water

MDPI

Article

Imputation of Ammonium Nitrogen Concentration in Groundwater Based on a Machine Learning Method

Wanlu Li [1,2], Xueyan Ye [1,2] and Xinqiang Du [1,2,*]

[1] Key Laboratory of Groundwater Resources and Environment, Ministry of Education, Jilin University, Changchun 130021, China; liwl19@mails.jlu.edu.cn (W.L.); yexy@jlu.edu.cn (X.Y.)
[2] College of New Energy and Environment, Jilin University, Changchun 130021, China
* Correspondence: duxq@jlu.edu.cn

Abstract: Ammonium is one of the main inorganic pollutants in groundwater, mainly due to agricultural, industrial and domestic pollution. Excessive ammonium can cause human health risks and environmental consequences. Its temporal and spatial distribution is affected by factors such as meteorology, hydrology, hydrogeology and land use type. Thus, a groundwater ammonium analysis based on limited sampling points produces large uncertainties. In this study, organic matter content, groundwater depth, clay thickness, total nitrogen content (TN), cation exchange capacity (CEC), pH and land-use type were selected as potential contributing factors to establish a machine learning model for fitting the ammonium concentration. The Shapley Additive exPlanations (SHAP) method, which explains the machine learning model, was applied to identify the more significant influencing factors. Finally, the machine learning model established according to the more significant influencing factors was used to impute point data in the study area. From the results, the soil organic matter feature was found to have a substantial impact on the concentration of ammonium in the model, followed by soil pH, clay thickness and groundwater depth. The ammonium concentration generally decreased from northwest to southeast. The highest values were concentrated in the northwest and northeast. The lowest values were concentrated in the southeast, southwest and parts of the east and north. The spatial interpolation based on the machine learning imputation model established according to the influencing factors provides a reliable groundwater quality assessment and was not limited by the number and the geographical location of samplings.

Keywords: ammonium nitrogen; spatial interpolation; machine learning; random forest; SHAP

Citation: Li, W.; Ye, X.; Du, X. Imputation of Ammonium Nitrogen Concentration in Groundwater Based on a Machine Learning Method. *Water* **2022**, *14*, 1595. https://doi.org/10.3390/w14101595

Academic Editors: Qiang Fu, Yongqiang Cao, Tianxiao Li and Mo Li

Received: 8 April 2022
Accepted: 12 May 2022
Published: 16 May 2022

1. Introduction

Groundwater is a component of water supply [1]. It is distributed in various natural and geological environments and is affected by numerous factors. Natural groundwater recharge in Asia has uneven spatial and temporal distribution. Since the 1970s, an increasing water demand has resulted in severe groundwater overdraft, water-level decline and water quality degradation in China [2,3].

There is growing concern about the consequences of nitrogen pollutants in groundwater since they are harmful to human and environmental health [4], for example, leading to noncarcinogenic health risk for adults and children due to the use of groundwater as drinking water [5,6]. Excessive nitrogen can cause soil acidification, eutrophication and greenhouse gas [7–10]. Therefore, the evaluation of the current situation of nitrogen contamination is of great significance to avoid groundwater quality degradation and contributes to the utilization of groundwater resources.

Based on the analysis of on-site sampling data combined with statistical methods, the distribution information of groundwater quality can be understood. Filling sampling gaps is traditionally carried out using spatial interpolation methods. However, the performance of the interpolation method depends not only on the method itself, but also on the quality

of the data. Due to the sparse and uneven characteristics of geological sampling points, the results of normal estimation have great uncertainty. Many researchers have attempted to improve the methods. For example, a decision tree for selecting an appropriate method was developed based on data availability and the features of the method [11]. A geological modeling method based on the dynamic normal estimation of sparse point clouds was proposed to improve accuracy [12]. Some researchers also pointed out that the optimal sampling design and spatial predicting method are both important for predictive mapping [13]. However, the determination of key control points of spatial variability that may be affected by multiple factors has become increasingly challenging in practice [14]. The improvement of methods can help to solve problems in the sampling data, but it also ignores the value of data to a certain extent. Therefore, not only the problem of insufficient or uneven data needs to be improved, but we also need to make full use of the information hidden in the existing data.

In recent years, machine learning (ML) methods have been remarkably useful in the extraction of essential information from data in the natural sciences, where the major goal is to obtain novel scientific insights and discoveries from observational or simulated data, particularly in cases where there are not enough data to understand the physical process of the system and relatively accurate prediction is required [15]. For example, a logistic regression (LR) model was built with independent variables, and the binary occurrence probability of nitrate in groundwater was predicted [16]. Multiple linear regression tree (MLR), classification, regression tree (CART), random forest (RF) and boosted regression tree (BRT) models were built to predict the spatial distribution of nitrate in groundwater [17]. A hybrid nonlinear machine learning model, BRT, was developed to interpolate and visualize the nitrate contamination of groundwater in California's Central Valley [18]. These studies have been proven to be valuable to obtain the distribution of groundwater quality. However, less research has been directed towards the interpretation of the results from machine learning models; that is, it is not clear which factors play a key role in groundwater quality, which is not conducive to pollution prevention and control in key polluted areas. By balancing interpretability and accuracy [19], interpretable machine learning methods that provide explanation for black-box model outputs are available for water research [20,21].

The current study took a typical irrigation area in Northeast China as an example to establish a machine learning model of ammonium concentration using environmental factors. Then, the SHAP method, which can explain the fitting results of the machine learning model for ammonium concentration values, was used to identify the significant influencing factors through feature importance and a dependency analysis. According to the significant influencing factors, a machine learning model was established for ammonium data imputation (Figure 1). The flowchart is shown in Figure 1. The results of this method will provide reliable information with which to assess ammonium pollution in groundwater. The method is not limited by the number of samples or geographic location. In addition, the machine learning imputation method establishes a link between groundwater ammonium and the main influencing factors, which may be beneficial for pollution prevention at key sites.

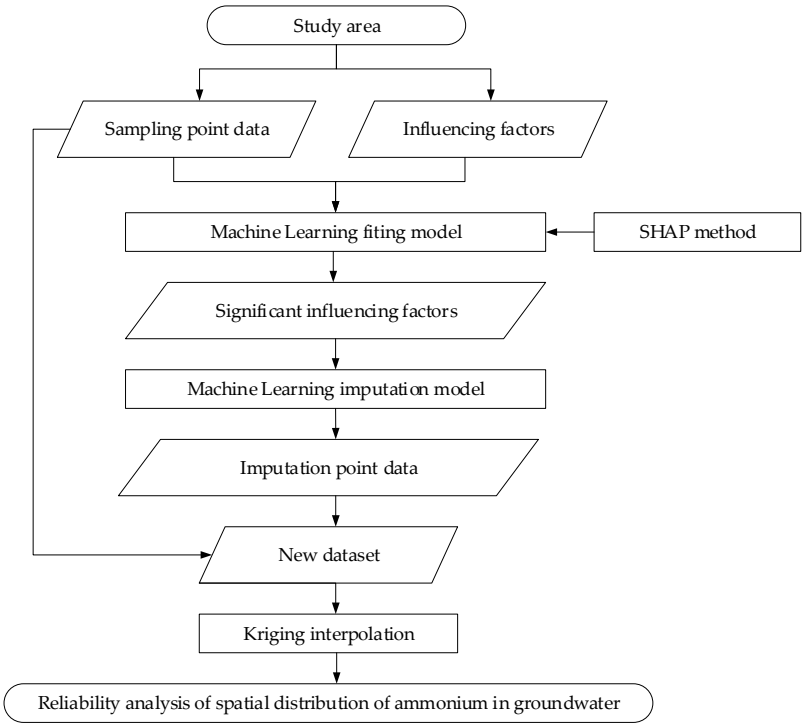

Figure 1. Flow chart.

2. Materials and Methods

2.1. Materials

2.1.1. Study Area

Located in the Sanjiang Plain, Heilongjiang Province (Figure 2), the Puyang irrigation district h a cold–temperate continental monsoon climate with an average annual temperature of −19.3 °C from January to 21.7 °C in July. The average evaporation is 694.4 mm, with a relative air humidity of 70–80%. The average annual precipitation is 535.5 mm, 70% of which occurs in June to September. The frost-free period is 116–154 days, and the seasonal frozen soil depth is 1.5–2.5 m.

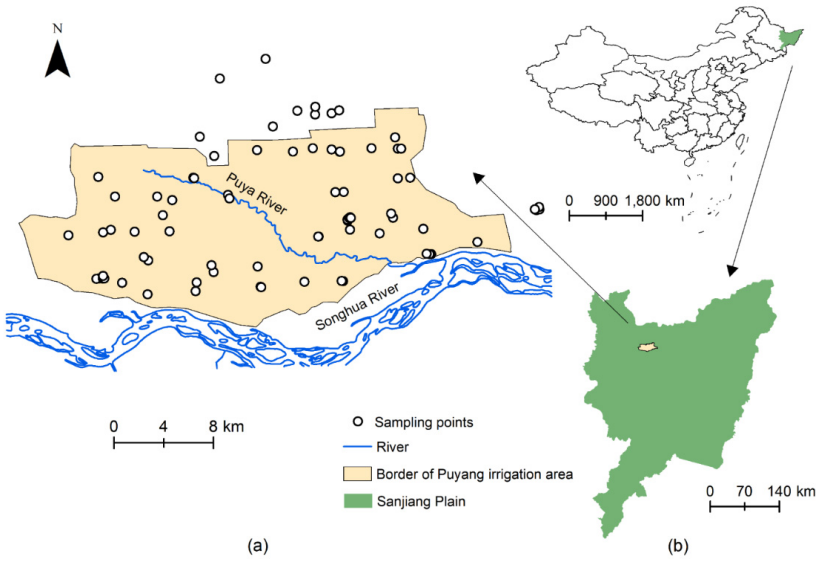

Figure 2. Distribution of sampling points (**a**) and location of the Puyang irrigation area (**b**).

The characteristic landform pattern consists of terraces and flood plains. Topographically, the study area is higher in the west and lower in the east, with an elevation of

64.0–70.0 m, and has a slope of 1/1000–1/3000 m. The Quaternary porous aquifer is composed of medium sand, medium-coarse sand and gravel, with a hydraulic conductivity of 9.6–16.1 m/d, and the buried depth is generally between 2 and 3 m. The unconfined aquifers and weakly confined aquifers mainly receive water from precipitation infiltration; lateral runoff replenishment; and discharge through lateral runoff, evaporation and exploitation.

2.1.2. Data

A total of 79 samples of shallow groundwater with groundwater wells mainly consisting of domestic wells and irrigation wells were sampled in August 2017 in the Puyang irrigation area (Figure 1). The inorganic nitrogen (NH_4^+, NO_3^- and NO_2^-) was analyzed using ion chromatography. This study focused on groundwater ammonium concentrations.

For the comprehensive consideration of groundwater ammonium-influencing factors and the existing relevant information in the study area, the following environmental factors for predicting groundwater ammonium concentration were selected: in soil samples, organic matter content, total nitrogen content (TN), cation exchange capacity (CEC) and pH. Land-use type information was obtained from the Resource and Environment Science and Data Center. In addition, groundwater depth and clay thickness were derived from historical data. Specific descriptions are presented in Table 1.

Table 1. Description of environmental factors related to the groundwater ammonium.

Factor	Description	Source	Point Number	Date	Resolution	Rmse	Place
Organic matter	Data from the special study on soil environmental quality	Kriging interpolation	457	October 2018	690 m	16.61	Songhua River-Naoli River Basin in Sanjiang Plain and surrounding areas
TN		Kriging interpolation	457	October 2018	690 m	0.70	
CEC		Kriging interpolation	457	October 2018	690 m	7.27	
pH		Kriging interpolation	457	October 2018	690 m	0.49	
Groundwater depth	Groundwater sampling point data	Kriging interpolation	275	August 2017	690 m	3.94	
Clay thickness	Historical data	Kriging interpolation	1614		690 m	2.64	
Land use	Data on the relevant website	Resource and Environment Science and Data Center		2018	1000 m		

Ammonium concentrations were selected as the label (the predicted item), while the remaining items, including organic matter content, groundwater depth, clay thickness, TN, CEC, pH and land-use type were selected as features that had an influence on the label.

2.2. Methods

2.2.1. Random Forest Regression Model

According to information from scikit-learn, which takes sample size and different problems into account to select the right estimator, the random forest regression model was applied to predict the continuous digital output for the ammonium concentration. The random forest regression model is an ensemble algorithm based on decision tree theory, which divides the data multiple times according to some cut-off values in the features, and creates many subsets to distinguish different samples [22,23]. The core purpose of decision tree algorithms is to select the best feature for a branch, such as ID3, C4.5 and CART [24]. Random forest consists of independent trees built by randomly selected features, averaging the results of each tree to determine the final output. It has an improved prediction accuracy compared with a single decision tree. Some additional benefits are that random forests are good at solving nonlinear problems, require no normalization or scaling of data and are insensitive to multicollinearity [25–29].

2.2.2. Model Interpretation

In recent years, Shapley Additive exPlanations (SHAP) demonstrated superior performance in uncovering the underlying phenomenon [30–32]. In this study, the Python SHAP library was used to understand the importance of the variables for predictions in the random forest models.

Proposed by Lundberg and Lee [33], the method based on the optimal Shapley values of alliance game theory is a common feature attribution mechanism that can explain artificial intelligence. The Shapley value of the feature is the weighted summation of the feature contribution to the output (prediction) of all possible feature combinations (Equation (1)):

$$\phi_j(val) = \sum_{S \subseteq \{x_1, x_2, \ldots, x_p\} \{x_j\}} \frac{|S|!(p - |S| - 1)!}{P!} \left(val\left(S \cup \{x_j\}\right) - val(S)\right) \quad (1)$$

where ϕ_j is the Shapley value of feature j, x is the feature of the instance, S is a subset of features and p is the number of features.

SHAP explains the output of an instance by computing the contribution of the feature to the prediction. It is represented as Equation (2):

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z'_j, \quad (2)$$

where g is the explanatory model, ϕ_j is the Shapley value for a feature j, M is the number of input features and $z' \in \{0, 1\}^M$ equals 1 when a feature is observed.

Tree SHAP, used for this study, is a variant of SHAP [34]. The influencing factors were identified using feature importance and a feature dependence analysis. Features with large absolute Shapley values are important. Interaction effects are captured using a feature dependence plot based on the SHAP interaction value [35]. The interaction value between feature i and feature j is defined as in Equation (3):

$$\phi_{i,j}(f, x) = \sum_{S \subseteq M \{I,J\}} \frac{|S|!(M - |S| - 2)!}{2(M - 1)!} \nabla_{i,j}(f, x, S), \quad (3)$$

where M represents all features, S is a subset of the features and x is the feature vector of the instance.

2.2.3. Kriging Interpolation Method

According to the instructions of Surfer, Cokriging uses a more densely sampled correlated secondary variable to help guide the estimation of the primary variable. Thus, the Kriging interpolation method was used to obtain the spatial distribution of the original groundwater ammonium concentration because the method is useful for interpolation with almost any type of data set. In order to qualitatively analyze the rationality of the machine learning imputation result, the kriging method was also applied to the ammonium concentration distribution mapping after data imputation.

Based on the theory of regionalized variables and semivariogram (Equation (4)), Kriging is a geostatistical analysis method that considers the size of the sample values, the spatial location and the distance between samples [36]. According to the original data and the structural characteristics of the variogram, the value of the unknown sampling point can be estimated.

$$\gamma(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [Z(x_i) - Z(x_i + h)]^2, \quad (4)$$

where $\gamma(h)$ is the semivariogram, N(h) is the logarithm of points at distance h, h is the distance between samples and $Z(x_i)$ is the value of the sample x_i.

3. Results and Discussion

3.1. Model Performance Evaluation

Parameter tuning needs to be conducted to maximize model performance. In this study, the mean square error (MSE) was selected to evaluate the model performance, which measures the difference between the prediction values and their corresponding actual

values. A good performance should result in low MSE values of close to 0. The index was calculated using Equation (5):

$$\mathrm{MSE} = \frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{y_i} - y_\mathrm{i})^2, \quad (5)$$

where n is the number of samples, i is each sample, $\mathrm{y_i}$ is the actual value of the sample point i and y_i is the value predicted by the model.

To analyze the influencing factors with a high performance machine learning model, random_state, the parameter related to the randomness of this algorithm, was adjusted with other parameters, which remained at default because the parameters associated with the structure of the decision tree (the underlying algorithm of random forest) are designed to overfit the data [37]; that is, these parameters ensure that the algorithm fits the data as closely as possible at their default values. Figure 3 shows that the minimum MSE is 0.017 when random_state is equal to 271, which indicates a good fit to the data. Figure 4 displays the predictions for the data. In summary, the model has good performance and can be used to influence a factor analysis.

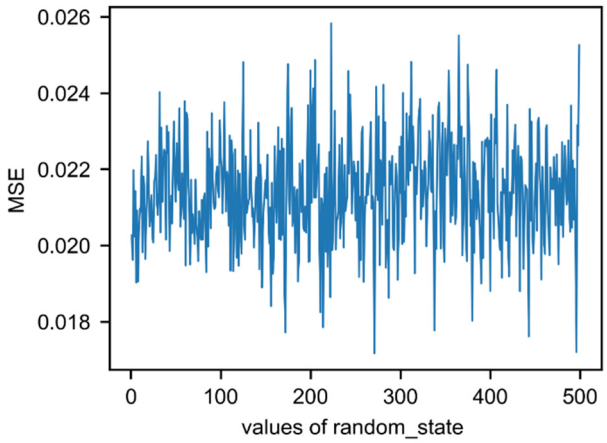

Figure 3. Curve of model performance with different parameter values.

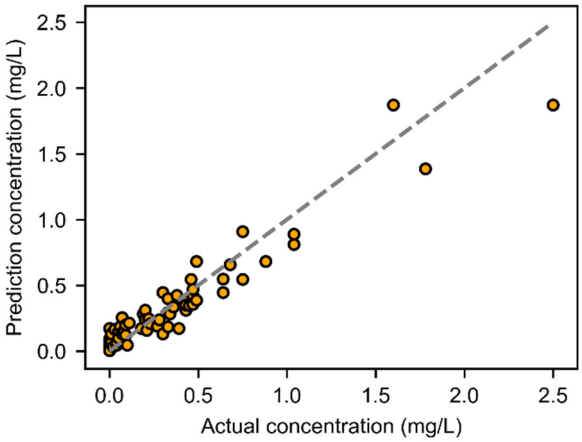

Figure 4. Scatter plot of ammonium concentration.

For the ammonium concentration prediction model, the max_depth, n_estimators and min_impurity_decrease parameters were selected for tuning using the grid search technique and 10-fold cross-validation (the data were randomly divided into 10 sub-samples, of which 9 sub-samples were used to train the model each time, and 1 sub-sample was used to test the model) to optimize model performance, with the aim to fit the original data as much as possible whilst avoiding fitting of the noise and random fluctuations,

resulting in reducing the generalization ability of the model for unseen data. It is noted that max_depth and n_estimators are usually the most effective parameters for random forest model performance [29,37]. Furthermore, min_impurity_decrease, stops a split if the level of reduction from the split is less than the entered amount. This parameter was tuned in the training process because it allows for more concise control over the tree structure. The optimal parameter combination is based on the MSE results of the test set during the grid search. In addition, according to previous related research, 10-fold cross-validation was used in this study [29,38].

As can be seen from the results, the selected optimal parameters values were as follows: n_estimators: 60; max_depth: 7; min_impurity_decrease: 0. The training and test score with an MSE of 0.02 and 0.09, respectively, suggest that the model has a low error level and is generalizable. Based on this result, the model seemed to perform well, and the predictive performance was good in the comprehensive evaluation.

Table 2 shows the information of the machine learning models. Considering the number of data and the machine learning method of this study, the trained machine learning model displayed high performance and can be used for further analysis. In order to reduce the errors and improve the performance, advanced machine learning algorithms should be selected on the basis of increasing the amount of data.

Table 2. Information of the machine learning models for ammonium concentration.

Model	Parameter	Parameter Value	MSE
Fitting model	random_state	271	0.017
Prediction model	max_depth	60	Training data MSE = 0.02 Test data MSE = 0.09
	n_estimators	7	
	min_impurity_decrease	0	

3.2. Analysis the Influencing Factors of Ammonium Concentration in Groundwater

3.2.1. Feature Importance

In the summary plot (Figure 5), the features are ordered according to their importance in the machine learning model output, and the colors represent the value of each feature. For the seven original features, the top four were selected in order of importance for analysis and further imputation-model construction in this study.

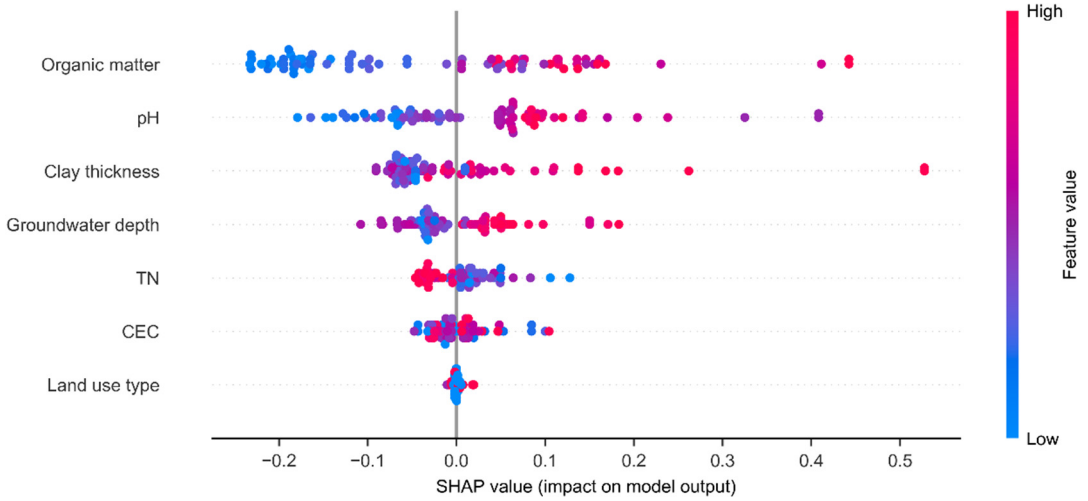

Figure 5. SHAP summary plot of the ammonium nitrogen concentration fitting model.

The organic matter content, clay thickness, groundwater depth and pH were regarded as the main influencing factors of this study according to the order of feature importance and the variance inflation factor (VIF) for multicollinearity detection (Table 3). Specifically, organic matter had the greatest impact on the model, while pH, clay thickness and

groundwater depth were the three next most important features (Figure 5). The higher values of these factors result in higher SHAP values, which means a higher probability that ammonium pollution has occurred. However, the observations of groundwater depth were theoretically unexpected, and it was speculated that the comprehensive result of the interactions between other factors were in accordance with feature dependency in SHAP theory [34], which is discussed further in the feature-dependency analysis section.

Table 3. VIF of the influencing factors.

Influencing Factor	Organic Matter	pH	Clay Thickness	Groundwater Depth
VIF	3.19	2.07	3.30	2.01

3.2.2. Spatial Distribution of the Influencing Factors

In order to perform an intuitive analysis of the influencing factors identified in the previous section, continuous maps for the spatial distribution were generated using Kriging methods. The spatial distribution of ammonium concentration is shown in Figure 6. The influencing factor results of Kriging interpolation based on the same 79 sampling points are shown in Figure 7. Table 4 shows the MSE results of the interpolation of the influencing factors.

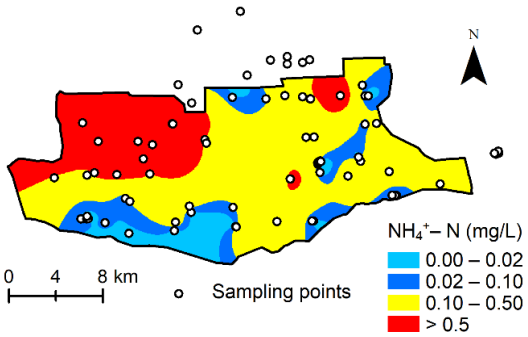

Figure 6. Spatial distribution of ammonium concentration in groundwater using Kriging.

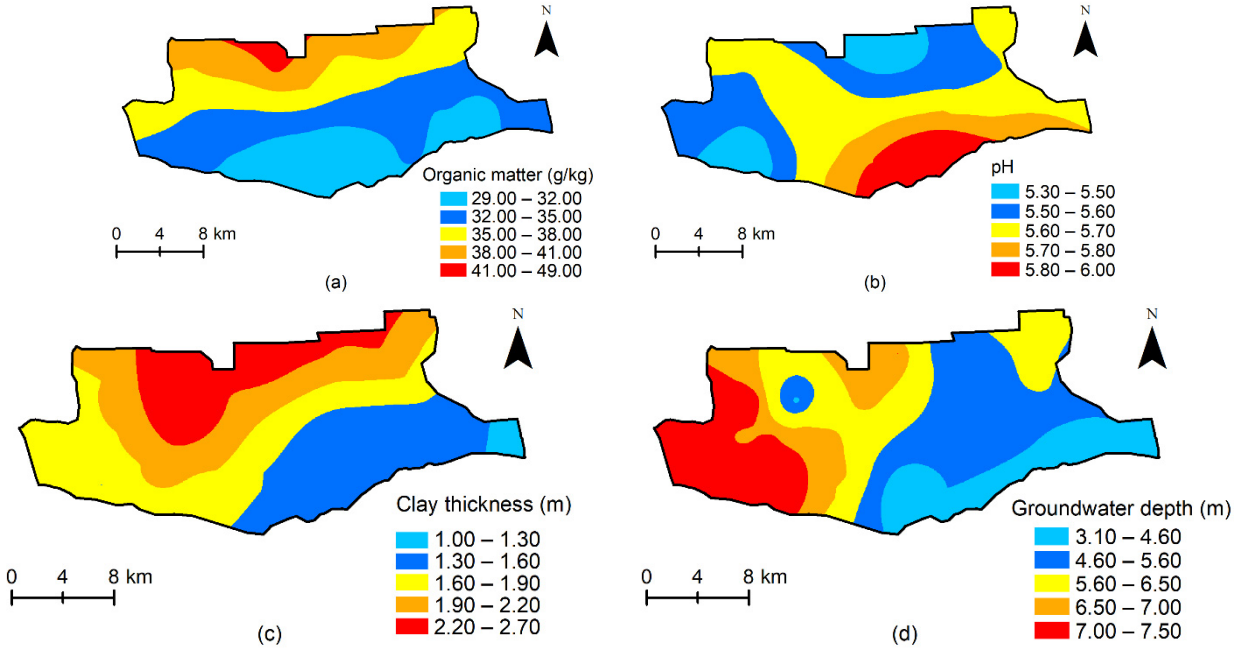

Figure 7. Spatial distribution of the influencing factors of organic matter (**a**), pH (**b**) clay thickness (**c**) and groundwater depth (**d**) using Kriging.

Table 4. MSE of Kriging interpolation result.

Organic Matter	pH	Clay Thickness	Groundwater Depth
1247.91	31.93	3.08	31.61

Figure 6 shows that the concentration of ammonium in the study area presented obvious spatial distribution characteristics, with values generally decreasing from northwest to southeast. The high-value areas were concentrated in the northwest and northeast districts, followed by the central, southeast, northwest and north of the overall area, while the organic matter content decreased from north to south, with the largest content in the north (Figure 7a). Together, Figures 6 and 7a show that the concentrations in the highest and lowest value areas of ammonium nitrogen were similar to the organic matter content, indicating that ammonium concentrations were related to the organic matter content. However, the overall distributions of the two were not exactly consistent. For example, the low organic matter content in the southeast correlated with a high ammonium concentration.

For pH, the tendency of pH to decrease from the south to north was obviously inconsistent with the ammonium concentrations (Figures 6 and 7b). It is noted that the ammonium concentrations were lower and the pH values were higher in the southeast edge of the study area, while in the northwest, the ammonium concentrations were higher and the pH values were lower.

With regard to the influencing factor of clay, the thickness in the north and southwest was greater than in other areas, and the thickness decreased from northwest to southeast. Except for the thicker clay and higher concentration in the northwest and a few northeastern districts, there was no significant distribution pattern in other locations (Figures 6 and 7c).

In addition, for the groundwater depth, the variation from deep to shallow can be summarized as follows: west > north > southeast (Figure 7d); ammonium nitrogen: northwest and northeast > central, south and southeast > southwest and some areas in the north and southeast. As seen in Figures 6 and 7d, the groundwater depths were lower in the southeast with higher ammonium nitrogen concentrations between 0.10 and 0.50 mg/L, while the depths were higher in the southwest with lower ammonium nitrogen concentrations. As for the north and central regions, the above-mentioned relations did not exist between the groundwater depth and ammonium nitrogen concentration.

Therefore, it can be concluded that the distributions of these four influencing factors were similar to the ammonium concentration in some areas, but this was not applicable to the entire irrigation area. It can be inferred that the ammonium concentration distribution was a comprehensive result of the interaction between the influencing factors. To further analyze the relationship between the ammonium concentration and distribution of the influencing factors in the overall irrigation area, the interactions were analyzed using SHAP feature dependency (Figure 8).

3.2.3. Feature Dependency

In accordance with the influencing factors of ammonium concentration in 79 groundwater samples in August 2017 in the irrigation area, the SHAP feature dependency plots were automatically color coded according to the strongest interaction for the influencing factors. Red dots represent higher values, and the blue dots represent lower values.

As shown in Figure 8a, a high organic matter content and TN correspond to larger SHAP values, which suggests that an increasing organic matter content and TN increase the risk of ammonium pollution.

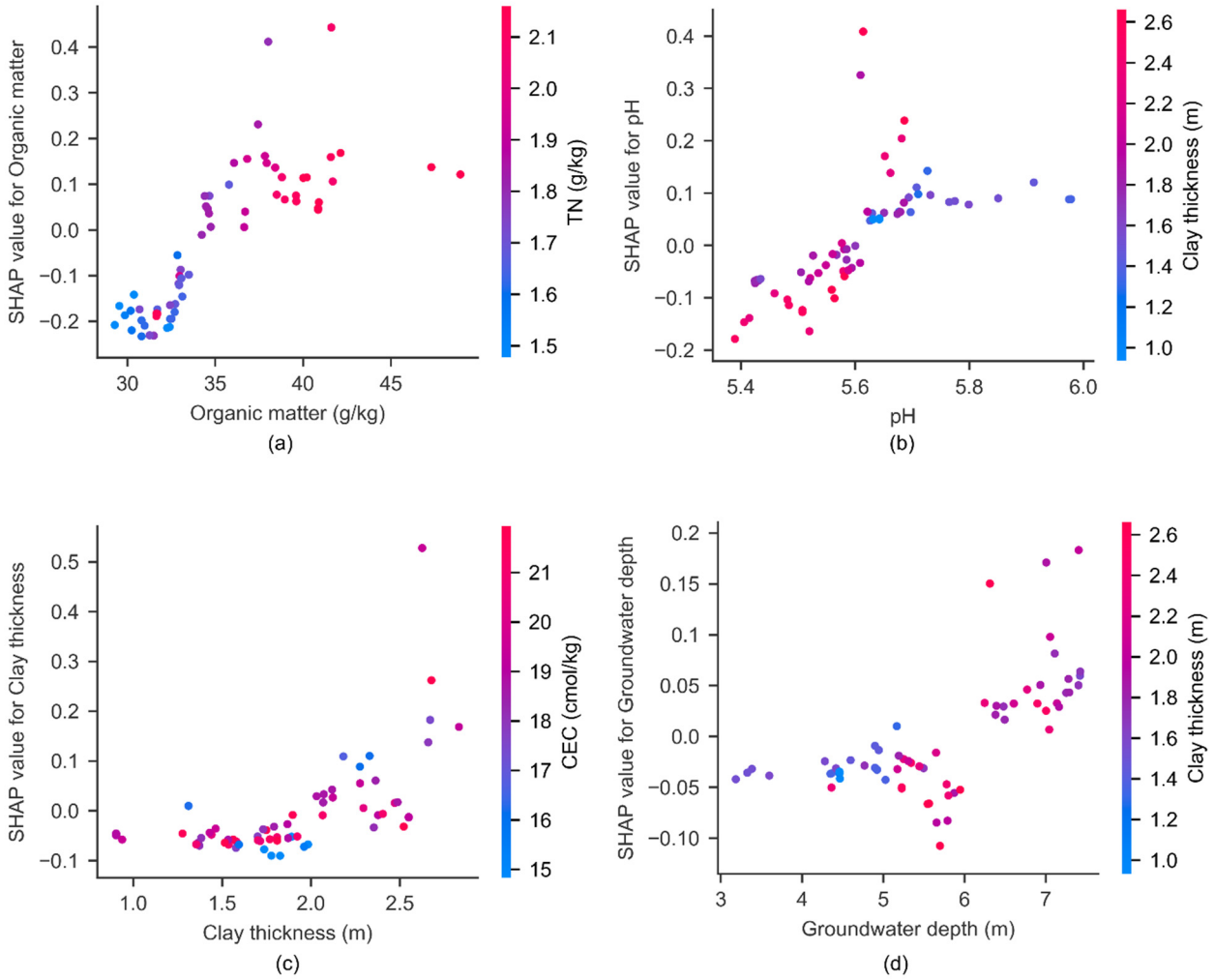

Figure 8. SHAP feature dependence plot of the influencing factors of organic matter (**a**), pH (**b**), clay thickness (**c**) and groundwater depth (**d**) with their strongest interaction.

Subsequently, the SHAP values tended towards 0.0 as the pH increased with a pH of less than approximately 5.6 (Figure 8b), which suggests that, upon increasing pH, the probability of ammonium pollution risk decreases. Conversely, when pH was higher than 5.6, the SHAP values were approximately 0.1 (Figure 8b), which means that increasing pH while clay thickness is low will not have a significant effect on the changes in ammonium nitrogen concentration.

Figure 8c shows the impact of clay thickness and CEC on NH_4^+ concentration. Increasing clay thickness caused higher ammonium concentration occurrence because the SHAP value increased significantly. Furthermore, the red points on the bottom of the figure, representing higher values of CEC, indicate that the ammonium concentration tended to be lower with a low clay thickness and high CEC.

Finally, the groundwater depth was selected to determine its impact on NH_4^+ concentration, and has the strongest interaction with clay thickness (Figure 8d). SHAP values were negative for points with a groundwater depth of below 6 m. The change in trend of the points indicated that increasing groundwater depth decreases the probability of a high ammonium concentration. In contrast, SHAP values were positive for points with a groundwater depth of above 6 m. It is speculated that at groundwater depths of greater than 6 m, ammonium concentrations may have been more susceptible to higher clay thicknesses in this study.

Some studies have shown that the mineralization of organic nitrogen from nitrogen fertilizers or natural soil is one of the potential sources of NH_4^+ [39], and there is a positive

correlation between NH_4^+ and DOC [5]. It has been demonstrated that the NH_4^+ absorption rate is high with a low nitrification rate in the range of a large DOC input in groundwater [40]. Therefore, organic matter is conducive to the enrichment of NH_4^+, and the process of mineralization is likely to affect the concentration of NH_4^+ in the irrigation area.

In general, pH affects groundwater concentrations and nitrogen by influencing the process of nitrification and denitrification [41]. Previous research shows that ammonium is abundant in groundwater when pH < 9.23 [42], and the decline in pH value leads to H+ and NH_4^+, forming a competitive relationship [43].

In terms of clay, the adsorption effect and the provision of a reducing environment to maintain the stability of ammonium nitrogen leads to thicker clay which contributes to ammonium accumulation [44,45]. As the main physical and chemical property that affects the removal of ammonium [46], an increase in CEC will reduce the NH_4^+ concentration mainly because the accumulation of nitrifying bacteria by particles helps to nitrify and degrade ammonium nitrogen. Moreover, high CEC is beneficial to the adsorption of ammonium in soil, leading to a decrease in the concentration of ammonium in groundwater.

For groundwater depth, generally, as the groundwater depth decreases, its effect on the concentration is enhanced, since a shallower depth reduces the vertical migration distance of nitrogen, resulting in reduced attenuation [47]. Meanwhile, ammonium nitrogen is more likely to dissolve through the interaction of shallow groundwater and soil [48].

3.3. Imputation of Ammonium Concentration in Groundwater

Uniform distribution was used as the main principle by considering the distribution of point choice. Using ArcGIS, 39 points were generated in the central district of the study area where sampling points were sparse (Figure 9). After the information on the influencing factors of these points was extracted, it was input to the model to predict the concentration of ammonium. Together with the original 79 sampling points, the dataset comprised 118 points. Figure 9 shows the distribution of ammonium concentration predicted by data imputation through machine learning, and it can be seen that the ammonium concentration generally decreased from northwest to southeast. The high-value areas were concentrated in the northwest and northeast. The low-value areas were concentrated in the southeast, southwest and parts of the east and north.

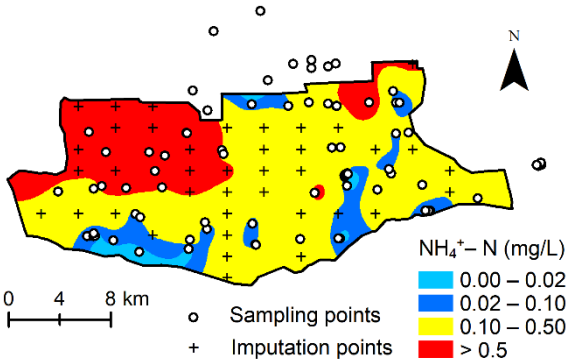

Figure 9. Spatial distribution of ammonium concentration in groundwater with data imputation using machine learning model.

3.4. Reliability Analysis of the Results of Points Imputation Using Machine Learning Model

This section presents the Kriging interpolation results when using original points and adding point data using the machine learning method. An analysis is presented for the level of performance of the machine learning imputation method.

Through the qualitative evaluation of the spatial interpolation methods using visual assessment, it was found that the two methods both provided satisfactory results for the overall decreasing trend of the NH_4^+ concentration from the northwest to the southeast (Figures 6 and 9), indicating that they both provide a good fit for most sampling points.

However, there were clear differences in that the machine learning method yielded higher concentration values in the south, southwest, northeast and parts of northwest, and lower concentration values in the east, northwest and central districts (Figure 10). In particular, the machine learning method identified northeast and central districts where the ammonium concentrations were greater than 0.5 mg/L, and also identified concentrations of less than 0.1 mg/L in the southeast (Figure 9). Further qualitative evaluation was performed on typical sites where ammonium concentrations increased and decreased significantly, which comprised eight points (Figure 10). The observations in Figure 9 are theoretically more reasonable than those in Figure 6 in that the points with higher ammonium concentrations (points 1, 2 and 3) generally corresponded to a lower pH, higher organic matter content and higher clay thickness (Table 5). Additionally, the points with lower ammonium concentrations (points 4–8) generally corresponded to a higher pH, a lower organic matter content and lower clay thickness (Table 5).

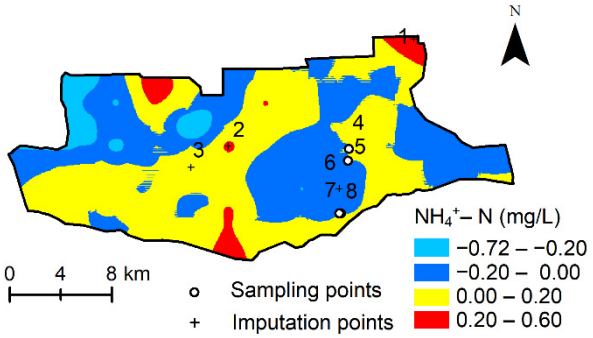

Figure 10. Difference in ammonium concentration between the machine learning imputation method and Kriging method.

Table 5. Values of influencing factors for typical points.

Point Number	Groundwater Depth (m)	Clay Thickness (m)	Organic Matter (g/kg)	pH
1	5.71	2.41	36.00	5.70
2	4.79	2.11	36.86	5.63
3	7.04	2.26	32.92	5.72
4	4.92	1.37	32.95	5.63
5	5.03	1.37	33.13	5.70
6	4.40	1.37	31.46	5.85
7	4.35	1.35	32.27	5.97
8	4.36	1.35	32.41	5.98

According to the quantitative evaluation of the results of the two methods, the interpolation seemed to perform better after the machine learning data imputation (machine learning MSE = 0.21; Kriging MSE = 0.27). Previous studies have shown that the Co-Kriging method, which introduces correlation into spatial modeling, is superior to the traditional Kriging method in terms of the reliability and accuracy of results for groundwater nitrogen distribution analysis [14,49]. It was speculated that higher accuracy would be obtained after machine learning data imputation if the sophisticated kriging method was applied to the groundwater ammonium interpolation.

In summary, the machine learning imputation method, which was based on the influencing factors from SHAP interpretation, yielded more reasonable results, thereby improving the accuracy of the Kriging method, and was not limited by the number of samples or geographic location. An additional benefit is that the machine learning imputation method can establish a link between groundwater ammonium nitrogen and the main influencing factors, which may be beneficial for pollution prevention at key sites.

4. Conclusions

The random forest ensemble model was established to model ammonium concentration using a set of relevant factors, whereas the random forest model was established based on the influencing factors for ammonium concentration data imputation. Spatial interpolation was performed to finally obtain the concentration distribution. Furthermore, the use of a machine learning model based on influencing factors for the spatial prediction of ammonium concentration values was compared to the Kriging interpolation result. The primary findings can be summarized as follows.

The organic matter feature was found to have a substantial impact on the concentration of ammonium, followed by the pH, clay thickness and groundwater depth. The ammonium concentration generally decreased from northwest to southeast. The highest values were concentrated in the northwest and northeast. The lowest values were concentrated in the southeast, southwest and parts of the east and north.

The concentration distribution analysis derived through the qualitative visual inspections showed that the results from the machine learning imputation method are more reasonable. The interpolation result seemed to perform better after machine learning data imputation. An additional advantage of machine learning imputation is that, given original sampling points, missing data can be infilled when there is little spatial connection between monitoring sites. The results suggest the good applicability of the model for ammonium concentration mapping in the study area. In summary, the machine learning imputation method yielded more reasonable results, improving the accuracy of the results, and was not limited by the number of samples or geographic location. In addition, the method can establish a link between groundwater ammonium and the main influencing factors, which may be beneficial for pollution prevention at key sites.

The limiting factors of this study include the relatively small study area selected with slight changes in environmental conditions, resulting in only minor differences between the two methods. This also led to insignificant machine learning imputation results. Additionally, the model accuracy needs to be improved by increasing the data size and applying advanced machine learning methods. In addition, the various soil data used in this study were derived from soil quality research and may not be applicable to other areas to more accurately detect ammonium content in groundwater.

In light of the fact that the variation law of groundwater quality is not easy to obtain, and that it would be desirable to analyze it on the basis of multi-source data, it is not unexpected that the robust behavior of prediction supports machine learning imputation techniques in the evaluation of groundwater quality. We recommend that the model performance be optimized by expanding the study area, and that the method be widely used with the ancillary support of multi-source big data, especially in groundwater quality assessments in industry.

Author Contributions: The research was conducted primarily by W.L., X.Y. and X.D.; the draft manuscript was primarily written by W.L. and X.D., and it was revised mainly by X.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Natural Science Foundation of China (41972247).

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Motlagh, A.M.; Yang, Z.J.; Saba, H. Groundwater quality. *Water Environ. Res.* **2020**, *92*, 1649–1658. [CrossRef]
2. Shen, D.J. Groundwater management in China. *Water Policy* **2015**, *17*, 61–82. [CrossRef]
3. Bierkens, M.F.P.; Wada, Y. Non-renewable groundwater use and groundwater depletion: A review. *Environ. Res. Lett.* **2019**, *14*, 063002. [CrossRef]
4. Zarama-Alvarado, S. The Challenges of Dealing with Nitrogen Pollutants in Groundwater. *Rev. Cient.* **2018**, *3*, 230–242. [CrossRef]

5. Norrman, J.; Sparrenbom, C.J.; Berg, M.; Nhan, D.D.; Jacks, G.; Harms-Ringdahl, P.; Nhan, P.Q.; Rosqvist, H. Tracing sources of ammonium in reducing groundwater in a well field in Hanoi (Vietnam) by means of stable nitrogen isotope (delta N-15) values. *Appl. Geochem.* **2015**, *61*, 248–258. [CrossRef]
6. Su, X.; Wang, H.; Zhang, Y. Health Risk Assessment of Nitrate Contamination in Groundwater: A Case Study of an Agricultural Area in Northeast China. *Water Resour. Manag.* **2013**, *27*, 3025–3034. [CrossRef]
7. Bacchus, S.T.; Barile, P.J. Discriminating sources and flowpaths of anthropogenic nitrogen discharges to Florida springs, streams and lakes. *Environ. Eng. Geosci.* **2005**, *11*, 347–369. [CrossRef]
8. Scherger, L.E.; Zanello, V.; Lexow, C. Impact of Urea and Ammoniacal Nitrogen Wastewaters on Soil: Field Study in a Fertilizer Industry (Bahia Blanca, Argentina). *Bull. Environ. Contam. Toxicol.* **2021**, *107*, 565–573. [CrossRef]
9. Lee, M.S.; Lee, K.K.; Hyun, Y.J.; Clement, T.P.; Hamilton, D. Nitrogen transformation and transport modeling in groundwater aquifers. *Ecol. Model.* **2006**, *192*, 143–159. [CrossRef]
10. Shi, W.M.; Yao, J.; Yan, F. Vegetable cultivation under greenhouse conditions leads to rapid accumulation of nutrients, acidification and salinity of soils and groundwater contamination in South-Eastern China. *Nutr. Cycl. Agroecosyst.* **2009**, *83*, 73–84. [CrossRef]
11. Li, J.; Heap, A.D. Spatial interpolation methods applied in the environmental sciences: A review. *Environ. Model. Softw.* **2014**, *53*, 173–189. [CrossRef]
12. Shi, T.D.; Zhong, D.Y.; Wang, L.G. Geological Modeling Method Based on the Normal Dynamic Estimation of Sparse Point Clouds. *Mathematics* **2021**, *9*, 1819. [CrossRef]
13. Zhao, Y.C.; Xu, X.H.; Tian, K.; Huang, B.A.; Hai, N. Comparison of sampling schemes for the spatial prediction of soil organic matter in a typical black soil region in China. *Environ. Earth Sci.* **2016**, *75*, 4. [CrossRef]
14. Du, Y.; Deng, Y.; Ma, T.; Shen, S.; Lu, Z.; Gan, Y. Spatial Variability of Nitrate and Ammonium in Pleistocene Aquifer of Central Yangtze River Basin. *Groundwater* **2020**, *58*, 110–118. [CrossRef]
15. Wang, M.X.; Liu, G.D.; Wu, W.L.; Bao, Y.H.; Liu, W.N. Prediction of agriculture derived groundwater nitrate distribution in North China Plain with GIS-based BPNN. *Environ. Geol.* **2006**, *50*, 637–644. [CrossRef]
16. Liu, C.W.; Wang, Y.B.; Jang, C.S. Probability-based nitrate contamination map of groundwater in Kinmen. *Environ. Monit. Assess.* **2013**, *185*, 10147–10156. [CrossRef]
17. Knoll, L.; Breuer, L.; Bach, M. Large scale prediction of groundwater nitrate concentrations from spatial data using machine learning. *Sci. Total Environ.* **2019**, *668*, 1317–1327. [CrossRef]
18. 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.; et al. A hybrid machine learning model to predict and visualize nitrate concentration throughout the Central Valley aquifer, California, USA. *Sci. Total Environ.* **2017**, *601*, 1160–1172. [CrossRef]
19. Mi, J.X.; Li, A.D.; Zhou, L.F. Review Study of Interpretation Methods for Future Interpretable Machine Learning. *IEEE Access* **2020**, *8*, 191969–191985. [CrossRef]
20. Althoff, D.; Bazame, H.C.; Nascimento, J.G. Untangling hybrid hydrological models with explainable artificial intelligence. *H_2Open J.* **2021**, *4*, 13–28. [CrossRef]
21. Thrun, M.C.; Ultsch, A.; Breuer, L. Explainable AI Framework for Multivariate Hydrochemical Time Series. *Mach. Learn. Knowl. Extr.* **2021**, *3*, 170–204. [CrossRef]
22. Loh, W.-Y. Classification and regression trees. *Wiley Interdiscip. Rev.-Data Min. Knowl. Discov.* **2011**, *1*, 14–23. [CrossRef]
23. Talekar, B.; Agrawal, S. A Detailed Review on Decision Tree and Random Forest. *Biosci. Biotechnol. Res. Commun.* **2020**, *13*, 245–248. [CrossRef]
24. Ture, M.; Tokatli, F.; Kurt, I. Using Kaplan-Meier analysis together with decision tree methods (C&RT, CHAID, QUEST, C4.5 and ID3) in determining recurrence-free survival of breast cancer patients. *Expert Syst. Appl.* **2009**, *36*, 2017–2026. [CrossRef]
25. Alves, L.G.A.; Ribeiro, H.V.; Rodrigues, F.A. Crime prediction through urban metrics and statistical learning. *Phys. A-Stat. Mech. Appl.* **2018**, *505*, 435–443. [CrossRef]
26. Wang, Z.L.; Lai, C.G.; Chen, X.H.; Yang, B.; Zhao, S.W.; Bai, X.Y. Flood hazard risk assessment model based on random forest. *J. Hydrol.* **2015**, *527*, 1130–1141. [CrossRef]
27. Tang, Z.P.; Mei, Z.; Liu, W.D.; Xia, Y. Identification of the key factors affecting Chinese carbon intensity and their historical trends using random forest algorithm. *J. Geogr. Sci.* **2020**, *30*, 743–756. [CrossRef]
28. le Maire, G.; Marsden, C.; Nouvellon, Y.; Grinand, C.; Hakamada, R.; Stape, J.L.; Laclau, J.P. MODIS NDVI time-series allow the monitoring of Eucalyptus plantation biomass. *Remote Sens. Environ.* **2011**, *115*, 2613–2625. [CrossRef]
29. Shin, K. Quantitative Precipitation Estimates Using Machine Learning Approaches with Operational Dual-Polarization Radar Data. *Remote Sens.* **2021**, *13*, 694. [CrossRef]
30. Politikos, D.V.; Petasis, G.; Katselis, G. Interpretable machine learning to forecast hypoxia in a lagoon. *Ecol. Inform.* **2021**, *66*, 101480. [CrossRef]
31. Wang, R.Z.; Kim, J.H.; Li, M.H. Predicting stream water quality under different urban development pattern scenarios with an interpretable machine learning approach. *Sci. Total Environ.* **2021**, *761*, 144057. [CrossRef]
32. Parsa, A.B.; Movahedi, A.; Taghipour, H.; Derrible, S.; Mohammadian, A. Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis. *Accid. Anal. Prev.* **2020**, *136*, 105405. [CrossRef]
33. Lundberg, S.M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 4–9 December 2017.

34. Lundberg, S.M.; Erion, G.; Chen, H.; Degrave, A.; Prutkin, J.M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.-I. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2020**, *2*, 56–67. [CrossRef]
35. Fujimoto, K.; Kojadinovic, I.K.; Marichal, J.L. Axiomatic characterizations of probabilistic and cardinal-probabilistic interaction indices. *Games Econ. Behav.* **2006**, *55*, 72–99. [CrossRef]
36. Negreiros, J.; Painho, M.; Aguilar, F.; Aguilar, M. Geographical Information Systems Principles of Ordinary Kriging Interpolator. *J. Appl. Sci.* **2010**, *10*, 852–867. [CrossRef]
37. Leiv, R.G.; Fernandez Anta, A.; Mancus, V.; Casari, P. A Novel Hyperparameter-Free Approach to Decision Tree Construction That Avoids Overfitting by Design. *IEEE Access* **2019**, *7*, 99978–99987. [CrossRef]
38. Bui, D.T.; Khosravi, K.; Karimi, M.; Busico, G.; Khozani, Z.S.; Nguyen, H.; Mastrocicco, M.; Tedesco, D.; Cuoco, E.; Kazakis, N. Enhancing nitrate and strontium concentration prediction in groundwater by using new data mining algorithm. *Sci. Total Environ.* **2020**, *715*, 136836. [CrossRef]
39. Shen, S.; Ma, T.; Du, Y.; Luo, K.W.; Deng, Y.M.; Lu, Z.J. Temporal variations in groundwater nitrogen under intensive groundwater/surface-water interaction. *Hydrogeol. J.* **2019**, *27*, 1753–1766. [CrossRef]
40. Lupon, A.; Denfeld, B.A.; Laudon, H.; Leach, J.; Sponseller, R.A. Discrete groundwater inflows influence patterns of nitrogen uptake in a boreal headwater stream. *Freshw. Sci.* **2020**, *39*, 228–240. [CrossRef]
41. Wang, L.S.; He, Z.B.; Li, J. Assessing the land use type and environment factors affecting groundwater nitrogen in an arid oasis in northwestern China. *Environ. Sci. Pollut. Res.* **2020**, *27*, 40061–40074. [CrossRef]
42. Zhao, S.; Zhang, B.J.; Zhou, N.Q. Effects of Redox Potential on the Environmental Behavior of Nitrogen in Riparian Zones of West Dongting Lake Wetlands, China. *Wetlands* **2020**, *40*, 1307–1316. [CrossRef]
43. Li, D.; Zhou, Y.; Long, Q.; Li, R.; Lu, C. Ammonia nitrogen adsorption by different aquifer media: An experimental trial for nitrogen removal from groundwater. *Hum. Ecol. Risk Assess.* **2020**, *26*, 2434–2446. [CrossRef]
44. Wang, C.; Wu, D.; Mao, X.; Hou, J.; Wang, L.; Han, Y. Estimating soil ammonium adsorption using pedotransfer functions in an irrigation district of the North China Plain. *Pedosphere* **2021**, *31*, 157–171. [CrossRef]
45. Wang, S.; Tang, C.; Song, X.; Yuan, R.; Wang, Q.; Zhang, Y. Using major ions and delta δ^{15}N-NO_3^- to identify nitrate sources and fate in an alluvial aquifer of the Baiyangdian lake watershed, North China Plain. *Environ. Sci.-Processes Impacts* **2013**, *15*, 1430–1443. [CrossRef]
46. Dong, Y.B.; Lin, H. Ammonia nitrogen removal from aqueous solution using zeolite modified by microwave-sodium acetate. *J. Cent. South Univ.* **2016**, *23*, 1345–1352. [CrossRef]
47. Almasri, M.N.; Kaluarachchi, J.J. Assessment and management of long-term nitrate pollution of ground water in agriculture-dominated watersheds. *J. Hydrol.* **2004**, *295*, 225–245. [CrossRef]
48. Rudzianskaite, A.; Sukys, P. Effects of groundwater level fluctuation on its chemical composition in karst soils of Lithuania. *Environ. Geol.* **2008**, *56*, 289–297. [CrossRef]
49. Huang, J.; Xu, J.; Liu, X.; Liu, J.; Ramsankaran, R.; Wang, L.; Su, W. Geospatial Based Assessment of Spatial Variation of Groundwater Nitrate Nitrogen in Shandong Intensive Farming Regions of China. *Sens. Lett.* **2012**, *10*, 491–500. [CrossRef]