



Article

# Predicting Heavy Metal Concentrations in Shallow Aquifer Systems Based on Low-Cost Physiochemical Parameters Using Machine Learning Techniques

Thi-Minh-Trang Huynh <sup>1</sup>, Chuen-Fa Ni <sup>1,2,\*</sup> , Yu-Sheng Su <sup>3,\*</sup> , Vo-Chau-Ngan Nguyen <sup>4</sup> , I-Hsien Lee <sup>1,2</sup> ,  
Chi-Ping Lin <sup>1,2</sup> and Hoang-Hiep Nguyen <sup>1</sup>

<sup>1</sup> Graduate Institute of Applied Geology, National Central University, Taoyuan 32001, Taiwan

<sup>2</sup> Center for Environmental Studies, National Central University, Taoyuan 32001, Taiwan

<sup>3</sup> Department of Computer Science and Engineering, National Taiwan Ocean University, Keelung 202301, Taiwan

<sup>4</sup> College of Environment and Natural Resources, Can Tho University, Can Tho 94000, Vietnam

\* Correspondence: nichuenfa@geo.ncu.edu.tw (C.-F.N.); ntoucsiesu@mail.ntou.edu.tw (Y.-S.S.)

**Abstract:** Monitoring ex-situ water parameters, namely heavy metals, needs time and laboratory work for water sampling and analytical processes, which can retard the response to ongoing pollution events. Previous studies have successfully applied fast modeling techniques such as artificial intelligence algorithms to predict heavy metals. However, neither low-cost feature predictability nor explainability assessments have been considered in the modeling process. This study proposes a reliable and explainable framework to find an effective model and feature set to predict heavy metals in groundwater. The integrated assessment framework has four steps: model selection uncertainty, feature selection uncertainty, predictive uncertainty, and model interpretability. The results show that Random Forest is the most suitable model, and quick-measure parameters can be used as predictors for arsenic (As), iron (Fe), and manganese (Mn). Although the model performance is auspicious, it likely produces significant uncertainties. The findings also demonstrate that arsenic is related to nutrients and spatial distribution, while Fe and Mn are affected by spatial distribution and salinity. Some limitations and suggestions are also discussed to improve the prediction accuracy and interpretability.

**Keywords:** Random Forest; heavy metals; groundwater quality; explainable artificial intelligence (XAI); prediction intervals



**Citation:** Huynh, T.-M.-T.; Ni, C.-F.; Su, Y.-S.; Nguyen, V.-C.-N.; Lee, I.-H.; Lin, C.-P.; Nguyen, H.-H. Predicting Heavy Metal Concentrations in Shallow Aquifer Systems Based on Low-Cost Physiochemical Parameters Using Machine Learning Techniques. *Int. J. Environ. Res. Public Health* **2022**, *19*, 12180. <https://doi.org/10.3390/ijerph191912180>

Academic Editors: Huiping Zeng, Yan Han and Yanan Cai

Received: 6 August 2022

Accepted: 20 September 2022

Published: 26 September 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Water quality sampling is essential to reflect the environmental status and governance efficiency. However, spare monitoring data and tedious analysis processes also inhibit the interpretation of ongoing pollution events, leading to broader contaminant transportation and treatment costs [1]. Heavy metals are not only among the toxic substances causing a high risk to ecological and human health but also need much time and complicated procedures to detect and remediate. Thus, developing a model to predict heavy metals will optimize the data collection process, transmission, and lab-based analyses [2,3], resulting in a faster response to contamination. Current technology advancements enable real-time water quality monitoring, but continuous monitoring at dense spatial scales will be a cost challenge. Most sensor technologies are expensive and need regular care but may measure a single or a few parameters [4]. Therefore, utilizing data that can be measured from one in situ multiparametric probe to predict heavy metals would be very applicable for setting up an effective monitoring system.

The practical approach to evaluating a model's feasibility relies on the performance of the model, including the data accessibility and the model's robustness, accuracy, and

uncertainty. The feasible model could support the stakeholders in deciding whether the target task is successful and applicable. Many studies have successfully applied cost-effective parameters and machine learning algorithms for predicting water quality indexes [5,6] or heavy metals [7,8]. However, the previous investigations usually focused on accessing model predictability for one data set without uncertainty evaluation. Evaluating a trained model based on historical data might be too optimistic because future data behavior could change with different degrees of uncertainty induced by natural or anthropogenic activities [9]. Therefore, comparing single values of evaluation metrics, such as goodness-of-fit and errors, could not reveal much about the prediction perturbation [10]. Modelers need more evidence beyond accuracy to convince the general public or managers to trust predicted values [11]. Thus, efforts need to be devoted to exploring how the changes in data input may lead to a different prediction, whether the cost-effective features can yield good predictions, and which factors drive the outcomes.

Furthermore, from a quick overview, studies that express interpretability in heavy metal prediction models have not yet been developed and assessed. The wisely used analyses, namely the correlation analysis, clustering analysis, and principal component analysis, have been applied to explore the causal relationship between groundwater heavy metals and other chemical compositions [12–17]. Those outcomes could explain the contamination sources or geochemistry conditions but could not provide information on how a predictive model could internally make decisions. The interpretable or transparent models with clear links between input and output are often insufficient to express highly non-linear relationships [18]. Although the black-box or opaque models are more accurate and reliable, operators and managers have found it difficult to trust and implement them [19]. Even transparent models occasionally require explanations for non-expert audiences [20]. A model with explainability will aid in expressing their results as convincing evidence for their real-life implementation, improvement, and transmission.

The study aimed to develop a scanning framework to test the feasibility and interpretability of a generalized machine learning model for heavy metal prediction. With the recent development of sensor technologies, low-cost multiparameter probes are widely used for measuring fundamental groundwater quality. However, determining various heavy metals remain challenging because of the cost-demand of field samplings and the associated laboratory work. The observations of multiparameter probes could facilitate real-time predictions of heavy metals. In the study, this modeling experiment focused on the feasibility of available physicochemical parameters by comparing a model's performance on data perturbation, feature changes, and prediction intervals. The proposed model can be reliable and accessible for practical applications if the low-cost parameters outcompete other predictors. Furthermore, the model interpretability is the supplementary information to enhance a user's understanding of the rationale behind the prediction or how much each predictor contributes to the estimated value. We expect the systematic assessment to provide general insight into the modeling issues of accuracy, reliability, and explainability to facilitate the trained models in groundwater quality monitoring and management.

## 2. Materials and Methods

### 2.1. Uncertainty Quantification

Several approaches have been proposed to quantify uncertainty, including the Monte Carlo methods [21–24], perturbation methods [25–28], or Bayesian methods [29–31]. In particular, most approaches can only measure specific uncertainty sources. The Monte Carlo methods focus on parameter uncertainty, Bayesian methods deal with input data uncertainty, and perturbation methods optimize model architecture robustness. These approaches require the known parameter distributions, a constraint in sparse sampling datasets. Nonetheless, bootstrapping does not need either assumptions about the data distribution or complex computations [32]. Bootstrapping is compatible with assessing the uncertainty of unknown or complicated data distribution and the insufficient sample size of a large-scale model. In practice, the bootstrap method has been applied in model

robustness evaluation by comparing the distribution of model performances [10,33] and prediction interval estimation [34–37].

In this computational experiment, we considered uncertainty in various sample distributions. The bootstrap scheme was set up 100 times, randomly splitting 75% of data for training and 25% for testing using different random states. One hundred times replicated sampling with replacement from the data pool will give 100 prediction results. Consequently, repeated in-samples (training data) were used for training models, whereas we only used out-samples (testing data) for uncertainty evaluation. The distribution of 100 evaluation metrics in the box and whisker plots can demonstrate the error range that each model produces. Similarly, the model's performance distribution on feature selection uncertainty was assessed on different feature groups of 100 bootstrap datasets. Finally, the models or predictors with the lowest bias range and highest goodness-of-fit were selected for the subsequent analysis.

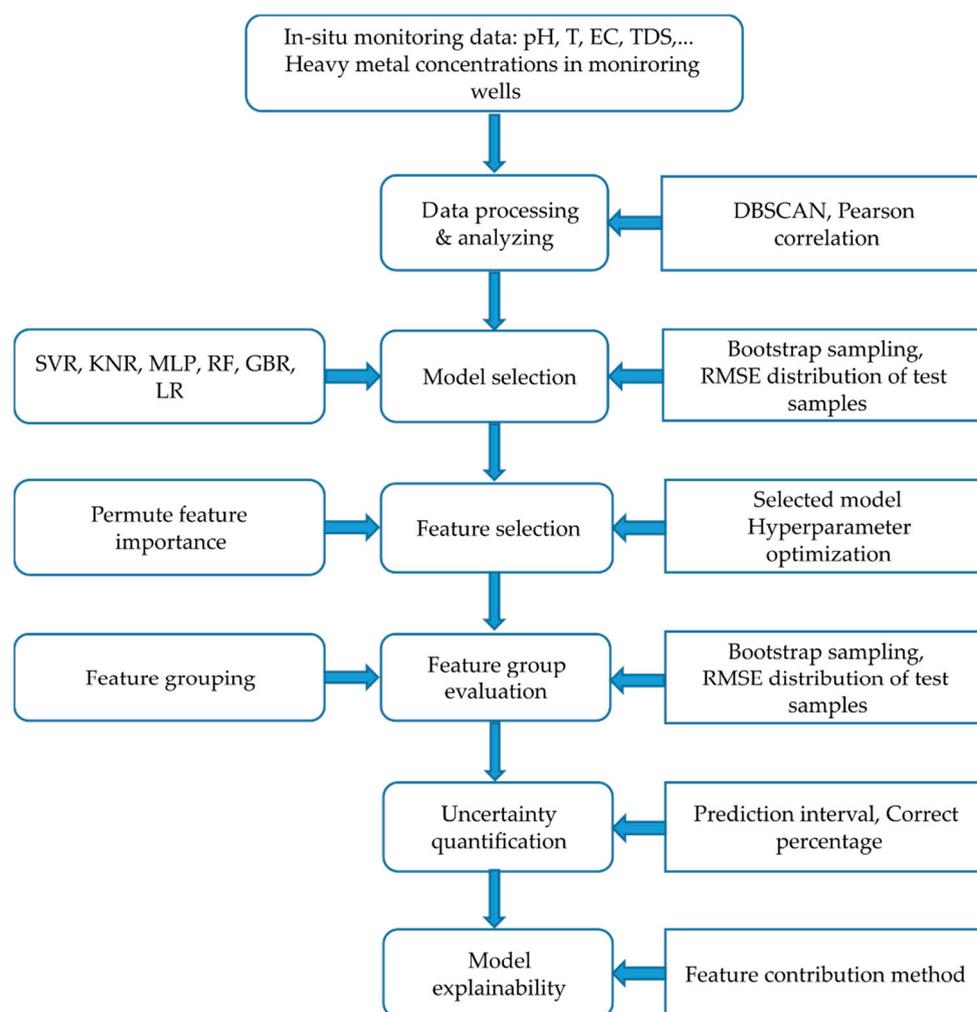
## 2.2. Interpretable Machine Learning

Explainable artificial intelligence (XAI) relates to using methods that explain how machine learning algorithms make decisions so that humans can understand them [38–41]. Concerning the scale of interpretability, global explanations determine which variables have the most predictive power, while local interpretations estimate how much each variable contributes to a prediction [42]. Permutation Feature Importance (PFI) is a global and model agnostic method that can be used for any model in selecting the most relevant features for predicted targets to decrease the required data and computational cost [43–46]. PFI only shows how sensitive each variable is but does not show how negatively or positively those features contribute to the model's output [47]. To obtain local explanations of a single prediction, many studies [48–52] have applied local interpretable model agnostic explanations (LIME) and Shapley additive Explanations (SHAP) because of their attractive visualization and multi-properties. However, SHAP-Kernel and SHAP-Deep Explainer are very slow and do not support some models. Feature coefficients provided by LIME are similar to the coefficients extracted from Bayesian linear regression and feature importance from decision trees [52]. For tree-based models, Treeinterpreter [53] outperformed SHAP-Tree Explainer [50] in terms of attribution accuracy computation cost [54].

## 2.3. Conceptual Framework

In this study, many popular machine learning algorithms such as Support Vector Regression (SVR), K-Nearest Neighbors (KNN), Feedforward Artificial Neural Network—Multilayer Perceptrons (MLP), Random Forest Regression (RFR), Gradient Boosting Regression (GBR) and Linear Regression (LR) are applied to achieve the objectives. Firstly, data discovery and pattern mining methods included clustering data by Density-based Spatial Clustering of Applications with Noise (DBSCAN) to find the similarity samples and outliers; we also analyzed multivariate statistics to discover the relationship between hydrochemical parameters. Secondly, the baseline model was trained with the most abundant cluster data, which was derived from DBSCAN, using all physicochemical parameters (Depth, Temp, EC, pH, TH, Cl, TDS, NO<sub>3</sub>, NH<sub>4</sub>, SO<sub>4</sub>, TOC) and two spatial features (X, Y geographic coordinate). We used a performance metric range from lower and upper bounds to express prediction uncertainty instead of a single value. A bootstrapping sampling was applied to generate sample variability for input perturbation. The distribution of model performance was evaluated to select which was the most suitable model. The best-performed model was selected for further analysis. Besides that, the selected model was optimized by a five-fold cross-validation of training data and the RandomsearchCV method to find the number of optimal features for each target. The tuned hyperparameters included max\_depth, min\_samples\_leaf, max\_features, min\_samples\_split, and n\_estimators. After optimizing, the models were applied to assess the feature contribution by ranking in PFI. Thirdly, we examined the feasibility of low-cost parameters for prediction by comparing the performance distribution between three feature combinations: (1) the first feature set consisted of

full parameters; (2) the second feature set was selected by an optimal number of features that have a high rank in PFI; (3) the third feature set followed the criteria of technical and economic efficiency that are quick and inexpensive to measure on-site, irrespective of its sensitivity. Finally, model uncertainty was quantified by the model prediction intervals (PI) and the percentage of expected values falling into the intervals. The added interpretability analysis also enhanced the model's trustworthiness. The general research procedure is illustrated in Figure 1. Although the applied methods in the framework are not highly complex, they are easy to use and computationally efficient. With the combination of all the procedures, we were able to quickly discover new insights into the capacity of the indicators and models.



**Figure 1.** A Conceptual research framework.

The study was conducted using Python scripting on the PyCharm integrated development environment. Scikit-learn [55], Matplotlib [56], Pandas [57], Numpy, Treeinterpreter, and other packages were used to perform the analysis, simulation, and output charts.

#### 2.4. Evaluation Metrics

There were three metrics used for the model's evaluation. The  $R^2$  (the coefficient of determination) is a typical metric for quantifying the variance in outputs of linear models [58], indicating the goodness of fit (Equation (1)). The  $R^2$  score ranges from  $-\infty$  to 1; thus, the closer to 1, the better the prediction is. Root mean squared error (RMSE) is the standard deviation of the residuals, computed as Equation (2). When compared to mean absolute error, RMSE can provide a more reliable error distribution

and sensitivity within a large sample size [59]. Moreover, the model's uncertainty is expressed by the Prediction Interval Coverage Probability (PICP) to find the possibility of expected values falling into the interval (Equation (3)) and the Mean Prediction Interval (MPI) (Equation (4)) to indicate the average width of all prediction intervals. The lower the MPI, the lower the uncertainties. If PICP is greater than the probability quantile, the uncertainty is overestimated, otherwise it is underestimated. PICP and MPI have been used in [22,60–62] to evaluate prediction uncertainties.

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (1)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (2)$$

$$PICP = \frac{\text{count}(N | Q_{low} \leq N \leq Q_{high})}{n} * 100\% \quad (3)$$

$$MPI = \frac{1}{n} \sum_{i=1}^n (Q_{high} - Q_{low}) \quad (4)$$

where  $y_i$  is the observed values, and  $\hat{y}_i$  is the estimated values. Notations  $Q_{low}$  represent lower quantile, and  $Q_{high}$  is for the upper quantile.

### 2.5. Study Site and Data Sources

Taiwan covers 35,808 square kilometers (35,801 km<sup>2</sup>), with 70% coverage of rugged and densely forested mountains as the spine in the central. The flatter area is located along the west coast, which is also densely inhabited. Uneven rainfall and steep-sloped rivers force residents to depend on groundwater resources significantly [63]. Since 2002, the Taiwan Environmental Protection Administration (EPA) has monitored groundwater with a relevant sampling frequency (monthly or seasonally, half-yearly) at different authority levels. All the monitoring data are updated on the EPA Taiwan website for public access, whether being single-well-, administrative- or watershed-level [64]. The study obtained physicochemical data of groundwater monitoring data from 453 wells over ten groundwater basins of Taiwan from 2000 to 2020. The location of monitored wells and the defined basins are shown in Figure 2.

In this study, the water quality records were downloaded from the Taiwan EPA website [65]. According to the Taiwan Groundwater Pollution Monitoring Standard, As and Mn concentrations for domestic supply and irrigation purposes are limited at 0.025 mg/L and 0.25 mg/L, respectively, whereas Fe values are 0.15 mg/L and 1.5 mg/L, respectively. The 2019 annual water quality monitoring report of Taiwan EPA revealed that only 53.2% of manganese (Mn) samples achieved the standard, while 73.4% of iron (Fe) is under-controlled [64]. A quick scan of collected data also shows that Mn, Fe, and As comprise levels of 27%, 10%, and 0.3% in samples, respectively, exceeding the standard levels, while other trace elements (Cr, Cd, Cu, Pb, and Zn) are much lower than the limits. Therefore, As, Fe, and Mn were targeted in this experiment. Fourteen physicochemical constituents were used for analysis. After the data preprocessing, there were 20,685 groundwater samples in the period of 2000–2020 used for modeling. Table 1 shows the statistical summary of the data.

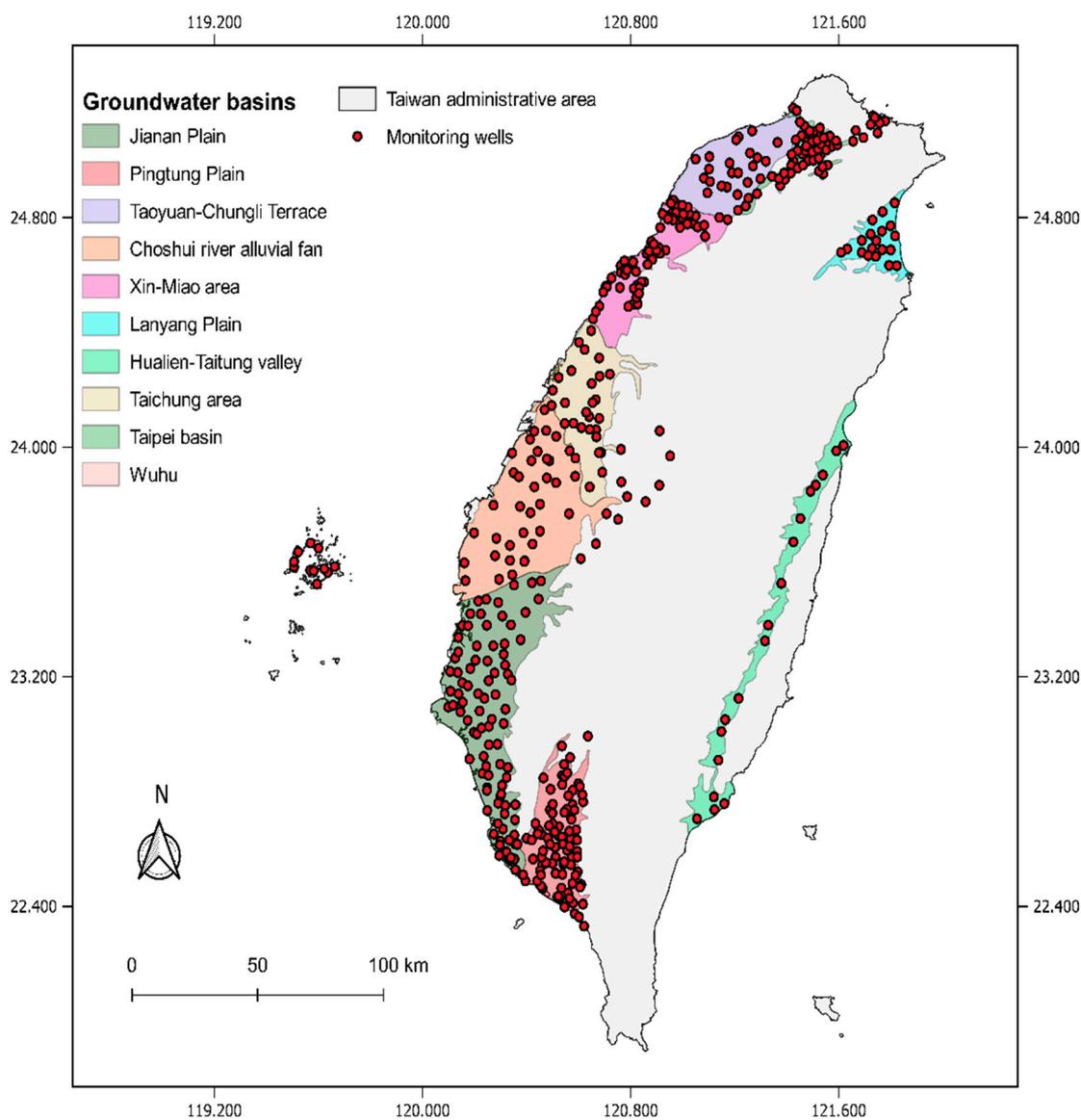


Figure 2. The locations of 453 observation wells in ten groundwater basins in Taiwan.

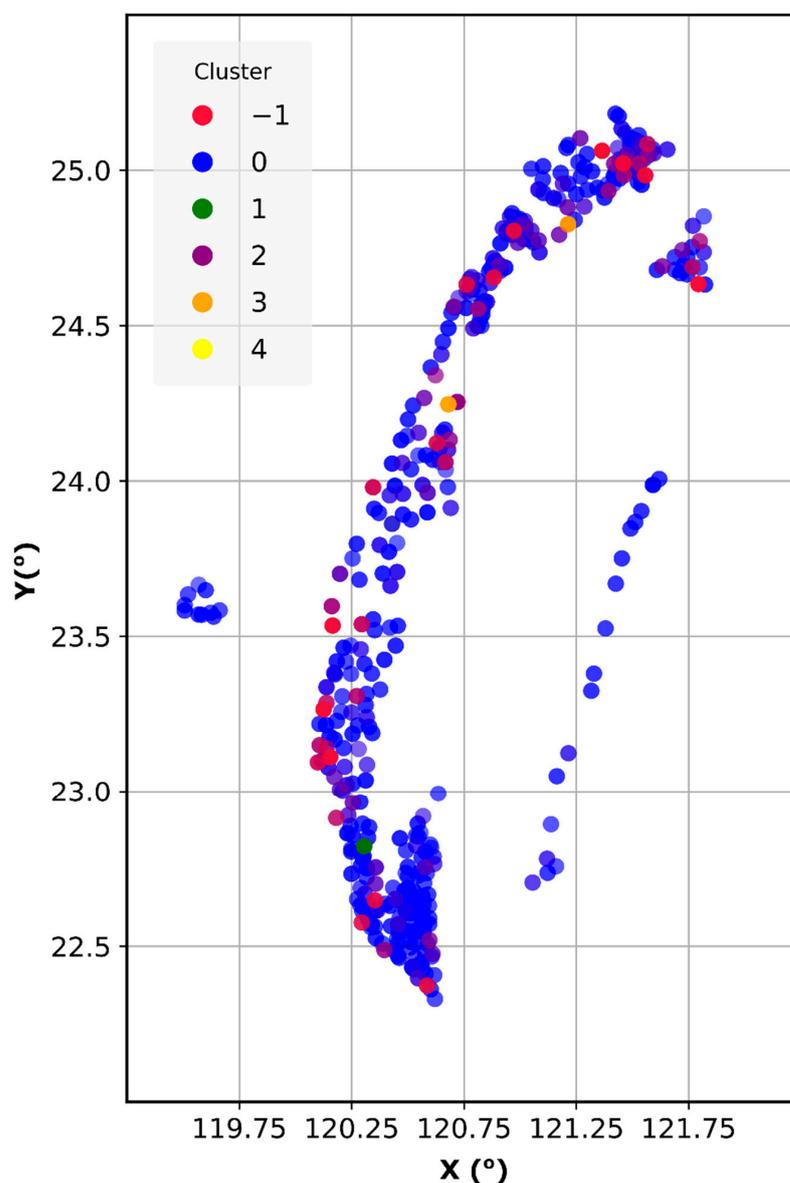
Table 1. Statistical summary of physicochemical parameters.

Variables	Description	Min <sup>a</sup>	Max <sup>b</sup>	Mean	STD <sup>c</sup>	Ske <sup>d</sup>	Kur <sup>e</sup>	Unit
Temp	Water temperature	18.600	33.200	26.524	1.704	−0.287	0.449	°C
Depth	Depth to water	0.000	39.627	4.944	4.569	2.777	10.014	m
EC	Electrical conductivity	2.000	65,800.000	1899.373	6205.603	6.419	43.874	μS/cm 25 °C
pH	pH	4.100	9.300	6.739	0.552	−0.922	1.799	-
TH	Total hardness	2.700	8390.000	416.436	735.115	6.207	43.697	mg/L
TDS	Total dissolved solids	4.100	52,300.000	1302.772	4502.321	6.689	48.265	mg/L
Cl	Chloride salt	0.500	27,800.000	454.055	2268.745	6.763	49.321	mg/L
NH <sub>4</sub>	Ammonia Nitrogen	0.001	20.000	0.791	1.704	3.9	19.951	mg/L
NO <sub>3</sub>	Nitrate Nitrogen	0.010	45.500	2.170	3.708	3.22	15.553	mg/L
SO <sub>4</sub>	Sulfate	0.500	4260.000	139.139	312.284	6.386	46.222	mg/L
TOC	Total organic carbon	0.020	15.800	1.922	1.549	2.453	9.442	mg/L
As	Arsenic	0.000	0.146	0.006	0.013	4.031	20.404	mg/L
Mn	Manganese	0.005	11.300	0.519	0.854	4.152	26.062	mg/L
Fe	Iron	0.000	58.800	1.526	4.161	5.821	44.764	mg/L

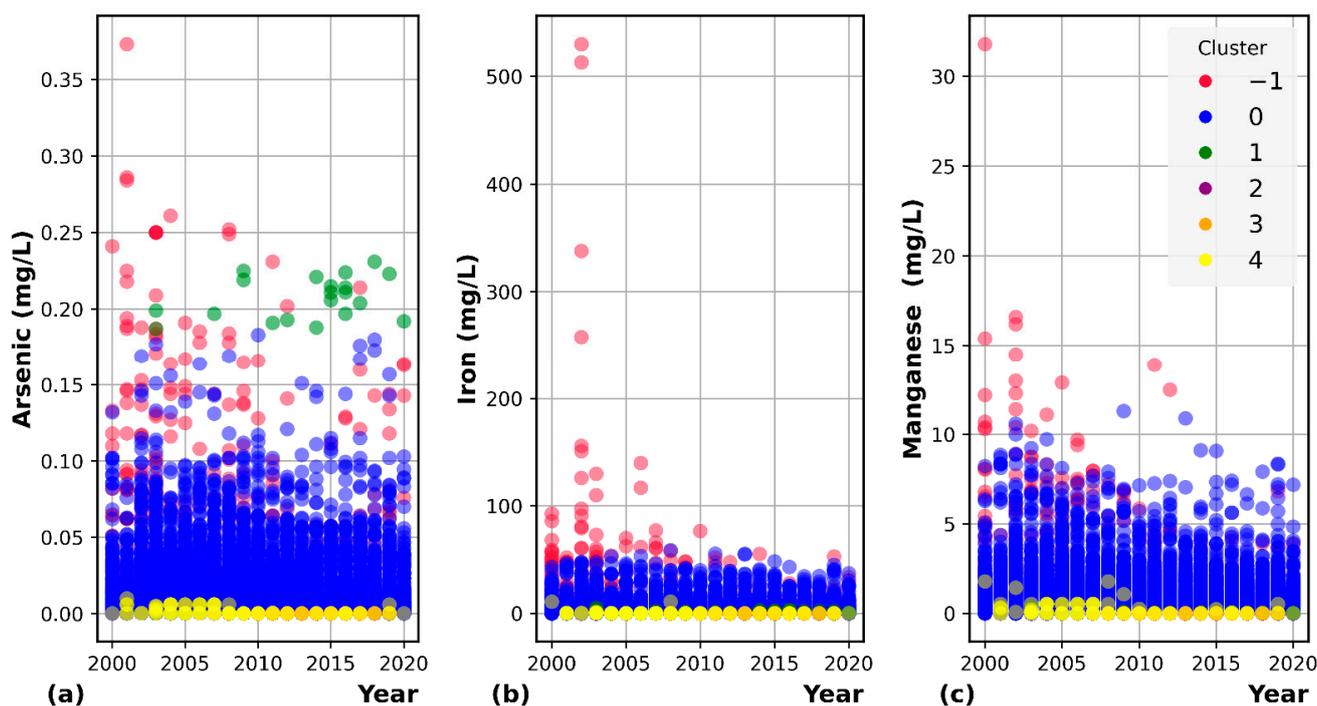
<sup>a</sup> minimum; <sup>b</sup> maximum; <sup>c</sup> standard deviation; <sup>d</sup> skewness coefficient; <sup>e</sup> kurtosis coefficient.

### 3. Data Analysis and Feature Engineering

The clustering analysis by DBSCAN shows the distribution of groundwater quality deviating through a spatiotemporal scale. The algorithm generated six unique clusters, including five clusters (from 0 to 4), based on their similarities and noise data (cluster-1). To illustrate how the clusters distribute across spatial scales, we added geographical coordinates and show them on the scatter chart in Figure 3. The concentration of target parameters (Mn, Fe, As) from each cluster is shown in Figure 4. There is a distinguishable characteristic of different clusters. For instance, Cluster 1 contains all extreme values considered outliers from most of the sampling locations. Cluster 0 can represent groundwater quality in Taiwan because it includes the most abundant samples in over ten groundwater basins. On the other hand, Cluster 1 to 4 have a smaller sample size, characterized by small As, Fe, and Mn concentrations in specific basins. Concerning temporal distribution, Cluster 0 is similarly distributed throughout the years, but minor clusters are scattered/distributed in some periods. In order to prepare data for As, Mn, and Fe prediction, Cluster 0 was used for training and testing the datasets, while all the other clusters with particular characteristics were filtered out.



**Figure 3.** Scatter plot showing the spatial distribution of derived clusters and outliers.



**Figure 4.** Scatter plot showing the temporal distribution of arsenic (a), iron (b), and manganese (c) in each cluster.

The Spearman correlation coefficient was applied to assess the interrelationship between parameters in Cluster 0 because this method can capture the non-linear relationship between variables. The coefficients are close to 1 for a similar trend and  $-1$  for an opposite trend, while nearly 0 means no relationship. Figure 5 shows a highlighted relationship between salinity indicators, namely EC, TDS, Cl, SO<sub>4</sub>, and TH. Moreover, salinity indicators have a slight opposite trend with X. It can be interpreted that the larger X values (from the west coast toward the east), the lower the salinity levels are. There was no clear relationship between Y and salinity. The water temperature slightly increases toward the south (decreasing Y) and the west (decreasing X). Considering the feature–target relationship, As, Mn, and Fe have a slight negative correlation with NO<sub>3</sub> and a positive correlation with NH<sub>4</sub>. The opposite trends of NO<sub>3</sub> and NH<sub>4</sub> can result from the oxygen concentration in water. Saltwater reduces dissolved oxygen, leading to an anoxic condition and denitrification. Therefore, As, Mn, and Fe may be related to the salinity level. Salinity had robust effects on Mn and Fe solubility [66]. Other features do not have a clear correlation with targets, but they may have non-linear relationships.

Preprocessing steps included filling missing values, removing bad attributes, adding features, resampling, and normalizing samples. In consideration of the spatial heterogeneity of samples, there is no other convenient information available except for locations. That is why longitude and latitude coordinates (X, Y) were also used as spatial feature inputs for prediction. To reduce the model complexity for point estimation, we shuffled the samples before the train–test partition and ignored the temporal dimension. Moreover, the sample size at each location should be over 30 samples for meaningful statistical analysis. As a result, from 403 wells of Cluster 0, we have 20,685 samples for training and testing. Finally, all input features were normalized into a fixed range between 0 and 1 to speed up the computation.

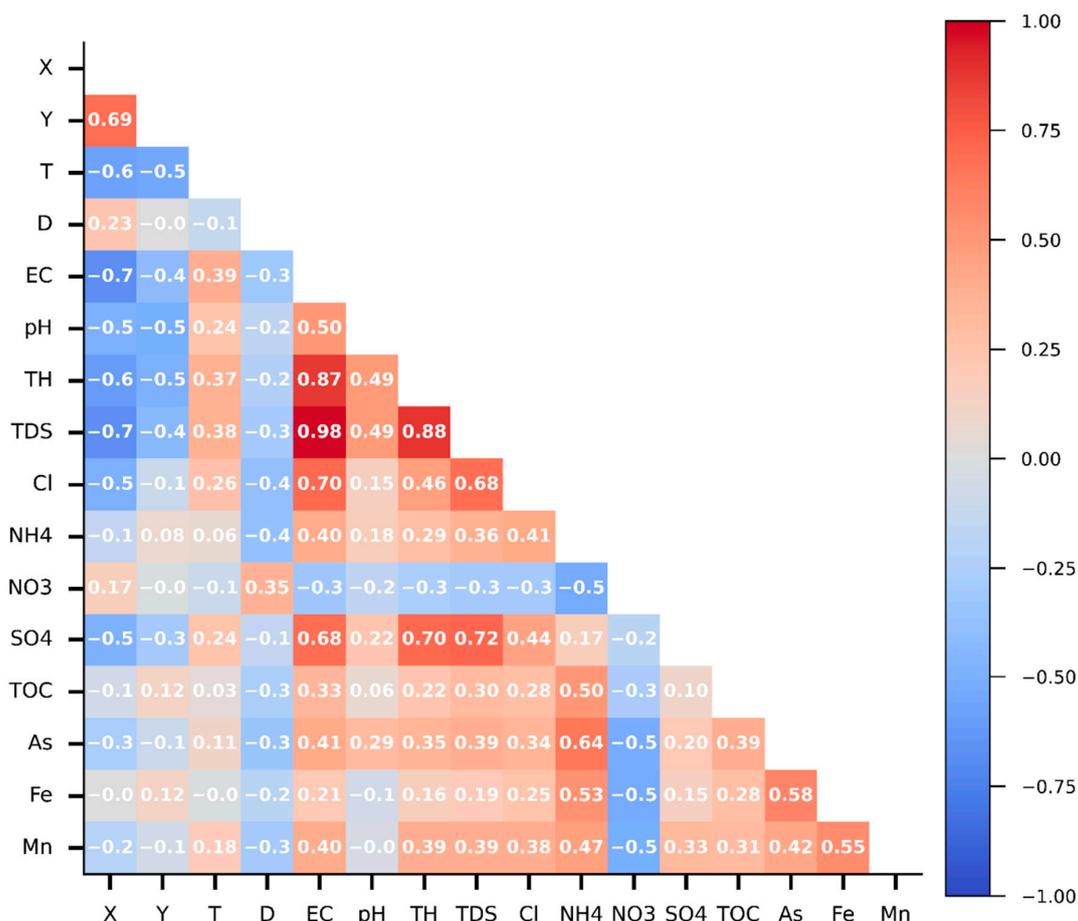


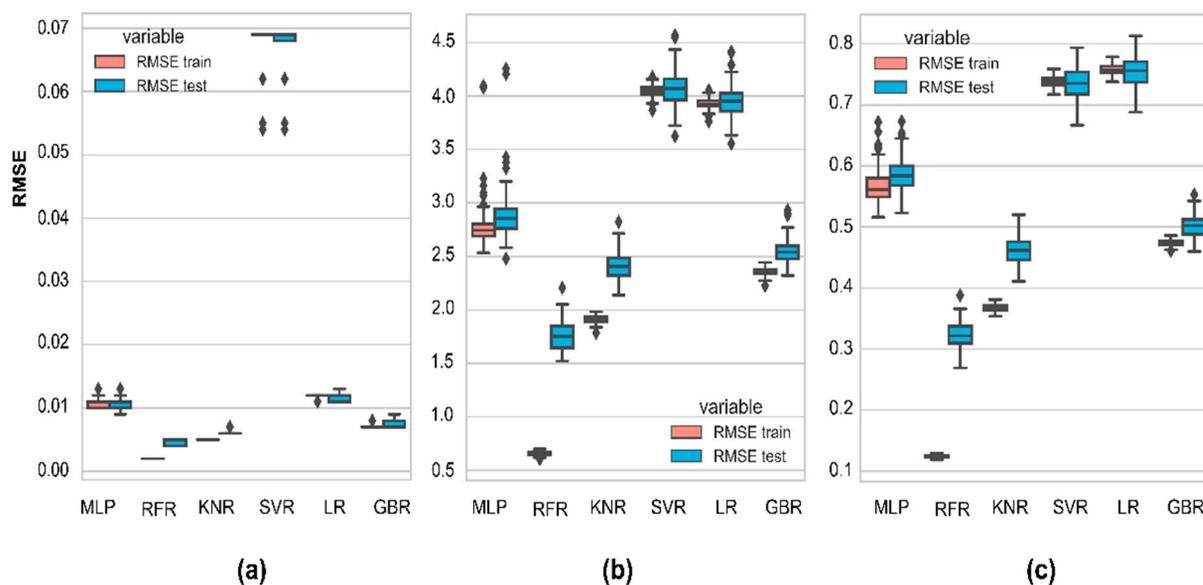
Figure 5. Spearman correlation matrix for all parameters.

### 4. Results

#### 4.1. Assessment of the model Predictability

It is necessary to evaluate model robustness by comparing result distribution since average metrics cannot precisely represent how stable the model performance is against new inputs. Initially, we trained six regression algorithms (MLP, RFR, KNR, SVR, LR, and GBR) to predict As, Fe, and Mn, using all features for 100 data sets to find which model is the most robust. Because  $R^2$  scores can have negative values, the RMSE values were used for plotting charts. The prediction biases of As, Fe, and Mn from 100 random datasets are shown in the box and whiskers plots in Figure 6 to illustrate the stochastic nature of the algorithm’s performance due to input changes.

Overall, the RMSE distribution of RFR in all target species (Figure 6a–c) has the lowest mean errors in the training and testing data sets, indicating that RFR produces fewer biases than other models. From As prediction (Figure 6b), all models (except for SVR) are robust to data perturbations because their interquartile ranges of errors are small and less susceptible to outliers. There are also tiny gaps between training and testing performances caused by the similar training and testing data distribution or generalized models. However, Fe prediction (Figure 6b) and Mn prediction (Figure 6c) experienced similar problems, with potential outliers resulting from noisy data sets or highly skewed distributions. The significant differences in the training and testing error distributions of RFR and KNR exhibit overfitting models. In fact, it is hard to build a model that is ideally fitted to all new data. Overfitting models learn noise rather than actual signals, but they can be improved through hyperparameter optimization or input regularization. That is why RFRs were considered even though they are overfitted. Those models with a low variance and high bias may need more data or features that we did not consider in this scope.



**Figure 6.** Boxplots of model performance (RMSE) on 100 datasets: (a) As prediction; (b) Fe prediction; (c) Mn prediction.

In addition, the testing  $R^2$  scores of the RFR models for all targets are the highest, followed by KNR and GBR (Table 2). All RFR models predicting As, Fe, and Mn have high average fitting scores (over 0.7), while LR and SVR seem unsuitable for the data. LR and SVR have average  $R^2$  scores lower than the satisfactory criteria ( $<0.5$ ). Therefore, RFR was selected to predict As, Fe, and Mn.

**Table 2.** Model performance on 100 testing data sets for different targets (mean  $R^2$  score  $\pm$  standard deviation).

Models	As Prediction	Fe Prediction	Mn Prediction
GBR	$0.65 \pm 0.02$	$0.59 \pm 0.03$	$0.62 \pm 0.02$
KNR	$0.73 \pm 0.02$	$0.60 \pm 0.03$	$0.65 \pm 0.03$
LR	$0.18 \pm 0.01$	$0.08 \pm 0.01$	$0.18 \pm 0.01$
MLP	$0.30 \pm 0.06$	$0.47 \pm 0.06$	$0.50 \pm 0.08$
RFR	$0.79 \pm 0.02$	$0.70 \pm 0.03$	$0.76 \pm 0.02$
SVR	$-26.32 \pm 1.88$	$0.02 \pm 0.01$	$0.21 \pm 0.02$

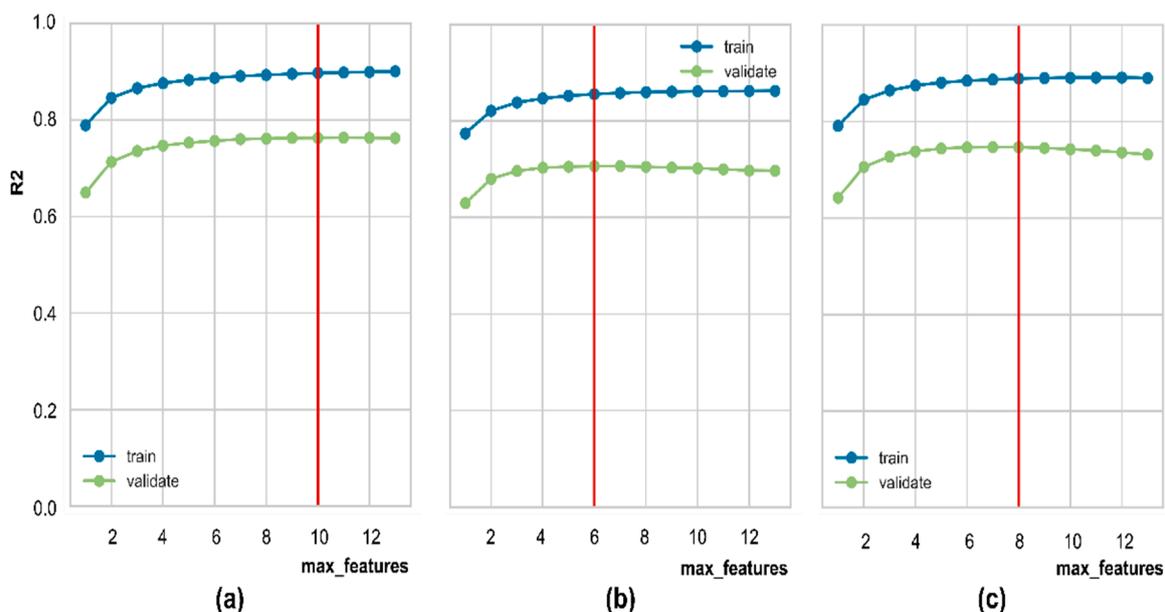
#### 4.2. Assessment of the Feature Predictability

The predictability of feature sets was evaluated by comparing performance metrics between feature groups toward each target. Then, the feature sets were selected through three steps.

Firstly, three RFR models for three targets were successively optimized by Random search CV and five-fold cross-validation to find the optimal values for hyperparameters. Some primary hyperparameters were used for tuning, namely `max_depth`, `min_samples_leaf`, `max_features`, `min_samples_split`, and `n_estimators`. The best configuration of each model is reported in Table 3. Figure 7 shows the validation curves for the number of features from each model. The red lines from Figure 7a–c are the optimized feature for As, Fe, and Mn models. For instance, the As model needs a maximum of ten features, while Fe and Mn models require six and eight features, respectively. Although the gaps between train and test errors can remain somewhat large at the stopping points, the validation scores will not decrease due to overfitting. An efficient model will require less input but perform satisfactorily for testing data. Fewer redundant features mean fewer opportunities to make decisions based on noise. Therefore, fewer features reduce the algorithm's complexity, resulting in faster model training [67].

**Table 3.** Random Forest Regressor hyperparameter optimization.

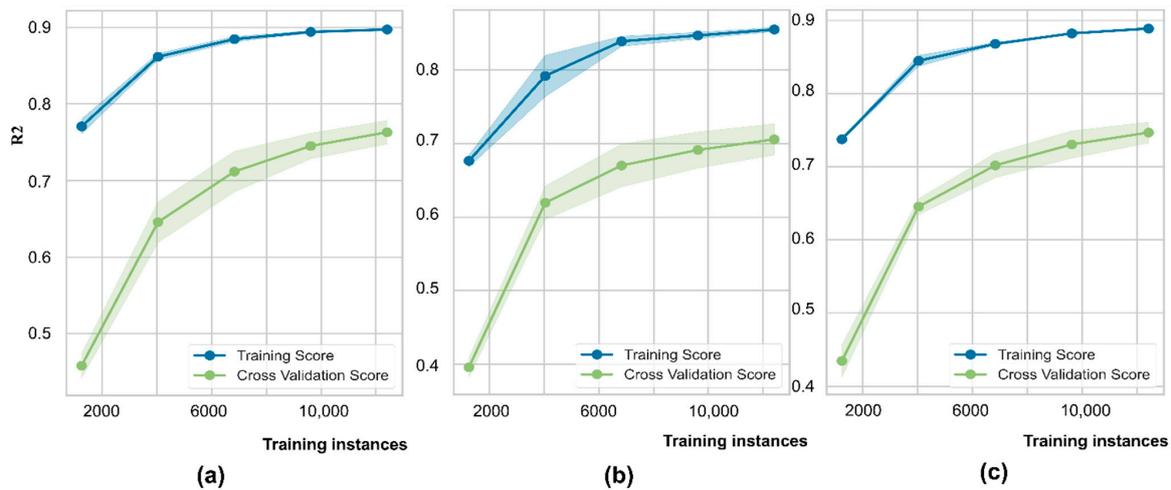
Hyperparameters	Description	As Model	Fe Model	Mn Model
min_samples_leaf	The lowest number of observations in a terminal node	4	4	4
max_features	Number of variables for the best split	10	6	8
min_samples_split	The lowest number of observations needed to split an internal node	6	4	8
n_estimators	Number of trees in a forest	1848	1727	1000
max_depth	The maximum depth of the tree	16	18	20



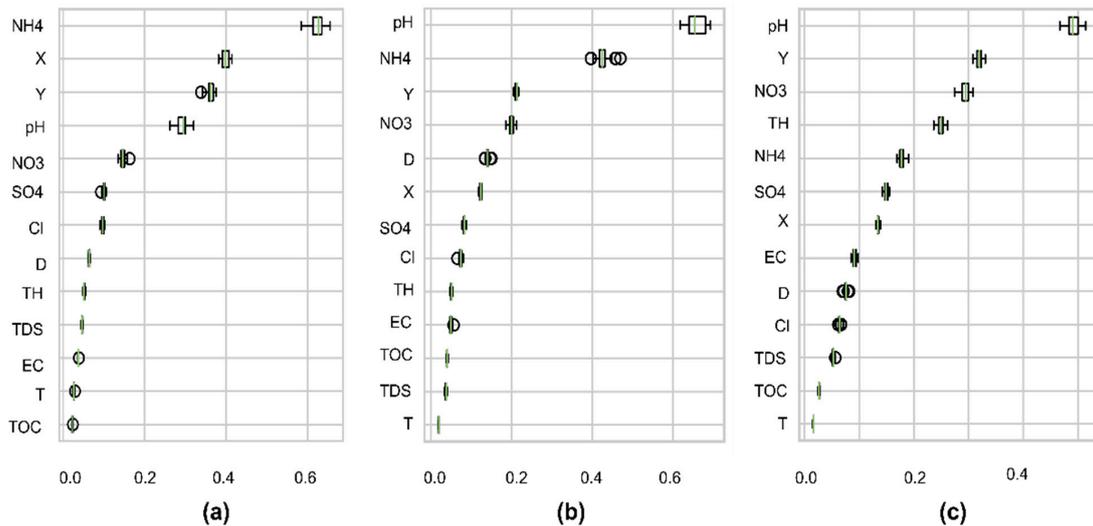
**Figure 7.** Validation curve of R<sup>2</sup> score versus max\_features: (a) As model; (b) Fe model; (c) Mn model. Vertical red lines indicate optimal max\_features.

The learning capability of the optimized models is expressed by learning curves, which show the behaviors of training scores and validation scores responding to sample numbers. From Figure 8, the training and validating curves have not yet converged as the data are increasing. Variability during cross-validation was higher than during training, showing that the models suffer from variance (overfitting) rather than bias. The dataset is so unrepresentative that the model cannot capture the statistical characteristics. Potentially, the validation scores could increase and would be closer to the training scores if more data are trained to generalize more effectively.

Secondly, as to which variable is selected for each group depends on the order in the feature importance process. Figure 9 reveals the sensitivity of each feature to the predicted targets. It can be seen that NH<sub>4</sub>, NO<sub>3</sub>, pH, X, and Y showed a strong effect in the three models, while TOC and Temp had a negligible impact on the results. As expected, the correlation of NH<sub>4</sub> and NO<sub>3</sub> with targets also affects the importance of analysis. pH, X, and Y have unclear correlations with targets but also have high importance. This result also proves that the most important features are not always the most statistically correlated features [46]. The ionic forms of heavy metals also relate to pH levels and oxidation–reduction conditions. For instance, increasing pH values cause the heavy metal precipitation to increase, whereas the solubility decreases [68–70]. The high ranks of X and Y show that spatial heterogeneity rather than physicochemical parameters influence the results. Additionally, RFR often gives low importance to those features with collinear relations, such as EC, TDS, although they may have physical meaning to the targets.



**Figure 8.** Learning curves of training versus 5-fold cross-validation: (a) As model; (b) Fe model; and (c) Mn model.



**Figure 9.** Permute importance distributions on training data: the vertical axis is the feature names, the horizontal axis is the prediction of score reduction when permuting that feature: (a) As model; (b) Fe model; and (c) Mn model.

Based on the optimized number of features required for each model in Figure 7 and feature ranks in Figure 9, we divided the input features into three groups: full features, important features, and low-cost features, as given in Table 4. Essential feature sets were selected from the highest downward until the required max features. The low-cost features should be quick and inexpensive to measure by one multiparameter probe; thus, a slow or expensive feature from the important feature sets was replaced. For instance, SO<sub>4</sub> is harder to measure by the same probe; it was replaced by EC (in the As model) or Cl (in the Mn model) as a low-cost feature. Finally, although important features from the Fe model can be measured easily by one multiparameter sensor, another low-cost feature set was created to evaluate.

Finally, we compared the performance distributions of the RFR between full features and reduced inputs on 100 testing data sets. This experiment aimed to find the most robust features for each model. The goodness-of-fit distribution was compared in Figure 10. In general, all feature sets successfully fit targets with a minimum R<sup>2</sup> score higher than 0.6. The As model performances by three feature groups were almost similar (Figure 10a), while the reduced feature sets of the Fe and Mn model caused a decrease in model performance (Figure 10b,c). The Wilcoxon signed-rank test was applied to calculate the performance

discrimination among benchmark models (full features) and other feature sets. The null hypothesis was that the performance of the paired feature sets is similar. If the *p*-value is more significant than 0.05, they have similar distributions; otherwise, they come from different distributions. As shown in Table 5, the important feature scores and the low-cost feature scores of As model are similar to the full feature scores; thus, using low-cost features is more beneficial. Although the Fe model’s important feature scores and low-cost feature scores are different from the full feature scores, the average low-cost feature score is higher and more robust to data changes in the Fe model. In Mn prediction, both important feature scores and low-cost feature scores have the same distribution and are worse than the full feature scores. Therefore, low-cost features can predict As, Fe and Mn.

Table 4. Feature grouping for models.

Feature Set	As Model	Fe Model	Mn Model
Full features	X, Y, pH, TH, EC, TDS, Cl, SO4, NO3, NH4, TOC, Temp, Depth		
Important features	NH4, X, Y, pH, NO3, SO4, Cl, Depth, TH, TDS	pH, NH4, Y, NO3, Depth, X	pH, Y, NO3, TH, NH4, SO4, X, EC
Low-cost features	NH4, X, Y, pH, NO3, EC, Cl, Depth, TH, TDS	pH, NH4, Y, EC, Depth, X	pH, Y, NO3, TH, NH4, Cl, X, EC

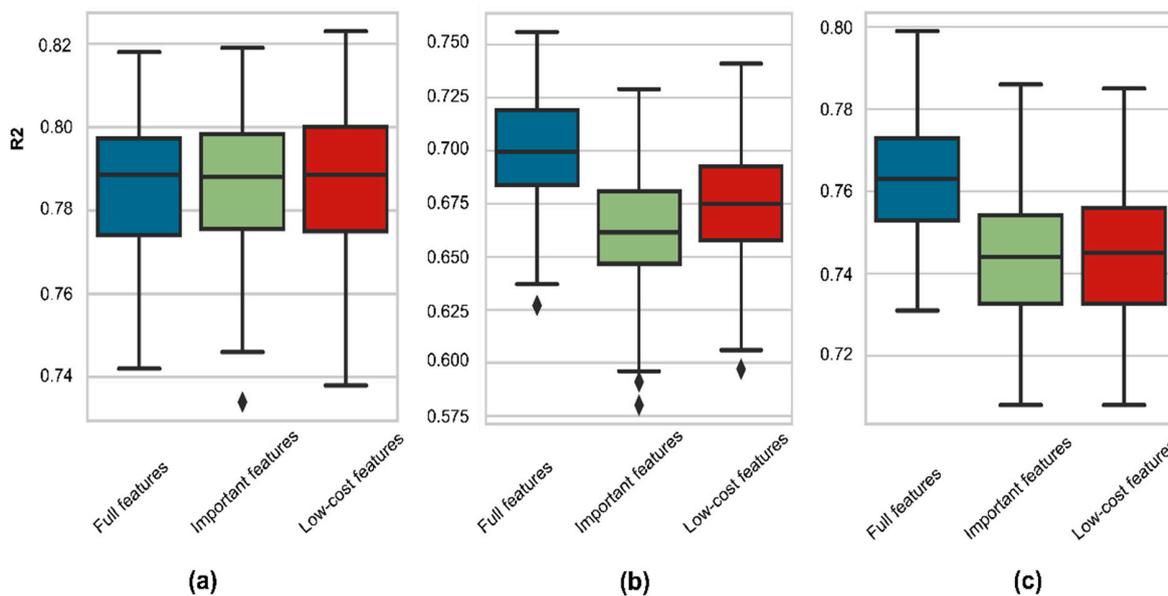


Figure 10. Boxplots of model performance ( $R^2$ ) on 100 random testing datasets: (a) As model; (b) Fe model; (c) Mn model.

Table 5. Results of Wilcoxon-signed rank test on  $R^2$  scores of 100 testing data sets: statistic (*p*-value).

Paired Tests	As Model	Fe Model	Mn Model
Full features—Important features	2065.000 (0.912)	0.000 (0.000)	0.000 (0.000)
Full features—Low-cost features	1883.000 (0.140)	6.000 (0.000)	0.000 (0.000)
Important features—Low-cost features	1396.500 (0.009)	661.500 (0.000)	1697.000 (0.277)

### 4.3. Quantification of Predictive Uncertainty

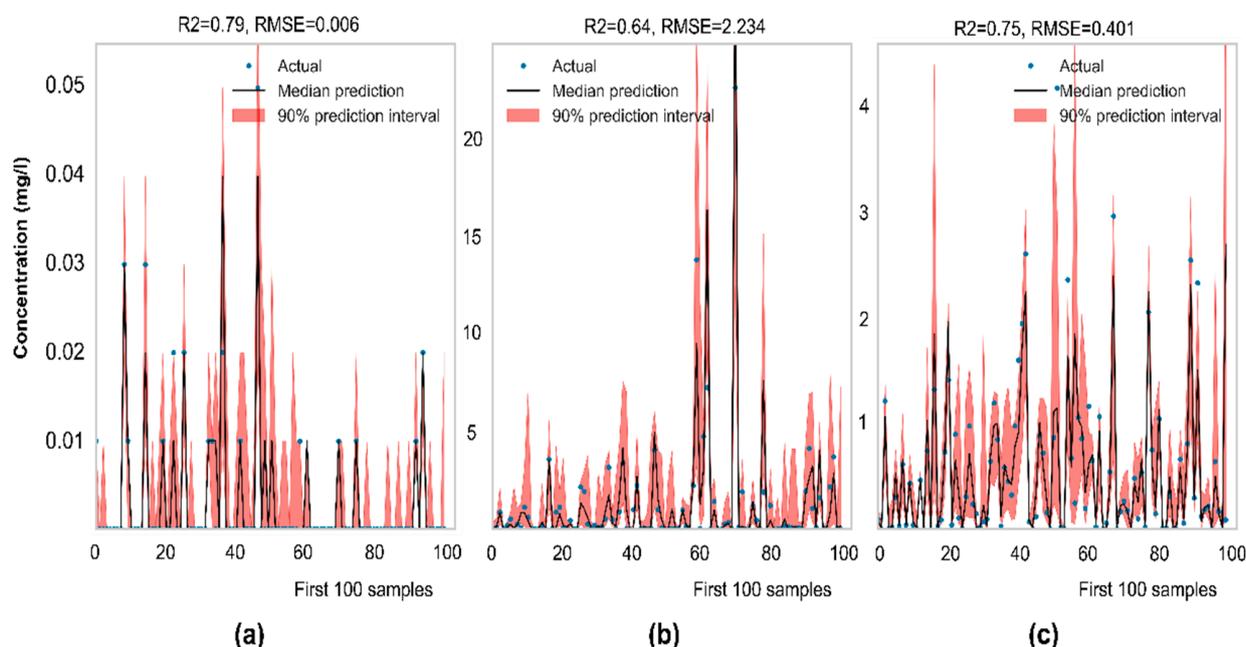
This evaluation explains how confident the models can perform on one data set. The prediction interval for a response is constructed from the results of single decision trees in the optimized RFR. In order words, each output value from the As model, Fe model, and Mn model was aggregated from 1848, 1727, and 1000 possibilities, respectively. We expected the PICP of 90% prediction interval to be around 90%. The lower values of MPI are the lower uncertainties. Table 6 summarizes the comparison of prediction uncertainties

generated by different feature sets. The full features produce the highest uncertainties among the three feature sets, whereas the low-cost features generate almost the lowest uncertainties. Very high PICPs of the As model by three feature sets indicate that the model overestimates the noise, while the Fe and Mn models underestimate the uncertainties. The real uncertainties may be higher than calculated, but the Mn models are more reliable than the Fe models. Hence, the low-cost features can predict heavy metals but need more regularization to improve the model's generalization.

**Table 6.** Coverage probability of 90% prediction intervals from different inputs: PICP (MPI).

Feature Sets	As Model	Fe Model	Mn Model
Full features	98.32 (0.0174)	80.53 (6.1914)	87.65 (1.4521)
Important features	98.05 (0.0154)	75.02 (4.8758)	86.31 (1.2356)
Low-cost features	98.07 (0.0152)	77.05 (4.9346)	86.41 (1.2340)

Uncertainty simulation results of low-cost features for three models are illustrated in Figure 11. Overall, the observed values tend to fall into the prediction intervals, masked by the red shading areas. The median predictive performance of the As model ( $R^2 = 0.80$ ,  $RMSE = 0.006$ ) is the highest, followed by the Mn model ( $R^2 = 0.75$ ,  $RMSE = 0.401$ ) and Fe model ( $R^2 = 0.65$ ,  $RMSE = 2.219$ ). It can be seen that the Fe model and Mn model cannot capture data patterns well, resulting in high variance. The reason may come from the lack of good predictors because the absence of causal links will limit machine learning algorithms from drawing desirable outcomes [11]. Besides, bias could be yielded from too few features or inference of false feature relationships [19]. Reducing features will ignore some essential features relevant to predictability and cause a significant impact on predictability. This is contrary to the Fe model. Even the full features also cause the largest uncertainties. This situation may result from unrepresentative data.



**Figure 11.** Visualization of observed data, median prediction line, and interval prediction by low-cost features: (a) As model, PICP = 97.97%, MPI = 0.015; (b) Fe model, PICP = 77.13%, MPI = 4.907; and (c) Mn model, PICP = 86.41%, MPI = 1.230.

#### 4.4. Interpretability of the Proposed Models

This analysis was aggregated from Treeinterpreter outputs. The predicted values were decomposed into a linear combination of feature contributions and biases to understand

how the model estimates those targets. The bias was assumed to be the same for training and testing data; thus, the differences in feature contributions will produce variability. Figure 12 shows the coefficient distribution for each target in training and testing data. These graphs describe the local interpretability for a single prediction. NO3 and NH4 have the most significant contribution range, while the other features are susceptible to outliers, indicating scattered or non-normal distributed data.

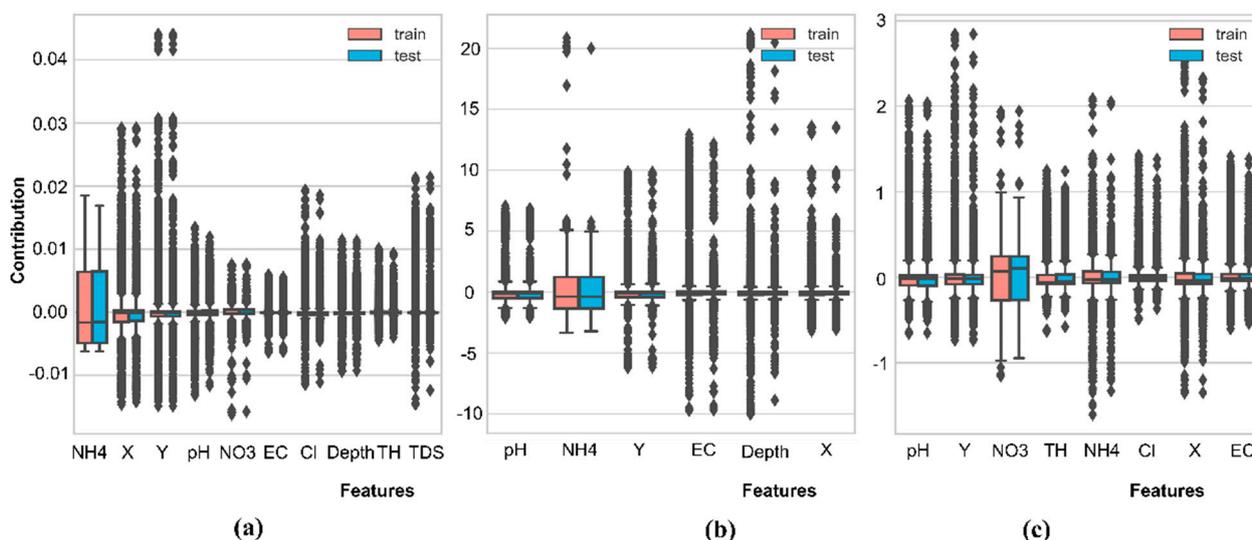


Figure 12. Distribution of local feature contributions for each predicted sample: (a) As model; (b) Fe model; (c) Mn model.

In general, the global feature contributions were calculated by averaging the local coefficients of each prediction (Table 7). Nutrients, spatial characteristics, and salinity highly contribute to As, Fe, and Mn predictions. Under more reduction conditions, Fe and Mn have more extensive concentration ranges [71]. These results are suitable with earlier correlation analyses and explanations. The dissimilar ranking in feature contributions in training and testing shows that either the decomposition method cannot explain the Fe and Mn models or that the training and testing sets come from different distributions. Moreover, the feature contribution ranking presented by the three models in the training stage is different from the ranking in feature importance (see Figure 9), indicating that the models do not generalize well [72]. In order to have a more efficient explanation of model behaviors, samples with different characteristics can be separated into different explanatory models or even different predictive models. Due to time and computation constraints, the interrelationship of features has not been calculated.

Table 7. Global feature contributions in each model.

Features	As Model		Fe Model		Mn Model	
	Train	Test	Train	Test	Train	Test
Cl	$4.33 \times 10^{-6}$	$-1.86 \times 10^{-5}$			$6.56 \times 10^{-4}$	$2.76 \times 10^{-3}$
Depth	$4.85 \times 10^{-6}$	$3.63 \times 10^{-5}$	$4.34 \times 10^{-3}$	$-7.28 \times 10^{-3}$		
EC	$4.49 \times 10^{-6}$	$-1.71 \times 10^{-6}$	$3.19 \times 10^{-3}$	$-3.52 \times 10^{-2}$	$4.19 \times 10^{-4}$	$-2.94 \times 10^{-3}$
NH4	$6.13 \times 10^{-6}$	$9.45 \times 10^{-5}$	$2.56 \times 10^{-3}$	$-6.71 \times 10^{-4}$	$3.62 \times 10^{-4}$	$-1.85 \times 10^{-3}$
NO3	$6.87 \times 10^{-6}$	$4.68 \times 10^{-5}$			$3.64 \times 10^{-4}$	$7.00 \times 10^{-3}$
pH	$6.86 \times 10^{-8}$	$8.14 \times 10^{-5}$	$1.22 \times 10^{-3}$	$-5.52 \times 10^{-3}$	$-2.52 \times 10^{-4}$	$-4.82 \times 10^{-3}$
TDS	$1.49 \times 10^{-6}$	$2.25 \times 10^{-5}$				
TH	$3.10 \times 10^{-6}$	$6.76 \times 10^{-6}$			$5.27 \times 10^{-4}$	$4.19 \times 10^{-3}$
X	$7.58 \times 10^{-7}$	$2.95 \times 10^{-5}$	$1.17 \times 10^{-3}$	$-4.67 \times 10^{-3}$	$2.76 \times 10^{-4}$	$-2.61 \times 10^{-3}$
Y	$-2.42 \times 10^{-6}$	$7.36 \times 10^{-5}$	$-2.55 \times 10^{-4}$	$-9.63 \times 10^{-3}$	$7.97 \times 10^{-5}$	$-5.02 \times 10^{-3}$

## 5. Discussion

Although low-cost features can potentially predict heavy metals through this evaluation framework, both model configuration and feature sets are not fine-tuned for their actual application. Therefore, some issues need to be considered to yield better results.

Firstly, the selected models are not finely optimized due to the shortcomings of the randomized search method, namely, not fully exploring the hyperparameter search space. That is why the models are overfitted. They can be improved by applying other exhaustive optimization methods, such as Bayesian optimization or Grid search. However, Random Search could be relatively close to optimal performance but requires less computation for large sample sizes and many hyperparameters [73,74]. It is still suitable for quick scanning over reasonable hyperparameter ranges.

Secondly, the feature selection method by importance rank may have an adverse effect. It gives a single impact on model performance, whereas the inter-relationships are hidden. Hence, the selected features cannot capture data characteristics. Even though we used all the feature sets, the training scores of Fe and Mn were not too high. In both the As and Fe models, uncertainties were unavoidable due to the inherent uncertainties of the hydrological process [75]. Koutsoyiannis, 2003 [76] also found that hydroclimatic processes produce more uncertainties than estimation because assumptions are often based on a stable climate. In order to understand the multi-scalar behavior of contaminated substances at multiple spatio-temporal scales, using the entropy technique and the Hurst exponent can be very helpful [77]. However, natural structural factors, including climate, soil and aquifer formation, sources of groundwater recharge, or human impacts, may interrupt the long-range correlation of hydroclimatic processes [78]. Vu et al., 2019 [79] and 2021 [80] also found that land use, especially in agriculture-dominated regions, has a higher potential for groundwater contamination; thus, land use can be a good indicator of groundwater characteristics. A simple set of predictors cannot perform as well as expected; thus, adding more features may improve the model's performance.

Moreover, the learning curves indicate that adding more samples or modifying the data structure may improve the model performance. A suitable dataset is more effective than many predictors and sample sizes [81]. Data regularization can also smooth the noise to get more generalized predictions. Due to training the models on bootstrapping samples, the Hurst phenomenon in hydrological processes was excluded. Analyzing Hurst–Kolmogorov dynamics to find independence structures [82] and cluster data in smaller groups and training by different models is another strategy to cope with data heterogeneity at different scales. Simplifying the models then necessitates further effort to explore data distribution, instance structures, and target analysis to enhance the scalability of the predictive models. However, training data subsets with different models will amplify computational load and management. Improving learning efficiency can be done by simultaneously computing services [83].

Lastly, uncertainty quantification and model interpretation show the heterogeneity in the data structure. By using gene entropy techniques and the Hurst exponent to understand the multi-scalar behavior of nitrate-N in groundwater, this approach should readily be transferable to other contaminated aquifers and catchments. A realized model should increase accuracy by increasing the complexity, such as adding multi-source data collection and internet of things (IoT), geographic information systems (GIS), and numerical models into a large-scale model [84] to have more precise simulation at each spatial scale. As a result, the model may lose interpretability. The model interpretation can also be processed before optimization to remove unnecessary features. The selection of quick scan techniques should be prioritized to save time and effort. Therefore, seeking a good model and indicators is an exhaustive trial–error process.

## 6. Conclusions

The study has developed a scanning framework to test the feasibility of candidate machine learning models, including Support Vector Regression (SVR), K—Nearest Neighbors

(KNN), Feedforward Artificial Neural Network—Multilayer Perceptrons (MLP), Random Forest Regression (RFR), Gradient Boosting Regression (GBR) and Linear Regression (LR) and interpretability of the selected model. The samplings and experiments for heavy metals in groundwater have been challenging tasks and have drawn much attention in environmental sciences. The main issues are the test costs that grow exponentially with the concentration accuracy of the target heavy metals. The prediction of heavy metals using low-cost water quality data can considerably benefit the monitoring efficiency of heavy metals in groundwater systems. The study evaluated the predictability of machine learning models, compared the performance of different water quality indicators, and explored the reliability and explainability of the model for practical applications.

This study observed that heavy metals in groundwater, namely As, Fe, and Mn, could be predicted by Random Forest using low-cost physicochemical water quality samples. Results showed that the As model using ten predictors (NH<sub>4</sub>, X, Y, pH, NO<sub>3</sub>, EC, Cl, Depth, TH, TDS) had low bias ( $R^2 = 0.80$ ) and low uncertainties (PICP = 97.97%, MPI = 0.015). The Mn model using eight features (pH, Y, NO<sub>3</sub>, TH, NH<sub>4</sub>, Cl, X, EC) yielded a relatively satisfactory fit ( $R^2 = 0.75$ ) and slightly high uncertainties (PICP = 86.41%, MPI = 1.230). The Fe model needed the fewest features (pH, NH<sub>4</sub>, Y, EC, Depth, X) but has not been generalized, leading to the lowest fitting scores ( $R^2 = 0.65$ ) and high uncertainties (PICP = 77.13%, MPI = 4.907). The feature contribution simulations showed that nutrients, salinity, and spatial factors strongly affect heavy metal behaviors.

In the study, the predicted values were decomposed into a linear combination of feature contributions and biases to understand how the model estimates those targets. Results showed that NO<sub>3</sub> and NH<sub>4</sub> had the most significant contribution range, while the other features were susceptible to outliers, indicating scattered or non-normal distributed data. Nutrients, spatial characteristics, and salinity highly contributed to As, Fe, and Mn predictions. Although the built models were not optimally generalized, the results were promising. This framework addressed how accurate and sensitive the model performances were, how confidently the predictions covered, where the uncertainty sources were coming from, and how a particular instance was predicted. The limitation of this experiment is that one generalized model cannot fit all data patterns in a large area. Further investigation for more features, namely vegetation cover, climate, topography, land deformation, or soil properties, may improve the results. Because the data are heterogeneous, it can be localized in smaller groups by time, space, or chemical characteristics (hardness, salinity) before training. Other techniques, like clustering, classification, and mixture models, can be combined to improve the regression performance. However, those findings help in assessing the practicability of the proposed model for groundwater quality inspection.

**Author Contributions:** Conceptualization, T.-M.-T.H., C.-F.N.; data acquisition and curation, writing—original draft, T.-M.-T.H.; methodology, writing—review and editing, C.-F.N., Y.-S.S., V.-C.-N.N., I.-H.L., C.-P.L. and H.-H.N. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was partially supported by the Ministry of Science and Technology, the Republic of China, under grants MOST 107-2116-M-008-003-MY2, MOST 109-2621-M-008-003, MOST 109-2625-M-008-006, MOST 111-2410-H-019-006-MY3, MOST 111-2622-H-019-001, MOST 110-2116-M-008-006, and MOST 111-2116-M-008-008.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Available on Github: <https://github.com/trangminhhuynh/GWQ-prediction> (accessed on 18 September 2022).

**Conflicts of Interest:** The authors declare that they have no conflict of interest.

## References

1. Vijayakumar, N.; Ramya, R. The Real Time Monitoring of Water Quality in IoT Environment. In Proceedings of the 2015 IEEE International Conference on Innovations in Information Technologies (ICCPCT), Embedded and Communication Systems, Coimbatore, India, 19–20 March 2015; pp. 16–19.
2. Syafrudin, M.; Alfian, G.; Fitriyani, N.L.; Rhee, J. Performance Analysis of IoT-Based Sensor, Big Data Processing, and Machine Learning Model for Real-Time Monitoring System in Automotive Manufacturing. *Sensors (Switzerland)* **2018**, *18*, 2946. [[CrossRef](#)]
3. Park, J.; Kim, K.T.; Lee, W.H. Recent Advances in Information and Communications Technology (ICT) and Sensor Technology for Monitoring Water Quality. *Water (Switzerland)* **2020**, *12*, 510. [[CrossRef](#)]
4. Saboe, D.; Ghasemi, H.; Gao, M.M.; Samardzic, M.; Hristovski, K.D.; Boscovic, D.; Burge, S.R.; Burge, R.G.; Hoffman, D.A. Real-Time Monitoring and Prediction of Water Quality Parameters and Algae Concentrations Using Microbial Potentiometric Sensor Signals and Machine Learning Tools. *Sci. Total Environ.* **2020**, *764*, 142876. [[CrossRef](#)]
5. Gholami, R.; Kamkar-Rouhani, A.; Doulati Ardejani, F.; Maleki, S. Prediction of Toxic Metals Concentration Using Artificial Intelligence Techniques. *Appl. Water Sci.* **2011**, *1*, 125–134. [[CrossRef](#)]
6. Ahmed, U.; Mumtaz, R.; Anwar, H.; Shah, A.A.; Irfan, R. Efficient Water Quality Prediction Using Supervised Machine Learning. *Water* **2019**, *11*, 2210. [[CrossRef](#)]
7. Cho, K.H.; Sthiannopkao, S.; Pachepsky, Y.A.; Kim, K.W.; Kim, J.H. Prediction of Contamination Potential of Groundwater Arsenic in Cambodia, Laos, and Thailand Using Artificial Neural Network. *Water Res.* **2011**, *45*, 5535–5544. [[CrossRef](#)]
8. Shafi, U.; Mumtaz, R.; Anwar, H.; Qamar, A.M.; Khurshid, H. Surface Water Pollution Detection Using Internet of Things. In Proceedings of the International Conference on Smart Cities: Improving Quality of Life Using ICT and IoT, HONET-ICT 2018, Islamabad, Pakistan, 8–10 October 2018; pp. 92–96.
9. Dunnington, D.W.; Trueman, B.F.; Raseman, W.J.; Anderson, L.E.; Gagnon, G.A. Comparing the Predictive Performance, Interpretability, and Accessibility of Machine Learning and Physically Based Models for Water Treatment. *ACS ES&T Eng.* **2021**, *1*, 348–356.
10. Lubke, G.H.; Campbell, I.; McArtor, D.; Miller, P.; Lunningham, J.; Berg, S.M. van den Assessing Model Selection Uncertainty Using a Bootstrap Approach: An Update. *Struct Equ Model.* **2017**, *24*, 230–245. [[CrossRef](#)] [[PubMed](#)]
11. Begoli, E.; Bhattacharya, T.; Kusnezov, D. The Need for Uncertainty Quantification in Machine-Assisted Medical Decision Making. *Nat. Mach. Intell.* **2019**, *1*, 20–23. [[CrossRef](#)]
12. Lu, Y.; Tang, C.; Chen, J.; Yao, H. Assessment of Major Ions and Heavy Metals in Groundwater: A Case Study from Guangzhou and Zhuhai of the Pearl River Delta, China. *Front. Earth Sci.* **2016**, *10*, 340–351. [[CrossRef](#)]
13. Wen, X.; Lu, J.; Wu, J.; Lin, Y.; Luo, Y. Influence of Coastal Groundwater Salinization on the Distribution and Risks of Heavy Metals. *Sci. Total Environ.* **2019**, *652*, 267–277. [[CrossRef](#)] [[PubMed](#)]
14. Yu, K.; Li, J.; Li, H.; Chen, K.; Lv, B.; Zhao, L. Statistical Characteristics of Heavy Metals Content in Groundwater and Their Interrelationships in a Certain Antimony Mine Area. *J. Groundw. Sci. Eng.* **2016**, *4*, 284–292.
15. Sun, L.; Peng, W.; Cheng, C. Source Estimating of Heavy Metals in Shallow Groundwater Based on UNMIX Model: A Case Study. *Indian J. Geo-Marine Sci.* **2016**, *45*, 756–762.
16. Lou, S.; Liu, S.; Dai, C.; Tao, A.; Tan, B.; Ma, G.; Chalov, R.S.; Chalov, S.R. Heavy Metal Distribution and Groundwater Quality Assessment for a Coastal Area on a Chinese Island. *Polish J. Environ. Stud.* **2017**, *26*, 733–745. [[CrossRef](#)]
17. Kanagaraj, G.; Sridhar, S.G.D.; Sangunathan, U.; Mohamed Rafik, M.; Balasubramanian, M.; Sakthivel, A.M.; Jenefer, S. Heavy Metal Concentration in Groundwater from Besant Nagar to Sathankuppam, South Chennai, Tamil Nadu, India. *Appl. Water Sci.* **2017**, *7*, 4651–4662.
18. Tjoa, E.; Guan, C. A Survey on Explainable Artificial Intelligence (XAI): Towards Medical XAI. *IEEE Trans. Neural Networks Learn. Syst.* **2020**, *32*, 4793–4813. [[CrossRef](#)]
19. Barredo Arrieta, A.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; Garcia, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *Inf. Fusion* **2020**, *58*, 82–115. [[CrossRef](#)]
20. Anguita-Ruiz, A.; Segura-Delgado, A.; Alcalá, R.; Aguilera, C.M.; Alcalá-Fdez, J. Explainable Artificial Intelligence (XAI) for the Identification of Biologically Relevant Gene Expression Patterns in Longitudinal Human Studies, Insights from Obesity Research. *PLoS Comput. Biol.* **2020**, *16*, e1007792. [[CrossRef](#)]
21. Zou, R.; Lung, W.-S.; Guo, H. Neural Network Embedded Monte Carlo Approach for Water Quality Modeling under Input Information Uncertainty. *J. Comput. Civ. Eng.* **2002**, *16*, 135–142. [[CrossRef](#)]
22. Knoll, L.; Breuer, L.; Bach, M. Nation-Wide Estimation of Groundwater Redox Conditions and Nitrate Concentrations through Machine Learning. *Environ. Res. Lett.* **2020**, *15*, 064004. [[CrossRef](#)]
23. Coulston, J.W.; Blinn, C.E.; Thomas, V.A.; Wynne, R.H. Approximating Prediction Uncertainty for Random Forest Regression Models. *Photogramm. Eng. Remote Sensing* **2016**, *82*, 189–197. [[CrossRef](#)]
24. Lee, I.H.; Ni, C.F.; Lin, F.P.; Lin, C.P.; Ke, C.C. Stochastic Modeling of Flow and Conservative Transport in Three-Dimensional Discrete Fracture Networks. *Hydrol. Earth Syst. Sci.* **2019**, *23*, 19–34. [[CrossRef](#)]
25. Ni, C.F.; Li, S.G.; Liu, C.J.; Hsu, S.M. Efficient Conceptual Framework to Quantify Flow Uncertainty in Large-Scale, Highly Nonstationary Groundwater Systems. *J. Hydrol.* **2010**, *381*, 297–307. [[CrossRef](#)]

26. Wong, E.; Kolter, J.Z. Learning Perturbation Sets for Robust Machine Learning. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual, 3–7 May 2021.
27. Jeddi, A.; Shafiee, M.J.; Karg, M.; Scharfenberger, C.; Wong, A. Learn2Perturb: An End-to-End Feature Perturbation Learning to Improve Adversarial Robustness. In Proceedings of the Computer Vision and Pattern Recognition; IEEE Computer Society: Washington, DC, USA, 2020.
28. Kaspchak, B.; Meißner, U.G. Neural Network Perturbation Theory and Its Application to the Born Series. *Phys. Rev. Res.* **2021**, *3*, 023223. [[CrossRef](#)]
29. Zhang, X.; Liang, F.; Srinivasan, R.; Van Liew, M. Estimating Uncertainty of Streamflow Simulation Using Bayesian Neural Networks. *Water Resour. Res.* **2009**, *45*, W2403. [[CrossRef](#)]
30. Chandra, R.; Azam, D.; Müller, R.D.; Salles, T.; Cripps, S. Bayeslands: A Bayesian Inference Approach for Parameter Uncertainty Quantification in Badlands. *Comput. Geosci.* **2019**, *131*, 89–101. [[CrossRef](#)]
31. McDermott, P.L.; Wikle, C.K. Bayesian Recurrent Neural Network Models for Forecasting and Quantifying Uncertainty in Spatial-Temporal Data. *Entropy* **2019**, *21*, 184. [[CrossRef](#)]
32. Tiwari, M.K.; Chatterjee, C. Uncertainty Assessment and Ensemble Flood Forecasting Using Bootstrap Based Artificial Neural Networks (BANNs). *J. Hydrol.* **2010**, *382*, 20–33. [[CrossRef](#)]
33. Chen, S.; Onnela, J.P. A Bootstrap Method for Goodness of Fit and Model Selection with a Single Observed Network. *Sci. Rep.* **2019**, *9*, 16674. [[CrossRef](#)]
34. Mentch, L.; Hooker, G. Quantifying Uncertainty in Random Forests via Confidence Intervals and Hypothesis Tests. *J. Mach. Learn. Res.* **2016**, *17*, 441–481.
35. Willcock, S.; Martínez-López, J.; Hooftman, D.A.P.; Bagstad, K.J.; Balbi, S.; Marzo, A.; Prato, C.; Sciandrello, S.; Signorello, G.; Voigt, B.; et al. Machine Learning for Ecosystem Services. *Ecosyst. Serv.* **2018**, *33*, 165–174. [[CrossRef](#)]
36. Barton, R.R.; Nelson, B.L.; Xie, W. Quantifying Input Uncertainty via Simulation Confidence Intervals. *INFORMS J. Comput.* **2014**, *26*, 74–87. [[CrossRef](#)]
37. Musil, F.; Willatt, M.J.; Langovoy, M.A.; Ceriotti, M. Fast and Accurate Uncertainty Estimation in Chemical Machine Learning. *J. Chem. Theory Comput.* **2019**, *15*, 906–915. [[CrossRef](#)] [[PubMed](#)]
38. Adadi, A.; Berrada, M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* **2018**, *6*, 52138–52160. [[CrossRef](#)]
39. Su, Y.S.; Wu, S.Y. Applying Data Mining Techniques to Explore User Behaviors and Watching Video Patterns in Converged IT Environments. *J. Ambient Intell. Humaniz. Comput.* **2021**. [[CrossRef](#)] [[PubMed](#)]
40. Su, Y.S.; Chou, C.H.; Chu, Y.L.; Yang, Z.Y. A Finger-Worn Device for Exploring Chinese Printed Text with Using CNN Algorithm on a Micro IoT Processor. *IEEE Access* **2019**, *7*, 116529–116541. [[CrossRef](#)]
41. Su, Y.S.; Ding, T.J.; Chen, M.Y. Deep Learning Methods in Internet of Medical Things for Valvular Heart Disease Screening System. *IEEE Internet Things J.* **2021**, *8*, 16921–16932. [[CrossRef](#)]
42. Neto, M.P.; Paulovich, F.V. Explainable Matrix-Visualization for Global and Local Interpretability of Random Forest Classification Ensembles. *IEEE Trans. Vis. Comput. Graph.* **2021**, *27*, 1427–1437. [[CrossRef](#)]
43. Altmann, A.; Tološi, L.; Sander, O.; Lengauer, T. Permutation Importance: A Corrected Feature Importance Measure. *Bioinformatics* **2010**, *26*, 1340–1347. [[CrossRef](#)]
44. Galkin, F.; Aliper, A.; Putin, E.; Kuznetsov, I.; Gladyshev, V.N.; Zhavoronkov, A. Human Microbiome Aging Clocks Based on Deep Learning and Tandem of Permutation Feature Importance and Accumulated Local Effects. *bioRxiv* **2018**.
45. Huang, N.; Lu, G.; Xu, D. A Permutation Importance-Based Feature Selection Method for Short-Term Electricity Load Forecasting Using Random Forest. *Energies* **2016**, *9*, 767. [[CrossRef](#)]
46. Yajima, H.; Derot, J. Application of the Random Forest Model for Chlorophyll-a Forecasts in Fresh and Brackish Water Bodies in Japan, Using Multivariate Long-Term Databases. *J. Hydroinformatics* **2018**, *20*, 191–205. [[CrossRef](#)]
47. Petkovic, D.; Altman, R.; Wong, M.; Vigil, A. Improving the Explainability of Random Forest Classifier – User Centered Approach. *HHS Public Access* **2018**, *23*, 204–215.
48. Elshawi, R.; Al-Mallah, M.H.; Sakr, S. On the Interpretability of Machine Learning-Based Model for Predicting Hypertension. *BMC Med. Inform. Decis. Mak.* **2019**, *19*, 146.
49. Ryo, M.; Angelov, B.; Mammola, S.; Kass, J.M.; Benito, B.M.; Hartig, F. Explainable Artificial Intelligence Enhances the Ecological Interpretability of Black-Box Species Distribution Models. *Ecography* **2021**, *44*, 199–205.
50. Lundberg, S.M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J.M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.I. Explainable AI for Trees: From Local Explanations to Global Understanding. *arXiv* **2019**, arXiv:1095.0461v1. [[CrossRef](#)]
51. Hall, P. On the Art and Science of Explainable Machine Learning: Techniques, Recommendations, and Responsibilities. In Proceedings of the KDD'19 XAI Workshop, Anchorage, AK, USA, 4–8 August 2019.
52. Jalali, A.; Schindler, A.; Haslhofer, B.; Rauber, A. Machine Learning Interpretability Techniques for Outage Prediction: A Comparative Study. In Proceedings of the European Conference on the Prognostics and Health Management Society, Turin, Italy, 1–3 July 2020; pp. 1–10.
53. Saabas, A. Treeinterpreter. Available online: <https://github.com/andosa/treeinterpreter> (accessed on 15 April 2020).

54. Sharma, P.; Mirzan, S.R.; Bhandari, A.; Pimpley, A.; AbhiramEswaran; Srinivasan, S.; Shao, L. Evaluating Tree Explanation Methods for Anomaly Reasoning: A Case Study of SHAP TreeExplainer and TreeInterpreter. In Proceedings of the Advances in Conceptual Modeling, Vienna, Austria, 3–6 November 2020; Grossmann, G., Ram, S., Eds.; Springer Nature Switzerland: Vienna, Austria, 2020.
55. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
56. Hunter, J.D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **2007**, *9*, 90–95. [[CrossRef](#)]
57. McKinney, W. Data Structures for Statistical Computing in Python. In Proceedings of the Proceedings of the 9th Python in Science Conference, Austin, TX, USA, 28 June–3 July 2010; Volume 445, pp. 56–61.
58. Deb, S. A Novel Robust R-Squared Measure and Its Applications in Linear Regression. *Adv. Intell. Syst. Comput.* **2017**, *532*, 131–142.
59. Chai, T.; Draxler, R.R. Root Mean Square Error (RMSE) or Mean Absolute Error (MAE)? -Arguments against Avoiding RMSE in the Literature. *Geosci. Model Dev.* **2014**, *7*, 1247–1250. [[CrossRef](#)]
60. Mazloumi, E.; Rose, G.; Currie, G.; Moridpour, S. Prediction Intervals to Account for Uncertainties in Neural Network Predictions: Methodology and Application in Bus Travel Time Prediction. *Eng. Appl. Artif. Intell.* **2011**, *24*, 534–542.
61. Seifi, A.; Ehteram, M.; Singh, V.P.; Mosavi, A. Modeling and Uncertainty Analysis of Groundwater Level Using Six Evolutionary Optimization Algorithms Hybridized with ANFIS, SVM, and ANN. *Sustain.* **2020**, *12*, 4023. [[CrossRef](#)]
62. Fox, E.W.; Ver Hoef, J.M.; Olsen, A.R. Comparing Spatial Regression to Random Forests for Large Environmental Data Sets. *PLoS One* **2020**, *15*, e0229509. [[CrossRef](#)] [[PubMed](#)]
63. Chang, F.J.; Huang, C.W.; Cheng, S.T.; Chang, L.C. Conservation of Groundwater from Over-Exploitation—Scientific Analyses for Groundwater Resources Management. *Sci. Total Environ.* **2017**, *598*, 828–838. [[CrossRef](#)]
64. EPA. *Environmental Water Quality Monitoring Annual Report*; Environment Protection Administration: Taipei City, Taiwan, 2020.
65. EPA Environmental Protection Administration. Available online: <https://ewq.epa.gov.tw/Code/?Languages=tw> (accessed on 13 April 2020).
66. Zhang, Z.; Xiao, C.; Adeyeye, O.; Yang, W.; Liang, X. Source and Mobilization Mechanism of Iron, Manganese and Arsenic in Groundwater of Shuangliao City, Northeast China. *Water (Switzerland)* **2020**, *12*, 534. [[CrossRef](#)]
67. Mahbooba, B.; Timilsina, M.; Sahal, R.; Serrano, M. Explainable Artificial Intelligence (XAI) to Enhance Trust Management in Intrusion Detection Systems Using Decision Tree Model. *Complexity* **2021**, 6634811. [[CrossRef](#)]
68. Ibrahim, N.I.M. Majmaah The Relations Between Concentration of Iron and the PH Ground Water (Case Study Zulfi Ground Water). *Int. J. Environ. Monit. Anal.* **2016**, *4*, 140.
69. Klingel, F. Potential of In-Situ Groundwater Treatment for Iron, Manganese and Arsenic Removal In. In Proceedings of the Proceeding of The 4th International Symposium Vietnam Water Cooperation Initiative for Water Security in a Changing Era, Hanoi, Vietnam, 19 October 2015.
70. Rajakovic, J.; Rajakovic Ognjanovic, V. Arsenic in Water: Determination and Removal Chapter. In *Arsenic-Analytical and Toxicological Studies Figure*; IntechOpen: London, UK, 2018.
71. Groschen, G.E.; Arnold, T.L.; Morrow, W.S.; Warner, K.L. *Occurrence and Distribution of Iron, Manganese, and Selected Trace Elements in Ground Water in the Glacial Aquifer System of the Northern United States*; USGS: Reston, VA, USA, 2009.
72. Molnar, C. *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*; Leanpub: Victoria, BC, Canada, 2019; p. 247.
73. Bergstra, J.; Bengio, Y. Random Search for Hyper-Parameter Optimization. *J. Mach. Learn. Res.* **2012**, *13*, 281–305.
74. Andradóttir, S. A Review of Random Search Methods. In *Handbook of Simulation Optimization*; Fu, M.C., Ed.; Springer Science+Business Media: New York, NY, USA, 2015; pp. 277–292.
75. Solomatine, D.P.; Shrestha, D.L. A Novel Method to Estimate Model Uncertainty Using Machine Learning Techniques. *Water Resour. Res.* **2009**, *45*, WR006839. [[CrossRef](#)]
76. Koutsoyiannis, D. Climate Change, the Hurst Phenomenon, and Hydrological Statistics. *Hydrol. Sci. J.* **2003**, *48*, 3–24. [[CrossRef](#)]
77. Dwivedi, D.; Mohanty, B.P. Hot Spots and Persistence of Nitrate in Aquifers across Scales. *Entropy* **2016**, *18*, 25. [[CrossRef](#)]
78. Lu, C.; Song, Z.; Wang, W.; Zhang, Y.; Si, H.; Liu, B.; Shu, L. Spatiotemporal Variation and Long-Range Correlation of Groundwater Depth in the Northeast China Plain and North China Plain from 2000~2019. *J. Hydrol. Reg. Stud.* **2021**, *37*, 100888. [[CrossRef](#)]
79. Vu, T.D.; Ni, C.F.; Li, W.C.; Truong, M.H. Modified Index-Overlay Method to Assess Spatial-Temporal Variations of Groundwater Vulnerability and Groundwater Contamination Risk in Areas with Variable Activities of Agriculture Developments. *Water (Switzerland)* **2019**, *11*, 2492. [[CrossRef](#)]
80. Vu, T.D.; Ni, C.F.; Li, W.C.; Truong, M.H.; Hsu, S.M. Predictions of Groundwater Vulnerability and Sustainability by an Integrated Index-Overlay Method and Physical-Based Numerical Model. *J. Hydrol.* **2021**, *596*, 126082. [[CrossRef](#)]
81. Machado, D.F.T.; Silva, S.H.G.; Curi, N.; Menezes, M.D. De Soil Type Spatial Prediction from Random Forest: Different Training Datasets, Transferability, Accuracy and Uncertainty Assessment. *Soil Plant Nutr.* **2019**, *76*, 243–254.
82. Dimitriadis, P.; Koutsoyiannis, D.; Iliopoulou, T.; Papanicolaou, P. A Global-Scale Investigation of Stochastic Similarities in Marginal Distribution and Dependence Structure of Key Hydrological-Cycle Processes. *Hydrology* **2021**, *8*, 59. [[CrossRef](#)]

83. Wang, M.; Fu, W.; He, X.; Hao, S.; Wu, X. A Survey on Large-Scale Machine Learning. *IEEE Trans. Knowl. Data Eng.* **2020**, *34*, 2574–2594. [[CrossRef](#)]
84. Su, Y.S.; Ni, C.F.; Li, W.C.; Lee, I.H.; Lin, C.P. Applying Deep Learning Algorithms to Enhance Simulations of Large-Scale Groundwater Flow in IoTs. *Appl. Soft Comput. J.* **2020**, *92*, 106298. [[CrossRef](#)]