GeoHealth

RESEARCH ARTICLE

10.1029/2021GH000585

Predicting the Distribution of Arsenic in Groundwater by a Geospatial Machine Learning Technique in the Two Most Affected Districts of Assam, India: The Public Health Implications

Bibhash Nath[1], **Runti Chowdhury**[2], **Wenge Ni-Meister**[1], **and Chandan Mahanta**[3]

[1]Department of Geography and Environmental Science, Hunter College of City University of New York, New York, NY, USA, [2]Department of Geological Sciences, Gauhati University, Guwahati, India, [3]Department of Civil Engineering, Indian Institute of Technology Guwahati, Guwahati, India

Key Points:

- A habitation-level predictive model was developed to identify arsenic risk area by using intrinsic and extrinsic predictor variables
- We identified a large area potentially unsafe for the installation of tube wells for water use and highlighted public health concerns
- We identified a potential target area for policymakers to consider immediately for the provision of treated, piped water to rural households

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

B. Nath,
bibhash12@gmail.com;
bibhash.nath86@myhunter.cuny.edu

Citation:

Nath, B., Chowdhury, R., Ni-Meister, W., & Mahanta, C. (2022). Predicting the distribution of arsenic in groundwater by a geospatial machine learning technique in the two most affected districts of Assam, India: The public health implications. *GeoHealth*, *6*, e2021GH000585. https://doi.org/10.1029/2021GH000585

Received 30 DEC 2021
Accepted 18 FEB 2022

Author Contributions:

Conceptualization: Bibhash Nath
Data curation: Bibhash Nath, Runti Chowdhury, Chandan Mahanta
Formal analysis: Bibhash Nath
Funding acquisition: Wenge Ni-Meister, Chandan Mahanta
Investigation: Bibhash Nath

Abstract

Arsenic (As) is a well-known carcinogen and chemical contaminant in groundwater. The spatial heterogeneity in As distribution in groundwater makes it difficult to predict the location of safe areas for tube well installations, consumption, and agriculture. Geospatial machine learning techniques have been used to predict the location of safe and unsafe areas of groundwater As. We used a similar machine learning technique and developed a habitation-level (spatial resolution 250 m) predictive model to determine the risk and extent of As >10 μg/L in groundwater in the two most affected districts of Assam, India, with an aim to advise policymakers on targeted interventions. A random forest model was employed in Python environments to predict the probabilities of As at concentrations >10 μg/L using intrinsic and extrinsic predictor variables, which were selected for their inherent relationship with As occurrence in groundwater. The relationships between predictor variables and proportions of As occurrences >10 μg/L follow the well-documented processes leading to As release in groundwater. We identified potential As hotspots based on a probability of ≥0.7 for As >10 μg/L, including regions not previously surveyed and extending beyond previously known As hotspots. Of the total land area (6,500 km^2), 25% was identified as a high-risk zone, with an estimated 155,000 people potentially consuming As through drinking water or cooking food. The ternary hazard probability map (showing high, moderate, and low risk for As >10 μg/L) could inform policymakers on establishing newer drinking water treatment plants and providing safe drinking water connections to rural households.

Plain Language Summary

We developed a habitation-level predictive model to identify the extent of arsenic risk in areas of the two most affected districts of Assam, India, based on an understanding the relationships between intrinsic and extrinsic predictor variables and As concentration in groundwater. We identified a large area potentially unsafe for the installation of tube wells for water use, which has implications for public health as, in many localities, groundwater is the sole drinking water source. We estimated that about 155,000 people have been potentially exposed to As concentrations >10 μg/L. The hazard probability map generated in this study could be used by policymakers for targeted well-testing campaigns, as it highlights where inhabitants must consider switching their wells for safe water access. As a mitigation strategy, we identified the region where authorities must consider providing treated piped water connections to rural households.

1. Introduction

Groundwater is one of the most critical natural resources and plays an important role in day-to-day human life and economic development. It is the largest freshwater resource on earth (Foster & Chilton, 2003). However, the presence of naturally occurring arsenic (As) in groundwater is known to affect fluvial sedimentary aquifers worldwide (Ravenscroft et al., 2009). The occurrence of As in groundwater is widespread in the Indian subcontinent, especially in West Bengal, India (McArthur et al., 2004; Nath et al., 2009) and Bangladesh (Chakraborti et al., 2002; van Geen et al., 2003). The use of As-contaminated groundwater for drinking and irrigation has caused widespread public health issues (Smith et al., 2000). The primary route of As exposure to humans is through drinking of water and consuming vegetables and food grains, particularly rice, contaminated by As (Rahman et al., 2011). Consequently, the World Health Organization (WHO) has set an As concentration safety guideline of a maximum of 10 μg/L in drinking water.

Methodology: Bibhash Nath
Resources: Wenge Ni-Meister
Software: Bibhash Nath
Validation: Bibhash Nath
Visualization: Bibhash Nath
Writing – original draft: Bibhash Nath
Writing – review & editing: Bibhash Nath, Runti Chowdhury, Wenge Ni-Meister, Chandan Mahanta

In India and Bangladesh alone, the consumption of groundwater with elevated As has probably killed hundreds of thousands of people prematurely and exposed millions more to a range of ailments (Flanagan et al., 2012). Arsenic in groundwater and its clinical manifestation was first reported in West Bengal, India, in 1984 and was later characterized as the "greatest mass poisoning in human history" by WHO (Saha, 1995). Three decades later, a survey of drinking water collected in the homes of many rural households documented that 40 million people in Bangladesh alone were still consuming drinking water with As concentrations >10 μg/L (BBS & UNICEF, 2014). This alarming finding reflects failures toward implementing effective intervention strategies. The installation of shallow, private wells has continued unabated but without sufficient knowledge of whether the wells are safe for use, since sedimentary structure is extremely complex and the distribution of As in the aquifers is extremely heterogeneous (McArthur et al., 2004; Nath et al., 2008; van Geen et al., 2003).

The occurrence of As in the Brahmaputra flood plain (BFP) groundwater, in Assam, India, was reported much later (Borah et al., 2009; Chetia et al., 2011; Singh, 2004). Singh (2004) reported As concentrations >50 μg/L in 20 of the 30 districts of Assam. The study in the Bongaigaon and Darrang districts of the BFP demonstrated As enrichment in groundwater ranging from 5 to 606 μg/L with 66% of the analyzed groundwaters containing an As concentration above the Indian drinking water standard and WHO guideline value of 10 μg/L (Enmark & Nordborg, 2007). Mahanta et al. (2015) reported As concentrations >10 μg/L in 29% of the tested wells based on a state-wide survey of 56,180 tube wells. Verma et al. (2016) and Choudhury et al. (2018) reported that the high As concentration in groundwater is associated with thick clay capping at the top of the aquifer. Choudhury et al. (2018) further suggested that the thick clay layer inhibited flushing of the aquifer, which resulted in high As concentrations in groundwater due to longer sediment-water interactions. Goswami et al. (2014) also reported elevated As concentrations in distinct locations in Majuli, a river island.

The sedimentological history of the BFP is quite similar to the As-contaminated regions of Bangladesh. Therefore, conclusions about the source, extent, and mobilization mechanism of As found in Bangladesh could be partly applicable to the BFP. However, the aquifer in the BFP region is much less perturbed from the agricultural use of groundwater (CGWB, 2013), which was one of the most critical factors in the development of high As in Bangladeshi groundwater (DPHE, 2001). A lack of resources has prevented the comprehensive blanket testing of As in groundwater in the state of Assam, India. Therefore, mapping the spatial extent of the aquifer contaminated with As is urgently needed, considering the hazards to human health posed by and the fact that inhabitants rely heavily on groundwater for their domestic water needs.

Here, we developed a machine learning algorithm to predict the local-scale distribution of As in groundwater by establishing a statistical relationship between As occurrences and environmental predictors (intrinsic and extrinsic). A systematic machine learning algorithm would not only predict the local-scale distribution of As in groundwater but also, through policy intervention, inform the villagers of the status of their wells, including whether their water is safe or unsafe for consumption.

Machine learning approaches have previously been adopted for groundwater contamination prediction studies in India, such as for the states of Gujarat (Wu et al., 2021) and Uttar Pradesh (Bindal & Singh, 2019), as well as for the entire country (Mukherjee et al., 2021; Podgorski et al., 2020). Likewise, this study focused on risk determination using geostatistical machine learning approaches to identify the extent of elevated in As concentrations in the two most affected districts of Assam, India, where a previous study showed that As concentrations in groundwater increased with distance from the river (Choudhury et al., 2018). This pattern of this As occurrence is the opposite of that in the downstream areas of Bangladesh (DPHE, 2001; van Geen et al., 2003). Such an observed pattern makes this study unique and highlights the challenges of in delineating the spatial extent of elevated groundwater As concentrations. Additionally, we conducted this study at the habitation level with a finer spatial resolution in a smaller area. This finer resolution study allows the model to characterize the groundwater As occurrences with greater certainty, produces a more accurate model prediction, and enables the proper implementation of mitigation measures for greater public health benefits.

2. Study Area

The study area, Jorhat and Golaghat district, lies on the southern bank of the Brahmaputra River in Assam, including the river island Majuli (a recently created district in Assam). The total area is approximately 6,500 km^2 (Figure 1), and the total population is 2.5 million. These areas have been reported to contain elevated As

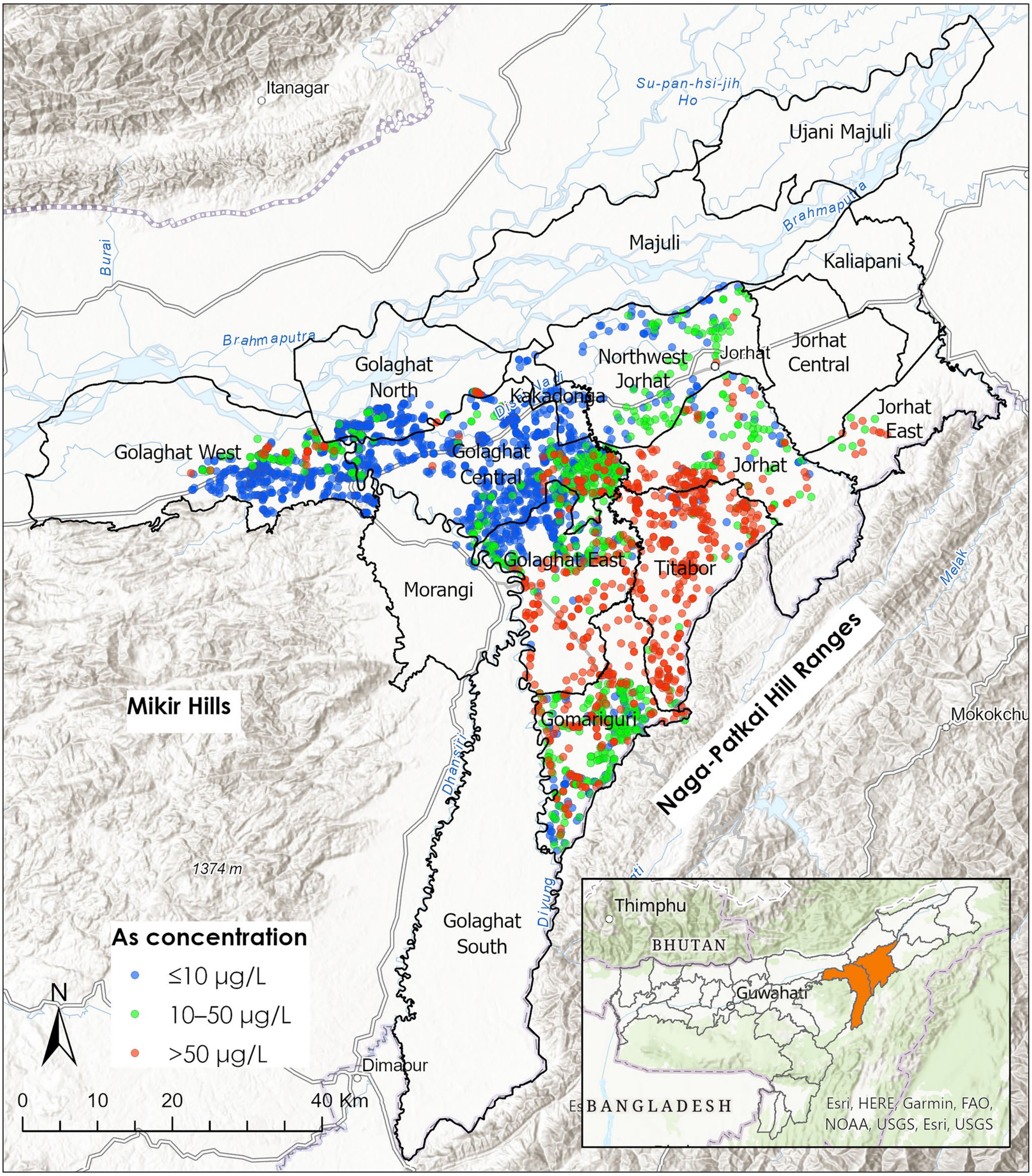

Figure 1. The study area map shows the location of measured As concentrations ($n = 3{,}600$) in groundwater. The administrative boundaries of sixteen subdistricts (blocks) are also shown. The inset map showing the study site within Assam, India.

concentrations and to be the most affected among the districts in Assam (Mahanta et al., 2015). The study area is a part of the Brahmaputra River floodplains and consist of younger and older alluvial sediment depositional environments. The area is encircled by the Brahmaputra River to the north, the Naga Patkai Hill range to the south, and the Mikir Hills to the west.

The average rainfall is 2,818 mm, the temperature ranges between 6°C and 38°C, and the mean relative humidity is between 92% and 98%. The Brahmaputra River, and its tributaries, Dhansiri, Bhogdoi, and Kakodonga, drain the study area (CGWB, 2013). The alluvial deposits, characterized by light to dark gray colored sands, silts, and clay, are mainly confined to the floodplain areas of the Brahmaputra River and its tributaries (CGWB, 2013). The inhabitants of the study area are highly dependent on groundwater to meet domestic water requirements, while irrigated water use is low (CGWB, 2013).

3. Methodology

3.1. Model Environment

The predictive model was implemented in the Python programming language using a random forest machine learning algorithm (Pedregosa et al., 2011). The probabilistic relationships between the intrinsic and extrinsic predictors and As concentration in groundwater were determined. The random forest model is an ensemble of decision trees. The decision tree algorithm uses a supervised learning method for classification and regression and is a non-parametric method (Rokach & Maimon, 2008). The role of a decision tree is to grow, as target variables are split into consecutive nodes by predictor variables with a conditional statement starting from a root node through to a decision node and ultimately to a leaf node where the decision is being made about an instance. The decision to split the target data set is based on the importance of a predictor variable and its associated conditions. The choice and condition of the predictors to split the data sets are based on how best a predictor decreases the randomness or entropy in the input target samples. The best split is selected based on the lowest Gini scores attained after the split.

In a random forest model, each tree uses a different randomly selected subset of predictor variables (typically the square root of the total number of predictor variables) and a random selection with the replacement of data rows (bootstrap aggregating or bagging). Because of data replacement, roughly one-third of the data are not used in growing a tree. Randomness is introduced during the creation of trees to promote uncorrelated forests, avoid multicollinearity among predictors and improve model performance (Ho, 1995). The randomness is created in such a way that each tree is grown using different sample data sets, including a different set of predictor variables. Each tree makes its prediction, and these predictions are then averaged in the case of regression or the majority votes in the case of classification to produce a single outcome. The random forest model takes care of multicollinearity among the predictors because not all variables are used simultaneously in decision trees (Podgorski et al., 2020).

3.2. Target Variable: Arsenic in Groundwater

We compiled the measured As concentrations in groundwater for the two most affected districts of Assam from two sources (Choudhury et al., 2018; Mahanta et al., 2015). The lack of available geolocated data from other areas have restricted this study to these areas. These data were generated through a combination of measurements, including in the field using test kits and in the laboratory using high-resolution inductively coupled plasma mass spectrometry. For modeling, the data were aggregated to grids of a spatial resolution of 250 m by using the geometric mean of concentrations. The extent of the grids was exactly the same as the GeoTIFFs of the predictor variables, such that the grids could be stacked exactly on top of the GeoTIFFs, which avoids grids intersecting with the predictor data sets' raster grid. The gridded As values were then converted into binary form by assigning all As concentrations meeting the WHO guideline of ≤10 μg/L to zero and all concentrations >10 μg/L to one. The binary conversion of the data was done to determine the extent and total number of people exposed to elevated As to aid policymakers in their decision-making regarding priority area for health intervention studies (Podgorski et al., 2020).

Table 1
The List of Predictor Variables Used in the Development of Random Forest Model

Predictors	Spatial resolution	References
Precipitation	30arc-sec (~1,000m)	Fick and Hijmans (2017)
Temperature	30arc-sec (~1,000m)	Fick and Hijmans (2017)
Potential evapotranspiration (PET)	30arc-sec (~1,000m)	Trabucco and Zomer (2018)
Aridity	30arc-sec (~1,000m)	Trabucco and Zomer (2018)
Topsoil organic carbon	250m	Poggio et al. (2021) (SoilGrids)
Subsoil organic carbon	250m	SoilGrids
Topsoil sand	250m	SoilGrids
Subsoil sand	250m	SoilGrids
Topsoil silt	250m	SoilGrids
Subsoil silt	250m	SoilGrids
Topsoil clay	250m	SoilGrids
Subsoil clay	250m	SoilGrids
Topsoil cation exchange capacity	250m	SoilGrids
Subsoil cation exchange capacity	250m	SoilGrids
Topsoil pH	250m	SoilGrids
Subsoil pH	250m	SoilGrids
Topsoil bulk density	250m	SoilGrids
Subsoil bulk density	250m	SoilGrids
Fluvisols	250m	SoilGrids
Topsoil coarse fragments	250m	SoilGrids
Subsoil coarse fragments	250m	SoilGrids
Elevation	30m	Farr and Kobrick (2000)
Slope	30m	Computed from elevation.
Distance to river	15arc-sec (~500m)	Lehner et al. (2008)
Topographic wetness index	500m	Hengl (2018)
Land use/land cover	Polygon	Roy et al. (2016)

Note. The data source and spatial resolution are provided.

3.3. Predictor Variables

The occurrence and spatial extent of groundwater As were modeled through a statistical relationship of the influence of intrinsic and extrinsic predictors and As concentrations within the study area. In total, 26 spatially continuous predictor variables (precipitation, temperature, potential evapotranspiration, aridity, topsoil– and subsoil–organic carbon, sand, silt, clay, cation exchange capacity, pH, bulk density, and coarse fragments, fluvisols, slope, elevation, distance to the river, topographic wetness index, and land use/land cover) were used in developing the model. A detailed list of predictor variables and the data sources are presented in Table 1. The chosen predictor variables can be broadly classified into climate variables, soil (topsoil and subsoil) characteristics, hydrology, and land use/land cover. The spatial resolution of the predictor variables was mostly 250 m, except for climate variables, which were 1,000 m, and distance to the river and topographic wetness index, which were both 500 m. We resampled those coarser-resolution data sets to finer resolution using spatial interpolation techniques.

The predictor variables were selected based on their relationships to the process of accumulation and dissolution of As in groundwater (Podgorski et al., 2020; Podgorski & Berg, 2020; Mukherjee et al., 2021). The soil parameters, estimated at a 2.0-m depth, can create geochemical environments favorable for As release. The enriched concentrations of organic carbon in soils favor the development of reducing conditions that result in As release in the aquifer. Flat surface topography indicates a low hydraulic gradient, sluggish groundwater flow, and extended sediment-water interactions (Nath et al., 2005). Young alluvial sediments, such as a high fluvisols probability, indicate the presence of Holocene sediments, which have been associated with high As concentrations (McArthur et al., 2011).

The values of the predictor variables were extracted at the centroid location of each grid pixel containing known (for model development) and unknown (for prediction) As concentrations. During the development of the random forest model, the environmental predictors were statistically evaluated (Pearson correlation coefficient) and verified the statistical significance (p-value <0.05) of the association with the percentage of grid-averaged As values exceeding 10 μg/L. This was done by organizing the data sets into 12 bins, each containing an identical number of observations. The percentage of As measurements >10 μg/L in each bin was then calculated. Sturges' formula ($1 + \log 2n$) was adopted to determine the optimal number of bins (Sturges, 1926). The statistical significance of the predictors (mean values) in each bin and the percentage of As concentrations >10 μg/L were checked before being included in the model. During model development, the data set was randomly divided into training (80%) and testing (20%) samples by preserving the ratio of high and low As values through stratified sampling.

3.4. Modeling and Validation

The random forest model was developed in the Python programming language using the "Scikit-learn 1.0.1" package (Pedregosa et al., 2011). The scikit-learn package is efficient and straightforward in predictive data analysis, and the user can tune several hyperparameters to improve the model's accuracy and efficiency. We developed the model by growing 1,000 trees and five predictors for each tree using a training data set. The testing data set was used to crossvalidate the model and determine the accuracy in the prediction of low (≤10 μg/L) and high (>10 μg/L) As concentrations. The probability values of the occurrence of an As concentration >10 μg/L in groundwater for the entire study area were determined by applying the model to 18 spatially continuous and statistically significant predictor variables.

A mean decrease in accuracy and a mean decrease in Gini scores were computed during the model development of each tree grown to evaluate the importance of predictor variables. The mean decrease accuracy and mean decrease Gini score were higher for the more important predictor variables used in the model. Gini node impurity, a measure of misclassification, was calculated for every split based on the probability of samples belonging to a single class. Predictors and the corresponding criterion that provided the best splits (i.e., the most information gained, or the least randomness attained) in the samples were selected first, which was based on the values of the lowest Gini score, that is, the largest difference in Gini scores before and after the splits. Gini purity indicates the homogeneity of the samples obtained after a split by a specific variable condition (Breiman et al., 1984). The importance of the variable increases as the Gini impurity decreases. The decrease in accuracy is calculated on out-of-bag (OOB) samples by randomly mixing the data for a particular predictor in the OOB samples. A predictor variable is considered to be highly important if the model's accuracy suffers upon removal of that variable.

Accuracy, sensitivity (i.e., true positive rate), and specificity (i.e., true negative rate) values were also calculated. A cutoff value was used to distinguish between high and low As hazard areas. The cutoff value was selected as the value at which the sensitivity and specificity of the model become equal. The predictive power of the model was evaluated by using the area under the receiver operating characteristic (ROC) curve. The standard deviations of classification outcome from each tree during the development of the forest were estimated for analyzing the model uncertainty. A low standard deviation indicates the greater certainty or predictive power of the model. Pearson residuals were also calculated to test model under- and over-predictions.

3.5. Estimations of the Total Area and Exposed Population

Population data were collected from the WorldPop website for 2020 at a spatial resolution of 1 km (www.worldpop.org). The total land area and the population exposed to elevated As concentrations (>10 μg/L) in drinking water were computed after generating ternary risk maps from the modeled probability.

Three risk levels were considered based on the cutoff values. The cutoff value was determined after plotting the sensitivity and specificity values for probabilities between 0 and 1 (Figure 2). The cutoff point was chosen as the point at which the sensitivity and specificity values intersect. Such a method avoid bias toward either a true positive rate or false positive rate (i.e., predicting high or low As concentrations, respectively; Podgorski et al., 2020).

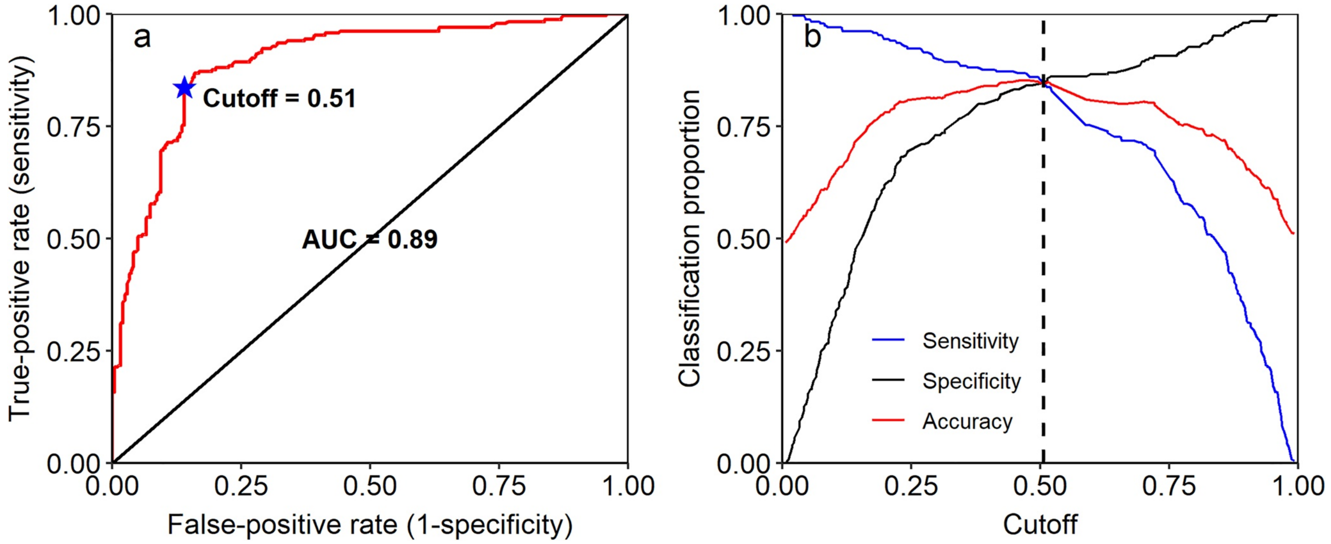

Figure 2. The classification strength. (a) ROC curve with an AUC of 0.89, which indicates the discriminative power of the random forest model. (b) True-positive rate (sensitivity), true-negative rate (specificity), and accuracy against the different cutoff values to identify the best conditions in predicting low and high As concentrations in the study area.

High (probability ≥0.7), moderate (probability > cutoff point but < 0.7), and low (probability < cutoff point) risk levels were adopted to identify the vulnerable zones (i.e., regions where the As exposure level is too high). This breakdown of vulnerable zones will aid in decision-making, with regard to prioritizing locations for immediate interventions, providing safe water access, and generating community awareness.

A three-level hazard area map was used to estimate the number of at-risk populations living in different subdistricts (i.e., community development blocks): the estimation was made by multiplying the total populations of each risk area with the modeled probabilities. The total at-risk populations were further refined based on the proportion of people living in urban and rural areas according to Integrated Management Information System (IMIS) reports (Integrated Management Information System, 2021) and the rate of use of untreated groundwater (0.25) in the study area. The rate of use of untreated groundwater use was determined based on surveys of homemade sand filters and whether such a filter lowered the As concentrations below 10 µg/L or not. We observed that 25% of the households were consuming drinking water with As concentrations >10 µg/L.

Table 2
The Confusion Matrix and Statistics of the Random Forest Model Developed Based on Eighteen Significant Predictor Variables

	Actual class	
Predicted class	0	1
0	208	36
1	37	198
Accuracy	0.848	
Cohen's kappa	0.695	
Sensitivity	0.846	
Specificity	0.849	
Positive predicted value	0.843	
Negative predicted value	0.852	
Prevalence	0.511	
AUC	0.89	

Note. The probability cutoff value is 0.51.

The coverage of piped water supply schemes (PWSS) was also assessed to better understand the percentage of households that have access to either safe or unsafe water in each subdistricts. The PWSS house connectivity data were retrieved directly from the IMIS reports (Integrated Management Information System, 2021).

4. Results and Discussion

4.1. Arsenic Prediction and the Strength of the Model Development

The results based on a 10-fold cross-validation on the test data set using the developed random forest model are provided in Table 2. The area under the ROC curve (AUC) was used to determine the strength of our binary (high and low) classification made by the random forest model. The AUC value generally ranges between 0.5 and 1. The values close to 1 indicate a perfect model, while a value of 0.5 indicates a no better than random chance occurrence. The AUC has been computed using different probability cutoff values (Fawcett, 2006). AUC values of 0.89 indicate excellent model performance and are comparable to several other groundwater quality predictions at both the regional and country scale. An AUC value of 0.84 was reported for fluoride

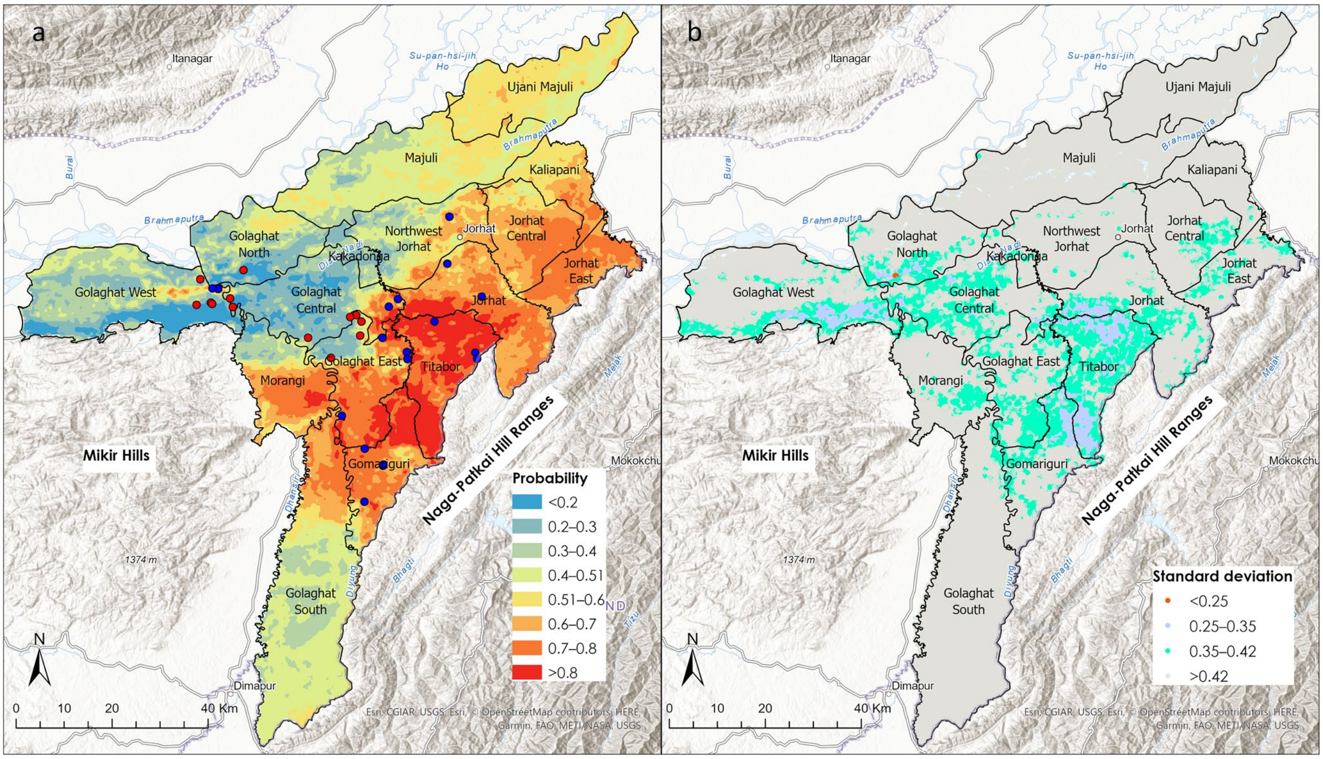

Figure 3. Maps showing the (a) Probability of As >10 μg/L in groundwater in the study area predicted by random forest model. The red filled circles represent grid locations of model underprediction, and blue filled circles represent grid locations of model overprediction based on Pearson residuals of predicted probability of As >10 μg/L, (b) Standard deviation of predicted class (either 0 or 1) in each grid location.

prediction in the entire country of India (Podgorski et al., 2018), while for As, AUC values of 0.71–0.83 have been reported in Gujarat (Wu et al., 2021), and 0.755 in Uttar Pradesh (Bindal & Singh, 2019). The test data set was also used to calculate our model's overall accuracy, which was 0.85, following 10-fold crossvalidation. In addition to that, the overall accuracy value was very close to the accuracy value computed with the OOB samples (0.83). The no information rate was 0.5042 (p-value $< 2.0 \times 10^{-16}$), which is significantly low compared to the accuracy obtained in this model. The no information rate is equivalent to the accuracy that could be achieved without a model (Podgorski et al., 2020). The no information rate is similar to the dominant class of the training data set, that is, the percentage of As ≤10 μg/L (50.5%) for this study. Cohen's kappa value was 0.695, indicating stronger reliability and substantial agreement of the model.

The conversion of As values from 3,600 point locations to spatially averaged 250-m grid locations produces high-density data points covering much of the study regions where 85% of the population resides. The grid-averaged 2,400 locations were found to distribute uniformly throughout the study area covering populated regions, suggesting any unlikely bias in the model prediction by providing excess weight to the environmental conditions in some areas (Podgorski et al., 2020). Such a situation can occur if the study area is very large, with an imbalance in sampling density (Podgorski et al., 2020). Compared to other studies, our study site is very small and limited to one major river basin (Bindal & Singh, 2019; Podgorski et al., 2020). Therefore, one single random forest model should effectively account for the heterogeneity in the geochemical environments and produce a reliable outcome. This reliable outcome has been made possible using large sets of predictor variables representing climate, hydrology, soil characteristics, and land use/land cover to define the different geochemical conditions, either favorable or unfavorable, for As release in the aquifer.

The predicted probability identified known As-rich areas near the foothill regions of the Naga-Patkai range (Figure 3a). The model also identified previously unknown areas, such as parts of the Golaghat South, Morangi, Jorhat East, Jorhat Central, and Kaliapani subdistricts. Chetia et al. (2011) reported As concentrations >10 μg/L in 8 out of 19 wells surveyed in Morangi subdistricts. Chetia (2010) also reported elevated As concentrations in the Borpathar and Sarupathar villages, located at the northern tip of the Golaghat South subdistricts, coinciding

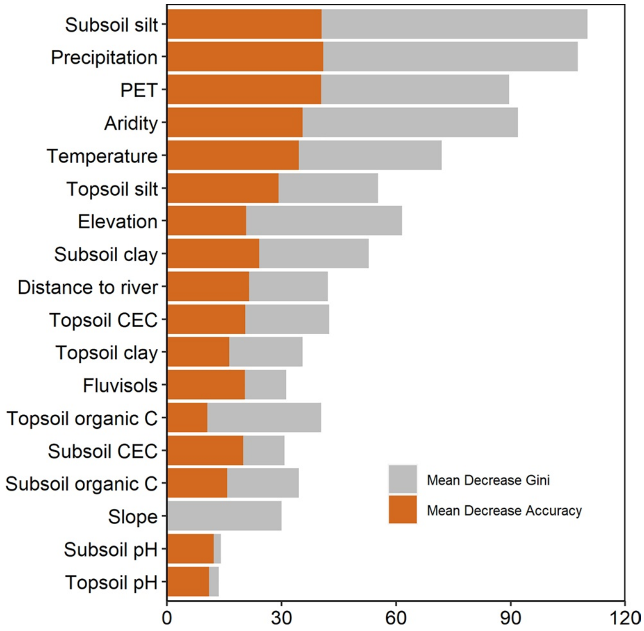

Figure 4. The importance of the predictor variables in relation to mean decrease in Gini and mean decrease in accuracy as calculated by the random forest model.

with areas of high predicted probabilities for As >10 μg/L. These areas lack sampling for As measurements, and no public geolocated records were available for use in the model. The model highlights the advantages of machine learning techniques that can handle large data sets with many predictors and still produce an excellent outcome by utilizing the statistical relationships of predictors and As concentrations (Podgorski et al., 2020; Podgorski & Berg, 2020). The predicted outcome can be effectively used in targeted well-testing campaigns in areas where the probability of As >10 μg/L is high to confirm potential threats to public health.

The standard deviation of the predicted classes for each of the 1,000 decision trees is shown in Figure 3b. The standard deviation values can be used to assess model uncertainty in the final predicted class. The data showed that the areas with a high predicted probability of As >10 μg/L have a low standard deviation in the predicted class. At the same time, the areas with slightly higher standard deviations are associated with medium to high predicted probability. These areas are in the Majuli, Ujani Majuli, and Golaghat South subdistricts, where the sampling for As in groundwater is lacking. Goswami et al. (2014) did report As concentrations >10 μg/L in Majuli and Ujani Majuli subdistricts. They observed isolated regions of elevated As concentrations that coincide with the regions of moderate to high probabilities determined by our model. However, due to the unavailability of location information, we could not incorporate these data into our model.

Pearson residuals of the predicted probability of As >10 μg/L were also used to identify the locations of model under and overpredictions. The results showed that the predictions made by the model show good agreement since the Pearson residuals in most of the grid locations were within the considered range (i.e., Pearson residuals between <2 and >−2) (Ayotte et al., 2006). But in a handful of grid locations, the Pearson residuals were either >2 or < −2, suggesting slight over and underpredictions by the model, respectively (Ayotte et al., 2006). These grid locations could form outliers within high and low As sites (i.e., low-high and high-low outliers) since the subsoil silt and topsoil clay concentrations (two important predictors) are comparable to the areas predicted to contain high and low As concentrations, respectively.

4.2. Association Between Environmental Predictors and the Spatial Distribution of Arsenic

The Gini impurity and accuracy scores averaged over all trees grown in the forest were used to determine the importance of the predictor variables (Figure 4). The results showed greater importance of subsoil silt and all climate variables (i.e., precipitation, potential evapotranspiration, aridity, and temperature) in the model prediction. In addition to those variables, topsoil silt, elevation, and subsoil clay showed strong importance in the model prediction. The least important variables for model prediction were subsoil pH, topsoil pH, and slope.

The strength of the relationships between predictor variables (the average in each bin) and As concentrations (the percentage of As measurements >10 μg/L) was tested using Pearson correlation coefficient (R) and p-value (95% confidence level) (Figure 5). The strongest R values were found with topsoil clay, subsoil clay, topsoil silt, subsoil silt, topsoil organic carbon, subsoil organic carbon, topsoil cation exchange capacity, subsoil cation exchange capacity, elevation, and fluvisols (Figure 5). These variables had R values >0.88 (either positive or negative) and showed statistical significance at p-values <0.05. In addition, precipitation, temperature, potential evapotranspiration, topsoil pH, and distance to the river also showed stronger relationships with the percentage of As measurements >10 μg/L. Subsoil and topsoil sand showed well-defined troughs, with the lowest proportion of As measurements >10 μg/L coinciding with mean values of 38 and 34 weight %, respectively (Figures 5i and 5j). These variables were also tested in developing the final model since the random forest algorithm is highly effective in capturing nonlinear relationships between predictor and response variables (Ryo & Rillig, 2017).

Our model depends strongly on the subsoil silt content, and a negative relationship with elevated As concentrations signifies the occurrence of high As concentrations in older floodplains. High subsoil silt content was mostly

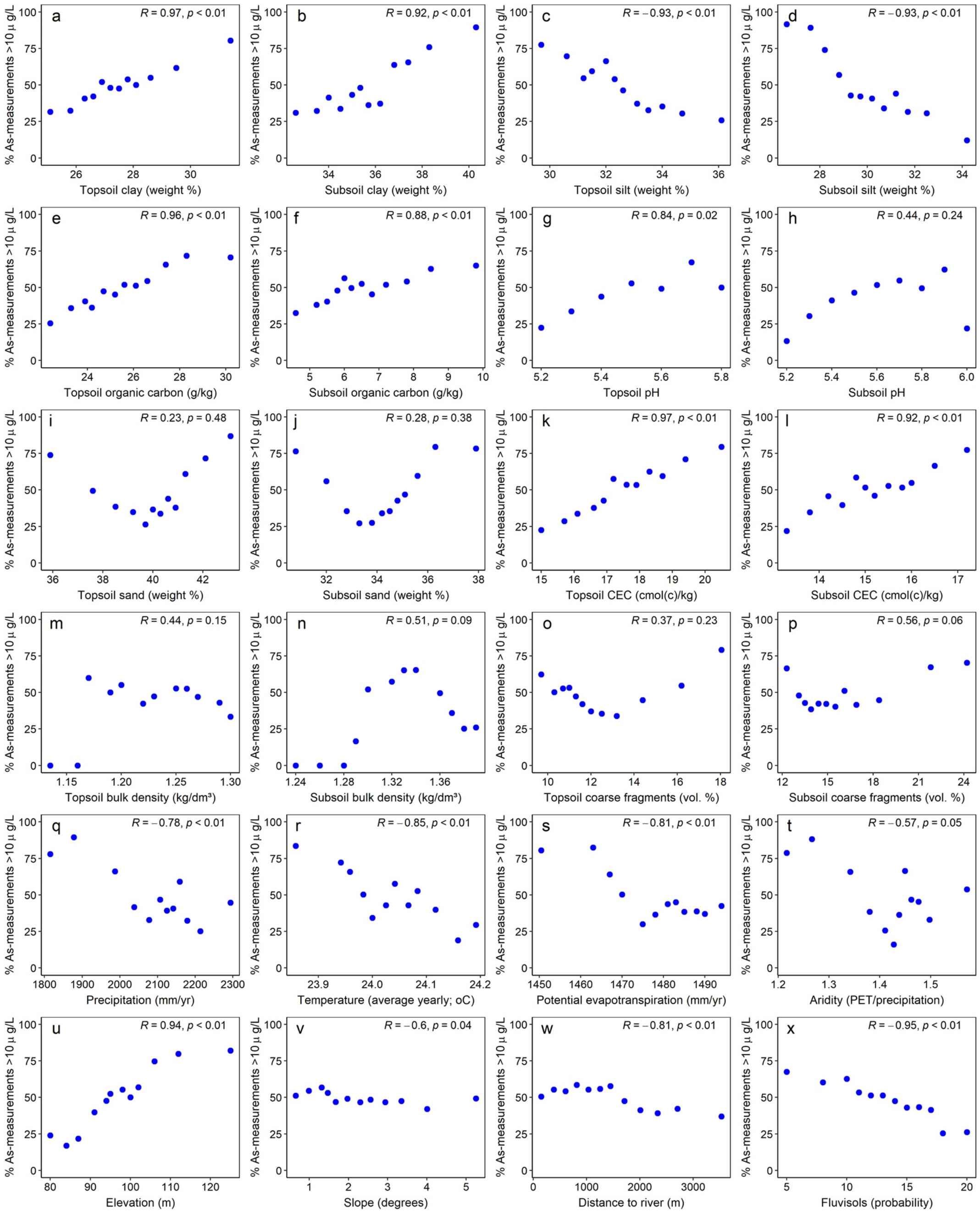

Figure 5. (a)–(x) The relationships between predictor variables and percentages of grid-averaged As concentrations >10 µg/L in 12 equally sized bins (except for topsoil and subsoil pH). Pearson correlation coefficients (*R*) with a statistically significant p-value (95% confidence level) are shown.

found to localize in the areas close to the current river channel (Figures S1 and S2 in Supporting Information S1). Higher silt content in soils indicates the deposition of fresh sediments by the river (Ahmed et al., 2004). Silt is highly reactive and produced by mechanical weathering; therefore, the presence of high silt content indicates more active sorption sites for As adsorption (Amini et al., 2008) and lack of As in the water. However, we observed a higher occurrence of subsoil clay further away from the current river channel and found it to be associated with elevated As concentrations(Figures S3 and S4 in Supporting Information S1). The observation of greater clay fractions and clay capping to greater aquifer depths supports the hypothesis of As leaching to groundwater and the lack of monsoonal dilution by rainwater. Such an observation increases the chance of extensive sediment-water

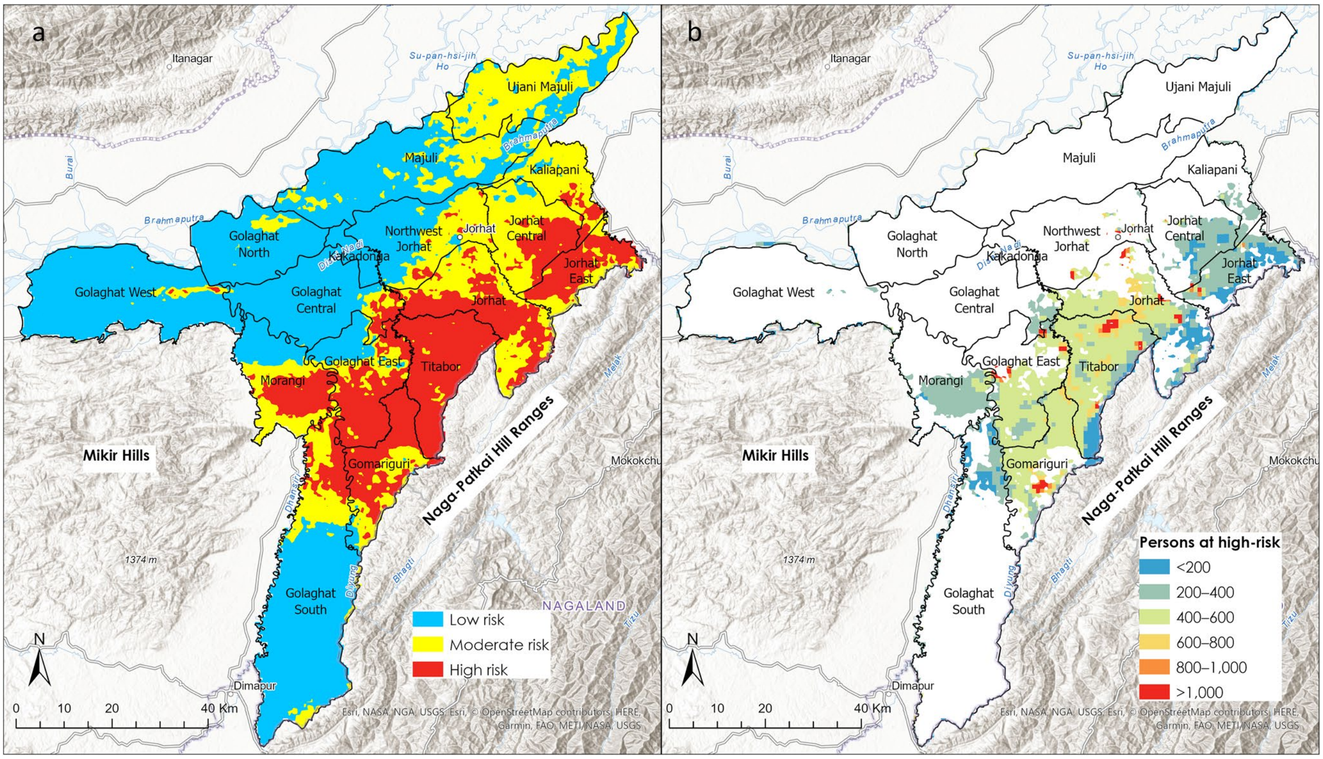

Figure 6. (a) Ternary As hazard probability map based on the predicted probability (high risk: ≥0.7, moderate risk: 0.51–0.7, and low risk: <0.51). (b) The density of population (per km^2) living in high-risk areas.

interactions and minimal flushing (Choudhury et al., 2018). Therefore, the composition of the aquifer materials is vital to the occurrence of As in groundwater (Bindal & Singh, 2019).

Our model shows the high importance of climate variables, which is consistent with the findings of Podgorski and Berg (2020), who suggested that climatic factors control the release of As in aquifers. Low precipitation and evapotranspiration are associated with a high percentage of As >10 μg/L in tube wells, suggesting less recharge and greater sediment-water interactions. Low precipitation favors the reduced flow of groundwater in the aquifer, and the lack of dilution effects by infiltrating rainwater increases the likelihood of As accumulation in groundwater over an extended time (Rodríguez et al., 2004). The combined effect of precipitation and evapotranspiration could favor As release due to saturated conditions and the formation of reducing environments for As release (Podgorski & Berg, 2020). The lack of flushing and dilution, as suggested by the higher radiocarbon ages of groundwater in high As sites, is consistent with the preservation of a larger pool of exchangeable As and, therefore, the maintenance of higher As concentrations (Choudhury et al., 2018).

Elevation shows moderate importance in our model (Figure S5 in Supporting Information S1). However, the relationship between elevation and the percentage of As >10 μg/L is highly significant and is positively correlated. This finding suggests that the high As in groundwater is occurring in elevated areas, that is, further away from the river channel, which is in contrast to the association between low elevation and As concentrations >10 μg/L in Bangladesh (Shamsudduha et al., 2008). Interestingly, our previous studies showed the occurrence of brown (oxidized) sands near the river channel, suggesting flushing of the aquifer near the river, while reduced (gray) sands were observed in As-enriched elevated terrain bordering the Naga-Patkai Hill ranges (Choudhury et al., 2018).

4.3. Arsenic-Hazard Probabilistic Zones

The estimated As risk areas and population exposed were computed based on the probability cutoff of 0.51 (Figure 6). The probability of As >10 μg/L was grouped into three categories (high, moderate, and low-risk zones). The population map shows a cluster of densely populated localities within the high-risk zone (Figure 6b),

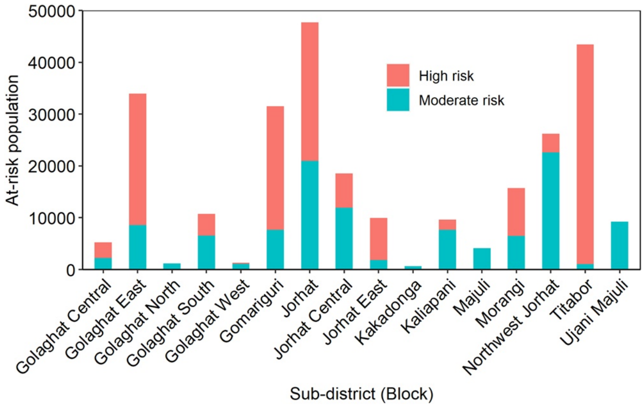

Figure 7. The total population, living in high and moderate risk areas, potentially exposed to As concentrations >10 μg/L in sixteen subdistricts of the study area.

indicating a potential significant public health concern. The high-risk zones are predominantly located south of the Brahmaputra River and beside the Naga-Patkai foothill ranges. The moderate risk areas typically extend from the high-risk zones toward the river, including the island of Majuli. The low-risk zone is mainly located near the river channel, yet isolated low-risk pockets are located in moderate-risk zones. In addition to that, a large part of the Golaghat South subdistrict is within the low-risk zone, which is bordered by Naga-Patkai hill ranges on the east and Mikir hills on the west. These areas are located at a higher elevation with greater slopes, and are thus favorable for greater aquifer flushing and unfavorable for As release to groundwater. Similarly, a study by Puzari et al. (2015) and NPCB (2018) reported As concentrations <10 μg/L in Dimapur, Nagaland, located at the southern tip of the Golaghat South subdistrict (Figure 3a). These studies support our prediction of a low probability of As >10 μg/L in groundwater in these areas.

The total population potentially exposed to As concentrations >10 μg/L is shown for each subdistrict (Figure 7). Based on the probability (a cutoff score of 0.51), a total of approximately 115,000 people in moderate-risk areas (probability scores between 0.51 and 0.70) and a further 155,000 people in high-risk areas (probability scores ≥0.70) may be directly or indirectly exposed to As concentrations >10 μg/L. Most of the exposed populations in high-risk areas live in four subdistricts: Titabor (a total of 42,000 people), Jorhat (a total of 26,750 people), Golaghat East (a total of 25,500 people), and Gomariguri (a total of 24,000 people). The people in these four subdistricts require immediate public health interventions: targeted well-testing and the provision of safe water access.

5. Public Health Implications

The results of this study effectively demarcate the high- and low-risk areas in the two most affected districts of Assam, as well as the moderate-risk areas in the district of Majuli, whose inhabitants are relatively poor. The hazard probability map and risk model of the occurrence of As concentrations in groundwater exceeding the WHO guideline of 10 μg/L will be helpful for policymakers in managing the aquifer sustainably, since agricultural use of groundwater is negligible compared to the As-contaminated areas in Bangladesh (CGWB, 2013; Zahid & Ahmed, 2006). In the short term, local administration can target the regions having the greatest As risk (probability ≥0.70) and can educate local inhabitants on the potential health effects of consuming high As drinking water. Moreover, a targeted well-testing campaign could help identify wells contaminated with As. Such a campaign would inform inhabitants as to whether they need to switch their drinking water sources to a safer tube well to reduce As exposure. Well-testing and subsequent well-switching have been effective in lowering population-level As exposure in Bangladesh (Jamil et al., 2019; van Geen et al., 2002). In the long term, policymakers, with support from the Government of India or other sources, could target exposed regions by implementing additional water treatment plants.

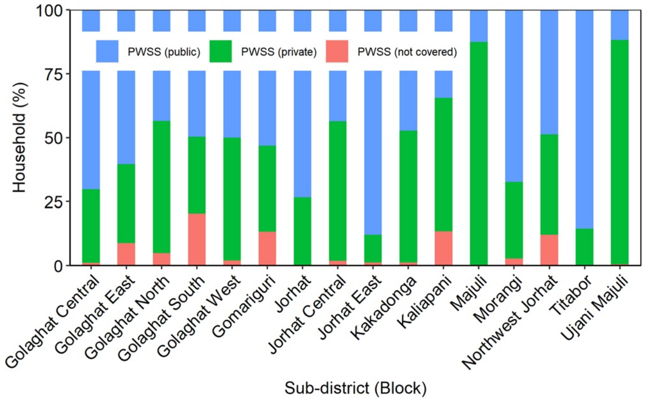

Figure 8. Percent household access to treated water through a piped water supply schemes (network public and private connections) and no access to treated water in sixteen subdistricts of the study area.

Based on the analysis of PWSS providing treated drinking water by the government, we identified that the high-risk subdistricts Gomariguri, Morangi, Golaghat East, and Golaghat South in Golaghat district are the least connected to household tap waters (Figure 8). Likewise, the Titabor, Jorhat East, and Jorhat subdistricts in Jorhat districts are the least connected to household tap waters. These findings suggest a greater risk of As exposure in rural households in high-risk areas due to the lack of household tap water connection. In the two moderate-risk subdistricts, Majuli and Ujani Majuli, more than 85% of households are connected to PWSS tap water that provides safe water access (Figure 8). This is a significant achievement on the part of the government in providing safe water access to rural inhabitants. Though treated piped water is provided through public tap water connections in most high-risk subdistricts, it is possible that a lack of awareness regarding the use of treated water and the distance to public tap water may increase the likelihood that private tube wells are installed to meet water needs. Therefore, it is important to increase awareness among the public in these high-risk subdistricts that it is possible to gain access to this safe drinking water via household connection. The local government could subsidize the costs associated with house connections, as many rural inhabitants might not be able to afford the costs. By doing so, the local governments would be able to accurately address the public health concerns and sustainably use groundwater by discouraging the installation of private tube wells.

6. Conclusion

The hazard probability map presented in this study provides insights into the locations and areas where inhabitants are potentially being exposed to elevated As concentrations in groundwater. The hazard probability map is also useful for finding appropriate locations to install community drinking water wells and treatment facilities to provide safe water access. The hazard probability map can inform policymakers on targeted well-testing campaigns for mitigation and highlights where inhabitants must consider testing their wells for As contamination. The habitation-level predictive model can be used to inform villagers and generate community awareness about the potential impact of elevated As in groundwater through active participation. As mitigation strategies, we identified areas where authorities must consider providing safe water connections to rural households.

The availability of groundwater As testing data from newer areas would undoubtedly help strengthen the predictive power of the model, particularly data from Morangi, Golaghat South, Kaliapani, and Majuli, where the data are almost absent for training the model. The hazard probability map presented here does not account for the role of aquifer depths in relation to the spatial pattern of As. In general, As concentration was found to vary with sediment age and depth of the tube wells. In addition, time could be incorporated into the model since As concentrations in tube wells change with the seasons. A spatio-temporal model could provide further insights into understanding aquifer contamination and modeling the threats to public health.

Conflict of Interest

The authors declare no conflicts of interest relevant to this study.

Data Availability Statement

The data and source code are available at Hydroshare and can be freely downloaded from https://doi.org/10.4211/hs.d4f4b7601c694667bdf62a7826cad1a6 (Nath, 2022).

Acknowledgments
The authors thank the Department of Geography and Environmental Science, Hunter College of the City University of New York, for providing the ArcGIS Pro 2.5 software package. This study is funded by NASA Grant No. 80NSSC20K1718.

References

Ahmed, K. M., Bhattacharya, P., Hasan, M. A., Akhter, S. H., Alam, S. M. M., & Bhuyian, M. A. H., et al (2004). Arsenic contamination in groundwater of alluvial aquifers in Bangladesh: An overview. *Applied Geochemistry*, *19*(2), 181–200. https://doi.org/10.1016/j.apgeochem.2003.09.006

Amini, M., Abbaspour, K. C., Berg, M., Winkel, L., Hug, S. J., Hoehn, E., et al. (2008). Statistical modeling of global geogenic arsenic contamination in groundwater. *Environmental Science and Technology*, *42*(10), 3669–3675. https://doi.org/10.1021/es702859e

Ayotte, J. D., Nolan, B. T., Nuckols, J. R., Cantor, K. P., Robinson, G. R., Baris, D., et al. (2006). Modeling the probability of arsenic in groundwater in New England as a tool for exposure assessment. *Environmental Science and Technology*, *40*(11), 3578–3585. https://doi.org/10.1021/es051972f

Bangladesh Bureau of Statistics (BBS) and United Nations Children's Fund (UNICEF). (2014). *Bangladesh multiple indicator cluster survey 2012–2013. Progotir pathey. Key district level findings*. BBS and UNICEF Bangladesh.

Bindal, S., & Singh, C. K. (2019). Predicting groundwater arsenic contamination: Regions at risk in highest populated state of India. *Water Research*, *159*, 65–76. https://doi.org/10.1016/j.watres.2019.04.054

Borah, K., Bhuyan, B., & Sarma, H. P. (2009). Lead, arsenic, fluoride, and iron contamination of drinking water in the tea garden belt of Darrang district, Assam, India. *Environmental Monitoring and Assessment*, *169*(1–4), 347–352. https://doi.org/10.1007/s10661-009-1176-2

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Chakraborti, D., Rahman, M. M., Paul, K., Chowdhury, U. K., Sengupta, M. K., Lodh, D., et al. (2002). Arsenic calamity in the Indian subcontinent what lessons have been learned? *Talanta*, *58*(1), 3–22. https://doi.org/10.1016/s0039-9140(02)00270-9

Choudhury, R., Nath, B., Khan, M. R., Mahanta, C., Ellis, T., & van Geen, A. (2018). *Impact of aquifer flushing across a 35 Km transect perpendicular to the upper Brahmaputra river in Assam, India*. Water Resources Research. https://doi.org/10.1029/2017WR022485

CGWB. (2013). *Central ground water board, ministry of water resources government of India, ground water information booklet. Golaghat and Jorhat districts*. Retrieved from http://cgwb.gov.in/

Chetia, M. (2010). *Water quality in Golaghat district of Assam India with special reference to arsenic contamination and its mitigation* [Ph.D. thesis, Gauhati University]. http://hdl.handle.net/10603/68234

Chetia, M., Chatterjee, S., Banerjee, S., Nath, M. J., Singh, L., Srivastava, R. B., & Sarma, H. P. (2011). Groundwater arsenic contamination in the Brahmaputra River basin: A water quality assessment in the Golaghat (Assam), India. *Environmental Monitoring and Assessment*, *173*(1–4), 371–385. https://doi.org/10.1007/s10661-010-1393-8

DPHE. (2001). *Arsenic contamination of groundwater in Bangladesh. Department of Public Health Engineering, British geological survey, and mott MacDonald*. BGS Technical Report WC/00/19 (4 volumes).

Enmark, G., & Nordborg, D. (2007). Arsenic in the groundwater of the Brahmaputra floodplains, Assam, India–Source, distribution and release mechanisms. Minor Field Study. (Vol. *131*, p. 35p). Committee of tropical ecology, Uppsala University (ISBN: 1653-5634).

Farr, T. G., & Kobrick, M. (2000). Shuttle radar topography mission produces a wealth of data [dataset]. Eos, Transactions, AGU, *81*, 583. https://doi.org/10.1029/eo081i048p00583

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Fick, S. E., & Hijmans, R. J. (2017). WorldClim 2: New 1km spatial resolution climate surfaces for global land areas [dataset]. International Journal of Climatology, *37*, 4302–4315. https://doi.org/10.1002/joc.5086

Flanagan, S. V., Johnston, R. B., & Zheng, Y. (2012). Arsenic in tube well water in Bangladesh: Health and economic impacts and implications for arsenic mitigation. *Bulletin of the World Health Organization*, *90*(11), 839–846. https://doi.org/10.2471/blt.11.101253

Foster, S. S. D., & Chilton, P. J. (2003). Groundwater: The processes and global significance of aquifer degradation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *358*(1440), 1957–1972. https://doi.org/10.1098/rstb.2003.1380

Goswami, R., Rahman, M. M., Murrill, M., Sarma, K. P., Thakur, R., & Chakraborti, D. (2014). Arsenic in the groundwater of Majuli – the largest river island of the Brahmaputra: Magnitude of occurrence and human exposure. *Journal of Hydrology*, *518*, 354–362. https://doi.org/10.1016/j.jhydrol.2013.09.022

Ho, T. K. (1995). Random decision forests. *Proceedings of the 3rd International Conference on Document Analysis and Recognition*, *1*, 278–282. Montreal, QC, Canada

Integrated Management Information System. (2021). *Jal jeevan mission reports*. Retrieved from https://ejalshakti.gov.in/IMISReports/IMISReportLogin.aspx

Hengl, T. (2018). Global DEM derivatives at 250 m, 1 km and 2 km based on the MERIT DEM (Version 1.0) [Dataset]. Retrieved from https://zenodo.org/record/1447210#.YftJvOrMKUk

Jamil, N. B., Feng, H., Ahmed, K. M., Choudhury, I., Barnwal, P., & van Geen, A. (2019). Effectiveness of different approaches to arsenic mitigation over 18 years in Araihazar, Bangladesh: Implications for national policy. *Environmental Science and Technology*, *53*(10), 5596–5604. https://doi.org/10.1021/acs.est.9b01375

Lehner, B., Verdin, K., & Jarvis, A. (2008). New global hydrography derived from spaceborne elevation data [Dataset]. Eos, Transactions, AGU, *89*, 93–94. https://doi.org/10.1029/2008eo100001

Mahanta, C., Choudhury, R., Basu, S., Hemani, R., Dutta, A., Barua, P. P., & Saikia, L. (2015). Preliminary assessment of arsenic distribution in Brahmaputra River basin of India based on examination of 56,180 public groundwater wells. In *Safe and sustainable use of arsenic-contaminated aquifers in the Gangetic Plain* (pp. 57–64). Springer Publishers. https://doi.org/10.1007/978-3-319-16124-2_4

McArthur, J. M., Banerjee, D. M., Hudson-Edwards, K. A., Mishra, R., Purohit, R., Ravenscroft, P., et al. (2004). Natural organic matter in sedimentary basins and its relation arsenic in anoxic ground water: The example of West Bengal and its worldwide implications. *Applied Geochemistry*, *19*(8), 1255–1293. https://doi.org/10.1016/j.apgeochem.2004.02.001

McArthur, J. M., Nath, B., Banerjee, D. M., Purohit, R., & Grassineau, N. (2011). Palaeosol control on groundwater flow and pollutant distribution: The example of arsenic. *Environmental Science and Technology*, *45*(4), 1376–1383. https://doi.org/10.1021/es1032376

Mukherjee, A., Sarkar, S., Chakraborty, M., Duttagupta, S., Bhattacharya, A., Saha, D., et al. (2021). Occurrence, predictors and hazards of elevated groundwater arsenic across India through field observations and regional-scale AI-based modeling. *Science of the Total Environment*, *759*, 143511. https://doi.org/10.1016/j.scitotenv.2020.143511

Nagaland Pollution Control Board (NPCB). (2018). *Water monitoringNational water quality monitoring programme (NWMP) stations in Nagaland*. Retrieved from https://npcb.nagaland.gov.in/water-monitoring/

Nath, B., Berner, Z., Basu Mallik, S., Chatterjee, D., Charlet, L., & Stueben, D. (2005). Characterization of aquifers conducting groundwaters with low and high arsenic concentrations: A comparative case study from West Bengal, India. *Mineralogical Magazine*, *69*(5), 841–854. https://doi.org/10.1180/0026461056950292

Nath, B., Stüben, D., Basu Mallik, S., Chatterjee, D., & Charlet, L. (2008). Mobility of arsenic in West Bengal aquifers conducting low and high arsenic. Part 1: Comparative hydrochemical and hydrogeological characteristics. *Applied Geochemistry*, *23*(5), 977–995. https://doi.org/10.1016/j.apgeochem.2007.11.016

Nath, B., Chakraborty, S., Burnol, A., Stueben, D., Chatterjee, D., & Charlet, L. (2009). Mobility of arsenic in the sub-surface environment: An integrated hydrogeochemical study and sorption model of the sandy aquifer materials. *Journal of Hydrology*, *364*(3–4), 236–248. https://doi.org/10.1016/j.jhydrol.2008.10.025

Nath, B. (2022). Predictive_model_Assam, HydroShare [Dataset/Software]. https://doi.org/10.4211/hs.d4f4b7601c694667bdf62a7826cad1a6

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-Learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Podgorski, J. E., Labhasetwar, P., Saha, D., & Berg, M. (2018). Prediction modeling and mapping of groundwater fluoride contamination throughout India. *Environmental Science and Technology*, *52*(17), 9889–9898. https://doi.org/10.1021/acs.est.8b01679

Podgorski, J., Wu, R., Chakravorty, B., & Polya, D. A. (2020). Groundwater arsenic distribution in India by machine learning geospatial modeling. *International Journal of Environmental Research and Public Health*, *17*(19), 7119. https://doi.org/10.3390/ijerph17197119

Podgorski, J., & Berg, M. (2020). Global threat of arsenic in groundwater. *Science*, *368*(6493), 845–850. https://doi.org/10.1126/science.aba1510

Poggio, L., de Sousa, L. M., Batjes, N. H., Heuvelink, G. B. M., Kempen, B., Ribeiro, E., & Rossiter, D. (2021). SoilGrids 2.0: Producing soil information for the globe with quantified spatial uncertainty[dataset]. Soils, *7*, 217–240. https://doi.org/10.5194/soil-7-217-2021

Puzari, A., Khan, P., Thakur, D., Kumar, M., Shanu, K., Chutia, P., & Ahmed, Z. (2015). Quality assessment of drinking water from Dimapur district of Nagaland and Karbi-Anglong district of Assam for possible related health hazards. *Current World Environment*, *10*(2), 634–640. https://doi.org/10.12944/cwe.10.2.29

Ravenscroft, P., Brammer, H., & Richards, K. S. (2009). *Arsenic pollution: A global synthesis*. Wiley-Blackwell.

Rahman, M. M., Asaduzzaman, M., & Naidu, R. (2011). Arsenic exposure from rice and water sources in the Noakhali district of Bangladesh. *Environmental Geochemistry and Health*, *3*, 1–10. https://doi.org/10.1007/s12403-010-0034-3

Rokach, L., & Maimon, O. (2008). *Data mining with decision trees: Theory and applications*. World Scientific Pub Co Inc. (ISBN: 978-9812771711).

Rodríguez, R., Ramos, J. A., & Armienta, A. (2004). Groundwater arsenic variations: The role of local geology and rainfall. *Applied Geochemistry*, *19*(2), 245–250. https://doi.org/10.1016/j.apgeochem.2003.09.010

Roy, P. S., Meiyappan, P., Joshi, P. K., Kale, M. P., Srivastav, V. K., Srivastava, S. K., et al. (2016). Decadal Land Use and Land Cover Classifications across India, 1985, 1995, 2005 [Dataset]. ORNL DAAC. https://doi.org/10.3334/ORNLDAAC/1336

Ryo, M., & Rillig, M. C. (2017). *Statistically reinforced machine learning for nonlinear patterns and variable interactions*. Ecosphere. https://doi.org/10.1002/ecs2.1976

Saha, K. C. (1995). Chronic arsenic dermatosis from tubewell water in West Bengal during 1983–87. *Indian Journal of Dermatology*, *40*, 1–12.

Shamsudduha, M., Uddin, A., Saunders, J. A., & Lee, M. K. (2008). Quaternary stratigraphy, sediment characteristics and geochemistry of arsenic-contaminated alluvial aquifers in the ganges-brahmaputra floodplain in Central Bangladesh. *Journal of Contaminant Hydrology*, *99*(1–4), 112–136. https://doi.org/10.1016/j.jconhyd.2008.03.010

Singh, A. K. (2004). Arsenic contamination in groundwater of North eastern India. Paper presented at Proceedings of 11th national symposium on hydrology with focal theme on water quality. National Institute of Hydrology. (pp. 255–262).

Smith, A. H., Lingas, E. O., & Rahman, M. (2000). Contamination of drinking-water by arsenic in Bangladesh: A public health emergency. *Bulletin of the World Health Organization*, *78*, 1093–1103.

Sturges, H. A. (1926). The choice of a class interval. *Journal of the American Statistical Association*, *21*(153), 65–66. https://doi.org/10.1080/01621459.1926.10502161

Trabucco, A., & Zomer, R. J. (2018). Global aridity index and potential evapo-transpiration (ET0) climate database v2; CGIAR consortium for spatial information (CGIAR-CSI) [Dataset]. https://doi.org/10.6084/m9.figshare.7504448.v2

van Geen, A., Ahsan, H., Horneman, A. H., Dhar, R. K., Zheng, Y., Hussain, I., et al. (2002). Promotion of well-switching to mitigate the current arsenic crisis in Bangladesh. *Bulletin of the World Health Organization*, *80*, 732–737.

van Geen, A., Zheng, Y., Versteeg, R., Stute, M., Horneman, A., Dhar, R. K., et al. (2003). Spatial variability of arsenic in 6000 tubewells in a 25 km^2 area of Bangladesh. *Water Resources Research*, *39*(5), 1140. https://doi.org/10.1029/2002wr001617

Verma, S., Mukherjee, A., Mahanta, C., Choudhury, R., & Mitra, K. (2016). Influence of geology on groundwater-sediment interactions in varied arsenic enriched tectono-morphic aquifers of the Brahmaputra River basin. *Journal of Hydrology*, *540*, 176–195. https://doi.org/10.1016/j.jhydrol.2016.05.041

Wu, R., Podgorski, J., Berg, M., & Polya, D. A. (2021). Geostatistical model of the spatial distribution of arsenic in groundwaters in Gujarat State, India. *Environmental Geochemistry and Health*, *43*(7), 2649–2664. https://doi.org/10.1007/s10653-020-00655-7

Zahid, Z., & Ahmed, S. R. U. (2006). *Groundwater resources development in Bangladesh: Contribution to irrigation for food security and constraints to sustainability, No H039306* (pp. 25–46). IWMI Books, Reports, International Water Management Institute.