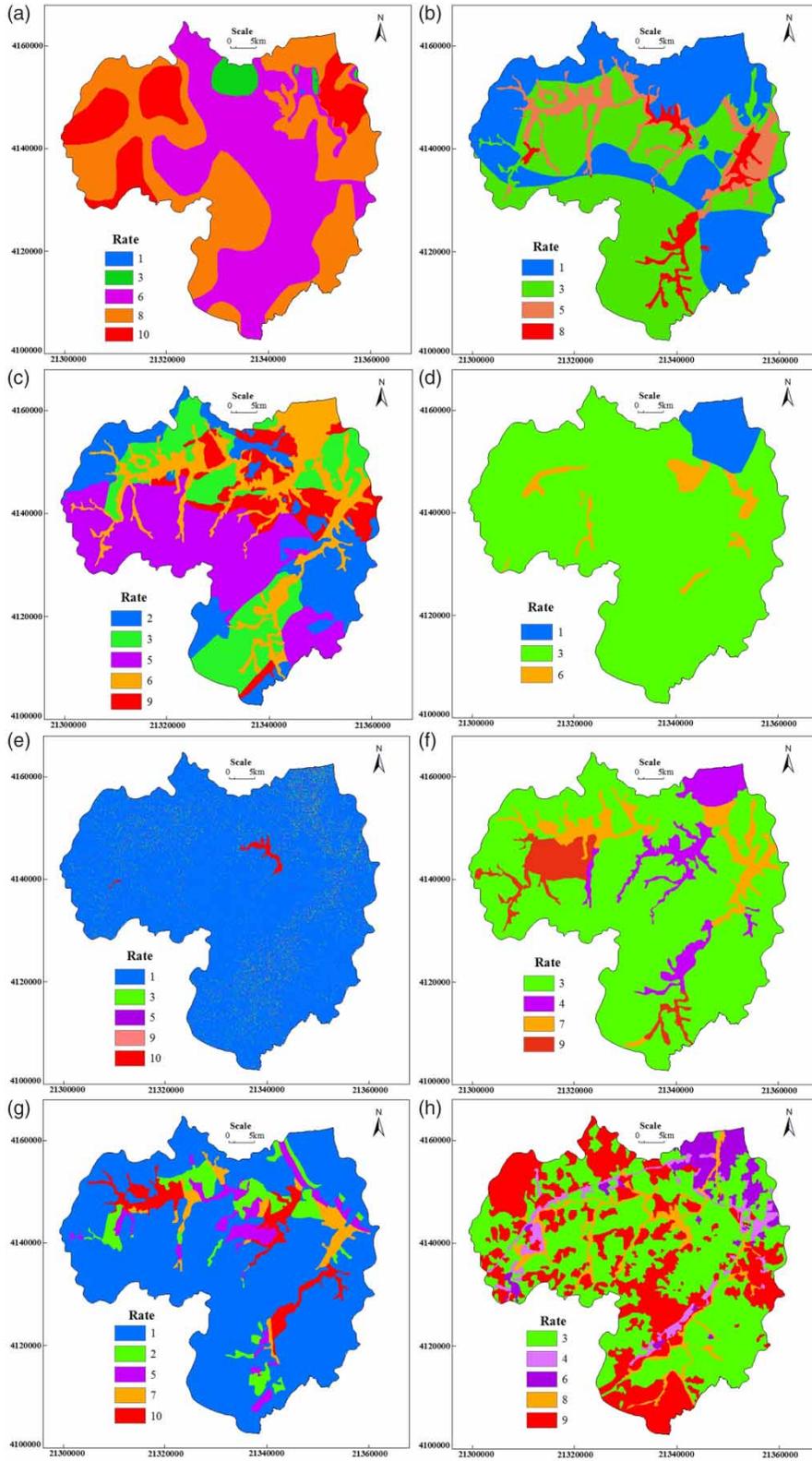


Figure 2 | Spatial distribution of the parameters: (a) depth to groundwater; (b) net recharge; (c) aquifer media; (d) soil media; (e) topography; (f) impact of vadose zone; (g) hydraulic conductivity; (h) land use.

Table 2 | Results of testing for DRASTIC model

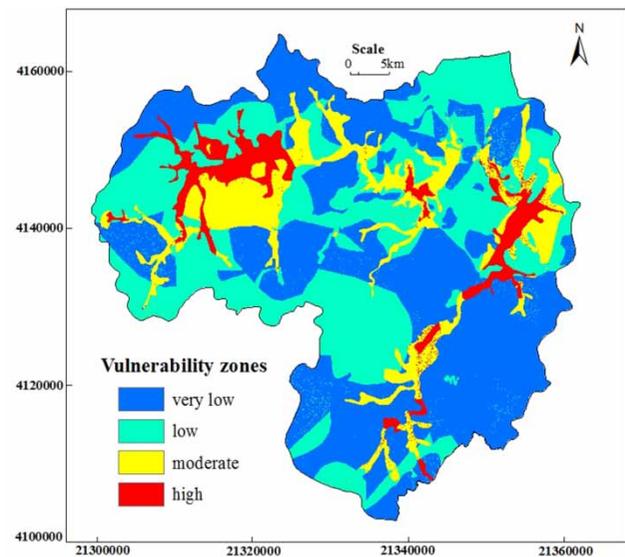
Factors	Unstandardized coefficients		Standardized coefficients			Collinearity statistics	
	B	Std error	Beta	T	Sig	TOL	VIF
D	0.087	0.043	0.130	2.025	0.045	0.979	1.022
R	0.129	0.052	0.219	2.492	0.014	0.520	1.923
A	0.038	0.037	0.067	1.006	0.317	0.892	1.122
S	0.145	0.068	0.143	2.121	0.036	0.880	1.136
T	0.225	0.077	0.252	2.940	0.004	0.547	1.830
I	0.112	0.046	0.203	2.430	0.017	0.572	1.748
C	0.074	0.038	0.176	1.958	0.053	0.497	2.011

Table 3 | Results of testing for DRASTIC-LU model

Factors	Unstandardized coefficients		Standardized coefficients			Collinearity statistics	
	B	Std error	Beta	T	Sig	TOL	VIF
D	0.042	0.041	0.062	1.034	0.304	0.920	1.087
R	0.102	0.048	0.173	2.134	0.035	0.512	1.954
A	0.042	0.034	0.075	1.219	0.226	0.891	1.122
S	0.148	0.062	0.147	2.373	0.020	0.880	1.136
T	0.180	0.071	0.202	2.546	0.012	0.536	1.866
I	0.102	0.042	0.185	2.412	0.018	0.571	1.753
C	0.075	0.035	0.178	2.168	0.033	0.497	2.011
LU	0.125	0.028	0.285	4.533	0.000	0.849	1.179

The higher VI value demonstrates higher vulnerability. The obtained VI values using the natural break classification method were divided into four vulnerability categories: high (49–84), moderate (84–109), low (109–138) and very low (138–185). The detailed groundwater vulnerability map was produced by the original model, which is clearly depicted in Figure 3.

The high and moderate degrees of vulnerability zones were mainly located within the alluvial deposit region, and covered 6.78% and 12.9% of the study area, respectively. The lithology of the high vulnerability zone was dominated by coarse sand and sand, whereas the rate of infiltration of contaminants from the surface as well as the transport in the aquifer were fast. The moderate-class region was mainly located at the edges of the foothill and structure development region, where pollution can easily infiltrate

**Figure 3** | The result of the DRASTIC model.

and be transported. The low and very low vulnerability sub-areas were mainly situated at the hilly region, lower infiltration and permeable unsaturated zones, and covered 38.87% and 41.45% of the total area, respectively. Compared with the distribution of 103 groundwater samples, there were some deviations between the vulnerability classifications and nitrate pollution.

DRASTIC-LU vulnerability map

Due to the land use applying different burdens to groundwater nitrate pollution, DRASTIC-LU was used to assess the groundwater's specific vulnerability. The original seven parameters from the DRASTIC and the land use layer were overlaid using Equation (2). The VIL value was 61–221, which was also divided into four classes: high (61–99), moderate (99–122), low (122–152), and very low (152–221), which covered 11.73%, 21.63%, 36.88%, and 29.76% of the basin, respectively. The corresponding results are shown in Figure 4.

Compared with the results of the DRASTIC model, the proportion of high and moderate degrees had obviously increased. The high and moderate vulnerability zones encompassed the previous subarea of the DRASTIC model and the enlargement region influenced by intensive human

activities. Considering the effect of land use on the groundwater nitrate pollution, a part of the original low and very low intrinsic vulnerability zones from DRASTIC transformed into moderate or low specific vulnerability classifications. The result of the DRASTIC-LU model suggested a better agreement with the nitrate contamination compared with that of the DRASTIC model. The use of land cover indeed influenced the occurrence of nitrate contamination in the basin.

Improved vulnerability map based on SVM

The rating scale and assigned weights of the DRASTIC and DRASTIC-LU models are subjective. In order to avoid the bias and combine with the obvious advantages of the DRASTIC model, the SVM method was introduced to improve the standard method.

As a supervised machine learning technique, the verification process of SVM included comparison of the vulnerability index obtained using DRASTIC + SVM and DRASTIC-LU + SVM with the nitrate concentration. The classified 103 groundwater samples were used to realize this machine learning process. The dataset was randomly partitioned into subsets, where 70% of the data were used to train and 30% to test the model (Naghibi et al. 2017).

The receiver operating characteristic (ROC) curve was used for evaluating and visualizing the performance of SVM. The ROC curve is typically generated by plotting the false positive rate against the true positive rate at different threshold settings (Asefa et al. 2006). The area under the curve (AUC) is an index, which provides a measure of the predictive capability. The AUC values were 0.801 and 0.837 for the modified DRASTIC and DRASTIC-LU models, respectively. Therefore, the accuracy of these two SVM models indicated high reliability (Huan et al. 2012). The final results of the modified model using SVM are presented in Figures 5 and 6.

The AI method made the classification region more decentralized and the rate of high and moderate vulnerability zones became bigger. The DRASTIC + SVM model showed visible variations, as compared with the original DRASTIC model: high, moderate, low, and very low classes, which covered 11.78%, 15.75%, 30.19%, and 42.28% of the basin, respectively. The results of DRASTIC + SVM

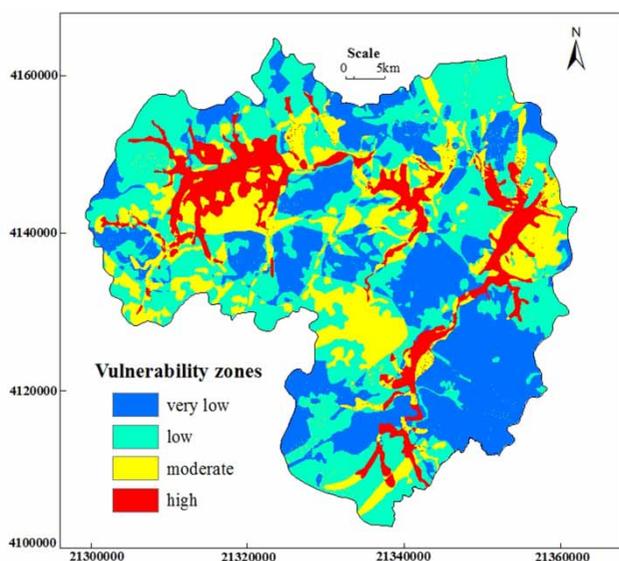


Figure 4 | The result of the DRASTIC-LU model.

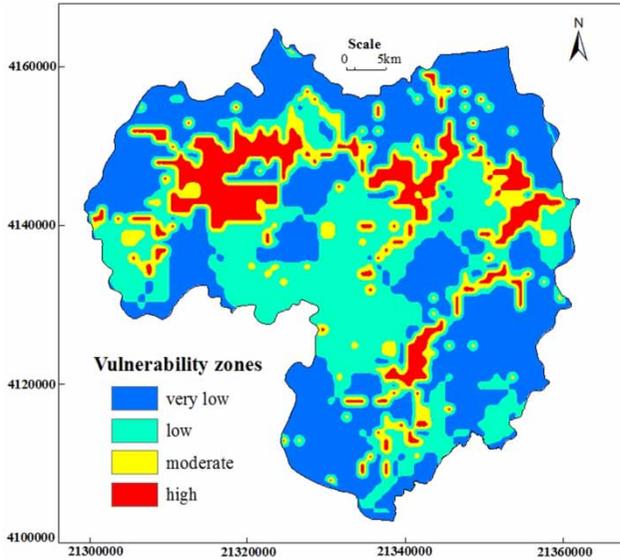


Figure 5 | The result of the DRASTIC model combined with SVM.

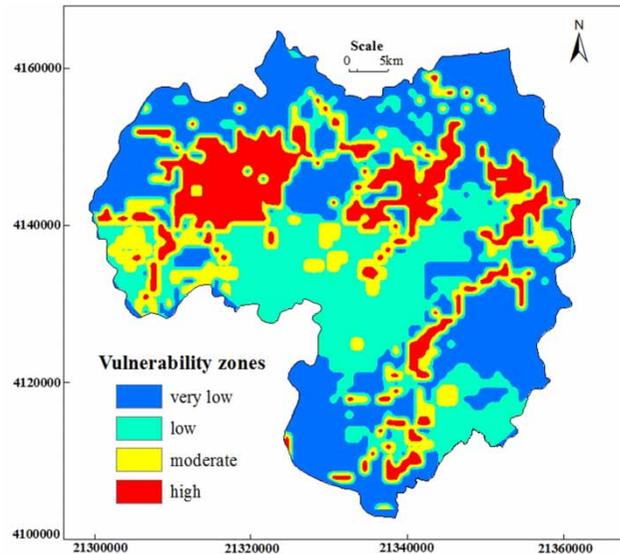


Figure 6 | The result of the DRASTIC-LU model combined with SVM.

were more consistent with the nitrate pollution. The DRASTIC-LU + SVM model showed high and moderate vulnerability categorized regions, accounting for 13.11% and 16.70% of the total area, respectively. Furthermore, 31.32% and 38.87% were divided into the low and very low vulnerability zones, respectively. The results of the improved DRASTIC-LU + SVM model were better than those of the DRASTIC-LU model.

Comparison of the nitrate distribution and vulnerability maps

For comparing the groundwater vulnerability assessment and improving the predicted results further, the classified collected 103 samples were used. Further details about the results of the different methods are presented in Table 4. According to the number of properly classified samples, the specific vulnerability results became generally more unified than the intrinsic vulnerability. Meanwhile, the integrated model using SVM became better than the common overlay and index method.

In order to make the results of different models comparable, the classification was further quantified: the values of high, moderate, low and very low classifications were set to be 4, 3, 2 and 1, respectively. The Root Mean Squared Error (RMSE) was introduced to compare the classification performance, using Equation (9). The smaller the values of RMSE, the better the prediction results were. The results of RMSE are presented in Table 4.

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (P_i - O_i)^2}{N}} \tag{9}$$

Table 4 | Comparison of different models with nitrate concentration

Model	Class	Number of water sample (based on the nitrate concentration)				RMSE
		Very low	Low	Moderate	High	
DRASTIC	Very low	23	9	3	1	0.853
	Low	16	11	3	2	
	Moderate	2	1	14	8	
	High	0	0	1	9	
DRASTIC-LU	Very low	19	3	0	1	0.757
	Low	19	14	1	1	
	Moderate	3	3	15	5	
	High	0	1	5	13	
DRASTIC + SVM	Very low	28	1	0	0	0.631
	Low	10	13	3	1	
	Moderate	3	6	17	4	
	High	0	1	1	15	
DRASTIC-LU + SVM	Very low	32	0	0	0	0.502
	Low	8	13	2	0	
	Moderate	1	5	19	2	
	High	0	2	0	18	

Here P_i and O_i are the prediction and target values, and n is the number of total samples.

As can be seen from Table 4, compared with the DRASTIC model, the DRASTIC-LU model was more suitable to assess nitrate pollution vulnerability, and the values of RMSE for the two models were 0.853 and 0.757, respectively. The modified SVM models indicated a similar phenomenon. The RMSE value of DRASTIC-LU + SVM (0.502) was smaller than the value of DRASTIC + SVM (0.631). Land use influences the distribution of groundwater nitrate pollution. Moreover, the results also show that the application of SVM significantly increases the correlation between the vulnerability and the nitrate pollution. Generally, the groundwater vulnerability assessment results obtained using the modified models were more reliable than the original DRASTIC model.

CONCLUSIONS

Due to rapid population and economic growth, groundwater nitrate pollution has become a serious and socio-economic issue. Assessment of groundwater nitrate pollution vulnerability is a vital tool for preventing the deterioration of groundwater resources. The classic DRASTIC model is a widely used and cost-effective tool. However, ratings and weights of conditioning factors are made through experiences of assessment experts. This paper presented a novel integrated model for the assessment of groundwater vulnerability. The DRASTIC model was firstly used as a standard method to analyze the intrinsic vulnerability. The standard method with the addition of land-use was improved to assess the specific vulnerability using the DRASTIC-LU model. Meanwhile, in order to avoid the disadvantages while using index-overlay methods, SVM was introduced to further improve the accuracy of the proposed models (namely the DRASTIC + SVM and DRASTIC-LU + SVM models). In addition, 103 groundwater samples were collected to validate and test the proposed methods.

Different from the results of the standard DRASTIC model, the improved methods were more suitable to assess the groundwater vulnerability to nitrate. Compared with

the DRASTIC model with the RMSE of 0.853, the DRASTIC-LU model obtained a smaller RMSE value of 0.757. Models combined with SVM exhibited similar results. The model including the land use factor was more precise than the original one. Meanwhile, the RMSE values of DRASTIC + SVM and DRASTIC-LU + SVM (0.631 and 0.502, respectively) were smaller than the RMSE values of the two corresponding overlay and index models. The vulnerability maps improved using the SVM for the original index method and were also more efficient.

Based on the assessment results, the high vulnerability zone was mostly distributed along the riverbeds, which is mostly dominated by coarse sand and sand. The moderate subarea was related to the lithology of the aquifer and unsaturated zone. The low and very low vulnerability subareas were mainly situated at the hilly region. Meanwhile, the groundwater pollution presented obvious spatial distribution characteristics, which are related to intensive human activities that should be taken into consideration. Generally, the study used a supervised AI technique to build the robust model that obviously improved the accuracy of assessment. The results of the assessment will provide essential information for future strategies on groundwater management and land use planning.

ACKNOWLEDGEMENTS

This research was supported by Natural Science Foundation of Science and Technology Department in Hebei Province (D2019403194), the Graduate Students Teaching Case of Hebei Province (KCJSZ2019090, KCJSZ2019092, and KCJSZ2021098), and the Teaching Research Project of Hebei GEO University (2018J28). The authors are indebted to the anonymous reviewers and the editors who have done a lot of work to improve the paper.

DATA AVAILABILITY STATEMENT

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

REFERENCES

- Arabgol, R., Sartaj, M. & Asghari, K. 2016 Predicting nitrate concentration and its spatial distribution in groundwater resources using support vector machines (SVMs) model. *Environmental Modeling & Assessment* **21** (1), 71–82.
- Asefa, T., Kembrowski, M., McKee, M. & Khalil, A. 2006 Multi-time scale stream flow predictions: the support vector machines approach. *Journal of Hydrology* **318** (1–4), 7–16.
- Caprario, J., Rech, A. S. & Finotti, A. R. 2019 Vulnerability assessment and potential contamination of unconfined aquifers. *Water Science & Technology: Water Supply* **19** (4), 1008–1016.
- Chen, W., Peng, J., Hong, H., Shahabi, H., Pradhan, B., Liu, J., Zhu, A. X., Pei, X. & Duan, Z. 2018 Landslide susceptibility modelling using GIS-based machine learning techniques for Chongren County, Jiangxi Province, China. *Science of the Total Environment* **626**, 1121–1135.
- Choubin, B., Mosavi, A., Alamlarloo, E. H., Hosseini, F. S., Shamshirband, S., Dashtekian, K. & Ghamisi, P. 2019 Earth fissure hazard prediction using machine learning models. *Environmental Research* **179**, 108770.
- Cortes, C. & Vapnik, V. 1995 Support-vector networks. *Machine Learning* **20** (3), 273–297.
- Deng, W., Yao, R., Zhao, H., Yang, X. & Li, G. 2019 A novel intelligent diagnosis method using optimal LS-SVM with improved PSO algorithm. *Soft Computing* **23** (7), 2445–2462.
- Deng, W., Xu, J., Song, Y. & Zhao, H. 2020 An effective improved co-evolution ant colony optimisation algorithm with multi-strategies and its application. *International Journal of Bio-Inspired Computation* **16** (3), 158–170.
- Huan, H., Wang, J. & Teng, Y. 2012 Assessment and validation of groundwater vulnerability to nitrate based on a modified DRASTIC model: a case study in Jilin City of Northeast China. *Science of the Total Environment* **440**, 14–23.
- Jia, Z., Bian, J., Wang, Y., Wan, H., Sun, X. & Li, Q. 2019 Assessment and validation of groundwater vulnerability to nitrate in porous aquifers based on a DRASTIC method modified by projection pursuit dynamic clustering model. *Journal of Contaminant Hydrology* **226**, 103522.
- Kim, Y. J. & Hamm, S.-Y. 1999 Assessment of the potential for groundwater contamination using the DRASTIC/EGIS technique, Cheongju area, South Korea. *Hydrogeology Journal* **7** (2), 227–235.
- Li, X., Ye, S., Wang, L. & Zhang, J. 2017 Tracing groundwater recharge sources beneath a reservoir on a mountain-front plain using hydrochemistry and stable isotopes. *Water Science & Technology: Water Supply* **17** (5), 1447–1457.
- Machiwal, D., Jha, M. K., Singh, V. P. & Mohan, C. 2018 Assessment and mapping of groundwater vulnerability to pollution: current status and challenges. *Earth-Science Reviews* **185**, 901–927.
- Naghibi, S. A., Ahmadi, K. & Daneshi, A. 2017 Application of support vector machine, random forest, and genetic algorithm optimized random forest models in groundwater potential mapping. *Water Resources Management* **31** (9), 2761–2775.
- Pacheco, F. A. L., Pires, L. M. G. R., Santos, R. M. B. & Sanches Fernandes, L. F. 2015 Factor weighting in DRASTIC modeling. *Science of the Total Environment* **505**, 474–486.
- Pradhan, B. 2013 A comparative study on the predictive ability of the decision tree, support vector machine and neuro-fuzzy models in landslide susceptibility mapping using GIS. *Computers & Geosciences* **51**, 350–365.
- Raghavendra, S. & Deka, P. C. 2014 Support vector machine applications in the field of hydrology: a review. *Applied Soft Computing Journal* **19**, 372–386.
- Rajput, H., Goyal, R. & Brighu, U. 2020 Modification and optimization of DRASTIC model for groundwater vulnerability and contamination risk assessment for Bhiwadi region of Rajasthan, India. *Environmental Earth Sciences* **79** (6), 136.
- Sajedi-Hosseini, F., Malekian, A., Choubin, B., Rahmati, O., Cipullo, S., Coulon, F. & Pradhan, B. 2018 A novel machine learning-based approach for the risk assessment of nitrate groundwater contamination. *Science of the Total Environment* **644**, 954–962.
- Secunda, S., Collin, M. L. & Melloul, A. J. 1998 Groundwater vulnerability assessment using a composite model combining DRASTIC with extensive agricultural land use in Israel's Sharon region. *Journal of Environmental Management* **54** (1), 39–57.
- Singha, S. S., Pasupuleti, S., Singha, S., Singh, R. & Venkatesh, A. S. 2019 A GIS-based modified DRASTIC approach for geospatial modeling of groundwater vulnerability and pollution risk mapping in Korba district, Central India. *Environmental Earth Sciences* **78** (21), 628.
- Tehrany, M. S., Pradhan, B. & Jebur, M. N. 2015 Flood susceptibility analysis and its verification using a novel ensemble support vector machine and frequency ratio method. *Stochastic Environmental Research and Risk Assessment* **29** (4), 1149–1165.
- Thapa, R., Gupta, S., Guin, S. & Kaur, H. 2018 Sensitivity analysis and mapping the potential groundwater vulnerability zones in Birbhum district, India: a comparative approach between vulnerability models. *Water Science* **32** (1), 44–66.
- Yoon, H., Jun, S. C., Hyun, Y., Bae, G. O. & Lee, K. K. 2011 A comparative study of artificial neural networks and support vector machines for predicting groundwater levels in a coastal aquifer. *Journal of Hydrology* **396** (1–2), 128–138.

First received 21 July 2020; accepted in revised form 7 January 2021. Available online 20 January 2021