RESEARCH ARTICLE

WATER RESOURCES

Global threat of arsenic in groundwater

Joel Podgorski[1,2]* and Michael Berg[1,3]*

Naturally occurring arsenic in groundwater affects millions of people worldwide. We created a global prediction map of groundwater arsenic exceeding 10 micrograms per liter using a random forest machine-learning model based on 11 geospatial environmental parameters and more than 50,000 aggregated data points of measured groundwater arsenic concentration. Our global prediction map includes known arsenic-affected areas and previously undocumented areas of concern. By combining the global arsenic prediction model with household groundwater-usage statistics, we estimate that 94 million to 220 million people are potentially exposed to high arsenic concentrations in groundwater, the vast majority (94%) being in Asia. Because groundwater is increasingly used to support growing populations and buffer against water scarcity due to changing climate, this work is important to raise awareness, identify areas for safe wells, and help prioritize testing.

The natural, or geogenic, occurrence of arsenic in groundwater is a global problem with wide-ranging health effects for humans and wildlife. Because it is toxic and does not serve any beneficial metabolic function, inorganic arsenic (the species present in groundwater) can lead to disorders of the skin and vascular and nervous systems, as well as cancer (*1*, *2*). The major source of inorganic arsenic in the diet is through arsenic-contaminated water, although ingestion through food, particularly rice, represents another important route of exposure (*3*). As a consequence, the World Health Organization (WHO) has set a guideline arsenic concentration of 10 μg/liter in drinking water (*4*).

At least trace amounts of arsenic occur in virtually all rocks and sediments around the world (*5*). However, in most of the large-scale cases of geogenic arsenic contamination in groundwater, arsenic accumulates in aquifers composed of recently deposited alluvial sediments. Under anoxic conditions, arsenic is released from the microbial and/or chemical reductive dissolution of arsenic-bearing iron(III) minerals in the aquifer sediments (*6*–*9*). Under oxidizing, high-pH conditions, arsenic can also desorb from iron and aluminum hydroxides (*10*). Furthermore, aquifers in flat-lying sedimentary sequences generally have a small hydraulic gradient, causing groundwater to flow slowly. This longer groundwater residence time allows dissolved arsenic to accumulate and its concentration to increase. Other processes responsible for arsenic release into groundwater include oxidation of arsenic-bearing sulfide minerals as well as release from arsenic-enriched geothermal deposits.

That arsenic is generally not included in the standard suite of tested water quality parameters (*11*) and is not detected by the human senses means that arsenic is regularly being discovered in new areas. Since one of the greatest occurrences of geogenic groundwater arsenic was discovered in 1993 in the Bengal delta (*5*, *12*, *13*), high arsenic concentrations have been detected all around the world, with hot spots including Argentina (*14*–*17*), Cambodia (*18*, *19*), China (*20*–*22*), India (*23*–*25*), Mexico (*26*, *27*), Pakistan (*28*, *29*), the United States (*30*, *31*), and Vietnam (*32*, *33*).

To help identify areas likely to contain high concentrations of arsenic in groundwater, several researchers have used statistical learning methods to create arsenic prediction maps based on available datasets of measured arsenic concentrations and relevant geospatial parameters. Previous studies have focused on Burkina Faso (*34*), China (*21*, *35*), South Asia (*29*, *36*), Southeast Asia (*37*), the United States (*31*, *38*, *39*), and the Red River delta in Vietnam (*33*), as well as sedimentary basins around the world (*40*). The predictor variables used in these studies generally include various climate and soil parameters, geology, and topography (table S3).

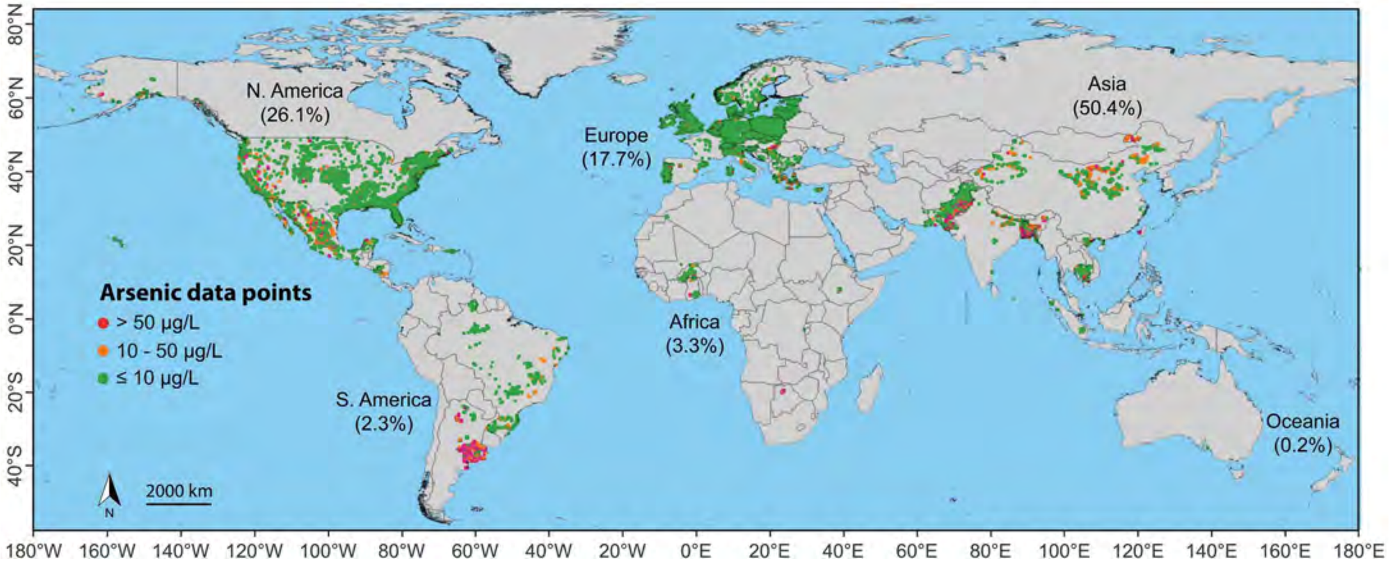


Fig. 1. Arsenic concentrations, excluding those known to originate from a depth greater than 100 m. Values are from the sources listed in table S1. The geographical distribution of data is indicated by continent.

[1] Department of Water Resources and Drinking Water, Eawag, Swiss Federal Institute of Aquatic Science and Technology, 8600 Dübendorf, Switzerland. [2] Department of Earth and Environmental Sciences, University of Manchester, Manchester M13 9PL, UK. [3] UNESCO Chair on Groundwater Arsenic within the 2030 Agenda for Sustainable Development and School of Civil Engineering and Surveying, University of Southern Queensland, Toowoomba, QLD 4350, Australia.
*Corresponding author. Email: joel.podgorski@eawag.ch (J.P.); michael.berg@eawag.ch (M.B.)

Taking advantage of the increasing availability of high-resolution datasets of relevant environmental parameters, we use statistical learning to model what to our knowledge is the most spatially extensive compilation of arsenic measurements in groundwater assembled, which makes a global model possible. To focus on health risks, we consider the probability of arsenic in groundwater exceeding the WHO guideline. For this, we have chosen the random forest method, which our preliminary tests showed to be highly effective in addressing this classification problem. We use the resulting model to produce the most accurate and detailed global prediction map to date of geogenic groundwater arsenic, which can be used to help identify previously unknown areas of arsenic contamination as well as more clearly delineate the scope of this global problem and considerably increase awareness.

Results

Random forest modeling

We aggregated data from nearly 80 studies of arsenic in groundwater (see table S1 for references and statistics) into a single dataset ($n >$ 200,000). Averaging into 1-km^2 pixels resulted in more than 55,000 arsenic data points for use in modeling based on groundwater samples not known to originate from greater than 100-m depth (Fig. 1).

To create the simplest and most accurate model, an initial set of 52 potentially relevant environmental predictor variables was iteratively reduced in consideration of their relative importance and impact on the accuracy of a succession of random forest models. The final selection of 11 predictor variables (table S2) includes several soil parameters (topsoil clay, subsoil sand, pH, and fluvisols), all of the climate variables (precipitation, actual and potential evapotranspiration, and combinations thereof, as well as temperature), and the topographic wetness index. By contrast, none of the geology variables proved to be statistically important. This is not to imply that geology does not play a role in geogenic arsenic accumulation, but rather that the particular geology variables tested were not as relevant as the other variables. This may be due to the coarse nature of the geological maps, which are standardized for the entire world. Although the number of predictor variables was reduced by nearly 80%, both the area

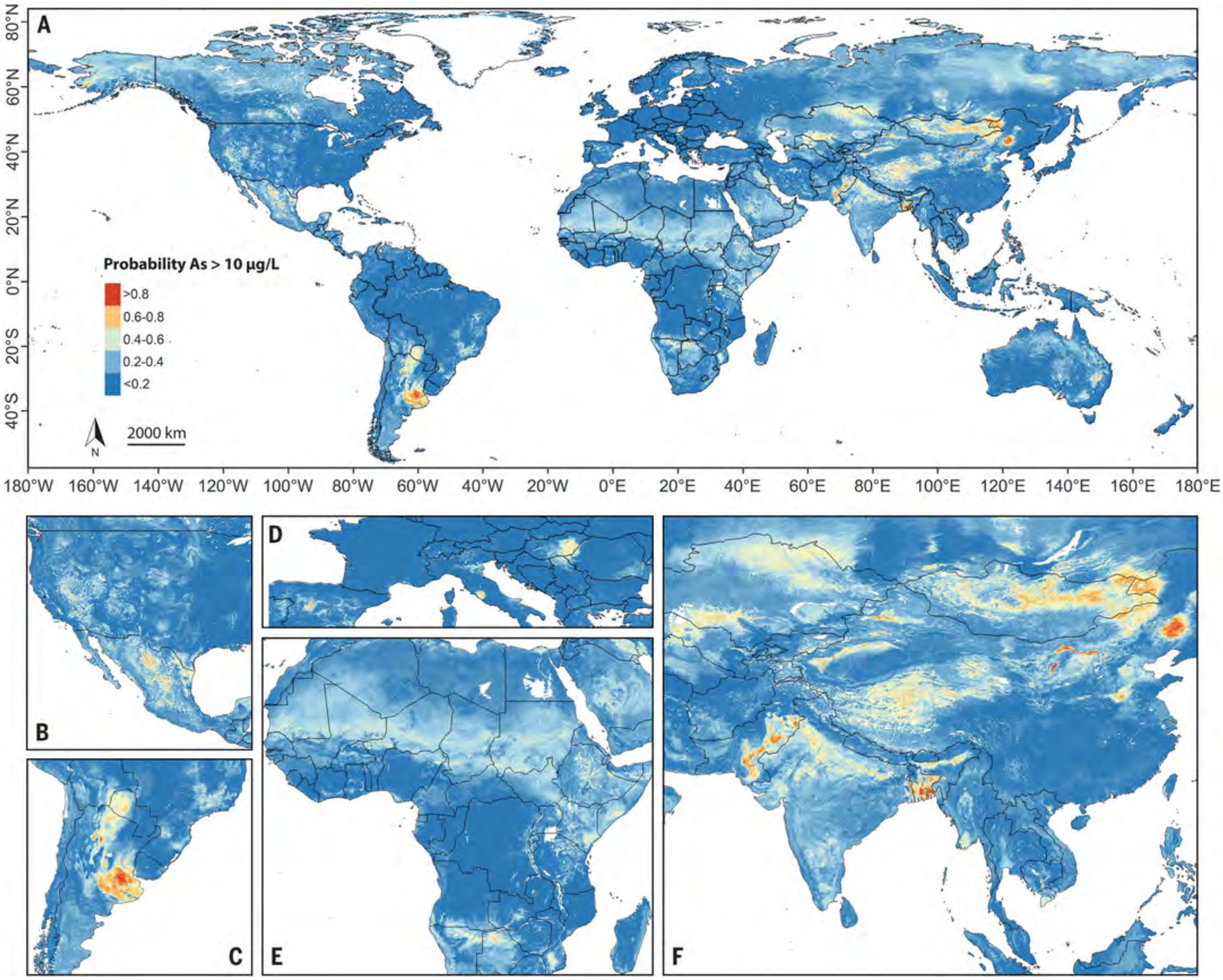


Fig. 2. Global prediction of groundwater arsenic. (**A** to **F**) Modeled probability of arsenic concentration in groundwater exceeding 10 µg/liter for the entire globe (A) along with zoomed-in sections of the main more densely populated affected areas (B) to (F). The model is based on the arsenic data points in Fig. 1 and the predictor variables in table S2. Figs. S2 to S8 provide more detailed views of the prediction map.

under the curve (AUC, 0.89) and Cohen's kappa statistic (0.55) remained unchanged.

The final random forest model was created based on the compiled global dataset of high and low arsenic concentrations along with the 11 predictor variables. The standard number of variables to be made available at each branch of each tree is between three and four (see methods). Because our tests showed the value of three performing better than four and higher values (though error and performance rates varied only within ~1%), we set this parameter to three. The global map produced from this model is displayed in Fig. 2A along with more detailed views of the more populated affected continental regions shown in Fig. 2, B to F. It indicates the probability of the concentration of arsenic in groundwater in a given 1-km^2 cell exceeding 10 μg/liter. The uncertainty of the model is inherent in the probabilities themselves, because they are simply the average of the votes or predictions of high or low values of each of the 10,001 trees grown. That is, each tree casts a vote of 0 or 1 ("no" or "yes" to As > 10 μg/liter) for each cell based on the values of the predictor variables in that cell. Figures S2 to S8 also provide more detailed views of the prediction map for each of the inhabited continents.

The importance of each of the 11 predictor variables in terms of mean decrease in accuracy and mean decrease in the Gini index is listed in fig. S1. Relative to the initial set of 52 variables, the values of these two statistics for most of the 11 final predictor variables appear to fall within a fairly narrow range, indicating comparable importance. Exceptions include fluvisols and soil pH, which have somewhat greater importance, and temperature, which, according to both statistics, is the least important of the 11 variables. Soil pH was also found to be an important predictor variable in arid, oxidizing environments in Pakistan (*29*). Although widespread arsenic dissolution occurs in Holocene fluvial sediments (*5*–*7*, *9*, *37*), this geological epoch has not been consistently mapped around the world. However, the global dataset of fluvisols provides a very suitable alternative (*29*), which may even be more appropriate because fluvisols by definition encompass recent fluvial sediments and not, for example, aeolian Holocene sediments that are generally not relevant for arsenic release. The generally high model importance of climate variables, as evidenced by them all being selected for the final model, highlights the strong control that climate has on arsenic release in aquifers. In particular, precipitation and evapotranspiration have a direct role in creating conditions conducive for arsenic release under reducing conditions (e.g., waterlogged soils) as well as high aridity associated with oxidizing, high-pH conditions.

Table 1. Confusion matrix and other statistics summarizing the results of applying the random forest model to the test dataset at a probability cutoff of 0.50.

Model output	Value
Predicted As ≤ 10 μg/liter	
Measured As ≤ 10 μg/liter	7710
Measured As > 10 μg/liter	555
Predicted As > 10 μg/liter	
Measured As ≤ 10 μg/liter	1394
Measured As > 10 μg/liter	2037
Sensitivity	0.79
Specificity	0.85
PPV	0.59
NPV	0.93
Prevalence	0.22
Balanced accuracy	0.82
Cohen's kappa	0.55
AUC	0.89

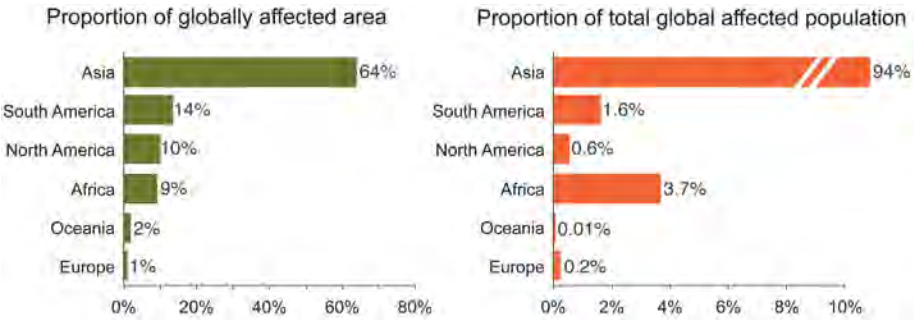


Fig. 3. Proportions of land area and population potentially affected by arsenic concentrations in groundwater exceeding 10 μg/liter by continent.

The performance of the random forest model on the test dataset (20% of the data, which was randomly selected while maintaining the relative distribution of high and low values) is summarized in the confusion matrix in Table 1. Despite a prevalence of high values (>10 μg/liter) of only 22% in the dataset, the model performs well in predicting both high values (sensitivity: 0.79) and low values (specificity: 0.85) at a probability cutoff of 0.50. The average of these two figures, known as balanced accuracy, is correspondingly high at 0.82. Likewise, the model's AUC, which considers the full range of possible cutoffs, has a very high value of 0.89 with the test dataset (Table 1). For comparison, the AUC of a random forest using all 52 original predictor variables is also 0.89.

The model was also tested on a dataset of more than 49,000 arsenic data points originating from known depths greater than 100 m (average 562 m, standard deviation 623 m). Although the model was not trained on any measurements from these depths and the fact that only surface parameters were used as predictor variables, the model nevertheless performed quite well in predicting the arsenic concentrations of these deep groundwater sources, as evidenced by an AUC of 0.77.

Regions and populations at risk

Areas predicted to have high arsenic concentrations in groundwater exist on all continents, with most being located in Central, South, and Southeast Asia; parts of Africa; and North and South America (Fig. 2 and figs. S2 to S8). Known areas of groundwater arsenic contamination are generally well captured by the global arsenic prediction map, for example, parts of the western United States, central Mexico, Argentina, the Pannonian Basin, Inner Mongolia, the Indus Valley, the Ganges-Brahmaputra delta, and the Mekong River and Red River deltas. Areas of increased arsenic hazard where little concentration data exist include parts of Central Asia, particularly Kazakhstan, Mongolia, and Uzbekistan; the Sahel region; and broad areas of the Arctic and sub-Arctic. Of these, the Central Asian hazard areas are better constrained, as evidenced by higher probabilities.

Probability threshold values of 0.57 from the sensitivity-specificity comparison and 0.72 from the positive predictive value (PPV)–negative predictive value (NPV) comparison were found using the full dataset (combined training and test datasets) of arsenic concentrations. The proportions of high modeled arsenic hazard by continent associated with each of these probabilities are shown in Fig. 3. Global maps of the potentially affected population in the risk areas, as determined by these two thresholds, are shown in Fig. 4. As described in the methods, these maps were then used to estimate the population potentially affected by drinking groundwater with arsenic concentrations exceeding 10 μg/liter.

The resulting global arsenic risk assessment indicates that about 94 million to 220 million people around the world (of which 85 to 90% are in South Asia) are potentially exposed to high concentrations of arsenic in groundwater from their domestic water supply (tables S4 and S5). This range is consistent with the previous most comprehensive literature compilations, that is, 140 million people (*41*) and 225 million people (*42*). Household groundwater-use statistics were not available for ~6 to 8% of the affected countries (depending on the cutoff), for which the less detailed statistics derived from the AQUASTAT database of the Food and Agriculture Organization of the United Nations were used instead (see methods for details). To determine the amount of error that using these more general groundwater-use statistics might introduce to the overall population figures, the global potentially affected populations were recalculated with these countries' (those lacking household groundwater-use statistics) groundwater-use rates set to the extreme values of 0 and 100%. Because this applied to relatively few countries and arsenic-affected areas, doing so affected the overall global population figures by an inconsequential amount (±0.1%), indicating that using the AQUASTAT groundwater-use rates, where necessary, is an acceptable approximation.

This estimate of risk takes into account only the proportion of households utilizing unprocessed groundwater and assumes uniform rates throughout the urban and nonurban areas of each country. The uncertainties of these rates are unknown. The population in each cell was reduced by the uncertainty of the cell's prediction, which is justified based on the heterogeneity inherent in the accumulation of arsenic in an aquifer, which is generally at a much finer scale than that of the 1-km^2 resolution of the arsenic hazard map. Because the arsenic prediction for a cell represents the average outcome for that cell, we can take the modeled probability as a first-order approximation of the proportion of an aquifer in that cell containing high arsenic concentrations. Only cells exceeding the probability threshold (i.e., 0.57 or 0.72) were considered. The global estimate of 94 million to 220 million people potentially affected by consuming arsenic-contaminated groundwater is

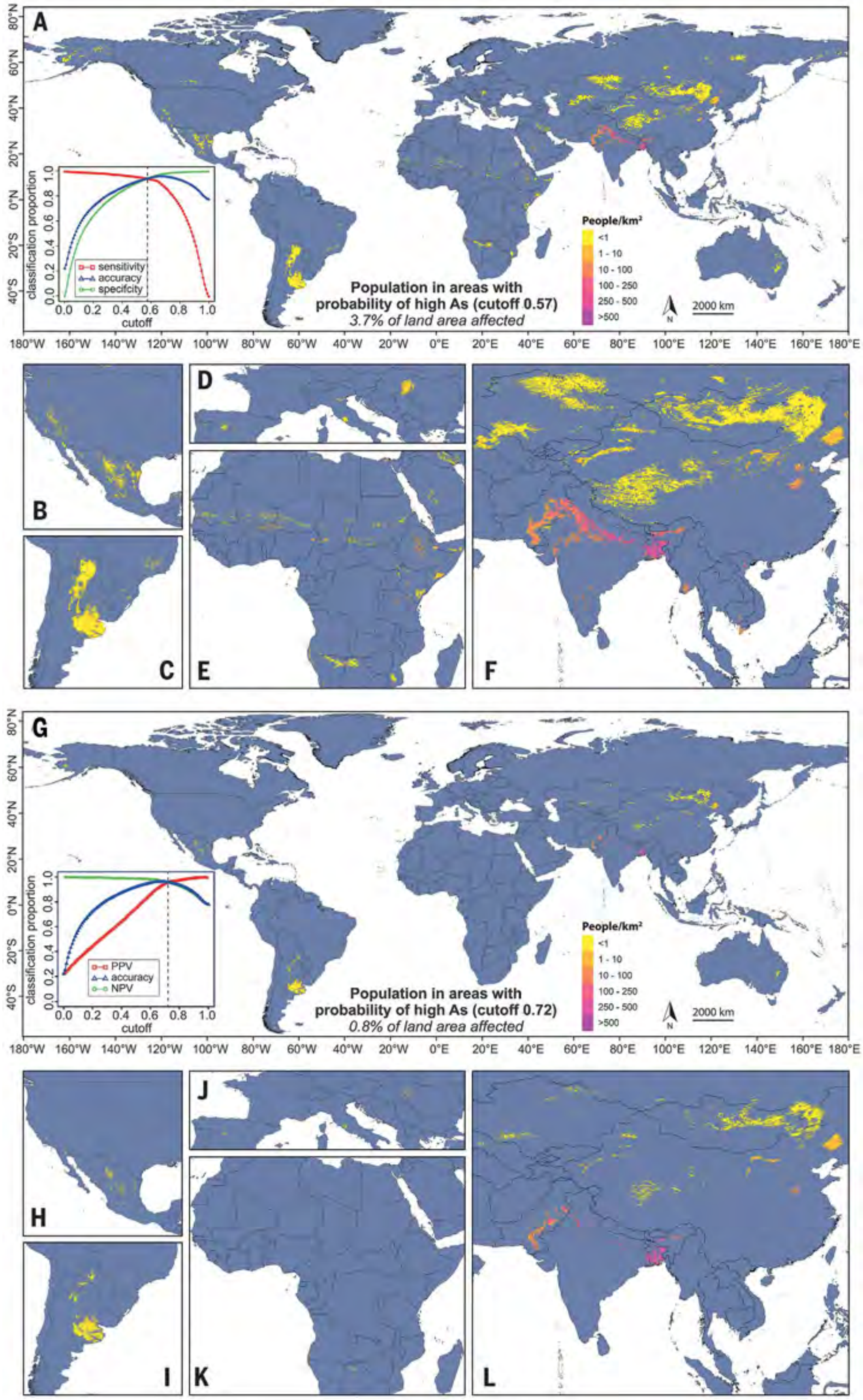

A
80°N
60°N
40°N
20°N
0°N
20°S
40°S
classification proportion
0.0 0.2 0.4 0.6 0.8 1.0
sensitivity
accuracy
specifcity
cutoff
Population in areas with
probability of high As (cutoff 0.57)
3.7% of land area affected
People/km²
<1
1 - 10
10 - 100
100 - 250
250 - 500
>500
N
2000 km
180°W 160°W 140°W 120°W 100°W 80°W 60°W 40°W 20°W 0°E 20°E 40°E 60°E 80°E 100°E 120°E 140°E 160°E 180°E
B
C
D
E
F
G
PPV
accuracy
NPV
Population in areas with
probability of high As (cutoff 0.72)
0.8% of land area affected
H
I
J
K
L

Fig. 4. Estimated population at risk. (**A** to **L**) Population in risk areas potentially containing aquifers with arsenic concentrations >10 μg/liter using probability cutoffs of 0.57 (A), at which sensitivity and specificity are equal [inset in (A)] as applied to the full (training and test) dataset, and 0.72 (G), at which PPV and NPV are equal [inset in (G)] using the full dataset. The detailed areas of Fig. 2 are also repeated here for both models (B) to (F) and (H) to (L).

broken down by continent and country in tables S4 and S5, respectively, and represents the most accurate and consistent global estimate available.

Discussion

The accuracy of the global groundwater arsenic prediction model presented here, as indicated, for example, with an AUC of 0.89 calculated with the test dataset, exceeds that found in previous arsenic prediction studies (table S3). The dominance of climate and soil parameters in the final model is indicative of their direct influence or at least strong association with the processes of arsenic accumulation in groundwater.

With respect to previous arsenic prediction maps of global sedimentary basins (*40*, *43*), the new model represents a substantial advancement on a few different levels. First, the new model presented here provides predictions for all areas of the inhabited continents, whereas the previous first-generation statistical model covered only about half of the land areas. In addition, a 10-fold increase in measurement points has allowed arsenic concentrations to be incorporated from many more areas of the globe. The greatly expanded availability and quality of global predictor datasets over the past 10 years has enabled new variables to be considered, such as soil type (e.g., fluvisols), as well as provided a 10- to 60-fold greater spatial resolution (i.e., 30 arc-sec versus 5 to 30 arc-min). However, the presence of high arsenic in groundwater at a given location is of course predicated on the existence of an aquifer in the first place, which may not be so in the case of unfractured solid rock, steep terrain, or very dry conditions. Models are only as good as the data on which they are based. As accurate as the new arsenic model is, it could be further improved as more arsenic data and more detailed predictor datasets come into existence.

Particularly in sedimentary aquifers, arsenic concentration is often highly dependent on depth, that is, on specific sedimentary sequences that differ in the concentration of arsenic in sediments as well as the geochemical conditions conducive to arsenic release. To better characterize this relationship in a given sedimentary basin, detailed depth information of groundwater samples would need to be incorporated in a separate basin-level study. Unfortunately, it is not feasible in a global-scale study to account for all of the diversity of the sedimentary basins of the world, especially because depth information of groundwater samples is often not available. As such, we have relied on a statistical analysis of model performance against depth ranges of samples (where present) to determine model sensitivity to depth.

Our approach in the risk assessment of potentially affected populations is relatively discerning and/or conservative. As such, the resulting population estimates may in some cases be lower than those found in earlier studies. One reason for this is that we used country-specific statistics of rural and urban domestic groundwater usage, which allowed us to subtract the proportion of the population that uses surface water, tap water, or other sources. This was not the case, for example, in a previous study of China that estimated that 19.6 million people were affected in the country (*21*), whereas our estimate is considerably lower at 4.3 million to 12.1 million. Furthermore, we consider only areas in which the probability of high arsenic exceeds the statistically determined cutoffs, that is, 0.57 and 0.72. Taking the United States as an example, applying this criterion left only 0.2 to 2% of the area of the country over which to sum the potentially affected population (≤0.21 million, this study). In a previous arsenic risk assessment of the United States (*31*), the entire country was used to estimate affected population (2.1 million), that is, not only the high-risk areas.

The actual proportion of groundwater usage varies spatially throughout a country, and so more detailed usage statistics beyond only urban versus rural would improve the accuracy of a risk assessment. In addition, more groundwater samples (ideally including depth information) from areas that currently have poor coverage would benefit future modeling efforts by allowing the model to be better adapted to those areas.

The presented arsenic probability maps should be used as a guide to further groundwater arsenic testing, for example, in Central Asia, the Sahel, and other regions of Africa. Only actual groundwater quality testing can definitively determine the suitability of groundwater with respect to arsenic, particularly because of small-scale (<1 km) aquifer heterogeneities that cannot be modeled with existing global datasets (*9*, *44*). The hazard maps highlight areas at risk and provide a basis for targeted surveys, which continue to be important. The already large number of people potentially affected can be expected to increase as groundwater use expands with a growing population and increasing irrigation, especially in the light of water scarcity associated with warmer and drier conditions related to climate change. The maps can also help aid mitigation measures, such as awareness raising, coordination of government and financial support, health intervention programs, securing alternative drinking water resources, and arsenic removal options tailored to the local groundwater conditions as well as social setting.

REFERENCES AND NOTES

1. A. H. Smith, E. O. Lingas, M. Rahman, *Bull. World Health Organ.* **78**, 1093–1103 (2000).
2. M. F. Hughes, *Toxicol. Lett.* **133**, 1–16 (2002).
3. D. Mondal *et al.*, *Environ. Geochem. Health* **32**, 463–477 (2010).
4. H. G. Gorchev, G. Ozolins, *WHO Chron.* **38**, 104–108 (1984).
5. P. Smedley, D. Kinniburgh, *Appl. Geochem.* **17**, 517–568 (2002).
6. R. Nickson *et al.*, *Nature* **395**, 338–338 (1998).
7. J. McArthur, P. Ravenscroft, S. Safiulla, M. Thirlwall, *Water Resour. Res.* **37**, 109–117 (2001).
8. M. Berg *et al.*, *Chem. Geol.* **249**, 91–112 (2008).
9. S. Fendorf, H. A. Michael, A. van Geen, *Science* **328**, 1123–1127 (2010).
10. M. I. Litter *et al.*, *Sci. Total Environ.* **676**, 756–766 (2019).
11. Y. Zheng, S. V. Flanagan, *Environ. Health Perspect.* **125**, 085002 (2017).
12. P. Bhattacharya, D. Chatterjee, G. Jacks, *Int. J. Water Resour. Dev.* **13**, 79–92 (1997).
13. A. van Geen *et al.*, *Water Resour. Res.* **39**, 1140 (2003).
14. H. B. Nicolli, J. M. Suriano, M. A. Gomez Peral, L. H. Ferpozzi, O. A. Baleani, *Environ. Geol. Water Sci.* **14**, 3–16 (1989).
15. P. Smedley, H. Nicolli, D. Macdonald, A. Barros, J. Tullio, *Appl. Geochem.* **17**, 259–284 (2002).
16. M. Blarasin, A. Cabrera, E. Matteoda, paper presented at the XXXIII IAH – 7° ALHSUD Congress, Zacatecas, Mexico, 11 to 15 October 2004.
17. M. Auge, G. E. Viale, L. Sierra, in *VIII Congreso Argentino de Hidrogeología: Aguas subterráneas recurso estratégico* (Editorial de la Universidad Nacional de La Plata, 2013), vol. 2, pp. 58–63.
18. M. Berg *et al.*, *Environ. Sci. Technol.* **35**, 2621–2626 (2001).
19. J. Buschmann, M. Berg, C. Stengel, M. L. Sampson, *Environ. Sci. Technol.* **41**, 2146–2152 (2007).
20. P. Smedley, M. Zhang, G. Zhang, Z. Luo, *Appl. Geochem.* **18**, 1453–1477 (2003).
21. L. Rodríguez-Lado *et al.*, *Science* **341**, 866–868 (2013).
22. Y. Zhou *et al.*, *Appl. Geochem.* **77**, 116–125 (2017).
23. D. Chatterjee, R. Roy, B. Basu, *Environ. Geol.* **49**, 188–206 (2005).
24. B. Nath, D. Stüben, S. B. Mallik, D. Chatterjee, L. Charlet, *Appl. Geochem.* **23**, 977–995 (2008).
25. B. A. Shah, *Curr. Sci.* **98**, 1359–1365 (2010).
26. B. Planer-Friedrich, "Hydrogeological and hydrochemical investigations in the Rioverde Basin, Mexico," thesis, Institute of Geology, University of Mining and Technology Freiberg (2000).
27. M. T. Alarcón-Herrera *et al.*, *J. Hazard. Mater.* **262**, 960–969 (2013).
28. R. Nickson, J. McArthur, B. Shrestha, T. Kyaw-Myint, D. Lowry, *Appl. Geochem.* **20**, 55–68 (2005).
29. J. E. Podgorski *et al.*, *Sci. Adv.* **3**, e1700935 (2017).
30. J. D. Ayotte, M. G. Nielsen, G. R. Robinson Jr., R. B. Moore, *Water Resour. Invest. Rep.* **99**, 4162 (1999).
31. J. D. Ayotte, L. Medalie, S. L. Qi, L. C. Backer, B. T. Nolan, *Environ. Sci. Technol.* **51**, 12443–12454 (2017).
32. M. Berg *et al.*, *Sci. Total Environ.* **372**, 413–425 (2007).
33. L. H. Winkel *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **108**, 1246–1251 (2011).
34. A. Bretzler *et al.*, *Sci. Total Environ.* **584–585**, 958–970 (2017).
35. Q. Zhang *et al.*, *J. Hazard. Mater.* **262**, 1147–1153 (2013).
36. S. Bindal, C. K. Singh, *Water Res.* **159**, 65–76 (2019).
37. L. Winkel, M. Berg, M. Amini, S. J. Hug, C. A. Johnson, *Nat. Geosci.* **1**, 536–542 (2008).
38. Q. Yang, H. B. Jung, R. G. Marvinney, C. W. Culbertson, Y. Zheng, *Environ. Sci. Technol.* **46**, 2080–2087 (2012).
39. N. Yang, L. H. Winkel, K. H. Johannesson, *Environ. Sci. Technol.* **48**, 5660–5666 (2014).
40. M. Amini *et al.*, *Environ. Sci. Technol.* **42**, 3669–3675 (2008).
41. P. Ravenscroft, H. Brammer, K. Richards, *Arsenic Pollution: A Global Synthesis* (Wiley, 2009), vol. 28.
42. S. Murcott, *Arsenic Contamination in the World* (IWA Publishing, 2012).
43. P. Ravenscroft, "Predicting the global extent of arsenic pollution of groundwater and its potential impact on human health," unpublished report prepared for UNICEF, December 2007.
44. Y. Zheng, *Curr. Environ. Health Rep.* **4**, 373–382 (2017).
45. J. Podgorski, M. Berg, Podgorski_and_Berg_2020. ERIC/open (2020); http://dx.doi.org/10.25678/0001ZT.

ACKNOWLEDGMENTS

We thank our colleagues A. Bretzler and C. Zurbrügg (Eawag), A. Steiner and S. Piers de Raveschoot (SDC), and D. A. Polya and R. Wu (University of Manchester) for their support, as well as the many providers of data, which were an essential component of this work. **Funding:** We thank the Swiss Agency for Development and Cooperation (project nos. 7F-09010.01.01 and 7F-09963.01.01) for long-term support and cofunding of this study, as well as a University of Manchester EPSRC IAA Impact Support Fund Award to D. A. Polya. **Author contributions:** J.P.: Methodology, Modeling, Writing–Original draft preparation; M.B.: Supervision, Writing–Reviewing and Editing. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** The modeling data, code, and raster output maps are available at ERIC/open (*45*). Arsenic concentration data points and hazard and risk maps are also available for viewing on the GIS-based Groundwater Assessment Platform (GAP), www.gapmaps.org.

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/368/6493/845/suppl/DC1
Methods
Figs. S1 to S11
Tables S1 to S6
References (*46*–*127*)

10 November 2019; accepted 3 April 2020
10.1126/science.aba1510

Science

Global threat of arsenic in groundwater

Joel Podgorski and Michael Berg

Science **368** (6493), 845-850.
DOI: 10.1126/science.aba1510

Dowsing for danger

Arsenic is a metabolic poison that is present in minute quantities in most rock materials and, under certain natural conditions, can accumulate in aquifers and cause adverse health effects. Podgorski and Berg used measurements of arsenic in groundwater from ~80 previous studies to train a machine-learning model with globally continuous predictor variables, including climate, soil, and topography (see the Perspective by Zheng). The output global map reveals the potential for hazard from arsenic contamination in groundwater, even in many places where there are sparse or no reported measurements. The highest-risk regions include areas of southern and central Asia and South America. Understanding arsenic hazard is especially essential in areas facing current or future water insecurity.

Science, this issue p. 845; see also p. 818

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. The title *Science* is a registered trademark of AAAS.

science.sciencemag.org/content/368/6493/845/suppl/DC1

Supplementary Materials for

Global threat of arsenic in groundwater

Joel Podgorski* and Michael Berg*

*Corresponding author. Email: joel.podgorski@eawag.ch (J.P.); michael.berg@eawag.ch (M.B.)

Published 22 May 2020, *Science* **368**, 845 (2020)
DOI: 10.1126/science.aba1510

This PDF file includes:

Methods
Figs. S1 to S11
Tables S1 to S6
References

This PDF file includes:

METHODS

Random forest modeling(*46*) was used with the R programming language(*47*) to determine statistical relationships between global independent or predictor variables of various environmental parameters and the dependent or target variable of arsenic concentration in groundwater. Random forests generate an ensemble of decision trees, which are models in which the target variable is split in consecutive nodes using a predictor variable and an associated cutoff that result in the greatest variance in the target variable at each node. Each tree in a random forest is different due to a randomly selected subset of predictor variables being made available at each node (typically the square root of the total number of variables)(*48*) as well as a random selection with replacement of data rows (bootstrap aggregating or bagging), consequently resulting in about one-third of the data not being used in a given tree(*46*). Adding these elements of randomness to the decision trees and averaging their results produces a model that is more stable against small changes in the data(*49*). The following sections describe the various steps that were taken in the modeling process.

Preparation of arsenic data

Measurements of arsenic concentration in groundwater were compiled from many different sources (Table S1). These data sources represent a combination of measurements from field test kits and sophisticated lab analyses, though the specific method that was used is often not clearly stated. Since the resolution of the predictor data sets used is 1 km, the geometric mean of concentrations falling within 1-km^2 pixels were used for the data points for modeling. These values were then converted into binary form by setting all arsenic concentrations meeting the WHO guideline of ≤10 µg/L to zero and all concentrations >10 µg/L to one. This was done in order to i) focus on the basic health question of groundwater being safe or unsafe for drinking and ii) mitigate differences in precision among the different analysis methods used by the various data sources. The values of the predictor variables were then found at the geographical coordinates of each data point. The dataset was then randomly divided into training (80%) and test (20%) datasets, each preserving the proportion of high and low values of the full dataset.

Evaluation of well depth

Since by necessity the predictor variables used in our geospatial modeling are all based on land surface data, we reason that arsenic concentrations originating from greater depths are likely less well explained by parameters at the surface. We therefore tuned the selection of concentration data by testing model performance with different data subsets based on the following reported well depths or lack thereof: 0-25 m, 0-50 m, 0-75 m, 0-100 m, 0-125 m, 0-150 m, 0-3700 m (all concentrations with well-depth information) and all data (with and without well depth). For each of these subsets, we ran a random forest model with the full set of predictor variables (see below). Since the proportion of high measurements (prevalence) varies among the different concentration subsets, we analyzed model performance using the Cohen's kappa coefficient(*50*), which indicates the accuracy of a classification model beyond what would be expected merely by chance. Although model performance peaks using concentration data ranging from 0-100 m depth, we found that also including data points with no depth information results in only a relatively minor decrease in Cohen's kappa while at the same time allowing us to use more than twice as much data (table S6) and thereby cover much more of the world. Over 200,000 arsenic concentration measurements with a nominal depth of up to 100 m as well as those without a specified depth (35% without depth information) were thereby aggregated by the process described above into more than 55,000 data points.

Selection of predictor variables.

A collection of 52 spatially continuous predictor variables with global coverage representing various climatic, geologic, soil and other parameters known or hypothesized to be related to the dissolution and accumulation of arsenic in groundwater was assembled (table S2). Table S3 lists the predictor variables used in other statistical modeling studies of groundwater arsenic contamination, which were used as an initial point of orientation in selecting variables. In order to remove poorly performing predictors and in the interest of creating the simplest best model, subsets of the initial 52 variables were iteratively produced through recursive feature elimination (RFE), whereby 20% of the least important predictors were removed in a series of random forest models. Importance here is defined as the decrease in the accuracy of a random forest model when the values of a variable are randomly reassigned over all cases. Variables were removed iteratively until only two remained. 5000 trees were grown for the first random forest iteration and 2000 trees for all subsequent iterations. The collection of 11 variables selected for use in the final model was that from the model with the least number of variables whose error rate was within one standard error of the random forest with the smallest error rate.

Random forest modeling and validation.

The random forest grew 10,001 trees using the training dataset and the variables found using the automated selection procedure described above. The binary outcomes of these trees were averaged such that the random forest model provides the probability of encountering groundwater arsenic greater than the 10 µg/L threshold for a given set of values of the predictor variables.

The model was then applied to the test dataset with its performance being evaluated by various statistics, including the Area Under the ROC (receiver operating characteristic (ROC)) Curve (AUC)(*51*). The AUC provides a single statistic characterizing the accuracy of predicting high values (sensitivity) and low values (specificity) and is found by applying many different cutoffs between 0 and 1 to the modeled probabilities. The cutoff is the value used to determine whether the modeled probabilities should be considered high or low. The AUC is the area beneath the curve drawn through the points on the plot of specificity versus 1-specificity. The value of the AUC generally ranges from 0.5 (equivalent to an uneducated guess) to 1 (perfect predictive accuracy).

The importance of the variables was also used to help assess the relative influence of the different predictor variables, as measured for each tree by the mean decrease in accuracy and mean decrease in Gini node impurity, and averaged over all trees. The accuracy test was performed on out-of-bag samples (those not randomly selected to grow a tree) by randomly resorting the values of a variable over all cases, such that a variable's importance is inversely proportional to its decrease in accuracy when the incorrect values of the variable are used. Node impurity as measured by the Gini index refers to how well the two classes (high or low arsenic) are split into two branches at a given node.

Impurity relates to the amount of mixing of the two classes within each branch, such that the lower the impurity associated with using a certain variable, the more effective that variable is in differentiating between the two classes.

Calculation of population affected.

The random forest model was used to estimate the number of people potentially exposed to high levels of geogenic arsenic in drinking water. The first step was to find probability cutoffs to use in classifying the arsenic groundwater hazard as being either high or low. For this purpose, we considered the probability at which sensitivity and specificity are equal as well as that at which the positive predictive value (PPV) and negative predictive value (NPV) are equal. These comparisons were carried out using probability intervals of 0.01 with the full dataset. That is, although the training dataset was used to generate the model and the test dataset used to verify the model, both were combined so as to utilize all available data to determine how to interpret the model.

Sensitivity and specificity are the rates at which high and low measured arsenic concentrations, respectively, are successfully identified by the model. On the other hand, PPV and NPV are the rates at which the high and low model predictions, respectively, are correct. The seemingly subtle differences between these two sets of statistics can be clarified by considering their definitions in terms of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN):

$$\text{Sensitivity} = \frac{TP}{TP+FN} \qquad (1)$$

$$\text{Specificity} = \frac{TN}{TN+FP} \qquad (2)$$

$$\text{Positive predictive value} = \frac{TP}{TP+FP} \qquad (3)$$

$$\text{Negative predictive value} = \frac{TN}{TN+FN} \qquad (4)$$

Comparisons of sensitivity-specificity and PPV-NPV generally yield different results, either of which could provide a sensible basis for deriving risk from the arsenic hazard map. Given the similar nature of these two sets of statistics, either considering the accuracy of the determination of high and low values or the accuracy of the predictions themselves, both were taken together to establish a range of reasonable probability cutoff values for deriving risk.

Once identified, the high hazard areas were then used to determine the populations at risk in these areas. Global population was taken from a 1-km resolution model of projected population in 2020 based on a “middle-of-the-road” socio-economic scenario with respect to current trends in environmental sustainability and distribution of wealth(*52*). Urban and non-urban populations in each country were then multiplied by the urban/non-urban rates of household groundwater use, as indicated in the most recent studies (up to past two decades) reporting groundwater-use statistics contained within country-level reports of the UNICEF/WHO Joint Monitoring Program (JMP)(*53*). These reports provide rates of consumption of unprocessed groundwater, as opposed to tap water (that may come from groundwater and have undergone some degree of treatment), rainwater, packaged water, surface water or other improved and non-improved sources, but they do not account for any arsenic filtration that may occur at a community water point or in the household. Urban areas were found by means of a global land use map(*54*). Where these household-level groundwater-use statistics were not available for a country, a single groundwater utilization rate was applied to both urban and non-urban areas that was calculated from countries' groundwater withdrawal rates provided in FAO's AQUASTAT database(*55*), where available. Finally, the value of each cell or pixel was reduced by multiplying the pixel's groundwater-consuming population by its probability of having high arsenic concentrations. The calculation of potentially affected population is summarized in the following equations:

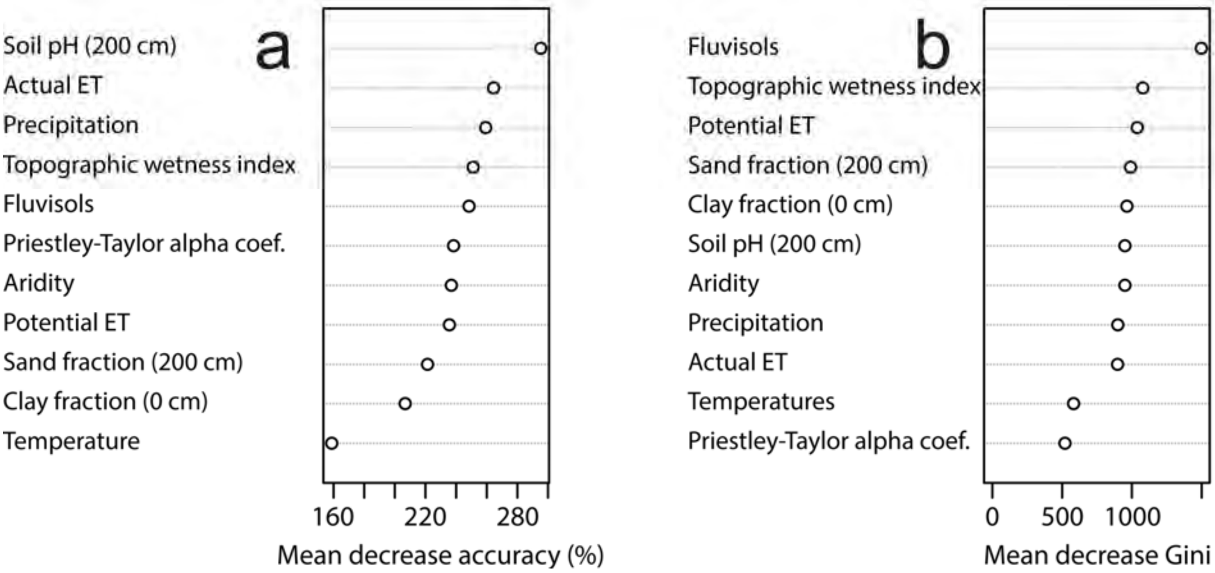
$$Pop_{\text{affect}} = Pop_{\text{affect,rural}} + Pop_{\text{affect,urban}} \tag{5}$$

$$Pop_{\text{affect,rural}}(Prob_{\text{As>10}}) = \begin{cases} Pop_{\text{rural}} \times GW_{\text{rural}} \times Prob_{\text{As>10}}, & Prob_{\text{As>10}} > Cut \\ 0, & Prob_{\text{As>10}} \leq Cut \end{cases} \tag{6}$$

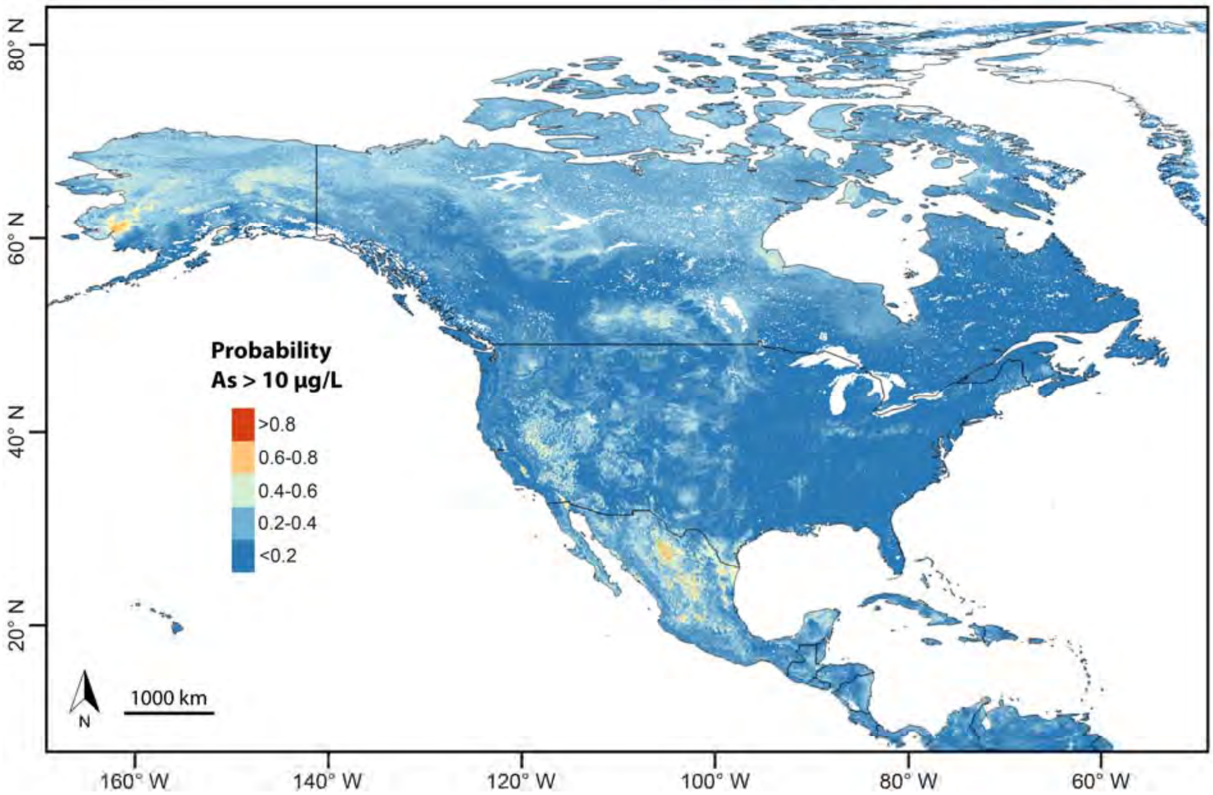
$$Pop_{\text{affect,urban}}(Prob_{\text{As>10}}) = \begin{cases} Pop_{\text{urban}} \times GW_{\text{urban}} \times Prob_{\text{As>10}}, & Prob_{\text{As>10}} > Cut \\ 0, & Prob_{\text{As>10}} \leq Cut \end{cases} \tag{7}$$

where:

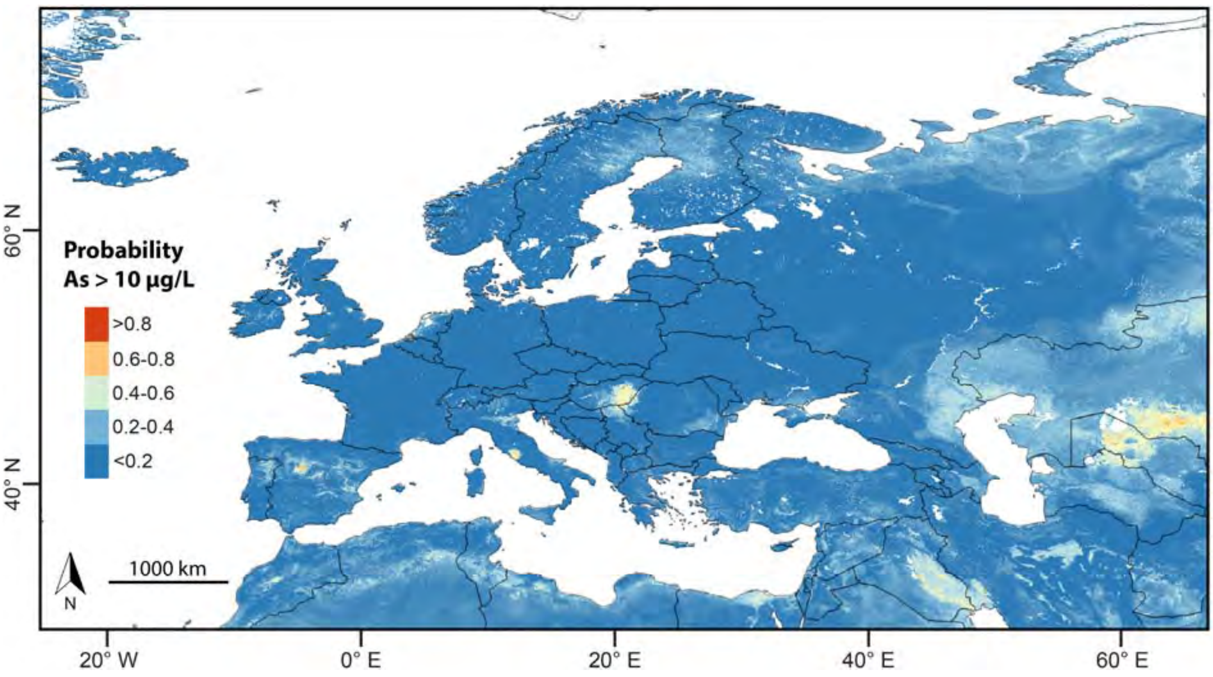
Pop_{affect}	is the potentially affected population
$Pop_{\text{rural(urban)}}$	is the rural (urban) population
GW	is the (rural or urban) proportion of household groundwater usage
$Prob_{\text{As>10}}$	is the probability of the concentration of arsenic exceeding 10 µg/L
Cut	is the probability cutoff used to distinguish between high and low risk areas



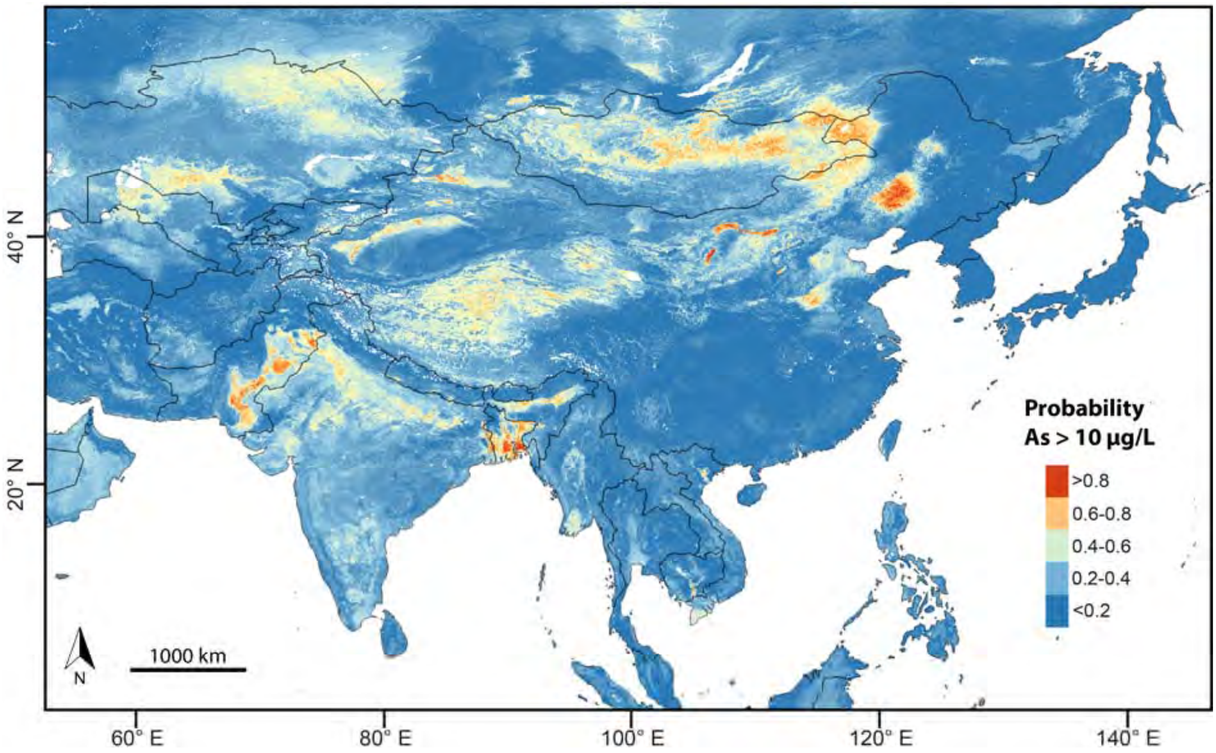
Supplementary Figure 1 | Importance of the predictor variables in the final random forest model. Both the (a) mean decrease in accuracy and (b) mean decrease in the Gini index are greater when the variable in question more strongly improves the model's accuracy or node impurity, respectively.



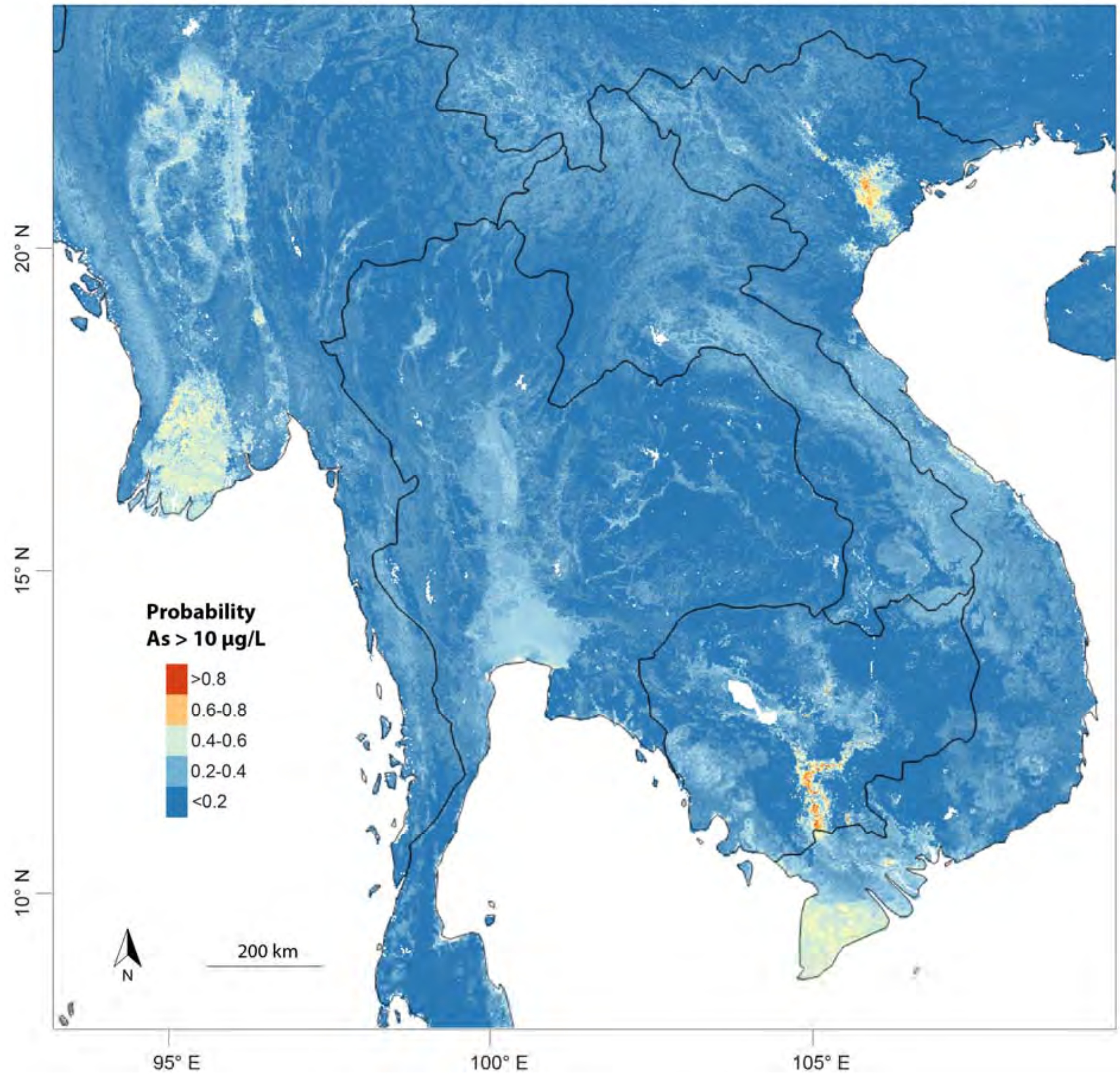
Supplementary Figure 2 | Arsenic in groundwater prediction map for North and Central America and the Caribbean.



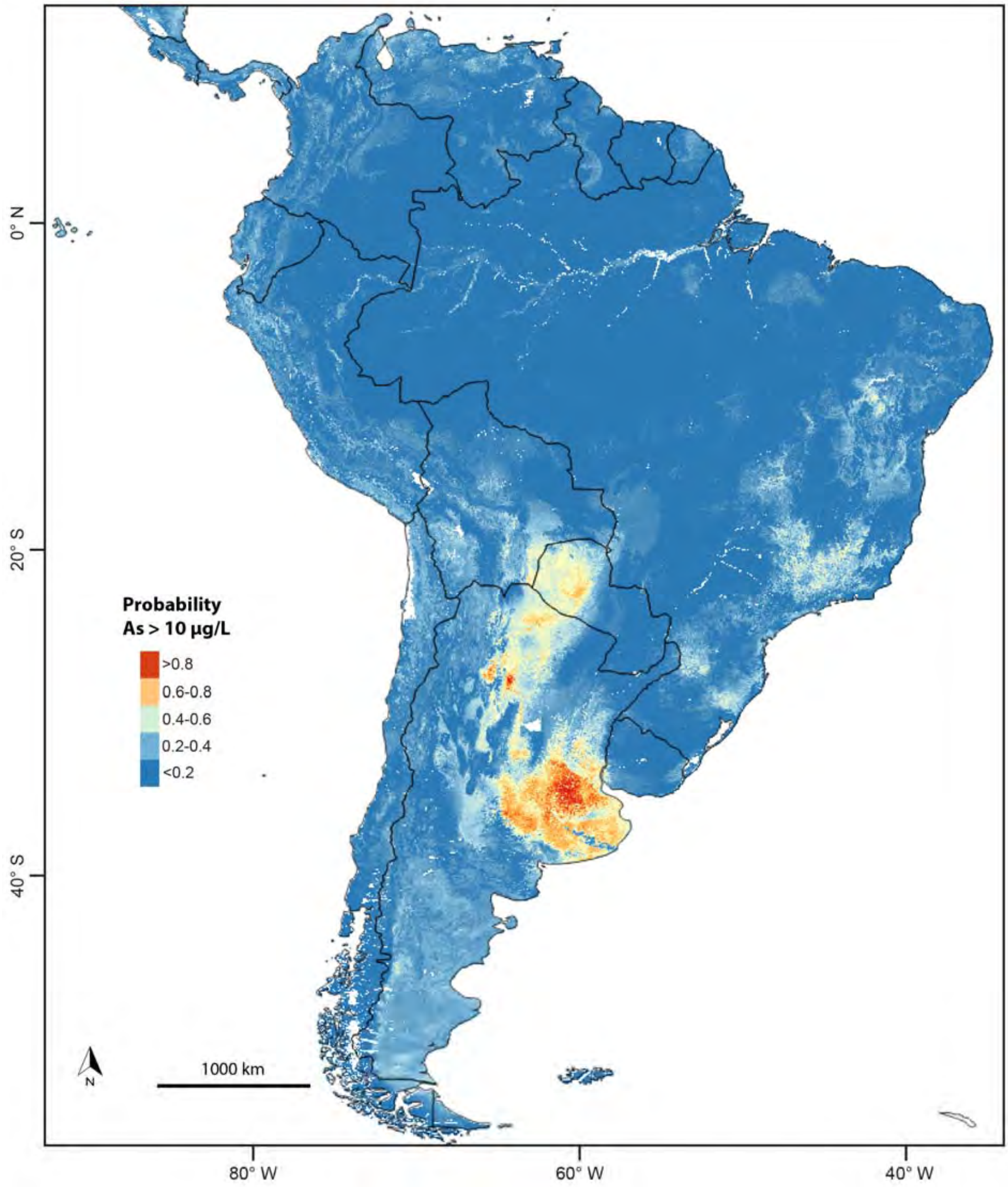
Supplementary Figure 3 | Arsenic in groundwater prediction map for Europe.



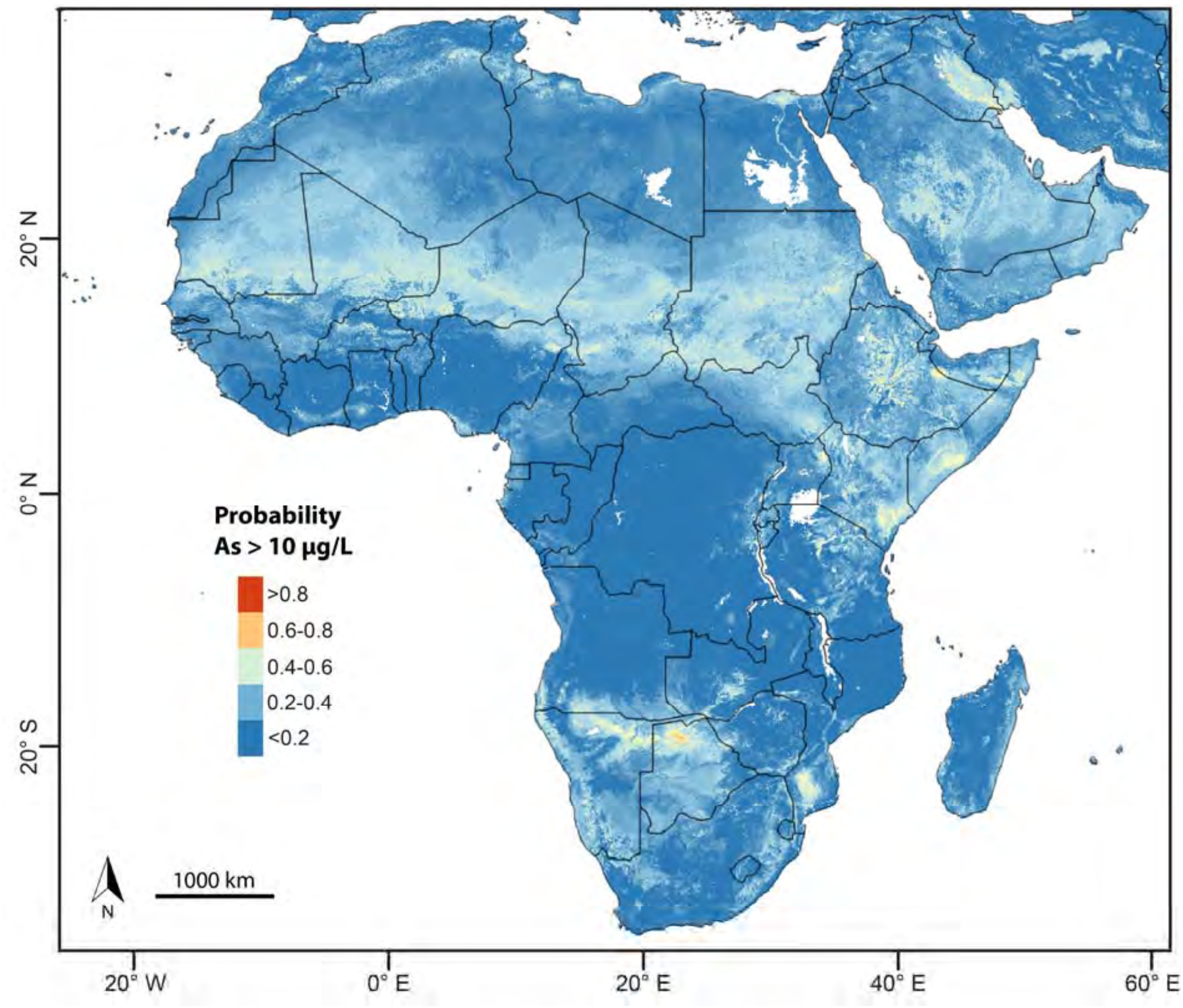
Supplementary Figure 4 | Arsenic in groundwater prediction map for central, South and East Asia.



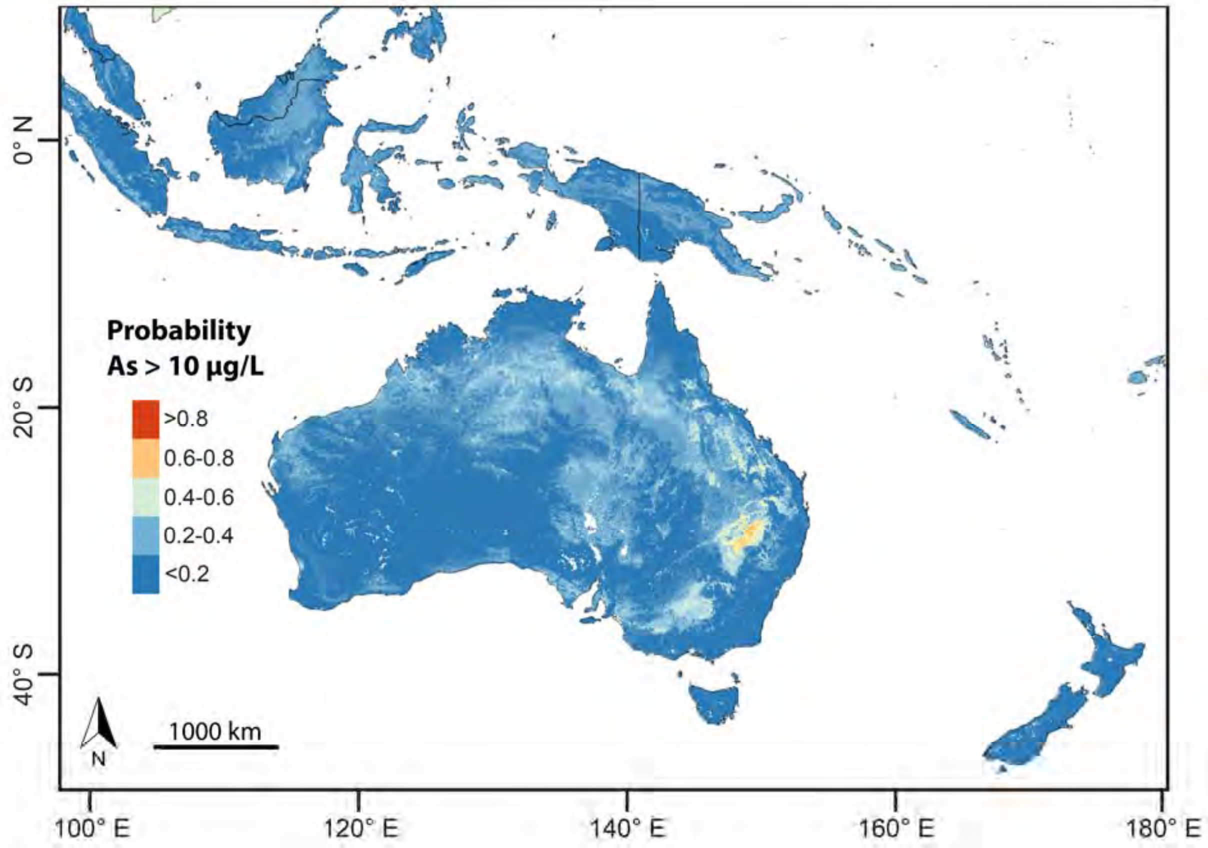
Supplementary Figure 5 | Arsenic in groundwater prediction map for continental Southeast Asia.



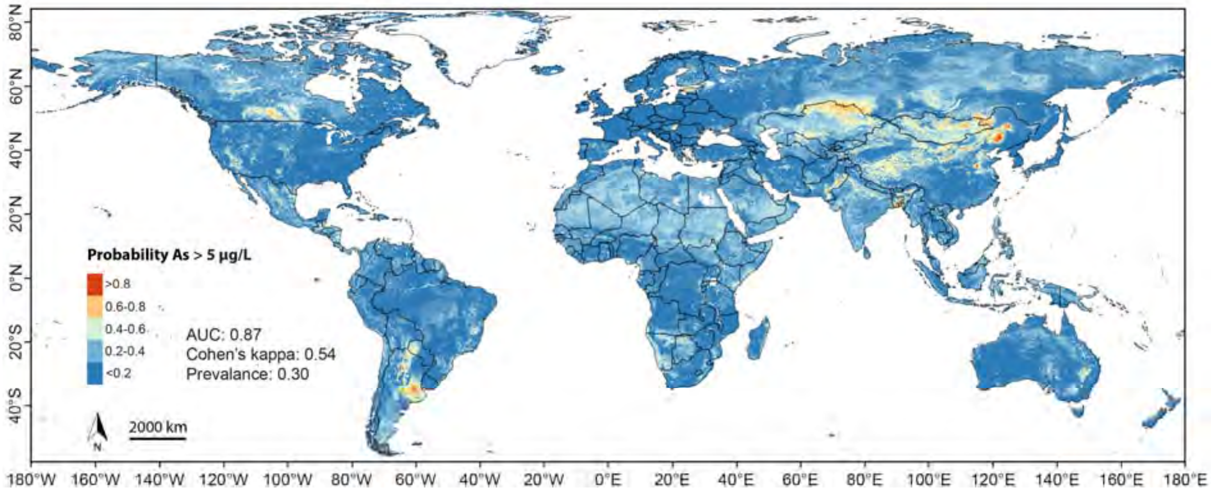
Supplementary Figure 6 | Arsenic in groundwater prediction map for South America.



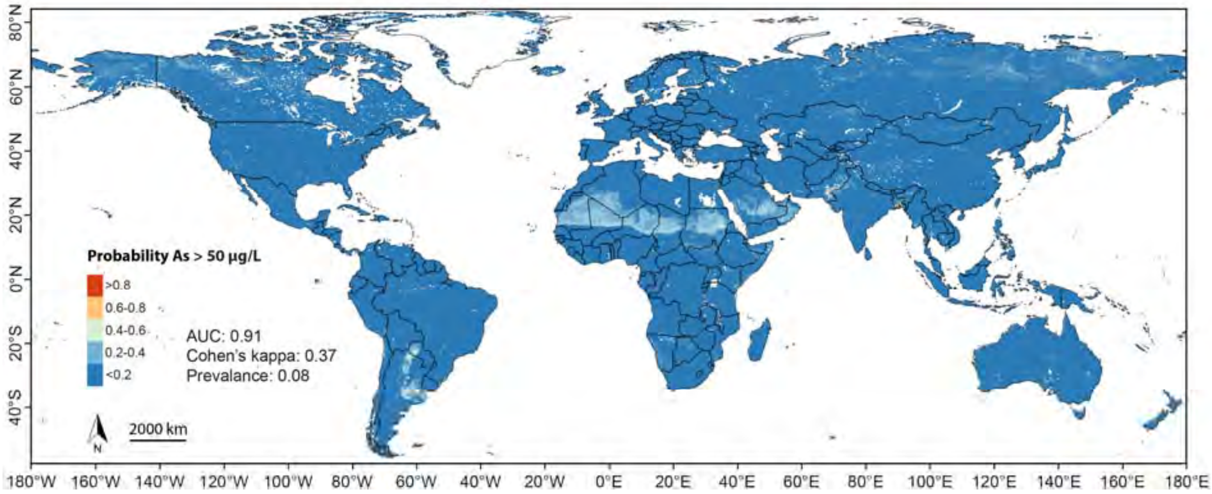
Supplementary Figure 7 | Arsenic in groundwater prediction map for Africa and the Arabian peninsula.



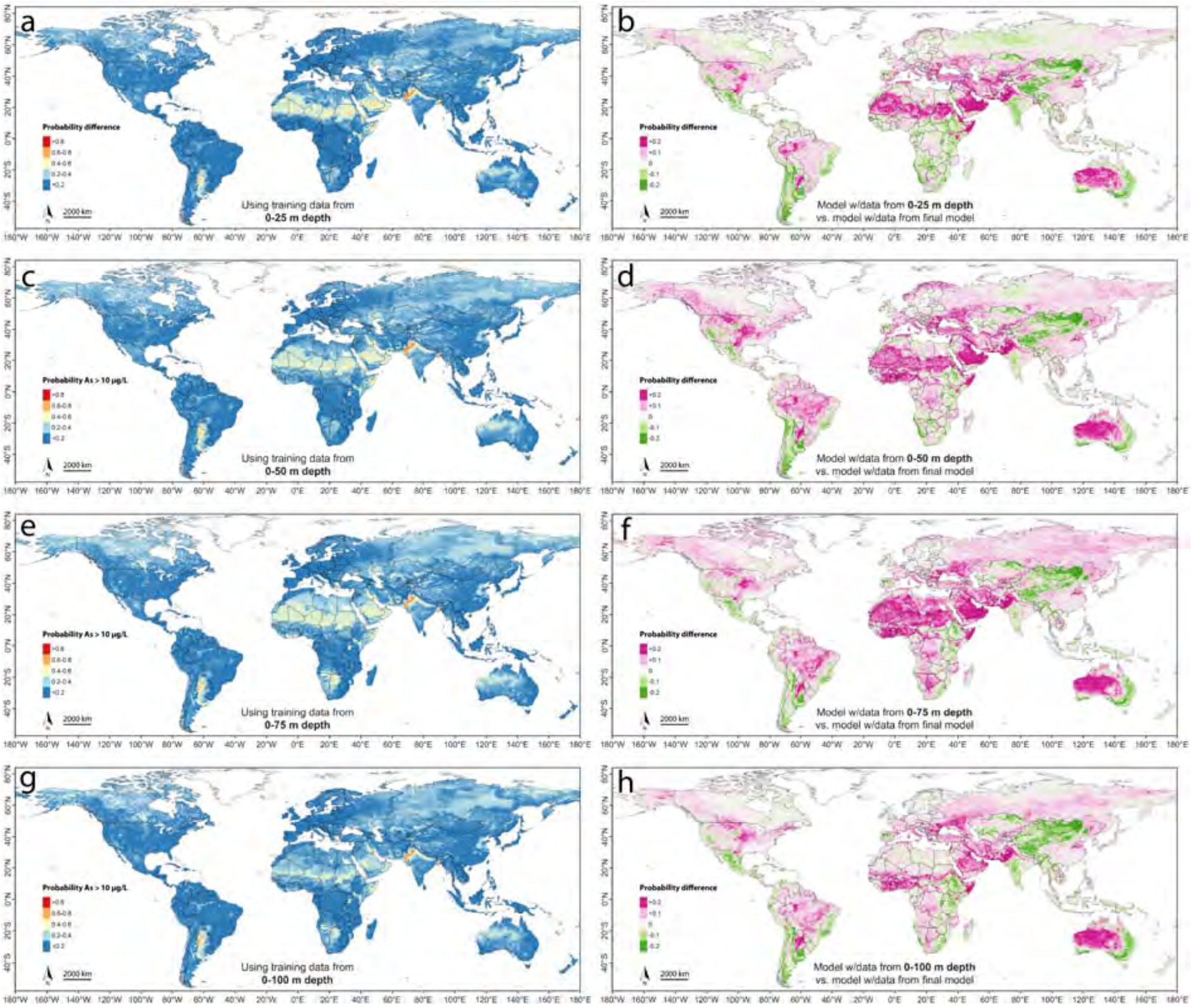
Supplementary Figure 8 | Arsenic in groundwater prediction map for Australia and southwest Pacific islands.



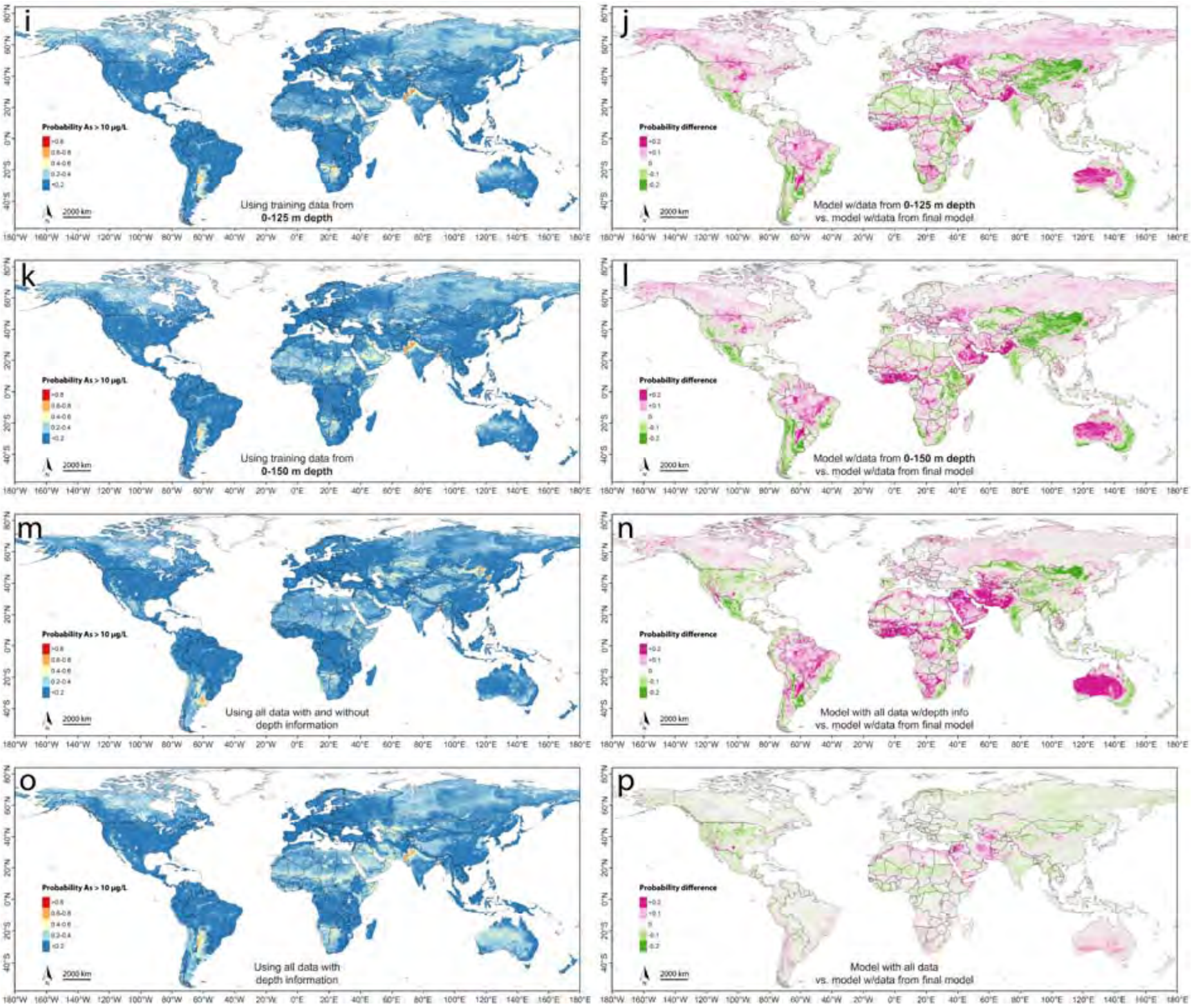
Supplementary Figure 9 | Modeled probability of arsenic concentration in groundwater exceeding 5 µg/L. This model was created using the same training/testing data and variables as the final model of the paper (Fig. 2).



Supplementary Figure 10 | Modeled probability of arsenic concentration in groundwater exceeding 50 µg/L. This model was created using the same training/testing data and variables as the final model of the paper (Fig. 2).



Supplementary Figure 11 | Prediction models using different depth ranges of data. All predictor variables were used with different depth ranges of predictor data (see Table S6). Maps in the right-side column show probability differences of the model in the left-side column relative to the model using data from 0-100m plus data without depth information (same as used in the final model, Fig. 2).



Supplementary Figure 11 (cont.)

Supplementary Table 1 | Groundwater arsenic measurements. Listed are all of the measurements considered for use in the study. For modelling, data were aggregated to 1-km^2 pixels corresponding to the predictor data and averaged. The proportion of aggregated measurements in the final model coming from each country is shown in the column, "% of final data coming from country".

Country	**n**	**Avg. Conc. (µg/L)**	**% >10 µg/L**	**% with depth**	**Avg. conc., aggregated points in final model (µg/L)**	**% of final data coming from country**
Afghanistan(*56*)	108	1.8 ± 2.8	1.9	100	1.8 ± 2.8	0.19
Algeria(*57*)	4	5.5 ± 1.7	0.0	100	5.5 ± 1.7	0.01
Argentina(*14-17, 58, 59*)	685	132.5 ± 723.1	59.9	20	139 ± 775	1.04
Australia(*60-63*)	215	30.5 ± 268.9	2.1	29	1.4 ± 9.2	0.25
Bangladesh(*64*)	4129	63 ± 139.7	47.3	100	62.2 ± 127	6.28
Belgium(*65*)	315	2.1 ± 1	0.6	0	2.1 ± 1	0.56
Bosnia and Herzegovina(*65*)	16	0.4 ± 0.9	0.0	0	0.4 ± 0.9	0.03
Botswana(*66*)	54	56 ± 92.3	53.8	13	49.9 ± 81.1	0.07
Brazil(*59, 67*)	1109	5.1 ± 3	5.0	0	5.7 ± 2.9	1.23
Bulgaria(*65*)	32	1.3 ± 1.5	0.0	0	1.3 ± 1.5	0.06
Burkina Faso(*34*)	1486	7.6 ± 33.8	15.4	0	7.3 ± 27.4	2.01
Cambodia(*18, 19, 68*)	42909	59.4 ± 131	17.6	3	21.3 ± 72.2	10.89
Canada(*69, 70*)	44	1.6 ± 2.6	2.4	41	1.5 ± 2.7	0.07
China(*20-22*)	3540	25 ± 106.1	34.8	7	23.7 ± 105	5.37
Croatia(*65*)	7	1.3 ± 0.4	0.0	0	1.2 ± 0.4	0.01
Cyprus(*65*)	99	2.1 ± 5.1	3.3	0	2.1 ± 5.2	0.16
Czech Republic(*65*)	702	2.2 ± 4.4	3.7	0	2.1 ± 4.3	1.10
Dem. Rep. Congo(*57*)	1	10.0	0.0	100	10.0	0.00
Denmark(*65*)	1383	1.9 ± 4.3	2.6	0	1.5 ± 2.7	0.88
Estonia(*65*)	68	5.2 ± 2.4	0.0	0	5.2 ± 2.4	0.09
Ethiopia(*71-73*)	155	12.1 ± 24.9	29.9	49	14.4 ± 28	0.21
France(*65*)	29	1.8 ± 1.5	0.0	0	1.8 ± 1.5	0.05
Germany(*65*)	860	1.8 ± 8.8	1.7	65	1.5 ± 7.4	1.16
Ghana(*74, 75*)	246	6 ± 32.7	5.0	7	5 ± 32.7	0.39
Greece(*65, 76*)	90	18.1 ± 22.9	44.3	73	17.7 ± 22.8	0.16
Hungary(*77*)	16	71 ± 69.4	40.9	25	39.1 ± 59.4	0.04
Iceland(*65*)	1	0.1	0.0	100	0.1	0.00
India(*23-25, 78-100*)	123436	55.6 ± 466.5	46.6	100	27.3 ± 75.8	19.47
Indonesia(*37, 101*)	485	2.5 ± 7.4	7.6	81	2.6 ± 8	0.42
Ireland(*65*)	88	3.4 ± 12	0.0	0	1.6 ± 1.6	0.07
Italy(*65*)	1440	6.5 ± 30.9	7.5	0	5.9 ± 26.9	2.50
Japan(*102*)	2	5.5 ± 6.4	0.0	0	5.5 ± 6.4	0.00
Latvia(*65*)	191	1.1 ± 1	0.0	0	1 ± 0.7	0.15
Lithuania(*65, 102*)	122	1.9 ± 3.9	0.0	0	1.3 ± 0.8	0.16
Malawi(*103*)	25	0 ± 0	0.0	100	0 ± 0	0.04
Mali(*57*)	1	0.0	0.0	100	1 ± 1.4	0.00
Mexico(*26, 104*)	1561	25.7 ± 303.6	33.9	2	18.4 ± 65.5	2.34

Morocco(*102*)	2	2.6 ± 2.1	0.0	0	2.6 ± 2.1	0.00
Myanmar(*105*)	55	69.6 ± 119.4	50.0	100	73.4 ± 118	0.01
Nepal(*106*)	7575	15.8 ± 61.5	26.8	91	12.8 ± 37.5	3.19
Netherlands(*65*)	196	5.3 ± 13.1	8.8	0	5.3 ± 13.2	0.35
Nicaragua(*59*)	388	21.9 ± 76.3	15.7	0	7.9 ± 15.3	0.37
North Macedonia(*65*)	35	1.4 ± 1.3	0.0	0	1.3 ± 1.1	0.06
Norway(*107, 108*)	477	0.8 ± 2	1.2	95	0.7 ± 1.9	0.61
Pakistan(*28, 29*)	1281	102.5 ± 123.1	49.5	31	64.2 ± 96.9	1.36
Peru(*109*)	56	26.6 ± 112	12.0	88	29 ± 118	0.09
Poland(*65*)	1130	2.8 ± 3.9	2.0	0	2.8 ± 4	1.80
Portugal(*59, 65*)	745	9 ± 46.4	7.6	0	6.7 ± 28.1	1.06
Romania(*77*)	56	29.3 ± 55.7	36.4	66	16.3 ± 40.1	0.04
Russia(*108*)	1	0.1	0.0	0	5.5 ± 3.4	0.01
Serbia(*65*)	77	3.9 ± 8.5	12.3	0	4.1 ± 8.7	0.13
Slovakia(*65*)	234	9.9 ± 108	4.1	1	10.4 ± 111	0.39
Slovenia(*65*)	56	0.4 ± 0.5	0.0	98	0.3 ± 0.5	0.10
South Africa(*110*)	56	6.1 ± 24.8	5.4	0	6.1 ± 24.8	0.10
Spain(*59*)	123	38.4 ± 50.6	49.4	0	38.8 ± 48.3	0.14
Sweden(*65, 111*)	595	1.6 ± 6.2	4.3	89	1.3 ± 6.1	0.53
Switzerland(*112, 113*)	1027	2.9 ± 9.4	5.8	7	2.8 ± 9.5	1.42
Tanzania(*114*)	48	2.3 ± 5	6.7	0	2.4 ± 5.2	0.08
United Kingdom(*65, 115*)	2804	3.1 ± 14.5	4.6	8	3.1 ± 14.9	4.64
United States (*30, 69*)	50625	8.8 ± 123	6.9	88	4.9 ± 43.7	24.43
Vietnam(*32, 33, 116*)	1140	31.5 ± 72.3	33.3	57	25.1 ± 60.2	1.60
TOTAL	**254436**	**42 ± 339**	**39.8**	**72.3**	**16 ± 102**	**100**

Supplementary Table 2 | Independent variables tested for modeling. The 11 variables ultimately used in the final model are indicated in bold.

Dataset	Resolution
Climate	
Actual evapotranspiration (AET)(*117*)	**30”**
Aridity (PET(*118*)/precipitation(*119*))	**30”**
Potential evapotranspiration (PET)(*118*)	**30”**
Precipitation(*119*)	**30”**
Priestley-Taylor alpha coefficient(*117*) (AET/PET)	**30”**
Temperature(*120*)	**30”**
Geology	
Carbonate sedimentary rocks(*121*)	polygon
Felsic igneous rocks(*121*)	polygon
Felsic plutonic rocks(*121*)	polygon
Felsic volcanic rocks(*121*)	polygon
Igneous rocks(*121*)	polygon
Intermediate igneous rocks(*121*)	polygon
Intermediate plutonic rocks(*121*)	polygon
Intermediate volcanic rocks(*121*)	polygon
Mafic igneous rocks(*121*)	polygon
Mafic plutonic rocks(*121*)	polygon
Mafic volcanic rocks(*121*)	polygon
Metamorphic rocks(*121*)	polygon
Non-carbonate sedimentary rocks(*121*)	polygon
Plutonic rocks(*121*)	polygon
Pyroclastic rocks(*121*)	polygon
Quaternary units(*122*)	polygon
Sedimentary rocks, all(*121*)	polygon
Sedimentary rocks, carbonate(*121*)	polygon
Sedimentary rocks, other(*121*)	polygon
Volcanic rocks(*121*)	polygon
Soil	
Andosols(*123*)	30”
Calcisols(*123*)	30”
Cation exchange capacity(*123*)	30”
Clay (weight percentage, 0 cm depth)(*123*)	**30”**
Clay (weight percentage, 200 cm depth)(*123*)	30”
Coarse fragments (volumetric percentage, 0 cm depth)(*123*)	30”
Coarse fragments (volumetric percentage, 200 cm depth)(*123*)	30”
Fluvisols(*123*)	**30”**
Gleysols(*123*)	30”
Hydrologic soil groups A and A/D (>90% sand and <10% clay)(*124*)	30”
Hydrologic soil groups B and B/D (50-90% sand and 10-20% clay)(*124*)	30”
Hydrologic soil groups C and C/D (<50% sand and 20-40% clay)(*124*)	30”
Hydrologic soil groups D and D/D (<50% sand and >40% clay)(*124*)	30”
Organic carbon content(*123*)	30”
pH (200 cm depth) (*123*)	**30”**
Sand (weight percentage, 0 cm depth)(*123*)	30”
Sand (weight percentage, 200 cm depth)(*123*)	**30”**

Silt (weight percentage, 0 cm depth)(*123*)	30”
Silt (weight percentage, 200 cm depth)(*123*)	30”
Soil and sedimentary deposit thickness(*125*)	30”
Solonchaks(*123*)	30”
Water capacity until wilting point(*123*)	30”
Other	
Surface slope(*126*)	30”
Topographic wetness index(*126*)	**30”**
Urbanization(*54*)	30”
Water table depth(*127*)	30”

Supplementary Table 3 | Summary of the predictor variables used in previous statistical learning classification models of arsenic concentrations in groundwater.

Country/region	Predictor variables	Geochemical setting	Reported AUC
Burkina Faso(*34*)	Metamorphic and igneous intrusive and extrusive rocks	hard rock	0.57-0.83
China(*21*)	Distance to rivers, gravity, Holocene sediments, river density, saline soils, slope, subsoil texture, topographic wetness index	arid-oxidizing/reducing	n/a
China (Shanxi Province)	Distance to rivers, gravity, saline soils, topographic index, topographic wetness index, vegetation index	reducing	n/a
Global sedimentary basins (*40*)	Aridity, carbon to nitrogen ration of subsoil, distance to rivers, distance to volcanoes/volcanic rocks, drainage condition, elevation, evapotranspiration, irrigation, precipitation, slope, soil drainage, subsoil/topsoil texture, subsoil organic carbon, subsoil soil pH, temperature	arid-oxidizing/reducing sedimentary basins	n/a
India (Uttar Pradesh)	Fluvisols, geology, groundwater level, land use, slope, soil organic carbon, soil texture	reducing	0.74
Pakistan(*29*)	Aridity, Holocene fluvial sediments, irrigated area, slope, soil organic carbon, Soil pH	arid-oxidizing	0.80
Southeast Asia(*37*)	Alluvial/deltaic/floodplain deposits, subsoil/topsoil texture, organic-rich deposits	reducing	~0.7
USA(*31*)	As C/Be C/Bi C/Mo C/Sb C-soil-horizon concentrations, base flow index, depth to bedrock/groundwater, elevation difference in watershed, geological age (Cambrian to Quaternary), drainage condition, intrusive/extrusive igneous rocks, land cover (crops), saline lake sediments, sand in soil, slope	arid-oxidizing/reducing, hard rock	0.82
USA (Maine)	Geology, water geochemistry	hard rock	n/a
USA (south Louisiana)	Distance to rivers, geology, soil texture	reducing	0.76
Vietnam (Red River Delta)(*32*)	Alluvial/deltaic deposits, medium-textured soils, organic-rich deposits	reducing	n/a

Supplementary Table 4 | Percentage of population potentially affected by consuming arsenic >10 µg/L from groundwater and area with high arsenic hazard by continent as a range of values based on the cutoffs of 0.57 and 0.72.

Continent	Area (km^2) with high As hazard	Population potentially affected
Asia	636,000-2,895,000 (1.41% – 6.44%)	90,800,000 – 206,800,000
Africa	15,000-591,000 (0.05% – 1.97%)	425,000 – 8,100,000
South America	345,000-849,000 (1.94% – 4.77%)	2,400,000 – 3,600,000
North America	59,000-448,000 (0.24% – 1.85%)	375,000 – 1,250,000
Europe	6,000-33,000 (0.06% – 0.34%)	102,000 – 525,000
Oceania	23,000-110,000 (0.28% – 1.35%)	<1,000
TOTAL	**1,084,000-4,926,000 (0.81% – 3.70%)**	**94,102,000 – 220,275,000**

Supplementary Table 5 | Potentially arsenic-affected population by country. Range is based on cutoffs of 0.57 and 0.72.

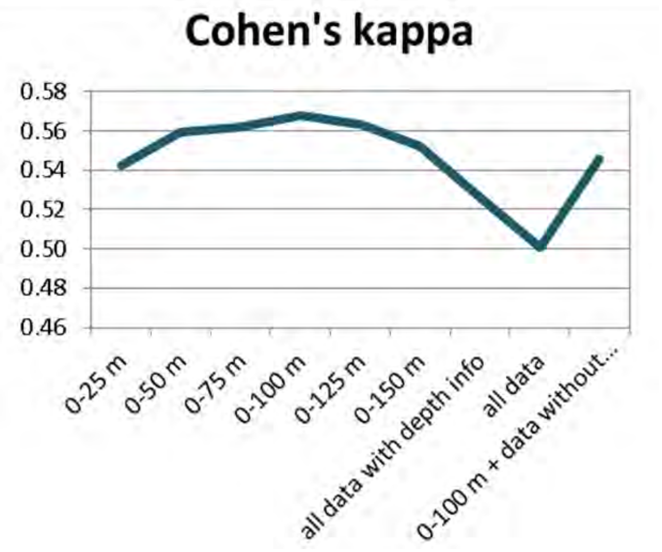
Country	Potentially affected population (10 µg/L)
Afghanistan	2 - 32,651
Algeria	7 - 9,451
Angola	224 - 16,551
Argentina	2,391,606 - 3,432,091
Australia	148 - 890
Austria	0 - 8
Bangladesh	51,371,880 - 69,146,550
Belgium	0 - 189
Belize	0 - 40
Benin	0 - 2,351
Bhutan	0 - 6,502
Bolivia	3,350 - 34,863
Botswana	753 - 5,901
Brazil	8,172 - 120,053
Bulgaria	0 - 2
Burkina Faso	21,996 - 274,577
Burundi	1,147 - 203,921
Côte d'Ivoire	0 - 1,048
Cambodia	278,774 - 524,256
Cameroon	0 - 139,050
Canada	0 - 429
Central African Rep.	0 - 1,787
Chad	135 - 255,304
Chile	0 - 48
China	4,308,100 - 12,149,940
Colombia	36 - 2,195
Congo	0 - 45
Croatia	0 - 4
Cuba	0 - 15,394
Cyprus	0 - 58
Czech Republic	0 - 24

Dem. Rep. Congo	1,289 - 49,228
Denmark	31 - 453
Djibouti	8 - 847
Dominican Rep.	1 - 840
Ecuador	0 - 6,050
Egypt	0 - 169,388
El Salvador	43 - 324
Eq. Guinea	0 - 9
Eritrea	10 - 21,001
Eswatini	0 - 197
Ethiopia	315,840 - 3,888,376
Finland	0 - 2
France	0 - 27
Gabon	0 - 1
Germany	0 - 3
Ghana	296 - 24,935
Greece	476 - 2,004
Guatemala	259 - 16,967
Haiti	107 - 18,485
Honduras	213 - 3,459
Hungary	29,942 - 164,158
India	17,527,410 - 90,347,280
Indonesia	9,170 - 67,830
Iran	221 - 22,954
Iraq	18,157 - 170,171
Ireland	0 - 63
Israel	0 - 171
Italy	10,214 - 23,797
Jordan	0 - 26
Kazakhstan	36,219 - 295,985
Kenya	37,439 - 405,190
Kuwait	5,255 - 61,458
Kyrgyzstan	378 - 3,147
Laos	0 - 133
Lesotho	0 - 15
Libya	5 - 2,988
Madagascar	58 - 45,337
Malawi	0 - 8,005
Mali	955 - 131,336
Mauritania	11 - 92,956
Mexico	353,877 - 977,231
Moldova	0 - 21
Mongolia	226,112 - 550,620
Morocco	103 - 44,357
Mozambique	280 - 73,235
Myanmar	19,659 - 1,859,850

N. Cyprus	0 - 88
Namibia	130 - 29,959
Nepal	315,985 - 858,837
Netherlands	64 - 3,646
Nicaragua	1,422 - 16,259
Niger	1,997 - 762,295
Nigeria	0 - 218,666
Norway	0 - 6
Oman	0 - 235
Pakistan	15,932,580 - 27,002,110
Palestine	0 - 67
Panama	0 - 189
Papua New Guinea	0 - 63
Paraguay	2,654 - 4,943
Peru	28 - 2,088
Philippines	0 - 11,501
Poland	0 - 38
Portugal	81 - 1,293
Qatar	0 - 211
Romania	13,444 - 73,196
Russia (including Asian part)	29,632 - 186,941
Rwanda	0 - 54,052
S. Sudan	0 - 28,000
Saudi Arabia	1 - 1,090
Senegal	0 - 9,303
Serbia	705 - 9,458
Slovakia	6 - 54
Somalia	1,266 - 330,639
Somaliland	10,746 - 49,714
South Africa	4 - 5,854
Spain	176 - 927
Sri Lanka	0 - 66
Sudan	1,036 - 127,991
Sweden	4 - 202
Switzerland	123 - 575
Syria	0 - 4,157
Taiwan	32,452 - 236,214
Tajikistan	3 - 133
Tanzania	24,938 - 471,639
Thailand	9 - 3,317
Tunisia	0 - 999
Turkey	0 - 809
Turkmenistan	46 - 63,572
Uganda	3,615 - 71,741
United Arab Emirates	0 - 84
United Kingdom	16,672 - 57,589

United States of America	21,837 - 207,249
Uruguay	30 - 255
Uzbekistan	6,264 - 229,305
Venezuela	4 - 5,428
Vietnam	730,240 - 3,151,414
Yemen	7 - 19,625
Zambia	12 - 25,110
Zimbabwe	57 - 44,987
TOTAL:	94,128,637 - 220,311,265

Supplementary Table 6 | Analysis of the effect on random forest model performance by the selection of concentration data based on depth range. All predictor variables were used in each model.

Depth range of concentration data	No. data points	Prevalence	Cohen's kappa
0-25 m	15,298	0.3804	0.5426
0-50 m	22,320	0.3423	0.5591
0-75 m	25,776	0.3192	0.5618
0-100 m	28,040	0.2993	0.5678
0-125 m	29,495	0.2902	0.563
0-150 m	30,649	0.2817	0.552
all data with depth info	56,801	0.2014	0.5262
all data	86,905	0.18387	0.5009
0-100 m + data without depth info	58,445	0.2217	0.5456
			

References and Notes

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

2. M. F. Hughes, Arsenic toxicity and potential mechanisms of action. *Toxicol. Lett.* **133**, 1–16 (2002). doi:10.1016/S0378-4274(02)00084-X Medline

3. D. Mondal, M. Banerjee, M. Kundu, N. Banerjee, U. Bhattacharya, A. K. Giri, B. Ganguli, S. Sen Roy, D. A. Polya, Comparison of drinking water, raw rice and cooking of rice as arsenic exposure routes in three contrasting areas of West Bengal, India. *Environ. Geochem. Health* **32**, 463–477 (2010). doi:10.1007/s10653-010-9319-5 Medline

4. H. G. Gorchev, G. Ozolins; WHO, WHO guidelines for drinking-water quality. *WHO Chron.* **38**, 104–108 (1984). Medline

5. P. Smedley, D. Kinniburgh, A review of the source, behaviour and distribution of arsenic in natural waters. *Appl. Geochem.* **17**, 517–568 (2002). doi:10.1016/S0883-2927(02)00018-5

6. R. Nickson, J. McArthur, W. Burgess, K. M. Ahmed, P. Ravenscroft, M. Rahman, Arsenic poisoning of Bangladesh groundwater. *Nature* **395**, 338–338 (1998). doi:10.1038/26387 Medline

7. J. McArthur, P. Ravenscroft, S. Safiulla, M. Thirlwall, Arsenic in groundwater: Testing pollution mechanisms for sedimentary aquifers in Bangladesh. *Water Resour. Res.* **37**, 109–117 (2001). doi:10.1029/2000WR900270

8. M. Berg, P. T. K. Trang, C. Stengel, J. Buschmann, P. H. Viet, N. Van Dan, W. Giger, D. Stüben, Hydrological and sedimentary controls leading to arsenic contamination of groundwater in the Hanoi area, Vietnam: The impact of iron-arsenic ratios, peat, river bank deposits, and excessive groundwater abstraction. *Chem. Geol.* **249**, 91–112 (2008). doi:10.1016/j.chemgeo.2007.12.007

9. S. Fendorf, H. A. Michael, A. van Geen, Spatial and temporal variations of groundwater arsenic in South and Southeast Asia. *Science* **328**, 1123–1127 (2010). doi:10.1126/science.1172974 Medline

10. M. I. Litter, A. M. Ingallinella, V. Olmos, M. Savio, G. Difeo, L. Botto, E. M. Farfán Torres, S. Taylor, S. Frangie, J. Herkovits, I. Schalamuk, M. J. González, E. Berardozzi, F. S. García Einschlag, P. Bhattacharya, A. Ahmad, Arsenic in Argentina: Occurrence, human health, legislation and determination. *Sci. Total Environ.* **676**, 756–766 (2019). doi:10.1016/j.scitotenv.2019.04.262 Medline

11. Y. Zheng, S. V. Flanagan, The case for universal screening of private well water quality in the U.S. and testing requirements to achieve it: Evidence from arsenic. *Environ. Health Perspect.* **125**, 085002 (2017). doi:10.1289/EHP629 Medline

12. P. Bhattacharya, D. Chatterjee, G. Jacks, Occurrence of arsenic-contaminated groundwater in alluvial aquifers from delta plains, eastern India: Options for safe drinking water supply. *Int. J. Water Resour. Dev.* **13**, 79–92 (1997). doi:10.1080/07900629749944

13. A. van Geen, Y. Zheng, R. Versteeg, M. Stute, A. Horneman, R. Dhar, M. Steckler, A. Gelman, C. Small, H. Ahsan, J. H. Graziano, I. Hussain, K. M. Ahmed, Spatial variability of arsenic in 6000 tube wells in a 25 km^2 area of Bangladesh. *Water Resour. Res.* **39**, 1140 (2003). doi:10.1029/2002WR001617

14. H. B. Nicolli, J. M. Suriano, M. A. Gomez Peral, L. H. Ferpozzi, O. A. Baleani, Groundwater contamination with arsenic and other trace elements in an area of the

Pampa, Province of Córdoba, Argentina. *Environ. Geol. Water Sci.* **14**, 3–16 (1989). doi:10.1007/BF01740581

15. P. Smedley, H. Nicolli, D. Macdonald, A. Barros, J. Tullio, Hydrogeochemistry of arsenic and other inorganic constituents in groundwaters from La Pampa, Argentina. *Appl. Geochem.* **17**, 259–284 (2002). doi:10.1016/S0883-2927(01)00082-8
16. M. Blarasin, A. Cabrera, E. Matteoda, paper presented at the XXXIII IAH – 7° ALHSUD Congress, Zacatecas, Mexico, 11 to 15 October 2004.
17. M. Auge, G. E. Viale, L. Sierra, in *VIII Congreso Argentino de Hidrogeología: Aguas subterráneas recurso estratégico* (Editorial de la Universidad Nacional de La Plata, 2013), vol. 2, pp. 58–63.
18. M. Berg, H. C. Tran, T. C. Nguyen, H. V. Pham, R. Schertenleib, W. Giger, Arsenic contamination of groundwater and drinking water in Vietnam: A human health threat. *Environ. Sci. Technol.* **35**, 2621–2626 (2001). doi:10.1021/es010027y Medline
19. J. Buschmann, M. Berg, C. Stengel, M. L. Sampson, Arsenic and manganese contamination of drinking water resources in Cambodia: Coincidence of risk areas with low relief topography. *Environ. Sci. Technol.* **41**, 2146–2152 (2007). doi:10.1021/es062056k Medline
20. P. Smedley, M. Zhang, G. Zhang, Z. Luo, Mobilisation of arsenic and other trace elements in fluviolacustrine aquifers of the Huhhot Basin, Inner Mongolia. *Appl. Geochem.* **18**, 1453–1477 (2003). doi:10.1016/S0883-2927(03)00062-3
21. L. Rodríguez-Lado, G. Sun, M. Berg, Q. Zhang, H. Xue, Q. Zheng, C. A. Johnson, Groundwater arsenic contamination throughout China. *Science* **341**, 866–868 (2013). doi:10.1126/science.1237484 Medline
22. Y. Zhou, Y. Zeng, J. Zhou, H. Guo, Q. Li, R. Jia, Y. Chen, J. Zhao, Distribution of groundwater arsenic in Xinjiang, PR China. *Appl. Geochem.* **77**, 116–125 (2017). doi:10.1016/j.apgeochem.2016.09.005
23. D. Chatterjee, R. Roy, B. Basu, Riddle of arsenic in groundwater of Bengal Delta Plain—Role of non-inland source and redox traps. *Environ. Geol.* **49**, 188–206 (2005). doi:10.1007/s00254-005-0011-5
24. B. Nath, D. Stüben, S. B. Mallik, D. Chatterjee, L. Charlet, Mobility of arsenic in West Bengal aquifers conducting low and high groundwater arsenic. Part I: Comparative hydrochemical and hydrogeological characteristics. *Appl. Geochem.* **23**, 977–995 (2008). doi:10.1016/j.apgeochem.2007.11.016
25. B. A. Shah, Arsenic-contaminated groundwater in Holocene sediments from parts of middle Ganga plain, Uttar Pradesh, India. *Curr. Sci.* **98**, 1359–1365 (2010).
26. B. Planer-Friedrich, *Hydrogeological and Hydrochemical Investigations in the Rioverde Basin, Mexico* (Verlag nicht ermittelbar, 2000).
27. M. T. Alarcón-Herrera, J. Bundschuh, B. Nath, H. B. Nicolli, M. Gutierrez, V. M. Reyes-Gomez, D. Nuñez, I. R. Martín-Dominguez, O. Sracek, Co-occurrence of arsenic and fluoride in groundwater of semi-arid regions in Latin America: Genesis, mobility and remediation. *J. Hazard. Mater.* **262**, 960–969 (2013). doi:10.1016/j.jhazmat.2012.08.005 Medline
28. R. Nickson, J. McArthur, B. Shrestha, T. Kyaw-Myint, D. Lowry, Arsenic and other drinking water quality issues, Muzaffargarh District, Pakistan. *Appl. Geochem.* **20**, 55–68 (2005). doi:10.1016/j.apgeochem.2004.06.004

29. J. E. Podgorski, S. A. M. A. S. Eqani, T. Khanam, R. Ullah, H. Shen, M. Berg, Extensive arsenic contamination in high-pH unconfined aquifers in the Indus Valley. *Sci. Adv.* **3**, e1700935 (2017). doi:10.1126/sciadv.1700935 Medline

30. J. D. Ayotte, M. G. Nielsen, G. R. Robinson Jr., R. B. Moore, Relation of arsenic, iron, and manganese in ground water to aquifer type, bedrock lithogeochemistry, and land use in the New England Coastal Basins. *Water Resour. Invest. Rep.* **99**, 4162 (1999).

31. J. D. Ayotte, L. Medalie, S. L. Qi, L. C. Backer, B. T. Nolan, Estimating the high-arsenic domestic-well population in the conterminous United States. *Environ. Sci. Technol.* **51**, 12443–12454 (2017). doi:10.1021/acs.est.7b02881 Medline

32. M. Berg, C. Stengel, T. K. Pham, H. V. Pham, M. L. Sampson, M. Leng, S. Samreth, D. Fredericks, Magnitude of arsenic pollution in the Mekong and Red River Deltas—Cambodia and Vietnam. *Sci. Total Environ.* **372**, 413–425 (2007). doi:10.1016/j.scitotenv.2006.09.010 Medline

33. L. H. Winkel, T. K. Pham, M. L. Vi, C. Stengel, M. Amini, T. H. Nguyen, H. V. Pham, M. Berg, Arsenic pollution of groundwater in Vietnam exacerbated by deep aquifer exploitation for more than a century. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 1246–1251 (2011). doi:10.1073/pnas.1011915108 Medline

34. A. Bretzler, F. Lalanne, J. Nikiema, J. Podgorski, N. Pfenninger, M. Berg, M. Schirmer, Groundwater arsenic contamination in Burkina Faso, West Africa: Predicting and verifying regions at risk. *Sci. Total Environ.* **584–585**, 958–970 (2017). doi:10.1016/j.scitotenv.2017.01.147 Medline

35. Q. Zhang, L. Rodriguez-Lado, J. Liu, C. A. Johnson, Q. Zheng, G. Sun, Coupling predicted model of arsenic in groundwater with endemic arsenism occurrence in Shanxi Province, Northern China. *J. Hazard. Mater.* **262**, 1147–1153 (2013). doi:10.1016/j.jhazmat.2013.02.017 Medline

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

37. L. Winkel, M. Berg, M. Amini, S. J. Hug, C. A. Johnson, Predicting groundwater arsenic contamination in Southeast Asia from surface parameters. *Nat. Geosci.* **1**, 536–542 (2008). doi:10.1038/ngeo254

38. Q. Yang, H. B. Jung, R. G. Marvinney, C. W. Culbertson, Y. Zheng, Can arsenic occurrence rates in bedrock aquifers be predicted? *Environ. Sci. Technol.* **46**, 2080–2087 (2012). doi:10.1021/es203793x Medline

39. N. Yang, L. H. Winkel, K. H. Johannesson, Predicting geogenic arsenic contamination in shallow groundwater of south Louisiana, United States. *Environ. Sci. Technol.* **48**, 5660–5666 (2014). doi:10.1021/es405670g Medline

40. M. Amini, K. C. Abbaspour, M. Berg, L. Winkel, S. J. Hug, E. Hoehn, H. Yang, C. A. Johnson, Statistical modeling of global geogenic arsenic contamination in groundwater. *Environ. Sci. Technol.* **42**, 3669–3675 (2008). doi:10.1021/es702859e Medline

41. P. Ravenscroft, H. Brammer, K. Richards, *Arsenic Pollution: A Global Synthesis* (Wiley, 2009), vol. 28.

42. S. Murcott, *Arsenic Contamination in the World* (IWA Publishing, 2012).

43. P. Ravenscroft, "Predicting the global extent of arsenic pollution of groundwater and its potential impact on human health," unpublished report prepared for UNICEF, December 2007.

44. Y. Zheng, Lessons learned from arsenic mitigation among private well households. *Curr. Environ. Health Rep.* **4**, 373–382 (2017). doi:10.1007/s40572-017-0157-9 Medline

45. J. Podgorski, M. Berg, Podgorski_and_Berg_2020. ERIC/open (2020); doi:10.25678/0001ZT.

46. L. Breiman, Random forests. *Mach. Learn.* **45**, 5–32 (2001). doi:10.1023/A:1010933404324

47. R Core Team, R: A language and environment for statistical computing (R Foundation for Statistical Computing, Vienna, Austria, 2014).

48. T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning* (Springer, ed. 2, 2008).

49. T. K. Ho, in *Proceedings of the Third International Conference on Document Analysis and Recognition* (IEEE, 1995), vol. 1, pp. 278–282.

50. M. L. McHugh, Interrater reliability: The kappa statistic. *Biochem. Med.* **22**, 276–282 (2012). doi:10.11613/BM.2012.031 Medline

51. T. Fawcett, An introduction to ROC analysis. *Pattern Recognit. Lett.* **27**, 861–874 (2006). doi:10.1016/j.patrec.2005.10.010

52. J. Gao, Global population projection grids based on shared socioeconomic pathways (SSPs), downscaled 1-km grids, 2010–2100. NASA Socioeconomic Data and Applications Center (SEDAC) (2019); doi:10.7927/H44747X4.

53. WHO/UNICEF Joint Monitoring Program (JMP), Global data on Water Supply, Sanitation and Hygiene (WASH) (JMP, 2019); https://washdata.org/data/household#!/.

54. M. A. Friedl, D. Sulla-Menashe, B. Tan, A. Schneider, N. Ramankutty, A. Sibley, X. Huang, MODIS Collection 5 global land cover: Algorithm refinements and characterization of new datasets. *Remote Sens. Environ.* **114**, 168–182 (2010). doi:10.1016/j.rse.2009.08.016

55. FAO, Food and Agriculture Organization of the United Nations (FAO), AQUASTAT main database (FAO, 2016); www.fao.org/nr/water/aquastat/data/query/index.html?lang=en.

56. R. E. Broshears, M. A. Akbari, M. P. Chornack, D. K. Mueller, B. C. Ruddy, “Inventory of ground-water resources in the Kabul Basin, Afghanistan,” (U.S. Geological Survey Scientific Investigations Report 2005-5090, USGS, 2005).

57. United Nations High Commissioner for Refugees (UNHCR), Borehole GIS portal (UNHCR, 2019); http://wash.unhcr.org/wash-gis-portal/.

58. M. E. Zabala, M. Manzano, L. Vives, Assessment of processes controlling the regional distribution of fluoride and arsenic in groundwater of the Pampeano Aquifer in the Del Azul Creek basin (Argentina). *J. Hydrol.* **541**, 1067–1087 (2016). doi:10.1016/j.jhydrol.2016.08.023

59. M. E. Morgada, M. Mateu, J. Bundschuh, M. I. Litter, Arsenic in the Iberoamerican region. The IBEROARSEN Network and a possible economic solution for arsenic removal in isolated rural zones. *e-Terra* **5**, 1–11 (2008).

60. K. Ivkovic, K. Watkins, R. Cresswell, J. Bauld, “A groundwater quality assessment of the fractured rock aquifers of the Piccadilly Valley, South Australia” (Australian Geological Survey Organisation, 1998).

61. J. Fitzgerald *et al.*, "Groundwater quality and environmental health implications, Anangu Pitjantjara Lands, South Australia" (Report, Bureau of Rural Sciences, Canberra, Australia1999), pp. 1–30.

62. S. Clohessy, "Perth shallow groundwater systems investigation: Lake Gwelup" (Hydrogeological Record Series, Department of Water, Perth, Australia, 2012).

63. R. M. Larsen, *A Groundwater Quality Assessment of the Jandakot Mound, Swan Coastal Plain, Western Australia* (Australian Geological Survey Organisation, 1998).

64. European Environment Agency (EEA), Waterbase – water quality (EEA, 2019); www.eea.europa.eu/data-and-maps/data/waterbase-water-quality-2.

65. D. Kinniburgh, P. Smedley, "Arsenic contamination of groundwater in Bangladesh" (British Geological Survey, 2001).

66. Water Resources Consultants, "Maun groundwater development project, Phase 2: Resources assessment and wellfield development, Final report" (TB 10/3/5/2000-2001, Department of Water Affairs, Republic of Botswana, 2004).

67. Geological Survey of Brazil (CPRM), SIAGAS (2017); http://siagasweb.cprm.gov.br.

68. Ministry of Rural Development of Cambodia, Cambodia WellMap (2015); www.cambodiawellmap.com/.

69. E. K. Read, L. Carr, L. De Cicco, H. A. Dugan, P. C. Hanson, J. A. Hart, J. Kreft, J. S. Read, L. A. Winslow, Water quality data for national-scale aquatic research: The Water Quality Portal. *Water Resour. Res.* **53**, 1735–1745 (2017). doi:10.1002/2016WR019993

70. D. R. Boyle, W. A. Spirito, S. W. Adcock, "Groundwater hydrogeochemical survey of central New Brunswick" (Open file 3306, Geological Survey of Canada, 1996).

71. C. Reimann, K. Bjorvatn, R. Tekle-Haimanot, Z. Melako, U. Siewers, "Drinking water quality, Rift Valley, Ethiopia" (Report 2002.0333, Geological Survey of Norway, 2002), p. 131.

72. A. Bretzler, K. Osenbrück, R. Gloaguen, J. S. Ruprecht, S. Kebede, S. Stadler, Groundwater origin and flow dynamics in active rift systems–A multi-isotope approach in the Main Ethiopian Rift. *J. Hydrol. (Amst.)* **402**, 274–289 (2011). doi:10.1016/j.jhydrol.2011.03.022

73. T. Rango, G. Bianchini, L. Beccaluva, R. Tassinari, Geochemistry and water quality assessment of central Main Ethiopian Rift natural waters with emphasis on source and occurrence of fluoride and arsenic. *J. Afr. Earth Sci.* **57**, 479–491 (2010). doi:10.1016/j.jafrearsci.2009.12.005

74. B. Kortatsi, J. Asigbe, G. A. Dartey, C. Tay, G. K. Anornu, E. Hayford, Reconnaissance survey of arsenic concentration in ground-water in south-eastern Ghana. *West Afr. J. Appl. Ecol.* **13**, 16–26 (2008). doi:10.4314/wajae.v13i1.40586

75. P. L. Smedley, Arsenic in rural groundwater in Ghana: Part special issue: Hydrogeochemical studies in sub-Saharan Africa. *J. Afr. Earth Sci.* **22**, 459–470 (1996). doi:10.1016/0899-5362(96)00023-1

76. I. A. Katsoyiannis, S. J. Hug, A. Ammann, A. Zikoudi, C. Hatziliontos, Arsenic speciation and uranium concentrations in drinking water supply wells in Northern Greece: Correlations with redox indicative parameters and implications for groundwater treatment. *Sci. Total Environ.* **383**, 128–140 (2007). doi:10.1016/j.scitotenv.2007.04.035 Medline

77. H. A. Rowland, E. O. Omoregie, R. Millot, C. Jimenez, J. Mertens, C. Baciu, S. J. Hug, M. Berg, Geochemistry and arsenic behaviour in groundwater resources of the

Pannonian Basin (Hungary and Romania). *Appl. Geochem.* **26**, 1–17 (2011). doi:10.1016/j.apgeochem.2010.10.006

78. S. Chandra, S. Ahmed, E. Nagaiah, S. K. Singh, P. Chandra, Geophysical exploration for lithological control of arsenic contamination in groundwater in Middle Ganga Plains, India. *Phys. Chem. Earth Parts ABC* **36**, 1353–1362 (2011). doi:10.1016/j.pce.2011.05.009

79. T. Ghosh, R. Kanchan, Geoenvironmental appraisal of groundwater quality in Bengal alluvial tract, India: A geochemical and statistical approach. *Environ. Earth Sci.* **72**, 2475–2488 (2014). doi:10.1007/s12665-014-3155-3

80. A. Mukherjee, A. E. Fryar, E. M. Eastridge, R. S. Nally, M. Chakraborty, B. R. Scanlon, Controls on high and low groundwater arsenic on the opposite banks of the lower reaches of River Ganges, Bengal basin, India. *Sci. Total Environ.* **645**, 1371–1387 (2018). doi:10.1016/j.scitotenv.2018.06.376 Medline

81. J. McArthur, P. Ravenscroft, D. M. Banerjee, J. Milsom, K. A. Hudson-Edwards, S. Sengupta, C. Bristow, A. Sarkar, S. Tonkin, R. Purohit, How paleosols influence groundwater flow and arsenic pollution: A model from the Bengal Basin and its worldwide implication. *Water Resour. Res.* **44**, W11411 (2008). doi:10.1029/2007WR006552

82. A. Mukherjee, A. E. Fryar, H. D. Rowe, Regional-scale stable isotopic signatures of recharge and deep groundwater in the arsenic affected areas of West Bengal, India. *J. Hydrol.* **334**, 151–161 (2007). doi:10.1016/j.jhydrol.2006.10.004

83. V. S. Chauhan, R. T. Nickson, D. Chauhan, L. Iyengar, N. Sankararamakrishnan, Ground water geochemistry of Ballia district, Uttar Pradesh, India and mechanism of arsenic release. *Chemosphere* **75**, 83–91 (2009). doi:10.1016/j.chemosphere.2008.11.065 Medline

84. D. Saha, S. Sahu, A decade of investigations on groundwater arsenic contamination in Middle Ganga Plain, India. *Environ. Geochem. Health* **38**, 315–337 (2016). doi:10.1007/s10653-015-9730-z Medline

85. D. P. Shukla, C. Dubey, N. P. Singh, M. Tajbakhsh, M. Chaudhry, Sources and controls of Arsenic contamination in groundwater of Rajnandgaon and Kanker District, Chattisgarh Central India. *J. Hydrol* **395**, 49–66 (2010). doi:10.1016/j.jhydrol.2010.10.011

86. J. P. Maity, B. Nath, C.-Y. Chen, P. Bhattacharya, O. Sracek, J. Bundschuh, S. Kar, R. Thunvik, D. Chatterjee, K. M. Ahmed, G. Jacks, A. B. Mukherjee, J.-S. Jean, Arsenic-enriched groundwaters of India, Bangladesh and Taiwan—Comparison of hydrochemical characteristics and mobility constraints. *J. Environ. Sci. Health A Tox. Hazard. Subst. Environ. Eng.* **46**, 1163–1176 (2011). doi:10.1080/10934529.2012.598711 Medline

87. S. Kar, J. P. Maity, J.-S. Jean, C.-C. Liu, B. Nath, H.-J. Yang, J. Bundschuh, Arsenic-enriched aquifers: Occurrences and mobilization of arsenic in groundwater of Ganges Delta Plain, Barasat, West Bengal, India. *Appl. Geochem.* **25**, 1805–1814 (2010). doi:10.1016/j.apgeochem.2010.09.007

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

89. B. Nath, S. J. Sahu, J. Jana, A. Mukherjee-Goswami, S. Roy, M. J. Sarkar, D. Chatterjee, Hydrochemistry of arsenic-enriched aquifer from rural West Bengal, India: A study of the arsenic exposure and mitigation option. *Water Air Soil Pollut.* **190**, 95–113 (2008). doi:10.1007/s11270-007-9583-x

90. R. A. Olea, N. J. Raju, J. J. Egozcue, V. Pawlowsky-Glahn, S. Singh, Advancements in hydrochemistry mapping: Methods and application to groundwater arsenic and iron concentrations in Varanasi, Uttar Pradesh, India. *Stochastic Environ. Res. Risk Assess.* **32**, 241–259 (2018). doi:10.1007/s00477-017-1390-3

91. M. Kumar, A. L. Ramanathan, M. M. Rahman, R. Naidu, Concentrations of inorganic arsenic in groundwater, agricultural soils and subsurface sediments from the middle Gangetic plain of Bihar, India. *Sci. Total Environ.* **573**, 1103–1114 (2016). doi:10.1016/j.scitotenv.2016.08.109 Medline

92. S. Chidambaram, R. Thilagavathi, C. Thivya, U. Karmegam, M. V. Prasanna, A. L. Ramanathan, K. Tirumalesh, P. Sasidhar, A study on the arsenic concentration in groundwater of a coastal aquifer in south-east India: An integrated approach. *Environ. Dev. Sustain.* **19**, 1015–1040 (2017). doi:10.1007/s10668-016-9786-7

93. S. Ghosh, P. Sar, Identification and characterization of metabolic properties of bacterial populations recovered from arsenic contaminated ground water of North East India (Assam). *Water Res.* **47**, 6992–7005 (2013). doi:10.1016/j.watres.2013.08.044 Medline

94. S. Sharma, J. Kaur, A. K. Nagpal, I. Kaur, Quantitative assessment of possible human health risk associated with consumption of arsenic contaminated groundwater and wheat grains from Ropar Wetand and its environs. *Environ. Monit. Assess.* **188**, 506 (2016). doi:10.1007/s10661-016-5507-9 Medline

95. B. A. Shah, Role of Quaternary stratigraphy on arsenic-contaminated groundwater from parts of Barak Valley, Assam, North–East India. *Environ. Earth Sci.* **66**, 2491–2501 (2012). doi:10.1007/s12665-011-1472-3

96. B. A. Shah, Role of Quaternary stratigraphy on arsenic-contaminated groundwater from parts of Middle Ganga Plain, UP–Bihar, India. *Environ. Geol.* **53**, 1553–1561 (2008). doi:10.1007/s00254-007-0766-y

97. B. A. Shah, Status of groundwater arsenic pollution of Mirzapur district in Holocene aquifers from parts of the Middle Ganga Plain, India. *Environ. Earth Sci.* **73**, 1505–1514 (2015). doi:10.1007/s12665-014-3501-5

98. L. Sailo, C. Mahanta, Arsenic mobilization in the Brahmaputra plains of Assam: Groundwater and sedimentary controls. *Environ. Monit. Assess.* **186**, 6805–6820 (2014). doi:10.1007/s10661-014-3890-7 Medline

99. D. Paul, S. K. Kazy, A. K. Gupta, T. Pal, P. Sar, Diversity, metabolic properties and arsenic mobilization potential of indigenous bacteria in arsenic contaminated groundwater of West Bengal, India. *PLOS ONE* **10**, e0118735 (2015). doi:10.1371/journal.pone.0118735 Medline

100. R. Nickson, C. Sengupta, P. Mitra, S. N. Dave, A. K. Banerjee, A. Bhattacharya, S. Basu, N. Kakoti, N. S. Moorthy, M. Wasuja, M. Kumar, D. S. Mishra, A. Ghosh, D. P. Vaish, A. K. Srivastava, R. M. Tripathi, S. N. Singh, R. Prasad, S. Bhattacharya, P. Deverill, Current knowledge on the distribution of arsenic in groundwater in five states of India. *J. Environ. Sci. Health Part A Tox. Hazard. Subst. Environ. Eng.* **42**, 1707–1718 (2007). doi:10.1080/10934520701564194 Medline

101. C. Marohn, A. Distel, G. Dercon, R. Tomlinson, M. Noordwijk, G. Cadisch, Impacts of soil and groundwater salinization on tree crop performance in post-tsunami Aceh Barat, Indonesia. *Nat. Hazards Earth Syst. Sci.* **12**, 2879–2891 (2012). doi:10.5194/nhess-12-2879-2012

102. UN Environment Programme (UNEP), “Water quality, 2005 state of the UNEP GEMS/Water Global Network and annual report” (UNEP, 2005).

103. M. Pritchard, T. Mkandawire, J. O’neill, Assessment of groundwater quality in shallow wells within the southern districts of Malawi. *Phys. Chem. Earth Parts ABC* **33**, 812–823 (2008). doi:10.1016/j.pce.2008.06.036

104. Inventario Nacional de Calidad del Agua (INCA), *Arsénico y fluoruro en agua: riesgos y perspectivas desde la sociedad civil y la academia en México*, L. M. Del Razo, J. M. Ledón, M. N. Velasco, Eds. (INCA, Mexico, 2020).

105. A. van Geen, K. H. Win, T. Zaw, W. Naing, J. L. Mey, B. Mailloux, Confirmation of elevated arsenic levels in groundwater of Myanmar. *Sci. Total Environ.* **478**, 21–24 (2014). doi:10.1016/j.scitotenv.2014.01.073 Medline

106. B. R. Shrestha, J. W. Whitney, K. B. Shrestha, Eds., “The state of arsenic in Nepal–2003” (National Arsenic Steering Committee, Environment and Publich Health Organization, Kathmandu, Nepal, 2004).

107. B. Frengstad, A. K. M. Skrede, D. Banks, J. R. Krog, U. Siewers, The chemistry of Norwegian groundwaters: III. The distribution of trace elements in 476 crystalline bedrock groundwaters, as analysed by ICP-MS techniques. *Sci. Total Environ.* **246**, 21–40 (2000). doi:10.1016/S0048-9697(99)00413-1 Medline

108. P. de Caritat, S. Danilova, Ø. J1ger, C. Reimann, G. Storrø, Groundwater composition near the nickel—Copper smelting industry on the Kola Peninsula, central Barents Region (NW Russia and NE Norway). *J. Hydrol.* **208**, 92–107 (1998). doi:10.1016/S0022-1694(98)00147-4

109. C. M. C. de Meyer, J. M. Rodríguez, E. A. Carpio, P. A. García, C. Stengel, M. Berg, Arsenic, manganese and aluminum contamination in groundwater resources of Western Amazonia (Peru). *Sci. Total Environ.* **607–608**, 1437–1450 (2017). doi:10.1016/j.scitotenv.2017.07.059 Medline

110. L. P. McCaffrey, J. P. Willis, “Distribution of fluoride-rich groundwater in the eastern and Mogwase regions of the Northern and North-West Provinces” (WRC Report No. 526/1/01, Water Research Commission Pretoria, 2001).

111. Geological Survey of Sweden, Data from environmental monitoring of groundwater (2007); www.sgu.se/produkter/geologiska-data/oppna-data/grundvatten-oppna-data/miljoovervakning-av-grundvatten/.

112. M. Haldimann, E. Pfammatter, P.-M. Venetz, P. Studer, V. Dudler, Occurrence of arsenic in drinking water of the canton of Valais. Part I: Overview of arsenic concentration and geographic distribution. *Mitt. Lebensmitteluntersuchung Hyg.* **96**, 89–105 (2005).

113. O. Deflorin, O. (2004). “Natürliche Radionuklide in Grundwässern des Kantons Graubünden,” thesis, Université de Neuchâtel (2004).

114. P. Smedley *et al.*, “Fluoride in groundwater from high-fluoride areas of Ghana and Tanzania” (British Geological Survey, 2002).

115. P. L. Smedley, W. M. Edmunds, Redox patterns and trace-element behavior in the East Midlands Triassic sandstone aquifer, UK. *Ground Water* **40**, 44–58 (2002). doi:10.1111/j.1745-6584.2002.tb02490.x Medline

116. J. Buschmann, M. Berg, C. Stengel, L. Winkel, M. L. Sampson, P. T. K. Trang, P. H. Viet, 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 (2008). doi:10.1016/j.envint.2007.12.025 Medline

117. A. Trabucco, R. Zomer, Global soil water balance geospatial database (CGIAR Consortium for Spatial Information, 2010); www.cgiar-csi.org [accessed January 2013].

118. A. Trabucco, R. J. Zomer, Global aridity index (global-aridity) and global potential evapo-transpiration (global-PET) geospatial database (CGIAR Consortium for Spatial Information, 2009).

119. R. J. Hijmans, S. E. Cameron, J. L. Parra, P. G. Jones, A. Jarvis, Very high resolution interpolated climate surfaces for global land areas. *Int. J. Climatol.* **25**, 1965–1978 (2005).

120. S. E. Fick, R. J. Hijmans, WorldClim 2: New 1-km spatial resolution climate surfaces for global land areas. *Int. J. Climatol.* **37**, 4302–4315 (2017).

121. J. Hartmann, N. Moosdorf, The new global lithological map database GLiM: A representation of rock properties at the Earth surface. *Geochem. Geophys. Geosyst.* **13**, Q12004 (2012). doi:10.1029/2012GC004370

122. U.S. Geological Survey–Energy Resources Program, Central Energy Resources Science Center, World geologic maps (2015); https://certmapper.cr.usgs.gov/data/apps/world-maps/.

123. T. Hengl, J. Mendes de Jesus, G. B. M. Heuvelink, M. Ruiperez Gonzalez, M. Kilibarda, A. Blagotić, W. Shangguan, M. N. Wright, X. Geng, B. Bauer-Marschallinger, M. A. Guevara, R. Vargas, R. A. MacMillan, N. H. Batjes, J. G. B. Leenaars, E. Ribeiro, I. Wheeler, S. Mantel, B. Kempen, SoilGrids250m: Global gridded soil information based on machine learning. *PLOS ONE* **12**, e0169748 (2017). doi:10.1371/journal.pone.0169748 Medline

124. C. W. Ross, L. Prihodko, J. Anchang, S. Kumar, W. Ji, N. P. Hanan, HYSOGs250m, global gridded hydrologic soil groups for curve-number-based runoff modeling. *Sci. Data* **5**, 180091 (2018). Medline

125. J. Pelletier *et al.*, Global 1-km gridded thickness of soil, regolith, and sedimentary deposit layers. ORNL DAAC (2016); doi:10.3334/ORNLDAAC/1304.

126. T. Hengl, Global DEM derivatives at 250 m, 1 km and 2 km based on the MERIT DEM, version 1.0, Zenodo (2018); doi:10.5281/zenodo.1447210.

127. Y. Fan, H. Li, G. Miguez-Macho, Global patterns of groundwater table depth. *Science* **339**, 940–943 (2013). doi:10.1126/science.1229881 Medline