Supplementary Materials for

Groundwater Arsenic Contamination Throughout China

Luis Rodríguez-Lado, Guifan Sun, Michael Berg, Qiang Zhang, Hanbin Xue, Quanmei Zheng, C. Annette Johnson*

*Corresponding author. E-mail: annette.johnson@eawag.ch

Published 23 August 2013, *Science* **341**, 866 (2013)
DOI: 10.1126/science.1237484

This PDF file includes:

Additional data is available through www.eawag.ch/repository/remarc

Probability map for As > 10 µg/L (raster data)
Binary risk map for As > 10 µg/L using the 0.46 probability threshold (raster data)
Population at risk map (raster data)
Map of model uncertainty (raster data)
As database used to calibrate the model (ESRI shapefile)
As database used to validate the model (ESRI shapefile)

Materials and Methods

Calibration dataset.

Arsenic data from 49,362 tested wells in 2,668 villages in the provinces of Inner Mongolia (n=683), Gansu (n=594), Shanxi (n=505), Ningxia (n=268), Henan (n=299) and Heilongjiang (n=319), obtained from the "Chinese National Survey Program", were used to calibrate the model (Fig. S1). The database provides information on the number of wells tested, the percentage of wells with arsenic concentrations above and below arsenic thresholds of 10, 50, 100 and 200 μg L^{-1} and the maximum and minimum values of arsenic concentrations found at each location. Additional geochemical information is not available. The arsenic measurements in this dataset have been aggregated to the spatial resolution of the auxiliary raster maps (1 km^2). We used the maximum arsenic concentration found at each 1 km^2 pixel for the statistical analyses. The aggregated point-data were binary-coded using the World Health Organization guideline value for As in drinking water (10 μg L^{-1}) as a threshold and used as a binary response variable in the logistic regression models. We could not calibrate the model to the Chinese standard threshold of 50 μg L^{-1} since arsenic concentrations above this threshold were only found in 7% of the samples.

Auxiliary rasters.

i) Topographic parameters. A Digital Elevation Model (DEM) at 500 meter resolution was retrieved from the Consortium for Spatial Information (http://www.cgiar-csi.org/) and aggregated to 1 km^2 resolution. The DEM was projected to a Lambert Conformal Conic coordinate system to calculate the raster maps for slope (radians) and Topographic Wetness Index (TWI, Fig. S2). The rasters were then back transformed to Latitude-Longitude (EPSG 4326). Topographic Wetness Index expresses the potential wetness in soils due to topography. It is calculated as a function of the upstream contributing area (Ac) and the slope (β) of the landscape.

$$TWI = \ln\left(\frac{A_c}{\tan\beta}\right)$$

TWI has already been identified as a good predictor for high soil-moisture zones due to topography *(31)*. Onishi et al. *(32)* found a high correlation between dissolved iron in surface waters and TWI, and used TWI as a macroscopic index to assess the production of dissolved iron in water. Löhr et al. *(33)* linked TWI to the presence of seasonal redox processes in the soils and sediments of a forested coastal catchment. Andersson and Nyberg *(34)* found a strong correlation between TWI and DOC concentrations in boreal catchments. Pei et al. *(35*) also found TWI to be the most significant terrain parameter correlated with the organic matter content of soils; thus, TWI may also account for the influence of organic carbon in triggering the release of As in reducing environments *(36-38*).

ii) Remote sensing information. We used a temporal series of eight images of the mean monthly "Enhanced Vegetation Index" (EVI) recorded with the TERRA "Moderate Resolution Imaging Spectroradiometer" (MODIS-TERRA). Each image is formed by a mosaic of 34 tiles covering the whole of Asia with a spatial resolution of 1 km^2 and a temporal resolution of three months over two consecutive years. These images were transformed by Principal Component Analysis to avoid redundancy in the information, and we tested their ability to explain the distribution of high As concentrations. The

First Component (EVI-PCA1), condensing 80% of the variability, was interpreted as an indicator of water availability, based on the response of vegetation, and was used as a proxy to differentiate climatic regions. The environmental interpretation for the remaining 7 components is unclear. None of these auxiliary variables was significant at the 95% level in the univariate logistic regression tests and they were not used in the model calculations.

iii) Hydrologic parameters. We used the river network of Asia provided by the "Digital Chart of the World" (http://www.fas.harvard.edu/~chgis/data/dcw/). This river network was used to derive raster maps of "Density or rivers" and "Distance to rivers" (Fig. S2).

iv) Earth's Gravitational Force. We used the data (μGal) at 1' resolution, resampled to 30 arc-second resolution, provided at the Technical University of Denmark *(39)*. This data (Fig. S2) was used as a proxy to identify sedimentary areas due to their lower mass in relation to solid (rock) substrates *(40-42)*.

v) Soil parameters. Information about subsoil texture and soil salinity has been obtained from the "Harmonized World Soil Database" (HWSD) (http://www.iiasa.ac.at/Research/LUC/External-World-soil-database/HTML/). It includes soil information at 1:1M scale for China as a 30 arc-second raster. These properties have been reclassified to binary maps (Fig. S2) as follows:

a) Subsoil texture: We used a binary classification of soil textures, in which all medium-textured soils (classes Clay Loam + Sandy Clay + Loam) were grouped together =1 and other textures = 0;

b) Salinity: Solonchaks, Solonetzs and soils with a salic phase = 1, other soils = 0.

vi) Geology. We used the digital geological map of China at 1:5M scale (http://pubs.usgs.gov/of/2001/of01-318/) to create a map of Holocene sediments (Fig. S2). In this map, desert gravels and aeolian sand deposits are also classified as Holocene sediments. However the risk of arsenic contamination in these areas is low due to the lack of water sources and to the stable population. We identified those features using both land use (http://ies.jrc.ec.europa.eu/global-land-cover-2000) and soil information (HWSD) and excluded them from the category of Holocene sediments.

vii) Population density. The dataset "Gridded Population of the World, Version 3" (GPWv3), created by the Center for International Earth Science Information Network *(43)*, provided estimates of human population for 2000 in the form of raster data at 2.5 arc-minute resolution. A proportional allocation gridding algorithm, utilizing more than 300,000 national and sub-national administrative units, is used to assign population values to cells in the raster. The population density grid (persons km^{-2}) is derived by dividing the population count grid by the land area.

Modelling procedures

Univariate logistic regression tests were conducted separately on each predictor to assess their ability to explain the arsenic binary-coded data in the calibration dataset. Those variables significant at the 95% confidence level in the tests were considered for inclusion in the subsequent multivariate logistic regression analyses. Holocene sediments, soil salinity, topographic wetness index, slope, distance to rivers, density of rivers and earth's gravity (Table S1) were significant in the univariate logistic regression tests and were thus included in the multivariate logistic regression analyses. Subsoil texture was

also included in the analyses because previously reported data indicate that fine and medium soil textures clearly affect the presence of high As concentrations due to limitations for drainage and water flux *(44)*.

The above eight retained proxies have environmental meaning to explain the natural As enrichment in groundwater: Quaternary (Holocene) sediments, together with Earth's Gravity, indicate the presence of large volumes of young floodplain and delta sediments, where most of the As affected areas in South-East Asia occur. Slope, Topographic Wetness Index (TWI) and Subsoil Texture allocate areas with reducing aquifer conditions due to topographic flatness, low hydraulic gradients (deltas, closed/semi-closed basins) and limitations for drainage and water flux. Saline soils identify environments with high pH and alkalinity where high As concentrations in arid/semiarid conditions have been reported *(11-15)*, while Distance to Rivers and Density of Rivers served as additional proxies indicating reducing aquifer conditions.

i) Classification algorithm. Single algorithm predictive models have been successfully used to estimate the spatial extent of arsenic contamination in Southeast Asia *(23–26)*. However, such models are generally sensitive to the number and location of the samples used for calibration. Ensemble models, combining several base models (ensemble members) into a single aggregated model (ensemble), have been proposed as an alternate modelling approach because they perform significantly better than single constituents *(45)*. We created an ensemble model by using multivariate stepwise logistic regression as the base classifier method to predict the spatial distribution of high arsenic occurrence rates (As >10 μg L^{-1}) in groundwaters. Logistic regression determines the existing relationship between a binary response Y (in this case as an As threshold) and a number of independent auxiliary variables $\{X_1,....,X_p\}$, using a logit link function of the form:

$$f(z) = 1/1 + e^{-z}$$

where:

$$z = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p$$

We used the 2,668 georeferenced As measurements in the calibration dataset, together with the eight previously reported relevant proxies, to calculate 100 equally-likely logistic regression models (the ensemble members) using, at each run, random subsets of the calibration dataset by sampling with replacement *(46)*. The auxiliary variables retained in each logistic regression ensemble member were automatically obtained through stepwise selection (both directions), using the Akaike Information Criterion (AIC). In addition, for each ensemble member, we calculated the overall internal accuracy of the model by cross-validation (the data samples used to build the member were randomly assigned to 20 different groups; each group was removed in turn, while the remaining data was used to re-fit the regression model and predict the deleted observations). We used the Hosmer & Lemeshow goodness-of-fit test *(47)* to verify all ensemble members. High p-values on this test ($p>0.05$) indicate that the model fits the data. The 48 ensemble members that passed this test (Fig. S3) were withheld to build the final ensemble model, which is the weighted average of the predictions from the 48 ensemble members according to their performance (the overall internal accuracy of each ensemble member was used as the weighting factor, so that members with higher prediction power have a greater influence on the final prediction).

$$p_{EM} = \sum_{k=1}^{M} w_k p_k / M$$

where p_{EM} is the final ensemble model probability and w_k and p_k are the single-model weights (internal cross-validation accuracy) and predicted probabilities, respectively.

All the statistical analyses were conducted on the normalized values of the auxiliary rasters.

ii) Variable importance. The importance of the independent variables in calculating high As probabilities was estimated by the values of the odds ratio $exp(\beta)$ from the univariate logistic regression models (Table S1, column 4). It expresses the changes in the response variable associated with 1 unit change in the predictor variable (the larger the value, the more influential the variable). Since the predictors were previously normalized, the values of $exp(\beta)$ are comparable. We also calculated the number of multivariate logistic regression members, after the stepwise method, in which each auxiliary variable was retained as significant (*Frequency*). It reflects the relative contribution of each variable to the ensemble model output (Table S1, column 5).

iii) Model validation. We validated the model with an external dataset consisting of 625 independent geo-referenced observations from a compilation of arsenic measurements in wells that included aggregated data from the "Chinese National Survey Program" in the Autonomous Province of Xinjiang (4,458 tested wells in 184 villages), 261 records from published data *(16, 18, 2-22, 48, 49)* and 180 wells from our own field surveys. The arsenic concentrations in this dataset were converted to binary data by considering the WHO guideline of 10 μg L^{-1} as the threshold, and they were then compared with the corresponding binary classes of high/low arsenic model probabilities considering an optimal cut-off value of 0.46. This cut-off threshold was calculated using the Receiver Operating Characteristic curve (ROC) (Fig. S4). At each location, the prediction is considered ≤10 μg L^{-1} if the predicted probability is <0.46. The model accuracy was estimated from the overall correct classification rate, model sensitivity (ability to correctly classify samples with As >10 μg L^{-1}) and model specificity (ability to correctly classify samples with As ≤10 μg L^{-1}) (Tables S3-S7).

Prediction errors are mainly related to errors in As measurements, the spatial uncertainty of the auxiliary variables (which do not fully represent the spatial heterogeneity at field scale), to the high spatial variability of As concentrations over short distances and to the exclusive use of surface parameters as predictors. Groundwater depth and water management practices are important parameters that are often overlooked when assessing groundwater As risk. The development of 3D geological models that include in-depth groundwater information is promising and may, in some cases, be more appropriate for evaluating the risk of arsenic contamination on local scales.

*iv) Model uncertainty***.** We used the standard deviation of the mean (s_m) to estimate the uncertainty of the weighted average ensemble model estimates. It was calculated as follows:

$$s_m = \frac{\sqrt{\sum_{i=1}^{M}(x_i - \bar{x})^2 / M}}{\sqrt{M}}$$

where x_i is each i^{th} ensemble member and x is the weighted average ensemble model. In this case, M is represented by the 48 ensemble members with p-values >0.05 for the Hosmer & Lemeshow goodness-of-fit test. The uncertainty map (Fig. S5) is illustrated in normalized s_m values to highlight the differences between areas.

Validation

The following equations were used for validation:

$$Overall \cdot accuracy = \frac{TP + TN}{TP + FP + FN + TN} \times 100;$$

$$Sensitivity = \frac{TP}{TP + FN} \times 100;$$

$$Specificity = \frac{TN}{TN + FP} \times 100;$$

The Cohen's Kappa Index of Agreement, KIA (Cohen, 1960), evaluates the strength of the model agreements in the binary classification matrix by taking into account the proportions of agreement expected due to chance.

$$KIA = \frac{P_{obs} - P_{\exp}}{1 - P_{\exp}}$$

where P_{obs} and $P_{\exp}$ are the observed and expected agreements, as follows:

$$P_{obs} = \frac{TP + TN}{N};\ P_{\exp} = \frac{(TP + FN)(TP + FP) + (FP + TN)(FN + TN)}{N^2};$$

Perfect and random agreements between model and reality are related to KIA values of 1 and 0 respectively. Otherwise, the model performance range from from “poor’ (<0.40), ‘fair’ (0.4–0.75) to ‘excellent’ (>0.75) *(50)*.

Table S2 shows the confusion matrix template used for the validation with the external dataset. Tables S3 shows the results of the whole data set. Tables S4 to S7 show the validation details for known reducing and oxidizing environments.

Supplementary Figures

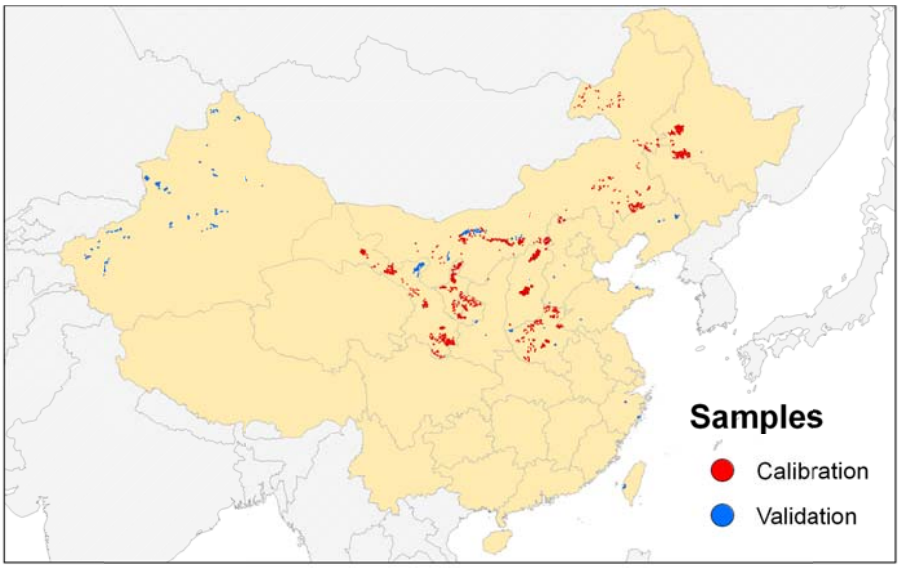

Fig. S1. Location of the arsenic samples in validation and calibration datasets.

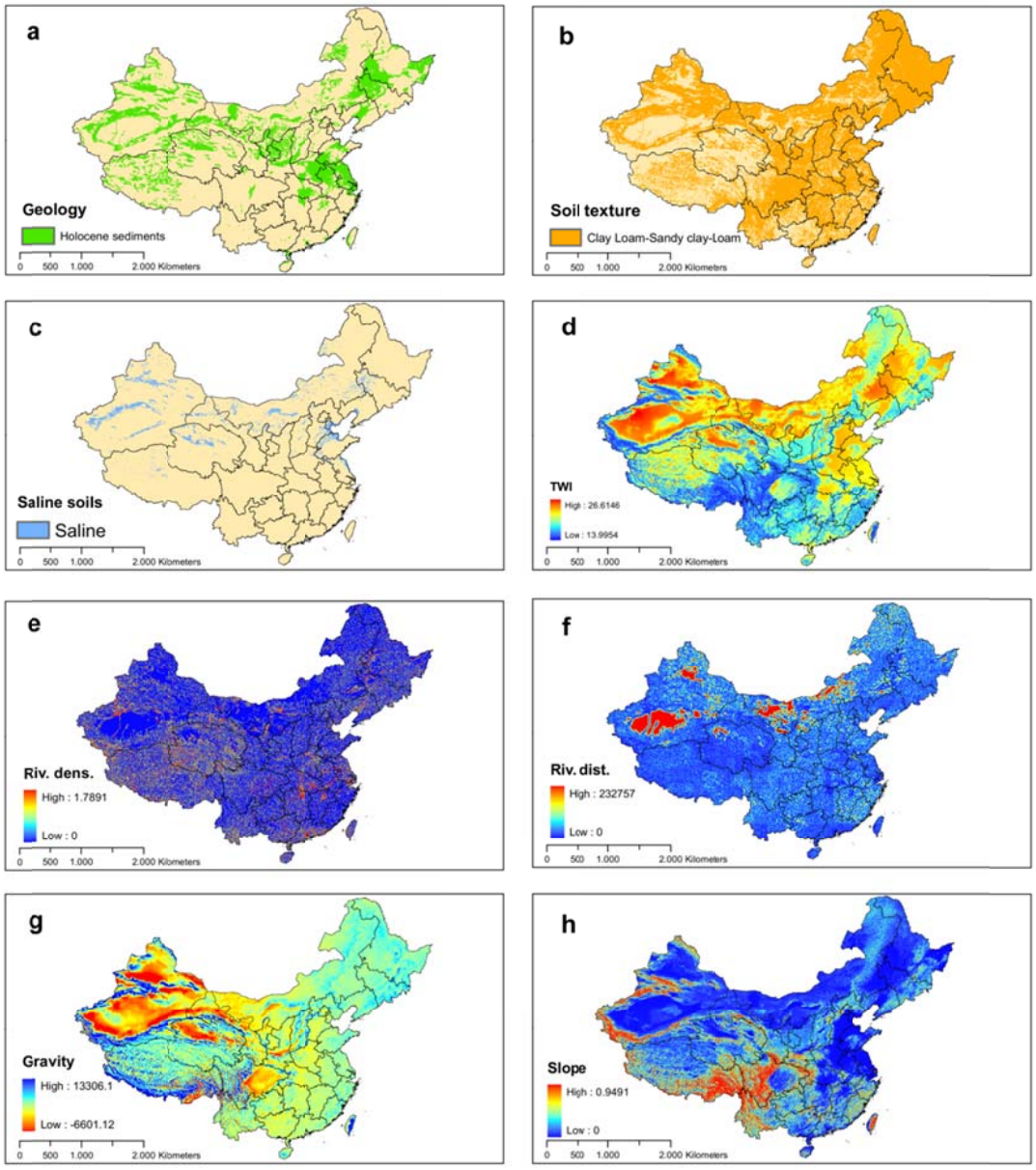

Fig. S2. Auxiliary variables included in the model.

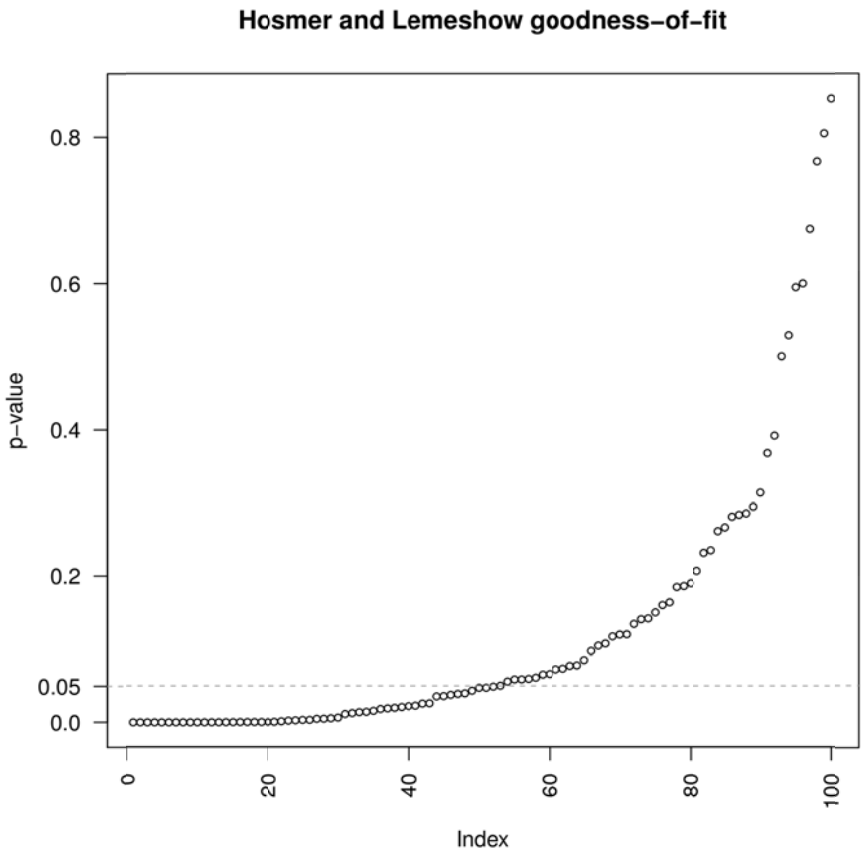

Fig. S3. Sorted p-values of the Hosmer and Lemeshow goodness-of-fit test for the 100 ensemble members. Only those members with p-values > 0.05 (n=48) were used to create the weighted average ensemble model.

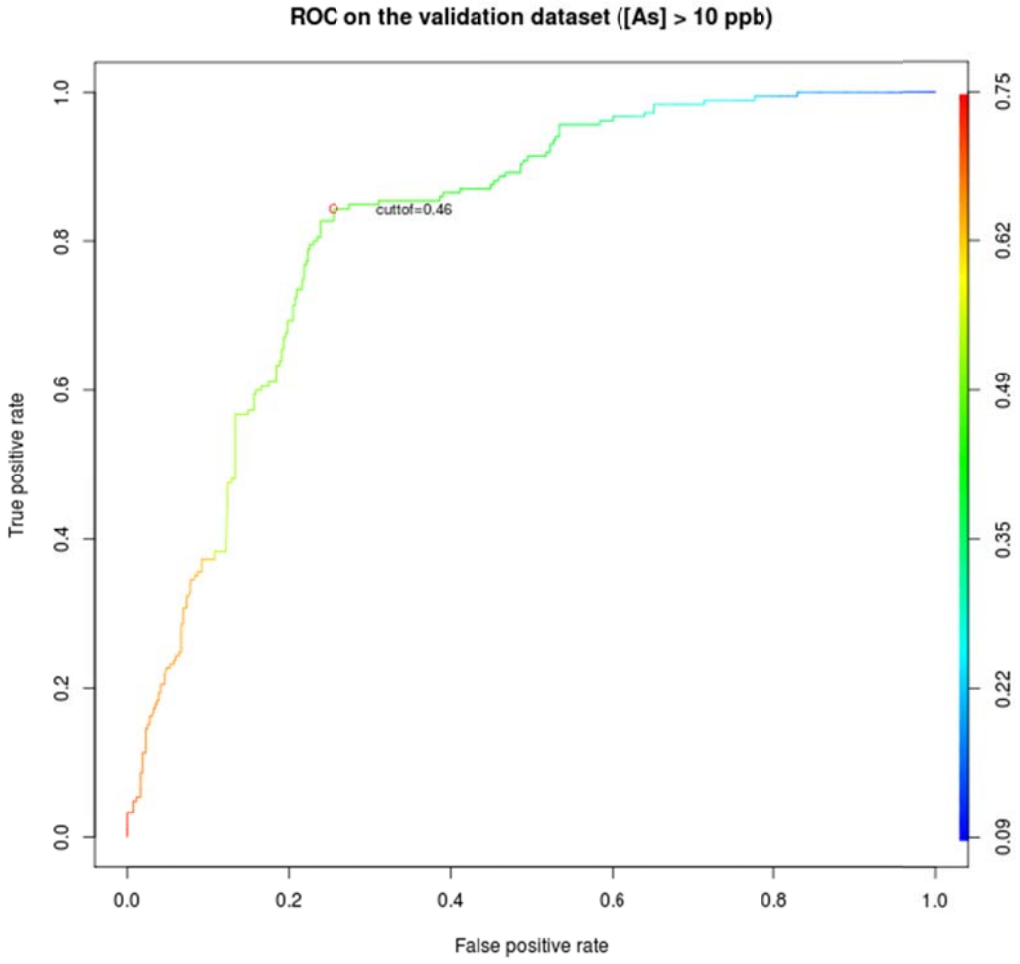

Fig. S4. Receiver Operating Characteristic curve (ROC) and optimal cutoff value.

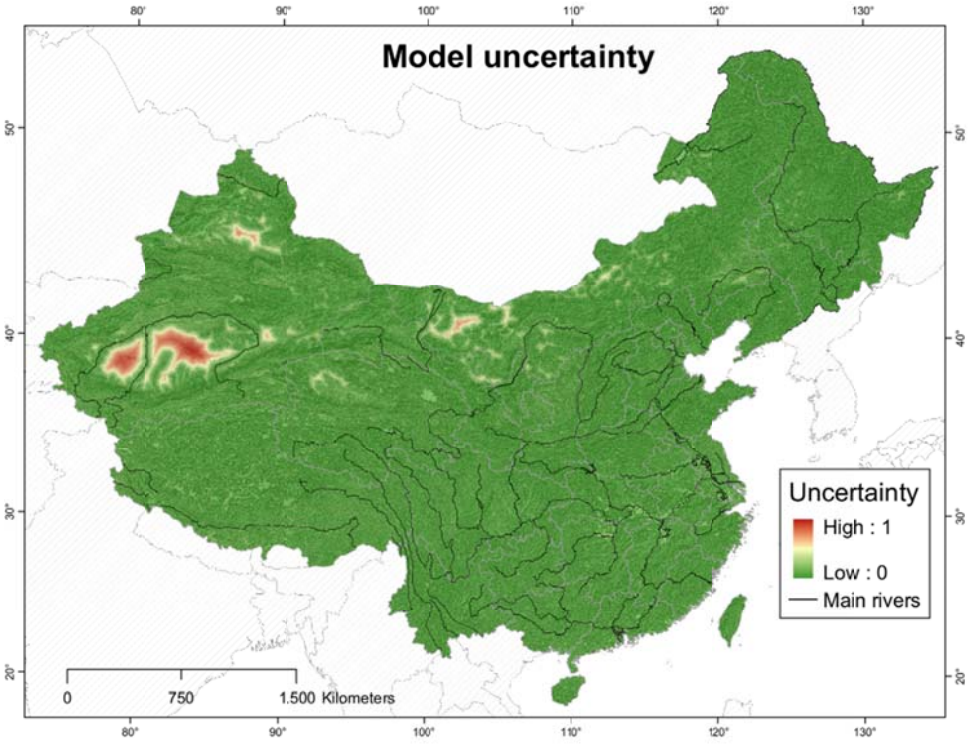

Fig. S5. Normalized model uncertainty. This map reveals high uncertainty in the predictions for desert areas in northern China.

Supplementary Tables

Table S1. Explanatory variables retained to build the logistic regression ensemble. Regression coefficients in the univariate logistic regression models (β). The relative importance of the variables is indicated by *exp*(β). *Frequency* represents the number of members in which each auxiliary variable has been retained after the stepwise method.

	Variable	*Type* [a]	β	*exp(β)* [b]	*Frequency*
Geology	Holocene sediments	cat.	0.91	2.49	48
Soils	Saline soils	cat.	1.04	2.83	48
	Subsoil texture	cat.	0.16	1.17	47
Topography	TWI	cont.	4.10	60.5	45
	Slope	cont.	-17.21	0.00	4
Hydrology	Density of rivers	cont.	1.02	2.79	7
	Distance to rivers	cont.	-11.53	0.00	3
Gravity	Gravity	cont.	-4.72	0.00	1

[a]cat.=categorical variable; cont.= continuous variable.
[b]change in the odds-ratio associated with a 1 unit change in the predictor variable.

Table S2. Confusion matrix template used for the validation with the external dataset.

		Predicted	
		>10 μg L^{-1}	≤10 μg L^{-1}
Observed	>10 μgL^{-1}	True Positive (TP)	False Negative (FN)
	≤10 μgL^{-1}	False Positive (FP)	True Negative (TN)

Table S3. Confusion matrix and validation results for the complete validation dataset.

Overall agreement = 77.12% Sensitivity = 83% Specificity = 74.54% KIA = 0.513 (Fair agreement). SE of KIA = 0.034; C.I. for KIA = 0.446 to 0.580 Observed agreements = 482 Agreements expected by chance = 331 (53%)		Predicted	
		>10 μg L^{-1}	≤10 μg L^{-1}
Observed	>10 μg L^{-1}	154 (TP)	31 (FN)
	≤10 μg L^{-1}	112 (FP)	328 (TN)

Table S4. Confusion matrix and validation results for Xinjiang province.

Overall agreement = 64% Sensitivity = 85% Specificity = 50%		Predicted	
		>10 μg L^{-1}	≤10 μg L^{-1}
Observed	>10 μg L^{-1}	59 (TP)	10 (FN)
	≤10 μg L^{-1}	57 (FP)	58 (TN)

Table S5. Confusion matrix and validation results for Hetao Plain and Huhhot Basin.

Overall agreement = 76% Sensitivity = 95% Specificity = 15%		Predicted	
		>10 μg L^{-1}	≤10 μg L^{-1}
Observed	>10 μg L^{-1}	78 (TP)	4 (FN)
	≤10 μg L^{-1}	22 (FP)	4 (TN)

Table S6. Confusion matrix and validation results for Minqin Basin and Chahaertan Oasis.

Overall agreement = 83% Sensitivity = 0% Specificity = 83%		Predicted	
		>10 μg L^{-1}	≤10 μg L^{-1}
Observed	>10 μg L^{-1}	0 (TP)	0 (FN)
	≤10 μg L^{-1}	13 (FP)	62 (TN)

Table S7. Confusion matrix and validation results for Liao-Ho Basin.

Overall agreement = 98% Sensitivity = 0% Specificity = 98%		Predicted	
		>10 μg L^{-1}	≤10 μg L^{-1}
Observed	>10 μg L^{-1}	0 (TP)	0 (FN)
	≤10 μg L^{-1}	1 (FP)	41 (TN)

References

1. J. Qiu, China to spend billions cleaning up groundwater. *Science* **334**, 745–745 (2011). doi:10.1126/science.334.6057.745

2. G. Sun, Arsenic contamination and arsenicosis in China. *Toxicol. Appl. Pharmacol.* **198**, 268–271 (2004). doi:10.1016/j.taap.2003.10.017 Medline

3. G. Sun, Y. Xu, Q. Zheng, S. Xi, Arsenicosis history and research progress in Mainland China. *Kaohsiung J. Med. Sci.* **27**, 377–381 (2011). doi:10.1016/j.kjms.2011.05.004 Medline

4. J. X. Guo, L. Hu, P. Z. Yand, K. Tanabe, M. Miyatalre, Y. Chen, Chronic arsenic poisoning in drinking water in Inner Mongolia and its associated health effects. *J. Environ. Sci. Health A* **42**, 1853–1858 (2007). doi:10.1080/10934520701566918 Medline

5. S. Li, T. Xiao, B. Zheng, Medical geology of arsenic, selenium and thallium in China. *Sci. Total Environ.* **421-422**, 31–40 (2012). doi:10.1016/j.scitotenv.2011.02.040 Medline

6. W. P. Tseng, H. M. Chu, S. W. How, J. M. Fong, C. S. Lin, S. Yeh, Prevalence of skin cancer in an endemic area of chronic arsenicism in Taiwan. *J. Natl. Cancer Inst.* **40**, 453–463 (1968). Medline

7. Y. Jin, C. Liang, G. He, J. Cao, [Study on distribution of endemic arsenism in China]. *J. Hygiene Res.* **32**, 519–540 (2003). (In Chinese) Medline

8. J. Liu, B. Zheng, H. V. Aposhian, Y. Zhou, M.-L. Chen, A. Zhang, M. P. Waalkes, Chronic arsenic poisoning from burning high-arsenic-containing coal in Guizhou, China. *Environ. Health Perspect.* **110**, 119–122 (2002). doi:10.1289/ehp.02110119 Medline

9. 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). doi:10.1016/S0883-2927(02)00018-5

10. Y. Deng, Y. Wang, T. Ma, Isotope and minor element geochemistry of high arsenic groundwater from Hangjinhouqi, the Hetao Plain, Inner Mongolia. *Appl. Geochem.* **24**, 587–599 (2009). doi:10.1016/j.apgeochem.2008.12.018

11. H. Guo, Y. Wang, Geochemical characteristics of shallow groundwater in Datong basin, northwestern China. *J. Geochem. Explor.* **87**, 109–120 (2005). doi:10.1016/j.gexplo.2005.08.002

12. H. Guo, B. Zhang, G. Wang, Z. Shen, Geochemical controls on arsenic and rare earth elements approximately along a groundwater flow path in the shallow aquifer of the Hetao Basin, Inner Mongolia. *Chem. Geol.* **270**, 117–125 (2010). doi:10.1016/j.chemgeo.2009.11.010

13. L. H. Zhang, Q. H. Guo, in *Water-Rock Interaction Volumes 1 & 2*, T. D. Bullen, Y. Wang, Eds. (Taylor & Francis, London, 2007), pp. 1299–1303.

14. P. L. 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

15. H. Hagiwara, J. Akai, K. Terasaki, T. Yoshimura, H. Luo, *Appl. Geochem.* **26**, 380–393 (2011).

16. X. Xie, T. M. Johnson, Y. Wang, C. C. Lundstrom, A. Ellis, X. Wang, M. Duan, Mobilization of arsenic in aquifers from the Datong Basin, China: Evidence from geochemical and iron isotopic data. *Chemosphere* **90**, 1878–1884 (2013). doi:10.1016/j.chemosphere.2012.10.012 Medline

17. J.-J. Lee, C.-S. Jang, C.-W. Liu, C.-P. Liang, S.-W. Wang, Determining the probability of arsenic in groundwater using a parsimonious model. *Environ. Sci. Technol.* **43**, 6662–6668 (2009). doi:10.1021/es900540s Medline

18. M. Currell, I. Cartwright, M. Raveggi, D. Han, Controls on elevated fluoride and arsenic concentrations in groundwater from the Yuncheng Basin, China. *Appl. Geochem.* **26**, 540–552 (2011). doi:10.1016/j.apgeochem.2011.01.012

19. W. M. Edmunds, J. Ma, W. Aeschbach-Hertig, R. Kipfer, D. P. F. Darbyshire, Groundwater recharge history and hydrogeochemical evolution in the Minqin Basin, North West China. *Appl. Geochem.* **21**, 2148–2170 (2006). doi:10.1016/j.apgeochem.2006.07.016

20. B. E. O. Dochartaigh, A. M. MacDonald, *Groundwater Degradation in the Chahaertan Oasis, Alxa League, Inner Mongolia* (British Geological Survey, Nottingham, 2006).

21. G. Yu, D. Sun, Y. Zheng, Health effects of exposure to natural arsenic in groundwater and coal in China: An overview of occurrence. *Environ. Health Perspect.* **115**, 636–642 (2007). doi:10.1289/ehp.9268 Medline

22. G. F. Sun, J. B. Pi, B. Li, X. Y. Guo, H. Yamauchi, T. Yoshida, in *Arsenic Exposure and Health Effects IV*, W. R. Chappell, C. O. Abernathy, R. L. Calderon, Eds. (Elsevier, Amsterdam, 2001), pp. 79–85.

23. 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

24. L. R. Lado, D. Polya, L. Winkel, M. Berg, A. Hegan, Modelling arsenic hazard in Cambodia: A geostatistical approach using ancillary data. *Appl. Geochem.* **23**, 3010–3018 (2008). doi:10.1016/j.apgeochem.2008.06.028

25. 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

26. 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). doi:10.1073/pnas.1011915108 Medline

27. J. Bian, J. Tang, L. Zhang, H. Ma, J. Zhao, Arsenic distribution and geological factors in the western Jilin province, China. *J. Geochem. Explor.* **112**, 347–356 (2012). doi:10.1016/j.gexplo.2011.10.003

28. S. Han, H. Zhang, M. Zhang, Hydrogeological and hydrochemical characterization of shallow high arsenic and deep low arsenic aquifers in Yinchuan Plain: A case study of deep aquifer development for domestic water supply. *Geochim. Cosmochim. Acta* **74**, A378 (2010).

29. M. N. Mead, Arsenic: In search of an antidote to a global poison. *Environ. Health Perspect.* **113**, A378–A386 (2005). doi:10.1289/ehp.113-a378 Medline

30. Y. Xia, J. Liu, An overview on chronic arsenism via drinking water in PR China. *Toxicology* **198**, 25–29 (2004). doi:10.1016/j.tox.2004.01.016 Medline

31. J. Boehner *et al*., in *Soil Classification 2001*, E. Micheli, F. Nachtergaele, L. Montanarella, Eds. (European Soil Bureau, Luxembourg, 2002), pp. 213–222.

32. T. Onishi, M. Yoh, H. Shibata, S. Nagao, M. Kawahigashi, V. Shamov, Topography as a macroscopic index for the dissolved iron productivity of different land cover types in the Amur River Basin. *Hydrol. Res. Lett.* **4**, 85–89 (2010). doi:10.3178/hrl.4.85

33. S. Löhr, M. Grigorescu, J. Hodgkinson, M. Cox, S. Fraser, Iron occurrence in soils and sediments of a coastal catchment: A multivariate approach using self organising maps. *Geoderma* **156**, 253–266 (2010). doi:10.1016/j.geoderma.2010.02.025

34. J. O. Andersson, L. Nyberg, Using official map data on topography, wetlands and vegetation cover for prediction of stream water chemistry in boreal headwater catchments. *Hydrol. Earth Syst.* **13**, 537–549 (2009). doi:10.5194/hess-13-537-2009

35. T. Pei, C.-Z. Qin, A.-X. Zhu, L. Yang, M. Luo, B. Li, C. Zhou, Mapping soil organic matter using the topographic wetness index: A comparative study based on different flow-direction algorithms and kriging methods. *Ecol. Indic.* **10**, 610–619 (2010). doi:10.1016/j.ecolind.2009.10.005

36. 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). doi:10.1016/j.apgeochem.2004.02.001

37. H. A. L. Rowland, R. L. Pederick, D. A. Polya, R. D. Pancost, B. E. Van Dongen, A. G. Gault, D. J. Vaughan, C. Bryant, B. Anderson, J. R. Lloyd, The control of organic matter on microbially mediated iron reduction and arsenic release in shallow alluvial aquifers, Cambodia. *Geobiology* **5**, 281–292 (2007). doi:10.1111/j.1472-4669.2007.00100.x

38. 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

39. O. B. Andersen, P. Knudsen, P. Berry, S. Canyon, The DNSC08 ocean wide altimetry derived gravity field. Presented at EGU-2008, Vienna, April 2008. Data available on request at www.space.dtu.dk/english/Research/Scientific_data_and_models/Global_Marine_Gravity_Field.

40. H. Granser, Three-dimensional interpretation of gravity data from sedimentary basins using an exponential density-depth function. *Geophys. Prospect.* **35**, 1030–1041 (1987). doi:10.1111/j.1365-2478.1987.tb00858.x

41. P. A. Cowie, G. D. Karner, Gravity effect of sediment compaction: Examples from the North Sea and the Rhine Graben. *Earth Planet. Sci. Lett.* **99**, 141–153 (1990). doi:10.1016/0012-821X(90)90078-C

42. A. Chappell, N. Kusznir, An algorithm to calculate the gravity anomaly of sedimentary basins with exponential density-depth relationships. *Geophys. Prospect.* **56**, 249–258 (2008). doi:10.1111/j.1365-2478.2007.00674.x

43. NASA Center for International Earth Science Information Network, *Gridded Population of the World, Version 3* (Socioeconomic Data and Applications Center, Columbia University, Palisades, NY, 2005; http://sedac.ciesin.columbia.edu/gpw).

44. Y. Guo, Y. Zhang, L. Xing, Y. Jia, Spatial variation in arsenic and fluoride concentrations of shallow groundwater from the town of Shahai in the Hetao basin, Inner Mongolia. *Appl. Geochem.* **27**, 2187–2196 (2012). doi:10.1016/j.apgeochem.2012.01.016

45. T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning—Data Mining, Inference, and Prediction* (Springer, New York, ed. 2, 2009).

46. B. Efron, R. J. Tibshirani, *An Introduction to the Bootstrap* (Chapman, New York, 1993).

47. D. W. Hosmer, S. Lemeshow, *Applied Logistic Regression* (Wiley, New York, 2000).

48. F. Shi, thesis, Kungliga Tekniska Högskolan (2004).

49. H. Guo, S. Yang, X. Tang, Y. Li, Z. Shen, Groundwater geochemistry and its implications for arsenic mobilization in shallow aquifers of the Hetao Basin, Inner Mongolia. *Sci. Total Environ.* **393**, 131–144 (2008). doi:10.1016/j.scitotenv.2007.12.025 Medline

50. J. L. Fleiss, *Statistical Methods for Rates and Proportions* (Wiley, New York, ed. 2, 1981).