This is an accepted manuscript version of an article published in:
Winkel, L., Berg, M., Amini, M., Hug, S. J., & Johnson, C. A. (2008). Predicting groundwater arsenic contamination in Southeast Asia from surface parameters. Nature Geoscience, 1, 536-542.
http://doi.org/10.1038/ngeo254

Predicting groundwater arsenic contamination in Southeast Asia from surface parameters

Lenny Winkel, Michael Berg*, Manouchehr Amini, Stephan J. Hug, C. Annette Johnson

Eawag, Swiss Federal Institute of Aquatic Science and Technology, 8600 Dübendorf, Switzerland.

*** Corresponding author. e-mail: michael.berg@eawag.ch**
Phone: +41-44-823 50 78; Fax: +41-44-823 50 28

2^{nd} revision submitted to *Nature Geoscience*
(Manuscript NGS-2008-01-00049A)

June 16, 2008

Arsenic contamination of groundwater resources threatens the health of millions of people worldwide, particularly in the densely populated river deltas of Southeast Asia. Although many arsenic-affected areas have been identified in recent years, a systematic evaluation of vulnerable areas remains to be carried out. Here we present maps pinpointing areas at risk of groundwater arsenic concentrations exceeding 10 μgL^{-1}. These maps were produced by combining geological and surface soil parameters in a logistic regression model, calibrated with 1756 aggregated and geo-referenced groundwater data points from the Bengal, Red River and Mekong deltas. We show that Holocene deltaic and organic-rich surface sediments are key indicators for arsenic risk areas and that the combination of surface parameters is a successful approach to predict groundwater arsenic contamination. Predictions are in good agreement with the known spatial distribution of arsenic contamination and further indicate elevated risks in Sumatra and Myanmar where no groundwater studies exist.

More than 100 million people worldwide ingest excessive amounts of arsenic (As) through drinking water contaminated from natural geogenic sources. Many Asian countries, in particular, are known to be affected by high groundwater As concentrations as a result of chemically reducing aquifer conditions: Bangladesh[1-10], India[3,11,12], China[13,14], Nepal[15], Cambodia[16-18] and Vietnam[17,19,20]. However, since As analysis is expensive and time-consuming, groundwater resources of many regions still remain to be tested. Therefore, maps pinpointing areas vulnerable to As contamination can guide households at risk of arsenic contamination, as well as scientists and policy-makers to initiate early mitigation measures and protect the populations from chronic As poisoning.

Though the exact chemical conditions and reactions leading to As mobilization are still under debate, it is generally accepted that microbial and/or chemical reductive dissolution of As-bearing iron minerals in the aquifer sediments[1,4,21] is the main cause for the release of As. Reducing conditions are often associated with the presence of natural (bio)degradable organic carbon embedded in sediments[9,11,22-25]. Other identified key characteristics of contaminated areas are rapidly buried, young (Holocene) sediments and low hydraulic gradients in flat and low-lying areas[8,18,26,27]. Ideally an As prediction model for groundwater should be based on parameters that indicate the key characteristics mentioned above in three dimensions. However, in the absence of a 3-dimensional spatially continuous database of aquifer conditions to depth, globally and regionally available (two dimensional) surface parameters can be used as indicators for As-enrichment in underlying aquifers[28,29].

In the past, several geostatistical interpolation methods (e.g. kriging) have been used to predict elevated As in groundwater on a regional scale[30-32]. However, for predictions of areas where no groundwater quality data exist, interpolation methods are not applicable[32] and models based on logistic regression are more appropriate[33,34]. In an expert-based statistical model to delineate areas at risk of groundwater As contamination on a coarse global scale, we found that geological information was of crucial importance[29]. Here we focus on an in-depth assessment of depositional environments in Southeast Asia. We use a logistic regression approach based on relationships between sedimentary information, soil maps and measured groundwater As data of Bangladesh, Cambodia and Vietnam, to assess the relative importance of the different surface proxies in these countries. We apply these relationships to set up prediction maps of As contamination in Southeast Asia including Indonesia (Sumatra) and other countries where groundwater quality data is scarce (Myanmar and Thailand). Furthermore, we verify the predicted risk in South Sumatra, where the groundwater has not previously been tested for the presence of As.

ARSENIC PREDICTION MODEL

Our model is based on three assumptions. First, sedimentary depositional environments are characterized by a unique combination of chemical, physical, and biological properties[35] and can serve as indicators (proxies) for chemical and physical conditions of the aquifers beneath the surface. Second, soil properties are proxies for present and past drainage conditions and they are also indicators of recent depositional environments. Third, soil textures, for example clay and silt are proxies for the chemical maturity of the sediments, where clay is more mature than silt. An important factor in the development of soil textures is topography[36], which allows the delineation of areas where the model is applicable.

GIS-datasets were established from digital elevation data, countrywide geological maps and global soil data (FAO), which were converted to a raster format using ArcGIS (ver. 9.2). An overview of GIS data used in this study is provided in Table S1 (see Supplementary Information). Because each geological map applied a different classification terminology, we created an uniformly classified geological map for all regions (Figure 1). Although Bangladesh does geographically not belong to Southeast Asia, it was included in the model because of the large number of available data points[5]. Statistical relations between As concentrations and 30 parameters related to soil properties, geology, climate, and hydrology (Table S2) were initially evaluated by stepwise regression. The six parameters exhibiting a significance >95% and two additional soil parameters were employed in the final model (see Methods and Table 1). Since young geological deposits and As groundwater contamination are rarely observed in areas with steeper slopes, and groundwater As concentration data are only available for regions with a flat topography (slope $<0.1° \approx 0.17\%$), areas with slopes >0.1° were excluded from the model (see Supplementary Figure S1).

The prediction required three steps: i) Aggregation and binary-coding of measured As concentrations to reduce spatial heterogeneities (dependent variables), ii) Logistic regression to obtain weighting coefficients of independent variables (see Methods section and corresponding results in Table 1), and, iii) Calculation of the probability of As contamination based on the threshold value of 10 $\mu g L^{-1}$. The spatial datasets considered as independent variables for the model are topographic data to delineate the model area, sedimentary depositional environments as a proxy for aquifer conditions, and soil variables as a proxy for drainage and chemical maturity of sediments (see the Methods section).

SURFACE PARAMETERS CONTRIBUTING TO THE MODEL

The weighting factors (λ) and significance of the eight independent variables retained in the final model are listed in Table 1. In general, variables describing sedimentary depositional environments have a larger contribution to the model than soil variables, presumably because geological variables come closest to describing As contamination in the aquifer itself. The presence of young deltaic deposits (λ=1.65) is a particularly significant indicator for As-contaminated aquifers. In Southeast Asia, delta initiation and progradation occurred simultaneously with the Holocene Climate Optimum[37]. Therefore, delta progradation resulted in the burial of organic matter at a high rate. The presence of relatively fresh organic carbon provides favourable conditions to establish reducing environments, which may lead to As enrichment in groundwater. Logistic regression confirms that organic-rich deposits (λ=1.11) play an important part in the model. Recent alluvial deposits (λ=0.55) are also indicative for elevated As concentrations in groundwaters (Table 1). Of the soil parameters, medium textured soils seem to be indicative of As-bearing aquifers evolved from rapid accumulation of young (Holocene) sediments. In contrast, the negative weighting coefficient (λ= -0.95) for floodplain deposits implies that these fine-grained deposits could overlie aquifers low in dissolved As (compare Figures 1 and 3). Fine-grained deposits (high clay content) thereby point to low-energy depositional environments with condensed sediments where Holocene aquifers are rare and where groundwater is likely drawn from older aquifers.

Pre-Holocene deposits, other Holocene deposits and tidal deposits (Figure 1) were found to be statistically insignificant ($p > 0.05$) and they were excluded from the model in the first stepwise regression (see Table S2 and Table 1). Tidal sediments are generally associated with aquifers abundant in sulfate that may be microbially reduced to sulfide and re-precipitate As[38,39]. They might serve as proxies for low-risk areas, but our As data in such aquifers were possibly too few to see such a relation.

INFERRING DEPOSITIONAL ENVIRONMENTS AT DEPTH FROM THE SURFACE

As mentioned in the introduction, our model is based on two- dimensional data (i.e. surface maps). Nevertheless, geological information (sedimentary depositional environments) inherently contains a three-dimensional component. The recent sedimentary history of major Southeast Asian basins is characterized by delta initiation, which occurred on a global scale at about 8500-6500 years BP and was principally controlled by the deceleration of sea-level rise[40]. Delta progradation resulted in the unconformable deposition of thick late Pleistocene-Holocene sediments on Pleistocene and older sediments, e.g. as incised-valley-fill deposits[8,18]. As-contaminated aquifers are mainly present in these Holocene aquifers, whereas deeper Pliocene-Pleistocene aquifers are, to a large extent, free of As[7,8,26]. The boundary between Pleistocene and Holocene sediments is not located at a constant depth[8] and this can lead to misclassifications, since our model inherently assumes that the underlying aquifer belongs to the same sedimentary depositional environment as on the surface.

Two situations exist where the environment at the surface does not reflect the geology at depth, i) when the As measurements used in the model were obtained from tubewells tapping deep (Pleistocene) aquifers because shallow (Holocene) aquifers are not present, or shallow groundwater is too saline for consumption (e.g. in coastal regions), and ii) when Holocene sediments were deposited at sedimentation rates to small to form a usable aquifer. In both situations Holocene depositional environments are present at the surface and high-risk areas are indicated, although measured As concentrations are low (false positive cases). Even though the probability maps would be improved by including the thickness of Holocene sediments, the absence of country-wide three-dimensional geological data rules this out. Furthermore, the complexity of aquifer heterogeneity at a local scale makes it inevitable that misclassifications occur.

PROBABILITY MAPS

Predicted areas at risk of As contamination agree well with known spatial contamination patterns, a finding which is supported by the results of the model classification indicating the performance of model prediction (Figure 2) and the Hosmer-Lemeshow goodness-of-fit statistics (Table S3). An absolute average deviation of 7.3% is found between expected and modelled probabilities of As being ≤10 or >10 μgL^{-1}. The model is further characterised by the receiver operating characteristics curve area of ~0.7 (Figure S2), which is a good result considering that neither depths of analyzed groundwater wells, nor aquifer hydrological data form part of the model.

The probability maps for As concentrations exceeding 10 μgL^{-1} are presented in Figure 3 and supplementary Figure S3 (probability map of whole Southeast Asia). The highest probabilities (0.7-0.8) for As contamination (>10 μgL^{-1}) are found in the south-central part of Bangladesh, with a value of 0.5-0.6 in the north-eastern Sylhet basin (Figure 3a). The probability of finding contaminated wells in the Red River delta reaches a value of 0.7 (Figure 3e). The sedimentary depositional environments present along the Mekong river differ from the deltaic environments of Bangladesh and the Red River delta in that organic-rich deposits are found at close distance of the modern Mekong and Bassac river courses, surrounded by extensive floodplain areas free of As[18] (compare Figures 1 and 3). The probability map for the Mekong delta shows values of up to 0.6 at close distance to the modern Mekong and Bassac river courses and adjacent swampy marshes (Figure 3c). In addition, the large floodplain of Lake Tonle Sap with organic-rich sediments is a risk area with probabilities ranging between 0.4-0.6.

To compare our predicted low-risk (probability ≤0.4) and high-risk areas (probability >0.4) (Figures 3b and 3d) with measured As contamination, a misclassification analysis was performed based on the 1756 aggregated and binary-coded (≤10 or >10 μgL^{-1}) As measurement data for the three deltas. We report 71% (1023 aggregated points), 59% (107 a.p.), and 75% (100 a.p.) of correctly classified cases for Bangladesh, Red River and Mekong deltas respectively (average 70%). In our study, the Red River and Mekong deltas have 6 and 9% false negative cases. In Bangladesh false negative cases (13%) were found specifically in wells lying at a close distance to rivers. The number of false negatives is outnumbered by false positives in all the three deltas with 17, 35 and 16% for Bangladesh, the Red River and Mekong deltas, respectively. Errors in the prediction can arise from the commonly reported well-to-well variability where wells with low As levels are often present at close distance from wells high in arsenic[5,6,18,19], as well as from uncertainties in measured As (estimated at 25%) leading to misclassifications of concentrations being close to the threshold of 10 μgL^{-1}. However, we interpret these misclassifications to be mainly an effect of modelling three-dimensional processes based on two-dimensional data.

Apart from the three deltas discussed above, our Southeast Asia probability map (supplementary Figure S3) also highlights risk areas that are largely unknown or unreported, particularly in Sumatra, Myanmar and Thailand (Figure 3f).

VERIFICATION OF PREDICTIONS FOR SUMATRA

According to the modelled probability, an area of about 100,000 km^2 at Sumatra's east coast is prone to high risk of As contamination (probability >0.4) (see Figure 4a and supplementary Figure S3). To validate the Sumatra prediction map, 97 groundwater samples were collected in 2007 in the Province of South Sumatra in the vicinity of Palembang (see Methods section). This area was chosen because it is at the border of a low- and high-risk area and not previously studied. Since there is no large variation in geological and topographical features along Sumatra's east coast, the study area is representative of the whole high risk area. Figure 4b shows As concentrations ($\leq 10\ \mu gL^{-1}$ and $>10\ \mu gL^{-1}$) measured in the groundwater, imposed on the binary probability map. The classification results for Sumatra (63% correctly classified, 36% false positive and 1% false negative) are comparable to those of the Bangladesh, Red River, and Mekong deltas. In total, 94% of the 50 tubewells located in the low-risk area have As levels below 10 μgL^{-1}. Of the 12 contaminated wells (As $>10\ \mu gL^{-1}$) 75% are positioned in the high-risk area. However, in the high-risk area the contaminated wells are clearly outnumbered by uncontaminated groundwater measurements (9 contaminated wells vs. 38 uncontaminated wells), for reasons explained below.

On average, both high and low risk areas in Sumatra are characterized by high DOC, NH_4^+, HCO_3^-, PO_4^{3-}, and Fe(II) and low SO_4^{2-} concentrations (Supplementary Table S5). At least two-thirds of the sampled groundwaters are reducing in nature, especially those located in the high-risk area. Since the chemical conditions of the aquifers would permit the reductive dissolution of As, other explanations for the overall low As concentrations must be considered.

To gain a greater insight into the characteristics of the aquifer, the local geology was examined laterally and in depth. The studied area is geologically young (Tertiary-Quaternary) (Figure 4c). The sediment contains deposits from peat swamp forests that developed during the Holocene era (past 5,000 to 10,000 years) and unconformably overlie older sediments[41]. The outline of the high-risk As-contamination zone mainly follows the outline of these deposits. The peat deposits are usually found ranging from 4-8 meters although depths of up to 24 metres have been reported[42]. However, the depths of sampled tubewells in Sumatra average 46 meters, which implies that most of them tap groundwaters from aquifers below the Holocene peat deposits. This shows that the prediction map is a useful tool for the identification of areas at risk of As contamination, but that understanding the local geology as a function of depth is of vital importance for specific areas.

ARSENIC CONTAMINATION IN THAILAND AND MYANMAR

The probability map in Figure 3f shows that the Chao Phraya basin in central Thailand and the Irrawaddy delta in Myanmar have a risk of elevated As being present in groundwater, whereas the Sittang basin (Myanmar) has a lower risk, and the Salween basin (Myanmar) has virtually no risk at all. The problem of As contamination in the Irrawaddy delta is partly known[43] although its spatial extent has not been investigated to date, which is particularly worrying considering the size of the area at risk. In the Chao Phraya basin in central Thailand, a groundwater survey was undertaken in 2001 using As field test kits to test wells with minimum depths of 80 m[44]. The range of measured As concentrations in the 37 tested wells was <1 μgL^{-1} to 100 μgL^{-1} with an average of 11 μgL^{-1}. These results correlate with our maps that predict a low to moderate As contamination. Indeed, North-South geological profiles across the basin[45] indicate maximal depths of 20 m for the pre-Holocene incisions implying that sampled tubewells draw water from Pleistocene or older aquifers.

In contrast to the slow sedimentation rates of the Sittang and Chao Phraya deltas, the Irrawaddy delta received massive amounts of sediments during the Cenozoic era[46]. The Irrawaddy River[47] still has a 10 times larger sediment load than the Chao Phraya River[45]. In 2002 the Departments of Medical Research and Health (Lower Myanmar), financially supported by UNICEF, conducted a groundwater sampling campaign in the Irrawaddy division[43]. In total, 99 groundwater samples (90 shallow tubewells and 9 dug wells) were collected in 25 villages, and As was quatified by atomic absorption spectroscopy. It was reported that 67% of the sampled wells had As levels >50 μgL^{-1}. These results show that the risk of As contamination in the Irrawaddy delta, as indicated by the probability map, should be taken seriously and that there is an urgent need to test shallow tubewells in other townships or districts in the high risk area.

ASSESSING ARSENIC ELSEWHERE

In the study presented here, we identified regions in Southeast Asia, based on As prediction maps, where tubewells should be tested for elevated As concentrations (>10 μgL^{-1}). Although the use of such maps in other scientific investigations (e.g. climate research) is a common practice, the prediction of As (and other groundwater conditions) is still an emerging technique in the field of natural groundwater contaminants. Our As model differs from earlier models in its ability to predict contamination in areas of unknown groundwater quality and on a sub-continental scale. The strength of the prediction lies rather in the combination of surface parameters than in the individual parameters. As an example, the Sittang basin and the vicinity of the lower Mekong river branches are both characterized by alluvial deposits, but only the latter has a modelled high-risk because of the contribution of soil properties. A limitation to be considered is that shallow, As-bearing groundwater can only be expected where Holocene aquifers are present. Where this is not the case, high risk areas may indicate the presence of reducing aquifers, but As concentrations in groundwater could be lower than predicted.

Our approach provides a blueprint for further modelling and mapping of As-tainted aquifers in and outside of Southeast Asia. The probability maps can further be improved when data with higher spatial resolution or in three dimensions becomes available, although we emphasize that it will not be possible to account for the local heterogeneities of aquifers. The presented prediction maps are a valuable and resource-saving tool that can serve both scientists and policy-makers to initiate early mitigation measures in order to protect the people from As-related health problems as well as to efficiently guide water resources management.

METHODS

Geological maps of Bangladesh, Cambodia, Thailand and Vietnam were available in digital format (Supplementary Table S1). Maps of Myanmar and Sumatra were digitized for this study. Geological maps of Malaysia and Laos were not available. The sedimentary depositional environments employed in the model as independent variables are deltaic deposits, organic-rich sediments (e.g. sediments deposited in marshy environments), floodplains, alluvial deposits, tidal deposits, other Holocene sediments and pre-Holocene sediments. Soil variables are percentages of silt, clay and sand in both the topsoil (0-30 cm) and subsoil (30-100 cm), and coarse, medium and fine soil textures.

We compiled >4600 data points of groundwater As-concentrations from Bangladesh (median well depth 35 m), the Mekong delta (Cambodia and Southern Vietnam, m.w.d. 39 m) and the Red River delta (Northern Vietnam, m.w.d. 30 m). The data originates from BGS and DPHE (Bangladesh, n=3534)[5], from Buschmann et al. (Mekong delta, n=352)[18,20], and from Berg et al. (Red River delta, n=720)[17,25,48] and was used without a restriction in well depths. To test the model, 97 tubewell were randomly sampled for this study in Sumatra at about 5-10 km intervals at an average sampling density of 1 sample per 54 km^2. This study area is positioned at a latitude of 2°872S to 3°911S and a longitude of 103°949E to 104°993E (sampling area 85 km by 65 km). Procedures of sampling and analysis were carried out as described in Buschmann et al.[18]. Concentrations of As and additional parameters measured in Sumatra groundwater samples are provided in Table S5 (Supplementary Information).

Arsenic concentrations are point measurements within vertical depths of the wells, while the other variables have coarser spatial resolution being generally greater than 30 arc seconds. Point data of measured As-concentrations were therefore aggregated using the geometric mean to a resolution of one point per pixel with a size of 5 arc minutes (~9.3 km at the equator), which is the pixel size of the global soil data (FAO). Aggregation resulted in a decreased dataset of 1756 pixel-based data points (Bangladesh 1443, Red River Delta 180, and Mekong delta 133 points). The aggregated point-data were binary-coded using the WHO guideline value for As in drinking water (10 μgL^{-1}) as a threshold. We acknowledge that several countries apply a guideline of 50 μgL^{-1}, but adopting this threshold would result in significantly less data points (992) for the calibration of the model. The binary variable of whether As concentrations exceed the WHO threshold (1) or not (0) was used as the dependent variable in this study.

Logistic regression was used to determine the weighting of the independent variables. This is a common statistical method used in environmental research and allows for the concurrent use of continuous and categorical variables[33,34]. The parameter P denotes the probability of As concentrations exceeding the WHO threshold. Logistic regression models log(odds), which is defined as the ratio of the probability that an event occurs to the probability that it fails to occur log(P/(1−P)), as a linear combination of independent variables[49].

$$\log(odds) = C + \sum_{i=1}^{n} \lambda_i X_i \qquad (1)$$

where C is the intercept of regression, X_i are independent variables, and λ_i are the weighting coefficients that were obtained using the maximum likelihood procedure[49]. Exponential values of coefficients, Wald statistics, and p-values (see Table 1) indicate the importance of each variable. Statistically insignificant independent variables were excluded from the model during each of the subsequent regression steps (Table S2). The threshold for maintaining a variable in the model was determined by the 95% significance level ($p < 0.05$). The silt contents in the subsoil and medium textured soils were kept in the model because of their good spatial match with known contaminated areas, which is supported by the presence of silty sands at the surface of regions exhibiting contaminated aquifers in West Bengal[50], and by reported elevated groundwater As concentrations in aquifers capped with fine surface material (silt and clay) in Bangladesh[10,27].

According to the calculated odds, the probability (P) of having an As concentration above 10 μgL^{-1} was calculated as follows:

$$P = \frac{\exp(C + \sum_{i=1}^{n} \lambda_i X_i)}{1 + \exp(C + \sum_{i=1}^{n} \lambda_i X_i)} \qquad (2)$$

In addition, we tested how successful the model predicted the number of contaminated cases for probability intervals of 0.1 (Figure 2 and Supplementary Information).

Misclassification occurs when either a point with As concentration >10 μgL^{-1} falls in an uncontaminated area (false negative) or an As concentration <10 μgL^{-1} falls in a contaminated area (false positive). Based on the model classification results (Figure 2 and Table S4), a probability threshold of 0.4 resulted in a minimum misclassification rate. The binary risk maps was hence categorized into low-risk areas (probability <0.4) and high risk areas (probability >0.4).

REFERENCES

1. Nickson, R. *et al.* Arsenic poisoning of Bangladesh groundwater. *Nature* **395**, 338 (1998).
2. Smith, A. H., Lingas, E. O. & Rahman, M. Contamination of drinking water by arsenic in Bangladesh: a public health emergency. *Bull. World Health Org.* **78**, 1093–1102 (2000).
3. Chowdhury, U. K. *et al.* Groundwater arsenic contamination in Bangladesh and West Bengal, India. *Environ. Health Perspect.* **108**, 393-397 (2000).
4. McArthur, J. M., Ravenscroft, P., Safiulla, S. & Thirlwall, M. F. Arsenic in groundwater: Testing pollution mechanisms for sedimentary aquifers in Bangladesh. *Water Resour. Res.* **37**, 109-117 (2001).

5. BGS and DPHE. *Arsenic contamination of groundwater in Bangladesh*. Eds. Kinniburgh, D. G. & Smedley, P. L. (British Geological Survey, Keyworth, U.K., 2001). www.bgs.ac.uk/arsenic/bangladesh.
6. van Geen, A. *et al.* Spatial variability of arsenic in 6000 tube wells in a 25 km(2) area of Bangladesh. *Water Resour. Res.* **39**(2003).
7. Ahmed, K. M. *et al.* Arsenic enrichment in groundwater of the alluvial aquifers in Bangladesh: an overview. *Appl. Geochem.* **19**, 181-200 (2004).
8. Ravenscroft, P., Burgess, W. G., Ahmed, K. M., Burren, M. & Perrin, J. Arsenic in groundwater of the Bengal Basin, Bangladesh: Distribution, field relations, and hydrogeological setting. *Hydrogeol. J.* **13**, 727-751 (2005).
9. Meharg, A. A. *et al.* Codeposition of organic carbon and arsenic in Bengal Delta aquifers. *Environ. Sci. Technol.* **40**, 4928-4935 (2006).
10. van Geen, A. *et al.* Flushing history as a hydrogeological control on the regional distribution of arsenic in shallow groundwater of the Bengal Basin. *Environ. Sci. Technol.* **42**, 2283-2288 (2008).
11. McArthur, J. M. *et al.* 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).
12. Ahamed, S. *et al.* Arsenic groundwater contamination and its health effects in the state of Uttar Pradesh (UP) in upper and middle Ganga plain, India: A severe danger. *Sci. Total Environ.* **370**, 310-322 (2006).
13. Smedley, P. L., Zhang, M., Zhang, G. & Luo, Z. Mobilisation of arsenic and other trace elements in fluviolacustrine aquifers of the Huhhot Basin, Inner Mongolia. *Appl. Geochem.* **18**, 1453-1477 (2003).
14. Yu, G. Q., Sun, D. J. & Zheng, Y. Health effects of exposure to natural arsenic in groundwater and coal in China: An overview of occurrence. *Environ. Health Perspect.* **115**, 636-642 (2007).
15. Shrestha, R. R. *et al.* Groundwater arsenic contamination, its health impact and mitigation program in Nepal. *Environ. Sci. Health, Part A* **38**, 185-200 (2003).
16. Polya, D. A. *et al.* Arsenic hazard in shallow Cambodian groundwaters. *Mineral. Mag.* **69**, 807-823 (2005).
17. Berg, M. *et al.* Magnitude of arsenic pollution in the Mekong and Red River Deltas - Cambodia and Vietnam. *Sci. Total Environ.* **372**, 413-425 (2007).

18. Buschmann, J., Berg, M., Stengel, C. & Sampson, M. L. Arsenic and Manganese Contamination of Drinking Water Resources in Cambodia: Coincidence of Risk Areas with Low Relief Topography. *Environ. Sci. Technol.* **41**, 2146–2152 (2007).
19. Berg, M. *et al.* Arsenic contamination of groundwater and drinking water in Vietnam: A human health threat. *Environ. Sci. Technol.* **35**, 2621-2626 (2001).
20. Buschmann, J. *et al.* Contamination of drinking water resources in the Mekong delta floodplains: Arsenic and other trace metals pose serious health risks to population. *Environ. Int.* **34**, doi:10.1016/j.envint.2007.12.025 (2008).
21. Ford, R. G., Fendorf, S. & Wilkin, R. T. Introduction: Controls on arsenic transport in near-surface aquatic systems. *Chem. Geol.* **228**, 1-5 (2006).
22. Harvey, C. F. *et al.* Arsenic mobility and groundwater extraction in Bangladesh. *Science* **298**, 1602-1606 (2002).
23. Islam, F. S. *et al.* Role of metal-reducing bacteria in arsenic release from Bengal delta sediments. *Nature* **430**, 68-71 (2004).
24. Rowland, H. A. L. *et al.* The control of organic matter on microbially mediated iron reduction and arsenic release in shallow alluvial aquifers, Cambodia. *Geobiology* **5**, 281-292 (2007).
25. Berg, M. *et al.* 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).
26. Smedley, P. L. & Kinniburgh, D. G. A review of the source, behaviour and distribution of arsenic in natural waters. *Appl. Geochem.* **17**, 517-568 (2002).
27. Stute, M. *et al.* Hydrological control of As concentrations in Bangladesh groundwater. *Water Resour. Res.* **43**(2007).
28. Twarakavi, N. K. C. & Kaluarachchi, J. J. Arsenic in the shallow ground waters of conterminous United States: Assessment, health risks, and costs for MCL compliance. *J. Am. Water Resour. Assoc.* **42**, 275-294 (2006).
29. Amini, M. *et al.* Statistical modeling of global geogenic arsenic contamination in groundwater. *Environ. Sci. Technol.* **42**, 3669-3675 (2008).
30. Goovaerts, P. *et al.* Geostatistical modeling of the spatial variability of arsenic in groundwater of southeast Michigan. *Water Resour. Res.* **41**(2005).

31. Lee, J. J., Jang, C. S., Wang, S. W. & Liu, C. W. Evaluation of potential health risk of arsenic-affected groundwater using indicator kriging and dose response model. *Sci. Total Environ.* **384**, 151-162 (2007).

32. Hossain, F., Hill, J. & Bagtzoglou, A. C. Geostatistically based management of arsenic contaminated ground water in shallow wells of Bangladesh. *Water Resour. Manag.* **21**, 1245-1261 (2007).

33. Twarakavi, N. K. C. & Kaluarachchi, J. J. Aquifer vulnerability assessment to heavy metals using ordinal logistic regression. *Ground Water* **43**, 200-214 (2005).

34. Ayotte, J. D. *et al.* Modeling the probability of arsenic in groundwater in New England as a tool for exposure assessment. *Environ. Sci. Technol.* **40**, 3578-3585 (2006).

35. Reading, H. G. (ed.) *Sedimentary Environments: Processes, Facies and Stratigraphy*, thrid edition (Blackwell Science, Oxford, 1996).

36. Tan, K. H. *Environmental Soil Science*, second edition (M. Dekker, New York, 2000).

37. Hori, K. *et al.* Delta initiation and Holocene sea-level change: example from the Song Hong (Red River) delta, Vietnam. *Sediment. Geol.* **164**, 237-249 (2004).

38. Kirk, M. F. *et al.* Bacterial sulfate reduction limits natural arsenic contamination in groundwater. *Geology* **32**, 953-956 (2004).

39. Lowers, H. A. *et al.* Arsenic incorporation into authigenic pyrite, bengal basin sediment, Bangladesh. *Geochim. Cosmochim. Acta* **71**, 2699-2717 (2007).

40. Stanley, D. J. & Warne, A. G. Worldwide Initiation of Holocene Marine Deltas by Deceleration of Sea-Level Rise. *Science* **265**, 228-231 (1994).

41. Wosten, J. H. M. *et al.* Interrelationships between hydrology and ecology in fire degraded tropical peat swamp forests. *Int. J. Water Resour. Dev.* **22**, 157-174 (2006).

42. Giesen, W. *Causes of Peatswamp Forest Degradation in Berbak National Park and Recommendations for Restoration, Water for Food & Ecosystems Programme*. (Arcadis Euroconsult, Arnhem, Holland, 2004).

43. Tun, K. M. A. *Report on the Assessment of Arsenic Content in Groundwater and the Prevalence of Arsenicosis in Thabaung and Kyonpyaw Townships, Ayeyarwaddy Division*. (Department of Medical Research, Yangoon, Myanmar, 2002).

44. Kohnhorst, A. Arsenic in groundwater in selected countries in south and southeast Asia: A review. *J. Trop. Med. Parasitol.* **28**, 73-82 (2005).

45. Tanabe, S. *et al.* Stratigraphy and Holocene evolution of the mud-dominated Chao Phraya delta, Thailand. *Quat. Sci. Rev.* **22**, 789-807 (2003).

46. Metivier, F., Gaudemer, Y., Tapponnier, P. & Klein, M. Mass accumulation rates in Asia during the Cenozoic. *Geophys. J. Int.* **137**, 280-318 (1999).

47. Robinson, R. A. J. *et al.* The Irrawaddy River sediment flux to the Indian Ocean: The original nineteenth-century data revisited. *J. Geol.* **115**, 629-640 (2007).

48. Berg, M. *et al.* Arsenic removal from groundwater by household sand filters: Comparative field study, model calculations, and health benefits. *Environ. Sci. Technol.* **40**, 5567-5573 (2006).

49. Kleinbaum, D. G. & Klein, M. *Logistic Regression: A Self-Learning Text*, second edition (Springer-Verlag, New York, 2002).

50. Pal, T., Mukherjee, P. K., Sengupta, S., Bhattacharyya, A. K. & Shome, S. Arsenic pollution in groundwater of West Bengal, India - An insight into the problem by subsurface sediment analysis. *Gondwana Res.* **5**, 501-512 (2002).

Correspondence and requests for materials should be addressed to Michael Berg.

ACKNOWLEDGEMENTS

We thank Thomas Rosenberg, Rudi Febriamansyah, M.A. Hayatuddin, and Erwin Nofyan for support during groundwater sampling in Sumatra; Caroline Stengel, Thomas Rüttimann, Madeleine Langmeier and Richard Illi for elemental analyses; Karim Abbaspour and Bas den Brok for helpful discussions; Rich Wildman and Helen Rowland for proofreading, and the anonymous reviewers for constructive comments.

Supplementary Information accompanies this paper

Competing financial interests statement

The authors declare that they have no competing financial interests.

Figure captions

Figure 1. Uniformly classified geological map of Southeast Asia. It indicates seven different sedimentary depositional environments in the mapped countries of Bangladesh, Myanmar, Thailand, Cambodia, Vietnam, and Sumatra (Indonesia).

Figure 2. Model classification results. The graph shows the sensitivity (true positives) and specificity (true negatives) of the model for different probability cutoff values. The full classification table is provided in the Supplementary Information (Table S4). A probability threshold of 0.4 was applied to delineate low- and high-risk areas in the binary risk maps shown in Figures 3b, 3d, and 4b (see Methods).

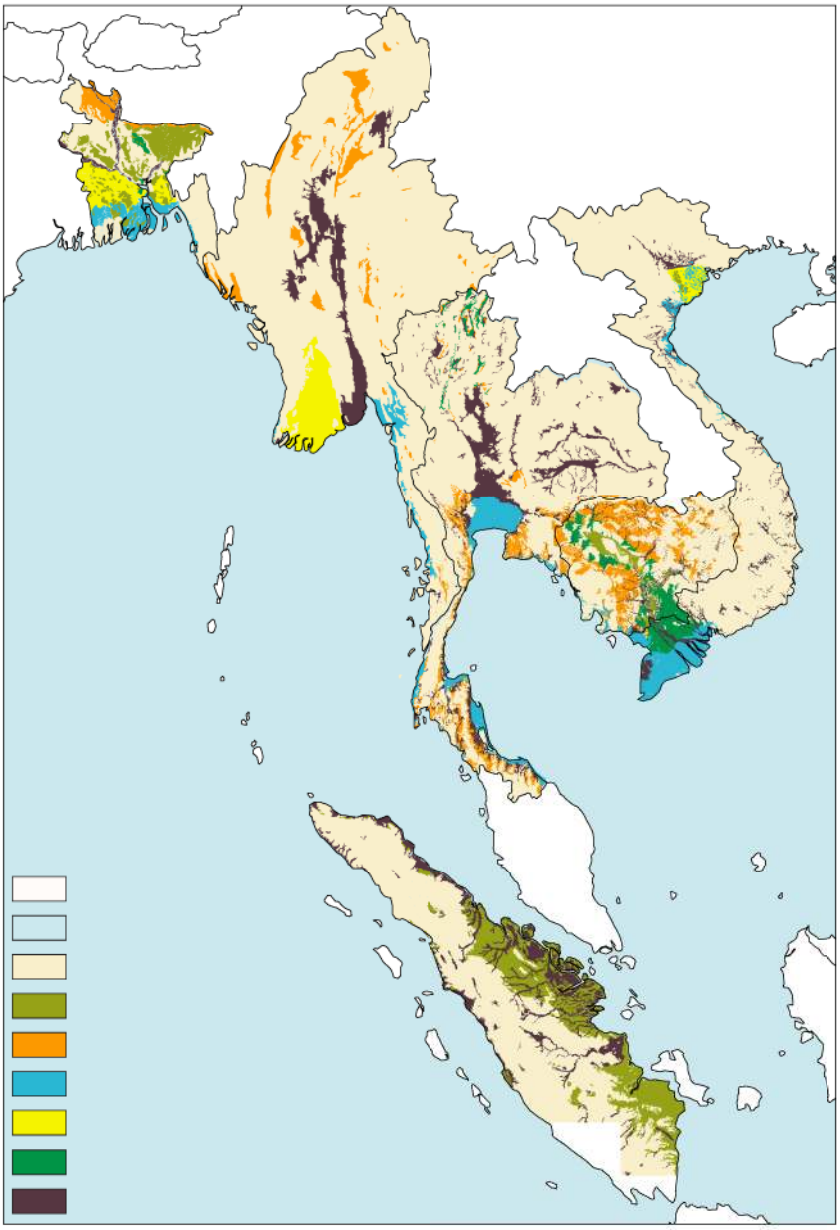
Figure 3. Modelled probability of As concentrations exceeding 10 μgL^{-1} under reducing aquifer conditions. a, Continuous probability map of Bangladesh. **b**, Binary map of Bangladesh (probability threshold 0.4) indicating high- and low-risk areas overlain by aggregated As concentrations. Areas where groundwater is mainly drawn from Pleistocene aquifers are sketched in brown. **c**, Continuous probability map of the Mekong delta (Cambodia and Vietnam). **d**, Binary risk map of the Mekong delta overlain by aggregated As concentrations. **e**, Continuous probability map of the Red River delta (Vietnam). **f**, Continuous probability map of the Irrawaddy delta (Myanmar) and Chao Phraya basin (Thailand).

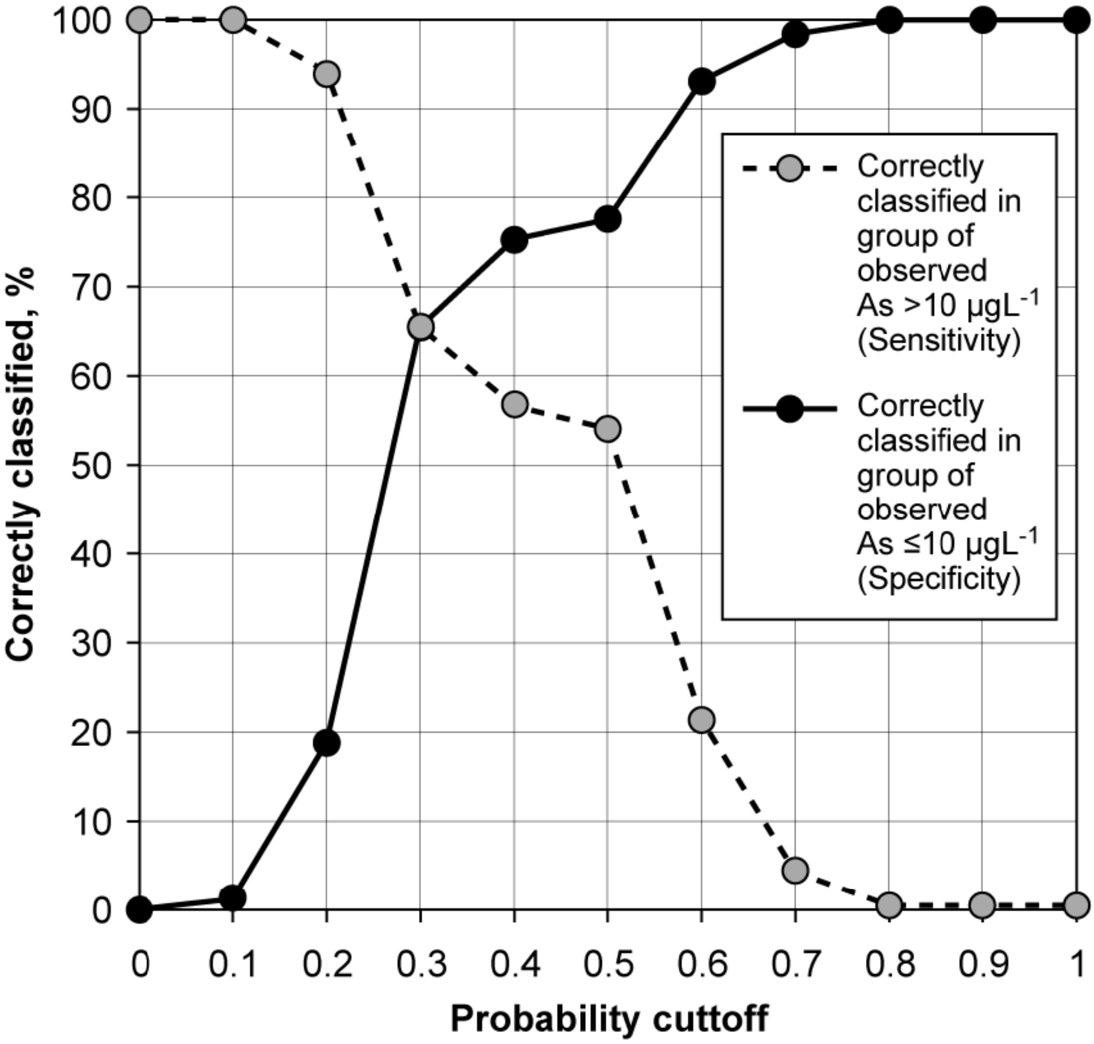
Figure 4. Maps of the model verification study area in Southeast Sumatra. a, Map of whole Sumatra depicting the probability of groundwater As concentrations exceeding 10 μg L^{-1}. The corresponding colour code is given in Figure 3. **b**, Binary risk map (probability cuttoff 0.4) and As concentrations measured for this study in the vicinity of Palembang (South Sumatra). Swampy areas (high-risk area) are scarcely

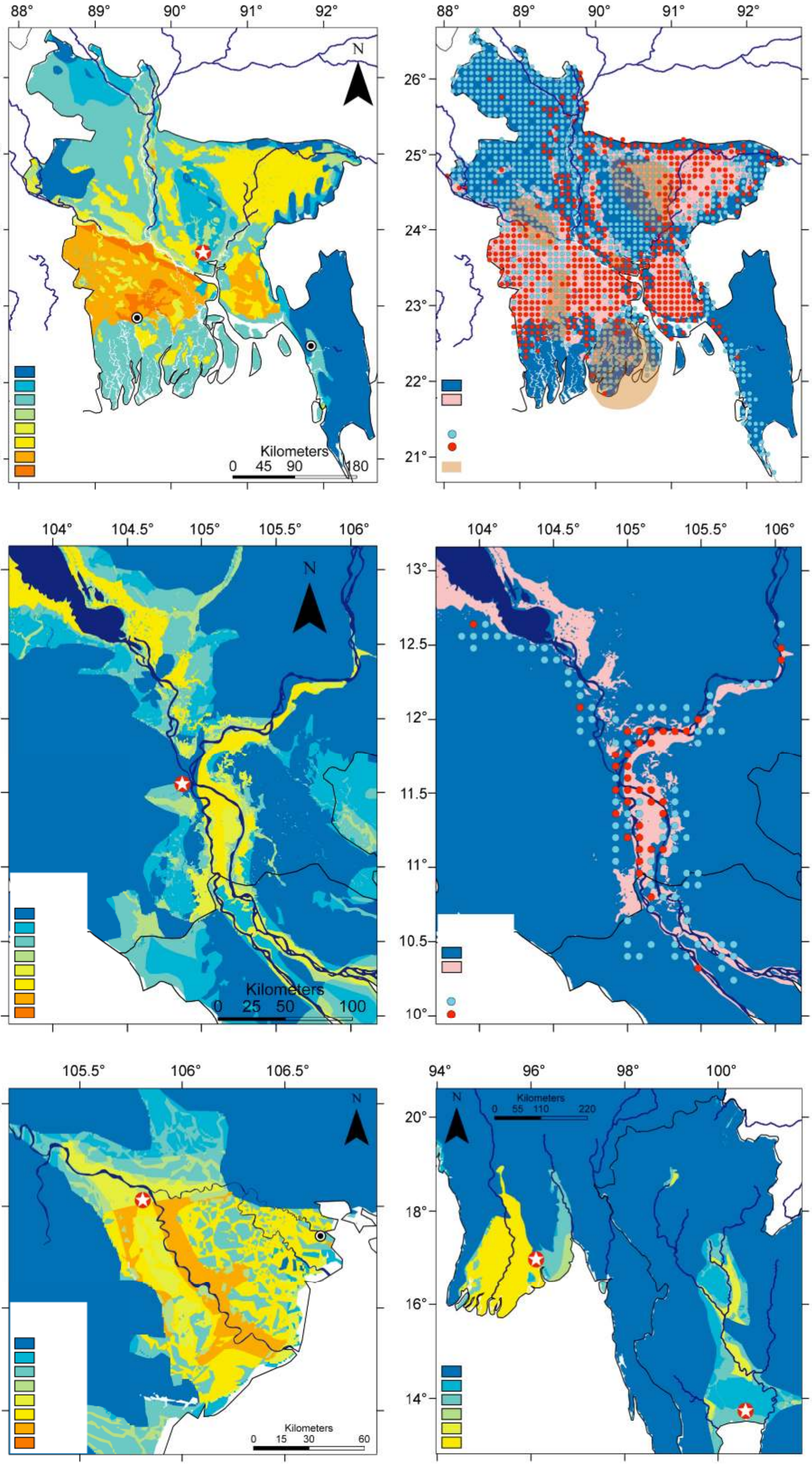
populated and the number of sampled tubewells in this area is therefore limited. **c**, Geological map (source: see Supplementary Table S1).

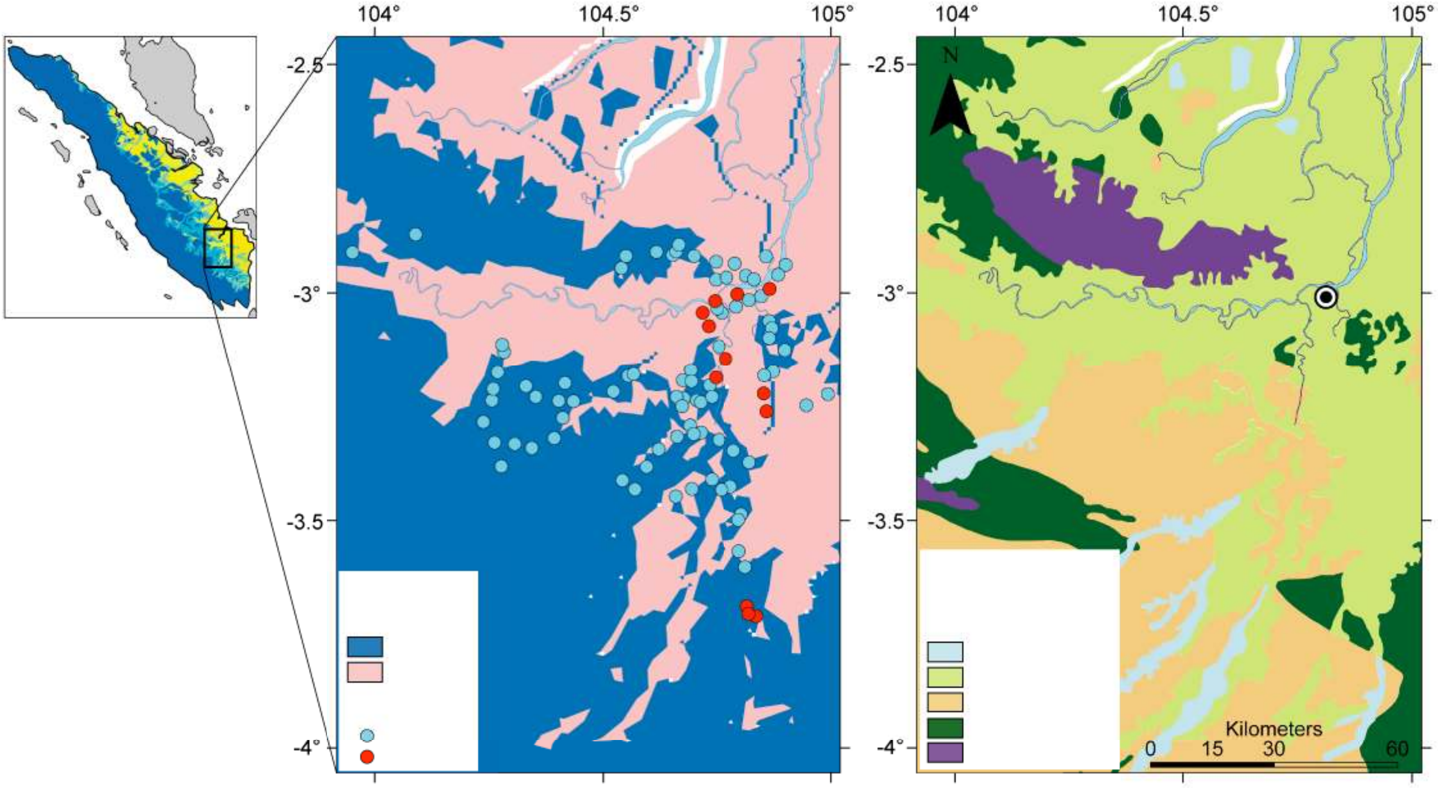
Table 1. Results of logistic regression analysis. Weighting coefficients of the independent variables (λ) were used to calculate probabilities of As contamination. Wald values (%) indicate the relative importance of the variables, and p-values the absolute significance, where a value <0.05 indicates a significance of at least 95%. Variables that were not statistically significant (based on a 95% level) were excluded from the model (Table S2), with exception of medium textured soils and silt in subsoil (see Methods). Excluded variables are: tidal deposits, other Holocene deposits, pre-Holocene deposits (see Figure 1), coarse textured soils, soil sand and clay contents, and climate.

Variables		λ	Wald	p-value
Sedimentary depositional environments	Deltaic deposits	1.65	71.55	<0.001
	Organic-rich deposits	1.11	72.20	<0.001
	Alluvial deposits	0.55	12.51	<0.001
	Floodplain deposits	– 0.95	2.32	0.010
Soil variables	Medium textured soils	0.60	7.80	0.128
	Fine textured soils	0.24	1.00	0.005
	Silt in subsoil	0.10	6.62	0.317
	Silt in topsoil	– 0.09	6.28	0.012
–	Regression constant	– 1.54	20.67	<0.001




100
90
80
70
60
50
40
30
20
10
0
Correctly classified, %
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Probability cuttoff
Correctly classified in group of observed As >10 μgL^{-1} (Sensitivity)
Correctly classified in group of observed As ≤10 μgL^{-1} (Specificity)


88°
89°
90°
91°
92°
26°
25°
24°
23°
22°
21°
N
Kilometers
0 45 90 180
104°
104.5°
105°
105.5°
106°
13°
12.5°
12°
11.5°
11°
10.5°
10°
Kilometers
0 25 50 100
105.5°
106°
106.5°
21.5
21°
20.5
20°
Kilometers
0 15 30 60
94°
96°
98°
100°
20°
18°
16°
14°
Kilometers
0 55 110 220


104°
104.5°
105°
-2.5
-3°
-3.5
-4°
104°
104.5°
105°
-2.5
-3°
-3.5
-4°
N
Kilometers
0
15
30
60