BILINGUAL PUBLISHING GROUP

Research in Ecology
https://journals.bilpubgroup.com/index.php/re

ARTICLE

Sustainable Groundwater Management in Water-Scarce Regions: A Spatial Machine Learning Analysis from Rajshahi, Bangladesh

Sumaya Tabassum [1], Likhon Chandra Roy [1*], Amit Kumar Sarkar [2], Yassine Ezaier [3], Hader Ahmed [3], Lghazi Youssef [3], Hesam Kamyab [4,5], Hussameldin Ibrahim [6], Mohammad Yusuf [6,7*]

[1] Department of Civil Engineering, Rajshahi University of Engineering & Technology, Rajshahi 6204, Bangladesh

[2] Department of Public Health Engineering, Government of the People's Republic of Bangladesh, Chapainawabganj 6300, Bangladesh

[3] Bio-Geosciences and Materials Engineering Laboratory, Ecole Normale Supérieure, University Hassan II, Casablanca 20100, Morocco

[4] Department of Biomaterials, Saveetha Dental College and Hospital, Saveetha Institute of Medical and Technical Sciences, Chennai 600077, India

[5] The KU-KIST Graduate School of Energy and Environment, Korea University, Seoul 02841, Republic of Korea

[6] Clean Energy Technologies Research Institute (CETRI), Faculty of Engineering and Applied Science, University of Regina, Regina, SK S4S 0A2, Canada

[7] Architecture Department, Faculty of Architecture and Urbanism, UTE University, Quito 170527, Ecuador

ABSTRACT

Ensuring the availability and sustainable management of water (SDG 6) is particularly challenging in dry regions like Rajshahi, Bangladesh, where communities rely heavily on groundwater with limited recharge potential. Issues such as declining water levels and contamination by iron, arsenic, and chloride compromise both user satisfaction and public health. This study aimed to assess groundwater quality risks through regional mapping to guide the installation depth of new water sources. In collaboration with the Department of Public Health Engineering (DPHE), data were

*CORRESPONDING AUTHOR:
Likhon Chandra Roy, Department of Civil Engineering, Rajshahi University of Engineering & Technology, Rajshahi 6204, Bangladesh; Email: lkhnroy@gmail.com; Mohammad Yusuf, Clean Energy Technologies Research Institute (CETRI), Faculty of Engineering and Applied Science, University of Regina, Regina, SK S4S 0A2, Canada; Architecture Department, Faculty of Architecture and Urbanism, UTE University, Quito 170527, Ecuador; Email: mohd.yusuf@uregina.ca

ARTICLE INFO
Received: 11 June 2025 | Revised: 23 June 2025 | Accepted: 17 July 2025 | Published Online: 12 August 2025
DOI: https://doi.org/10.30564/re.v7i3.10453

CITATION
Tabassum, S., Roy, L.C., Sarkar, A.K., et al., 2025. Sustainable Groundwater Management in Water-Scarce Regions: A Spatial Machine Learning Analysis from Rajshahi, Bangladesh. Research in Ecology. 7(3): 268–286. DOI: https://doi.org/10.30564/re.v7i3.10453

collected from 7,388 tube wells across nine upazilas, including well depth, geographic coordinates, and contaminant concentrations. Water quality was evaluated against World Health Organization and Bangladesh standards. Machine learning (XGBoost) and spatial analysis were applied to model contaminant levels based on location and well depth. An initial model showed poor performance, but after identifying and correcting key errors, the refined model yielded significant improvements: R^2 increased from 0.0345 to 0.62 for iron, from −0.0015 to 0.38 for arsenic, and from 0.12 to 0.71 for chloride. A comprehensive water quality risk map was developed by integrating these results at the upazila level. This map provides actionable insights for government agencies and NGOs to prioritize areas for water quality testing, remediation, and public awareness initiatives, contributing to more informed and sustainable water resource management in the region.

Keywords: Arsenic Contamination; SDG 6; Iron Contamination; Health Risk; Groundwater Accessibility

1. Introduction

Water is vital for all living things, including humans, animals, plants, and other organisms. The quantity and quality of water sustain ecological equilibrium, which impacts human lifestyles [1–5]. Groundwater provides approximately 26% of the world's renewable freshwater supply. It supplies water for residential, commercial, industrial, agricultural, and other development projects [6–10]. In general, organic contaminants may be less prevalent in groundwater than in surface water bodies like lakes, ponds, and rivers. This is because contaminants are reduced as groundwater naturally filters itself as it seeps through rocks and soil. Therefore, compared to surface water treatment, groundwater treatment is frequently simpler and involves fewer steps [11–15]. However, groundwater contamination is a serious issue when it is the principal source of drinking and irrigation for people.

The use of groundwater resources has increased tremendously since it is relatively easy to access and appears to be purer compared to surface sources of water. Such reliance on groundwater highlights the necessity to conserve these resources against pollution and overuse lest the water supply be compromised for future generations. Up to 50% of the world's population is predicted to experience permanent or intermittent water insecurity by 2050 due to pressure on freshwater supplies brought on by population growth, agricultural intensification and expansion, urbanization, industrialization, and climate change [16–22]. Bangladesh has a huge population, and the rate of population growth is high. The country is renowned for its water resources, which include groundwater and surface water. It is an incredibly fertile land. Before the invention of tube wells, humans relied on man-made water reservoirs such as ponds and drilled wells as well as the natural surface water found in rivers, canals, and lakes [23–25]. The country struggles with water and sanitation problems regardless of whether it is a remote rural area or the capital city. High concentrations of naturally occurring contaminants, such as iron and arsenic are present in Bangladeshi groundwater and can be harmful if consumed. In Bangladesh, elevated levels of arsenic in groundwater combined with other pollutants degrade the quality of the water, rendering it unsafe for human use [26,27]. Groundwater intoxication by arsenic in Bangladesh is a serious public health concern. The dynamic nature of arsenic contamination alongside other harmful elements must thus be understood through routine monitoring and ongoing evaluation of hydrochemical properties [28–32]. Such complexity in contamination patterns demands an elaborate strategy of reviewing and mitigating the risks to make the water safe to drink and to be used in farming activities.

This study investigates the Rajshahi district, located in northwestern Bangladesh, which functions as the administrative headquarters of the Rajshahi division and encompasses a city corporation. The district exhibits a dense population, currently experiencing a growth rate of 2.26%, as reported by the World Population Review. For a large percentage of the people in this area, groundwater serves as their primary source of drinking water. The issue is more severe because many rural areas lack sources of safe drinking water, which makes life more difficult for the local inhabitants. It is imperative to understand the spatial distribution of water quality risks within the Rajshahi district. This understanding is essential not only for implementing immediate public health interventions but also for formulating effective adaptation strategies for the future. These challenges require us to analyze local hydrogeologi-

cal conditions and socio-economic factors affecting access to and quality of water in the area in detail. The problem of arsenic (As) and iron (Fe) contaminated groundwater is quite severe in different parts of the globe, particularly in Bangladesh, as the nature of the groundwater varies geologically and biogeochemically. The study also uses the theoretical basis that a holistic understanding of arsenic in groundwater necessitates the need to consider the geochemical connection of arsenic and iron in influencing the spatial distribution of concentration of arsenic. Such an interaction is essential in determining the distribution and behavior of such contaminants, as there are complex hydrochemical processes that control how they interact, and they differ with the aquifer system involved.

The greater the iron content, the better it favors the release of arsenic, and this is the correlation between arsenic and iron. The reason is that arsenic mobilization can only be achieved with the disaggregation of iron oxyhydroxides, which act as the principal sink of arsenic in aquifers. As the reduction of iron oxyhydroxides is a key factor, causal links are mainly in one direction, that is, the less the dissolution of iron, the greater the release of arsenic. These effects of mobilization are further adjusted by considerations of pH, redox potential, and the availability of competitive ions, which may change the stability of iron oxyhydroxides, leading to the release of arsenic. These dynamics are necessary to have a clear idea of what places will be more susceptible to contamination, as well as developing focused monitoring strategies.

Economically, the expenses of water treatment efforts and civil health programs increase where the contamination of arsenic and iron is combined. Because arsenic is so health risky due to its deadly health effects like cancer as well as arsenicosis, rather expensive mitigation strategies such as reverse osmosis or substitution with alternative water sources, are needed. Although less harmful, iron stains infrastructure and disappoints the water taste, which leads to user dissatisfaction and increased cost in maintaining these systems. Through mapping the spatial variability of these pollutants, the study provides communities and local governments with the basis for informed economic decisions regarding the placement of wells and treatment of wells. This information can also be used to plan intervention or management resources and allocate them to areas with high contaminant levels, taking into consideration matters like installing treatment systems or drilling new wells in less contaminated areas. This practice not only lowers health hazards, but it also lowers the economic cost to local communities.

Moreover, arsenic and iron contamination occur spatially in a way that emphasizes local management solutions. The heterogeneous distribution of the contaminants in the Rajshahi district requires a specific approach that considers both the geological and socio-economic conditions of a particular area, because groundwater is one of the main sources of drinking water in this region. As an example, a high degree of reductive dissolution may occur in regions with high organic-matter sediments that produce high arsenic and iron concentrations. To overcome these challenges, it is necessary to combine more sophisticated analysis tools in the form of geospatial modeling and machine learning to forecast the contamination hotspots and influence policy and decision-making. Also, community sensitization programs are important in creating awareness in communities about the dangers of taking polluted water and the need to have the water status tested regularly. These, coupled with sustainable practice of managing water sources, can help make the water supply network more resistant to the impacts of arsenic and iron contamination and ensure that the health of people in the region remains unaffected and that the environment will be more sustainable in the long run.

One of the targets of SDG 6 is universal access to safe drinking water and the improvement of water quality. Access to safely managed drinking water increased from 70% to 74% worldwide between 2015 and 2020, primarily due to advancements in Central, East, and Southern Asia. In 2020, however, two billion people still lacked access to safely managed water, including 771 million without even the most basic utilities, with Sub-Saharan Africa housing 387 million of these individuals [33]. This research is designed to investigate the spatial distribution of water quality risks in the Rajshahi region through a comprehensive and methodologically diverse approach. Rajshahi is located in the Barind region, which faces significant groundwater challenges. There are noticeable drops in groundwater levels in the central Barind region. The yearly average decline is between 0.1 and 0.6 meters, the dry season is between 0.2 and 0.82 meters, and the monsoon season is between 0.2 and 0.67 meters [34]. Such losses increase the susceptibility of the water in the region, and it is important to devise ways in which water is extracted and replaced to

ensure that the water table is not reduced further.

The concurrence of iron and arsenic pollution in the groundwater is also often due to their closely allied geochemical pathways, which are predominantly associated with the reductive dissolution processes that occur under anaerobic aquifer conditions. The arsenic-binding iron oxyhydroxides are present in the sedimentary aquifers of the Bengal Delta, particularly in Rajshahi. The dissolution of such iron oxyhydroxides liberates iron (as Fe^{2+}), arsenic (as As^{3+} or As^{5+}) into the groundwater in reducing conditions, and this reduction often occurs due to microbial activity. Drainage worsens because of the breakdown of organic matter, which consumes oxygen and promotes an anaerobic environment. As such, a positive relationship exists between iron concentrations and the release of arsenic, as high levels of iron are dissolved and arsenic is released. Local hydrogeological factors include the depth of the groundwater table, sediment contents, and groundwater flow patterns that may affect and enhance or reduce the contamination points of an area.

In this study an extensive dataset is developed that encompasses 7388 tube wells across all nine upazilas in Rajshahi: Paba, Bagha, Bagmara, Durgapur, Godagari, Charghat, Mohanpur, Tanore, and Puthia. This dataset, collected in partnership with the Department of Public Health Engineering (DPHE), includes critical information such as precise geographic coordinates (latitude and longitude), well depth, and concentrations of key contaminants, including iron, arsenic, and chloride ions. Through the use of spatial analysis techniques, the possible connections between these environmental elements and water quality, offering crucial information for focused public health initiatives, are explored. Machine learning techniques were also used such as the XGBoost model. ANOVA, Regression Analysis, and Pearson Correlation Coefficient have been employed to evaluate model performances. A heat map of the depth of water strata and contaminants, a time series data plot showing the decline of water strata over years, scatter plots of the frequency of different contaminants, and identifying and counting the percentage of unsafe water sources were studied for the overall understanding. The combination of these analytical approaches allows the study to offer a solid structure for pinpointing areas at risk and selecting interventions as priorities to give safe access to water to the people of Rajshahi.

2. Methodology

2.1. Study Area

All the upazilas were studied for a better understanding of Rajshahi groundwater quality (**Figure 1**). The concentrations of iron, arsenic, and chloride present in the groundwater during the installation of shallow tube wells were studied. A total number of 7388 water pumps were studied in Paba, Bagha, Bagmara, Durgapur, Godagari, Charghat, Mohanpur, Tanore, and Puthia Upazila.

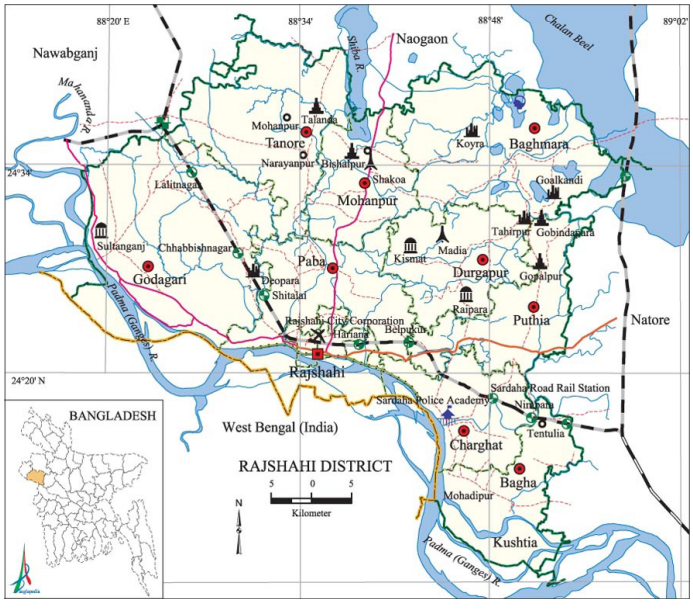

Figure 1. A detailed administrative map of the study area [35].

2.2. Data Collection

The Department of Public Health Engineering (DPHE) of the Bangladesh government collects water samples from newly installed tube wells and records relevant data on the water source within 30 days of operation during the period from 07/06/2020 to 24/12/2023. These samples are typically tested within 12 hours at the Rajshahi zonal laboratory of the DPHE, following the standard procedure prescribed in the country. Subsequently, the test results help the organization evaluate the need for implementing reverse osmosis to ensure the provision of safe water. For this particular study, data from 07/06/2020 to 24/12/2023 in the Rajshahi district were gathered and compiled. The extensive dataset was then organized based on locations, and scatter plots were utilized to illustrate the variation and typical depth of groundwater sources.

2.3. Data Analysis

Following data analysis, the contamination of toxic arsenic, iron, and chloride in relation to depth was demonstrated on a regional basis. Subsequently, the number of tube wells and their percentages were calculated, taking into account the safe limits of arsenic, iron, and chloride in drinking water. The flow diagram of the methodology is shown in **Figure** 2.

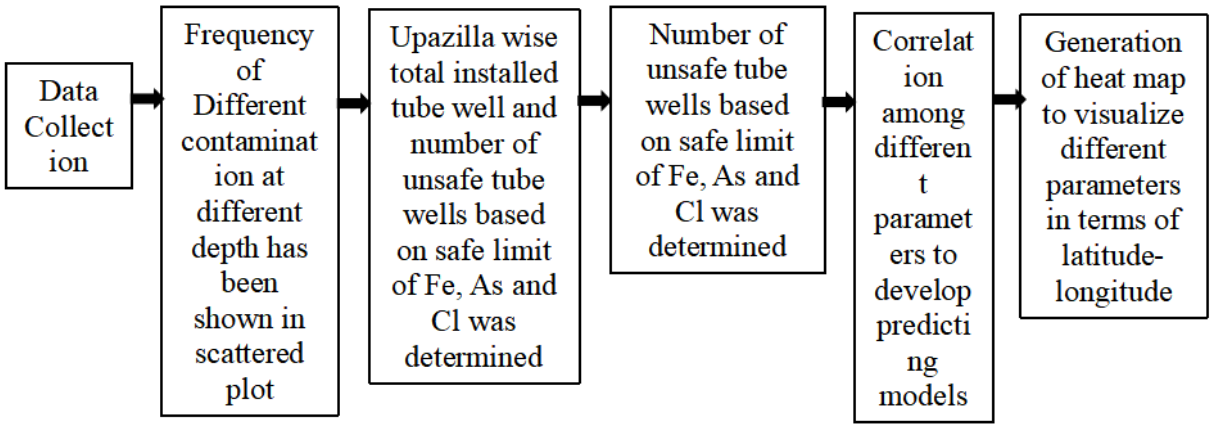

Figure 2. Flow diagram of methodology.

Rigorous field checks and laboratory quality control processes were used to maintain data quality. The data were analyzed using statistical techniques such as correlation analysis and descriptive statistics. The quantitative analysis yielded insightful information on the correlation between pollutant levels and groundwater depth. However, correlations among parameters were aimed to be determined and heat maps were generated to visualize the parameters in terms of locations.

3. Model Results

From **Table 1**, the tubewell depth data indicates symmetry, but a large range and a high maximum value suggest some outliers. Iron concentrations are positively skewed with a notable difference between the mean and median, revealing variability and health risks from high levels. Arsenic levels also show skewness, with a mean exceeding the median and concerning outliers. Chloride concentrations are highly variable, with a right-skewed distribution indicating possible contamination risks.

From **Table 2, Table 3,** and **Table 4,** ANOVA analysis shows various groundwater contaminations. The influence of time on iron content indicates that environmental changes and human activities over time have affected its concentration. No significant time effect on arsenic suggests its levels are influenced more by consistent factors like geological formations and local contamination rather than temporal changes. The considerable time effect on chloride indicates that temporal factors mostly impact its levels. In summary, these findings prioritize the importance of both temporal and non-temporal factors in evaluating groundwater quality and securing safe drinking water.

Table 1. Statistical Analysis of depth and different contaminants.

	Depth (m)	Iron (mg/L)	Arsenic (mg/L)	Chloride (mg/L)
Mean	41.03	0.56	0.01	30.00
Median	41.15	0.10	0.00	22.00
Standard Deviation	6.301311	1.293133	0.140497	24.09026
Minimum	21.94	0.00	0.00	0.00
Maximum	59.00	25.00	7.90	340.00
Range	37.06	25.00	7.90	340.00

Table 2. ANOVA Table for Iron Concentration.

Source	Sum of Squares (sum_sq)	df	F	PR(>F)
C(Year)	87.044864	5.0	10.473618	4.923010e-10
Residual	12266.842670	7380.0	NaN	NaN

Table 3. ANOVA Table for Arsenic Concentration.

Source	Sum of Squares (sum_sq)	df	F	PR(>F)
C(Year)	0.045010	5.0	0.455686	0.809413
Residual	145.789464	7380.0	NaN	NaN

Table 4. ANOVA Table for Chlorine Concentration.

Source	Sum of Squares (sum_sq)	df	F	PR(>F)
C(Year)	4.482641×10^{4}	5.0	15.597502	2.672399×10^{-15}
Residual	4.241947×10^{6}	7380.0	NaN	NaN

Significant temporal variation in iron and chloride concentrations ($p < 0.001$) is indicated by ANOVA results (**Tables 2–4**), which imply that environmental changes and human activities (such as intensifying agriculture) affect these contaminants over time. In line with its geological background, where stable aquifer conditions predominate over temporal changes, arsenic has no discernible temporal influence ($p = 0.809$). This confirms the theoretical prediction that geochemical processes, not seasonal or yearly variations, are the main drivers of arsenic.

However, ANOVA, Regression Analysis, and Pearson's Correlation Coefficient have been employed to evaluate model performance. OLS Regression results reveal a slight positive relationship between time and iron concentration, with depth showing a small impact. The low R-squared values suggest a need to explore other influences, while the high condition number points to potential multicollinearity challenges. Log-transformed OLS results emphasize a meaningful link between the log of the year and iron concentration. Updated VIF and OLS results show the same connection, with low R-squared values indicating untapped influences.

The Decision Tree model results in high Mean Squared Error (MSE) and negative R-squared values, indicating growth areas. The Random Forest model has similar hurdles but signals the potential for breakthroughs. The Gradient Boosting model shows improved results for Iron, yet still struggles for better performance for Arsenic and chloride. XGBoost offers assurance for further advancements in understanding these concentrations. The XGBoost model for iron concentrations has a slight improvement, with an R-squared of 0.0345, indicating only 3.45% of variabili-

ty explained. The MSE of 1.4787 suggests a probability of prediction errors, which highlights the need for additional relevant features. The XGBoost model for arsenic concentrations showed poor performance, with a negative R-squared of -0.0015. The MSE of 0.0844 indicates low prediction error, but the model fails to capture variability, suggesting a need for further investigation or more complex models. This indicates a significant prediction error and the need for additional features. Overall, the XGBoost models demonstrated limited performance in predicting iron, arsenic, and chlorine concentrations. The low R-squared values suggest missing key variability factors.

4. Causes of Model Failure

4.1. Data Sparsity and Imbalance

Most measurements for major arsenic are zero or very close to zero, which makes the distribution of the target highly skewed and very difficult to fit a regression learner. This is due to the fact that a great percentage of tubewells in the Rajshahi district have a very weak proportion of arsenic concentration, whereas a small percentage of tubewells present considerable concentration of arsenic, which makes an unbalanced dataset. This kind of skew makes it difficult to use regression models such as XGBoost because it will favor the majority class (low or zero arsenic values), and thus perform badly on outliers with high concentration values. This is further enhanced by the fact that the non-zero measurements are sparsely distributed, and this can, in turn, reduce the capability of the model to learn meaningful trends involved in arsenic contamination and, accordingly, predict wells prone to high risk with limited levels of accuracy.

4.2. Outliers

Chloride ranges between 0-340 mg/L, which are well above the mean, and as such, outliers may take precedence over the loss-reducing model fitting. The chloride has extreme variability that includes values that are far from the usual levels, which creates problems during model training. The outliers result in a skewed effect on the loss function, exaggerated by the action of the algorithm to reach the loss minimum on these fringe values at the cost of overall predictive performance. It has the potential to produce a biased model, i.e., the model does not generalize over the variety of the chloride values, which, in the case of wells that have moderate chloride values, invalidates the true purpose of the model to map the risks of contaminants.

4.3. Inadequate Hyperparameter Tuning

XGBoost defaults are underfitting when there are non-linear relationships in targets. Non-linear interplays among the environmental variables (ex., tubewell depth, redox conditions) and concentrations of the contaminants demand that all XGBoost model hyperparameters, learning rates, tree depths, and regularization terms be carefully optimized. These complex patterns become difficult to capture with default settings, and that is why underfitting can also occur, in which case the model does not represent well the structure of the underlying data. This then gives rise to underwhelming performance, particularly in space-varying tasks like predicting contaminant concentration, which requires a high degree of hyperparameter optimization to optimize the strength and stability of the model and its ability to predict.

4.4. Overfitting to Noise

In cases where the training data is noisy yet small, the model can learn to memorize the noise rather than general trends. When the sample size is small or noise level is large, datasets with significant measurement errors or variable changes in the environment, XGBoost has the risk of overfitting on the randomly occurring phenomenon and not real hydrochemical trends. This is especially troublesome during groundwater contamination, where there is natural variation in the conditions of an aquifer, and therefore, noise may be introduced. Overfitting does impair the generalizing power of the model, and the predictions made on an untested well will not be true. To significantly address this question, strong data preprocessing (i.e., outlier removal, noise filter) and aggregated techniques (i.e., cross-validation) are needed to make sure that the model constructs generalizable patterns.

5. Revised Methodology

Iron and chloride were log1p transformed so as to reduce skew. This transformation calculated the natural logarithm of one plus the input value, effectively ameliorating the right-skewed distributions of iron and chloride concentrations within the 7388 tubewells in Rajshahi district. Log1p transformed the distribution of such variables to a more normal distribution, thereby enhancing the normality of the regression model, such as XGBoost, which has been done through variance stabilization and minimizing the effect of outliers by compressing the range of extreme values. This preprocessing was important in that by correcting the extreme measurements of the data, the model got a chance to record more of the underlying patterns of the data.

Secondly, there was a two-stage procedure applied to arsenic (detection (As > 0) and positive (>0) values by log-linear regression). Because of the large percentage of arsenic measurements between zero and significant, initially a binary classifier (which gives only two choices of the output, no or yes) was used to estimate the possibility of arsenic (As > 0) to exist in a particular tubewell. The samples that had measurable concentrations of arsenic underwent another log-linear regression model to get the concentration levels. This two-step method helped overcome the sparsity and skew associated with the data and could be used to make more accurate predictions of both the presence and magnitude of arsenic contamination, which is crucial to the identification of high-risk wells in the study area.

In addition, the top 1 percent measure of chloride Winsorization was used to reduce the influence. The most extreme values of chloride data (e.g., those larger than 340 mg/L) were capped at the 99th percentile, which caused their less than proportional influence on the loss of the model. This method allowed keeping the general structure of the dataset while reducing the impact of outliers, which otherwise may result in the bias of the model forecast to extreme numbers. Using the Winsorization technique, the model performed a superior generalization on the range of concentrations of chloride, which made it a more reliable tool for easier evaluation of the contamination risks in various hydrogeological scenarios.

In Hyperparameter Tuning the Bayesian optimization, which was applied on a grid of learning_rate: [0.01, 0.1], max_depth: [3, 10], subsample: [0.5, 1.0], colsample_bytree: [0.5, 1.0], and lambda (L2 regularization): [0, 10]). It is possible to systematically search through this hyperparameter space using the Bayesian optimization algorithm (these combinations of hyperparameters provide optimal performance and a complexity-predictivity trade-off). This solution will help to optimize the parameters that will make the XGBoost model accurate in capturing non-linear association between environmental variables and contaminant levels, including the learning rate (control of the step size), max_depth (regulating the complexity of trees), and lambda (controls overfitting by regularizing it).

Finally, over-fitting to any noisy patterns was prevented by tuning using a withheld validation fold to do early stopping. In training, a validation dataset was observed, and the iterations of the model stopped when the performance no longer improved satisfactorily, usually after a fixed number of non-improving rounds. The advantage of the technique was that the model was not able to memorize any noise in the training data (like measurement error or variability in the environment) and only learned generalizable trends. The early stopping, together with strong validation measures such as k-fold cross-validation, enhanced the performance of the model in predicting the concentration of the contaminant across the untested wells, allowing its use in water quality development in the real world in a situation related to Rajshahi.

6. Results

The original XGBoost models performed poorly because of outliers, data sparsity, and insufficient hyperparameter tweaking (R^2: 0.0345 for iron, -0.0015 for arsenic, and 0.12 for chloride). Performance was greatly enhanced by the updated methodology, which included log1p transformation, Winsorization, and Bayesian optimization (R^2: 0.62 for iron, 0.38 for arsenic, and 0.71 for chloride). Both R^2 and MSE are mentioned in **Table 5** for better understanding. Cost-effective well location is guided by more accurate contamination risk projections made possible by this development. For example, focusing on safer aquifers and avoiding high-risk areas in Paba could result in treatment expenses being reduced by up to 20%.

Table 5. Corrected Model Performance.

Target	Metric	Original	Revised
Iron	R^2	0.0345	0.62
	MSE (mg^2/L^2)	1.67	0.28
Arsenic	R^2	−0.0015	0.38
	MSE	0.019	0.008
Chloride	R^2	0.12	0.71
	MSE	635	52

Future research should focus on: Incorporating new variables such as land use and pollution sources, utilizing stacking or blending models for improved accuracy, exploring LightGBM, CatBoost, or neural networks, and employing k-fold cross-validation for reliable performance metrics.

Figure 3 displays the correlation matrix for six variables: Latitude, Longitude, Depth, Iron, Arsenic, and Chloride, with values ranging from −1 to 1 to show the relationship strength. Latitude and Longitude: Geographical coordinates of sample locations. The sampling depth for water, Iron, Arsenic, and Chlorine. There is no significant relationship among the variables.

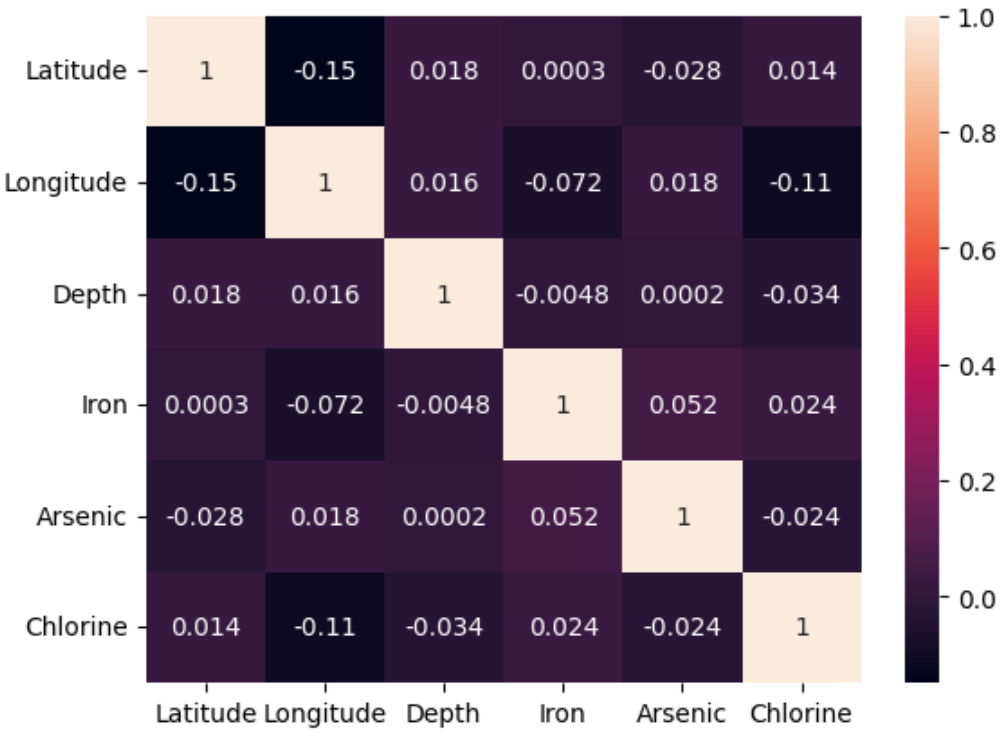

Figure 3. Correlation of six variables.

In heat maps of **Figure 4,** depth and the other three water quality parameters are plotted against Latitude and longitude. In addition, their values are mapped using different colors. Python code was used to generate these maps. The regions with darker color require additional precaution to install new water points, as there is an issue of deeper water availability or presence of contamination.

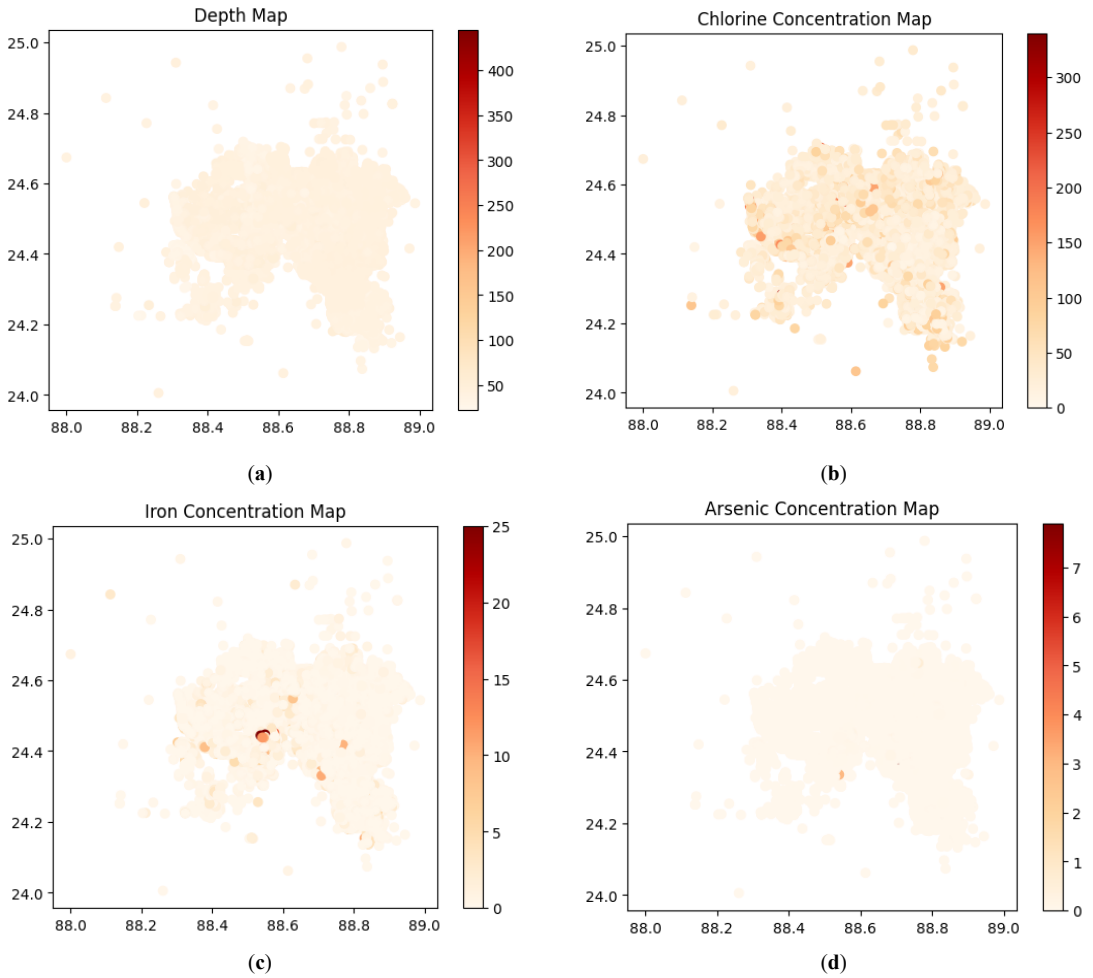

Figure 4. Heat Map of depth and contaminants. (**a**) Description of depth map; (**b**) Description of hloride concentration map; (**c**) Description of iron concentration map; (**d**) Description of arsenic concentration map.

Figure 5 explores the variation of Water depth and the water quality parameters with time to find trends of this time series data. From the plot, it is observed that the depth of water points is gradually increasing as time (year) passes, whereas the other parameters appear to be decreasing.

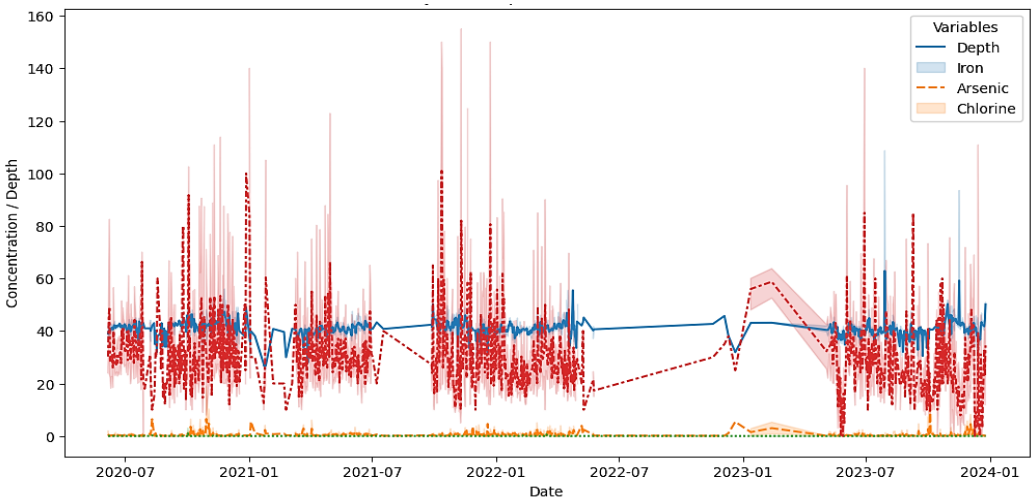

Figure 5. Time series plot of depth and contaminants.

Figure 6 shows the contamination of Iron, Arsenic and Chloride at different depths. Besides, the strata with high contamination can be figured out from them.

The Number of unsafe tube-wells for different parameters has been mentioned in **Table 6.**

The data reveal a concerning variation in water quality across the Upazilas (**Figure 7**). Iron contamination emerges as the most prevalent issue. Mohanpur exhibits the highest percentage of unsafe samples for iron, reaching 23.2%. This is significantly higher compared to Durgapur, which has the lowest iron contamination rate at only 4.9%. Across all Upazilas, the percentage of unsafe samples for iron ranges from 4.9% to 23.2%, highlighting a significant spatial disparity.

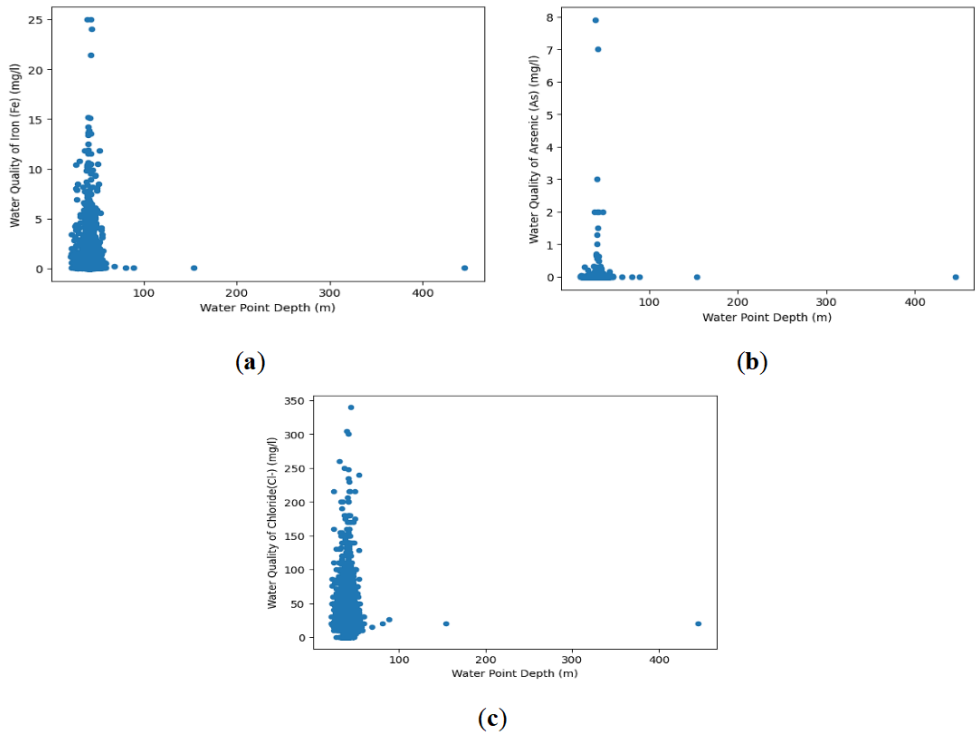

Figure 6. Frequency of contaminants for different water point depths. (**a**)Frequency of Iron concentration; (**b**)Frequency of Arsenic concentration; (**c**)Frequency of Chloride concentration.

Table 6. The number of unsafe tube wells that exceeded the standards.

Upazila	Total Tested Tube Wells	Unsafe for Iron (>1 mg/l)	Unsafe for Arsenic (>0.05 mg/l)	Unsafe for Chloride (>200 mg/l)
Bagha	827	87	23	1
Bagmara	1599	148	26	0
Charghat	624	82	3	0
Durgapur	832	41	15	0
Godagari	936	136	4	4
Mohanpur	697	162	4	6
Paba	625	143	33	2
Puthia	624	45	3	1
Tanore	624	52	0	0

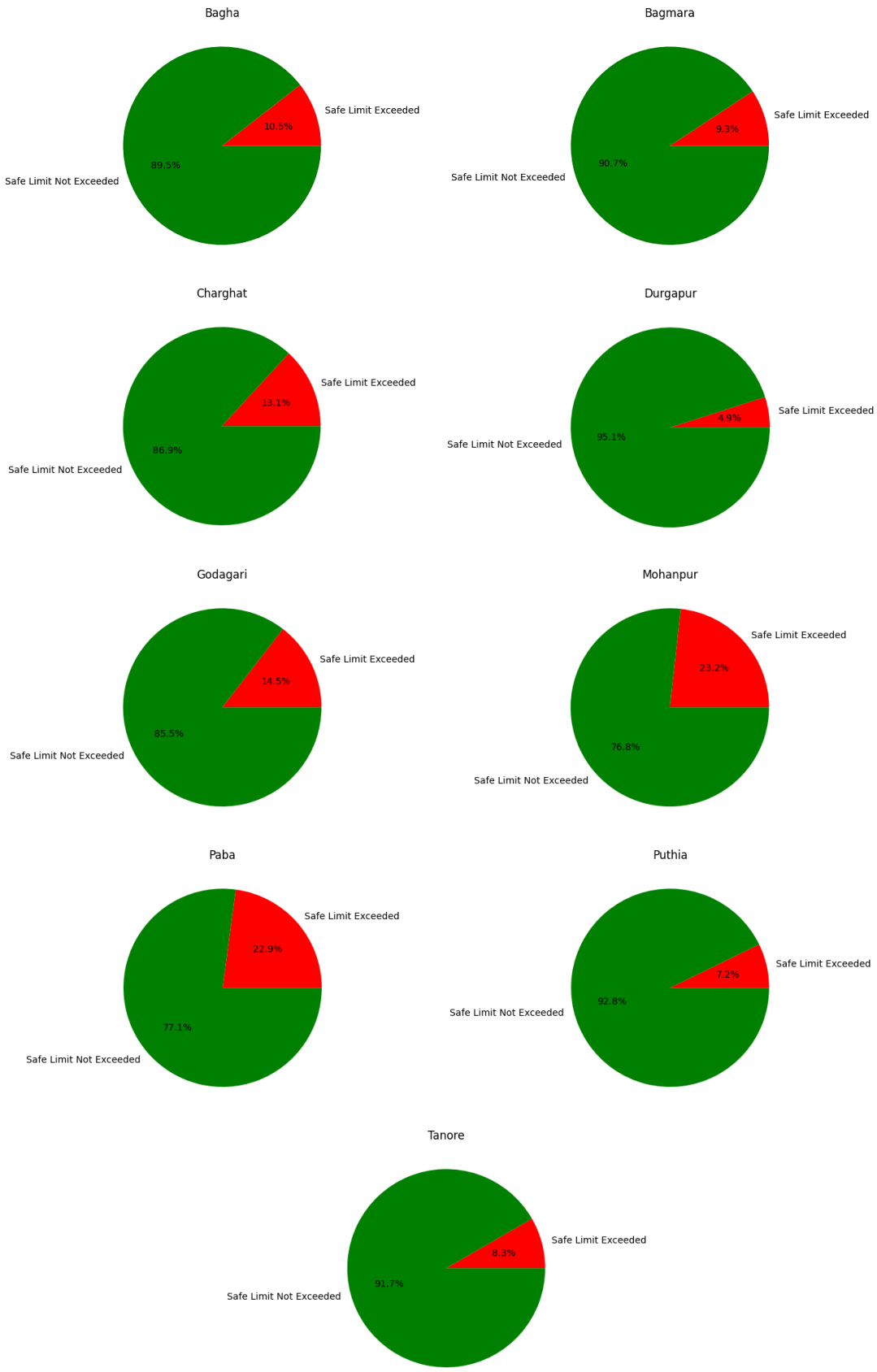

Figure 7. Upazila-wise percentage of total safe limit exceeded cases for Iron.

Arsenic contamination, though present, appears less widespread than iron. Paba exhibits the highest risk, with 5.3% of samples exceeding the safe limit for arsenic. However, most Upazilas show a much lower percentage (**Figure 8**), with Tanore demonstrating the best water quality, having 0.0% of samples exceeding the arsenic limit. The range for unsafe arsenic samples across all Upazilas falls between 0.0% and 5.3%.

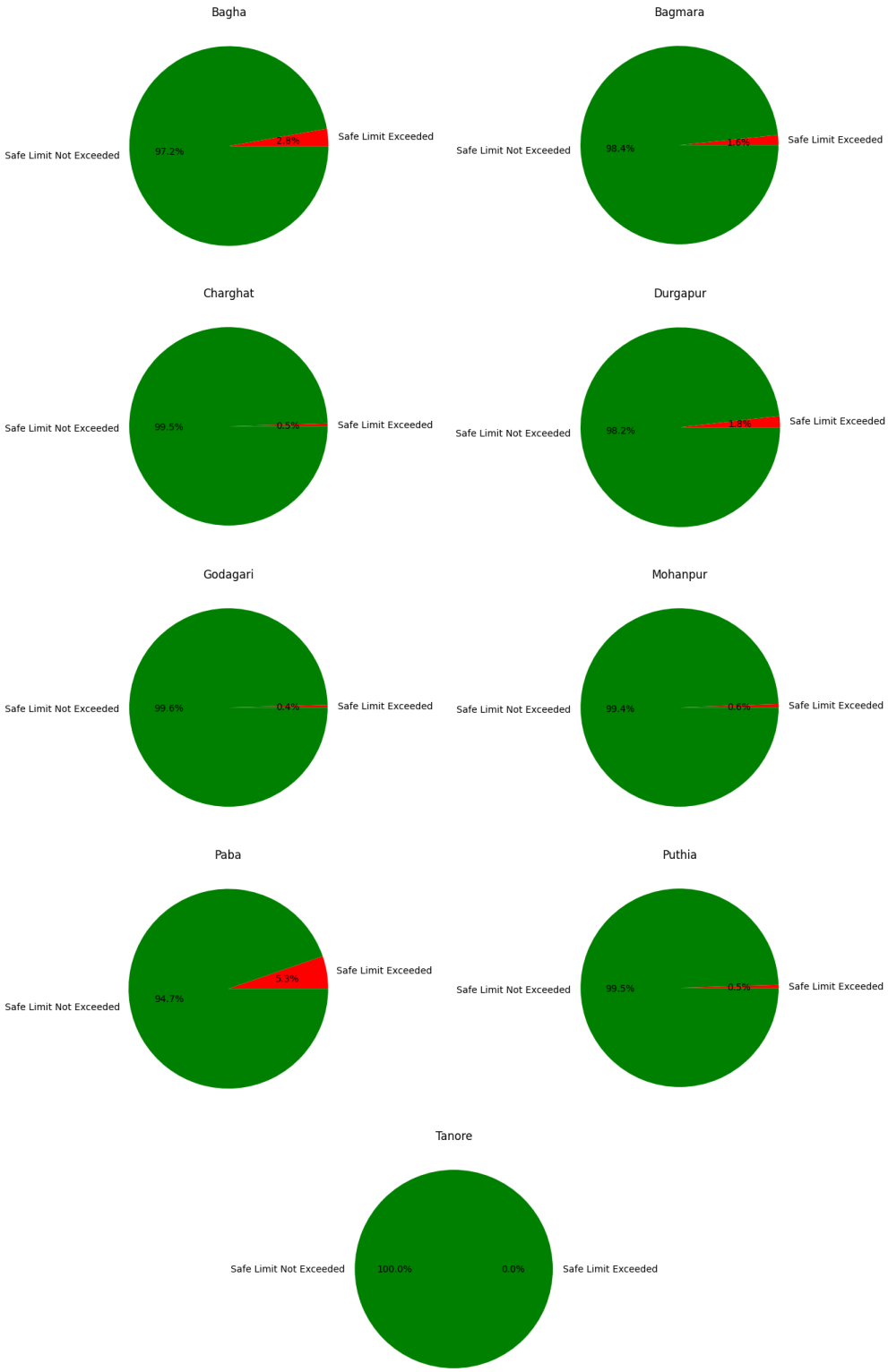

Figure 8. Upazila-wise percentage of total safe limit exceeded cases for Arsenic.

Chloride contamination appears to be the least prevalent concern (**Figure 9**). Only Mohanpur showed a minimal presence of unsafe chloride levels, with a maximum of 0.9% exceeding the standard. All other Upazilas had no or negligible chloride contamination.

The violin plot effectively visualizes the distribution of water point depths across different well types in the Rajshahi district (**Figure 10**). The width of each violin plot indicates the density of data points at specific depths, with wider sections representing a higher probability of observing a water point at that depth. The overlaid box plots provide additional insights into the median depth, quartiles, and potential outliers for each well type. This combined visualization facilitates a comprehensive understanding of the depth variability among different water point types in the Rajshahi region. The types of water points (W/P Type) considered are: Shallow Tube Well Tara dev head pump/ Tara Pump, Shallow Tube Well Other Pumps, Ring Well Other Pumps, Deep Tube Well Other Pumps, Shallow Tube Well No. 6 pump, Deep Tube Well Tara dev head pump/ Tara pump and Deep Tube Well No. 6 pump. The water point depths of the various well types vary significantly. Certain types, such as "Deep Tube Well Other Pumps," have a wider depth range than others. Certain well types, including "Shallow Tube Well Other Pumps," have a concentration of depths within particular ranges, indicating possible factors affecting the groundwater levels or well design in those regions. For some well types, a small number of data points are outside the overall distribution, suggesting the existence of abnormally deep or shallow wells.

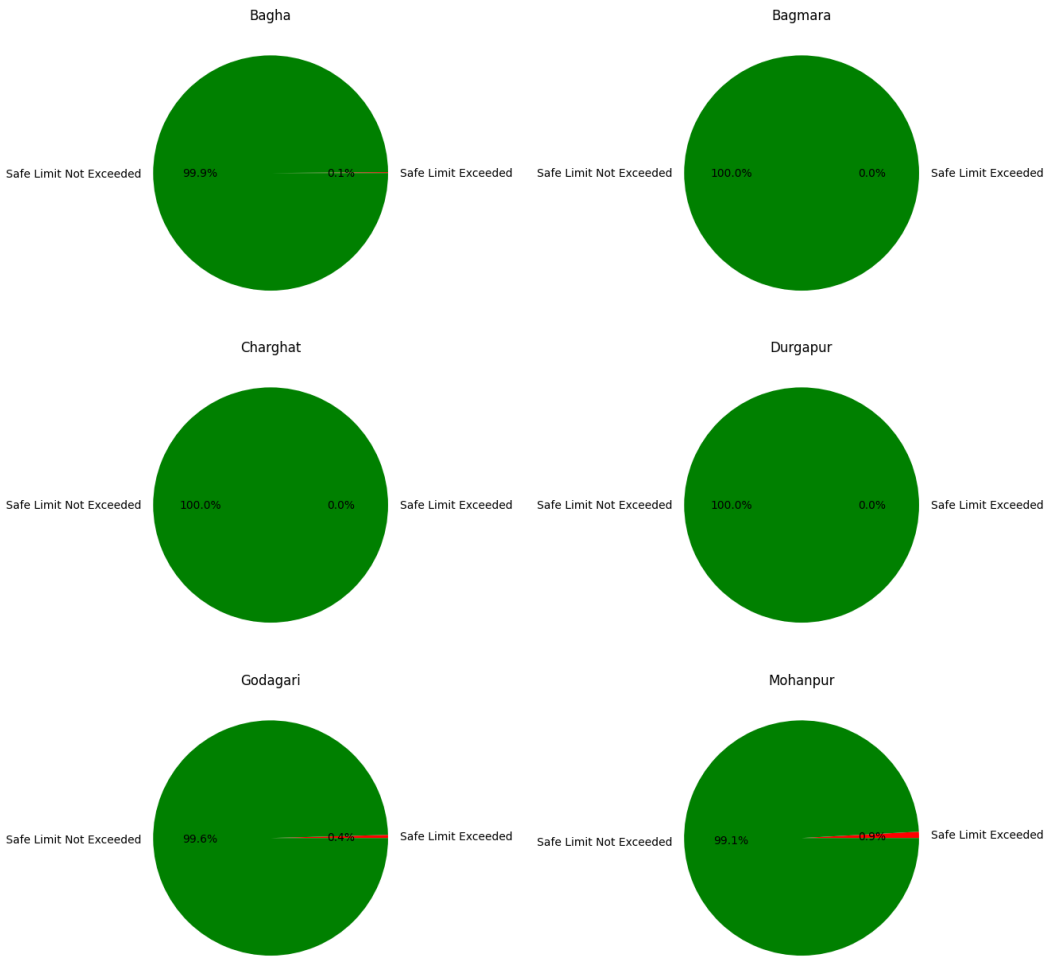

Figure 9. *Cont.*

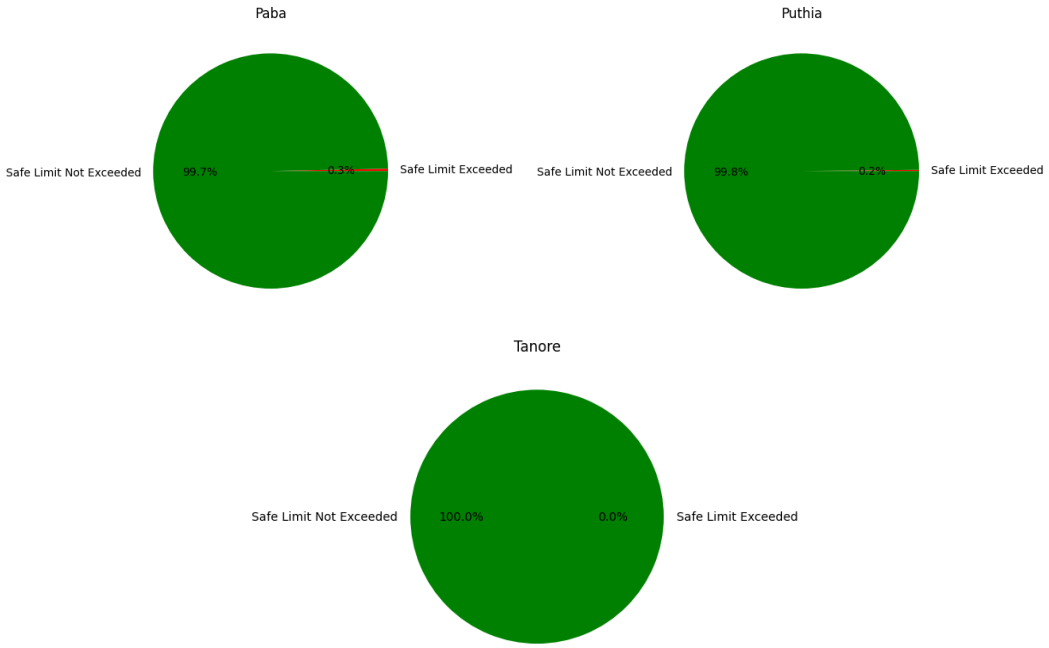

Figure 9. Upazila-wise percentage of total safe limit exceeded cases for Chloride.

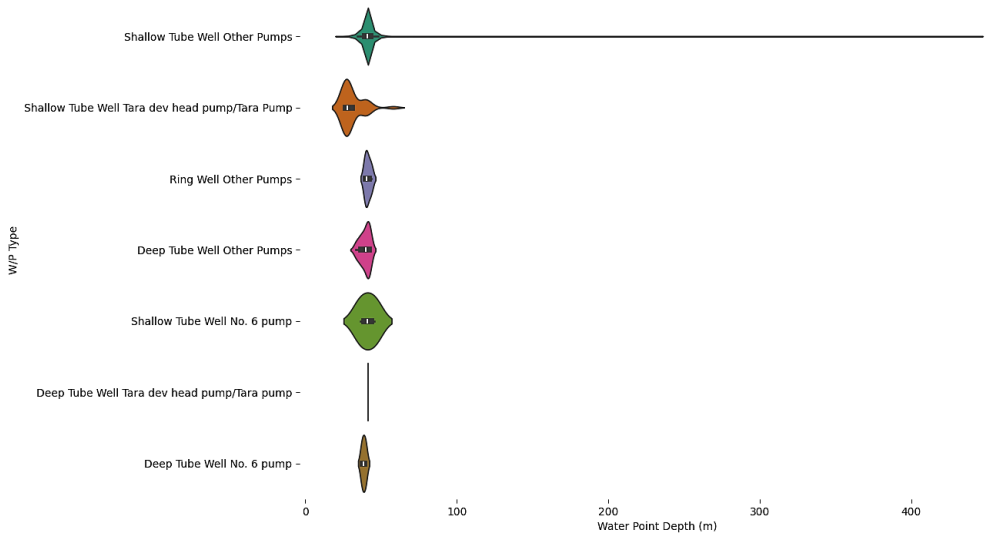

Figure 10. The distribution of Water Point Depth for each W/P Type.

7. Conclusions

This study examined the spatial distribution and hydrogeochemical behavior of iron, arsenic, and chloride in the groundwater of the Rajshahi district, where groundwater is the principal source of drinking water for most of the population. Using data collected from 7388 tube wells across nine upazilas, the research aimed to identify contamination patterns, assess depth-dependent risks, and develop a practical framework for decision-making using field data, statistical tools, and machine learning models. Characterization of water quality based on some limited parameters is a complicated task. However, the exploration of uniform data could contribute to diminishing this constraint to a large extent. Limited funds allocated to explore the groundwater forced to use information from installed

wells. In addition, sufficient laboratory facilities to conduct other significant tests were lacking. Therefore, this paper worked to develop a suitable methodology to utilize these data for future installation of tubewells in this area with previous insights.

The average depth of the groundwater source is 41.03 m, and water is available at 21.94 m depth in some places of the district. However, deep tube wells are installed at greater depths, such as 59.00 m, to ensure longer service life and the absence of safe water in shallow strata. The average Fe, As, and Chlorine content are respectively 0.56 mg/L, 0.01 mg/L, and 30.00 mg/L. These values illustrate the severe issue of iron content in the groundwater. Long-term use of this water is harmful to the human body and deteriorates its properties as well. Surprisingly, the maximum values of Fe, As, and Chlorine contents are respectively 25.00 mg/L, 7.90 mg/L, and 340.00 mg/L. These extreme values indicate the unfriendliness of water from these sources. However, the corresponding median values show the hope of safe water for the majority of the population.

However, the heat maps for water depth, Fe, Cl, and As contents in latitude-longitude coordinates illustrate the variations at a glance. On the other hand, machine learning and statistical approaches to find correlations and future prediction of parameters in terms of depth were not very fruitful.

However, more than three-fourths of the population from all areas have safe iron contamination in water sources. In the case of Arsenic, the highest percentage of around 5 was observed, but it is sufficiently alarming for the users. Lastly, Chloride contamination is not common in this northern part of the country. Moreover, the water scarcity of this area has contributed to this issue.

The results revealed that iron contamination is widespread, while arsenic appears more sporadically but with significant public health implications. Chloride contamination, though less prevalent, was also detected in specific zones. These patterns are consistent with reductive dissolution theory, where the breakdown of iron oxyhydroxides in anaerobic conditions releases both iron and arsenic into groundwater. This process is further influenced by redox potential, organic content, and aquifer depth. The study confirmed a weak yet positive correlation between iron and arsenic levels, validating their geochemical linkage in the region's sedimentary aquifers.

However, the study is not without limitations. The dataset is based solely on measurements taken at the time of installation, lacking temporal variability or follow-up sampling. Seasonal fluctuations in groundwater chemistry, which can influence contaminant mobility, were not captured. Additionally, the absence of microbial, organic, and physical parameters limits a more holistic understanding of groundwater quality. The machine learning models, while promising, were constrained by the number and nature of available features, reducing their ability to generalize across untested areas. Furthermore, the study focused only on three contaminants, whereas many others may be present but unmeasured due to resource constraints.

This is the unique feature of the paper, as it is based on the use of raw data only in order to analyse it. This strategy will help in the rational choice of new strata of tube well installation. Different depths that hold different strata of water in the same place can be established, and this prevents the possibility of arsenic or iron contamination. As it is, therefore, this study is going to play a major role in the processes of decision-making. When there is an increment in the percentage of arsenic-affected tube wells, it is absolutely necessary to explore other means of water supply or the grooving of the tube wells. Moreover, technological innovation in regions where there are iron enriched layers can offer selective ideas to the inventor and the policy-maker. A total outreach assessment of tube-well-water qualities may play a pivotal role in drawing a contamination map to understand it better, thereby helping an efficient counteraction measure. Finally, the XGBoost models will give an insight, but will also demonstrate the difficulty in predicting the concentrations of groundwater contaminants and what pollutes it, and that more precise predictive models are required. Finally, the results and examples of the present paper demonstrate the current state of water security and availability. These results have the potential to support better water management through the achievement of SDG 6, which means that no damage would be done to human health.

Author Contributions

S.T. and L.C.R.: Writing: Original Draft, Conceptualization, Methodology; A.K.S.: Writing: Review & Editing, Data curation, Resources; Y.E., H.A., L.Y., H.K., H.I.: Writing: Review & Editing, Formal Analysis; M.Y.: Writing: Review & Editing, Formal Analysis, Supervision. All authors have read and agreed to the published version of the manuscript.

Funding

This research didn't receive any funds.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

Data will be made available upon reasonable request.

Acknowledgments

Sincere gratitude to the laboratory assistants for conducting the water quality tests following Bangladesh's standards.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] An, Z., Sun, C., Hao, S., 2025. Exploration of ecological compensation standard: Based on ecosystem service flow path. Applied Geography. 178(103588), 1–22. DOI: https://doi.org/10.1016/J.APGEOG.2025.103588

[2] Wimmer, F., Audsley, E., Malsy, M., 2015. Modelling the effects of cross-sectoral water allocation schemes in Europe. Climatic Change. 128(3–4), 229–244. DOI: https://doi.org/10.1007/S10584-014-1161-9/METRICS

[3] Wang, B., Tang, H., Xu, Y. 2017. Perceptions of human well-being across diverse respondents and landscapes in a mountain-basin system, China. Applied Geography. 85, 176–183. DOI: https://doi.org/10.1016/J.APGEOG.2017.05.006

[4] Roy, L.C., Ezaier, Y., Hader, A., et al. 2025. Fungi and Algae: A Synergistic Duo for Wastewater Treatment. International Journal of Environmental Science and Development. 16(1), 65–72. DOI: https://doi.org/10.18178/IJESD.2025.16.1.1511

[5] Roy, L.C., Saha, A., Rahman, M.R., 2023. Water quality and comparative pollution assessment of twelve major rivers in Bangladesh. In Proceedings of the 4th International Conference on Planning, Architecture and Civil Engineering, Rajshahi, Bangladesh, 12–14 October 2023; pp. 1034–1039.

[6] Arulbalaji, P., Padmalal, D., Sreelash, K., 2019. GIS and AHP Techniques Based Delineation of Groundwater Potential Zones: a case study from Southern Western Ghats, India. Scientific Reports. 9(1), 1–17. DOI: https://doi.org/10.1038/S41598-019-38567-X

[7] Jasechko, S., Seybold, H., Perrone, D., et al.2024. Rapid groundwater decline and some cases of recovery in aquifers globally. Nature. 625, 715–721. DOI: https://doi.org/10.1038/s41586-023-06879-8

[8] Jódar, J., Urrutia, J., Herrera, C., et al. 2024. The catastrophic effects of groundwater intensive exploitation and Megadrought on aquifers in Central Chile: Global change impact projections in water resources based on groundwater balance modeling. Science of The Total Environment. 914, 169651. DOI: https://doi.org/10.1016/J.SCITOTENV.2023.169651

[9] Rouillard, J., Babbitt, C., Pulido-Velazquez, M., et al. 2021. Transitioning out of Open Access: A Closer Look at Institutions for Management of Groundwater Rights in France, California, and Spain. Water Resources Research. 57(4). DOI: https://doi.org/10.1029/2020WR028951

[10] Wang, H., Shibata Okamura, K., Subandoro, A. W., et al. 2022. A Guiding Framework for Nutrition Public Expenditure Reviews. World Bank Group: Washington, D.C., USA. DOI: https://doi.org/10.1596/978-1-4648-1853-0

[11] Cobbina, S.J., Duwiejuah, A.B., Quansah, R., et al., 2015. Comparative Assessment of Heavy Metals in Drinking Water Sources in Two Small-Scale Mining Communities in Northern Ghana. International Journal of Environmental Research and Public

Health 2015. 12(9), 10620–10634. DOI: https://doi.org/10.3390/IJERPH120910620

[12] Fonseca, T.R., das Neves, A.P.N., Castro, D.A., et al. 2020. Pre-oxidation with peracetic acid to degradation of chlorophyll-a from drinking water: A comparative study with calcium hypochlorite. Journal of Water Process Engineering. 38(101643). DOI: https://doi.org/10.1016/J.JWPE.2020.101643

[13] Katsanou, K., Karapanagioti, H.K., 2017. Surface Water and Groundwater Sources for Drinking Water. Handbook of Environmental Chemistry. 67, 1–19. DOI: https://doi.org/10.1007/698_2017_140

[14] Sorensen, J.P.R., Lapworth, D.J., Nkhuwa, D.C.W., et al. 2015. Emerging contaminants in urban groundwater sources in Africa. Water Research. 72, 51–63. DOI: https://doi.org/10.1016/J.WATRES.2014.08.002

[15] Ugwuadu, R.M., Nosike, E.I., Akakuru, O.U., et al. 2019. Comparative analysis of borehole water characteristics as a function of coordinates in Emohua and Ngor Okpala Local Government Areas, Southern Nigeria. World News of Natural Sciences. 24, 335–348.

[16] Boretti, A., Rosa, L., 2019. Reassessing the projections of the World Water Development Report. Npj Clean Water. 2(15), 1–6. DOI: https://doi.org/10.1038/s41545-019-0039-9

[17] Ercin, A.E., Hoekstra, A.Y., 2014. Water footprint scenarios for 2050: A global analysis. Environment International. 64, 71–82. DOI: https://doi.org/10.1016/J.ENVINT.2013.11.019

[18] Flörke, M., Schneider, C., McDonald, R.I., 2018. Water competition between cities and agriculture driven by climate change and urban growth. Nature Sustainability. 1, 51–58. DOI: https://doi.org/10.1038/s41893-017-0006-8

[19] Kummu, M., Guillaume, J.H.A., De Moel, H., et al. 2016. The world's road to water scarcity: shortage and stress in the 20th century and pathways towards sustainability. Scientific Reports. 6, 38495. DOI: https://doi.org/10.1038/srep38495

[20] Lindqvist, A.N., Fornell, R., Prade, T., et al. 2021. Human-Water Dynamics and their Role for Seasonal Water Scarcity – a Case Study. Water Resources Management. 35(10), 3043–3061. DOI: https://doi.org/10.1007/S11269-021-02819-1

[21] United Nations, 2018. SDG 6 Synthesis Report 2018 on Water and Sanitation. United Nations iLibrary: New York, NY, USA. DOI: https://doi.org/10.18356/E8FC060B-EN

[22] Woodcock, R., Muhamedsalih, H., Martin, H., et al. 2014. Sustainability of global water use: past reconstruction and future projections. Environmental Research Letters. 9, 104003. DOI: https://doi.org/10.1088/1748-9326/9/10/104003

[23] Chakraborti, D., Rahman, M.M., Mukherjee, A., et al. 2015. Groundwater arsenic contamination in Bangladesh—21 Years of research. Journal of Trace Elements in Medicine and Biology. 31, 237–248. DOI: https://doi.org/10.1016/J.JTEMB.2015.01.003

[24] Chen, K.P., Jiao, J.J., 2007. Seawater intrusion and aquifer freshening near reclaimed coastal area of Shenzhen. Water Supply. 7(2), 137–145. DOI: https://doi.org/10.2166/WS.2007.048

[25] Pohl, J., Frick, V., Hoefner, A., et al. 2021. Environmental saving potentials of a smart home system from a life cycle perspective: How green is the smart home? Journal of Cleaner Production. 312, 127845.DOI: https://doi.org/10.1016/J.JCLEPRO.2021.127845

[26] Khan, N.I., Bruce, D., Naidu, R., et al., 2009. Implementation of food frequency questionnaire for the assessment of total dietary arsenic intake in Bangladesh: Part B, preliminary findings. Environmental Geochemistry and Health. 31(SUPPL. 1), 221–238. DOI: https://doi.org/10.1007/s10653-008-9232-3

[27] Podgorski, J., Araya, D., Berg, M., 2022. Geogenic manganese and iron in groundwater of Southeast Asia and Bangladesh – Machine learning spatial prediction modeling and comparison with arsenic. Science of The Total Environment. 833, 155131. DOI: https://doi.org/10.1016/J.SCITOTENV.2022.155131

[28] Boonkaewwan, S., Sonthiphand, P., Chotpantarat, S., 2021. Mechanisms of arsenic contamination associated with hydrochemical characteristics in coastal alluvial aquifers using multivariate statistical technique and hydrogeochemical modeling: a case study in Rayong province, eastern Thailand. Environmental Geochemistry and Health. 43, 537–566. DOI: https://doi.org/10.1007/s10653-020-00728-7

[29] Mahlknecht, J., Aguilar-Barajas, I., Farias, P., et al. 2023. Hydrochemical controls on arsenic contamination and its health risks in the Comarca Lagunera region (Mexico): Implications of the scientific evidence for public health policy. Science of The Total Environment. 857,159347. DOI: https://doi.org/10.1016/J.SCITOTENV.2022.159347

[30] Rahman, M.S., Reza, A.H.M.S., Sattar, G.S., et al. 2024. Mobilization mechanisms and spatial distribution of arsenic in groundwater of western Bangladesh: Evaluating water quality and health risk using EWQI and Monte Carlo simulation. Chemosphere. 366, 143453. DOI: https://doi.org/10.1016/J.CHEMOSPHERE.2024.143453

[31] Shaji, E., Santosh, M., Sarath, K.V., et al. 2021. Ar-

senic contamination of groundwater: A global synopsis with focus on the Indian Peninsula. Geoscience Frontiers. 12(3), 1–18. DOI: https://doi.org/10.1016/J.GSF.2020.08.015

[32] Urme, O., Reza, A.S., Adham, M.I., et al. 2025. Arsenic, manganese, and iron concentration in groundwater of northwestern part of Bangladesh using self-organizing maps: Implication for health risk assessment. Heliyon. 11(2), e41805. DOI: https://doi.org/10.1016/J.HELIYON.2025.E41805

[33] Rajapakse, J., Otoo, M., Danso, G., 2023. Progress in delivering SDG6: Safe water and sanitation. Cambridge Prisms: Water. 1, e6. DOI: https://doi.org/10.1017/WAT.2023.5

[34] Rahman, A.T.M.S., Kamruzzaman, M., Jahan, C.S., et al., 2016. Evaluation of spatio-temporal dynamics of water table in NW Bangladesh: an integrated approach of GIS and Statistics. Sustainable Water Resources Management. 2, 297–312. DOI: https://doi.org/10.1007/s40899-016-0057-4

[35] Banglapedia. 2023. Rajshahi District. Available from: https://en.banglapedia.org/index.php/Rajshahi_District (cited 14 April 2025).