Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

Geogenic manganese and iron in groundwater of Southeast Asia and Bangladesh – Machine learning spatial prediction modeling and comparison with arsenic

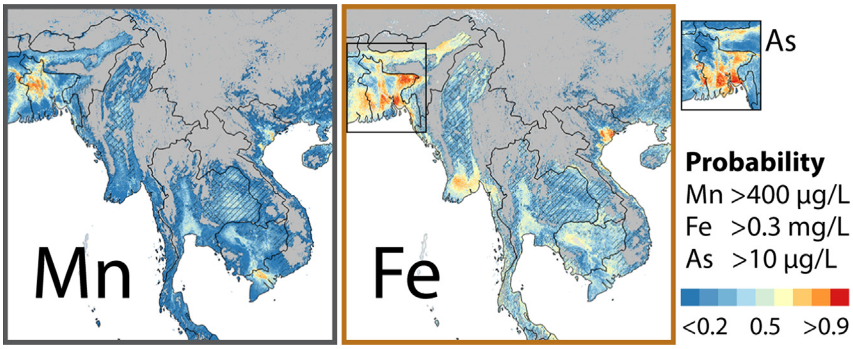
Joel Podgorski *, Dahyann Araya, Michael Berg

Eawag, Swiss Federal Institute of Aquatic Science and Technology, Department Water Resources and Drinking Water, 8600 Dübendorf, Switzerland

HIGHLIGHTS

- Machine learning modeling Mn/Fe in groundwater in Southeast Asia and Bangladesh
- Hazard of high Mn concentrations immediately adjacent to areas with high arsenic
- Mn/Fe prediction maps highlight health risks and are useful for water providers.

GRAPHICAL ABSTRACT



ARTICLE INFO

Editor: Pauline Smedley

Keywords:
Groundwater quality
Drinking water
Human health
Random forest modeling
Generalized boosted regression modeling
Reducing groundwater

ABSTRACT

Naturally occurring, geogenic manganese (Mn) and iron (Fe) are frequently found dissolved in groundwater at concentrations that make the water difficult to use (deposits, unpleasant taste) or, in the case of Mn, a potential health hazard. Over 6000 groundwater measurements of Mn and Fe in Southeast Asia and Bangladesh were assembled and statistically examined with other physicochemical parameters. The machine learning methods random forest and generalized boosted regression modeling were used with spatially continuous environmental parameters (climate, geology, soil, topography) to model and map the probability of groundwater Mn > 400 μg/L and Fe > 0.3 mg/L for Southeast Asia and Bangladesh. The modeling indicated that drier climatic conditions are associated with a tendency of elevated Mn concentrations, whereas high Fe concentrations tend to be found in a more humid climate with elevated levels of soil organic carbon. The spatial distribution of Mn > 400 μg/L and Fe > 0.3 mg/L was compared and contrasted with that of the critical geogenic contaminant arsenic (As), confirming that high Fe concentrations are often associated with high As concentrations, whereas areas of high concentrations of Mn and As are frequently found adjacent to each other. The probability maps draw attention to areas prone to elevated concentrations of geogenic Mn and Fe in groundwater and can help direct efforts to mitigate their negative effects. The greatest Mn hazard is found in densely populated northwest Bangladesh and the Mekong, Red and Ma River Deltas of Cambodia and Vietnam. Widespread elevated Fe concentrations and their associated negative effects on water infrastructure pose challenges to water supply. The Mn and Fe prediction maps demonstrate the value of machine learning for the geospatial prediction modeling and mapping of groundwater contaminants as well as the potential for further constituents to be targeted by this novel approach.

* Corresponding author.
E-mail address: joel.podgorski@eawag.ch (J. Podgorski).

http://dx.doi.org/10.1016/j.scitotenv.2022.155131
Received 29 October 2021; Received in revised form 25 March 2022; Accepted 5 April 2022
Available online 08 April 2022

1. Introduction

Groundwater is the primary source of drinking water for a large proportion of the world's population, with an estimated 2.5 billion people relying solely on it for drinking water (WWAP, 2015). Manganese, one of the most abundant metals in Earth's crust, can occur naturally in groundwater from the dissolution of manganese oxides, silicates and carbonates within rocks and soil. It varies in concentration around the world, and it is usually found associated with iron-bearing water (Kohl and Medlar, 2006).

Even at very low concentrations in drinking water, manganese and iron can damage infrastructure by forming coatings on water pipes, which can then flake off and make the taste and color of water unpleasant (Kohl and Medlar, 2006; Sly et al., 1990). At low concentrations, manganese is essential for human health; however, it has been associated with adverse health effects at higher concentrations. In 2004, the World Health Organization (WHO) recommended a health-oriented guideline value of 400 μg/L for manganese in drinking water and an aesthetic and taste-oriented threshold of 0.3 mg/L for iron (WHO, 2004). However, manganese exposure from water consumption is generally lower than from food consumption, and iron does not pose a threat to human health (WHO, 2003). The manganese value was discontinued in 2011 as a global reference to guide public water policies as it was found not to be a health threat in concentrations found in drinking water (WHO, 2011). However, the WHO does encourage countries to establish their own standards and regulations (WHO, 2017). For example, the US EPA lifetime health advisory, the USGS health-based screening level and the standard of the Indian Bureau of Standards all use 300 μg/L for manganese (EPA U, 2004). Nevertheless, there is an ongoing debate on the neurotoxic effects on human health from exposure to excessive levels of manganese in drinking water (Bouchard et al., 2007; Bouchard et al., 2011; Claus Henn et al., 2017; Haynes et al., 2015; Iyare, 2019; Kondakis et al., 1989; Rahman et al., 2021; Sahni et al., 2007; Schullehner et al., 2020; Wasserman et al., 2006; WHO, 2011; WHO, 2017; Woolf et al., 2002). For example, exposure to concentrations less than the 400 μg/L guideline have been reported to negatively impact the neurological development of children (Bouchard et al., 2011; Schullehner et al., 2020). Also, the consumption of manganese through drinking water can adversely affect neurological health, similarly to Parkinson's syndrome (Holzgraefe et al., 1986; Perl and Olanow, 2007; WHO, 2011).

Despite the health-related concerns and the high levels of manganese in drinking water that have been reported in Asia (Bacquart et al., 2012; Bacquart et al., 2015; Buschmann et al., 2008; Ghosh et al., 2020; Wasserman et al., 2006; Winkel et al., 2011), Africa (Amoako et al., 2011), South America (Carretero and Kruse, 2015; de Meyer et al., 2017), North America (Bouchard et al., 2007; Dion et al., 2018; Johnson et al., 2018; Spangler and Spangler, 2009), Europe (Homoncik et al., 2010; Kondakis et al., 1989; Roccaro et al., 2007) and Australia (Koppi et al., 1996), only a few isolated studies have spatially predicted areas prone to manganese contamination in groundwater (Erickson et al., 2021b; Johnson et al., 2018; Thapa et al., 2018). However, groundwater contamination modeling using environmental predictor variables has proven valuable in identifying potentially contaminated areas of concern, thereby helping how to prioritize the testing of groundwater sources (e.g. DeSimone et al., 2020; DeSimone and Ransom, 2021; Erickson et al., 2021a; Erickson et al., 2021b; Huang et al., 2021; Podgorski and Berg, 2020; Podgorski et al., 2020; Sajedi-Hosseini et al., 2018; Wu et al., 2021; Zhong et al., 2021).

Manganese and iron can be released from their constituent mineral phases in the aquifer matrix into groundwater along the same redox chain as arsenic (Buschmann et al., 2008; Van Geen et al., 2008; Ying et al., 2017). Since redox conditions may vary smoothly over broad regions, high manganese concentrations in groundwater may be found near areas with high arsenic concentrations in anoxic groundwater (de Meyer et al., 2017; Erickson et al., 2021b; Ying et al., 2017). This association between arsenic and manganese has been explored across Southeast Asia (Phan et al., 2019; Richards et al., 2017; Winkel et al., 2011; Ying et al., 2017), a global hotspot for arsenic. However, there are no studies for the region that explore this relationship spatially.

Here we investigate the geochemical and environmental conditions associated with high concentrations of manganese and iron in Southeast Asia and Bangladesh. We do so through comparison with other physicochemical parameters as well as by using machine learning (ML) to model these elements, utilizing both groundwater chemistry and environmental parameters. The occurrence of manganese and iron are then compared with that of the highly toxic geogenic contaminant arsenic.

2. Materials and methods

2.1. Groundwater chemistry measurements and geospatial predictor variables

Georeferenced measurements of manganese (Mn; n = 6122) and iron (Fe; n = 6107) in groundwater in Southeast Asia as well as 14 other concurrently measured physicochemical parameters were assembled from multiple published sources (Table S1). These other parameters were used to help examine the conditions leading to high concentrations of dissolved Mn/Fe and include: ammonium (NH_4), arsenic (As), bicarbonate (HCO_3), chloride (Cl), dissolved oxygen (O_2), electrical conductivity (EC), nitrate (NO_3), pH, phosphate (PO_4), redox potential (Eh), sodium (Na), sulfate (SO_4), water temperature and well depth. These groundwater quality measurements stem from five general areas: Bangladesh, the Irrawaddy Delta (Myanmar), Mekong Delta (Cambodia and Vietnam), Red and Ma River Deltas (northern Vietnam) and Sumatra (Indonesia). The Mn and Fe measurements are plotted in Fig. 1, and descriptive statistics are given in Table 1.

In addition, a total of 57 spatially continuous environmental parameters were also assembled from publicly available global datasets for use as predictors of high concentrations of manganese and iron (Table S2). These predictors generally fall into the categories of climate, geology, land use, soil properties or topography and were chosen based on their use in related studies for the prediction of manganese (DeSimone and Ransom, 2021; Erickson et al., 2021b) or arsenic (Ayotte et al., 2017; Bretzler et al., 2017; Erickson et al., 2021b; Podgorski and Berg, 2020; Podgorski et al., 2020; Podgorski et al., 2017; Wu et al., 2021). Nearly all of these variables were available with a resolution of either 7.5″ or 30″, which roughly correspond to 250 m and 1 km at the equator, respectively.

2.2. Analysis of groundwater chemistry and depth distribution

In order to help identify associations and potential geochemical relationships, Kendall rank correlations were calculated between manganese and iron and the 14 other groundwater parameters As, Cl, EC, Eh, HCO_3, Na, NH_4, NO_3, O_2, pH, PO_4, SO_4, temperature and well depth. As opposed to Pearson correlation, which quantifies the linearity between two parameters, a Kendall rank correlation instead evaluates their ranked order, such that a high rank correlation coefficient can ensue from two parameters that vary proportionally though non-linearly with each other. To explore the vertical dimension of the presence or absence of Mn, Fe and As, which lie along the same redox reaction chain, the moving averages of their proportions exceeding the WHO guidelines of Mn > 400 μg/L, Fe > 0.3 mg/L and As >10 μg/L (WHO, 2011) were plotted against well depth using a 20-m averaging window for each of the five main locations.

2.3. Machine-learning modeling

2.3.1. Relationships of physicochemical groundwater parameters

Machine learning (ML) was used to further explore relationships between Mn and Fe concentrations and the groundwater chemistry as well as to create prediction maps of Mn and Fe using the spatially continuous environmental parameters. The measured concentrations of Mn and Fe were converted into binary format according to the previously mentioned thresholds of Mn > 400 μg/L and Fe > 0.3 mg/L. The modeling described below therefore classifies the presence or absence of high Mn (>400 μg/L) or

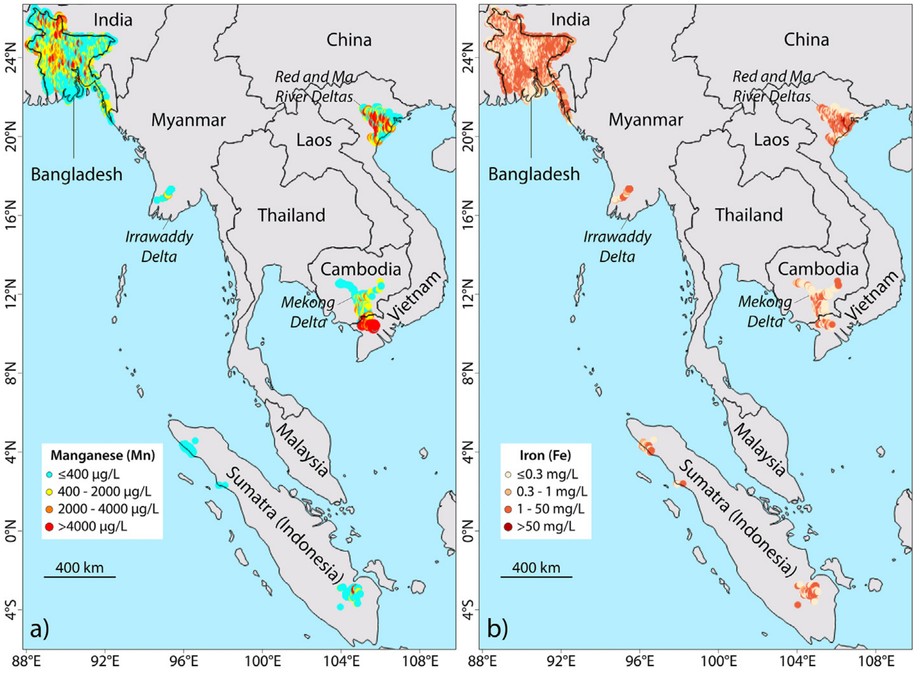


Fig. 1. Data points of groundwater measurements of a) manganese (n = 6122) and b) iron (n = 6107) used in this study. The data sources are listed in Table 1.

Table 1
Descriptive statistics of manganese and iron measurements in groundwater, sorted by region.

	Mn				Fe				Source
Region	n	Mean ± Std (µg/L)	Median (µg/L)	Proportion > 400 µg/L	n	Mean ± Std (mg/L)	Median (mg/L)	Proportion > 0.3 mg/L	
Bangladesh	4213	580 ± 750	321	0.45	4209	3.2 ± 5.0	1.1	0.66	(BGS and DPHE, 2001; Hoque et al., 2014)
Irrawaddy Delta (Myanmar)	55	411 ± 499	246	0.29	55	6.3 ± 5.9	5.1	0.85	(Van Geen et al., 2014)
Mekong Delta (Cambodia/Vietnam)	351	1444 ± 3582	400	0.48	351	2.7 ± 5.9	0.1	0.41	(Buschmann et al., 2008)
Red and Ma River Deltas (Vietnam)	1028	848 ± 1386	390	0.49	1028	10.5 ± 18.6	3.9	0.70	(Berg et al., 2001; Buschmann et al., 2008; Winkel et al., 2011)
Sumatra (Indonesia)	475	127 ± 410	16	0.06	464	2.5 ± 5.8	0.5	0.60	(Marohn et al., 2012; Winkel et al., 2008)
Total	6122	638 ± 1235	300	0.42	6107	4.4 ± 9.4	1.1	0.65	

high Fe (>0.3 mg/L) and was carried out with the R programming language (R Core Team, 2014). (This choice of units corresponds to those typically used in reporting these constituents.)

For all of the models, the original dataset was randomly split into training (80%) and testing (20%) subsets and stratified to maintain the proportion of high and low measurements in the full dataset. In order to ensure that the majority class of either high or low measurements does not dominate a model, the majority class was under-sampled in the training dataset to match the size of the minority class (Podgorski et al., 2018). The process of training a model on 80% of the data and testing it on the other 20% was repeated 100 times and the results subsequently averaged.

For modeling with the physicochemical parameters, the random forest (RF) (Breiman, 2001) algorithm was implemented using the randomForest package in R. A random forest grows many different decision trees by first randomly sampling the training dataset with replacement and then considering only a random subset of the predictor variables at each branch or node, which divides the target variable as heterogeneously as possible. This "forest" of decision trees is then averaged to produce the random forest model. For the modeling here, 5000 trees were grown for each random forest, which could be run in a reasonable amount of time and was not any worse than using more iterations. For the number of predictors made available at each node, the default value of the rounded down square root of the total number of predictors (e.g. $\sqrt{15} \approx 3$) was used.

Despite the physicochemical parameters potentially being indicative of the presence of Mn/Fe, they are point data and cannot be used for continuous spatial prediction, i.e. the creation of a probability map, as they are known only where the Mn/Fe measurements are already available. Therefore, in order to create prediction maps of manganese and iron, ML modeling was employed with the spatially continuous environmental predictor variables (Table S2). Although both sets of variables can be used to better understand the geochemical conditions under which the dissolution of Mn/Fe takes place, only spatially continuous variables can be used for creating a prediction map.

2.3.2. Spatially continuous prediction modeling of Mn and Fe

The spatial prediction modeling of Mn and Fe was carried out using both RF and generalized boosted regression modeling (GBM) (Ridgeway and Ridgeway, 2004). The results from the two different processes were then combined to form a final model according to their relative performance (see *Prediction maps* below). This was done in order to create a more robust final model by emphasizing the areas where the results of the two methods agree and placing less weight on where they do not.

As opposed to the RF models with the physicochemical variables, the RF geospatial models were created with subsets (different for Mn and Fe) of the statistically most important predictor variables. This was done using recursive feature elimination with the varSelRF package and opting for the least number of variables that result in an error rate within one standard deviation of the lowest cross-validation error rate with the out-of-bag (OOB) data points (Diaz-Uriarte and de Andrés, 2005). This was done to reduce the relatively large number of initial independent variables (n = 57) and help identify the most important variables as well as create additional diversity in the model outcomes.

GBM is also a tree-based method and works by creating an ensemble of decision trees that are grown successively and improved by reducing the error of the ensemble of previous trees. The overall result consists of the average of the final full set of trees. It uses decision trees with relatively few branches, which are considered to be weak learners that make small reductions in errors. This helps reduce overfitting and results in a more robust model, which is further promoted by utilizing a random subset of the training dataset for each tree. As such, three to seven decision-tree branches typically work well with GBM (Hastie et al., 2008).

The GBM was set up by first tuning the various modeling parameters using the Caret package (Kuhn, 2008). These parameters are the total number of trees, interaction depth or number of branches, shrinkage factor or learning rate and the minimum number of observations in a node. The GBM models were developed using the full set of 57 spatial predictor variables (Table S2) and randomly selecting one-half (default value) of the training dataset for growing each tree.

2.4. Model verification and assessment

Various metrics were used for assessing the classification of the binary target variable, i.e. Mn > 400 μg/L or Fe > 0.3 mg/L:

(1) Sensitivity $\frac{TP}{P}$
(2) Specificity $\frac{TN}{N}$
(3) Balanced accuracy $\frac{sensitivity+specificity}{2}$

where *TP* is true positives, *P* is total positive cases, *TN* is true negatives, *N* is total negative cases

(4) Kappa $\frac{p_0-p_e}{1-p_e}$

where p_0 is the agreement between predicted and actual values, and p_e is the expected agreement based on chance agreement.

The values of the metrics outlined above are typically reported using a probability cutoff of 0.5. However, another metric used that takes account of all possible cutoff values is the area under the ROC (receiver operating characteristic) curve (AUC). The ROC curve plots the true positive rate (sensitivity) against the false positive rate for many different probability cutoffs between 0 and 1. The AUC is simply the area beneath this curve, which ranges from 0 (model always incorrect), through 0.5 (equivalent to a random guess), to 1 (model always correct).

The relative importance of the predictor variables in contributing to each model was assessed by randomizing each variable in turn and measuring the resulting reduction in prediction performance (Breiman, 2001). To further help interpret the effect of each independent variable on the model outcome of the dependent variable, partial dependence plots (PDP) were produced of the predictors used in the random forest models. A PDP plots the model response for changes in a given predictor while holding all other variables constant at their average values.

2.5. Prediction maps

The entire dataset (training and testing) was used in creating a single model of each model type (GBM and RF) for each target variable (Mn > 400 μg/L and Fe > 0.3 mg/L). Each of these was then applied to its predictor variables to produce the spatially continuous probability prediction maps of Mn and Fe exceeding the threshold values. The final hazard prediction maps were then generated by averaging the GBM and RF prediction maps and weighting according to their respective AUC values.

In order to determine the parts of the prediction maps that can be considered more reliable, an assessment was made of the diversity of predictor data sampled by the concentration data points. This was then compared with the distribution of values of the predictor datasets across the entire study region to locate where the conditions are reasonably similar to those associated with the data points and thereby identify where the model can be most trusted. This was carried out using the CAST package (Meyer and Pebesma, 2021) by calculating a dissimilarity index (DI) for all points (pixels) within the study area. The DI is based on the distance in predictor space between two points (dissimilarity), weighted by variable importance, and represents the minimum distance to a training data point standardized by the average dissimilarity among all training points and can range from 0 to infinity. The DI was calculated separately for the RF and GBM models and combined according to their respective AUC values, analogously to the prediction maps themselves. This combined DI was then applied to the prediction maps as follows:

DI 0–1: prediction displayed in full.

DI 1–2: prediction displayed with hatch marks and considered "less reliable".

DI 2–3: prediction displayed with double hatch marks and considered "least reliable".

DI >3: prediction completely masked.

In addition, the uncertainty (or consistency) of the models as represented by the coefficient of variation (ratio of standard deviation to mean) was calculated by generating a prediction map from each of the 100 model iterations from cross-fold validation and then calculating their mean and standard deviation.

3. Results

3.1. Analysis of groundwater chemistry and depth distribution

The Kendall rank correlations among the concentrations of Mn, Fe and As for all of the locations are shown in Fig. 2. The correlations among all of the physicochemical parameters are in Fig. S1. At all of the sites, there is a significant correlation between the concentrations of Fe and As. Mn and As as well as Mn and Fe are positively correlated in Bangladesh, Indonesia and the Red and Ma River Deltas in northern Vietnam. However, in the Mekong Delta, Mn and As are slightly negatively correlated and Mn and Fe show no significant correlation.

The depth profiles of Mn, Fe and As are shown in Fig. 3. In Bangladesh, the concentrations of Mn, Fe and As exceeding their respective thresholds in groundwater are all high near the surface and decrease to about 50 m

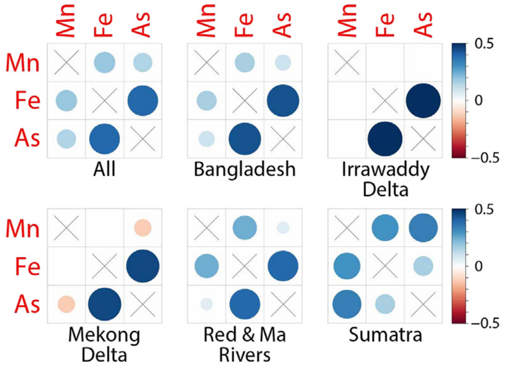


Fig. 2. Kendall rank correlations between manganese (Mn), iron (Fe) and arsenic (As) at the 0.05 significance level. Blank cells indicate the correlation did not meet the 0.05 significance level.

depth (Fig. 3a). Below this depth, Mn concentrations continue to decrease and stay low to the limit of the depths sampled. Fe and As are strongly correlated and show both sharp increases and decreases between about 50–200 m depth (Fig. 3a, b). Between 200 and 250 m depth, both Mn and Fe show a sharp increase and then decrease.

In the Mekong Delta (Fig. 3c) and the Red and Ma River Deltas of northern Vietnam (Fig. 3d), a strong negative correlation is seen between Mn and As in the shallowest aquifer to about 25 m depth. In the Mekong Delta, Fe concentrations stay roughly constant over this interval, whereas Fe in the Red and Ma River Deltas also decreases somewhat. At depths greater than about 50 m at these two locations, the concentrations of Mn, Fe and As often appear to vary together, sometimes offset by up to about ten vertical meters. Analysis of the vertical distribution of dissolved Mn, Fe and As in Myanmar (Fig. 3b) is hampered somewhat by limited vertical resolution due to a relatively small dataset. In Sumatra the data show variation in all three elements to about 100 m depth, often in the same sense.

The correlation analysis between the presence in groundwater of high Mn (>400 μg/L) and Fe (>0.3 mg/L) with each other as well as with high As (>10 μg/L) at the five regions (Fig. 2) confirms that high Fe is often found in connection with high As (Van Geen et al., 2008). This is also the case in the RF model using physicochemical predictor variables, which shows that As is generally positively associated with Fe (Fig. 4g). Although high concentrations of Fe can be detrimental to water infrastructure, high concentrations of As can be disastrous for health. While the presence of Fe may be manifested visually or through taste, even very high concentrations of arsenic remain undetected by our senses. Furthermore, due to the hazardous concentrations of As generally being quite low (e.g. μg/L rather than mg/L), it's analysis is relatively difficult and, as such, is often not undertaken. Therefore, the relatively easily detected presence of Fe in groundwater in Southeast Asia provides an indication that the given source should also be tested for As (Biswas et al., 2012; Hoque et al., 2012).

The relationship between Mn and As is less clear. Although it was found that Mn is sometimes positively correlated with As, it is generally weaker than the correlation between Fe and As (Fig. 2). In the measurements available from Indonesia, this relationship is actually stronger, but in Myanmar Mn and As are negatively correlated.

Considering the depth profiles of averaged high Mn, Fe and As (Fig. 3), the pattern emerges that while Fe and As often fluctuate together over depth, Mn and As frequently vary in the opposite sense. In addition, a peak in Mn may occur vertically offset from a peak in As, for example, in the Red and Ma River Deltas of Vietnam (Fig. 3d).

3.2. Random forest modeling with physicochemical variables

Random forest models of high Mn and Fe using the 14 other physicochemical parameters as predictors (performance results in Table S3) were created in order to examine the relationships between these parameters and Mn and Fe. The importance of the variables in terms of mean decrease in accuracy as well as the PDPs of four predictors (As, Eh, NO_3 and PO_4) that show opposite effects on high concentrations of Mn and Fe are shown in Fig. 4. Figs. S2 and S3 contain the PDPs of all the predictors. This analysis confirms known mechanisms. Because these parameters are point data, they could not be used for creating a probability map.

3.3. ML modeling with spatially continuous variables

For the RF modeling of high Mn and Fe using the spatially continuous predictor variables, eight predictors for Mn and 15 for Fe were selected. Their importance, expressed as mean decrease in accuracy, is shown in Fig. 5. The PDPs of these variables (Figs. S4 and S5) indicate that a drier climate and flatter terrain are associated with high Mn concentrations in groundwater in Southeast Asia, while more humid conditions and greater soil organic carbon predict high Fe concentrations. Some of the more important independent variables are displayed in Fig. S6.

As shown in Table 2, the GBM and RF models of high Mn using the spatially continuous variables perform equally well, whereas the RF model for Fe somewhat outperforms the GBM model. The GBM and RF models were

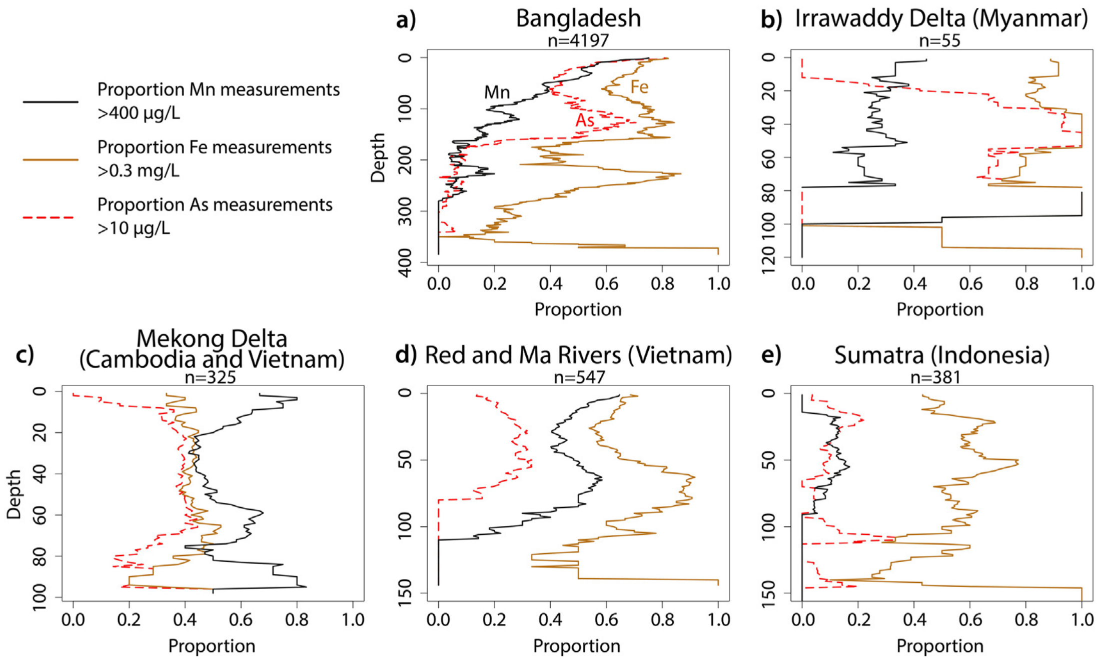


Fig. 3. Proportion of manganese (Mn), iron (Fe) and arsenic (As) measurements exceeding their respective thresholds of 400 μg/L, 0.3 mg/L and 10 μg/L with depth. The proportions were calculated for every meter of depth using a window of 20 m (between 10 m shallower and 10 m deeper) for a) Bangladesh, b) the Irrawaddy Delta (Myanmar), c) the Mekong Delta (Cambodia and Vietnam), d) the Red and Ma River Deltas of Vietnam and e) Sumatra (Indonesia). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

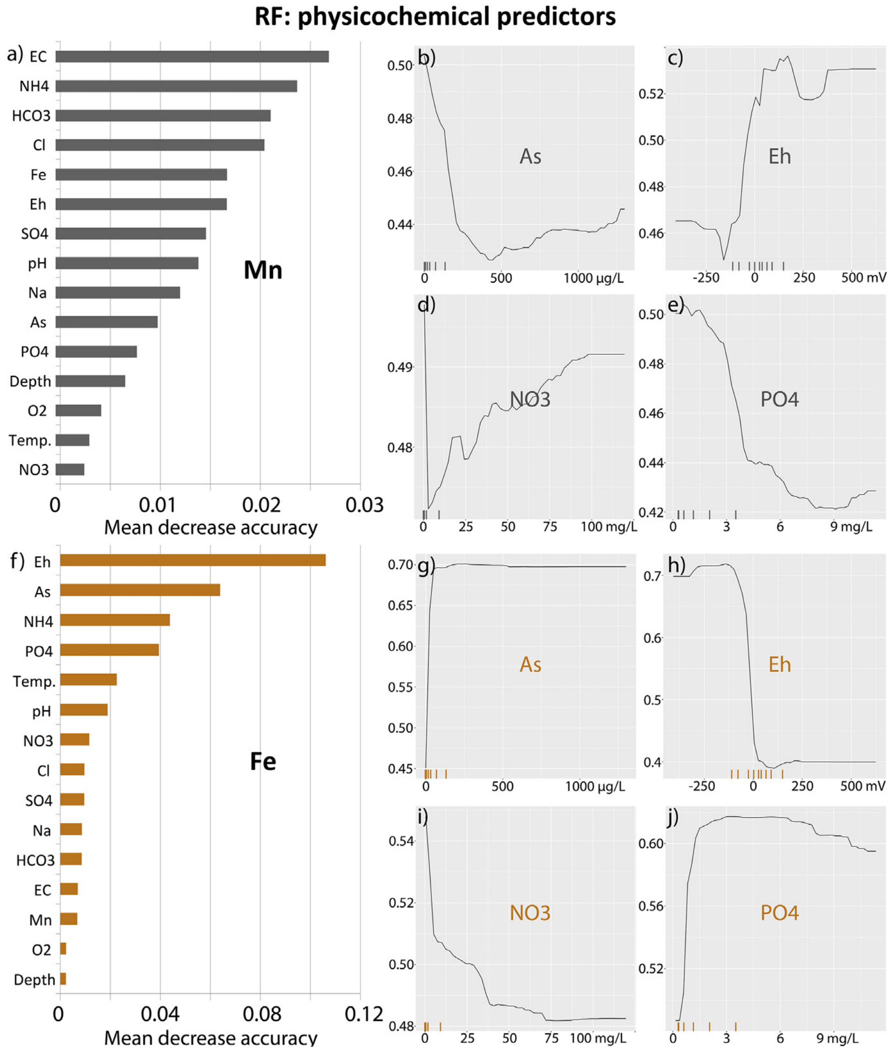


Fig. 4. Results of random forest (RF) models for Mn > 400 µg/L (a–e) and Fe > 0.3 mg/L (f–j) using physicochemical parameters as predictor variables. Variable importance (a, f) is in terms of the mean decrease in accuracy when each variable's values are randomly sorted. Partial dependence plots (PDP) of predictor variables are shown of the critical geogenic contaminant arsenic (b, g), redox potential (Eh) (c, h), nitrate (d, i) and phosphate (e, j). For example, the PDP of As for the Mn model (b) indicates that higher As concentrations lead to a lower probability of Mn > 400 µg/L. The tick marks at the bottom of each PDP indicate the distribution of the data in deciles.

averaged to form the combined final probability maps (Fig. 6a, b), weighting by their respective AUC values (Table 2). The respective dissimilarity indexes were likewise averaged according to the AUC and applied to the manganese and iron probability maps. For comparison, Fig. 6c contains a section of the global groundwater prediction map of As >10 µg/L (Podgorski and Berg, 2020) for the same area of Southeast Asia. For closer inspection, all three of these maps are displayed in a larger format in Figs. S7–S9. Maps of the coefficient of variation for the RF and GBM models of Mn and Fe are shown in Fig. S10.

4. Discussion

Areas of elevated Mn groundwater hazard (Fig. 6a) include northwest Bangladesh, the middle Irrawaddy basin in Myanmar, the plains along the Chao Phraya River in Thailand, the lowlands of eastern Thailand and northwest Cambodia, the Mekong Delta of Cambodia and Vietnam and the Red and Ma River Deltas of northern Vietnam. The greatest hazard of Mn > 400 µg/L is found in Bangladesh, the Red and Ma River Deltas (Fig. 7a) and Mekong Delta (Fig. 7c), all of which are very densely populated areas. The proportion of people using untreated groundwater is, on average, 84% in Bangladesh, 40% in Cambodia and 45% in Vietnam (JMP, 2019). It is therefore imperative that the chemical testing of drinking water wells in these areas include Mn.

Compared with Mn, there are many more predicted areas of Fe concentrations exceeding the associated WHO guideline of 0.3 mg/L (Fig. 6b). Although not representing a health threat, excess Fe in groundwater makes water undesirable to use and can negatively impact water infrastructure. As such, the Fe probability map is important for water resource planning and the siting of new water utilities.

Because Mn, Fe and As get released along the same redox chain, areas of an aquifer that have high concentrations of one element may be found next to areas with high concentrations in another one. This is clearly observed in the Red and Ma River Deltas, where areas of high Mn hazard (Fig. 7a) are located immediately adjacent to or slightly overlapping with areas with high As hazard (Fig. 7b). The same situation between Mn and As hazard areas exists in the Mekong Delta (Fig. 7c, d). That high Mn correlates less well with high As than does high Fe means that groundwater sources that have already been tested and determined to be free from high As concentrations may still contain high concentrations of Mn that pose a health risk. The Mn and Fe prediction maps (Fig. 6a, b) can therefore be used by water resource managers and well operators to help identify potentially problematic areas in terms of high concentrations of either of these elements as well as their association with As. Specifically, the prediction

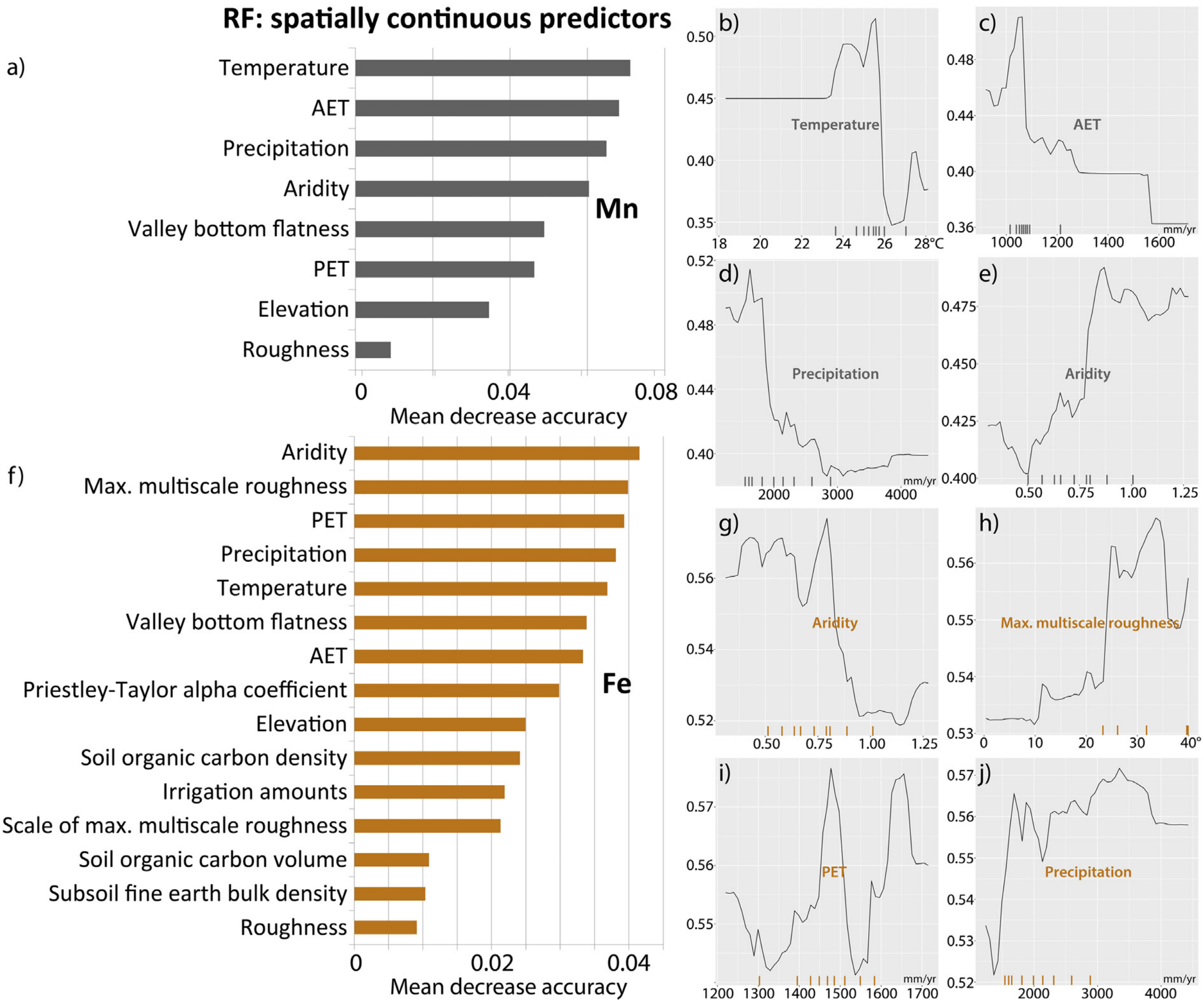


Fig. 5. Results of random forest (RF) models for Mn > 400 μg/L (a–e) and Fe > 0.3 mg/L (f–j) using spatially continuous parameters as predictor variables. Variable importance (a, f) is in terms of the mean decrease in accuracy when each variable's values are randomly sorted. Partial dependence plots (PDP) are shown of the four most important predictor variables from each model (b–e for Mn; g–j for Fe). The tick marks at the bottom of each PDP indicate the distribution of the data in deciles.

Table 2
Performance of models produced by the generalized boosted regression modeling (GBM) and random forest (RF) algorithms for manganese (Mn) and iron (Fe) in groundwater. Each of the results is based on 100-fold cross validation, with each model being grown with 5000 trees using an 80%/20% training/test data split. The probability cutoff was determined for each fold by finding the cutoff at which sensitivity equals specificity, which by definition also equals balanced accuracy.

	Mn (GBM)	Mn (RF)	Fe (GBM)	Fe (RF)
Prob. cutoff	0.52 ± 0.01	0.48 ± 0.02	0.49 ± 0.01	0.53 ± 0.01
Balanced accuracy	0.72 ± 0.02	0.72 ± 0.01	0.70 ± 0.01	0.72 ± 0.01
Kappa	0.44 ± 0.03	0.44 ± 0.02	0.40 ± 0.03	0.43 ± 0.02
AUC	0.80 ± 0.01	0.80 ± 0.01	0.76 ± 0.01	0.79 ± 0.01

maps of high Mn and Fe are to be used in prioritizing areas that require further testing of groundwater resources and are not a substitute for actual testing.

We have demonstrated that it is feasible to create spatial prediction models of the occurrence of high concentrations of the geogenic groundwater contaminants of Mn and Fe. Although they have generally received much less attention than the geogenic contaminants of As and F, Mn and Fe are nevertheless important from the standpoint of infrastructure damage as well as human health, in the case of Mn. These maps also highlight the possibility that other groundwater contaminants or constituents having associations with spatially continuous surface parameters (geogenic or even anthropogenic) may be suitable for ML-based spatial modeling.

4.1. Modeling limitations

Although the available groundwater quality data from the region are located in clusters (Fig. 1), the applicability of the spatial model to intermediate areas could be estimated by comparing the values of the predictor variables with those associated with the training data points. The dissimilarity index (DI) calculated and used to progressively mask areas of the prediction maps with greater dissimilarity to the training data thereby gives a sense of where the model results are more reliable (Fig. 6), which largely corresponds to lower elevation areas (mainly flood plains) that are also similar in terms of the other main predictor variables (Fig. S6).

In general when modeling, it is important that the training data sample an adequate diversity of values in the predictor data in order for the latter to be able to provide much information about the former. With regard to this aspect as well as being able to make predictions for new areas, clustered data can provide a challenge as closely spaced data points are generally associated with a small range of values of environmental variables. A semi-quantitative assessment, such as that made here with the DI, therefore helps determine the degree to which the available data can be applied elsewhere.

Complementary to this, the coefficient of variation calculated from the maps generated during cross validation (Fig. S10) often shows greater variability and therefore more uncertainty in areas with a higher DI. It is also

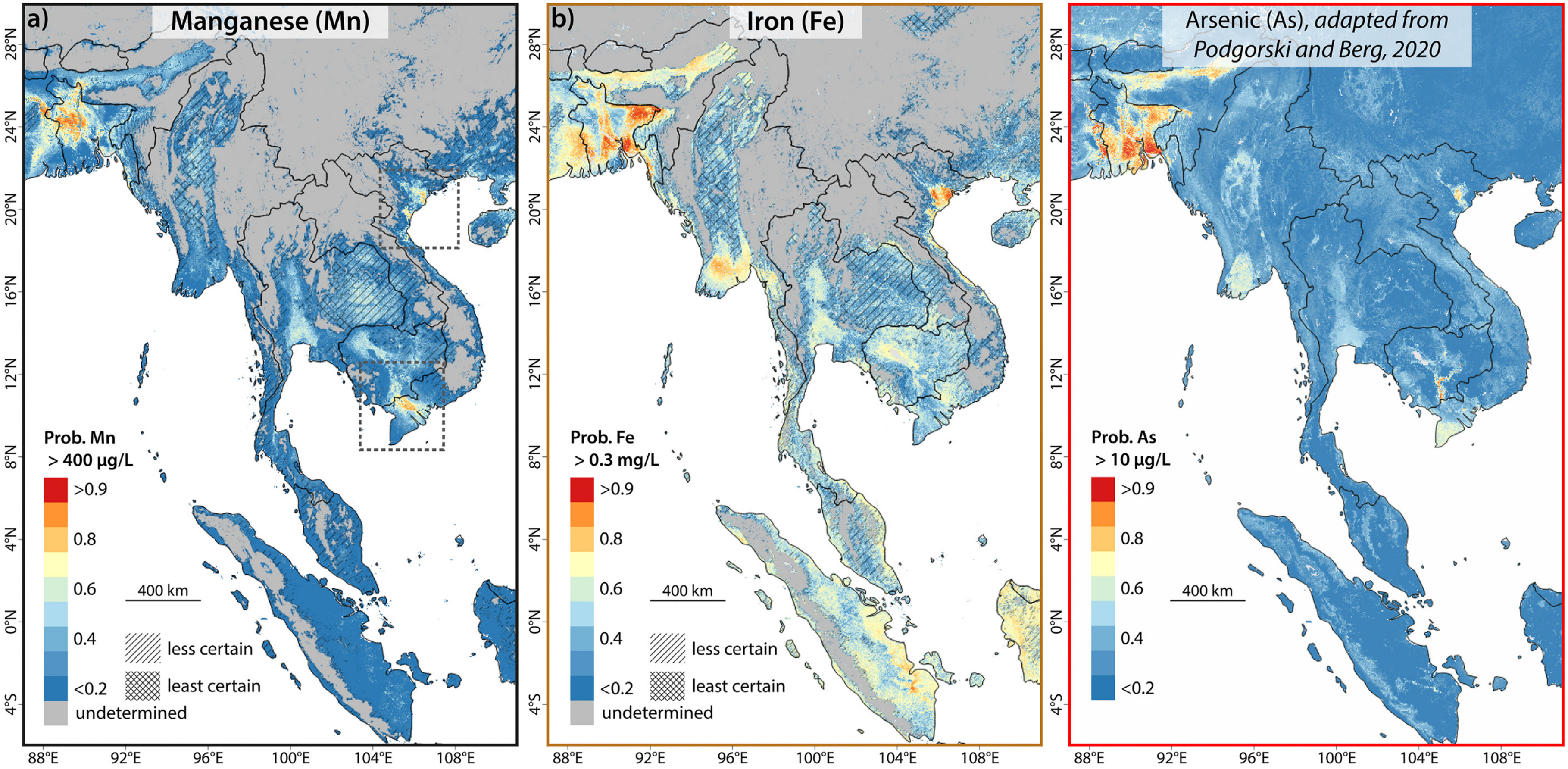


Fig. 6. Spatial probability maps of Southeast Asia and Bangladesh produced of a) manganese >400 μg/L and b) iron >0.3 mg/L in groundwater as well as c) an existing arsenic prediction map adapted from Podgorski and Berg, 2020. Less reliable sections of a) and b) are indicated with hatch marks; other areas that are considered too dissimilar to the training data are masked in gray. The dashed boxes in a) indicate the areas depicted in Fig. 7. Larger versions of each probability map are found in the Supplementary Materials (Figs. S7–S9).

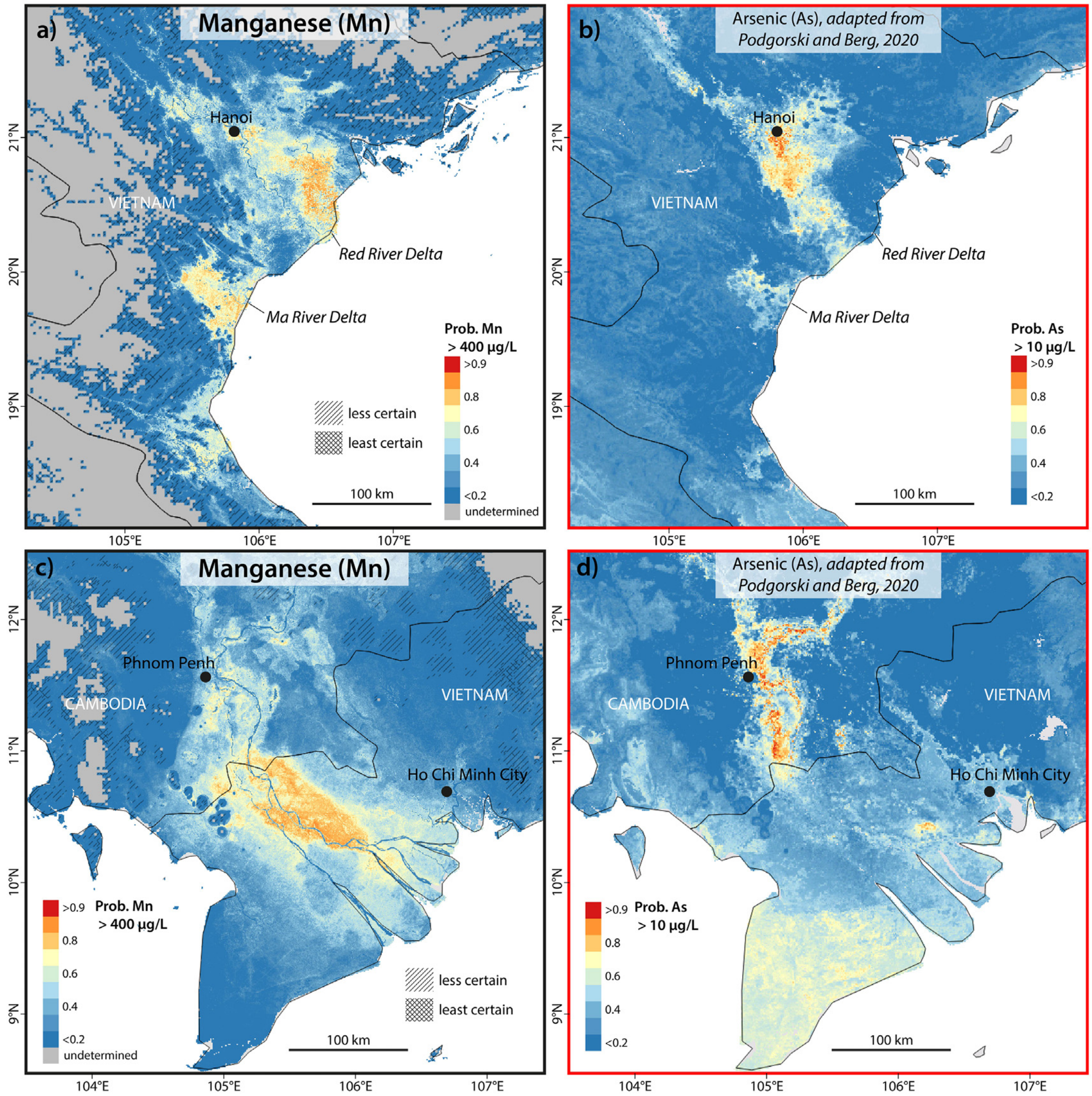


Fig. 7. Detailed views of the modeled Mn hazard map of a) the Red and Ma River Deltas in northern Vietnam next to b) the modeled As hazard of the same area from Podgorski and Berg (2020). Also shown are c) Mn and d) As in the Mekong Delta of Cambodia and Vietnam. These areas are indicated in the map in Fig. 6a. Less reliable sections of a) and c) are indicated with hatch marks; other areas that are considered too dissimilar to the training data are masked in gray. Both examples clearly show the presence of high hazard areas of Mn immediately adjacent to or slightly overlapping high hazard areas of As. Due to both areas being densely populated and relying heavily on groundwater, there is a high risk of exposure to Mn concentrations (in addition to As) in groundwater that are hazardous to health. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

apparent from this analysis that the overall variability among the GBM results is considerably greater than that from RF.

CRediT authorship contribution statement

Joel Podgorski - Conceptualization; Data curation; Formal analysis; Funding acquisition; Investigation; Methodology; Project administration; Software; Validation; Visualization; Roles/Writing – original draft; Writing – review & editing.

Dahyann Araya – Resources; Roles/Writing – original draft; Writing – review & editing.

Michael Berg – Conceptualization; Funding acquisition; Project administration; Supervision; Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank the Swiss Agency for Development and Cooperation (7F-09963.01.01) for their long-term support and co-funding of this study. We would also like to acknowledge our colleague Anja Bretzler for preliminary investigations and discussions on this topic.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2022.155131.

References

Amoako, J., Karikari, A., Ansa-Asare, D., 2011. Physico-chemical quality of boreholes in Densu Basin of Ghana. Appi Water Sci 1, 41–48.

Ayotte, J.D., Medalie, L., Qi, S.L., Backer, L.C., Nolan, B.T., 2017. Estimating the high-arsenic domestic-well population in the conterminous United States. Environ. Sci. Technol. 51, 12443–12454.

Bacquart, T., Bradshaw, K., Frisbie, S., Mitchell, E., Springston, G., Defelice, J., et al., 2012. A survey of arsenic, manganese, boron, thorium, and other toxic metals in the groundwater of a West BengalIndia neighbourhood. Metallomics 4, 653–659.

Bacquart, T., Frisbie, S., Mitchell, E., Grigg, L., Cole, C., Small, C., et al., 2015. Multiple inorganic toxic substances contaminating the groundwater of myingyan township, Myanmar: arsenic, manganese, fluoride, iron, and uranium. Sci. Total Environ. 517, 232–245.

Berg, M., Tran, H.C., Nguyen, T.C., Pham, H.V., Schertenleib, R., Giger, W., 2001. Arsenic contamination of groundwater and drinking water in Vietnam: a human health threat. Environmental Science & Technology 35, 2621–2626.

BGS, DPHE, 2001. Arsenic contamination of groundwater in Bangladesh. In: Kinniburgh, D.G., Smedley, P.L. (Eds.), British Geological Survey Technical Report WC/00/19. British Geological Survey, Keyworth.

Biswas, A., Nath, B., Bhattacharya, P., Halder, D., Kundu, A.K., Mandal, U., et al., 2012. Testing tubewell platform color as a rapid screening tool for arsenic and manganese in drinking water wells. Environ. Sci. Technol. 46, 434–440.

Bouchard, M., Laforest, F., Vandelac, L., Bellinger, D., Mergler, D., 2007. Hair manganese and hyperactive behaviors: pilot study of school-age children exposed through tap water. Environ. Health Perspect. 115, 122–127.

Bouchard, M.F., Sauvé, S., Barbeau, B., Legrand, M., Brodeur, M.-È., Bouffard, T., et al., 2011. Intellectual impairment in school-age children exposed to manganese from drinking water. Environ. Health Perspect. 119, 138–143.

Breiman, L., 2001. Random forests. Machine learning 45, 5–32.

Bretzler, A., Lalanne, F., Nikiema, J., Podgorski, J., Pfenninger, N., Berg, M., et al., 2017. Groundwater arsenic contamination in Burkina Faso, West Africa: predicting and verifying regions at risk. Sci. Total Environ. 584, 958–970.

Buschmann, J., Berg, M., Stengel, C., Winkel, L., Sampson, M.L., Trang, P.T.K., et al., 2008. Contamination of drinking water resources in the Mekong delta floodplains: arsenic and other trace metals pose serious health risks to population. Environ. Int. 34, 756–764.

Carretero, S., Kruse, E., 2015. Iron and manganese content in groundwater on the northeastern coast of the Buenos Aires provinceArgentina. Environmental Earth Sciences 73, 1983–1995.

Claus Henn, B., Bellinger, D.C., Hopkins, M.R., Coull, B.A., Ettinger, A.S., Jim, R., et al., 2017. Maternal and cord blood manganese concentrations and early childhood neurodevelopment among residents near a mining-impacted superfund site. Environ. Health Perspect. 125, 067020.

de Meyer, C.M., Rodríguez, J.M., Carpio, E.A., García, P.A., Stengel, C., Berg, M., 2017. Arsenic, manganese and aluminum contamination in groundwater resources of Western Amazonia (Peru). Sci. Total Environ. 607, 1437–1450.

DeSimone, L.A., Ransom, K.M., 2021. Manganese in the northern Atlantic coastal plain aquifer system, eastern USA—modeling regional occurrence with pH, redox, and machine learning. J. Hydrol. Reg. Stud. 37, 100925.

DeSimone, L.A., Pope, J.P., Ransom, K.M., 2020. Machine-learning models to map pH and redox conditions in groundwater in a layered aquifer system, northern Atlantic coastal plain, eastern USA. J. Hydrol. Reg. Stud. 30, 100697.

Diaz-Uriarte, R., de Andrés, S.A., 2005. Variable selection from random forests: application to gene expression data. arXiv (TR009), q-bio/0503025 https://doi.org/10.48550/arXiv.q-bio/0503025.

Dion, L.-A., Saint-Amour, D., Sauvé, S., Barbeau, B., Mergler, D., Bouchard, M.F., 2018. Changes in water manganese levels and longitudinal assessment of intellectual function in children exposed through drinking water. Neurotoxicology 64, 118–125.

EPA U, 2004. Drinking Water Health Advisory for Manganese, Washington, DC.

Erickson, M.L., Elliott, S.M., Brown, C.J., Stackelberg, P., Ransom, K.M., Reddy, J.E., 2021a. Machine learning predicted redox conditions in the glacial aquifer system, northern continental United States. Water Resour. Res. 57, e2020WR028207.

Erickson, M.L., Elliott, S.M., Brown, C.J., Stackelberg, P.E., Ransom, K.M., Reddy, J.E., et al., 2021b. Machine-learning predictions of high arsenic and high manganese at drinking water depths of the glacial aquifer system, northern continental United States. Environ. Sci. Technol. 55, 5791–5805.

Ghosh, G.C., Khan, M.J.H., Chakraborty, T.K., Zaman, S., Kabir, A.E., Tanaka, H., 2020. Human health risk assessment of elevated and variable iron and manganese intake with arsenic-safe groundwater in jashoreBangladesh. Scientific reports 10, 1–9.

Hastie, T.T., Robert, Friedman, Jerome, 2008. The Elements of Statistical Learning. 2nd ed. Springer.

Haynes, E.N., Sucharew, H., Kuhnell, P., Alden, J., Barnas, M., Wright, R.O., et al., 2015. Manganese exposure and neurocognitive outcomes in rural school-age children: the communities actively researching exposure study (Ohio, USA). Environ. Health Perspect. 123, 1066–1071.

Holzgraefe, M., Poser, W., Kijewski, H., Beuche, W., 1986. Chronic enteral poisoning caused by potassium permanganate: a case report. J. Toxicol. Clin. Toxicol. 24, 235–244.

Homoncik, S.C., MacDonald, A.M., Heal, K.V., Dochartaigh, B.É.Ó., Ngwenya, B.T., 2010. Manganese concentrations in Scottish groundwater. Sci. Total Environ. 408, 2467–2473.

Hoque, M., McArthur, J., Sikdar, P., 2012. The palaeosol model of arsenic pollution of groundwater tested along a 32 km traverse across West BengalIndia. Science of the Total Environment 431, 157–165.

Hoque, M.A., McArthur, J.M., Sikdar, P.K., Ball, J.D., Molla, T.N., 2014. Tracing recharge to aquifers beneath an Asian megacity with Cl/Br and stable isotopes: the example of Dhaka, Bangladesh. Hydrogeology Journal 22 (7), 1549–1560.

Huang, R., Ma, C., Ma, J., Huangfu, X., He, Q., 2021. Machine learning in natural and engineered water systems. Water Res. 117666.

Iyare, P., 2019. The effects of manganese exposure from drinking water on school-age children: a systematic review. Neurotoxicology 73, 1–7.

JMP, 2019. Global Data on Water Supply, Sanitation and Hygiene (WASH). 2019. WHO/UNICEF Joint Monitoring Program (JMP).

Johnson, C.D., Nandi, A., Joyner, T.A., Luffman, I., 2018. Iron and manganese in groundwater: using kriging and GIS to locate high concentrations in Buncombe CountyNorth Carolina. Groundwater 56, 87–95.

Kohl, P.M., Medlar, S.J., 2006. Occurrence of manganese in drinking water and manganese control. J. Am. Water Works Assoc. 1–435.

Kondakis, X.G., Makris, N., Leotsinidis, M., Prinou, M., Papapetropoulos, T., 1989. Possible health effects of high manganese concentration in drinking water. Arch. Environ. Health 44, 175–178.

Koppi, A.J., Edis, R., Field, D.J., Geering, H.R., Klessa, D.A., Cockayne, D.J., 1996. Rare earth element trends and cerium-uranium-manganese associations in weathered rock from koongarra, Northern TerritoryAustralia. Geochimica et Cosmochimica Acta 60, 1695–1707.

Kuhn, M., 2008. Building predictive models in R using the caret package. J. Stat. Softw. 28, 1–26.

Marohn, C., Distel, A., Tomlinson, R., Noordwijk, M.V., Cadisch, G., 2012. Impacts of soil and groundwater salinization on tree crop performance in post-tsunami Aceh Barat, Indonesia. Natural Hazards and Earth System Sciences 12, 2879–2891.

Meyer, H., Pebesma, E., 2021. Predicting into unknown space? Estimating the area of applicability of spatial prediction models. Methods Ecol. Evol. 12, 1620–1633.

Perl, D.P., Olanow, C.W., 2007. The neuropathology of manganese-induced parkinsonism. J. Neuropathol. Exp. Neurol. 66, 675–682.

Phan, V.T., Bernier-Latmani, R., Tisserand, D., Bardelli, F., Le Pape, P., Frutschi, M., et al., 2019. As release under the microbial sulfate reduction during redox oscillations in the upper Mekong delta aquifers, Vietnam: a mechanistic study. Sci. Total Environ. 663, 718–730.

Podgorski, J., Berg, M., 2020. Global threat of arsenic in groundwater. Science 368, 845–850.

Podgorski, J.E., Eqani, S.A.M.A.S., Khanam, T., Ullah, R., Shen, H., Berg, M., 2017. Extensive arsenic contamination in high-pH unconfined aquifers in the Indus Valley. ScienceAdvances 3.

Podgorski, J.E., Labhasetwar, P., Saha, D., Berg, M., 2018. Prediction modeling and mapping of groundwater fluoride contamination throughout India. Environ. Sci. Technol. 52, 9889–9898.

Podgorski, J., Wu, R., Chakravorty, B., Polya, D.A., 2020. Groundwater arsenic distribution in India by machine learning geospatial modeling. Int. J. Environ. Res. Public Health 17, 7119.

Rahman, M.F., Mahmud, M.J., Sadmani, A.A., Chowdhury, A.I., Anderson, W.B., Bodruzzaman, A.B., et al., 2021. Previously unrecognized potential threat to children from manganese in groundwater in rohingya refugee camps in Cox's Bazar, Bangladesh. Chemosphere 266, 129128.

Richards, L.A., Magnone, D., Sovann, C., Kong, C., Uhlemann, S., Kuras, O., et al., 2017. High resolution profile of inorganic aqueous geochemistry and key redox zones in an arsenic bearing aquifer in Cambodia. Sci. Total Environ. 590, 540–553.

Ridgeway, G., Ridgeway, M.G., 2004. The gbm Package. 5. R Foundation for Statistical Computing, Vienna, Austria.

Roccaro, P., Barone, C., Mancini, G., Vagliasindi, F., 2007. Removal of manganese from water supplies intended for human consumption: a case study. Desalination 210, 205–214.

Sahni, V., Léger, Y., Panaro, L., Allen, M., Giffin, S., Fury, D., et al., 2007. Case report: a metabolic disorder presenting as pediatric manganism. Environ. Health Perspect. 115, 1776–1779.

Sajedi-Hosseini, F., Malekian, A., Choubin, B., Rahmati, O., Cipullo, S., Coulon, F., et al., 2018. A novel machine learning-based approach for the risk assessment of nitrate groundwater contamination. Sci. Total Environ. 644, 954–962.

Schullehner, J., Thygesen, M., Kristiansen, S.M., Hansen, B., Pedersen, C.B., Dalsgaard, S., 2020. Exposure to manganese in drinking water during childhood and association with attention-deficit hyperactivity disorder: a nationwide cohort study. Environ. Health Perspect. 128, 097004.

Sly, L., Hodgkinson, M., Arunpairojana, V., 1990. Deposition of manganese in a drinking water distribution system. Appl. Environ. Microbiol. 56, 628–639.

Spangler, A.H., Spangler, J.G., 2009. Groundwater manganese and infant mortality rate by county in North Carolina: an ecological analysis. EcoHealth 6, 596–600.

Team, R.Core, 2014. In: RFfS, Computing (Ed.), R: A Language and Environment for Statistical Computing Vienna, Austria.

Thapa, R., Gupta, S., Kaur, H., Mandal, R., 2018. Assessment of manganese contamination in groundwater using frequency ratio (FR) modeling and GIS: a case study on burdwan district, West BengalIndia. Modeling Earth Systems and Environment 4, 161–174.

Van Geen, A., Radloff, K., Aziz, Z., Cheng, Z., Huq, M., Ahmed, K., et al., 2008. Comparison of arsenic concentrations in simultaneously-collected groundwater and aquifer particles from Bangladesh, India, Vietnam, and Nepal. Appl. Geochem. 23, 3244–3251.

Van Geen, A., Win, K.H., Zaw, T., Naing, W., Mey, J.L., Mailloux, B., 2014. Confirmation of elevated arsenic levels in groundwater of Myanmar. Science of the Total Environment 478, 21–24.

Wasserman, G.A., Liu, X., Parvez, F., Ahsan, H., Levy, D., Factor-Litvak, P., et al., 2006. Water manganese exposure and children's intellectual function in araihazarBangladesh. Environmental health perspectives 114, 124–129.

WHO, 2003. Iron in drinking-water. WHO Guidelines for Drinking-water Quality. Backgr. Doc. Dev. WHO Guidel. Drink. Qual. Who/Sde/Wsh/03.04/08-2. 4.

WHO, 2004. Guidelines for Drinking-water Quality. 1. World Health Organization.

WHO, 2011. Guidelines for drinking-water quality. WHO chronicle. 38, pp. 104–108.

WHO, 2017. Guidelines for Drinking-water Quality: First Addendum to the Fourth Edition.

Winkel, L., Berg, M., Stengel, C., Rosenberg, T., 2008. Hydrogeological survey assessing arsenic and other groundwater contaminants in the lowlands of Sumatra, Indonesia. Applied Geochemistry 23, 3019–3028.

Winkel, L.H., Trang, P.T.K., Lan, V.M., Stengel, C., Amini, M., Ha, N.T., et al., 2011. Arsenic pollution of groundwater in Vietnam exacerbated by deep aquifer exploitation for more than a century. Proc. Natl. Acad. Sci. 108, 1246–1251.

Woolf, A., Wright, R., Amarasiriwardena, C., Bellinger, D., 2002. A child with chronic manganese exposure from drinking water. Environ. Health Perspect. 110, 613–616.

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.

WWAP, 2015. The United Nations World Water Development Report 2015: Water for a Sustainable World. 1. UNESCO publishing.

Ying, S.C., Schaefer, M.V., Cock-Esteb, A., Li, J., Fendorf, S., 2017. Depth stratification leads to distinct zones of manganese and arsenic contaminated groundwater. Environ. Sci. Technol. 51, 8926–8932.

Zhong, S., Zhang, K., Bagheri, M., Burken, J.G., Gu, A., Li, B., et al., 2021. Machine learning: new ideas and tools in environmental science and engineering. Environ. Sci. Technol. 55 (19), 12741–12754.