Environmental Science & Technology

Article

Cite This: Environ. Sci. Technol. 2018, 52, 9889−9898

pubs.acs.org/est

Prediction Modeling and Mapping of Groundwater Fluoride Contamination throughout India

Joel E. Podgorski,*,[†] Pawan Labhasetwar,[‡] Dipankar Saha,[§,∇] and Michael Berg*,[†]

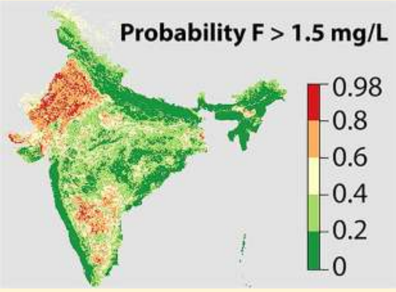
[†]Eawag, Swiss Federal Institute of Aquatic Science and Technology, Department Water Resources and Drinking Water, 8600 Dübendorf, Switzerland

[‡]Water Technology and Management Division, CSIR-NEERI, Nehru Marg, Nagpur 440020, India

[§]Central Ground Water Board, Government of India, Faridabad 121001, India

Supporting Information

ABSTRACT: For about the past eight decades, high concentrations of naturally occurring fluoride have been detected in groundwater in different parts of India. The chronic consumption of fluoride in high concentrations is recognized to cause dental and skeletal fluorosis. We have used the random forest machine-learning algorithm to model a data set of 12 600 groundwater fluoride concentrations from throughout India along with spatially continuous predictor variables of predominantly geology, climate, and soil parameters. Despite only surface parameters being available to describe a subsurface phenomenon, this has produced a highly accurate prediction map of fluoride concentrations exceeding 1.5 mg/L at 1 km resolution throughout the country. The most affected areas are the northwestern states/territories of Delhi, Gujarat, Haryana, Punjab, and Rajasthan and the southern states of Andhra Pradesh, Karnataka, Tamil Nadu, and Telangana. The total number of people at risk of fluorosis due to fluoride in groundwater is predicted to be around 120 million, or 9% of the population. This number is based on rural populations and accounts for average rates of groundwater consumption from nonmanaged sources. The new fluoride hazard and risk maps can be used by authorities in conjunction with detailed groundwater utilization information to prioritize areas in need of mitigation measures.



INTRODUCTION

With much of India consisting of arid and semiarid regions, the country relies heavily on groundwater to support its growing economy and population. Subsidized electricity for farmers has also led to substantial increases in groundwater pumping for irrigation since 1980, with groundwater now accounting for more than 60% of irrigation water and 85% of rural drinking water supplies in India.[1] Furthermore, satellite observations have detected a chronic overexploitation as well as drastic reduction of groundwater resources.[2,3] India has two main aquifer types: alluvial aquifers of the Gangetic plains where groundwater depth is generally less than 10 m and hard-rock aquifers elsewhere where groundwater depths can reach in excess of 60 m.[4,5] Given that groundwater plays a key role in the provisioning of water in India, considerable focus needs to be placed not only on the availability of groundwater but also its quality. Although groundwater is typically free of the microbiological contamination that is widespread in surface waters, various natural, or geogenic, chemical contaminants can be a problem. One of the most important of these in India is fluoride.[6]

Fluoride is reported to be beneficial for good dental health, with the optimal amount in drinking water generally being in the narrow range of 0.5−1.0 mg/L.[7] However, dental fluorosis or skeletal or nonskeletal fluorosis can result from chronic excess exposure. The World Health Organization (WHO) has set a maximum concentration guideline of 1.5 mg/L for drinking water,[8] although the volume of water consumed, other sources of fluoride and nutritional deficiencies should also be considered. For example, low intake of vitamin C, micronutrients, and calcium are associated with a greater incidence of fluorosis.[9] Although India maintains a permissible fluoride limit of 1.5 mg/L, it has set a target concentration of 1.0 mg/L, at least in part to account for the large quantities of water that people must drink given the country's hot climate.[10]

Fluorine is the 13th most abundant element in the earth's crust and can enter into groundwater as fluoride (F^-) as the result of geochemical interactions with fluoride-bearing minerals such as micas, hornblende, pyroxene, and apatite in rocks and sediments as well as from atmospheric deposition.[11] High fluoride concentrations in groundwater are often associated with rocks with low calcium content or alkaline groundwater dominated by sodium bicarbonate.[12] The residence time of groundwater, climatic conditions such as evapotranspiration and precipitation and soil pH and type can also affect fluoride dissolution.[12−17] In general, carbonate rocks act as a sink for fluoride, and measured fluoride concentrations have been found to be proportional to soil pH.[11,15,18] Irrigation has been shown to increase fluoride levels in groundwater due

Received: March 29, 2018
Revised: June 19, 2018
Accepted: July 25, 2018
Published: July 27, 2018

Table 1. Sources of Groundwater Fluoride Measurements

data source	no. of data	max./avg. conc. F (mg/L)	location	well type(s)
Central Ground Water Board[40]	10 272	26.0/0.60 ± 0.85	throughout India	dug wells, bore/tube wells
Central Ground Water Board[39]	2052	65.0/2.09 ± 2.88	throughout India	dug wells, bore/tube wells
Gupta et al.[41]	228	9.90/1.34 ± 1.35	Gujarat	dug wells, geothermal springs, hand-pumps, and tubewells
Hazarika and Bhuyan[42]	48	1.06/0.67 ± 0.18	Assam	ringwells and tubewells

to an associated increase in alkaline and sodic soils.[11] Anthropogenic sources may also play a role, such as through the application of phosphate-based fertilizers, which typically contain high levels of fluoride.[19]

High fluoride concentrations are found in groundwater on all inhabited continents around the world.[12,20] Elevated fluoride concentrations have been measured in groundwater in India since the first part of the 20th century.[21] Since then, many other studies have detected high fluoride levels in the northwest,[22−25] in the south[26−28] and in the east,[29,30] including the Gangetic Plains.[31,32] It has been estimated on the basis of dental fluorosis surveys in schools that about 62 million Indians experience the effects of fluorosis due to consuming water with high fluoride concentrations.[33]

In order to determine the extent of contamination to then aid its mitigation, geostatistical modeling with prediction variables is sometimes carried out. As opposed to the much more common practice of interpolating among known concentrations, modeling with predictor variables, including machine learning techniques, can create a more accurate model by finding relationships with the factors directly responsible for or related to the accumulation of the contaminant in question. This strategy has already been used on fluoride in groundwater on a global scale[13] and been widely applied to other contaminants, notably arsenic.[34−38]

In this paper, we use the random forest method to model fluoride concentrations in groundwater samples from throughout India to create a predictive model of areas in which fluoride concentrations exceed the WHO guideline of 1.5 mg/L. The purpose of this model is to determine the extent of fluoride contamination and to guide the localization and remediation of the problem. We also identify some of the key parameters related to the natural accumulation of fluoride in groundwater in India.

■ MATERIALS AND METHODS

Groundwater Samples. Groundwater quality analyses from throughout India were compiled almost exclusively from Central Ground Water Board of India (CGWB) surveys in 2010[39] and 2013[40] but also augmented with georeferenced samples from two other published sources (Table 1). A general lack of geographical coordinates prevented further samples from other studies being incorporated. The CGWB collected these samples at the end of the dry season shortly before the onset of the monsoon. Depending on the state, this was either in April or May. In total, 12 600 fluoride concentrations were compiled from all sources and are plotted with topography in Figure 1. In brief, the CGWB surveys collected groundwater samples from dug wells (85%) and bore/tube wells (15%) following a standardized protocol for sample collection and using a 1 L sampling size. The sampling wells are the same as those used in ongoing water-level monitoring, which were selected based on being regularly used and as evenly spaced as possible. The CGWB prefers using dug wells in their national monitoring network for the simplicity of measuring the level of the groundwater table.

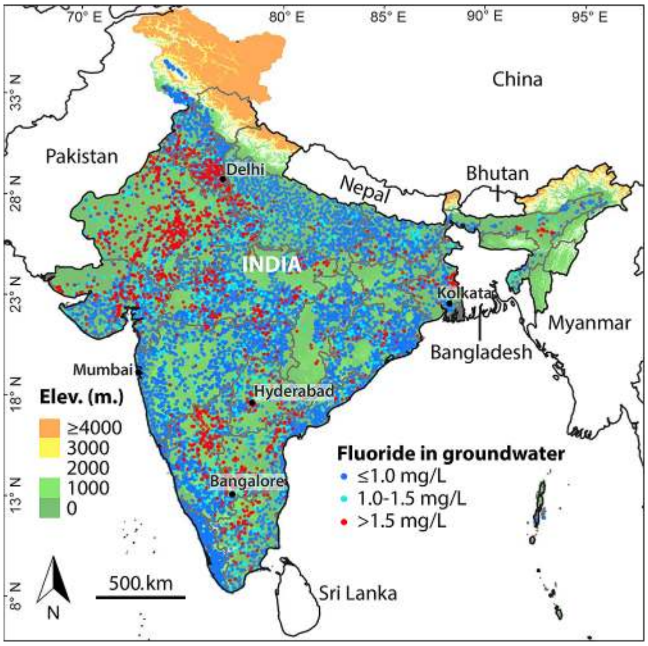


Figure 1. Fluoride concentrations from the sources listed in Table 1 (n = 12 600) and topography. This image was created in part using QGIS version 2.14.1-Essen, available under CC BY-SA 3.0 from https://www.qgis.org/en/site/.

Aside from the geographical coordinates of wells, other parameters were generally not available. Out of the total of 12 600 data points, 2895 (23%) exceed the local Indian guideline of 1.0 mg/L, 1704 (14%) exceed the WHO guideline of 1.5 mg/L and 394 (3%) are greater than 3.5 mg/L (see Figure S1 of the Supporting Information, SI).

Data Set Preparation. Prior to modeling, the data points of measured fluoride concentrations were assigned to one-square-kilometer pixels, which is the finest resolution of the independent variables used. If more than one measurement was available within a pixel, then the geometric mean was taken. This averaging reduced the number of data points from 12 600 to 11 320 for use in modeling. Of these, 1449 data points were greater than 1.5 mg/L (13%) and 9871 were equal to or less than 1.5 mg/L (87%). The data set was then converted into high and low classes by assigning zero to all fluoride concentrations ≤1.5 mg/L and one to all concentrations >1.5 mg/L. The resulting data set was then randomly split into training (80%) and testing (20%) data sets that maintain the same ratio of low and high values as in the full data set.

Variable Selection for Statistical Modeling. Twenty-five independent variables were selected for potential use in statistical modeling based on established or presumed relationships with the release and accumulation of fluoride in groundwater (Table 2). These are generally geology, soil, or

Table 2. Coefficients and Significance of Correlations and Univariate Logistic Regressions (Normalized) Based on the WHO Guideline for Fluoride in Drinking Water of 1.5 mg/L[a]

variable	resolution	correlation (p)	logistic regression (p)
Actual evapotranspiration (AET)[44]	30″	$-0.918\ (1.37 \times 10^{-06})$	$-5.26\ (1.42 \times 10^{-189})$
Aridity (PET[45]/precipitation[46])	30″	$0.956\ (2.85 \times 10^{-08})$	$8.78\ (4.30 \times 10^{-119})$
Calcisols[47]	30″	$0.829\ (1.35 \times 10^{-04})$	$5.98\ (7.74 \times 10^{-81})$
Carbonate sedimentary rocks[48]	polygon	n/a	$1.23\ (3.06 \times 10^{-04})$
Clay fraction (1.5m depth)[47]	30″	$-0.568\ (2.73 \times 10^{-02})$	$-2.18\ (5.25 \times 10^{-34})$
Cropland[49]	30″	n/a	$0.284\ (6.44 \times 10^{-06})$
Felsic igneous rocks[48]	polygon	n/a	$0.351\ (8.18 \times 10^{-05})$
Mafic igneous rocks[48]	polygon	n/a	$-0.622\ (5.07 \times 10^{-14})$
Noncarbonate sedimentary rocks[48]	polygon	n/a	$0.148\ (8.64 \times 10^{-03})$
Potential evapotranspiration (PET)[45]	30″	$0.908\ (2.93 \times 10^{-06})$	$7.24\ (1.87 \times 10^{-78})$
Precipitation[46]	30″	$-0.685\ (4.81 \times 10^{-03})$	$-13.2\ (4.32 \times 10^{-135})$
Sand fraction (1.5m depth)[47]	30″	$0.847\ (6.77 \times 10^{-05})$	$3.66\ (1.41 \times 10^{-64})$
Silt fraction (1.5m depth)[47]	30″	$-0.948\ (7.42 \times 10^{-08})$	$-2.84\ (1.30 \times 10^{-44})$
Slope[50]	30″	$-0.673\ (5.92 \times 10^{-03})$	$-4.24\ (4.46 \times 10^{-07})$
Soil pH[47,51,52]	30″	$0.893\ (7.51 \times 10^{-06})$	$4.95\ (4.39 \times 10^{-76})$
Evapotranspiration (ET), MODIS 2000−2013[53]	30″	*$-0.128\ (6.49 \times 10^{-01})$*	*$-0.328\ (1.84 \times 10^{-02})$*
Flow accumulation[50]	30″	*$0.101\ (7.21 \times 10^{-01})$*	*$-53.5\ (1.63 \times 10^{-01})$*
Gypsisols[47]	30″	*$0.456\ (8.74 \times 10^{-02})$*	*$1.87\ (2.66 \times 10^{-26})$*
Histosols[47]	30″	*$-0.210\ (4.53 \times 10^{-01})$*	*$-2.27\ (1.71 \times 10^{-11})$*
Irrigation amounts[54]	*0.5°*	$0.175\ (5.33 \times 10^{-01})$	*$0.322\ (3.51 \times 10^{-03})$*
Irrigated area[55]	*5′*	$0.460\ (8.46 \times 10^{-02})$	*$0.373\ (8.26 \times 10^{-05})$*
Metamorphic rocks[48]	*polygon*	*n/a*	*$0.0596\ (3.58 \times 10^{-01})$*
Soil cation exchange capacity[47]	30″	*$-0.331\ (2.28 \times 10^{-01})$*	*$-0.795\ (7.99 \times 10^{-06})$*
Solonchaks[47]	30″	$0.430\ (1.09 \times 10^{-01})$	*$2.64\ (1.10 \times 10^{-19})$*
Temperature[46]	30″	$0.200\ (4.76 \times 10^{-01})$	*$1.13\ (5.08 \times 10^{-03})$*

[a] Fifteen datasets met the 95% confidence interval for both statistics and were used for modeling. Those that did not are indicated in italics. Correlations are plotted in Figure S2.

climate parameters but also take into account anthropogenic effects such as land use and irrigation. The values of the independent variables were extracted from the associated GIS data sets (Table 2) at each fluoride measurement point and compiled in a table.

A subset of the initial set of variables was selected based on the statistical significance of relationships to fluoride concentrations meeting or exceeding the WHO guideline of 1.5 mg/L. Correlations were measured between each independent variable and the proportion of fluoride measurements greater than 1.5 mg/L for bins of each variable (Table 2 and Figure S2). The number of bins was determined using Sturges' formula,[43] which in this case was 15. The number of bins was in turn used to determine a fixed number of data values for each bin. Univariate logistic regressions were also run using the threshold of 1.5 mg/L (Table 2). Variables for which the correlation and/or univariate logistic regression did not fall within the 95% confidence interval (p-value ≤ 0.05) were not used in modeling. A modeling table was then assembled consisting of the fluoride concentration measurements along with the corresponding values of the remaining 15 potentially significant independent variables (Figure S3).

Random Forest Modeling. The random forest machine-learning algorithm generates an ensemble of decision trees, which are classification models that consecutively split a dependent variable (e.g., fluoride) on independent variables (nodes) and cutoff values that result in the greatest variance of the dependent variable.[56] The resulting decision tree model can then be used to predict a binary class on the basis of the values of the associated independent variables. All modeling was carried out using the R statistical programming language.[57]

Random forests grow many trees and introduce randomness to their development so as to cause trees to develop differently and thus utilize different combinations of the information contained in the input data set, ultimately resulting in a more robust model. By taking the average of the class-prediction results of the ensemble of trees, a random forest model is less sensitive to the starting conditions and avoids problems of overfitting and correlated variables masking each other, as often occurs in a single decision tree.[58] Randomness is introduced in two ways in the growing of random forest trees: (1) random selection with replacement of all data rows, which results in roughly one-third of the data (e.g., each fluoride measurement and its associated independent variables) not being selected for a given tree and referred to as out-of-bag samples (OOB),[56] and (2) a greatly restricted number of randomly selected variables made available at each node,[59] which in our case was chosen as the square root of the total number of independent variables (i.e., 3).

Since our data set of measured fluoride concentrations contains considerably more low than high values (more that are less than or equal to 1.5 mg/L) and to avoid a bias toward low values, the training data set was randomly downsampled for the growing of each tree by restricting the number of data rows made available with a low-value target variable to 1162, which is the number of high-value target variables and thus provides a balanced data set for modeling. Using 2324 data rows for each tree, a total of 1001 trees were grown to produce the random forest model. In this way, the information from all measurements was incorporated in the model, while also avoiding the problem of overemphasizing low concentrations.

DOI: 10.1021/acs.est.8b01679

Two ways to determine the accuracy of a class-prediction model include calculating its true-positive rate (sensitivity) and true-negative rate (specificity).[60] When using the model as a predictive tool, the cutoff corresponding to where sensitivity and specificity are equal results in an evenly balanced predictive accuracy of both high and low values and generally corresponds to the highest overall accuracy over all cutoffs if the testing data set is evenly balanced between high and low values. Plotting both sensitivity and specificity for the full range of model cutoff values from 0 to 1 produces a receiver operating characteristic (ROC) curve and the associated area under the ROC curve (AUC) value, which generally ranges between 0.5 (no predictive capability) and 1 (perfect predictive capability).[60]

To assess the relative importance of the predictor variables, the values of each variable in all OOB samples were randomly reassigned and the mean decrease in prediction accuracy was calculated. The greater the decrease in accuracy, the greater is the inferred importance of the variable. In addition, the mean decrease in Gini node impurity[61] was calculated over all splits for all trees. Gini purity refers to the homogeneity of the subnodes created by splitting on a given variable. The higher the average decrease in Gini node impurity, the more significant is the variable.

Multivariate Logistic Regression. The same predictor variables of the random forest model were used in 1001 logistic regressions to verify the random forest model and provide an assessment of the relationships between the predictor variables and the occurrence of fluoride concentrations exceeding 1.5 mg/L. As in the random forest model, the imbalance between the number of high and low concentration values was rectified in each logistic regression by downsampling the data rows of the larger class (low values) to match the number of high values, which is 1162 in the training data set. Since logistic regression results can be difficult to correctly interpret in the presence of strong collinearity among the predictor variables, the variables aridity, clay fraction, and soil pH were removed for this analysis on the basis of correlations exceeding 0.8 (Table S1). These variables are strongly correlated with actual evapotranspiration, sand fraction, and precipitation, respectively. All other modeling procedures, including the retaining or discarding of models based on the Hosmer-Lemeshow goodness-of-fit test[62] and the weighting and averaging of coefficients, are described by Podgorski et al.[37]

■ RESULTS AND DISCUSSION

Random Forest Model. The 15 variables initially used for modeling (Table 2) were refined in a further step by considering the importance of each individual independent variable. The carbonate sedimentary rocks variable was removed on the basis of the mean decrease in accuracy of when its values were randomly rearranged in the OOB samples. This variable was the only one to produce a mean increase in accuracy with this test, which indicates that it likely has no real relationship to the target variable.

The final random forest model was produced for the entire country as well as the neighboring countries of Bangladesh, Bhutan, Nepal, and Sri Lanka at a 1 km resolution (Figure 2a), which is derived from the resolution of the predictor data sets. The same model was applied to these immediately adjacent countries to take advantage of likely similar conditions. The model achieved an AUC of 0.84 on the test data set (Figure 3a) and an accuracy of 0.78 using a cutoff value of 0.44, which was determined by the best trade-off between sensitivity and specificity, i.e., where they are equal (Figure 3b). The accuracy at this cutoff is less than that at higher cutoff values due to the test data set being dominated by low values (87%). As a consequence, a cutoff of 1 would predict 0 for all cases and thereby yield 87% accuracy with the test data set, however all of the high values would be incorrectly classified.

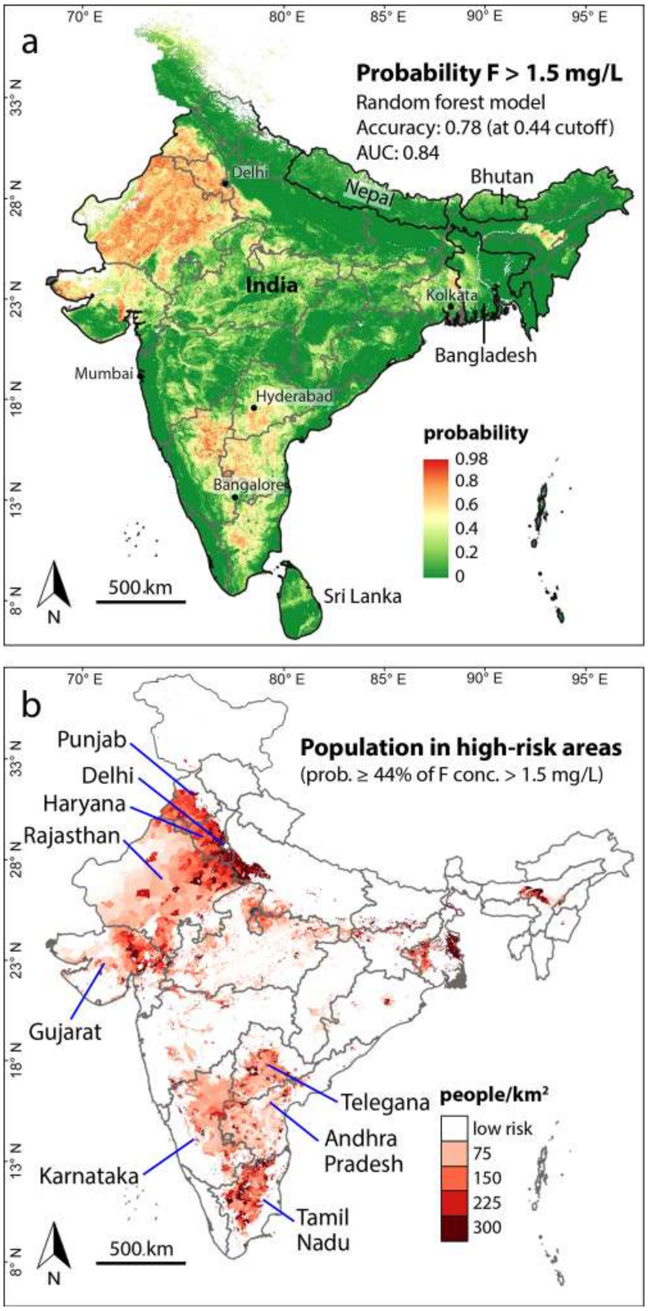


Figure 2. Random forest model of areas of aquifers in India with fluoride concentrations exceeding the WHO guideline of 1.5 mg/L. (a) Probability or mean of the 1001 random forest trees, including the neighboring countries of Bangladesh, Bhutan, Nepal, and Sri Lanka, and (b) population density of areas with a high probability (≥44%) of fluoride >1.5 mg/L (based on 2015 population statistics). This image was created in part using QGIS version 2.14.1-Essen, available under CC BY-SA 3.0 from https://www.qgis.org/en/site/.

The accuracy of 0.78 using the test data set is comparable to the internal OOB accuracy of 0.80. These two error estimates

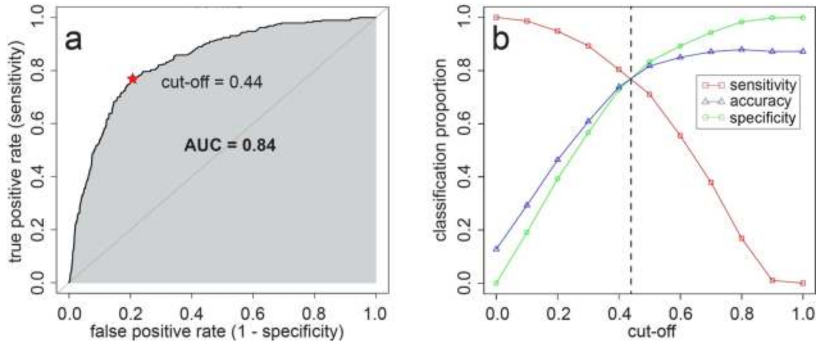


Figure 3. Random forest modeling results (a) ROC curve (b) sensitivity, specificity, and accuracy plotted against cutoff.

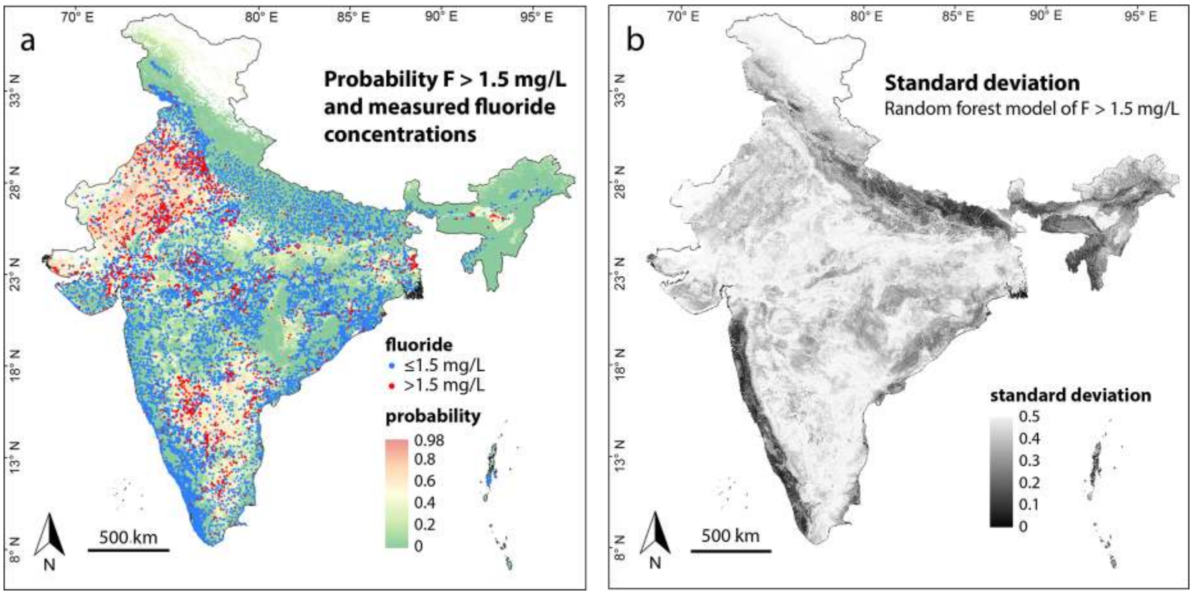


Figure 4. Validation of random forest model. (a) Probability plotted with the 11 320 fluoride concentrations of the training and testing data sets and (b) standard deviation of the mean shown in (Figure 2a). This image was created in part using QGIS version 2.14.1-Essen, available under CC BY-SA 3.0 from https://www.qgis.org/en/site/.

are consistent with each other and indicate that splitting the full data set into testing and training data sets did not create any fundamental differences between the two. The overall accuracy of the random forest model on the entire data set of 11 320 points that together comprise the training and test data sets (Figure 4a) is 0.91, which was found using a cutoff of 0.59 corresponding to an even trade-off between sensitivity and specificity. It is noteworthy that this model with its associated high accuracy was achieved using only surface parameters as predictors, although the target groundwater concentrations may originate from a considerable depth. Spatially continuous 3D data sets, such as of hydrogeology, could possible render a significant improvement to the model, if such data sets were to exist.

Prediction Map. The random forest model identifies parts of 23 states and territories having a high hazard of fluoride concentration in groundwater greater than 1.5 mg/L, comprising a total of 28% of the country. Two large areas of the country are strongly affected: the northwestern states/territories of Delhi, Gujarat, Haryana, Punjab, and Rajasthan and the southern states of Andhra Pradesh, Karnataka, Tamil Nadu, and Telangana. More fragmented pockets of predicted high fluoride concentrations are spread throughout the central and eastern parts of the country. The percentage of area in neighboring countries found to have a high fluoride hazard is as follows: Bangladesh (5%), Bhutan (0%), Nepal (0%), and Sri Lanka (4%).

Model Uncertainty. The probabilities of the random forest model shown in Figure 2a are the averages of the 1001 individual trees grown for each pixel. Since each tree can have an outcome of either 0 or 1, the corresponding standard deviations are a quadratic function of the mean values (Figure S4a). As a consequence, low standard deviations are associated with high and low means or probabilities. Figure 4b displays the model standard deviations, which are an indicator of the model's certainty. That is, the less variation in the votes of the trees for a given pixel, the stronger the decision of the model in selecting a high or low value.

The lowest standard deviations, i.e., greatest model certainty, are found throughout much of the Western Ghats (SW coast), Ganges River plain (north-central India), and far eastern India, which is also where the lowest probabilities are located. The histogram in Figure S4b confirms that the model contains more low probabilities than high ones. Other low probability areas, such as southern Gujarat or east-central India have somewhat higher standard deviations than in the aforementioned areas. Model certainties associated with the high probabilities of Figure 2a are greater in Delhi, Haryana,

Table 3. Number of High (>1.5 mg/L) and Total Measurements as well as Area and Population of India with High Probability (≥44%) of F > 1.5 mg/L by State/Territory[a]

state/territory	measurements high/total (%)	area classified as high risk	pop. in high risk area in 2015[66] (% of total)
Andaman and Nicobar	0/50 (0%)	0%	0 (0%)
Andhra Pradesh	99/564 (17.6%)	51%	7 516 234 (14%)
Arunachal Pradesh	0/11 (0.0%)	0%	0 (0%)
Assam	22/127 (17.3%)	15%	2 033 836 (6%)
Bihar	19/385 (4.9%)	4%	1 051 588 (1%)
Chandigarh	0/3 (0.0%)	0%	0 (0%)
Chhattisgarh	22/360 (6.1%)	6%	455 672 (2%)
Dadra and Nagar Haveli	0/5 (0.0%)	0%	0
Daman and Diu	27/103 (26.2%)	20%	19 306 (7%)
Delhi	2/66 (3.0%)	45%	1 894 174 (11%)
Goa	159/767 (20.7%)	0%	0 (0%)
Gujarat	127/364 (34.9%)	50%	10 988 520 (17%)
Haryana	0/41 (0.0%)	81%	8 152 992 (30%)
Himachal Pradesh	3/223 (1.3%)	0%	2952 (0%)
Jammu and Kashmir	9/133 (6.8%)	10%	185 938 (1%)
Jharkhand	187/1079 (17.3%)	9%	1 302 178 (4%)
Karnataka	2/712 (0.3%)	44%	9 260 131 (14%)
Kerala	244/1912 (12.8%)	0%	15 510 (0%)
Madhya Pradesh	34/1476 (2.3%)	23%	7 343 093 (9%)
Maharashtra	34/950 (3.6%)	9%	3 271 034 (3%)
Manipur	0/4 (0.0%)	1%	108 283 (3%)
Meghalaya	48/276 (17.4%)	0%	79 (0%)
Mizoram	477/1335 (35.7%)	0%	0 (0%)
Nagaland	67/394 (17.0%)	9%	153 065 (8%)
Orissa	39/300 (13.0%)	3%	552 800 (0%)
Puducherry	0/22 (0.0%)	0%	0 (0%)
Punjab	35/684 (5.1%)	56%	5 008 113 (17%)
Rajasthan	48/254 (18.9%)	81%	25 190 370 (33%)
Sikkim	0/50 (0%)	0%	0 (0%)
Tamil Nadu	99/564 (17.6%)	41%	10 890 790 (14%)
Telangana	0/11 (0.0%)	48%	7 332 835 (19%)
Tripura	22/127 (17.3%)	0%	0 (0%)
Uttar Pradesh	19/385 (4.9%)	13%	7 952 795 (4%)
Uttaranchal	0/3 (0.0%)	0%	9605 (0%)
West Bengal	22/360 (6.1%)	23%	8 307 174 (9%)
total	0/5 (0.0%)	28%	118 999 065 (9%)

[a] The population numbers are from non-urban areas and account for 46% of rural areas not having access to piped water and 85% utilizing groundwater for drinking.

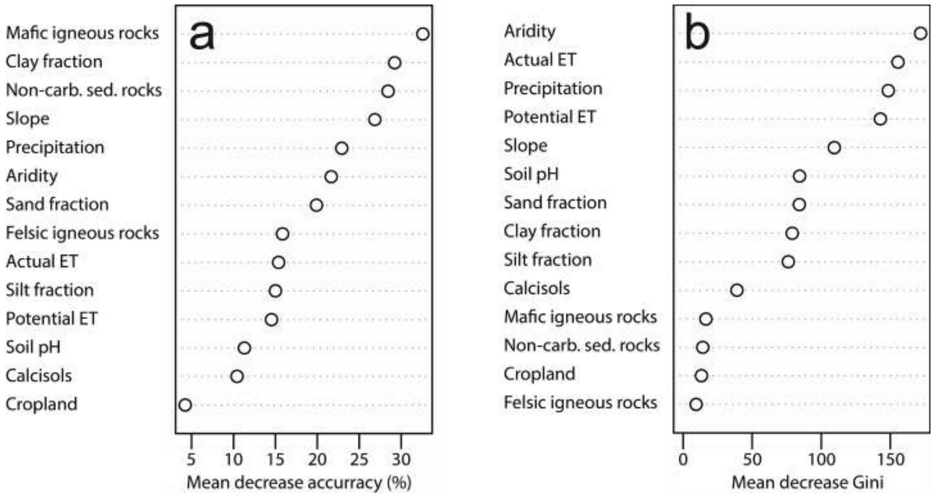


Figure 5. Measures of random forest variable importance. (a) Mean decrease in accuracy in OOB samples when using random values of each variable. (b) Mean decrease in Gini node impurity for each variable.[61]

Punjab, and Rajasthan than in the high-probability areas of Gujarat or the south or east (Figure 3b). This highlights that more spatially detailed water sample testing may be required in these latter areas. Likewise, the model is also less certain in the areas of low to moderate probability throughout the rest of the country.

Population in High Hazard Areas. Figure 2b shows the density of population using unmanaged groundwater sources in high hazard areas. Accordingly, the total population potentially exposed to fluoride in groundwater exceeding 1.5 mg/L (greater than or equal to 44%) is nearly 120 million (118 999 065). This is based on 2015 population counts,[63] considers only nonurban areas[49] and takes into account that approximately 46% of rural populations do not have access to piped water[64] and that 85% of rural populations consume groundwater for drinking. Table 3 lists the proportion of land area of each state/territory that is classified as high hazard and the number and percentage of population potentially at risk as well as the number of high (>1.5 mg/L) to total measurements used in modeling. The figure of approximately 120 million people at risk could be further refined with more detailed statistics on piped water, which is generally ensured to contain a low concentration of fluoride. Furthermore, consideration of the nutritional status of populations in different areas could highlight differences in susceptibility to fluorosis. The prediction model confirms the presence of high fluoride concentrations found in other recently published studies, e.g., the Sonbhadra district of Uttar Pradesh,[29,30] the Bhilwara district,[22] Pokhran area,[24] and Thar Desert[23,25] of Rajasthan, the Chandauli-Varanasi region in Uttar Pradesh,[65] the Nalgonda district of Telangana,[26] the Guntur district[27] and Anantapur district[28] of Andhra Pradesh, and the Jamui district of Bihar.[32]

Predictor Variables. Figure 5 contains rankings of the importance of the predictor variables in the random forest model. The mean decrease in accuracy associated with randomizing a variable in OOB samples (Figure 5a) and the mean decrease in Gini node impurity (Figure 5b) are consistent in assessing the relative importance for roughly half of the predictor variables: aridity and precipitation (high importance); slope, sand fraction, and silt fraction (moderate importance); and calcisols and cropland (low importance).

Logistic Regression Model. To better assess the relationships between the predictor variables and high fluoride concentrations, the same 11 of the predictor variables from the random forest model (all except aridity, clay fraction, and soil pH) were used in 1001 logistic regressions, 452 of which passed the goodness-of-fit test. Each variable was standardized by subtracting the variable's minimum value and then dividing by the difference between its high and low values, resulting in each variable ranging between 0 and 1. The composite logistic regression model also has a high AUC of 0.78, although accuracy is only 0.72 (at a cutoff of 0.66). The weighted coefficients and standard deviations of each variable along with their frequencies in the individual logistic regressions are listed in Table 4. Five variables appear in most or all of the 452 acceptable logistic regressions: AET, PET, precipitation, mafic igneous rocks, and noncarbonate sedimentary rocks. These are shown in Figure S3.

Main Predictor Variables. The correlations and logistic regressions (Tables 2 and 4) clearly indicate that drier conditions are associated with the occurrence of high fluoride concentrations. That is, AET and precipitation negatively predict high fluoride concentrations, whereas PET positively predicts fluoride. As per the logistic regression, the two most prominent geology variables, mafic igneous rocks and non-carbonate sedimentary rocks, in general negatively predict the presence of high fluoride concentrations (except in Rajasthan) and are also two of the most important variables in the random forest model according to the accuracy test (Figure 5a). Felsic igneous rocks generally positively predict high fluoride concentrations in the logistic regressions, however this variable was present roughly only two-thirds as frequently as the other two geology variables. Although it would be expected that carbonate sedimentary rocks would be positively related to high fluoride concentrations, very little of such rocks is found in India. As such, this variable proved to be ineffective in the initial random forest modeling of fluoride.

Table 4. Coefficients, Standard Deviations and Frequencies of the Standardized Predictor Variables of 1001 Logistic Regressions Run Using a Threshold of 1.5 mg/L for Fluoride[a]

variable	coefficient	standard deviation	freq in logistic regressions
(Intercept)	−1.06	0.82	452
Actual ET	−1.96	0.47	407
Calcisols	2.05	0.61	225
Cropland	0.23	0.04	113
Felsic igneous rocks	0.43	0.11	289
Mafic igneous rocks	−0.66	0.15	450
Noncarb. sed. rocks	−0.41	0.10	442
Potential ET	2.96	0.60	451
Precipitation	−5.17	1.15	452
Sand fraction	1.35	0.38	316
Silt fraction	0.80	0.49	67
Slope	−1.97	1.25	59

[a] 452 logistic regressions passed the Hosmer-Lemeshow goodness-of-fit test, which, along with the frequencies, was used to weight the coefficients and standard deviations.

As a test, the same five dominant variables of the logistic regression were used to create a random forest model in the same manner as described previously. The AUC of the random forest model with five variables was 0.81 (versus 0.84 with 14 variables) and the accuracy (at a cutoff of 0.41) was 0.74 (versus 0.78 with 14 variables). That is, the 14-variable random forest model (Figure 2a) outperforms the random forest model using only the five dominant variables of the logistic regression, which itself is the least-performing of the three models. The areas of high hazard as predicted by each of these models are shown in Figure S5.

The positive though weak correlation with cropland may be related to the downward leaching of fluoride contained in phosphate fertilizers and/or an increase in sodicity brought about by irrigation.[67] However, neither the amount of irrigation nor the land area under irrigation appeared to be significant in our initial screening of variables and were therefore excluded from further consideration. As expected, soil pH positively correlates with high fluoride concentrations (Table 2), with soil pH itself being strongly inversely correlated with precipitation (Table S1). Slope is inversely related to high fluoride and is a proxy for the hydrological gradient and thereby the residence time of groundwater, which increases the interaction time between fluoride-bearing minerals and aquifer pore water.[12] This is particularly relevant when the water is saturated with calcite, which is consistent with the positive coefficient found for calcisols.

Implications. Our fluoride prediction map and associated population risk assessment represent a significant improvement in understanding of both the detailed locations of high fluoride concentrations in groundwater as well as the number of people potentially affected. This is largely due to the fact that government reports and other studies on the countrywide occurrence of fluoride generally provide information only at the district level. For example, Chakraborti et al.[68] calculated 411.4 million people at risk based on the total population living in districts where fluoride had been measured in groundwater greater than 1.5 mg/L. The present study is much more precise by applying the derived statistical model to produce maps at a resolution of 1 km^2. Since the average size of a district is 4650 km^2, this represents an improvement in resolution of over 3 orders of magnitude. Furthermore, we have utilized the best-available statistics to account for groundwater usage and access to managed water sources. In addition to identifying large rural populations at risk of fluorosis, the newly produced maps can aid in prioritizing the application of mitigation funding in areas where the fluoride hazard is elevated and groundwater is being extracted at a high rate relative to recharge. The latter is relevant since the groundwater being pumped in such circumstances has generally had a longer residence time in the aquifer during which to accumulate fluoride.

Once safe and unsafe water sources have been identified, effective mitigation measures can be taken. If wells with high and low fluoride concentrations exist within close proximity to one another, then the simplest measure is to close the unsafe well or restrict its use to only nonconsumption activities, such as washing. Other possibilities include the dilution of fluoride-contaminated groundwater with surface water or the harvesting and use of rainwater. If such options do not exist or are not feasible, then various membrane and adsorption filtering methods are available.[69]

While our fluoride prediction map (Figure 2a) provides a highly accurate determination of the locations of high-fluoride containing aquifers and highlights areas of particular concern, it is not intended to be a definitive guide to safe and unsafe water sources. Especially since small-scale aquifer heterogeneities exist that cannot be modeled with the available data and resolution, particularly in hard-rock aquifers, as well as a lack of depth information, it is still necessary to test individual wells to determine the concentration of fluoride. In this regard, the presented fluoride prediction map can be invaluable in determining where to focus testing efforts and resources.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.est.8b01679.

Fluoride concentration data, predictor variables, and modeling results (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

*E-mail: joel.podgorski@eawag.ch (J.E.P.).
*E-mail: michael.berg@eawag.ch (M.B.).

ORCID

Joel E. Podgorski: 0000-0003-2522-1021
Michael Berg: 0000-0002-7342-4061

Notes

The authors declare no competing financial interest.
$^\nabla$D.S.: Former member.

■ ACKNOWLEDGMENTS

We acknowledge our colleague A. Johnson who initiated this collaboration but passed away in 2015. We thank the Swiss Agency for Development and Cooperation for their financial support as well as the Swiss Federal Institute of Aquatic Science and Technology (Eawag), the Indian National Environmental Engineering Research Institute (NEERI), and the Indian Central Ground Water Board (CGWB).

■ REFERENCES

(1) World Bank Deep Wells and Prudence. *Towards Pragmatic Action for Addressing Groundwater Overexploitation in India*; Washington, DC, 2010.

(2) Tiwari, V.; Wahr, J.; Swenson, S., Dwindling groundwater resources in northern India, from satellite gravity observations. *Geophys. Res. Lett.* **2009**, *36* (18) DOI: 10.1029/2009GL039401.

(3) Rodell, M.; Velicogna, I.; Famiglietti, J. S. Satellite-based estimates of groundwater depletion in India. *Nature* **2009**, *460* (7258), 999.

(4) Suhag, R. *Overview of Ground Water in India*; PRS Legislative Research: http://www.prsindia.org/, 2016.

(5) Saha, D.; Marwaha, S.; Mukherjee, A. Groundwater Resources and Sustainable Management Issues in India. In *Clean and Sustainable Groundwater in India*; Springer: Singapore, 2017.

(6) Saha, D.; Shekhar, S.; Ali, S.; Vittala, S. S.; Raju, N. J. Recent Hydrogeological Research in India. *Proc. Indian Natl. Sci. Acad., Part A* **2016**, *82* (3), 787−803.

(7) WHO Expert Committee on Oral Health Status Fluoride Use, Fluorides and Oral Health: *Report of the WHO Expert Committee on Oral Health Status and Fluoride Use*; World Health Organization: 1994; Vol. *846*.

(8) World Health Organization *Guidelines for Drinking-Water Quality*, 4th ed.; World Health Organization: 2011.

(9) Reddy, G. S.; Srikantia, S. Effect of dietary calcium, vitamin C and protein in development of experimental skeletal fluorosis. I. Growth, serum chemistry, and changes in composition, and radiological appearance of bones. *Metab., Clin. Exp.* **1971**, *20* (7), 642−649.

(10) Shah, S.; Bandekar, K. IS 10500: 91 Drinking Water compared to WHO Guidelines (1993) *J. Indian Waterworks Assoc.* **1998**, *30*, 179−184.

(11) Jacks, G.; Bhattacharya, P.; Chaudhary, V.; Singh, K. Controls on the genesis of some high-fluoride groundwaters in India. *Appl. Geochem.* **2005**, *20* (2), 221−228.

(12) Edmunds, W. M.; Smedley, P. L. Fluoride in natural waters. In *Essentials of Medical Geology*; Springer: Dordrecht, 2013; pp 311−336.

(13) Amini, M.; Mueller, K.; Abbaspour, K. C.; Rosenberg, T.; Afyuni, M.; Møller, K. N.; Sarr, M.; Johnson, C. A. Statistical modeling of global geogenic fluoride contamination in groundwaters. *Environ. Sci. Technol.* **2008**, *42* (10), 3662−3668.

(14) Ayoob, S.; Gupta, A. K. Fluoride in drinking water: a review on the status and stress effects. *Crit. Rev. Environ. Sci. Technol.* **2006**, *36* (6), 433−487.

(15) Saxena, V.; Ahmed, S. Inferring the chemical parameters for the dissolution of fluoride in groundwater. *Environ. Geol.* **2003**, *43* (6), 731−736.

(16) Hudak, P. F. Fluoride levels in Texas groundwater. *J. Environ. Sci. Health, Part A: Toxic/Hazard. Subst. Environ. Eng.* **1999**, *34* (8), 1659−1676.

(17) Valenzuela-Vasquez, L.; Ramirez-Hernandez, J.; Reyes-Lopez, J.; Sol-Uribe, A.; Lazaro-Mancilla, O. The origin of fluoride in groundwater supply to Hermosillo City, Sonora, Mexico. *Environ. Geol.* **2006**, *51* (1), 17−27.

(18) Chhabra, R.; Singh, A.; Abrol, I. Fluorine in Sodic Soils 1. *Soil Science Society of America Journal* **1980**, *44* (1), 33−36.

(19) Chaney, R. L. Food safety issues for mineral and organic fertilizers. In *Advances in Agronomy*; Academic Press, 2012; Vol. *117*, pp 51−116.

(20) Ali, S.; Thakur, S. K.; Sarkar, A.; Shekhar, S. Worldwide contamination of water by fluoride. *Environ. Chem. Lett.* **2016**, *14* (3), 291−315.

(21) Shortt, H.; Pandit, C.; Raghavachari, R. S. T. Endemic fluorosis in the Nellore district of South India. *Indian Med. Gazette* **1937**, *72* (7), 396.

(22) Hussain, J.; Husain, I.; Arif, M. Fluoride contamination in groundwater of central Rajasthan, India and its toxicity in rural habitants. *Toxicol. Environ. Chem.* **2013**, *95* (6), 1048−1055.

(23) Singh, C. K.; Mukherjee, S. Aqueous geochemistry of fluoride enriched groundwater in arid part of Western India. *Environ. Sci. Pollut. Res.* **2015**, *22* (4), 2668−2678.

(24) Singh, C. K.; Rina, K.; Singh, R.; Shashtri, S.; Kamal, V.; Mukherjee, S. Geochemical modeling of high fluoride concentration in groundwater of Pokhran area of Rajasthan, India. *Bull. Environ. Contam. Toxicol.* **2011**, *86* (2), 152−158.

(25) Singh, C. K.; Kumari, R.; Singh, N.; Mallick, J.; Mukherjee, S. Fluoride enrichment in aquifers of the Thar Desert: controlling factors and its geochemical modelling. *Hydrological Processes* **2013**, *27* (17), 2462−2474.

(26) Reddy, A.; Reddy, D.; Rao, P.; Prasad, K. M. Hydrogeochemical characterization of fluoride rich groundwater of Wailpalli watershed, Nalgonda District, Andhra Pradesh, India. *Environ. Monit. Assess.* **2010**, *171* (1−4), 561−577.

(27) Rao, N. S.; Subrahmanyam, A.; Rao, G. B. Fluoride-bearing groundwater in Gummanampadu sub-basin, Guntur district, Andhra Pradesh, India. *Environ. Earth Sci.* **2013**, *70* (2), 575−586.

(28) Padhi, S.; Muralidharan, D. Fluoride occurrence and mobilization in geo-environment of semi-arid Granite watershed in southern peninsular India. *Environ. Earth Sci.* **2012**, *66* (2), 471−479.

(29) Raju, N. J. Prevalence of fluorosis in the fluoride enriched groundwater in semi-arid parts of eastern India: Geochemistry and health implications. *Quaternary International* **2017**, *443*, 265−278.

(30) Raju, N. J.; Dey, S.; Gossel, W.; Wycisk, P. Fluoride hazard and assessment of groundwater quality in the semi-arid Upper Panda River basin, Sonbhadra district, Uttar Pradesh, India. *Hydrol. Sci. J.* **2012**, *57* (7), 1433−1452.

(31) Saha, D.; Sarangam, H.; Dwivedi, S. N.; Bhartariya, K. J. Evaluation of hydrochemical processes in the arsenic contaminated alluvial aquifers in parts of Mid-Ganga basin, Bihar, India, Eastern India. *Environ. Earth Sci.* **2010**, *61*, 799−811.

(32) Kumar, S.; Venkatesh, A.; Singh, R.; Udayabhanu, G.; Saha, D. Geochemical signatures and isotopic systematics constraining dynamics of fluoride contamination in groundwater across Jamui district, Indo-Gangetic alluvial plains. *Chemosphere* **2018**, *205*, 493−505.

(33) Andezhath, S. K.; Ghosh, G. Fluorosis management in India: the impact due to networking between health and rural drinking water supply agencies. *IAHS Publication (International Association of Hydrological Sciences)* **2000**, *260*, 159−165.

(34) Winkel, L.; Berg, M.; Amini, M.; Hug, S. J.; Johnson, C. A. Predicting groundwater arsenic contamination in Southeast Asia from surface parameters. *Nat. Geosci.* **2008**, *1* (8), 536−542.

(35) Amini, M.; Abbaspour, K. C.; Berg, M.; Winkel, L.; Hug, S. J.; Hoehn, E.; Yang, H.; Johnson, C. A. Statistical modeling of global geogenic arsenic contamination in groundwater. *Environ. Sci. Technol.* **2008**, *42* (10), 3669−3675.

(36) Rodríguez-Lado, L.; Sun, G.; Berg, M.; Zhang, Q.; Xue, H.; Zheng, Q.; Johnson, C. A. Groundwater arsenic contamination throughout China. *Science* **2013**, *341* (6148), 866−868.

(37) Podgorski, J. E.; Eqani, S. A. M. A. S.; Khanam, T.; Ullah, R.; Shen, H.; Berg, M. Extensive arsenic contamination in high-pH unconfined aquifers in the Indus Valley. *Sci. Advances* **2017**, *3*, (8).e1700935

(38) Ayotte, J. D.; Medalie, L.; Qi, S. L.; Backer, L. C.; Nolan, B. T. Estimating the high-arsenic domestic-well population in the conterminous United States. *Environ. Sci. Technol.* **2017**, *51* (21), 12443−12454.

(39) Jha, B. Ground water quality in shallow aquifers of India. *Central Ground Water Board, Ministry of Water Resources*; Govt. of India 2010.

(40) Gupta, S. *Annual Report 2013−14*; Government of India Central Ground Water Board: Faridabad, India, 2015.

(41) Gupta, S.; Deshpande, R.; Agarwal, M.; Raval, B. Origin of high fluoride in groundwater in the North Gujarat-Cambay region, India. *Hydrogeol. J.* **2005**, *13* (4), 596−605.

(42) Hazarika, S.; Bhuyan, B. Fluoride, arsenic and iron content of groundwater around six selected tea gardens of Lakhimpur District, Assam, India. *Arch Appl. Sci. Res.* **2013**, *5* (1), 57−61.

(43) Sturges, H. A. The choice of a class interval. *J. Am. Stat. Assoc.* **1926**, *21* (153), 65−66.

(44) Trabucco, A.; Zomer, R. *Global Soil Water Balance Geospatial Database*. CGIAR Consortium for Spatial Information, *Published online, available from the CGIAR-CSI GeoPortal at*: http://www.cgiar-csi.org (*last access: January* 2013) 2010.

(45) Trabucco, A.; Zomer, R. J. *Global Aridity Index (Global-Aridity) And Global Potential Evapo-Transpiration (Global-PET) Geospatial Database*; CGIAR Consortium for Spatial Information, 2009.

(46) Hijmans, R. J.; Cameron, S. E.; Parra, J. L.; Jones, P. G.; Jarvis, A. Very high resolution interpolated climate surfaces for global land areas. *International journal of climatology* **2005**, *25* (15), 1965−1978.

(47) *ISRIC—World Soil Information, SoilGrids: an automated system for global soil mapping*. Available for download at http://soilgrids1km.isric.org. In 2013.

(48) Hartmann, J.; Moosdorf, N. The new global lithological map database GLiM: A representation of rock properties at the Earth surface. *Geochem., Geophys., Geosyst.* **2012**, *13*, (12). DOI: 10.1029/2012GC004370

(49) Friedl, M. A.; Sulla-Menashe, D.; Tan, B.; Schneider, A.; Ramankutty, N.; Sibley, A.; Huang, X. MODIS Collection 5 global land cover: Algorithm refinements and characterization of new datasets. *Remote sensing of Environment* **2010**, *114* (1), 168−182.

(50) Lehner, B.; Verdin, K.; Jarvis, A. HydroSHEDS Technical Documentation. In *World Wildlife Fund US*: Washington, DC., 2006; p Available at http://hydrosheds.cr.usgs.gov.

(51) FAO/IIASA/ISRIC/ISS-CAS/JRC *Harmonized World Soil Database* (version 1.2). In FAO, Rome, Italy and IIASA, Laxenburg, Austria, 2012.

(52) *IGBP-DIS, SoilData(V.0) A Program for Creating Global Soil-Property Databases*; IGBP Global Soils Data Task, France,1998.

(53) Mu, Q.; Heinsch, F. A.; Zhao, M.; Running, S. W. Development of a global evapotranspiration algorithm based on MODIS and global meteorology data. *Remote sensing of Environment* **2007**, *111* (4), 519−536.

(54) Vörösmarty, C. J.; Léveque, C.; Revenga, C.; Bos, R.; Caudill, C.; Chilton, J.; Douglas, E.; Meybeck, M.; Prager, D.; Balvanera, P. Fresh water. *Millennium Ecosystem Assessment* **2005**, *1*, 165−207.

(55) Siebert, S.; Henrich, V.; Frenken, K.; Burke, J. *Global Map of Irrigation Areas*, version 5. In Rheinische Friedrich-Wilhelms-University, B., Germany/Food and Agriculture Organization of the United Nations, Rome, Italy, Ed. Version: 2013.

(56) Breiman, L. Random forests. *Machine learning* **2001**, *45* (1), 5−32.

(57) R Core Team. *R: A Language and Environment for Statistical Computing*; Vienna, Austria, 2014.

(58) Ho, T. K. In *Random Decision Forests, Document Analysis and Recognition, 1995, Proceedings of the Third International Conference on, 1995*; IEEE: 1995; pp 278−282.

(59) Hastie, T. T.; Friedman, J. R. *The Elements of Statistical Learning*, 2nd ed.; Springer: New York, 2008.

(60) Fawcett, T. An introduction to ROC analysis. *Pattern recognition letters* **2006**, *27* (8), 861−874.

(61) Breiman, L. *Classification and Regression Trees*; Routledge: New York, 1984.

(62) Hosmer, Jr, D. W.; Lemeshow, S. *Applied Logistic Regression*; John Wiley & Sons: 2004.

(63) Center for International Earth Science Information Network—CIESIN—Columbia University, Gridded Population of the World, Version 4 (GPWv4): Population Count Adjusted to Match 2015 Revision of UN WPP Country Totals. In *NASA Socioeconomic Data and Applications Center (SEDAC)*; Palisades, NY, 2017.

(64) Reddy, K. N. *Revised guidelines of National Water Quality Sub-Mission*; Government of India, Ministry of Drinking Water and Sanitation: New Delhi, 2017.

(65) Singh, S.; Raju, N. J.; Ramakrishna, C. Evaluation of groundwater quality and its suitability for domestic and irrigation use in parts of the Chandauli-Varanasi region, Uttar Pradesh, India. *J. Water Resour. Prot.* **2015**, *7* (07), 572.

(66) Center for International Earth Science Information Network—CIESIN—Columbia University, Gridded Population of the World, Version 4 (GPWv4): Population Density Adjusted to Match 2015 Revision UN WPP Country Totals. In *NASA Socioeconomic Data and Applications Center (SEDAC)*; Palisades, NY, 2016.

(67) Rengasamy, P.; Olsson, K. Sodicity and soil structure. *Aust. J. Soil Res.* **1991**, *29* (6), 935−952.

(68) Chakraborti, D.; Das, B.; Murrill, M. T. Examining India's groundwater quality management. *Environ. Sci. Technol.* **2011**, *45* (1), 27−33.

(69) Mohapatra, M.; Anand, S.; Mishra, B. K.; Giles, D. E.; Singh, P. Review of fluoride removal from drinking water. *J. Environ. Manage.* **2009**, *91* (1), 67−77.