ENVIRONMENTAL STUDIES

Extensive arsenic contamination in high-pH unconfined aquifers in the Indus Valley

Joel E. Podgorski,[1]* Syed Ali Musstjab Akber Shah Eqani,[2,3] Tasawar Khanam,[2] Rizwan Ullah,[4] Heqing Shen,[3] Michael Berg[1]*

Arsenic-contaminated aquifers are currently estimated to affect ~150 million people around the world. However, the full extent of the problem remains elusive. This is also the case in Pakistan, where previous studies focused on isolated areas. Using a new data set of nearly 1200 groundwater quality samples throughout Pakistan, we have created state-of-the-art hazard and risk maps of arsenic-contaminated groundwater for thresholds of 10 and 50 μg/liter. Logistic regression analysis was used with 1000 iterations, where surface slope, geology, and soil parameters were major predictor variables. The hazard model indicates that much of the Indus Plain is likely to have elevated arsenic concentrations, although the rest of the country is mostly safe. Unlike other arsenic-contaminated areas of Asia, the arsenic release process in the arid Indus Plain appears to be dominated by elevated-pH dissolution, resulting from alkaline topsoil and extensive irrigation of unconfined aquifers, although pockets of reductive dissolution are also present. We estimate that approximately 50 million to 60 million people use groundwater within the area at risk, with hot spots around Lahore and Hyderabad. This number is alarmingly high and demonstrates the urgent need for verification and testing of all drinking water wells in the Indus Plain, followed by appropriate mitigation measures.

INTRODUCTION

The trace element arsenic (As) is found throughout Earth's crust and hydrosphere (*1*). In particular, arsenic can strongly affect groundwater quality through natural geogenic leaching processes from host rocks and sediments (*2*–*5*). The general geochemical conditions that lead to mobilization of arsenic into groundwater are characterized by one or more of the following features: reducing (*6*–*8*) environments, arid oxidizing environments with elevated pH (*1*, *9*, *10*), geothermal activity (*11*, *12*) and/or oxidative weathering of sulfide minerals (*13*, *14*). Aquifers within Holocene sediments are particularly susceptible to arsenic enrichment due to the sediments' limited time of exposure to groundwater flushing such that the sediments continue to hold a relative abundance of mobilizable arsenic within its grains (*15*, *16*). Arsenic concentrations can also increase due to a low hydrological gradient, resulting in sluggish groundwater flow (*1*), as well as a strongly arid environment that leads to evaporative concentration (*11*, *17*).

Regular consumption of water containing high concentrations of arsenic can have adverse health effects, including skin disorders, lung cancer, and cardiovascular diseases (*18*, *19*). In actuality, arsenic-contaminated water is one of the most serious global health threats, with currently estimated 150 million people relying on arsenic-contaminated groundwater (*5*). The permissible concentration of arsenic in drinking water set by the World Health Organization (WHO) is 10 μg/liter, whereas the guideline in Pakistan is 50 μg/liter (*20*).

To determine where best to apply the limited resources for groundwater testing, geostatistical modeling can identify areas likely to be affected by arsenic contamination by finding statistically significant relationships measured arsenic concentrations and environmental predictors (*21*–*25*). As opposed to indicator kriging, such an approach takes into account the relevant physical processes of contaminant release and accumulation. This also has the advantage of being able to use spatially continuous predictor data sets to identify areas of high arsenic hazard, where groundwater quality data are lacking. Although this method can efficiently predict the occurrence of contamination on a large scale, it is generally ineffective at the scale of individual wells due to small-scale aquifer heterogeneities that are undetectable at the surface. Winkel *et al.* (*25*) explored the use of three-dimensional (3D) geological information in modeling, which, although more accurate, showed that models based on 2D geological information can effectively predict elevated concentrations of arsenic in groundwater.

Here, we investigate and model the distribution of arsenic in Pakistan, which faces critical water quality challenges. Although microbial contamination presents the most immediate health threat and causes one-third of all deaths in the country (*20*, *26*), arsenic and other toxic metals pose a significant health hazard through chronic exposure (*27*–*30*). While the full health effects of arsenic in Pakistan are not yet known, various studies over the past decade have uncovered arsenic-related skin disorders (*29*, *31*) and high levels of arsenic in blood and hair samples (*32*, *33*) from people living in predominantly rural areas with high exposure to arsenic in groundwater. Food crops in the Sindh (*34*, *35*) and Punjab (own measurements) provinces also indicate a potentially severe health threat due to plant uptake of arsenic via irrigation water extracted from shallow Holocene aquifers.

Numerous small-scale local studies, generally at the village level, have reported high arsenic concentrations in groundwater up to hundreds of micrograms per liter, primarily in the provinces of Punjab and Sindh (*20*, *30*, *36*–*41*). However, a lack of resources in the country has prevented the comprehensive evaluation of arsenic in groundwater (*30*). Considerable arsenic contamination has also been reported in other South and East Asian countries, for example, India, Bangladesh, Cambodia, and Vietnam (*42*–*46*). Shallow small-scale and family-based hand and motorized pumps have long been a major source of drinking water in the Indus Plain and are as widespread in Pakistan as in those other arsenic-affected regions of Asia. Higher-volume pumping with tube wells became popular throughout Pakistan in the 1960s and is used primarily not only for irrigation but also for municipal water supplies (*47*).

[1]Eawag, Swiss Federal Institute of Aquatic Science and Technology, Department Water Resources and Drinking Water, 8600 Dübendorf, Switzerland. [2]Public Health and Environment Division, Department of Biosciences, COMSATS Institute of Information Technology, Islamabad, Pakistan. [3]CAS Key Laboratory of Urban Environment and Health, Institute of Urban Environment, Chinese Academy of Sciences, Xiamen, China. [4]Department of Zoology, Mirpur University of Science and Technology, Azad Jammu and Kashmir, Pakistan.
*Corresponding author. Email: joel.podgorski@eawag.ch (J.E.P.); michael.berg@eawag.ch (M.B.)

Pakistan is characterized by the flat-lying Indus Plain in the east; the Himalaya, Karakoram, and Hindu Kush mountain ranges in the north; hill regions in the northwest; and the Baluchistan plateau in the west. With the exception of the temperate northwest, the climate is semiarid to arid. The most significant aquifers of the country are found in the Indus Plain, which is composed of up to 300 m of Quaternary alluvial deposits and permeable soils low in organic content (*48*), with groundwater yields in the range of 100 to 300 m^3/hour to 150-m depth (*49*). Windblown sands generally dominate in the neighboring desert regions (groundwater yields of 10 to 50 m^3/hour), and permeable gravels of limited extent can be found in the northwest (*49*). Because of its abundant water resources and fertile soils, the Indus Plain of Pakistan hosts extensive agricultural production and a population of over 100 million people, including the major cities of Karachi, Islamabad, Lahore, and Hyderabad (Fig. 1). On account of a highly arid climate in the Indus Plain, extensive irrigation uses groundwater resources and a widespread canal system that distributes water from the Indus River and its main tributaries across the adjacent plains.

The morphology and age of flat-lying, Holocene fluvial sediments along the Indus River and tributaries are similar to those of the well-known arsenic-affected areas of the Ganges/Brahmaputra Rivers in India and Bangladesh (*5*), the Red River in Vietnam (*45*), and the Mekong River in Cambodia and Vietnam (*46*). A chemically reducing environment generally dominates in the aquifers along these rivers, which is generally due to an abundance of organic material along with a limited supply of oxygen, and results in the desorption of arsenic from iron oxy(hydr)oxides. Depleted oxygen levels can come about, for example, due to an impermeable near-surface silt and/or clay layer that prevents contact of the aquifer with the atmosphere.

The aquifers of the Indus Plain, however, are generally unconfined and have hydraulic connectivity with the surface (*48*, *50*). This results, for example, in a strong connection between surface water of the Indus Basin Irrigation System and the underlying aquifer (*37*, *51*). Since the introduction of widespread irrigation, the water table has risen significantly with accompanying waterlogging and groundwater salinization (*37*).

Rather than being a detailed geochemical investigation, this study focuses on risk determination based on our new groundwater quality data set and has produced the first-ever statistically based arsenic hazard model and health risk map for Pakistan. Furthermore, the main geochemical conditions of arsenic release were assessed in conjunction with various environmental variables.

RESULTS

Groundwater quality measurements

Our data set of measured arsenic concentrations is displayed in Fig. 1, and 11 other measured species are also shown in figs. S1 and S2. Table S1 summarizes all of these measurements and well depth. The average pH is 7.67 ± 0.45, and the average total dissolved solids (TDS) is 556 ± 557 mg/liter.

High arsenic concentrations (>10 μg/liter) exist mainly along the Indus River and its tributaries. Very high arsenic concentrations (>200 μg/liter) were measured primarily in the southern half of the Indus Plain. Overall iron concentrations are low, averaging 0.05 ± 0.13 mg/liter and not exceeding 1.9 mg/liter. The highest concentration of iron measured in a water sample with arsenic greater than 10 μg/liter is only 0.86 mg / liter (table S1). Furthermore, low nitrate levels (average, 2.7 mg/liter; median, 1.3 mg/liter) in the wells with arsenic >10 μg/liter

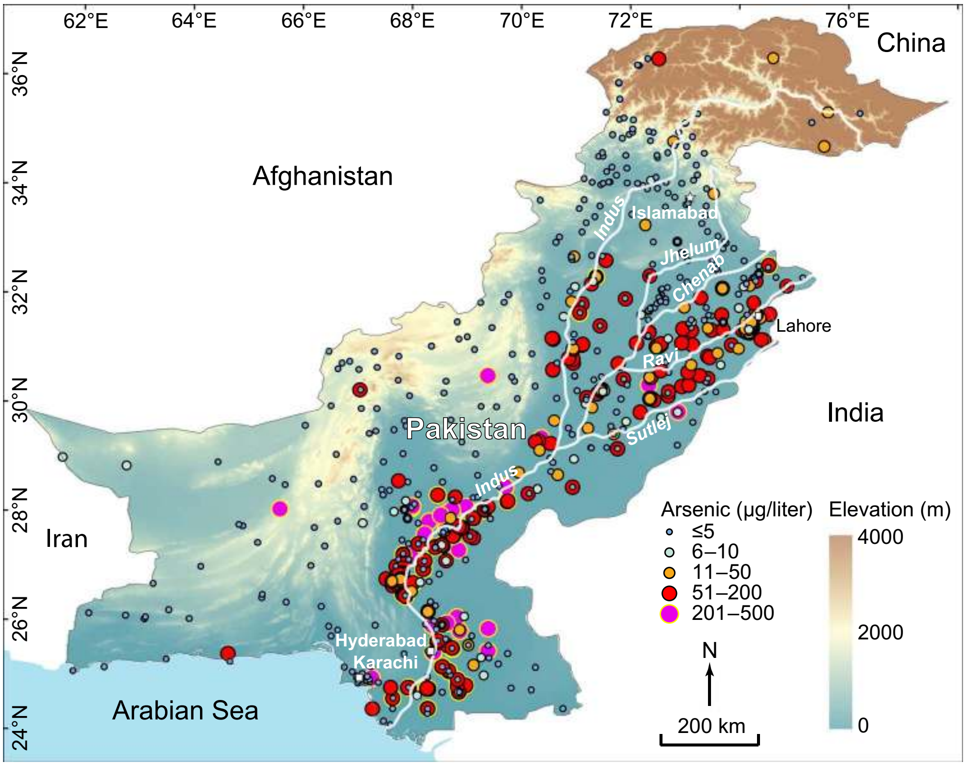

Fig. 1. Arsenic concentrations measured in Pakistan groundwater. Arsenic exceeds the WHO guideline of 10 μg/liter in large parts of the Indus plain. The green to brown coloring illustrates the topography. The Indus River and its major tributaries as well as the major cities are indicated. The samples were collected for this study (*n* = 1184) between 2013 and 2015.

generally imply ongoing nitrate reduction but limited iron-reducing aquifer conditions. Outside of the Indus Plain where arsenic is mostly below 10 μg/liter, nitrate concentrations are considerably higher (see fig. S2) and indicative of less reducing aquifer conditions.

Depth was recorded for ~30% of the measurements and ranges from 3 to 70 m (mean, 17 m; median, 12 m). However, there is no statistically significant correlation between depth and arsenic concentration. With regard to the other sampling sites, it is generally assumed from local knowledge that depths of family-based wells are in the range of 9 to 30 m and thus access the uppermost, unconfined aquifer (*37*, *48*). Forty-one other deeper (>120 m) water samples (not included in the modeling) collected from Lahore municipal water supply tube wells ranged in arsenic concentration from <1 to 85 μg/liter (average, 23 μg/liter; median, 25 μg/liter), despite the use of arsenic filters with these water sources.

Hazard probability model

In total, 232 of the 1000 logistic regression runs did not pass the Hosmer-Lemeshow goodness-of-fit test (*52*), leaving 768 runs from which a single set of coefficients was calculated. Of the nine variables used in the logistic regressions, the following were retained by stepwise selection: fluvisols, Holocene fluvial sediments, slope, soil organic carbon, and soil pH. The average weighted coefficients based on normalized variables are listed in table S2. Steeper slopes (>0.1°) and soil organic carbon were negative indicators (inverse correlation) of high arsenic levels in groundwater, whereas fluvisols, soil pH, and Holocene fluvial sediments were positive indicators (positive correlation). With the exception of soil organic carbon, the number of variables that passed the goodness-of-fit test is nearly equal.

The plots in Fig. 2 summarize the classification strength of the model at different cutoff values, which helps in finding the best cutoff to optimize the rates of correct positive and negative classification. Figure 2A shows the receiver operating characteristic (ROC) curve (*53*) using the entire data set of 743 aggregated points (original measurements averaged into 1-km × 1-km pixels). The area under the ROC curve (AUC) (possible values between 0.5 and 1) is 0.80, which shows that the logistic regression does a good job of correctly classifying high and low arsenic concentrations (The ROC and AUC indicate a model's classification rates of high and low values at different probability cutoff values). Figure 2B plots accuracy, sensitivity (true-positive rate), and specificity (true-negative rate) against cutoff, the latter two being equal at a cutoff of 0.60. The modeled arsenic probability map is shown in Fig. 3A, with the binary data points of measured arsenic concentrations used in the analysis being plotted. Table 1 compares the AUC and Akaike information criterion (AIC) (see Materials and Methods) (*54*) of the final model with other logistic regression analyses using manually selected, fixed predictor variables.

Health risk model

Combining the area of high hazard (≥60% probability) in Fig. 3A with population figures for 2016 indicates that approximately 88 million people live within the modeled hazard area. The population figures were calculated using census data from 2010 (*55*) multiplied by subsequent annual population growth rates (*56*). Using the assumption that about 60 to 70% of the population throughout Pakistan relies on groundwater for its drinking water (*57*, *58*), this number reduces to ~50 million to 60 million people potentially affected (Fig. 3B). Furthermore, arsenic remediation is minimal in rural areas (*20*, *30*), and our study showed that remediation in the municipal water supply system of the urban center of Lahore is only moderately effective in meeting the local arsenic concentration standard (57% of samples below 50 μg/liter) and does not attain the WHO standard (1% of samples below 10 μg/liter).

The above analysis was also carried out using a threshold of 50 μg/liter (fig. S3B and table S3), which is Pakistan's official health guideline for arsenic in drinking water. Although the probabilities of exceeding this guideline are lower, the cutoff where sensitivity and specificity are equal is also lower (fig. S5). Since less than 10% of the measured data points have concentrations between 10 and 50 μg/liter, the physical area encompassing this range of concentrations is small and the number of people potentially affected by arsenic-contaminated water ultimately remains approximately 50 million to 60 million (fig. S6).

DISCUSSION

pH-induced arsenic release and accumulation in unconfined aquifers

Analysis of data points with As > 10 μg/liter suggests that reducing conditions are generally not dominant in the Indus Plain, with nitrate averaging 2.7 mg/liter and ranging up to about 28 mg/liter (figs. S1 and S2 and table S1). The average dissolved concentration of iron (Fig. 4A

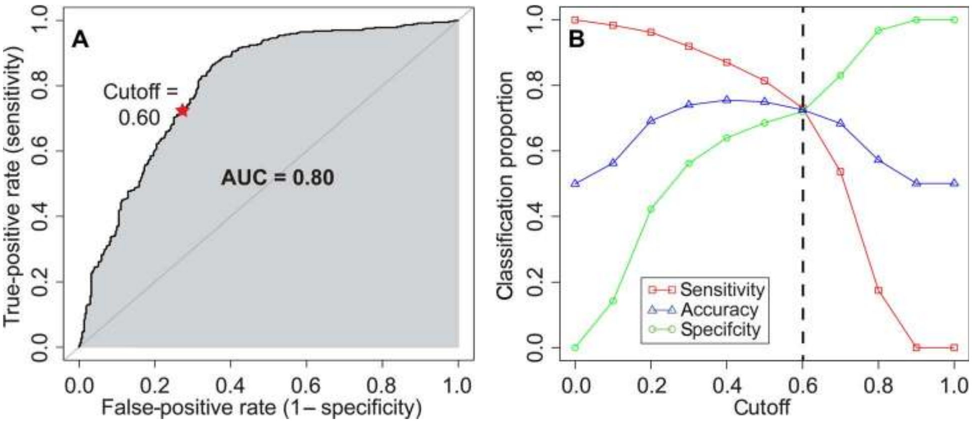

Fig. 2. Statistics of the classification strength of the logistic regression analysis results using the threshold of 10 μg/liter (WHO As guideline) applied to the entire set of 743 aggregated data points. (**A**) ROC curve with an AUC of 0.80, which indicates the discriminative power of the logistic equation. (**B**) Sensitivity (true-positive rate), accuracy, and specificity (true-negative rate) versus cutoff value.

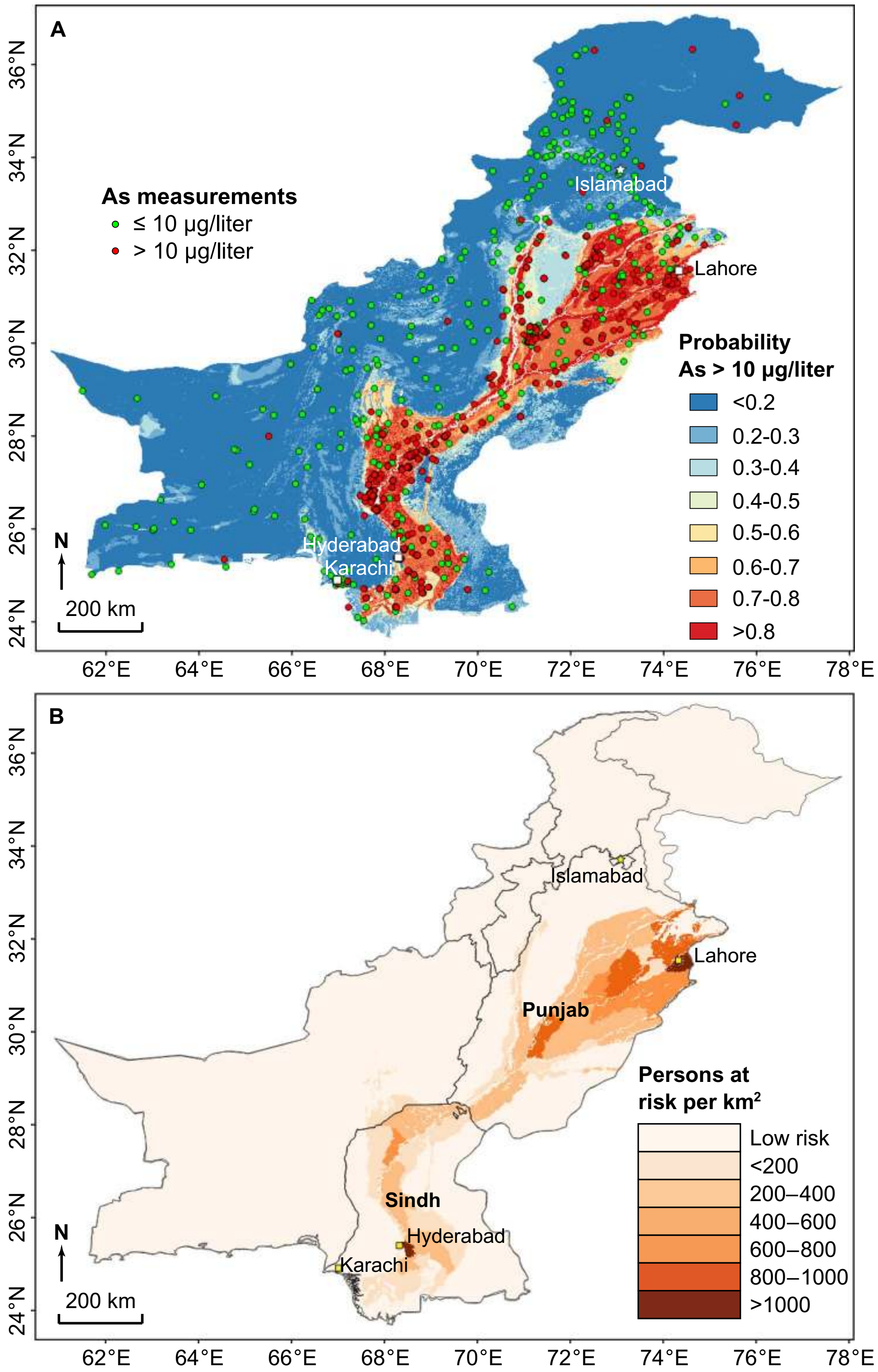

Fig. 3. Arsenic prediction and risk models. (**A**) Probability (hazard) map of the occurrence of arsenic concentrations in groundwater exceeding the WHO As guideline of 10 μg/liter along with the aggregated arsenic data points used in modeling (*n* = 743) (see fig. S3B for the hazard map using 50 μg/liter). (**B**) Density of population at risk of high levels of arsenic in groundwater using the WHO As guideline of 10 μg/liter. Figure was based on 2016 population figures and a 60 to 70% groundwater utilization rate (see text). The estimated number of people potentially affected is ~50 million to 60 million, with hot spots around Lahore and Hyderabad.

and table S1) in high-arsenic samples is low at 0.09 mg/liter, and high arsenic concentrations are negatively correlated ($R = -0.778$; Table 2) with soil organic carbon (Fig. 4B). However, some areas do locally show a high correlation between high arsenic and iron, suggesting a reducing environment.

Table 1. Comparison of AUC and AIC results of logistic regression analyses of models with fixed variables (models A to E) along with the final model (model F) shown in Fig. 3A, which was achieved by stepwise variable selection. A higher AUC shows better model prediction performance, whereas a lower AIC is indicative of a simpler, more effective model. The associated hazard maps are provided in fig. S4.

Model	Predictor variable(s)	AUC	AIC
A	Fluvisols (probability)	0.69	787 ± 5
B	Irrigated area	0.72	739 ± 7
C	Aridity, slope (binary, 0.1°)	0.73	716 ± 8
D	Slope (binary, 0.1°), soil pH	0.77	688 ± 9
E	Aridity, Holocene fluvial sediments (binary), slope (binary, 0.1°)	0.77	664 ± 8
F	Fluvisols (probability), Holocene fluvial sediments (binary), soil organic carbon, soil pH, slope (binary, 0.1°)	0.80	644 ± 9

The percentage of high values of arsenic (>10 μg/liter) correlates very strongly with soil pH ($R = 0.977$; Fig. 4C), which is consistently elevated throughout the Indus Plain (Fig. 4D). This implies pH-induced desorption and corroborates the findings of Farooqi *et al.* (*51*) for a site near Lahore, Punjab. Although soil pH throughout the Indus Plain is generally between 8.0 and 8.5, the average pH of groundwater samples with arsenic concentrations greater than 10 μg/liter is only 7.6, with no significant correlation. This indicates that arsenic release due to high-pH desorption may be occurring predominantly in the uppermost sediments before being transported downward via infiltration to the aquifer. Depth is not correlated with arsenic concentration over the range of available depth measurements (3 to 70 m) of the unfiltered water samples used in modeling, and, as previously mentioned, deeper (>120 m) municipal tube wells from Lahore, which are filtered for arsenic, also exhibit elevated arsenic concentrations.

Aridity (precipitation/PET) is also well correlated with high arsenic values ($R = -0.779$; Table 2), which is consistent with the process of evaporative concentration as suggested by Rasool *et al.* (*59*) for Mailsi, Punjab and by Brahman *et al.* (*40*) for the Tharparkar District, Sindh. The latter study also found a predominance of the As(V) species relative to the reduced As(III) species. The very strong correlation of irrigated area with arsenic contamination ($R = 0.967$; Table 2) could be a consequence of the role of irrigation in evaporative concentration and/or soil alkalization with associated arsenic desorption. Although these finding are consistent with arsenic release caused by oxidizing and/or elevated-pH dissolution, the process of reductive dissolution may be responsible locally, particularly as a result of industrial or human/animal organic waste in urban areas (*37*) or intensive agricultural activity.

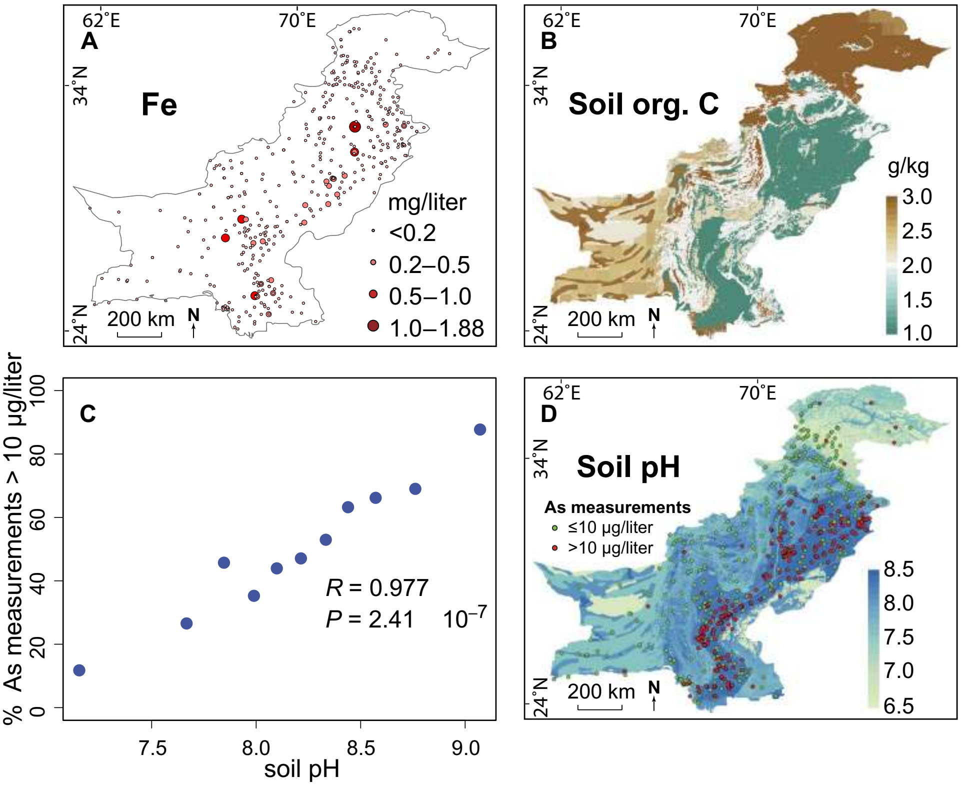

Fig. 4. Indicators of geochemical environment. (**A**) Measured iron concentrations ($n = 458$). (**B**) Soil organic carbon (*70*) used as a predictor variable. (**C**) Correlation of groundwater samples exceeding the WHO As guideline of 10 μg/liter against soil pH (in 11 bins; see Materials and Methods). (**D**) Soil pH (*70*) used as a predictor variable, shown with the arsenic measurements above and below 10 μg/liter.

Table 2. Data sets evaluated for use as predictor variables in the logistic regression analysis. Correlations were found using the percentage of measurements exceeding 10 μg/liter in 11 bins of equal member-size across the range of each variable (see Materials and Methods). *P* values of logistic regression are based on univariate analyses. An asterisk indicates data sets that were not significant and therefore removed from the full logistic regression analysis. n/a, not applicable.

Data set	Resolution	Correlation (*P*)	Logistic regression (*P*)
Potential evapotranspiration (PET) (*72*, *73*)	30″	0.730 (<0.05)	<0.05
Precipitation (*74*)	30″	−0.776 (<0.05)	<0.05
Aridity [precipitation (*74*)/ PET (*72*, *73*)]	30″	−0.779 (<0.05)	<0.05
Irrigated area % (*75*)	5′	0.967 (<0.05)	<0.05
Slope (binary, 0.1°) (*76*)	30″	n/a	<0.05
Fluvisol probability (%) (*70*, *77*, *78*)	30″	0.704 (<0.05)	<0.05
Soil organic carbon (*70*, *77*, *78*)	30″	−0.778 (<0.05)	<0.05
Soil pH (*70*, *77*, *78*)	30″	0.977 (<0.05)	<0.05
Soil clay % (*70*)*	30″	−0.338 (>0.05)	>0.05
Soil silt % (*70*)*	30″	-7.22×10^{-2} (>0.05)	>0.05
Holocene fluvial sediments (binary) (*69*)	Polygon	n/a	<0.05

*Data sets that were not significant and therefore removed from the full logistic regression analysis.

Model proxies

The relationships between the predictor variables and the process of arsenic enrichment in groundwater are generally straightforward. A flat surface topography, such as in the Indus Plain, implies a low hydraulic gradient resulting in sluggish groundwater flow, thereby allowing arsenic concentrations to gradually increase in the groundwater. High soil pH can drive arsenic desorption and is indicative of an evaporative environment, which further raises arsenic concentrations.

The data set of predicted fluvisols (young alluvial sediments) indicates an environment similar to that of Holocene sediments but specifically indicates an alluvial setting. As such, it could be an effective predictor variable where geological maps do not specify the Holocene epoch. High levels of soil organic carbon are a common driver of reducing conditions resulting in arsenic release (*5*). The inverse relationship in our model appears to be a consequence of the presence of an arid climate resulting in minimal natural growth of vegetation and accumulation of organic carbon in the soil of the Indus Plain, which leaves the other factors mentioned to be responsible for the enrichment of arsenic. This further implies that reducing conditions are not predominant in the Indus Plain.

By design, the stepwise method of variable selection produces the model with the best combination of model simplicity and likelihood, as measured by the AIC. After our initial identification of potential predictor variables (see Materials and Methods), the stepwise method was allowed to automatically identify the final set of predictors. To test the efficacy of this approach, we ran other logistic analyses using fixed combinations of variables that would be expected to effectively predict elevated arsenic concentrations based on their known relationships or associations with arsenic enrichment in groundwater. These results (Table 1) confirm that the stepwise method of variable selection in this case produces a statistically superior model to that which may be achieved by manually selecting variables based only on expert knowledge. In any case, the hazard maps produced using these fixed selections of variables are largely similar in that they place the highest probability of high arsenic enrichment in the Indus Plain (fig. S4).

These results highlight the fact that the variables used generally have their highest absolute values, and therefore greatest effect on resultant models, in the Indus Plain. However, independence and causality are not clear, particularly with regard to irrigation. For example, the strong positive correlation of high arsenic concentrations and irrigated area (*R* = 0.967; Table 2), as well as the relatively good performance of the logistic regression using only irrigated area (Table 1, model B), could simply be a coincidence of extensive irrigation being carried out on flat fluvial sediments, where there is a regular supply of water and nutrients from the rivers, as well as a high demand for irrigation due to very arid conditions. However, it appears reasonable that the increased evapotranspiration from irrigation water as well as its slow infiltration through the alkaline topsoil and young alluvial sediments may contribute to an increase in groundwater arsenic concentration.

Implications of hazard and risk models

The risk map of Fig. 3B showcases the severity of the problem of groundwater arsenic contamination in Pakistan. Regardless of using a model based on a guideline of 10 or 50 μg/liter (fig. S3), the area of high probability of arsenic contamination encompasses most of the Indus Plain in the province of Sindh and around the Indus tributaries in the province of Punjab (Jhelum, Chenab, Ravi, and Sutlej rivers), where population densities are particularly high. Although the estimate of 60 to 70% groundwater utilization is only a rough estimate based on available published figures and subject to potentially large regional variations, it remains clear that a great number of people are potentially exposed to dangerously high levels of arsenic. Our estimate of ~50 million to 60 million people potentially affected in the Indus Plain is as high as that for the Ganges-Brahmaputra Delta in Bangladesh and India (~50 million) (*3*, *42*, *60*) and exceeds that for China (~20 million) (*24*) and the Red River Delta in Vietnam (~7 million) (*25*). Furthermore, our estimate is about four times higher than that of Rabbani *et al.* (*36*) for the Indus River, which did not consider its tributaries in densely populated Punjab province and was made on the basis of distance from the river in a ~120-km × 40-km area in Sindh province. In addition, there are pockets of high measured arsenic concentrations within the Indus Plain that fall just below the statistically determined 60% probability cutoff used to calculate high-risk areas, which, if anything, would make our estimate somewhat conservative and only further highlights the severity of the problem.

First and foremost, the risk map indicates the need for widespread testing of drinking water wells in the Indus Plain to help safeguard the long-term health of its population. Because of an inherent high degree of small-scale spatial variability of geogenic arsenic contamination (*61*–*63*), wells should be tested individually so that measures can be implemented for those most severely affected.

Mitigation requires action at several levels, including awareness raising, emergency solutions, coordination of government and financial support, health intervention programs, alternative resources of drinking water (for example, deep wells) (*41*, *64*), and arsenic removal options (*41*, *65*, *66*). Ultimately, any treatment options must be socially acceptable and tailored to the local groundwater composition (*67*).

MATERIALS AND METHODS

Groundwater sampling

Between 2013 and 2015, groundwater samples were collected from 1184 sites throughout the country, mainly from hand and motor pumps as well as municipal and agricultural water supply tube wells. The site selection of sampling was designed to be evenly spatially distributed and was based on individual union council/tehsil (sub-administrative governmental units), for which topographical district maps obtained from the Survey of Pakistan were used.

To obtain representative samples of groundwater, we purged hand pumps with one stroke for every 30 cm of depth and ran electric pumps for 10 min before sampling. Groundwater samples were then immediately filtered on-site using 0.45-mm cellulose acetate filters contained in Millipore Sterivex syringe capsules. During fieldwork, representative samples were carefully handled to help ensure high quality of subsequent analyses and reduce cross contamination during sampling. Precleaned 1-liter polyethylene bottles were used for sample collection. All bottles were first rinsed with deionized water before sampling, and two aliquots of each sample (acidified and non-acidified) were taken on site. For acidified samples used for the analysis of arsenic (As) and other trace metals, a few drops of concentrated nitric acid (HNO_3) were added to reduce the pH of water samples to <2. The basic water quality parameters of pH, electrical conductivity (EC), bicarbonate (HCO_3^-), TDS, sulfate (SO_4^-), nitrate (NO_3^-), chloride (Cl^-), and fluoride (F^-), as well as the elements calcium (Ca), magnesium (Mg), and iron (Fe) were also measured (figs. S1 and S2). pH, temperature, dissolved oxygen (DO), and EC were measured at the time of sample collection, as were the geographical coordinates, using GPS (Global Positioning System). pH, EC, DO, and TDS of all water samples were measured using a W2015 pH/EC meter and a DO meter (Sinowell). The sealed samples were stored at 4°C in portable coolers before transportation to the laboratory for further analysis.

Groundwater samples were analyzed for calcium (Ca^{2+}) and magnesium (Mg^{2+}) by volumetric titration with EDTA (0.05 mol/liter) and alkalinity (measured as HCO_3) by volumetric titration with 0.1 HCl. NO_3^- was determined spectrophotometrically using an ultraviolet-visible (UV-Vis) spectrophotometer (Shimadzu, model UV 1601) at a wavelength of 220 nm. SO_4^{2-} values were determined by gravimetric analysis as $BaSO_4$. Chloride (Cl^-) was determined by titration (American Public Health Association 1998). Fluoride was measured using an ion chromatograph (ICS-3000, Thermo Fisher).

Arsenic and other elements were analyzed in the acidified samples with an Agilent 7500cx inductively coupled plasma mass spectrometer. Calibration solutions were prepared using multi-element stock solutions of 100 ppm (parts per million). The operating conditions were as follows: radio frequency power at 1510 W, carrier gas at 1.1 liter/min, makeup gas at 0.10 liter/min, helium gas flow at 3.5 ml/min, and nebulizer pump at 0.1 rps. The standard stock solutions mixed with elements (100 μg/ml; GSB 04-1767-2004) were obtained from the National Center of Analysis and Testing for Nonferrous Metals and Electronic Materials.

A quality control (QC) sample was prepared and injected after every 15th sample to check instrument stability. The QC sample was prepared by mixing aliquots of each sample and was therefore representative of the entire sample set. There was less than 15% variation in the metal concentrations of the QC samples. Spiked samples were also prepared in the same manner as the water samples. Some samples were spiked with As and other elements before analysis at the final two levels of 10 and 20 ng/ml. Moreover, to confirm analytical performance and adequate precision, we used standard reference solutions of analytical-grade chemicals with 99.9% spectroscopic purity (Merck). Twice-distilled water was used throughout the analyses. By means of blank and duplicate samples, reproducibility of the analytical data was found to be within 5% and the analytical error was estimated to be <10%. Working solutions were prepared on a daily basis by appropriate dilution of a standard stock solution with a mixture of 65% HNO_3 and H_2O [1:3 (v/v)]. The order in which all samples were run was randomized so as to reduce the error from any injection of artifacts or changes in instrument sensitivity during the analysis sequence.

Statistical modeling

Because it is of primary importance whether or not the arsenic concentration in groundwater poses a health hazard, we used logistic regression analysis to model arsenic concentration in groundwater being above or below the WHO guideline of 10 and 50 μg/liter. We used our own groundwater quality data ($n = 1184$) as well as those from the previously mentioned studies (*20*, *30*, *37*–*41*, *51*, *59*), which constitute 69 additional samples in total.

The data were first aggregated into 1-km squares by taking the geometric mean of measurements falling within each square, which resulted in 743 data points. The 1-km × 1-km dimensions correspond to the finest resolution (30″) of data set used in the logistic regression analysis (Table 2). Because the measured arsenic concentrations in groundwater were used as the dependent variable in the logistic regressions, they were assigned the value of 1 if the concentration is greater than the WHO guideline of 10 μg/liter (49.8% of data points) or 0 if the concentration is less than or equal to 10 μg/liter (50.2% of data points).

Logistic regression uses a logistic function, which takes on independent variables that can range between negative infinity and positive infinity and produces an outcome between 0 and 1 (*52*)

$$P = \frac{1}{1 + e^{-t}} \tag{1}$$

where $0 \le P \le 1$, and

$$t = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n \tag{2}$$

where $x_1 \ldots x_n$ are the independent variables and $\beta_1 \ldots \beta_n$ are the corresponding regression coefficients. P is interpreted as the probability of the dependent variable being 0 or 1. Logistic regression analyses were conducted using the generalized linear model function of R (*68*).

To aid comparison of different analyses, the AIC (*54*) provides a relative comparison in terms of the trade-off between complexity and goodness of fit

$$\mathrm{AIC} = 2k - 2\ln(L) \tag{3}$$

where k is the number of parameters and L is the likelihood. For a given suite of analyses, the one with the lowest AIC provides the best combination of modeling performance and simplicity.

Variable selection

Eleven predictor variables were considered for use in the logistic regressions (Table 2). These included PET, precipitation, aridity (precipitation/PET), temperature, irrigated land area, surface slope, the probability of fluvisols, soil organic carbon, soil pH, clay content, silt content, and Holocene fluvial sediments. All variables were continuous except for slope (binary above or below 0.1°) and the existence of Holocene fluvial sediments. The latter was taken from the Geological Map of Pakistan (*69*) and divided into two sections: (i) Holocene streambed, floodplain, and fluvial terrace sediments and (ii) all other Quaternary and older units (fig. S7E). This differentiation was based on the prevalence of elevated arsenic concentrations in Holocene fluvial sediments (*1*, *23*, *24*, *45*). The other variables were chosen for their direct or indirect relationship to the process of arsenic enrichment in groundwater. For example, surface topography may be related to the groundwater flow rate, and the soil parameters, estimated at 1.5-m depth (*70*), are assumed to have an effect at least on shallow aquifers (*51*).

To help identify effective predictor variables for logistic regression, we calculated linear (Pearson) correlations between each of the continuous variables and the occurrence of high arsenic concentrations (Table 2 and fig. S8). These were made using the percentage of arsenic measurements greater than 10 μg/liter for 11 bins across the range of each independent variable. The number of bins was determined using Sturges' formula (*71*), and the bin width was varied to have each bin contain the same number of members. The data were consolidated as such to focus on the fundamental criterion of whether or not the concentration of arsenic poses a health hazard. In addition, a univariate logistic regression was run with each variable, and the significance of the coefficient was assessed through its P value (Table 2). On the basis of the evaluation of P values from both analyses at the 95% confidence level ($P = 0.05$), clay content and silt content were removed from further consideration.

Logistic regression analysis

The data set was randomly divided into subsets of 80% for training and 20% for testing. Logistic regression analysis was then run on the training subset using a stepwise selection of variables (both directions), which removes or adds variables based on their improvement to the AIC. The result was then applied to the testing subset, and the Hosmer-Lemeshow goodness-of-fit test (*52*) was used to indicate the accuracy of predictions against the testing subset at the 95% confidence level. To avoid introducing bias due to the particular selection of training and testing subsets, we repeated this process 1000 times. Analyses were retained if they passed the Hosmer-Lemeshow goodness-of-fit test, and these results were used as a weighting in averaging the variable coefficients. If an analysis did not contain a certain variable, a zero was used instead. This technique provided stability to the final model and emphasized parameters that more strongly predict the occurrence of high arsenic concentrations.

The performance of the final set of averaged coefficients was evaluated on the entire data set by plotting sensitivity (rate of true positives, or success rate of predicting high concentrations) against specificity (rate of true negatives, or success rate of predicting low concentrations) for all possible cutoff levels in an ROC curve (*53*). The AUC was then calculated to evaluate the performance of the logistic regression (Fig. 2A). The AUC indicates the ability of a model to correctly classify positive and negative cases and generally ranges from 0.5 to 1, where 0.5 corresponds to an analysis that predicts the data no better than would a random selection and 1 corresponds to an analysis that perfectly predicts the observations.

For comparison, other logistic regression analyses (using a threshold of 10 μg/liter) were run using fixed combinations of variables that would be expected to effectively predict elevated arsenic accumulation. These were also run 1000 times, and the coefficients of the analyses passing the Hosmer-Lemeshow goodness-of-fit test were combined in a weighted average using the test results.

Generation of hazard and risk models

The coefficients of the final logistic regression were used to generate a hazard probability map of groundwater arsenic concentrations exceeding either 10 or 50 μg/liter. The cutoff to distinguish between areas of high and low hazard was selected where the sensitivity and specificity of the analysis are equal (Fig. 2B).

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/3/8/e1700935/DC1

fig. S1. Maps of spatial distribution and values of all chemical parameters.
fig. S2. Grids of all measured chemical parameters.
fig. S3. Hazard maps of logistic regression models using thresholds of 10 and 50 μg/liter.
fig. S4. Maps of other well-fitting logistic regression models.
fig. S5. Statistics of logistic regression model using threshold of 50 μg/liter.
fig. S6. Density of population at risk using logistic regression models with thresholds of 10 and 50 μg/liter.
fig. S7. Predictor data sets used in the final model.
fig. S8. Graphs of correlation between predictor variables and percentage of As measurements >10 μg/liter.
table S1. Summary statistics of all measured parameters.
table S2. Coefficients, SDs, and frequencies of predictor variables in logistic regressions with a threshold of 10 μg/liter.
table S3. Coefficients, SDs, and frequencies of predictor variables in logistic regressions with a threshold of 50 μg/liter.

REFERENCES AND NOTES

1. P. L. Smedley, D. G. Kinniburgh, A review of the source, behaviour and distribution of arsenic in natural waters. *Appl. Geochem.* **17**, 517–568 (2002).
2. 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).
3. R. Nickson, J. McArthur, W. Burgess, K. Matin Ahmed, P. Ravenscroft, M. Rahman, Arsenic poisoning of Bangladesh groundwater. *Nature* **395**, 338 (1998).
4. J. M. McArthur, D. M. Banerjee, K. A. Hudson-Edwards, R. Mishra, R. Purohit, P. Ravenscroft, A. Cronin, R. J. Howarth, A. Chatterjee, T. Talukder, D. Lowry, S. Houghton, D. K. Chadha, Natural organic matter in sedimentary basins and its relation to arsenic in anoxic ground water: The example of West Bengal and its worldwide implications. *Appl. Geochem.* **19**, 1255–1293 (2004).
5. P. Ravenscroft, H. Brammer, K. Richards, *Arsenic Pollution: A Global Synthesis* (John Wiley & Sons, 2009), vol. 28.
6. R. T. Nickson, J. M. McArthur, P. Ravenscroft, W. G. Burgess, K. M. Ahmed, Mechanism of arsenic release to groundwater, Bangladesh and West Bengal. *Appl. Geochem.* **15**, 403–413 (2000).
7. C. B. Dowling, R. J. Poreda, A. R. Basu, S. L. Peters, P. K. Aggarwal, Geochemical study of arsenic release mechanisms in the Bengal Basin groundwater. *Water Resour. Res.* **38**, 12-1–12-18 (2002).
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).
9. L. M. Del Razo, M. A. Arellano, M. E. Cebrián, The oxidation states of arsenic in well-water from a chronic arsenicism area of northern Mexico. *Environ. Pollut.* **64**, 143–153 (1990).
10. J. D. Ayotte, D. L. Montgomery, S. M. Flanagan, K. W. Robinson, Arsenic in groundwater in eastern New England: Occurrence, controls, and human health implications. *Environ. Sci. Technol.* **37**, 2075–2083 (2003).
11. A. H. Welch, D. Westjohn, D. R. Helsel, R. B. Wanty, Arsenic in ground water of the United States: Occurrence and geochemistry. *Ground Water* **38**, 589–604 (2000).

12. J. G. Webster, D. K. Nordstrom, in *Arsenic in Groundwater: Geochemistry and Occurrence*, A. H. Welch, K. G. Stollenwerk, Eds. (Springer-Verlag, 2003), pp. 101–126.
13. D. Langmuir, *Aqueous Environmental Geochemistry* (Prentice-Hall, 1997).
14. M. E. Schreiber, J. A. Simo, P. G. Freiberg, Stratigraphic and geochemical controls on naturally occurring arsenic in groundwater, eastern Wisconsin, USA. *Hydrogeol. J.* **8**, 161–176 (2000).
15. J. M. McArthur, P. Ravenscroft, S. Safiulla, M. F. Thirlwall, Arsenic in groundwater: Testing pollution mechanisms for sedimentary aquifers in Bangladesh. *Water Resour. Res.* **37**, 109–117 (2001).
16. D. Postma, F. Larsen, N. T. Thai, P. T. K. Trang, R. Jakobsen, P. Q. Nhan, T. V. Long, P. H. Viet, A. S. Murray, Groundwater arsenic concentrations in Vietnam controlled by sediment age. *Nat. Geosci.* **5**, 656–661 (2012).
17. S. Gao, J. Ryu, K. K. Tanji, M. J. Herbel, Arsenic speciation and accumulation in evapoconcentrating waters of agricultural evaporation basins. *Chemosphere* **67**, 862–871 (2007).
18. C. O. Abernathy, Y.-P. Liu, D. Longfellow, H. V. Aposhian, B. Beck, B. Fowler, R. Goyer, R. Menzer, T. Rossman, C. Thompson, M. Waalkes, Arsenic: Health effects, mechanisms of actions, and research issues. *Environ. Health Perspect.* **107**, 593–597 (1999).
19. A. H. Milton, Z. Hasan, S. M. Shahidullah, S. Sharmin, M. D. Jakariya, M. Rahman, K. Dear, W. Smith, Association between nutritional status and arsenicosis due to chronic arsenic exposure in Bangladesh. *Int. J. Environ. Health Res.* **14**, 99–108 (2004).
20. A. Azizullah, M. N. K. Khattak, P. Richter, D.-P. Häder, Water pollution in Pakistan and its impact on public health—A review. *Environ. Int.* **37**, 479–497 (2011).
21. J. D. Ayotte, B. T. Nolan, J. R. Nuckols, K. P. Cantor, G. R. Robinson Jr., D. Baris, L. Hayes, M. Karagas, W. Bress, D. T. Silverman, J. H. Lubin, Modeling the probability of arsenic in groundwater in New England as a tool for exposure assessment. *Environ. Sci. Technol.* **40**, 3578–3585 (2006).
22. 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).
23. 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).
24. 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).
25. L. H. E. Winkel, P. T. K. Trang, V. M. Lan, C. Stengel, M. Amini, N. T. Ha, P. H. Viet, 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).
26. World Bank–Strategic Country Environmental Assessment, *Pakistan Strategic Country Environmental Assessment: Main Report* (World Bank–Strategic Country Environmental Assessment, 2006).
27. H. V. Aposhian, M. M. Aposhian, Arsenic toxicology: Five questions. *Chem. Res. Toxicol.* **19**, 1–15 (2006).
28. M. F. Hughes, Biomarkers of exposure: A case study with inorganic arsenic. *Environ. Health Perspect.* **114**, 1790–1796 (2006).
29. Z. Fatmi, I. Azam, F. Ahmed, A. Kazi, A. B. Gill, M. M. Kadir, M. Ahmed, N. Ara, N. Z. Janjua; Core Group for Arsenic Mitigation in Pakistan, Health burden of skin lesions at low arsenic exposure through groundwater in Pakistan. Is river the source? *Environ. Res.* **109**, 575–581 (2009).
30. A. K. Bhowmik, A. Alamdar, I. Katsoyiannis, H. Shen, N. Ali, S. M. Ali, H. Bokhari, R. B. Schäfer, S. A. Eqani, Mapping human health risks from exposure to trace metal contamination of drinking water sources in Pakistan. *Sci. Total Environ.* **538**, 306–316 (2015).
31. T. G. Kazi, M. B. Arain, J. A. Baig, M. K. Jamali, H. I. Afridi, N. Jalbani, R. A. Sarfraz, A. Q. Shah, A. Niaz, The correlation of arsenic levels in drinking water with the biological samples of skin disorders. *Sci. Total Environ.* **407**, 1019–1026 (2009).
32. K. D. Brahman, T. G. Kazi, H. I. Afridi, J. A. Baig, S. S. Arain, F. N. Talpur, A. G. Kazi, J. Ali, A. H. Panhwar, M. B. Arain, Exposure of children to arsenic in drinking water in the Tharparkar region of Sindh, Pakistan. *Sci. Total Environ.* **544**, 653–660 (2016).
33. M. Bibi, M. Z. Hashmi, R. N. Malik, Human exposure to arsenic in groundwater from Lahore district, Pakistan. *Environ. Toxicol. Pharmacol.* **39**, 42–52 (2015).
34. K. D. Brahman, T. G. Kazi, J. A. Baig, H. I. Afridi, A. Khan, S. S. Arain, M. B. Arain, Fluoride and arsenic exposure through water and grain crops in Nagarparkar, Pakistan. *Chemosphere* **100**, 182–189 (2014).
35. M. Arain, T. G. Kazi, J. A. Baig, M. K. Jamali, H. I. Afridi, A. Q. Shah, N. Jalbani, R. A. Sarfraz, Determination of arsenic levels in lake water, sediment, and foodstuff from selected area of Sindh, Pakistan: Estimation of daily dietary intake. *Food Chem. Toxicol.* **47**, 242–248 (2009).
36. U. Rabbani, G. Mahar, A. Siddique, Z. Fatmi, Risk assessment for arsenic-contaminated groundwater along River Indus in Pakistan. *Environ. Geochem. Health* **39**, 179–190 (2017).
37. R. T. Nickson, J. M. McArthur, B. Shrestha, T. O. Kyaw-Myint, D. Lowry, Arsenic and other drinking water quality issues, Muzaffargarh District, Pakistan. *Appl. Geochem.* **20**, 55–68 (2005).
38. A. Farooqi, H. Masuda, N. Firdous, Toxic fluoride and arsenic contaminated groundwater in the Lahore and Kasur districts, Punjab, Pakistan and possible contaminant sources. *Environ. Pollut.* **145**, 839–849 (2007).
39. J. A. Baig, T. G. Kazi, A. Q. Shah, G. A. Kandhro, H. I. Afridi, M. B. Arain, M. K. Jamali, N. Jalbani, Speciation and evaluation of Arsenic in surface water and groundwater samples: A multivariate case study. *Ecotoxicol. Environ. Saf.* **73**, 914–923 (2010).
40. K. D. Brahman, T. G. Kazi, H. I. Afridi, S. Naseem, S. S. Arain, N. Ullah, Evaluation of high levels of fluoride, arsenic species and other physicochemical parameters in underground water of two sub districts of Tharparkar, Pakistan: A multivariate study. *Water Res.* **47**, 1005–1020 (2013).
41. T. Ahmad, M. A. Kahlown, A. Tahir, H. Rashid, Arsenic an emerging issue: Experiences from Pakistan, in *Proceedings of the 30th WEDC International Conference* (2004), pp. 459–466.
42. U. K. Chowdhury, B. K. Biswas, T. R. Chowdhury, G. Samanta, B. K. Mandal, G. C. Basu, C. R. Chanda, D. Lodh, K. C. Saha, S. K. Mukherjee, S. Roy, S. Kabir, Q. Quamruzzaman, D. Chakraborti, Groundwater arsenic contamination in Bangladesh and West Bengal, India. *Environ. Health Perspect.* **108**, 393 (2000).
43. L. Charlet, D. A. Polya, Arsenic in shallow, reducing groundwaters in southern Asia: An environmental health disaster. *Elements* **2**, 91–96 (2006).
44. H. Guo, D. Wen, Z. Liu, Y. Jia, Q. Guo, A review of high arsenic groundwater in Mainland and Taiwan, China: Distribution, characteristics and geochemical processes. *Appl. Geochem.* **41**, 196–217 (2014).
45. 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).
46. J. Buschmann, M. Berg, C. Stengel, L. Winkel, M. L. Sampson, P. T. 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).
47. A. S. Qureshi, P. G. McCornick, A. Sarwar, B. R. Sharma, Challenges and prospects of sustainable groundwater management in the Indus Basin, Pakistan. *Water Resour. Manag.* **24**, 1551–1569 (2010).
48. D. W. Greenman, W. V. Swarzenski, G. D. Bennett, *Ground-Water Hydrology of the Punjab, West Pakistan, with Emphasis on Problems Caused by Canal Irrigation* (Government Printing Office, 1967).
49. P. Smedley, *Groundwater Quality: Pakistan* (British Geological Survey, 2001), 6 pp.
50. F. Bender, H. Raza, D. Bannert, Geology of Pakistan: Gebruder Borntraeger (1995).
51. A. Farooqi, H. Masuda, R. Siddiqui, M. Naseem, Sources of arsenic and fluoride in highly contaminated soils causing groundwater contamination in Punjab, Pakistan. *Arch. Environ. Contam. Toxicol.* **56**, 693–706 (2009).
52. D. W. Hosmer Jr., S. Lemeshow, *Applied Logistic Regression* (John Wiley & Sons, ed. 2, 2004).
53. T. Fawcett, An introduction to ROC analysis. *Pattern Recognit. Lett.* **27**, 861–874 (2006).
54. H. Akaike, A new look at the statistical model identification. *IEEE Trans. Autom. Control.* **19**, 716–723 (1974).
55. Pakistan Bureau of Statistics, *Population Census* (Government of Pakistan, 2010).
56. World Bank, Population growth (annual %) (World Bank Open Data, 2016); http://data.worldbank.org/indicator/SP.POP.GROW?locations=PK [accessed 12 October 2016].
57. B. A. Chandio, M. Abdullah, M. A. Tahir, in *Proceedings of the National Workshop on Quality of Drinking Water* (Pakistan Council of Research in Water Resources, 1998), pp. 14–18.
58. N. Iqbal, F. Hossain, H. Lee, G. Akhter, Satellite gravimetric estimation of groundwater storage variations over Indus Basin in Pakistan. *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.* **9**, 3524–3534 (2016).
59. A. Rasool, A. Farooqi, S. Masood, K. Hussain, Arsenic in groundwater and its health risk assessment in drinking water of Mailsi, Punjab, Pakistan. *Hum. Ecol. Risk Assess. Int. J.* **22**, 187–202 (2016).
60. 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).
61. *Arsenic Contamination of Groundwater in Bangladesh*, D. G. Kinniburgh, P. L. Smedley, Eds. (British Geological Survey, 2001).
62. 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).
63. 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).

64. M. F. Ahmed, S. Ahuja, M. Alauddin, S. J. Hug, J. R. Lloyd, A. Pfaff, T. Pichler, C. Saltikov, M. Stute, A. van Geen, Ensuring safe drinking water in Bangladesh. *Science* **314**, 1687–1688 (2006).
65. C. K. Jain, R. D. Singh, Technological options for the removal of arsenic with special reference to South East Asia. *J. Environ. Manage.* **107**, 1–18 (2012).
66. J. G. Hering, I. A. Katsoyiannis, G. A. Theoduloz, M. Berg, S. J. Hug, Arsenic removal from drinking water: Experiences with technologies and constraints in practice. *J. Environ. Eng.* **143**, 03117002 (2017).
67. S. J. Hug, O. X. Leupin, M. Berg, Bangladesh and Vietnam: Different groundwater compositions require different approaches to arsenic mitigation. *Environ. Sci. Technol.* **42**, 6318–6323 (2008).
68. R Development Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2014).
69. M. Abu, M. S. Bakr, R. O. Jackson, *Geological Map of Pakistan, Scale: 1:2,000,000* (Geological Survey of Pakistan, 1964).
70. International Soil Reference and Information Centre, *World Soil Information* (International Soil Reference and Information Centre, 2013).
71. H. A. Sturges, The choice of a class interval. *J. Am. Stat. Assoc.* **21**, 65–66 (1926).
72. R. J. Zomer, D. A. Bossio, A. Trabucco, L. Yuanjie, D. C. Gupta, V. P. Singh, *Trees and Water: Smallholder Agroforestry on Irrigated Lands in Northern India* (Research Report 122, Internation Water Management Institute, 2007).
73. R. J. Zomer, A. Trabucco, D. A. Bossio, L. V. Verchot, Climate change mitigation: A spatial analysis of global land suitability for clean development mechanism afforestation and reforestation. *Agr. Ecosyst. Environ.* **126**, 67–80 (2008).
74. 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).
75. Food and Agriculture Organization (FAO), *FAOSTAT Database Collections* (FAO of the United Nations, 2013).
76. U.S. Geological Survey, *30sec GRID: Conditioned DEM* (U.S. Geological Survey, 2013).
77. Food and Agriculture Organization of the United Nations and the International Institute for Applied Systems Analysis, *Harmonized World Soil Database* (Food and Agriculture Organization of the United Nations and the International Institute for Applied Systems Analysis, 2012).
78. International Geosphere-Biosphere Programme–Data Information System, *A Program for Creating Global Soil-Property Databases* (International Geosphere-Biosphere Programme–Data Information System, 1998).

Acknowledgments: We acknowledge our colleague A. Johnson who initiated this collaboration but passed away in 2015. **Funding:** We thank the Swiss Agency for Development and Cooperation (7F-09010.01), the Swiss Federal Institute of Aquatic Science and Technology (Eawag), the Higher Education Commission of Pakistan (project no. PM-IPFP/HRD/HEC/2012/3599), the National Natural Science Foundation of China (Research Fellowship for International Young Scientists, no. 21450110419), and the Institute of Urban Environment of the Chinese Academy of Sciences (CAS President's International Fellowship for Postdoctoral Researchers, no. 2016PB009) for their support in carrying out this project. **Author contributions:** S.A.M.A.S.E. and H.S. designed the groundwater sampling campaigns. S.A.M.A.S.E., T.K., and R.U. conducted the groundwater sampling and performed the laboratory analyses. J.E.P. designed and carried out the statistical analyses. J.E.P. and M.B. formulated the geochemical interpretations and designed the figures. J.E.P. wrote the paper with contributions from M.B., S.A.M.A.S.E., and H.S. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Additional data related to this paper may be requested from the authors.

Submitted 20 March 2017
Accepted 28 July 2017
Published 23 August 2017
10.1126/sciadv.1700935

Citation: 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).

ScienceAdvances

Extensive arsenic contamination in high-pH unconfined aquifers in the Indus Valley

Joel E. Podgorski, Syed Ali Musstjab Akber Shah Eqani, Tasawar Khanam, Rizwan Ullah, Heqing Shen and Michael Berg

Sci Adv 3 (8), e1700935.
DOI: 10.1126/sciadv.1700935

Science Advances (ISSN 2375-2548) is published by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. 2017 © The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government Works. The title *Science Advances* is a registered trademark of AAAS.

advances.sciencemag.org/cgi/content/full/3/8/e1700935/DC1

Supplementary Materials for

Extensive arsenic contamination in high-pH unconfined aquifers in the Indus Valley

Joel E. Podgorski, Syed Ali Musstjab Akber Shah Eqani, Tasawar Khanam, Rizwan Ullah, Heqing Shen, Michael Berg

Published 23 August 2017, *Sci. Adv.* **3**, e1700935 (2017)
DOI: 10.1126/sciadv.1700935

This PDF file includes:

Supplementary Materials

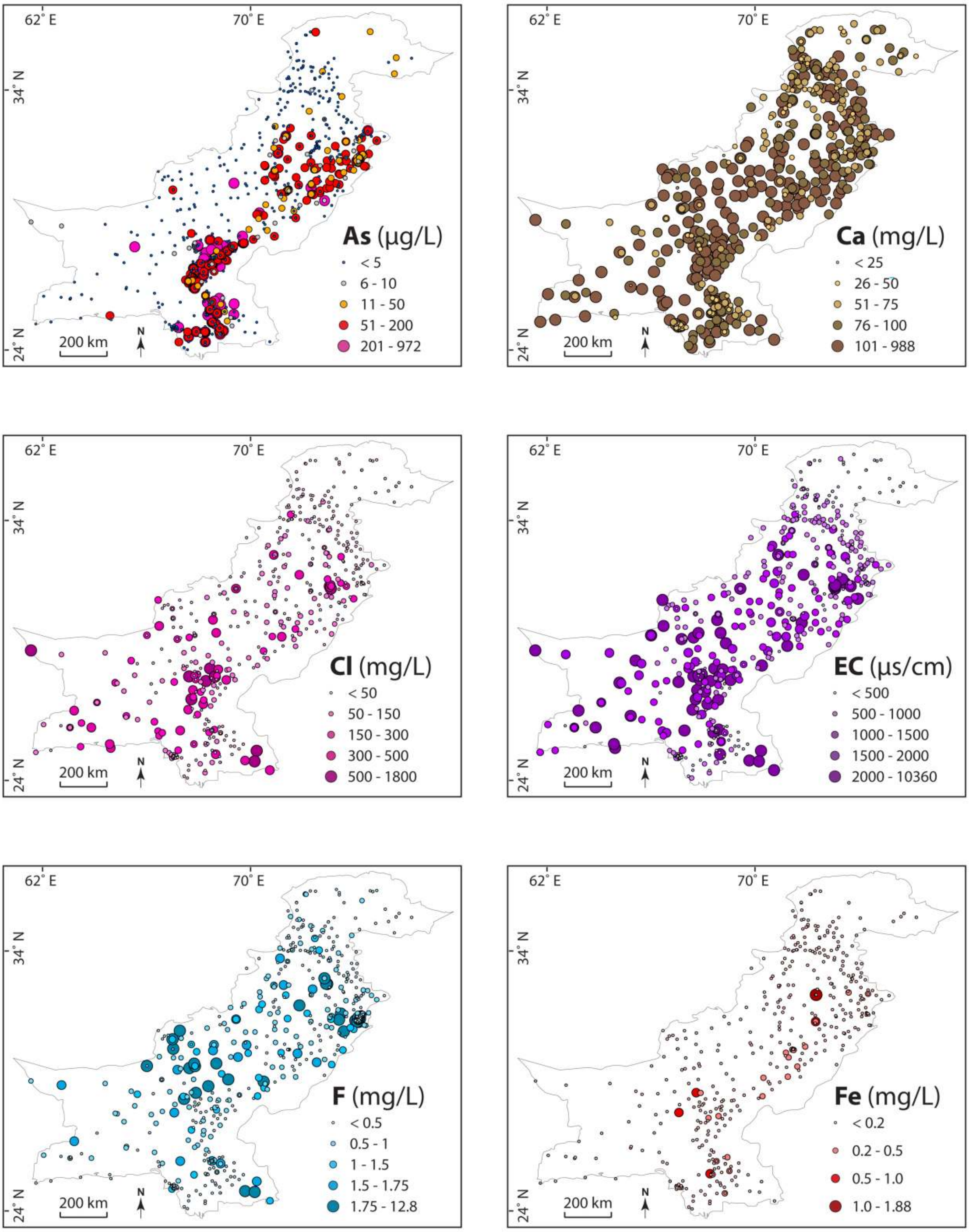

fig. S1. Maps of spatial distribution and values of all chemical parameters. Maps of Pakistan depicting the spatial distribution of all chemical parameters measured for this study in groundwater samples that were collected throughout the country (n=458).

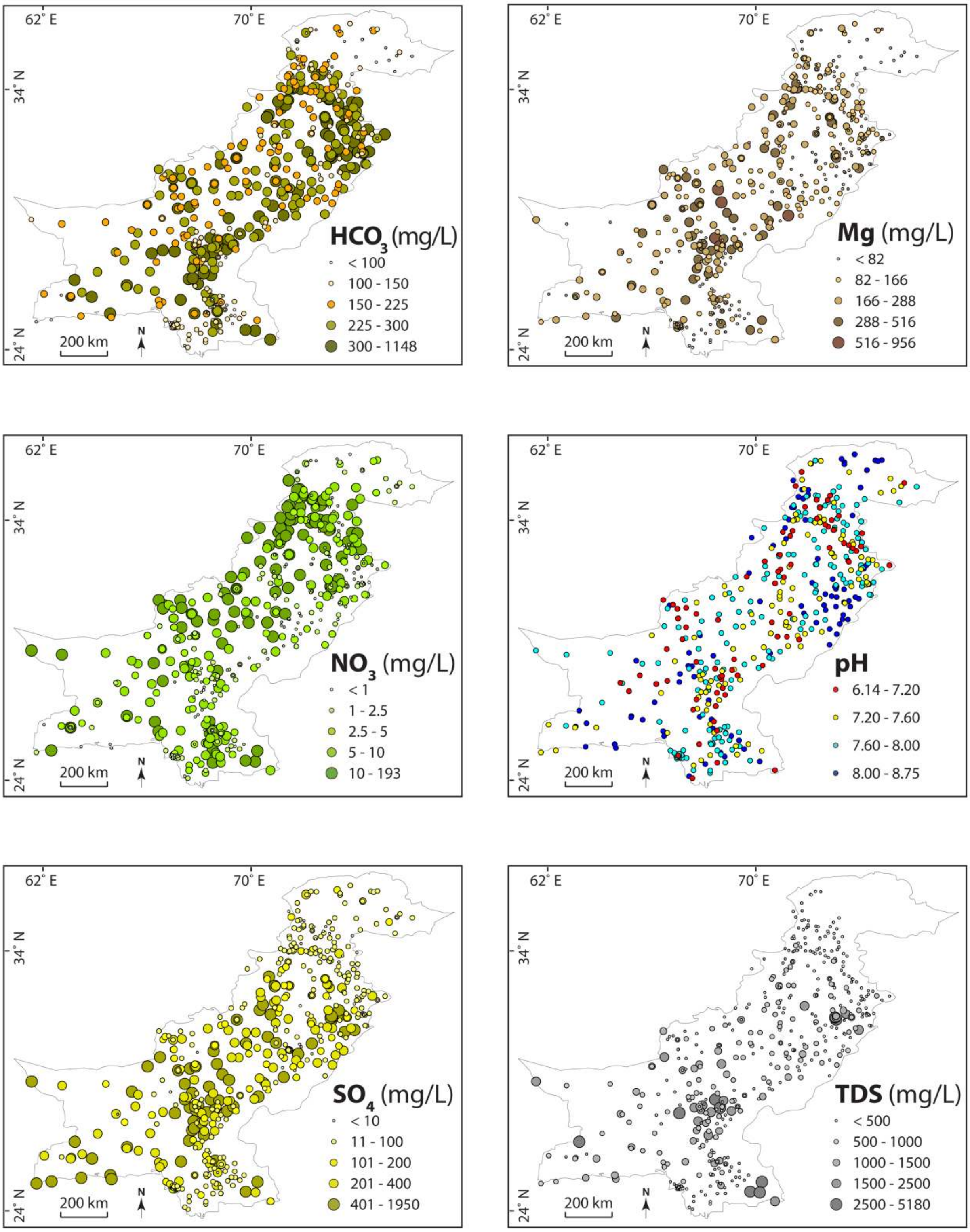
62° E
70° E
34° N
24° N
200 km
N
HCO3 (mg/L)
< 100
100 - 150
150 - 225
225 - 300
300 - 1148
Mg (mg/L)
< 82
82 - 166
166 - 288
288 - 516
516 - 956
NO3 (mg/L)
< 1
1 - 2.5
2.5 - 5
5 - 10
10 - 193
pH
6.14 - 7.20
7.20 - 7.60
7.60 - 8.00
8.00 - 8.75
SO4 (mg/L)
< 10
11 - 100
101 - 200
201 - 400
401 - 1950
TDS (mg/L)
< 500
500 - 1000
1000 - 1500
1500 - 2500
2500 - 5180

fig. S1 (cont.)

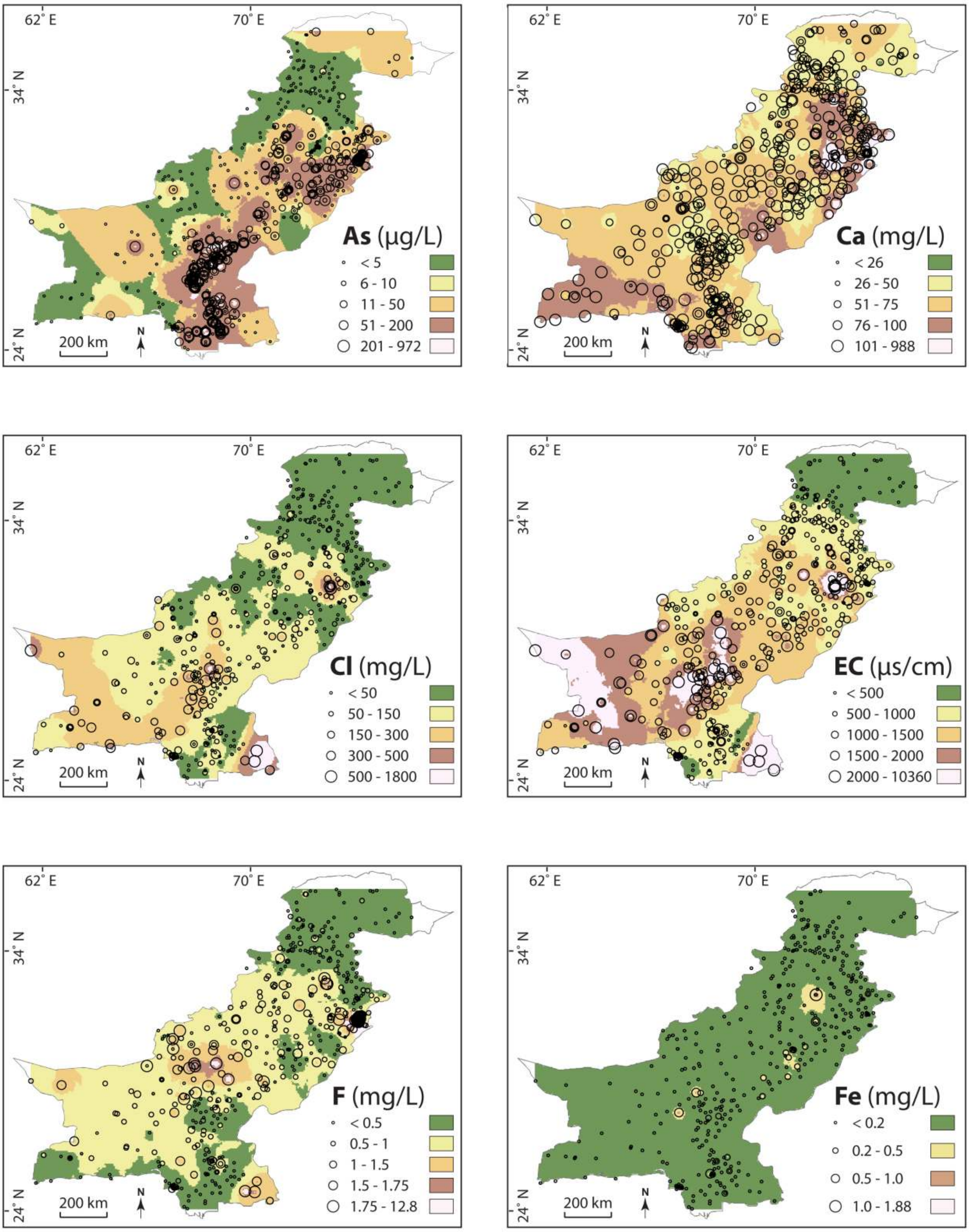

fig. S2. Grids of all measured chemical parameters. Inverse-distance-weighting grids of all chemical parameters measured for this study in groundwater samples that were collected throughout the country (n=458). The interpolation used a distance exponent of 1 and a variable search radius (encompassing 12 points).

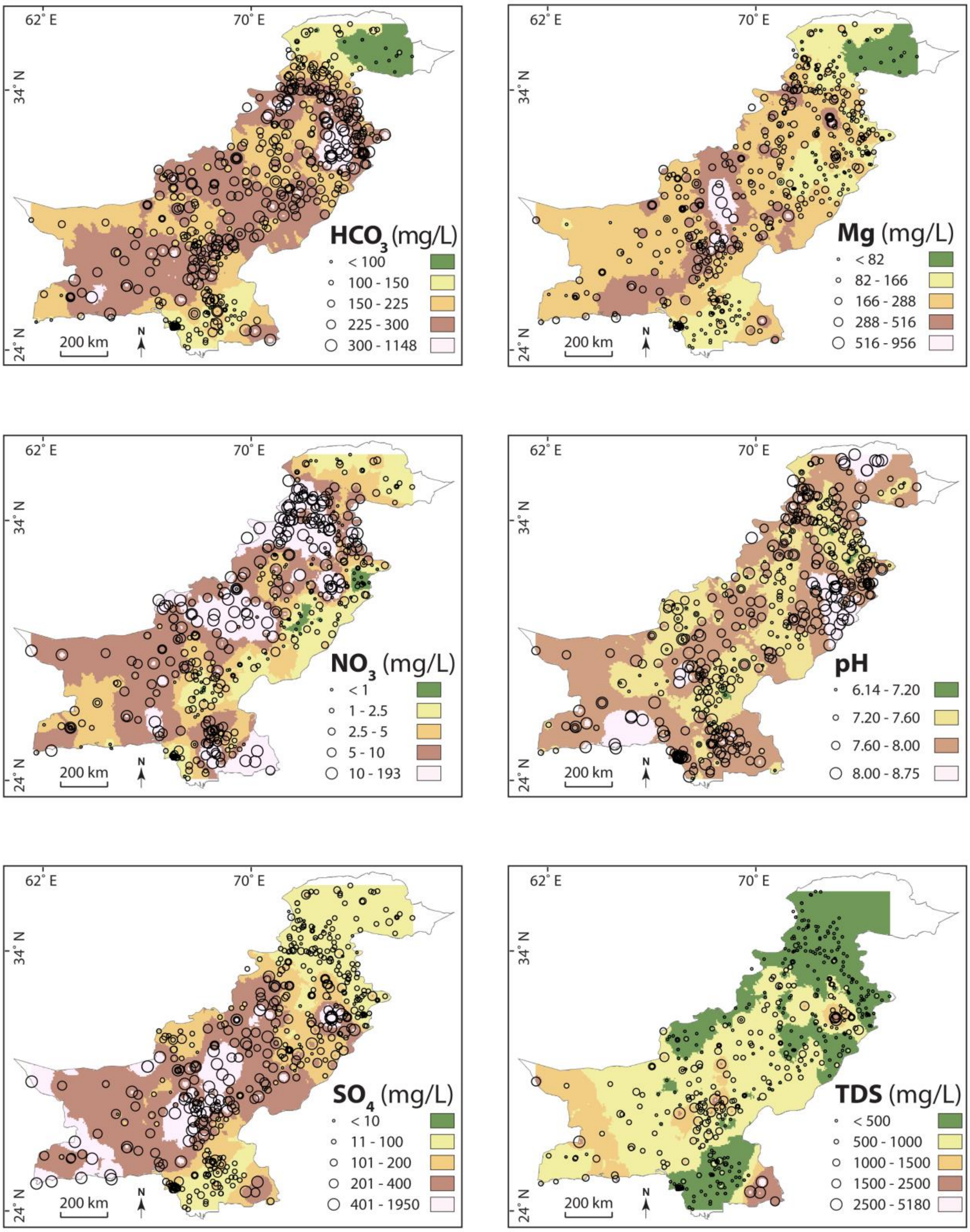
62° E
70° E
34° N
24° N
200 km
N
HCO_3 (mg/L)
< 100
100 - 150
150 - 225
225 - 300
300 - 1148
Mg (mg/L)
< 82
82 - 166
166 - 288
288 - 516
516 - 956
NO_3 (mg/L)
< 1
1 - 2.5
2.5 - 5
5 - 10
10 - 193
pH
6.14 - 7.20
7.20 - 7.60
7.60 - 8.00
8.00 - 8.75
SO_4 (mg/L)
< 10
11 - 100
101 - 200
201 - 400
401 - 1950
TDS (mg/L)
< 500
500 - 1000
1000 - 1500
1500 - 2500
2500 - 5180

fig. S2. (cont.)

table S1. Summary statistics of all measured parameters. Count, range, mean and median of arsenic and 11 other parameters measured in groundwater and well depth. Statistics are provided for all measurements as well as for the groupings of arsenic measurements within and exceeding 10 µg/L.

		As (µg/L)	**Ca (mg/L)**	**Cl (mg/L)**	**EC (µs/cm)**	**F (mg/L)**	**Fe (mg/L)**	**HCO_3 (mg/L)**	**Mg (mg/L)**	**NO_3 (mg/L)**	**pH**	**SO_4 (mg/L)**	**TDS (mg/L)**	**Depth (m)**
ALL	Count	1184	483	483	485	580	445	483	483	459	485	483	480	343
	Range	0-500	6-988	0.4-1800	24-10360	0-12.8	0-1.88	0.68-1148	4-956	0-193	6.14-9.18	1-1950	3.6-5180	3-70
	Mean	100±104	116±89	86±167	1100±1110	0.8±1.10	0.05±0.13	224±113	143±119	7.4±13.9	7.67±0.45	209±269	556±557	17±11
	Median	100	92	30	763	0.5	0.01	216	114	3.3	7.71	105	386	12
As ≤ 10 µg/L	Count	399	396	397	398	397	363	398	396	379	398	396	396	0
	Range	0-10	6-988	0.4-1800	24-10360	0-3.7	0-1.88	0.68-1148	4-784	0-193	6.32±9.18	1-1950	3.6-5180	n/a
	Mean	1±3	114±88	86±163	1102±1102	0.6±0.6	0.04±0.13	222±118	146±114	8.4±15.0	7.68±0.45	209±268	556±553	n/a
	Median	0	91	32	762	0.4	0.01	209	116	3.8	7.72	105	385	n/a
As > 10 µg/L	Count	785	87	86	87	183	82	85	87	80	87	87	84	343
	Range	12-500	18-656	1.5-1350	178-7240	0-12.8	0-0.86	64-400	8-956	0-28.4	6.14-8.30	8-1600	88.8-3620	3-70
	Mean	150±94	126±97	88±186	1089±1149	1.2±1.7	0.09±0.15	234±86	129±138	2.7±3.9	7.62±0.44	212±278	559±579	17±11
	Median	126	94	27	786	0.7	0.04	242	86	1.3	7.70	100	394	12

table S2. Coefficients, SDs, and frequencies of predictor variables in logistic regressions with a threshold of 10 µg/liter. Weighted coefficients, standard deviations and frequency of variables retained in the original 1000 logistic regression runs using a threshold of 10 µg/L. (WHO As-guideline).

Variable	Coefficient	Std. dev.	Freq. in logistic regressions
(intercept)	-3.75	0.48	768
fluvisols (probability)	1.39	0.23	768
Holocene fluvial sed. (binary)	1.16	0.10	768
slope (binary, 0.1°)	-0.67	0.10	766
soil organic carbon	-1.73	0.56	610
soil pH	4.21	0.58	768

table S3. Coefficients, SDs, and frequencies of predictor variables in logistic regressions with a threshold of 50 µg/liter. Weighted coefficients, standard deviations and frequency of variables retained in the original 1000 logistic regression runs using a threshold of 50 µg/L (Pakistan As-guideline).

Variable	Coefficient	Std. dev.	Freq. in analyses
(intercept)	-4.94	0.43	860
slope (binary, 0.1°)	-0.53	0.11	802
aridity (prec./PET)	-3.69	0.62	859
fluvisols (probability)	1.25	0.23	859
soil pH	5.14	0.54	860
irrigated area %	0.74	0.15	817
Holocene fluvial sed. (binary)	0.82	0.13	860

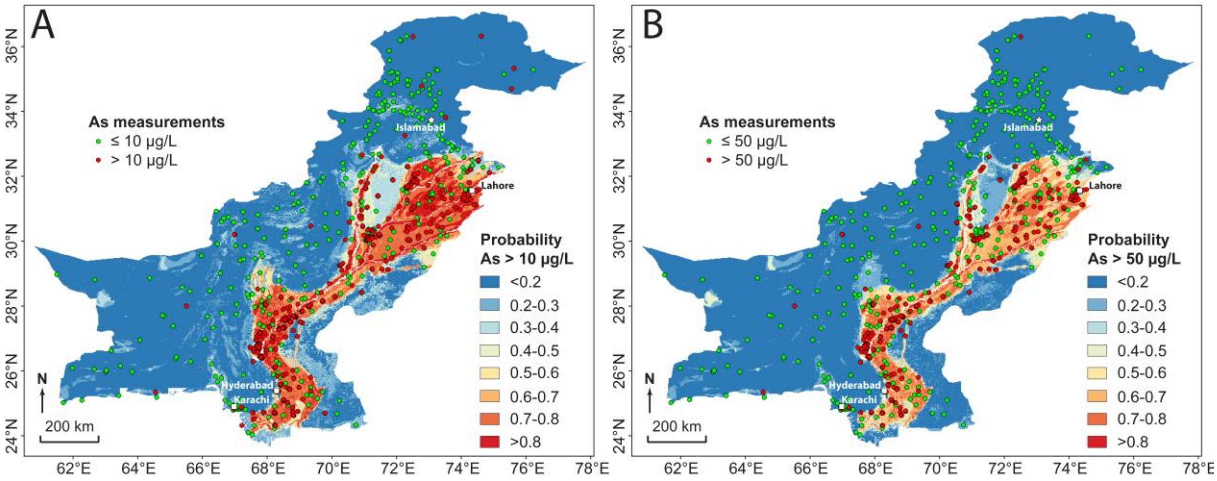

fig. S3. Hazard maps of logistic regression models using thresholds of 10 and 50 µg/liter. Probability (hazard) maps of the occurrence of arsenic concentrations in groundwater exceeding (**A**) the WHO guideline of 10 µg/L and (**B**) the Pakistan As-guideline of 50 µg/L along with the aggregated arsenic data points used in modeling (n=743).

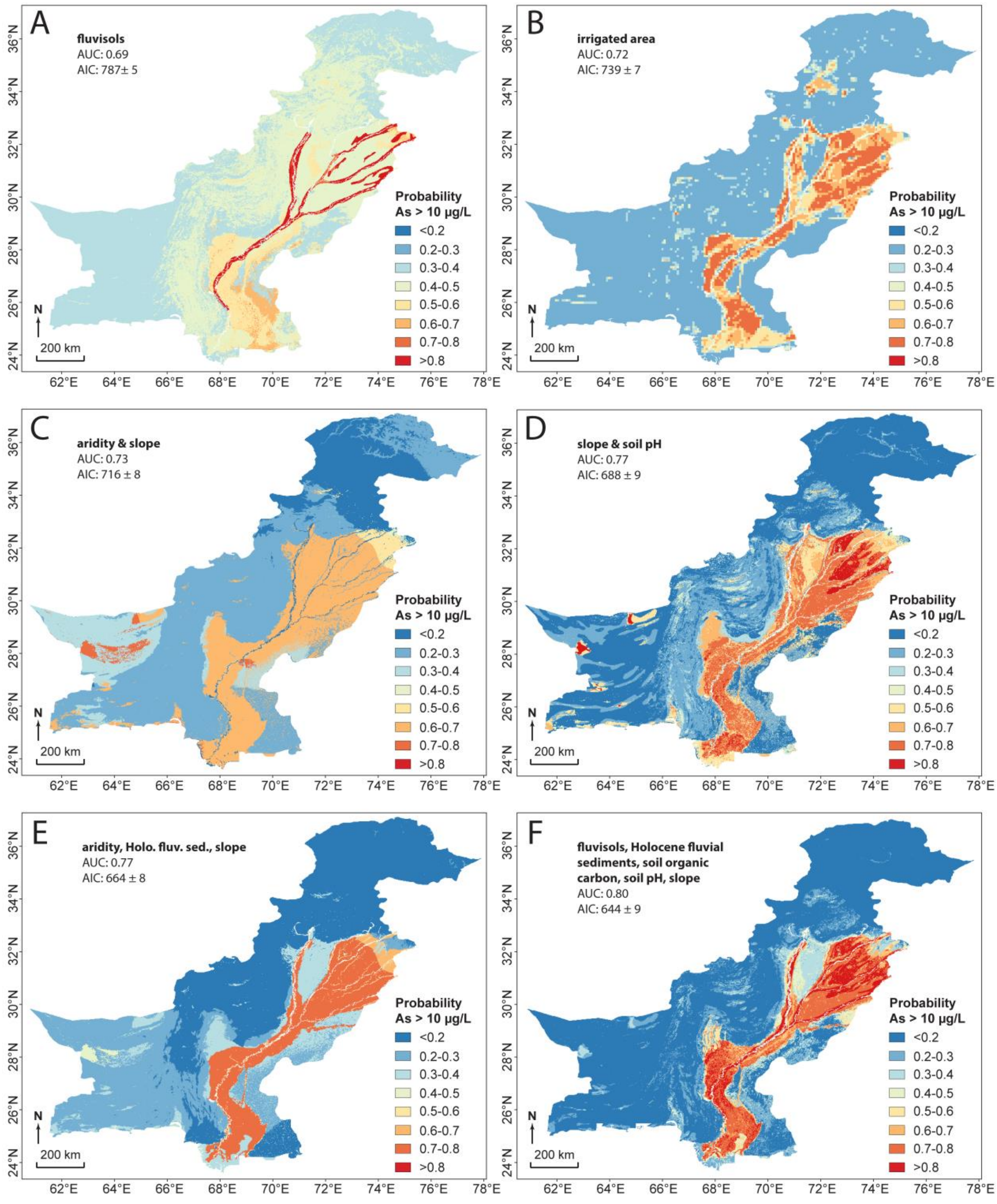

fig. S4. Maps of other well-fitting logistic regression models. Plausible hazard models using the predictor variables of (**A**) fluvisols, (**B**) irrigated area, (**C**) aridity and slope (binary, 0.1°), (**D**) slope (binary, 0.1°) and soil pH, (**E**) aridity, Holocene fluvial sediments and slope (binary, 0.1°), (**F**) fluvisols, Holocene fluvial sediments, soil organic carbon, soil pH and slope (binary, 0.1°) (final model in paper).

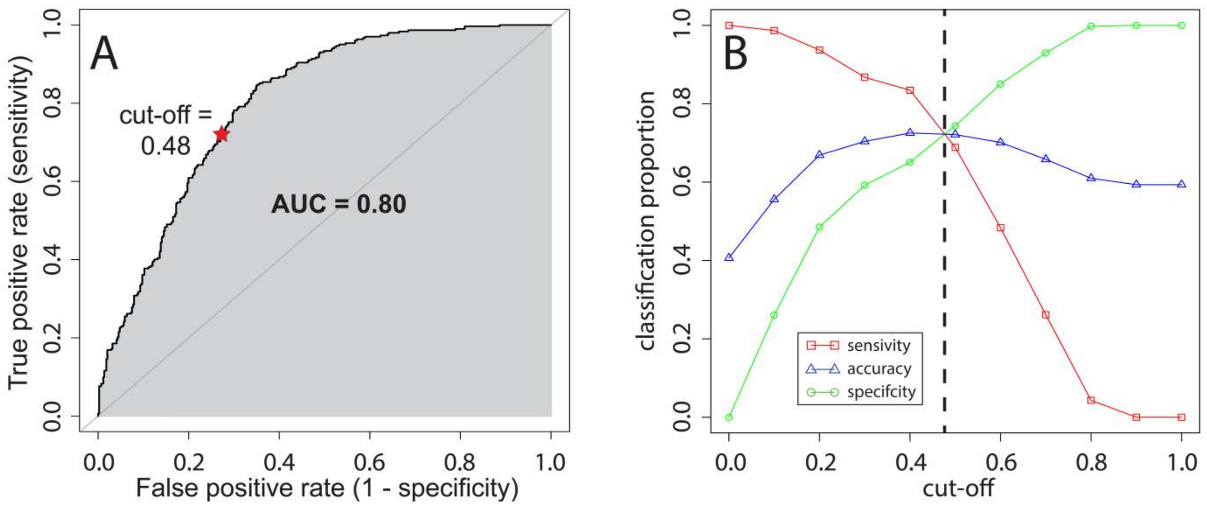

fig. S5. Statistics of logistic regression model using threshold of 50 µg/liter. Statistics of the of the classification strength of the logistic regression analysis results using the threshold of 50 µg/L (Pakistan As-guideline) applied to the entire set of 743 data points: (**A**) the ROC curve (AUC of 0.80) and (**B**) sensitivity (true positive rate), accuracy and specificity (true negative rate) versus cut-off value.

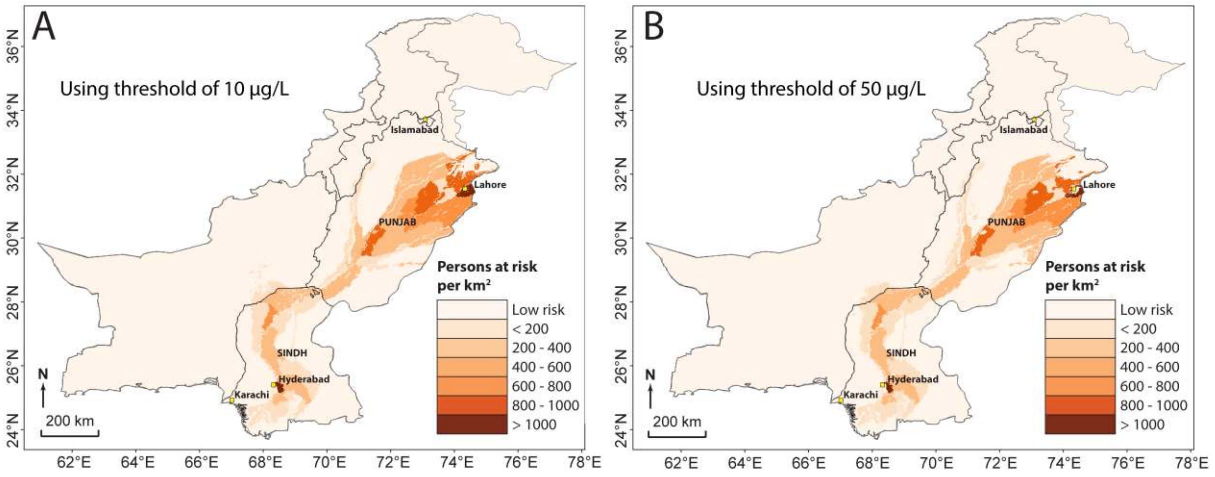

fig. S6. Density of population at risk using logistic regression models with thresholds of 10 and 50 µg/liter. Density of population at risk of high levels of arsenic in groundwater using (**A**) the WHO guideline of 10 µg/L and (**B**) Pakistan As-guideline of 50 µg/L. The population risk in both cases falls in the range of 50–60 million people, which is based on 2016 population estimates and a 60–70% groundwater utilization rate.

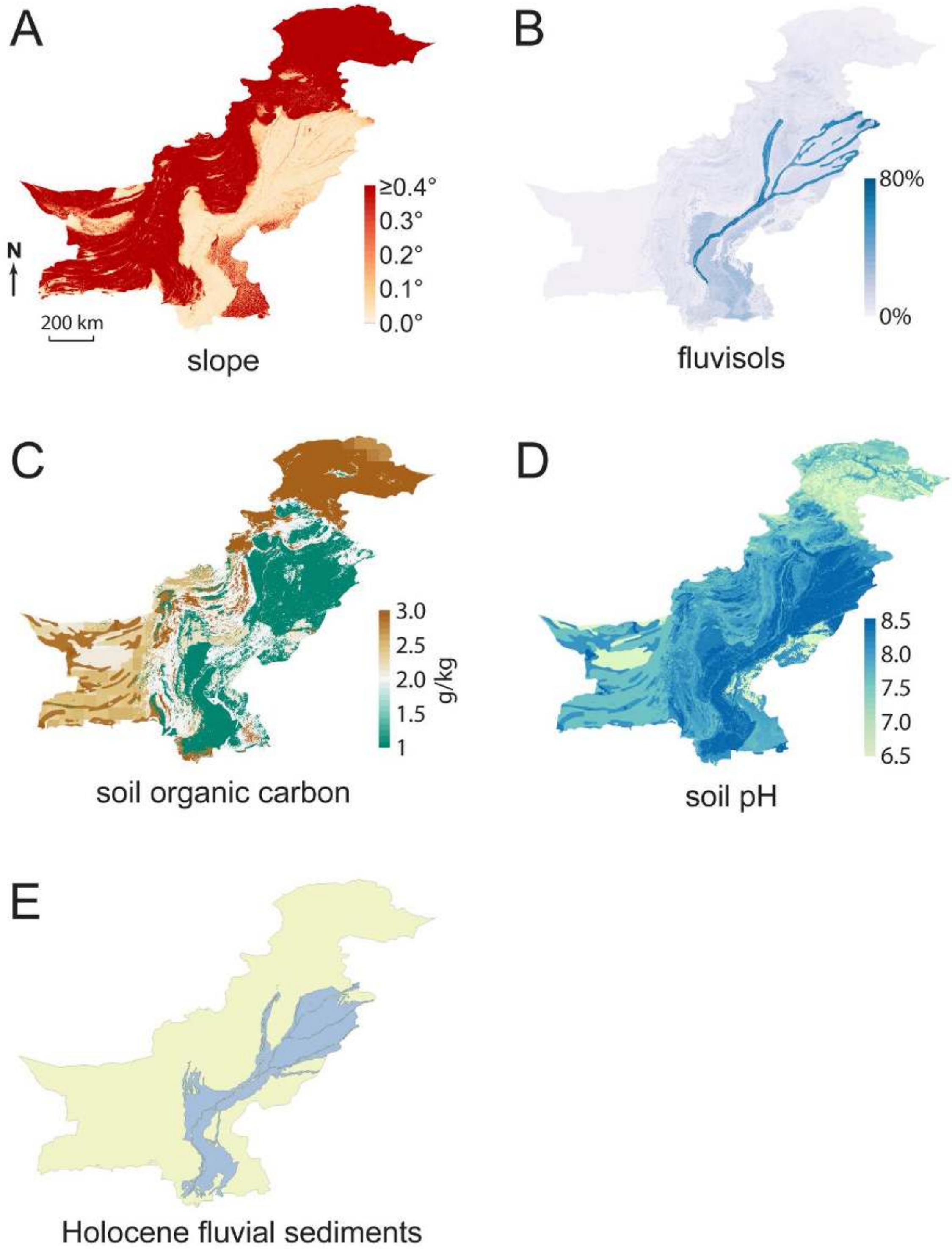

fig. S7. Predictor data sets used in the final model. Predictor datasets used in the final model: (**A**) Slope derived from the SRTM 30 arc-second DEM (*65*). (**B**) Predicted occurrence of the presence of fluvisols (*59*, *66*, *67*). (**C**) Estimated soil organic carbon at 150 cm depth (*59*, *66*, *67*). (**D**) Estimated soil pH at 150 cm depth (*59*, *66*, *67*). (**E**) Holocene fluvial sediments (*68*).

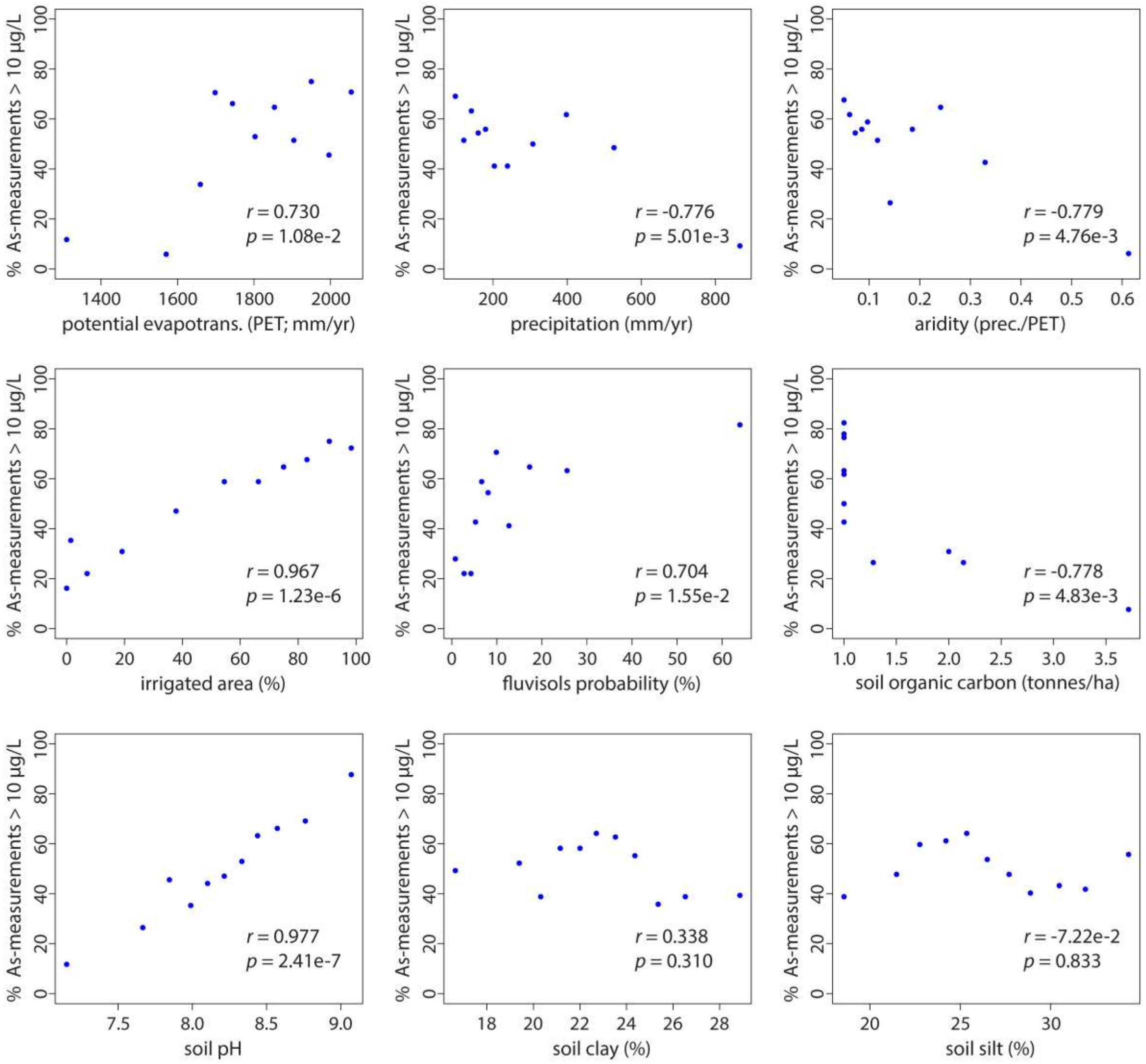

fig. S8. Graphs of correlation between predictor variables and percentage of As measurements >10 μg/liter. Correlation of the percentage of groundwater samples exceeding the WHO As-guideline of 10 μg/L against continuous predictor variables. Data were placed into 11 bins of variable width to allow each bin to contain the same number of members.