Science of the Total Environment, Volume 688, Issue October, 2019, pp. 855-866
DOI:10.1016/j.scitotenv.2019.06.320

Predicting uncertainty of machine-learning models for modelling nitrate pollution of groundwater using quantile regression and UNEEC methods

Omid Rahmati [a,b], Bahram Choubin[c], Abolhasan Fathabadi[d], Frederic Coulon[e], Elinaz Soltani[f], Himan Shahabi[g], Eisa Mollaefar[h], John Tiefenbacher[i], Sabrina Cipullo[e], Baharin Bin Ahmad[j], Dieu Tien Bui [k,*]

[a] Geographic Information Science Research Group, Ton Duc Thang University, Ho Chi Minh City, Viet Nam

[b] Faculty of Environment and Labour Safety, Ton Duc Thang University, Ho Chi Minh City, Viet Nam

[c] Faculty of Natural Resources, University of Tehran, Karaj, Iran

[d] Department of Range and Watershed Management, Gonbad Kavous University, Gonbad Kavous, Golestan Province, Iran

[e] Cranfield University, School of Water, Energy and Environment, Cranfield, MK43 0AL, UK

[f] Department of Natural Resources and Environmental Engineering, College of Agriculture, Shiraz University, Shiraz, Iran

[g] Department of Geomorphology, Faculty of Natural Resources, University of Kurdistan, Sanandaj, Iran

[h] Department of Natural Resources and Management of Golestan Province, Iran

[i] Department of Geography, Texas State University, San Marcos, TX 78666, USA

[j] Faculty of Built Environment and Surveying, Universiti Teknologi Malaysia (UTM), 81310 Johor Bahru, Malaysia

[k] Institute of Research and Development, Duy Tan University, Da Nang 550000, Viet Nam

Corresponding author's Email address: Dieu.T.Bui@usn.no

Abstract

Although estimating the uncertainty of models used for modelling nitrate contamination of groundwater is essential in groundwater management, it has been generally ignored. This issue motivates this research to explore the predictive uncertainty of machine-learning (ML) models in this field of study using two different residuals uncertainty methods: quantile regression (QR) and uncertainty estimation based on local errors and clustering (UNEEC). Prediction-interval coverage probability (PICP), the most important of the statistical measures of uncertainty, was used to evaluate uncertainty. Additionally, three state-of-the-art ML models including support vector machine (SVM), random forest (RF), and *k*-nearest neighbor (*k*NN) were selected to spatially model groundwater nitrate concentrations. The models were calibrated with nitrate concentrations from 80 wells (70% of the data) and then validated with nitrate concentrations from 34 wells (30% of the data). Both uncertainty and predictive performance criteria should be considered when comparing and selecting the best model. Results highlight that the *k*NN model is the best model because not only did it have the lowest uncertainty based on the PICP statistic in both the QR (0.94) and the UNEEC (in all clusters, 0.85–0.91) methods, but it also had predictive performance statistics (RMSE=10.63, R^2= 0.71) that were relatively similar to RF (RMSE= 10.41, R^2= 0.72) and higher than SVM (RMSE= 13.28, R^2= 0.58). Determining the uncertainty of ML models used for spatially modelling groundwater-nitrate pollution enables managers to achieve better risk-based decision making and consequently increases the reliability and credibility of groundwater-nitrate predictions.

Keywords: Groundwater pollution; Uncertainty assessment; Nitrate concentration; Machine learning; GIS.

1. Introduction

Nitrate pollution of groundwater can have severe impacts on human health, issues such as cancer (e.g., esophageal, lymphatic, and gastric cancers) and methemoglobinemia in infants and pregnant women (Suthar et al., 2009; Panagopoulos et al., 2011; Sajil et al., 2014), and on the environment, such as causing ecological disruption and eutrophication throughout the hydrological system

(Neshat et al., 2015; Wheeler et al., 2015; Stelzer and Scott, 2018). Groundwater-pollution modelling can aid managers of water resources and environmental protection in their quests to prevent groundwater pollution and to improve its quality (Almasri, 2008; Takizawa, 2008; Locatelli et al., 2019).

Recent studies have adopted novel approaches to assess groundwater contamination and to map nitrate hazards using different machine-learning (ML) and/or data-mining models including; artificial neural networks (ANN) (Panagopoulos et al., 2011; Ostad-Ali-Askari et al., 2017), boosted regression trees (BRT) (Ransom et al., 2017), support-vector machines (SVM) (Rodriguez-Galiano et al., 2018), random forests (RF) (Anning et al., 2012; Nolan et al., 2014; Rodriguez-Galiano et al. 2014, 2018; Wheeler et al., 2015), classification and regression trees (CART) (Rodriguez-Galiano et al., 2018), Dempster–Shafer (DS) (Rahmati and Melesse, 2016), and multivariate discriminant analysis (MDA) (Sajedi-Hosseini et al., 2018). Nolan et al. (2015) compared the capability of BRT, ANN, and Bayesian networks (BN) to predict nitrate concentration (in shallow groundwater of the Central Valley, California), but did not investigate or consider uncertainties. Sajedi-Hosseini et al. (2018) used ensemble modelling, including the three ML models BRT, MDA, and SVM, to assess and produce groundwater-pollution qualitative maps from a dataset that merely indicated the presence or absence of pollution. Ransom et al. (2017) developed a hybrid ML model combining both numerical and empirical outputs for the Central Valley Textural Model (CVTM) and the BRT groundwater reduction-oxidation (redox) model to predict nitrate concentration in the Central Valley aquifer, California. Results from the hybrid model which included 25 predictors (final model) provided a higher accuracy compared to ordinary kriging, universal kriging, and multiple linear regression. In another study, Messier et al. (2019) used the RF model for classification modelling using transformed nitrate values assigned to three categories of <1 mg/L, 1–5 mg/L, and ≥5 mg/L to predict groundwater nitrate of 22,000 private wells in North Carolina. The RF classification model had superior performance than censored maximum-likelihood regression (CMLR), RF, SVM, ANN, OK, and gradient-boosted machine (GBM) methods. Juntakut et al. (2019) estimated long-term nitrate concentrations in groundwater using the CART model in eastern Nebraska. The CART model achieved success in terms of both nitrate prediction and identification of the potential factors associated with higher nitrate-contamination zones. Finally, Knoll et al. (2019) compared the performance of four ML including MLR, CART, RF, and BRT for predicting nitrate concentrations of groundwater in Hesse state, Germany. Their RF model outperformed the others.

Although most of these studies generally improved the ability to predict nitrate concentration in groundwater, to the best of our knowledge, the assessment of uncertainty associated with groundwater pollution modelling has been disregarded. Indeed, the above studies only evaluated models' performance and disregarded models' uncertainties. It is well known that uncertainty is inherent in modelling (Solomatine and Shrestha, 2009), and therefore it is critical to report it transparently in decision-support tools (Uusitalo et al., 2015).

There are various sources of uncertainty; it can be related to predictors, model parameters, and model structure, etc. (Solomatine and Shrestha, 2009). Importantly, most of the uncertainty-assessment methods deal only with single sources of uncertainty. For instance, Bayesian methods only analyze the uncertainty associated with input data and Monte Carlo methods only assess the uncertainty in parameters (Solomatine and Shrestha, 2009). Since the contribution of different sources of errors is not completely known and separating their roles is often difficult, especially in hydrogeology, an overall assessment of uncertainty is, in practice, feasible. Understanding the total model uncertainty rather than the uncertainty resulting from individual sources is more important for decision-makers, particularly those in water resources management (Solomatine and Shrestha, 2009). In this paper we consider two methods, quantile regression (QR) (Basset and Koenker, 1978) and uncertainty estimation based on local errors and clustering (UNEEC) (Shrestha et al., 2006; Shrestha and Solomatine, 2006), to quantify the uncertainty of modelling groundwater pollution. Although both methods account for all sources of uncertainty, they differ in their methodological complexity. QR and UNEEC have been used in a broad range of applications: hydrological studies (e.g., Weerts et al., 2011; López López et al., 2014; Dogulu et al., 2015), economics (Taylor, 2007; Kudryavtsev, 2009), meteorology (Friederichs and Hense, 2007; Cannon, 2011), wind forecasting (Møller et al., 2008), and agriculture (Barnwal and Kotani, 2013). To predict groundwater-nitrate concentration, three state-of-the-art ML models – SVM, RF, and *k*NN were used to model groundwater-nitrate concentrations spatially. The main objectives of this research are to: 1) quantify the predictive uncertainty of different ML models –SVM, RF, and *k*NN– to model groundwater-nitrate concentrations with QR and UNEEC; 2) determine the most robust model in terms of predictive uncertainty and capability; and 3) assess the relationships between geo-environmental factors and groundwater-nitrate concentration.

2. Material and Methods

2.1 Study area and data sets

The study area is the Andimeshk-Dezful region, Khuzestan province, Iran (Fig. 1). The region covers an area of 2464.75 km^2 between 48°01' and 48°46' E and 31°58' and 32°33' N and contains approximately 385,000 residents. Unconsolidated surface material in the region derives primarily from the Quaternary as low-level pediment-fans and valley-terrace deposits (Qft2). This region is part of the Zagros Structural Zone (Heyvaert and Baeteman, 2007; Rahmati and Melesse, 2016). The climate is semi-arid with about 341 mm of annual precipitation. Summer is usually hot and dry. Winter is when the greatest portion of the region's precipitation falls (about 200 mm). Mean daily minimum temperature is 7.5 °C in winter and the mean daily maximum temperature is 46 °C in summer. In Iran, groundwater is the primary water source; over 85% is used for drinking water and for agriculture (irrigation). The study of the groundwater pollutants like nitrate can aid decision makers' control and management of water quality.

Groundwater-nitrate concentrations were measured by the Iranian Department of Water Resources Management (IDWRM) at 114 locations during May 2017 (Fig. 1). The highest nitrate values (≥75 mg/l) are in the southern parts of the region. There are patches in the western and northwestern sections of the study area that have nitrate concentrations that exceed the standard (≤50 mg/l) for safe drinking water (WHO, 2011). Data describing several geo-environmental variables were compiled for the study region: elevation (m), hydraulic conductivity (m/s), distance from stream (m), lineament density (km/km^2), and land use.

Fig. 1 Here

2.2 Methodology

The geo-environmental variables (Fig. 2) as inputs for modelling the groundwater nitrate concentration using three ML methods including SVM, RF, and kNN (these models are described in the section 2.2.2). The models were calibrated and validated (with a ratio 70 to 30) using the target value of nitrate concentration and values of the predictive factors at the location of each well. After ensuring the models' performance, groundwater-nitrate concentrations were predicted for the other parts of the region (areas without recorded nitrate concentrations). More details of the methods used are described below.

2.2.1 Groundwater nitrate conditioning factors

These five groundwater-nitrate conditioning factors –elevation, hydraulic conductivity (K), distance from stream (DFS), lineament density, and land use– were input as potential predictors of nitrate concentrations.

Elevation: A digital elevation model (DEM), with pixel size 10×10 m was obtained from IDWRM. The region slopes from north to south; elevation varies from 253 to 16 m asl (Fig. 2a).

Hydraulic conductivity (K): Hydraulic conductivity influences subsurface flow rates, groundwater recharge, and the mobility of contaminants in the saturated zone (Bouwer, 2002; Jiang et al., 2010). Hydraulic conductivity data were obtained from the IDWRM and estimated based on a combination of pumping test (with piezometers and observation wells) and geoelectrical measurements (Fig. 2b).

Distance from stream (DFS): The interface between surface water and groundwater in rivers and streams is an active area of nitrate removal and retention (Hedin et al., 1998), so DFS could be a good predictor of nitrate concentrations. DFS was calculated using the DEM and the Euclidean tool in ArcGIS, the maximum DFS in this study was 11705 m (Fig. 2c).

Lineament density: Lineament density indirectly reflects groundwater potential as lineaments generally denote a permeable zone (Magesh et al., 2012). Detection of lineaments, surface-subsurface structures (*e.g.,* geologic faults, fractures, etc.), is regarded as necessary for groundwater studies (*e.g.,* Gupta and Srivastava, 2010; Oh et al., 2011; Nampak et al., 2014). Lineaments have been defined by Hobbs (1904) as "significant lines of landscape that reveal the hidden architecture of the rock basement." Lineament structures are formed by a variety of geological and geomorphological processes and their nature are related to faulting in Earth's crust (Jordan and Schott, 2005; Nkono et al., 2018). In this study, the automatic method (MATLAB-based code) introduced by Soto-Pinto (2013) was used to detect lineaments patterns in Landsat 8 images (2015-2016) and to produce a density map using the line-density tool in ArcGIS software (Fig. 2d).

Land use: A land use and land cover map of the study was obtained from IDWRM (Fig. 2e). Surfaces are classified as bare land, dry farming, range land, riparian zone, urban, wetland, and irrigated agriculture; this last category covers the largest portion of the study area (Fig. 2e).

Fig. 2 Here

2.2.2 Machine learning models

Three ML models (SVM, RF, and *k*NN) were selected that are commonly used in groundwater-pollution modelling (e.g., Rodriguez-Galiano et al., 2018). It is beyond the scope of this study to evaluate and compare the performance of the ML models used here, however, to quantify uncertainties of the models, their predictive performance needs to be determined.

Support vector machine (SVM): SVM is a technique that uses statistical learning theory (Dixon and Candade, 2008), first introduced by Vapnik et al. (1997). It is one of the most cogent prediction methods using the dimension theory of Vapnik Chervonenk and the structural-risk minimization principle, and it can be used to solve problems in quadratic programming (Cortes and Vapnik, 1995). This classification method is a non-parametric statistical monitoring method (Mountrakis et al., 2011) that forms and reforms the boundaries of classes using an optimization algorithm

(Sajedi-Hosseini et al., 2018). In this study, the most popular kernel function (i.e., radial) was implemented with the e1071 package (Meyer et al., 2017) in R software. The 'tune' function tunes the kernel parameters with a grid search of parameter ranges. In this study, the best values for the parameters Gamm and Cost were 0.041 and 9.19, respectively. They were determined using the 'tune' function in e1071.

Random forest (RF): RF, an ensemble-tree method developed by Breiman (2001), can identify linear and nonlinear relationships between variables for classification and regression objectives (Elith et al., 2008). For regression objectives, RF can accurately produce the conditional mean of a dependent variable. It generates many decision trees and aggregates the predictions through bootstrap aggregation by averaging the predictions obtained from multiple decision trees (Liaw and Wiener, 2002; Hastie et al., 2009). In this study, RF was deployed in R software using the 'randomforest' package (Liaw and Wiener, 2002). The key parameters of the models, the number of trees and the best size of the nodes, were optimized with the objective function of root mean squared error (RMSE). The RF model explicitly measures the importance of variables through two metrics: the mean decrease in the Gini Index (GI) and the percentage of increase in RMSE (Hollister et al., 2016). Since the GI has a bias in its calculation of variable importance (Strobl et al., 2007; Hollister et al., 2016), we measured the importance of variables through the percentage of increase in RMSE using RF. The results of RF indicate that the optimum number of trees considered is 2000 and the best node size is 13.

The k-nearest neighbor (kNN): *k*NN is a non-parametric model able to identify non-linear and complex relationships among observations (McRoberts et al., 2007; Mansuy et al., 2014). In this method, a metric (Euclidean distance) is used to measure the similarity of distances to the target. In *k*NN there are two parameters, nearest neighbors (*k*) and the power term (*p*), that are used to design this approach. The value of *k* can be determined from a reference dataset and input variables, while *p* is based on weight–distance relationships and measures the degree of similarity of the contribution of each *k* to the simulation output (Botula et al., 2013). A trial-and-error methodology was used to find the optimal value of the *k*NN model parameters to predict the nitrate concentration. The best value for the power parameter in this study is 2.25 and for the number of nearest-neighbors is 14 (based on the objective function of RMSE).

2.2.3 Accuracy assessment

After calibration, the models were validated with the 30% of cases that were not used for training. Model accuracy was evaluated using RMSE and the coefficient of determination (R^2). Moreover, a graphical comparison was conducted using Taylor diagrams (Taylor, 2001), which enable visualization of the models' performances using correlation coefficients, RMSE, and standard deviations (SD) (Choubin et al., 2017).

2.2.4 Uncertainty assessment

The quantile regression (QR) and the uncertainty estimation based on local errors and clustering (UNEEC) methods were used to assess the predictive uncertainty of the models. These methods evaluate the model residuals and consider all sources of uncertainty, which is in contrast to the classic methods (such as Monte Carlo-based methods) in which the estimate usually regards only one source of uncertainty (Solomatine and Shrestha, 2009).

The QR was originally developed by Basset and Koenker (1978) for economics applications. It can be used to determine the distribution error. QR is a linear statistical method for estimating the quantiles conditional functions, of the prediction and distribution, based on possible causal relationships within the entire data set (Koenker and Hallock, 2001). The method describes the conditional quantiles distribution as functions of observed covariates and does not make any presumptions about the shape of the distribution the data. In this method, for each quantile τ, there is a linear relationship between the observed (y) and predicted ($\hat{y}$) data as Eq. 1:

$$y = a_\tau \hat{y} + b_\tau \qquad (1)$$

where $a_\tau$and b_τ are the slope and intercept of the QR, which are calculated by minimizing the sum of residuals (Eq. 2):

$$min \sum_{j=1}^{J} \rho_\tau \left(y_j - \left(a_\tau \hat{y}_j + b_\tau\right)\right) \qquad (2)$$

where $y_j$and $\hat{y}_j$are jth sample from a dataset, and ρ_τ is the QR function of the τth quantile:

$$\rho_\tau(\in_j) = \begin{Bmatrix} (\tau - 1). \in_j, & \in_j < 0 \\ \tau. \in_j, & \in_j \geq 0 \end{Bmatrix} \qquad (3)$$

The QR function (Eq. 3) is used for the residuals ($\in_j$) which are the differences between the observed and predicted data for the selected quantile τ (here quantiles of 5% to 95%). So, Eq. 3 can be used for calculation of any quantile τ. To estimate uncertainty using QR, all individual ML methods were trained using the training dataset (inputs and output variables) and the outputs (nitrate-concentration values) for all cells in the study area were calculated. Then, for each desired quantile (i.e., 0.05 and 0.95), the QR model was calibrated using the predicted nitrate values (from each ML model) as input values and the observed nitrate values from training data set as output values. Finally, the desired quantiles (i.e., 0.05 and 0.95) of the nitrate values were calculated with the calibrated QR model based on the predicated nitrate values (by each ML model) as inputs for the whole study area. In this study, the QR was conducted in the R software using the 'quantreg' package (Koenker, 2013).

UNEEC, a non-linear regression model, was introduced by Shrestha et al. (2006) and Shrestha and Solomatine (2006) and can be used to estimate the error-distribution quantiles. UNEEC deduces the residual uncertainty relying on the status of the simulated system with a clustering method. Fuzzy c-means clustering, a soft-clustering method, has the capacity to reduce the uncertainty in identifying the members of a cluster (Dodangeh et al., 2014). Therefore, it was used in the UNEEC method. There are several steps that occur (Solomatine and Shrestha, 2009): (i) training a ML model based on the predictors (inputs) and the nitrate values (outputs) and calculated residuals, (ii) clustering the input vectors (the values of the predictors for each sample) and associated residuals using fuzzy c-means clustering, (iii) constructing the empirical probability distribution function (pdf) of the model errors for each cluster. To construct the empirical probability distribution function, the clusters were sorted, in ascending order, by the value of the error in each, then, using Eq. 4, quantiles (for example pth quantile) were estimated:

$$ec_i^p = \epsilon_t, \qquad t: \sum_{k=1}^{t} \mu_{i,k} < p \sum_{t=1}^{n} \mu_{i,t} \qquad (4)$$

where ϵ_t is error correspond to observation t, t is maximum value that satisfied the above inequality, $\mu_{i,t}$ is membership value of tth observation to cluster i and ec_i^p is pth quantile associated with cluster i.

The next step (iv), involved the calculation of the membership values of each input vector of the training and testing datasets (in each cluster) and the estimation of the associated quantiles of residuals using Eq. 5:

$$e_t^p = \frac{\sum_{i=1}^{c} \mu_{i,t}^{2/m} ec_i^p}{\sum_{i=1}^{c} \mu_{i,t}^{2/m}} \qquad (5)$$

where e_t^p is value of pth quantile of errors for tth input vector, ec_i^p is value of pth quantile of errors for cluster i, $\mu_{i,t}$is the membership function of the tth input vector for cluster i, m is the smoothing exponential coefficient, and c total number of the clusters. (vi) the prediction interval of the model output (nitrate concentration) was constructed with Eq. 6 for each cell in study area:

$$y_t^p = \hat{y}_t + e^p \qquad (6)$$

where y_t^p is pth quantile for tth output data (nitrate concentration). Quantiles 5% (U^5) and 95% (U^{95}) are necessary to estimate 90% prediction interval. In the current research, UNEEC was run with MATLAB software.

There are several statistical measures of uncertainty to evaluate and compare performances of QR and UNEEC methods. In this study, two statistics, mean prediction interval (MPI) and prediction interval coverage probability (PICP), were used as suggested by Shrestha and Solomatine (2006). MPI is the average of the widths of the prediction intervals, where the lower values of MPI indicate lower uncertainty (i.e., a value of zero indicates no uncertainty) (Eq. 7). PICP is the probability that the observed values are within the prediction intervals (between 5% to 95%); each is computed for a significance level of 1– α (e.g., 90 %) (Eq. 8). The method with a PICP near the confidence level (i.e., 90 % with some tolerance) is the best method. MPI and PICP values are calculated as:

$$MPI = \frac{1}{n}\sum_{t=1}^{n}(PL_t^{upper} - PL_t^{lower}) \qquad (7)$$

$$PICP = \frac{1}{n}\sum_{t=1}^{n} C\,, \qquad C = \begin{Bmatrix} 1, PL_t^{lower} < y_t < PL_t^{upper} \\ 0, otherwise \end{Bmatrix} \qquad (8)$$

where y_t is observed value, PL_t^{lower} and PL_t^{upper} are lower and upper prediction limits respectively.

The PICP is the more important measurement of uncertainty as it indicates the number of observations that fall within the estimated interval (Dogulu et al., 2015). Therefore, MPI is used as a supplementary metric: between models with similar PICP values, the one with a lower MPI is regarded as the better model (Muthusamy et al., 2016).

3. Results and discussion

3.1 Spatial prediction of nitrate concentrations in groundwater

The groundwater-nitrate concentration maps generated by SVM, RF, and *k*NN show similar spatial patterns; they each predict high nitrate concentrations in the southern part of the study area (Fig. 3). The spatial detail of the models differs. The SVM model produced nitrate concentrations between 11 and 104 mg/l (Fig. 3a). Similar to the SVM, the RF model (Fig. 3b) predicted nitrate concentrations increasing from north to south with levels ranging from 20 to 92 mg/l. The *k*NN model also predicted nitrate concentrations increasing from north and east to the south with amounts from 18 to 101 mg/l (Fig. 3c).

Fig. 3 Here

3.3 Uncertainty assessment

Uncertainty bands for each ML model were determined using UNEEC and QR methods (Figures 4 and 5). More observations fall within the estimated interval in *k*NN model than in the other two.

In addition, the results of the UNEEC method and the values of PICP and MPI for all clusters were summarized (Table 1). As explained above, the PICP metric is the most important value to assess uncertainty. In the testing step, the *k*NN model had the lowest uncertainty in each cluster (PICP=0.85–0.91), followed by SVM (PICP= 0.76–0.79) and RF (PICP=0.65–0.68). However, the PICP was closer to the 90 % confidence level for *k*NN model than for either SVM or RF. Since the PICP values for the models were not equivalent, the MPI measure was not considered in judging the certainty of the models.

The QR, like the UNEEC method, calculated that the *k*NN model had the lowest uncertainty (PICP= 0.94) compared to SVM (PICP=0.74) and RF (PICP=0.59) (Table 2). Since the PICP measurements for the three models are very different, there was no need to compare the MPI values. Thus, based on the results of both the UNEEC and QR methods, the *k*NN model contained less uncertainty than did the other models.

Figure 4 Here

Figure 5 Here

Table 1 Here

Table 2 Here

3.4 Evaluating prediction performance

The goodness-of-fit and predictive performance of the models were also quantified using RMSE and R^2 metrics (Table 3). In the training step, the RF model produced a better prediction of groundwater-nitrate concentrations (RMSE=4.69, R^2=0.96) than did SVM (RMSE= 8.76, R^2= 0.82) and *k*NN (RMSE= 10.85, R^2=0.74). The goodness-of-fit of the model shows how well the model fit the training dataset. The prediction and generalization abilities of the model cannot be evaluated using the goodness-of-fit of the model because it is measured by the data that were used to calibrate the model (Henseler and Sarstedt, 2013). The predictive performance (i.e., the accuracy of the model in the testing step) reflects the ability of the model to accurately predict. Results indicated that the RF (RMSE= 10.41, R^2= 0.72) performance was slightly better than that of the *k*NN model (RMSE= 10.63, R^2= 0.71) and the SVM (RMSE= 13.28, R^2= 0.58). The visualization of the models' performance using the Taylor diagram also confirmed these results (Fig. 6). According to the Taylor criteria (i.e., correlation, standard deviation, and RMSE), the RF and the *k*NN had higher correlations with observed nitrate concentrations and lower RMSE compared to the did the SVM model.

Fig. 6 Here

3.5. A ranking of the models

There are numerous ML models; each has weaknesses, strengths and assumptions. In reality there is no single model that is absolutely correct and always best among the suite of models (Elith et al., 2002). Similarly, different model structures can produce different results (Goetz et al., 2015). Furthermore, models with similar predictive performance levels do not necessarily have similar uncertainties; this attribute may affect environmental management and water resources planning decisions. Therefore, it is difficult to identify the best ML algorithm. In this section, both predictive performance and uncertainty criteria were considered simultaneously to rank the models.

Although the RF model had the highest performance for predicting nitrate concentrations, it contained the greatest uncertainty (i.e., it ranked third). Though there are advantages to using RF (e.g., Anning et al., 2012; Knoll et al., 2019), there are some limitations: i) it underestimates high values and overestimates low values; and ii) it cannot predict beyond the range of response values because of the averaging that it does in all regression trees (Horning, 2010; Hengl et al., 2015; Cheng et al., 2019; Shiferaw et al., 2019). That the uncertainty of models provides insight into groundwater-nitrate pollution management is important. This study provides a practical analysis of predictive performance and uncertainty of ML models to shed light on the spatial modelling of groundwater-nitrate concentration.

SVM had the lowest predictive performance. But the SVM had lower uncertainty than the RF model, but higher than the *k*NN. This result confirms the results of Rodriguez-Galiano et al. (2018) which compared the performance of RF, SVM, and CART models for spatial predictions of nitrate concentrations in groundwater. Their results indicated that the SVM was the least accurate of the three.

The *k*NN can be regarded as the best model (of the three) for spatially modelling groundwater-nitrate pollution as its predictive performance was similar to that of the RF model (i.e., which had the highest predictive performance), but its uncertainty (based on UNEEC and QR analyses) was the lowest of the three. These results affirm those of McRoberts (2012) and Zhang et al. (2013) who demonstrated that *k*NN produces small error ratios and good error distributions. An advantage of the *k*NN is that it does not need to prescribe detailed solutions to the input–output mapping (Liu et al., 2016). Another advantage is its non-parametric nature, making it well suited to analyze non-linear and complex relationships (Nemes et al., 2006; Abedi et al., 2018). Researchers have used the *k*NN because of its capacity to predict a large set of attributes simultaneously (e.g., Mittal et al., 2018; Kuang et al., 2019; Lee et al., 2019). Therefore, it is a cost- and time-effective modelling approach to use in spatially extensive regions (Beaudoin et al., 2014). Though it has these advantages, there are also drawbacks to using the *k*NN model. An important issue is the optimal number of nearest neighbors (*k*) that can affect classification patterns (Jung et al., 2013). Another disadvantage is the underestimation and overestimation of values in the extremes of the range (Magnussen et al., 2010; Beaudoin et al., 2014).

3.5. Variable importance

One of the main advantages of the RF model is it enables assessment of the importance of the predictive factors used in the modelling process. Variable importance is assessed using the calculation of the index of the percentage of increase of MSE (Figure 7). The higher the MSE percentage, the higher is the importance of the variable considered. Results clearly showed that hydraulic conductivity and elevation are the two most important variables for predicting nitrate concentrations in groundwater with MSE equal to 117% and 95%, respectively, after removal of the variable from the modelling.

Fig. 7 Here

These findings are consistent with those of Peña-Haro et al. (2001) who found that the hydraulic conductivity factor has a strong influence on the spatial and temporal migration of nitrate concentration in groundwater and, therefore, on the optimal N-fertilizer use rate. Regional groundwater flow and the stream–aquifer interaction strongly depend on hydraulic conductivity (Erdal and Cirpka, 2016). However, it is a spatially variable factor and is related to porosity and aquifer characteristics (Salamon et al., 2007; Vrettas and Fung, 2015). Therefore, hydraulic conductivity is an important factor for groundwater management and should be discerned using different geostatistical techniques and field investigations (pumping tests, etc.) (Zhou et al., 2014). Moreover, elevation regulates the flushing of nitrates, and it has an important effect on leaching and export (from soil), and transference of nitrate (Creed and Band, 1998; Jiang et al., 2012).

Lower elevations have more opportunities than high elevations to allow infiltration of nitrate into an aquifer, as flat topography encourages water accumulation and allows more time to infiltrate (Sahoo et al., 2016; Shrestha and Luo, 2018). Also, agricultural lands are more often at lower elevations. Nitrogen from agricultural land is therefore more likely available in areas with low elevation. In the other words, highlands are not as favorable for agricultural activities, and thus see lower amounts of nitrogen fertilizer; less nitrate contamination can occur in these areas. Other important variables were land use, distance from stream, and lineament density (Fig. 7). Therefore, the N-fertilizer application would be best guided by the spatial variation of hydraulic conductivity and topographical characteristics.

4. Conclusions

Although several ML models have been used to spatially modelling nitrate concentrations in groundwater, the uncertainty of these models has not yet been investigated in this field of study. After evaluating the predictive capabilities of three ML models (*k*NN, SVM, and RF), the uncertainty of each was determined using the QR and UNEEC methods. The following conclusions can be drawn:

- The results demonstrate that in an evaluation of models in terms of both predictive performance and uncertainty, the determination that a model is the absolute best remains critical. In this study, the predictive performance of the *k*NN (RMSE= 10.63, R^2= 0.71) was similar to the RF (RMSE= 10.41, R^2= 0.72) and more than the SVM (RMSE= 13.28, R^2= 0.58), but its uncertainty determined with the QR and UNEEC methods was lowest. Although the predictive performance of the RF model was superior (slightly better than the *k*NN), it was the inferior model in terms of the uncertainty. Therefore, the *k*NN is the relatively best model for predicting nitrate concentrations in groundwater, considering both its predictive performance and its level of uncertainty.

- Spatial prediction of nitrates showed that it was strongly correlated with the highest hydraulic conductivity and the lowest elevations. The low elevations and high hydraulic conductivity increase the leaching and transfer of the nitrates from the surface and subsurface to groundwater in these regions.
- Because they use algebraic calculations, both QR and UNEEC methods have low running times. Hence, both can be easily used to estimate predictive uncertainty in ML and data-mining models when modelling nitrate concentration of groundwater.
- In this study, the proportions of the training (and validation) datasets (i.e., of well sampling points) was selected according to the literature: 70% for training and 30% for validation purposes. However, the training dataset size may affect model performance and predictive uncertainties. Therefore, it is recommended that further research be conducted that uses other training sample proportions to determine its effects on predictive uncertainty.
- Although the results of the ML models used in this study were good or excellent, the maps produced cannot be regarded as representative for seasonal or interannual fluctuations. Due to a lack of continuous sampling of nitrate concentrations, assessing seasonal and interannual fluctuations of the concentrations is not possible. It was a main limitation of this research, hence, further studies focusing on the role of spatio-temporal variations of nitrate and the attendant uncertainties is suggested.

Acknowledgments

We thank Iranian Department of Water Resources Management (IDWRM) and Department of Geological Survey of Iran (GSI) for providing necessary data and maps. We greatly appreciate the assistance of the associate editor, Prof. José Virgílio Cruz, and anonymous reviewers for their constructive comments that helped us to improve the paper. This research was partially supported by the Geographic Information Science Research Group, Ton Duc Thang University, Ho Chi Minh City, Viet Nam, and Universiti Teknologi Malaysia (UTM) based on Research University Grant (Q.J130000.2527.17H84).

References

Abedi, R., Bonyad, A.E., Moridani, A.Y. and Shahbahrami, A., 2018. Evaluation of IRS and Landsat 8 OLI imagery data for estimation forest attributes using k nearest neighbour non-parametric method. International Journal of Image and Data Fusion, 9(4), 287-301.

Almasri, M. N. 2008. Assessment of intrinsic vulnerability to contamination for Gaza coastal aquifer, Palestine. Journal of Environmental Management, 88(4), 577-593.

Anning, D. W., Paul, A. P., McKinney, T. S., Huntington, J. M., Bexfield, L. M., Thiros, S. A. 2012. Predicted nitrate and arsenic concentrations in basin-fill aquifers of the southwestern United States. US Department of the Interior, US Geological Survey.

Barnwal, P. and Kotani, K., 2013. Climatic impacts across agricultural crop yield distributions: An application of quantile regression on rice crops in Andhra Pradesh, India. Ecological Economics, 87, 95-109.

Bassett, G., Koenker, R., 1978. Asymptotic Theory of Least Absolute Error Regression. J. Am. Stat. Assoc. 73, 618–622. https://doi.org/10.1080/01621459.1978.10480065

Beaudoin, A., Bernier, P.Y., Guindon, L., Villemaire, P., Guo, X.J., Stinson, G., Bergeron, T., Magnussen, S. and Hall, R.J., 2014. Mapping attributes of Canada's forests at moderate resolution through k NN and MODIS imagery. Canadian Journal of Forest Research, 44(5), 521-532.

Botula, Y. D., Nemes, A., Mafuka, P., Van Ranst, E., Cornelis, W. M. 2013. Prediction of water retention of soils from the humid tropics by the nonparametric k-nearest neighbor approach. Vadose zone journal, 12(2).

Bouwer, H. 2002. Artificial recharge of groundwater: hydrogeology and engineering. Hydrogeology Journal, 10(1), 121-142.

Breiman, L. 2001. Random forests. Machine learning, 45(1), 5-32.

Cannon, A.J., 2011. Quantile regression neural networks: Implementation in R and application to precipitation downscaling. Computers & geosciences, 37(9), 1277-1284.

Cheng, X., Boiyo, R., Zhao, T., Xu, X., Gong, S., Xie, X. and Shang, K., 2019. Climate modulation of Niño3. 4 SST-anomalies on air quality change in southern China: Application to seasonal forecast of haze pollution. Atmospheric Research, 225, 157-164.

Choubin, B., Malekian, A., Samadi, S., Khalighi-Sigaroodi, S., Sajedi-Hosseini, F. 2017. An ensemble forecast of semi-arid rainfall using large-scale climate predictors. Meteorological Applications, 24(3), 376-386.

Cortes, C., Vapnik, V. 1995. Support-vector networks. Machine learning, 20(3), 273-297.

Creed, I.F. and Band, L.E., 1998. Export of nitrogen from catchments within a temperate forest: evidence for a unifying mechanism regulated by variable source area dynamics.*Water Resources Research*, *34*(11), pp.3105-3120.

Dixon, B., Candade, N. 2008. Multispectral landuse classification using neural networks and support vector machines: one or the other, or both?. International Journal of Remote Sensing, 29(4), 1185-1206.

Dodangeh, E., Soltani, S., Sarhadi, A., Shiau, J.T. 2014. Application of L-moments and Bayesian inference for low-flow regionalization in Sefidroud basin, Iran. Hydrological Process, 28, 1663–1676.

Dogulu, N., López López, P., Solomatine, D.P., Weerts, A.H. and Shrestha, D.L., 2015. Estimation of predictive hydrologic uncertainty using the quantile regression and UNEEC methods and their comparison on contrasting catchments. Hydrology and Earth System Sciences, 19(7), 3181-3201.

Elith, J., Burgman, M.A. and Regan, H.M., 2002. Mapping epistemic uncertainties and vague concepts in predictions of species distribution. Ecological modelling, 157(2-3), 313-329.

Elith, J., Leathwick, J. R., Hastie, T. 2008. A working guide to boosted regression trees. Journal of Animal Ecology, 77(4), 802-813.

Erdal, D. and Cirpka, O.A., 2016. Joint inference of groundwater–recharge and hydraulic–conductivity fields from head data using the ensemble Kalman filter. Hydrology and Earth System Sciences, 20(1), 555-569.

Friederichs, P. and Hense, A., 2007. Statistical downscaling of extreme precipitation events using censored quantile regression. Monthly weather review, 135(6), 2365-2378.

Goetz, J.N., Brenning, A., Petschko, H. and Leopold, P., 2015. Evaluating machine learning and statistical prediction techniques for landslide susceptibility modeling. Computers & geosciences, 81, 1-11.

Gupta, M., Srivastava, P. K. 2010. Integrating GIS and remote sensing for identification of groundwater potential zones in the hilly terrain of Pavagarh, Gujarat, India. Water International, 35(2), 233-245.

Hedin, L.O., von Fischer, J. C., Ostrom, N. E., Kennedy, B. P. Brown, M. G., Robertson, G. P. 1998. Thermodynamic constraints on nitrogen transformations and other biogeochemical processes at soil-stream interfaces. Ecology, 79, 684-703.

Hengl, T., Heuvelink, G.B., Kempen, B., Leenaars, J.G., Walsh, M.G., Shepherd, K.D., Sila, A., MacMillan, R.A., de Jesus, J.M., Tamene, L. and Tondoh, J.E., 2015. Mapping soil properties of Africa at 250 m resolution: Random forests significantly improve current predictions. PloS one, 10(6), p.e0125814.

Henseler, J. and Sarstedt, M., 2013. Goodness-of-fit indices for partial least squares path modeling. Computational Statistics, 28(2), 565-580.

Heyvaert, V. M. A., Baeteman, C. 2007. Holocene sedimentary evolution and palaeocoastlines of the Lower Khuzestan plain (southwest Iran). Marine Geology, 242(1-3), 83-108.

Hobbs, W. H. 1904. Lineaments of the Atlantic border region. Bulletin of the Geological Society of America, 15(1), 483-506.

Hollister, J.W., Milstead, W.B. and Kreakie, B.J., 2016. Modeling lake trophic state: a random forest approach. Ecosphere, 7(3).

Horning, N., 2010, December. Random Forests: An algorithm for image classification and generation of continuous fields data sets. In Proceedings of the International Conference on Geoinformatics for Spatial Infrastructure Development in Earth and Allied Sciences, Osaka, Japan (Vol. 911). http://wgrass.media.osaka-cu.ac.jp/gisideas10/papers/04aa1f4a8beb619e7fe711c29b7b.pdf

Jiang, R., Woli, K.P., Kuramochi, K., Hayakawa, A., Shimizu, M. and Hatano, R., 2012. Coupled control of land use and topography on nitrate-nitrogen dynamics in three adjacent watersheds. *Catena*, *97*, pp.1-11.

Jiang, X. W., Wan, L., Cardenas, M. B., Ge, S., Wang, X. S. 2010. Simultaneous rejuvenation and aging of groundwater in basins due to depth-decaying hydraulic conductivity and porosity. Geophysical Research Letters, 37(5).

Jordan, G., Schott, B. 2005. Application of wavelet analysis to the study of spatial pattern of morphotectonic lineaments in digital terrain models. A case study. Remote Sensing of Environment, 94(1), 31-38.

Jung, J., Kim, S., Hong, S., Kim, K., Kim, E., Im, J. and Heo, J., 2013. Effects of national forest inventory plot location error on forest carbon stock estimation using k-nearest neighbor algorithm. ISPRS journal of photogrammetry and remote sensing, 81, 82-92.

Juntakut, P., Snow, D.D., Haacker, E.M. and Ray, C., 2019. The long term effect of agricultural, vadose zone and climatic factors on nitrate contamination in the Nebraska's groundwater system. Journal of contaminant hydrology, 220, pp.33-48.

Knoll, L., Breuer, L. and Bach, M., 2019. Large scale prediction of groundwater nitrate concentrations from spatial data using machine learning. Science of The Total Environment, 668, pp. 1317-1327.

Koenker, R. 2013. quantreg: Quantile Regression, R package version 5.05, available at: http://CRAN.R-project.org/package=quantreg.

Koenker, R., Hallock, K. F. 2001. Quantile regression. Journal of economic perspectives, 15(4), 143-156.

Kuang, L., Yan, H., Zhu, Y., Tu, S. and Fan, X., 2019. Predicting duration of traffic accidents based on cost-sensitive Bayesian network and weighted K-nearest neighbor. Journal of Intelligent Transportation Systems, pp.1-14. https://doi.org/10.1080/15472450.2018.1536978.

Kudryavtsev, A.A., 2009. Using quantile regression for rate-making. Insurance: Mathematics and Economics, 45(2), 296-304.

Lee, T.R., Wood, W.T. and Phrampus, B.J., 2019. A Machine Learning (kNN) Approach to Predicting Global Seafloor Total Organic Carbon. Global Biogeochemical Cycles, 33(1), pp.37-46.

Liaw, A., Wiener, M. 2002. Classification and regression by randomForest. R news, 2(3), 18-22.

Liaw, A., Wiener, M. 2002. The randomforest package. R news, 2(3), 18-22.

Liu, K., Li, Z., Yao, C., Chen, J., Zhang, K. and Saifullah, M., 2016. Coupling the k-nearest neighbor procedure with the Kalman filter for real-time updating of the hydraulic model in flood forecasting. International Journal of Sediment Research, 31(2), 149-158.

Locatelli, L., Binning, P.J., Sanchez-Vila, X., Søndergaard, G.L., Rosenberg, L. and Bjerg, P.L., 2019. A simple contaminant fate and transport modelling tool for management and risk assessment of groundwater pollution from contaminated sites. Journal of contaminant hydrology, 221, 35-49.

López López, P., Verkade, J.S., Weerts, A.H. and Solomatine, D.P., 2014. Alternative configurations of quantile regression for estimating predictive uncertainty in water level forecasts for the upper Severn River: a comparison. Hydrology and Earth System Sciences, 18(9), 3411-3428.

Magesh, N. S., Chandrasekar, N., Soundranayagam, J. P. 2012. Delineation of groundwater potential zones in Theni district, Tamil Nadu, using remote sensing, GIS and MIF techniques. Geoscience Frontiers, 3(2), 189-196.

Magnussen, S., Tomppo, E. and McRoberts, R.E., 2010. A model-assisted k-nearest neighbour approach to remove extrapolation bias. Scandinavian Journal of Forest Research, 25(2), 174-184.

Mansuy, N., Thiffault, E., Paré, D., Bernier, P., Guindon, L., Villemaire, P., et al. 2014. Digital mapping of soil properties in Canadian managed forests at 250 m of resolution using the k-nearest neighbor method. Geoderma, 235, 59-73.

McRoberts, R. E., Tomppo, E. O., Finley, A. O., Heikkinen, J. 2007. Estimating areal means and variances of forest attributes using the k-nearest neighbors technique and satellite imagery. Remote Sensing of Environment, 111(4), 466-480.

McRoberts, R.E., 2012. Estimating forest attribute parameters for small areas using nearest neighbors techniques. Forest Ecology and Management, 272, 3-12.

Melchers RE. 1987. Structural reliability: analysis and prediction. Chichester, West Sussex, U.K.: Ellis Horwood Ltd.

Messier, K.P., Wheeler, D.C., Flory, A.R., Jones, R.R., Patel, D., Nolan, B.T. and Ward, M.H., 2019. Modeling groundwater nitrate exposure in private wells of North Carolina for the Agricultural Health Study. Science of The Total Environment, 655, pp.512-519.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C. C., Lin, C. C. 2017. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. Resource Document. R package 1.6–7.

Mittal, K., Aggarwal, G. and Mahajan, P., 2018. Performance study of K-nearest neighbor classifier and K-means clustering for predicting the diagnostic accuracy. International Journal of Information Technology, pp.1-6. https://doi.org/10.1007/s41870-018-0233-x.

Møller, J.K., Nielsen, H.A. and Madsen, H., 2008. Time-adaptive quantile regression. Computational Statistics & Data Analysis, 52(3), 1292-1303.

Mountrakis, G., Im, J., Ogole, C. 2011. Support vector machines in remote sensing: A review. ISPRS Journal of Photogrammetry and Remote Sensing, 66(3), 247-259.

Muthusamy, M., Godiksen, P.N., Madsen, H., 2016. Comparison of different configurations of quantile regression in estimating predictive hydrological uncertainty. Procedia engineering, 154, 513-520.

Nampak, H., Pradhan, B., Manap, M. A. 2014. Application of GIS based data driven evidential belief function model to predict groundwater potential zonation. Journal of Hydrology, 513, 283-300.

Nemes, A., Rawls, W.J. and Pachepsky, Y.A., 2006. Use of the nonparametric nearest neighbor approach to estimate soil hydraulic properties. Soil Science Society of America Journal, 70(2), 327-336.

Neshat, A., Pradhan, B., Javadi, S. 2015. Risk assessment of groundwater pollution using Monte Carlo approach in an agricultural region: an example from Kerman Plain, Iran. Computers, Environment and Urban Systems, 50, 66-73.

Nkono, C., Liégeois, J. P., Demaiffe, D. 2018. Relationships between structural lineaments and Cenozoic volcanism, Tibesti swell, Saharan metacraton. Journal of African Earth Sciences.

Nolan, B. T., Fienen, M. N., Lorenz, D. L. 2015. A statistical learning framework for groundwater nitrate models of the Central Valley, California, USA. Journal of Hydrology, 531, 902-911.

Nolan, B. T., Gronberg, J. M., Faunt, C. C., Eberts, S. M., Belitz, K. 2014. Modeling nitrate at domestic and public-supply well depths in the Central Valley, California. Environmental science & technology, 48(10), 5643-5651.

Oh, H. J., Kim, Y. S., Choi, J. K., Park, E., Lee, S. 2011. GIS mapping of regional probabilistic groundwater potential in the area of Pohang City, Korea. Journal of Hydrology, 399(3-4), 158-172.

Ostad-Ali-Askari, K., Shayannejad, M., Ghorbanizadeh-Kharazi, H. 2017. Artificial neural network for modeling nitrate pollution of groundwater in marginal area of Zayandeh-rood River, Isfahan, Iran. KSCE Journal of Civil Engineering, 21(1), 134-140.

Panagopoulos, Y., Makropoulos, C., Baltas, E., Mimikou, M. 2011. SWAT parameterization for the identification of critical diffuse pollution source areas under data limitations. Ecological Modelling, 222(19), 3500-3512.

Peña-Haro, S., Pulido-Velazquez, M. and Llopis-Albert, C., 2011. Stochastic hydro-economic modeling for optimal management of agricultural groundwater nitrate pollution under hydraulic conductivity uncertainty. Environmental Modelling & Software, 26(8), 999-1008.

Peña-Haro, S., Pulido-Velazquez, M., Llopis-Albert, C. 2011. Stochastic hydro-economic modeling for optimal management of agricultural groundwater nitrate pollution under hydraulic conductivity uncertainty. Environmental Modelling & Software, 26(8), 999-1008.

Rahmati, O., Melesse, A. M. 2016. Application of Dempster–Shafer theory, spatial analysis and remote sensing for groundwater potentiality and nitrate pollution analysis in the semi-arid region of Khuzestan, Iran. Science of the Total Environment, 568, 1110-1123.

Ransom, K.M., Nolan, B.T., Traum, J.A., Faunt, C.C., Bell, A.M., Gronberg, J.A.M., Wheeler, D.C., Rosecrans, C.Z., Jurgens, B., Schwarz, G.E. and Belitz, K., 2017. A hybrid machine learning model to predict and visualize nitrate concentration throughout the Central Valley aquifer, California, USA. Science of the Total Environment, 601, pp.1160-1172.

Rodriguez-Galiano, V. F., Luque-Espinar, J. A., Chica-Olmo, M., Mendes, M. P. 2018. Feature selection approaches for predictive modelling of groundwater nitrate pollution: An evaluation of filters, embedded and wrapper methods. Science of the Total Environment, 624, 661-672.

Rodriguez-Galiano, V., Mendes, M. P., Garcia-Soldado, M. J., Chica-Olmo, M., Ribeiro, L. 2014. Predictive modeling of groundwater nitrate pollution using Random Forest and multisource variables related to intrinsic and specific vulnerability: A case study in an agricultural setting (Southern Spain). Science of the Total Environment, 476, 189-206.

Sahoo, P.K., Kim, K. and Powell, M.A., 2016. Managing groundwater nitrate contamination from livestock farms: implication for nitrate management guidelines. Current Pollution Reports, 2(3), pp.178-187.

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.

Salamon, P., Fernandez‐Garcia, D. and Gómez‐Hernández, J.J., 2007. Modeling tracer transport at the MADE site: the importance of heterogeneity. *Water resources research*, *43*(8).

Shiferaw, H., Bewket, W. and Eckert, S., 2019. Performances of machine learning algorithms for mapping fractional cover of an invasive plant species in a dryland ecosystem. Ecology and evolution, 9(5), 2562-2574.

Shrestha, A. and Luo, W., 2018. Assessment of Groundwater Nitrate Pollution Potential in Central Valley Aquifer Using Geodetector-Based Frequency Ratio (GFR) and Optimized-DRASTIC Methods. *ISPRS International Journal of Geo-Information, 7*(6), p.211.

Shrestha, D. L., Rodriguez, J., Price, R. K., Solomatine, D. P. 2006. Assessing model prediction limits using fuzzy clustering and machine learning. In Proc. 7th Int. Conf. On Hydroinformatics (pp. 4-8).

Shrestha, D. L., Solomatine, D. P. 2006. Machine learning approaches for estimation of prediction interval for the model output. Neural Networks, 19(2), 225-235.

Solomatine, D. P. Shrestha, D. L. 2009. A novel method to estimate model uncertainty using machine learning techniques. Water Resources Research, 45(12).

Soto-Pinto, C., Arellano-Baeza, A., Sánchez, G. 2013. A new code for automatic detection and analysis of the lineament patterns for geophysical and geological purposes (ADALGEO). Computers & geosciences, 57, 93-103.

Stelzer, R. S., Scott, J. T. 2018. Predicting Nitrate Retention at the Groundwater-Surface Water Interface in Sandplain Streams. Journal of Geophysical Research: Biogeosciences, 123(9), 2824-2838.

Strobl, C., Boulesteix, A.L., Zeileis, A. and Hothorn, T., 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics*, *8*(1), p.25.

Suthar, S., Bishnoi, P., Singh, S., Mutiyar, P. K., Nema, A. K., Patil, N. S. 2009. Nitrate contamination in groundwater of some rural areas of Rajasthan, India. Journal of hazardous materials, 171(1-3), 189-199.

Takizawa, S. (Ed.). 2008. Groundwater management in Asian cities: Technology and policy for sustainability (Vol. 2). Springer Science & Business Media.

Taylor, J.W., 2007. Forecasting daily supermarket sales using exponentially weighted quantile regression. European Journal of Operational Research, 178(1), pp.154-167.

Taylor, K. E. 2001. Summarizing multiple aspects of model performance in a single diagram. Journal of Geophysical Research: Atmospheres, 106(D7), 7183-7192.

Uusitalo, L., Lehikoinen, A., Helle, I., Myrberg, K. 2015. An overview of methods to evaluate uncertainty of deterministic models in decision support. Environmental Modelling & Software, 63, 24-31.

Vapnik, V., Mukherjee, S. 2000. Support vector method for multivariate density estimation. In Advances in neural information processing systems (pp. 659-665).

Vrettas, M.D. and Fung, I.Y., 2015. Toward a new parameterization of hydraulic conductivity in climate models: Simulation of rapid groundwater fluctuations in Northern California. Journal of Advances in Modeling Earth Systems, 7(4), 2105-2135.

Weerts, A.H., Winsemius, H.C. and Verkade, J.S., 2011. Estimation of predictive hydrological uncertainty using quantile regression: examples from the National Flood Forecasting System (England and Wales). Hydrology and Earth System Sciences, 15(1), 255-265.

Wheeler, D. C., Nolan, B. T., Flory, A. R., DellaValle, C. T., Ward, M. H. 2015. Modeling groundwater nitrate concentrations in private wells in Iowa. Science of the Total Environment, 536, 481-488.

World Health Organization (WHO) 2011. Guidelines for Drinking-water Quality. fourth ed. http://www.who.int/water_sanitation_health/publications/2011/dwq_guidelines/en/.

Zhang, L., Liu, Q., Yang, W., Wei, N. and Dong, D., 2013. An improved k-nearest neighbor model for short-term traffic flow prediction. Procedia-Social and Behavioral Sciences, 96, 653-662.

Zhou, H., Gómez-Hernández, J.J. and Li, L., 2014. Inverse methods in hydrogeology: Evolution and recent trends. Advances in Water Resources, 63, 22-37.

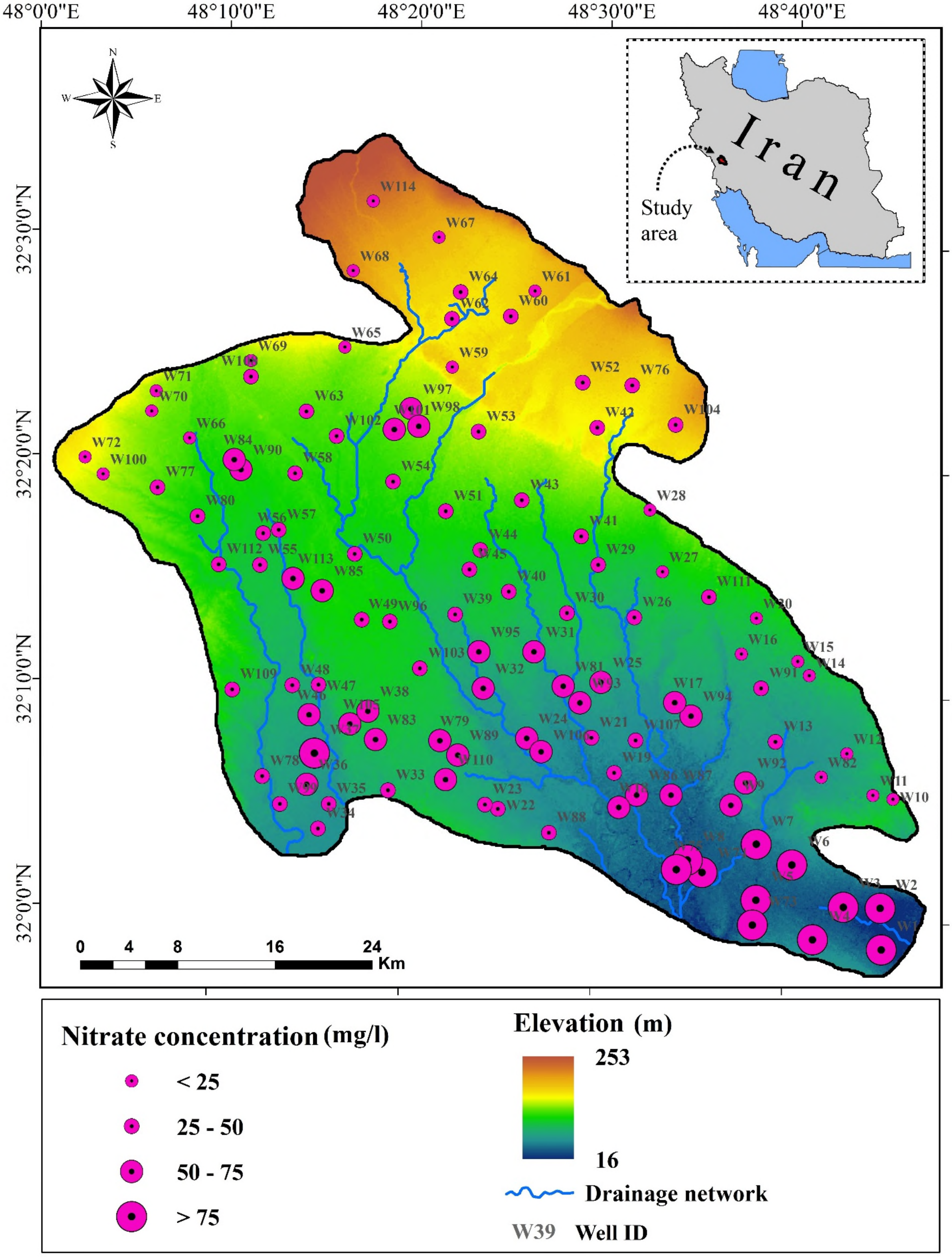
48°0'0"E
48°10'0"E
48°20'0"E
48°30'0"E
48°40'0"E
32°30'0"N
32°20'0"N
32°10'0"N
32°0'0"N
N
W
E
S
Iran
Study area
0
4
8
16
24
Km
W114
W67
W68
W64
W61
W62
W60
W65
W69
W108
W71
W70
W59
W97
W63
W101
W98
W102
W53
W52
W76
W42
W104
W66
W72
W84
W90
W100
W58
W77
W54
W43
W28
W80
W51
W56
W57
W41
W44
W112
W55
W113
W50
W45
W29
W27
W85
W40
W111
W49
W96
W39
W30
W26
W20
W95
W31
W16
W15
W14
W103
W32
W81
W25
W91
W109
W48
W47
W38
W17
W94
W46
W105
W83
W79
W24
W106
W21
W107
W13
W89
W12
W78
W36
W110
W19
W92
W82
W33
W86
W87
W9
W11
W35
W10
W23
W22
W34
W88
W7
W6
W3
W2
W4
W1
Nitrate concentration (mg/l)
< 25
25 - 50
50 - 75
> 75
Elevation (m)
253
16
Drainage network
W39
Well ID

Fig. 1 Map of the study area (Andimeshk-Dezful, Iran) showing the location of the 114 well sampling points (W01 – W114).

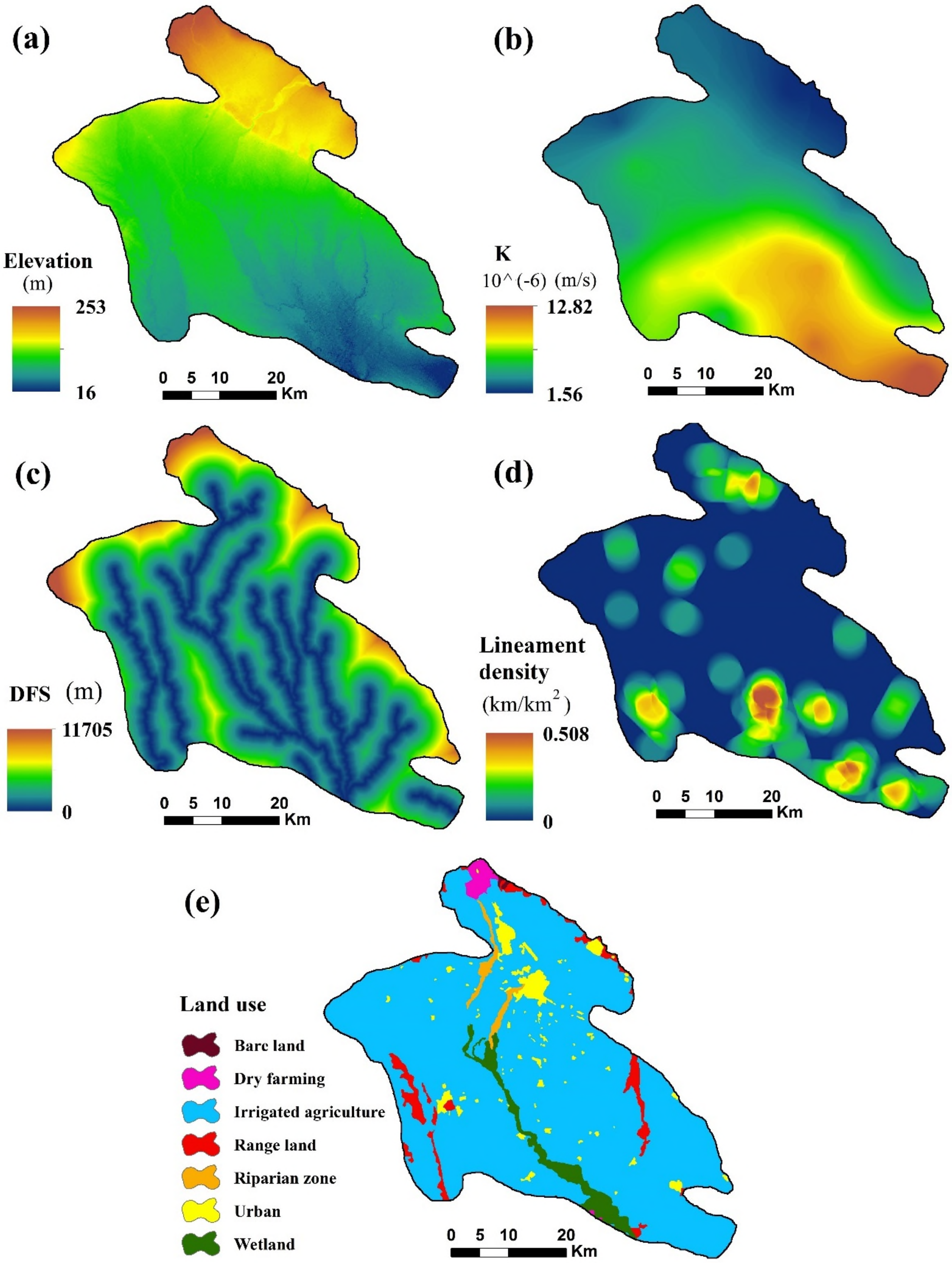

Fig. 2 The groundwater nitrate predictive variables: (a) elevation, (b) hydraulic conductivity (K), (c) distance from stream (DFS), (d) lineament density, and (e) land use.

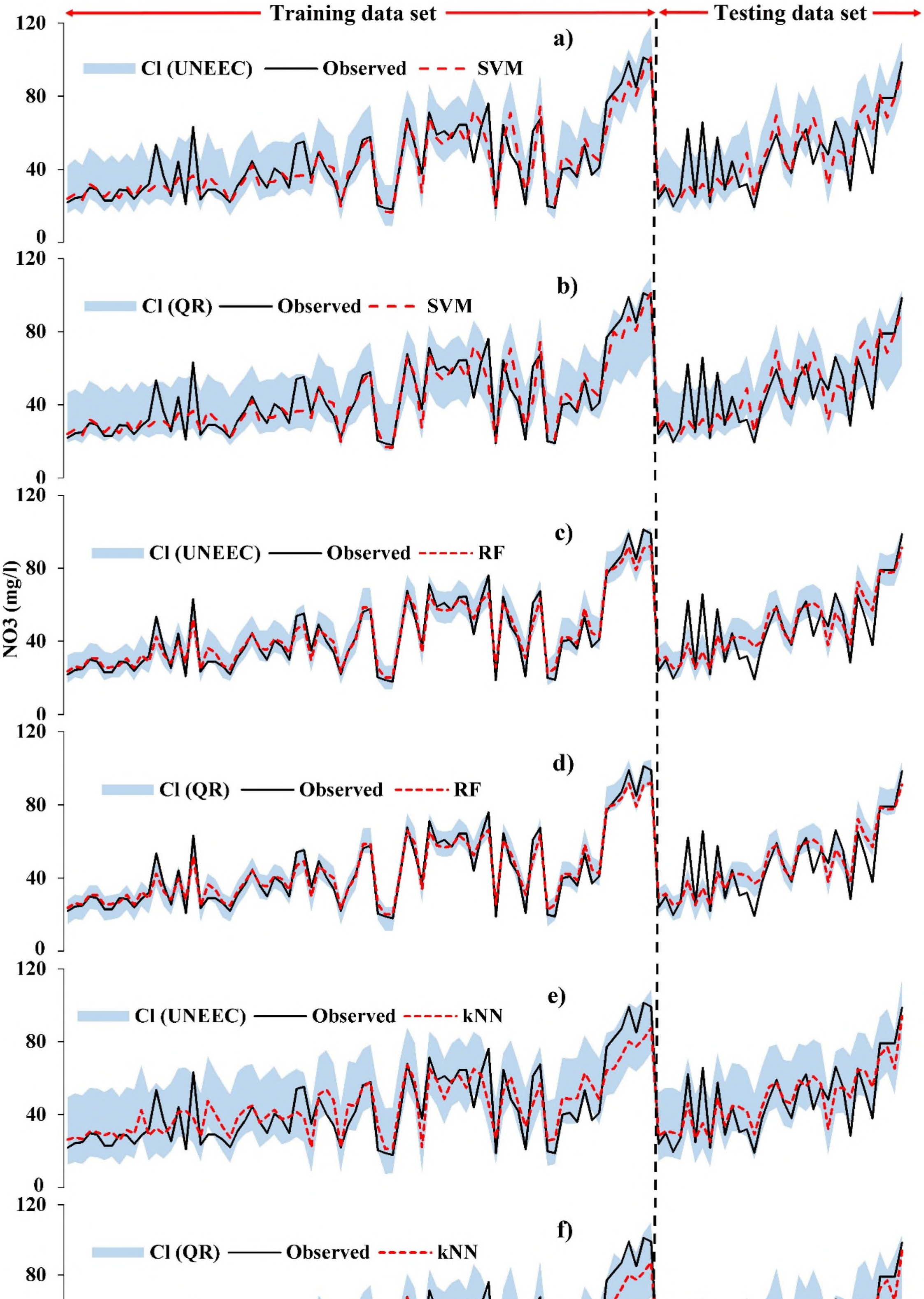
Training data set
Testing data set
a)
CI (UNEEC)
Observed
SVM
b)
CI (QR)
Observed
SVM
c)
CI (UNEEC)
Observed
RF
d)
CI (QR)
Observed
RF
e)
CI (UNEEC)
Observed
kNN
f)
CI (QR)
Observed
kNN
NO3 (mg/l)
0
40
80
120
1
11
21
31
41
51
61
71
81
91
101
111
Well number

Fig. 3 Observed versus modeled (SVM, RF, and *k*NN) Nitrate with 90 % confidential level (Cl) using UNEEC and QR methods. (Blue area is uncertainty band width)

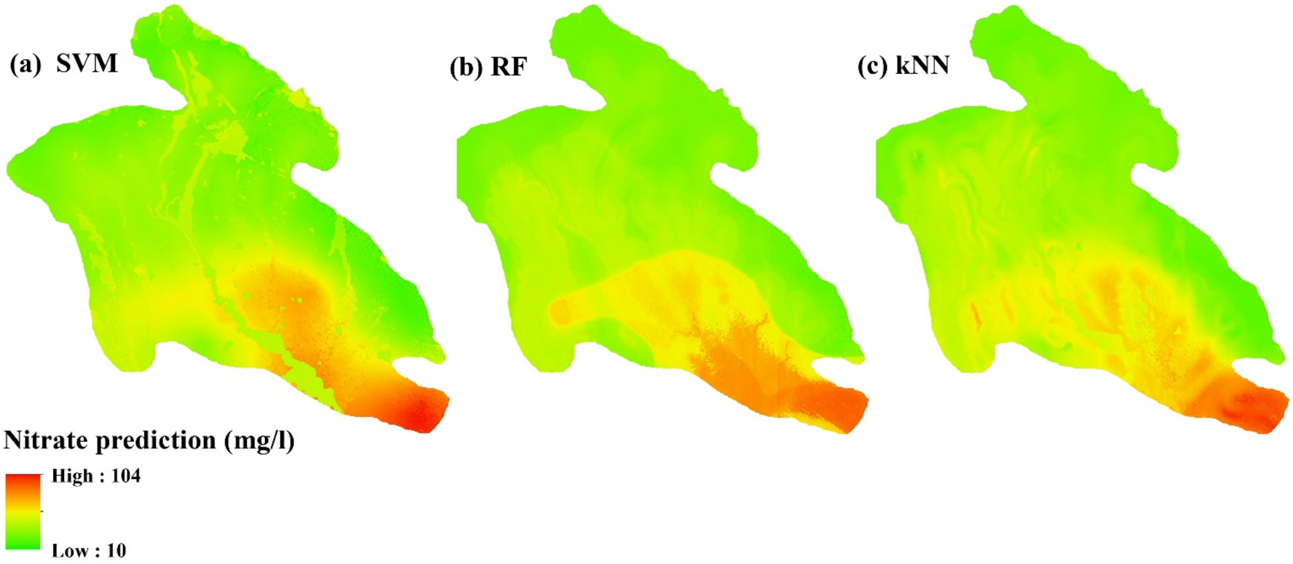

Fig. 4 Spatial prediction of nitrate concentrations (mg/l) using the SVM (a), RF (b) and *k*NN (c) models.

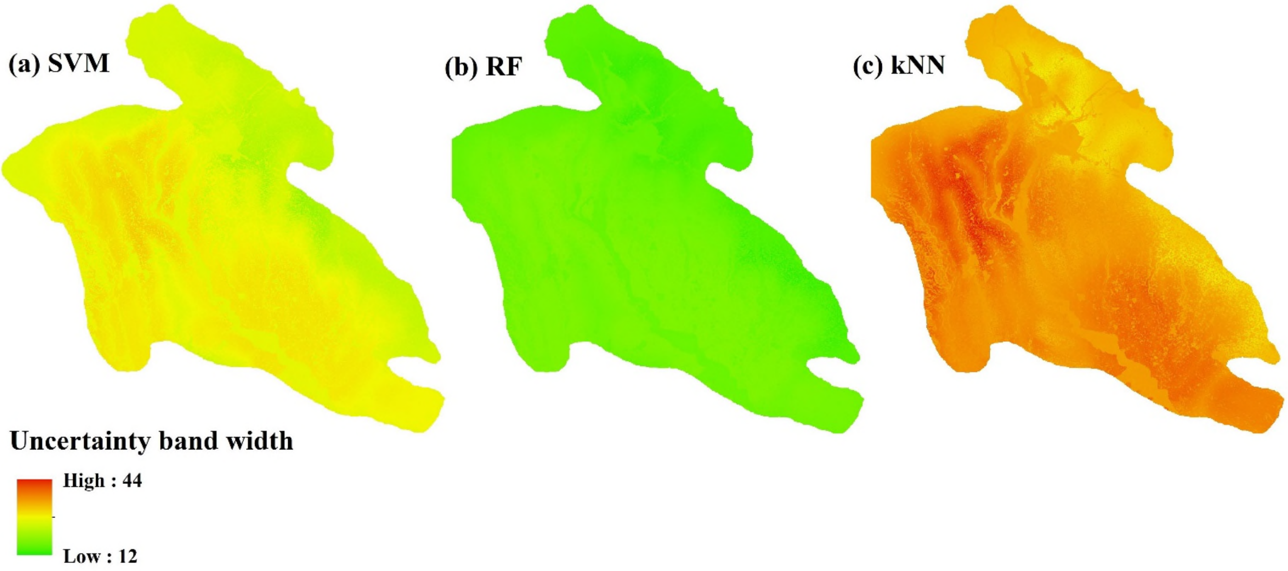

Fig. 5 Uncertainty band-width calculated by the UNEEC method.

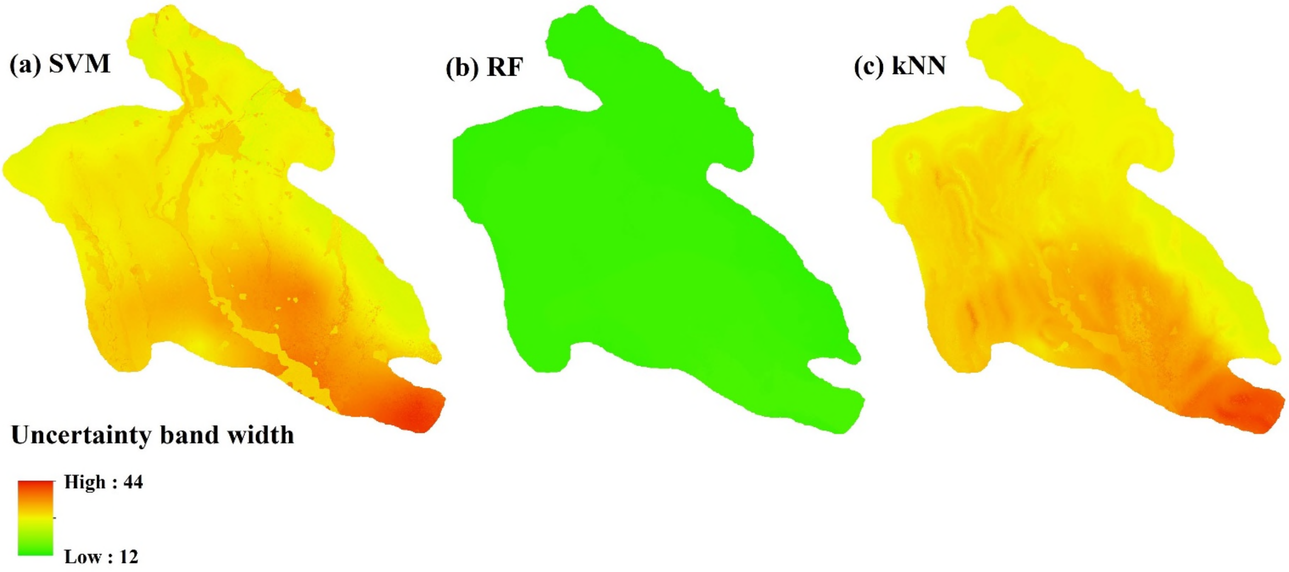

Fig. 6 Uncertainty band-width calculated by the QR method.

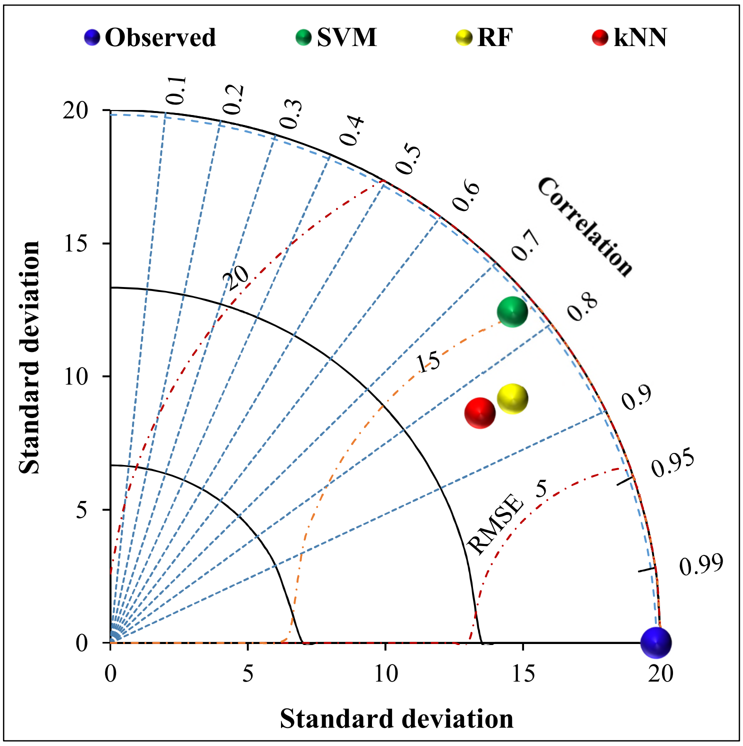

Fig. 7 Comparison of the models' performance using the Taylor diagram

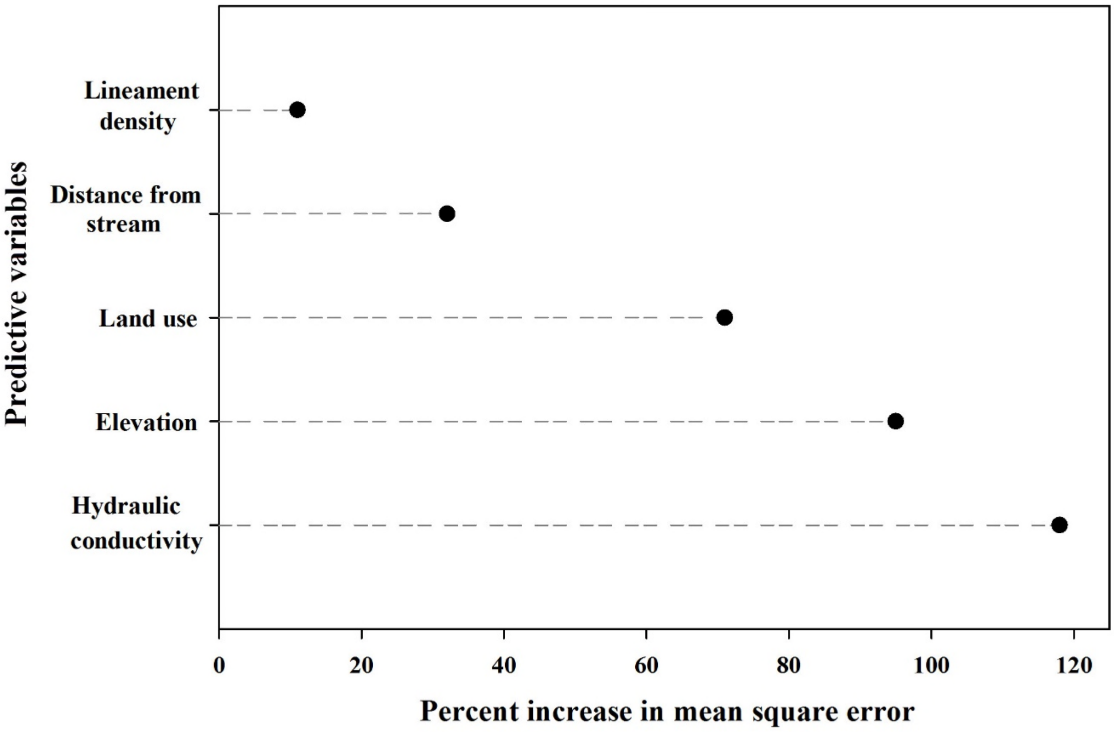

Fig. 8 Importance plot for predictive variables based on the RF model.

Table 1 Summary of the model performance

Model	Training		Testing	
	RMSE	R^2	RMSE	R^2
SVM	8.76	0.82	13.28	0.58
RF	4.69	0.96	10.41	0.72
*k*NN	10.85	0.74	10.63	0.71

Table 2 Uncertainty results using the UNEEC method

Model	Uncertainty statistic	Train				Test			
		Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 2	Cluster 3	Cluster 4	Cluster 5
SVM	MPI	30.23	27.63	28.32	28.62	30.17	27.92	28.36	29.74
	PICP	0.93	0.91	0.90	0.89	0.79	0.79	0.76	0.79
RF	MPI	17.84	16.71	16.93	16.41	17.83	16.82	17.10	16.66
	PICP	0.91	0.88	0.88	0.88	0.68	0.65	0.65	0.65
*k*NN	MPI	36.02	36.08	37.21	35.62	35.91	36.30	37.52	36.13
	PICP	0.91	0.93	0.94	0.93	0.85	0.88	0.88	0.91

SVM: Support vector machine; RF: Random forest; *k*NN: *k*-nearest neighbor; MPI: mean prediction interval; PICP: prediction interval coverage probability.

Table 3 Uncertainty results using the QR method

Model	Uncertainty statistic	Train	Test
SVM	MPI	30.77	31.76
	PICP	0.93	0.74
RF	MPI	13.52	13.59
	PICP	0.93	0.59
*k*NN	MPI	30.77	31.76
	PICP	0.93	0.74

SVM: Support vector machine; RF: Random forest; *k*NN: *k*-nearest neighbor; MPI: mean prediction interval; PICP: prediction interval coverage probability.