Making the most of imprecise measurements: Changing patterns of arsenic concentrations in shallow wells of Bangladesh from laboratory and field data

Yuling Yao* Rajib Mozumder† Benjamin Bostick‡ Brian Mailloux‡
Charles F. Harvey§ Andrew Gelman† Alexander van Geen‡

17 Jan 2021

Abstract

Millions of people in Bangladesh drink well water contaminated with arsenic. Despite the severity of this heath crisis, little is known about the extent to which groundwater arsenic concentrations change over time: Are concentrations generally rising, or is arsenic being flushed out of aquifers? Are spatially patterns of high and low concentrations across wells homogenizing over time, or are these spatial gradients becoming more pronounced? To address these questions, we analyze a large set of arsenic concentrations that were sampled within a 25 km^2 area of Bangladesh over time. We compare two blanket survey collected in 2000–2001 and 2012–2013 from the same villages but relying on a largely different set of wells. The early set consists of 4574 accurate laboratory measurements, but the later set poses a challenge for analysis because it is composed of 8229 less accurate categorical measurements conducted in the field with a kit. We construct a Bayesian model that jointly calibrates the measurement errors, applies spatial smoothing, and describes the spatiotemporal dynamic with a diffusion-like process model. Our statistical analysis reveals that arsenic concentrations change over time and that their mean dropped from 110 to 96 µg/L over 12 years, although one quarter of individual wells are inferred to see an increase. The largest decreases occurred at the wells with locally high concentrations where the estimated Laplacian indicated that the arsenic surface was strongly concave. However, well with initially low concentrations were unlikely to be contaminated by nearby high concentration wells over a decade. We validate the model using a posterior predictive check on an external subset of laboratory measurements from the same 271 wells in the same study area available for 2000, 2014, and 2015.

Keywords: measurement error, process model, smoothing spline, spatiotemporal dynamic, water pollution

1. Introduction

Elevated levels of arsenic (As) in water pumped from shallow wells and consumed without treatment pose a serious threat to public health across many villages of South and Southeast Asia (Fendorf et al., 2010; Flanagan et al., 2012). For lack of alternatives such as water treatment or a piped supply of safe water, rural residents of these regions with an unsafe well lowered their exposure to As mostly by switching their consumption to a neighbor's well that is low in As (van Geen et al., 2002; Chen et al., 2007). This is often a viable option because the sub-surface distribution of As in groundwater is highly heterogeneous. Well switching assumes, however, that wells are tested for As and that well users remember the test result, neither of which is necessarily the case. Another potential concern is that As levels in well water could potentially change over time. The few available studies relying on long-term monitoring indicate that groundwater As levels are by-and-large stable, but noteworthy exceptions have also been documented (Dhar et al., 2008; McArthur et al., 2010; Mihajlov et al., 2020). Groundwater flow patterns in shallow (<30 m deep) aquifers of

*Department of Statistics, Columbia University.
†Lamont-Doherty Earth Observatory of Columbia University.
‡Environmental Sciences, Barnard College.
§Department of Civil and Environmental Engineering, Massachusetts Institute of Technology.

the region have been highly perturbed and accelerated by irrigation pumping to grow rice during the winter season (Harvey et al., 2002, 2006). It is therefore plausible to anticipate some changes over time, including the possibility of convergence of spatially highly variable As concentrations towards a local mean due to physical mixing and dispersion triggered by the daily turning on and off of numerous irrigation pumps within a given area. Such a convergence of As concentrations could pose a risk to residents relying on shallow wells that were low in As when they were tested and are rarely if ever tested again thereafter.

To understand the evolution of spatial and temporal pattern of As in groundwater perturbed by irrigation pumping, we focus on a particularly well-studied $25\,\mathrm{km}^2$ area of Araihazar upazila (subdistrict) in Bangladesh. Groundwater samples from a total of 4827 shallow, mostly privately installed wells was sampled a first time in 2000–2001 and analyzed for As in the laboratory (van Geen et al., 2003). Most of these wells were replaced or re-installed by individual households in the subsequent decade (van Geen et al., 2014). In 2012–2013, 8228 shallow wells in the same area were tested, but this time with the ITS Econo-Quick Arsenic field kit (George et al., 2012) that is less precise and bins results in discrete categories (Figure 1). A subset of wells were tested with the field kit and in the laboratory and can be used for intercalibration (Mozumder, 2019). Although the Econo-Quick field kit is cost-effective (about USD 0.30 per test) and sufficiently accurate for large-scale monitoring, their limited precision poses a challenge when trying to evaluate the stability of the distribution of As in shallow groundwater over time.

This paper seeks to quantify the changing pattern over a decade of As concentrations in a perturbed shallow aquifer from a combination of laboratory measurements and noisier test kits. The contribution of our work is threefold. First, we designed a flexible Bayesian model that incorporates a before-after comparison with different precision and measurement error. Such modeling could be useful more generally when applied to field kits in environmental data analysis (e.g., Korfmacher and Dixon, 2007; Landes et al., 2019). Second, we assessed the direction and magnitude of groundwater mixing via a differential-equation-based dynamic. The inference result suggests that wells with initially low As concentration are unlikely to be contaminated by its high As neighbors, which validates the current recommendation for households with elevated arsenic levels to switch to a neighbor's safe well. Third, as a byproduct of the Bayesian modeling, we calibrated noisy individual field kit test results communicated to households.

Several projects have collected As data in Araihazar upazila, at different times and in different ways as follows.

1. *Blanket surveys.* In 2000–2001, a sample of size $n_1 = 4574$ of groundwater from all wells in the study area was analyzed by graphite-furnace atomic absorption (GFAA) in a laboratory (van Geen et al., 2003). A subset of samples at the detection limit of $5\,\mu\mathrm{g/L}$ for As by this method was subsequently re-analyzed by inductively-coupled plasma mass spectrometry (ICPMS, Cheng et al., 2004). We denote the reading of the i-th well as y_{1i}. In the second blanket survey conducted in 2012–2013, a group of trained workers tested $n_2 = 8229$ wells for As in the same area using the ITS Econo-Quick field kit. They recorded visual readings of a colored test strip relative to a reference scale showing 9 discrete levels (0, 10, 25, 50, 100, 200, 300, 500, $1000\,\mu\mathrm{g/L}$). We denote the discrete field-kit measurement of well-i as w_{2i}. Most wells in the second survey were are different from wells sampled during the first study because of replacements and new installations over a decade (van Geen et al., 2014). The well index i therefore does not corresponds the same unit in two surveys. The location of wells from both surveys was recorded with handheld GPS receivers with an accuracy of 10–50 m and is presented in a local Cartesian coordinate pair $(x_{ti}^{\mathrm{E}}, x_{ti}^{\mathrm{N}})$ for each sampled well. The

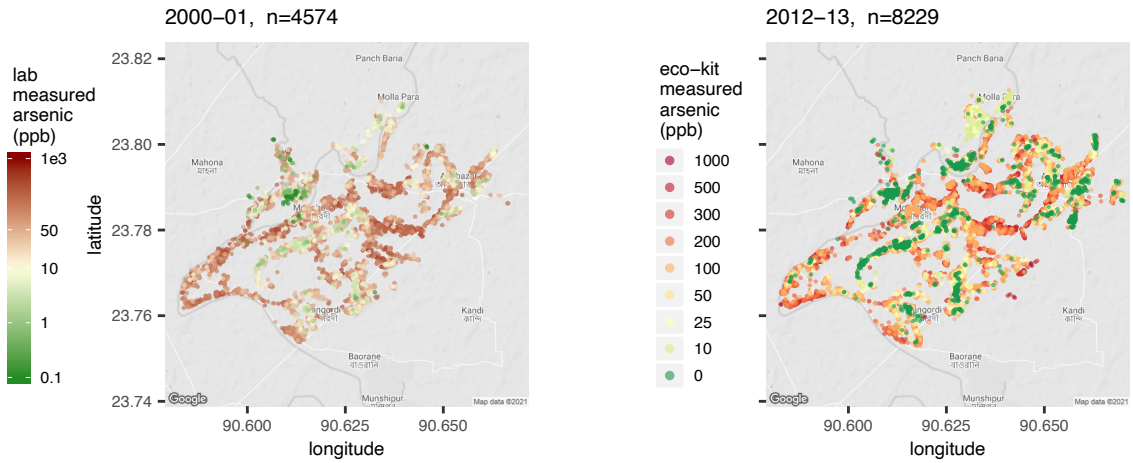

Figure 1: *In 2000–2001, a sample of size* $n_1 = 4574$ *of wells within a* $25\,\mathrm{km}^2$ *area of Bangladesh were analyzed by GFAA in laboratory. The second survey took place in 2012–2013 where* $n_2 = 8229$ *wells in the same area were measured by Econo-Quick field kit based on visual readings of a test strip at 9 discrete levels (0, 10, 25, 50, 100, 200, 300, 500, 1000 ppb). Because they were measured with different measurement precision and at different sampling locations, these two datasets are not directly comparable.*

depth of the well, as reported by households who paid for each section of pipe that went into the well's installation and therefore remember, is d_{ti} in meters.

2. *Quality-control sample.* The inaccuracy of field kits is not only attributable to discretization of visual readings but reflects also potential groundwater matrix effects on the reading, the occasional wetting of the test strip, and differences in handling. For calibration, we saved $n_{\mathrm{cal}} = 944$ water samples from the second blanket survey and analyzed them by ICPMS. We denote the lab and kit values by y_i^{cal} and w_i^{cal}, $1 \le i \le n_{\mathrm{cal}}$ respectively.

3. *Resampled subset.* A subset of 271 remaining wells from the 2000 blanket survey were re-identified in 2014 and 2015 with reasonable certainty based on the combination of a metal identification tag from 2000 that was still attached to the stem of the handpump along with a consistent location and installation year based on a conversation with the household that owns the well. These wells were therefore sampled a total of three times in 2000, 2014, and 2015 and are identified and labeled by consistent indices $i = 1, \ldots, 271$. The results from ICPMS analysis of all these samples, including re-analysis of the original set from 2000 for maximum consistency, are denoted by year t and well i as $\tilde{y}_{ti}$, for $t = 2000, 2014, 2015$. The depth of the i-th well is $\tilde{d}_i$ measured in 2000.

The remainder of this paper is organized as follows. In section 2 we first explore the 271 resampled wells for which kit-measurement error is not a major issue. In section 3.1 we consider the measurement error models that calibrate the discrete field kit measurement. We propose a Bayesian hierarchical model in section 3.2 to infer spatiotemporal patterns from the noisy kit readings. Section 4 presents the inference result of the blanket survey data, which reveals a general decline trend in As concentration across the study region. We conclude that physical homogenization is unlikely to be the driving force in this dynamic and hence do not need to be overly concerned

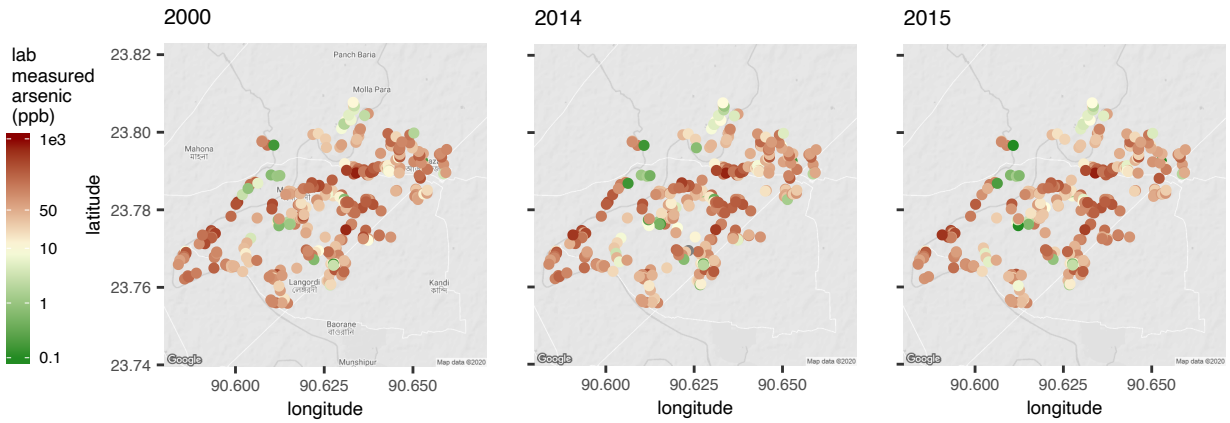

Figure 2: *308 wells in the 2000 blanket survey were re-identified in 2014 and measured in labs, among which 271 wells were tested again in 2015. Each dot in the graph represents one well.*

about shallow low arsenic wells being affected by high As wells nearby. We discuss limitations and implications of our finding in section 5.

2. Exploring the resampled wells

Because of the discreteness nature and large measurement error in the blanket survey, we first explore the accurate lab measurements in the resampled 271 wells. Besides data exploration, we later reuse this dataset as an external validation. Figure 2 visualizes the spatial distribution of arsenic values in the subset over time. All three maps show clusters of high, mixed, and low As concentrations that persisted over time. The sample average As concentrations over time for the 271 resampled wells was 100 µg/L in 2000–2001 and 90 µg/L in 2014 and 2015.

2.1. Regression to the mean vs. groundwater mixing

The main public health concern is whether physical water mixing could imperil wells that were initially safe when tested. To start, we run linear regressions on 3 pairs: 2014 versus 2000, 2015 versus 2014, and 2015 versus 2000 on the resampled dateset (Figure 3). The regression is in log scale to avoid the analysis driven by extremely large values. The posterior means of the regression coefficients are 0.97, 0.98, and 0.96 for these three pairs, and their 95% confidence intervals do not overlap with 1. The observed arsenic level in low-level wells tends to increase and in high-level wells tends to decrease in all three time periods. However, this pattern akin to regression to the mean could be the result of either measurement noise or natural fluctuations in each sampling period and is not necessarily an indication that physical mixing and homogenization is taking place.

In Figure 4, we compare log As changes over time. The well-wise changes in 2000–2014 are positively correlated with those in 2000–2015 but negatively correlated with 2014–2015. This evidence suggests measurement error rather than physical mixing as the underlying cause of convergence of groundwater As concentrations. Both the short term and long term changes show large variations, with multiplicative shifts ranging from 20 to 1/20.

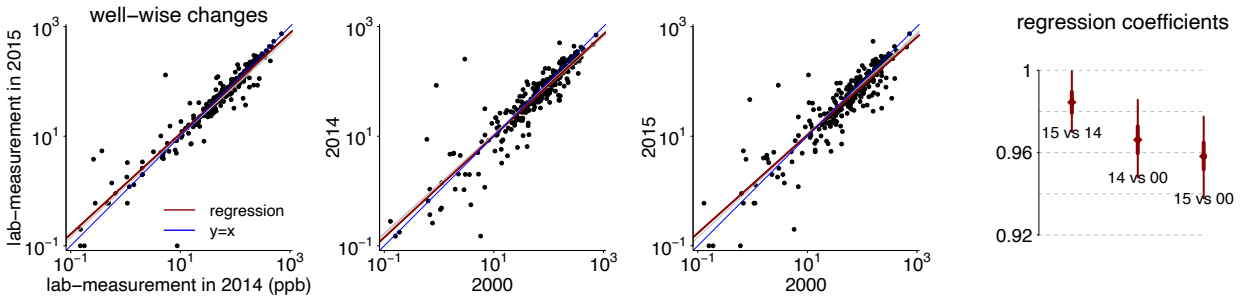

Figure 3: *"Regression to the mean." Lower values tend to increase and higher values tend to decrease. However, this does not separate the observational noise and the potential water mixing.*

2.2. Spatial smoothing using a Gaussian process and autoregression using a spline

To try to separate the pattern of water mixing, we decompose the observed log arsenic values $\log \tilde{y}_{ti} (t = 2000, 2014, 2015, i = 1, \ldots, 271)$ into a spatial baseline arsenic level (the expected value conditional on the location), denoted by θ, the depth term, and residuals ϵ.

$$\log \tilde{y}_{ti} = \theta_{t,i} + \beta_{\text{depth}}(\tilde{d}_i - d_0) + \epsilon_{ti}, \ \epsilon_{ti} \sim \text{normal}(0, \sigma_S). \tag{1}$$

The spatial-dependent parameter $\theta_{t,i}$ is the baseline value in well i and year t, which we will specify in the next paragraph. The depth term describes the dependence of arsenic concentration on well depth via a linear form $\beta_{\text{depth}}(\tilde{d}_i - d_0)$, where $d_0 = 15.29$ meters is the pre-calculated mean depth of all wells. The residual ϵ contains measurement errors, short-term fluctuations, or any other unobserved factors that can not be explained by the spatial distribution of wells. We model it by a normal distribution with standard deviation σ_S (the *short*-term).

We model baseline arsenic concentration θ as a function of location $x \in \mathbb{R}^2$: the two dimensional location coordinate. That is $\theta_{t,i} = \theta_t(x_i)$. We place a Gaussian process prior on $\theta_{2000}(x)$, with a squared exponential kernel,

$$\theta_{2000}(x) \sim \mathcal{GP}(\mu, K), \quad K(x_1, x_2) = \alpha \exp\left(-\frac{||x_1 - x_2||^2}{\rho^2}\right), \tag{2}$$

where $||x_1 - x_2||^2$ is the Euclidean distance between two locations x_1 and x_2. The Gaussian process is flexible to characterize the spatial pattern.

From the summary statistics (mean, median, standard deviation, median absolute deviation) of the well-level arsenic-changes (Figure 4), the mean of the change from 2014 to 2015 (the left panel) is small compared to the change between 2000 and 2014. Therefore, we assume

$$\theta_{2015,i} = \theta_{2014,i}, \quad i = 1, \ldots, 271. \tag{3}$$

This approximation is computationally necessary because the problem is non-identified otherwise.

Changes in well-water As concentrations driven by physical mixing would eventually lead to homogenization independently of observational noise. In order to approximate this mechanism, we consider an autoregression from θ_{2000} to θ_{2014}. A strong negative autocorrelation will support the water mixing hypothesis such that higher As values would drop and lower As values would increase.

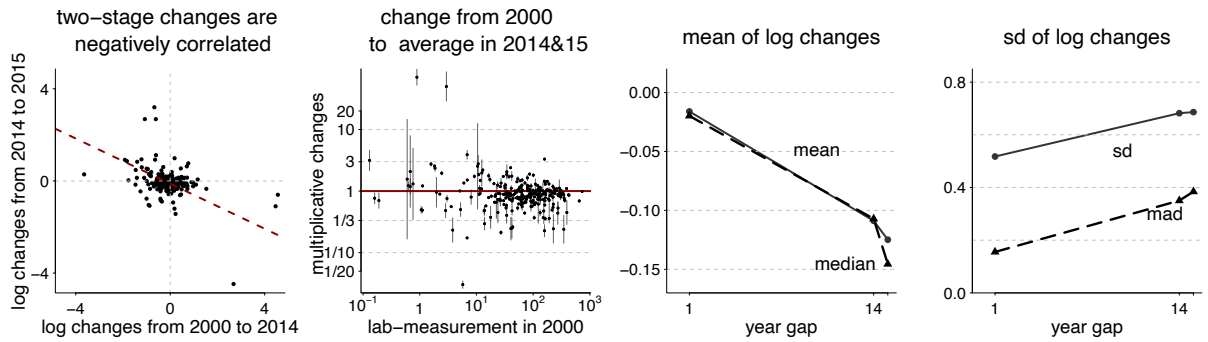

Figure 4: *(1) The comparison of log change in 2000–2014 versus 2014–2015 per well. (2) The multiplicative change from 2000 to 2014 or 2015 per well as a function of the the initial values. The two endpoints are the 2000–2014 and 2000–2015 changes and the dot is the average. (3–4): Summary statistics (mean, median, standard deviation, median absolute deviation) of the well-level arsenic-changes as a function of year gaps.*

In the exploration data analysis (Figure 3), we model a linear autoregression. To make it more flexible, we model the baseline change by a cubic spline of the initial value:

$$\theta_{2014,i} - \theta_{2000,i} = \beta_0 \theta_{2000,i} + \sum_{l=1}^{L} \beta_l B_l(\theta_{2000,i}) + \text{normal}(0, \sigma_L), \quad i = 1, \ldots, 271. \tag{4}$$

where $\{B_l\}_{l=1}^{L}$ is a collection of cubic spline basis functions. Unlike typical spline smoothing directly applied to data, here we are fitting a spline regression on the latent variable θ, whose range is unknown. Nevertheless, from the observational model (1), it is reasonable to expect that $\theta_{t,i}$ and $\log \tilde{y}_{t,i}$ have a similar range. Hence we choose the internal knots of the cubic spline to the $(0.1, 0.2, \ldots, 0.9)$ quantile of all observed $\log \tilde{y}_{t,i}$. We use a random-walk prior as on spline coefficients β_l to encourage smoothness, and place a weakly-informative prior on remaining parameters:

$$\begin{aligned} &\beta_0, \beta_1 \sim \text{normal}(0, 1), \ \beta_{l+1} \sim \text{normal}(\beta_l, 0.5), \\ &\sigma_S, \sigma_L \sim \text{InvGamma}(3, 3), \ \mu \sim \text{normal}(4, 1). \end{aligned} \tag{5}$$

We sample from this joint posterior distribution in Stan (Stan Development Team, 2020), a platform for Bayesian modeling and computation.

Inference results for 271 resampled wells

The left panel of Figure 4 displays the posterior means, 50%, and 95% confidence intervals of the autoregression cubic spline in (4), i.e., $\beta_0 \theta_{2000} + \sum_{l=1}^{L} \beta_l B_l(\theta_{2000})$ as a function of θ_{2000}. It represents the expected value of log As changes in 2014 and 2015, given a well's initial baseline arsenic value in 2000. This result suggests that large initial values were likely to drop over time. Perhaps more surprisingly, As concentrations at the low end of the range were more likely to decline as well, whereas intermediate values did not change appreciably.

Conditional on the the baseline As level in 2000, there is still a large amount of uncertainty of how the distribution of concentrations evolve due to both the long term regression noise σ_L and the short term observational noise σ_S. To help the interpretation of this uncertainty, we simulate

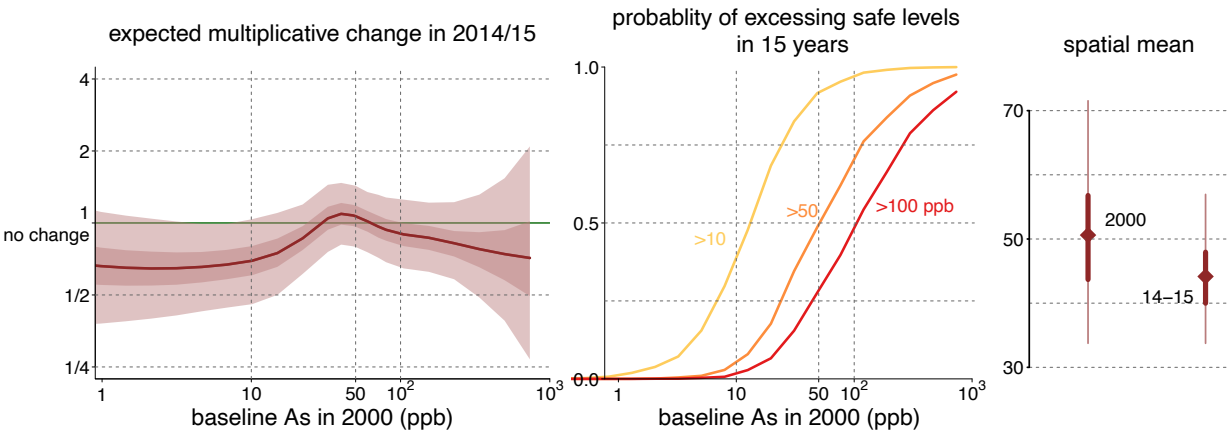

Figure 5: *Left: posterior mean and intervals of the predicted change in 2014, given the well-arsenic level in 2000. Middle: the probability that a well would excess safety values in 15 years as a function of initial measurements. Right: the global mean of* $\theta_{2000,i}$ *is* $\log 51$ *and the mean of* $\theta_{2014,i}$ *is* $\log 44$. *We label the y-axis by the usual linear scale (ppb) for readability.*

in the middle panel a series of posterior draws to compute the probability of a realized laboratory test in a well exceeding certain threshold in 15 years as a function of its initial baseline value: a well with initial value 10 ppb has a 40% chance to exceed 10 ppb and 4% chance to exceed 50 ppb in 2014's measurement. This graph provides an accessible way to communicate decision making: despite all the temporal variation readily exists, a well is unlikely to change from 10 ppb to 50 ppb in 14 years. In other words, a clearly safe/unsafe well (smaller than 10 or larger than 100 ppb) is nearly ensured to keep safe/unsafe and should be encouraged to switch to/away from. Mailloux et al. (2020) reach a similar conclusion using another approach as well as additional well-water arsenic data from different regions.

The right panel shows the fitted overall spatial mean of the latent baseline level θ in 2000 and 2014, with their 95% confidence interval overlapped. This is in line with the 95% confidence interval in the left panel showing an overall trend of mildly decreasing arsenic concentrations without ruling out that there was no change.

The data also confirm the gradual increase in As concentrations with depth. The coefficient β_{depth} is estimated to be 0.03 (95% CI $(0.01, 0.05)$) per meter.

The observational noise and suggestions of a "regression to the mean" do not exclude the possibility that the underlying As concentrations are truly converging due to physical water mixing. A limitation of the resampled wells is their small sample size. In the next section, we model the underlying stochastic process using two blanket surveys conducted in 2000–2001 and 2012–2013 that are much larger but partially measured by less accurate field kits.

3. A closer look at the two large surveys

3.1. Kit calibration using the quality-control samples

The discrete kit readings w_2 in 2012–2013 are not directly comparable to the more precise laboratory measurements y_1 in the 2000–2001 survey.

In the quality-control data, we observe both the kit measurement w_i^{cal} and the lab measurement y_i^{cal} for $i = 1, \ldots, 944$. The first two columns of Figure 6 compares the face value of eco-kit and

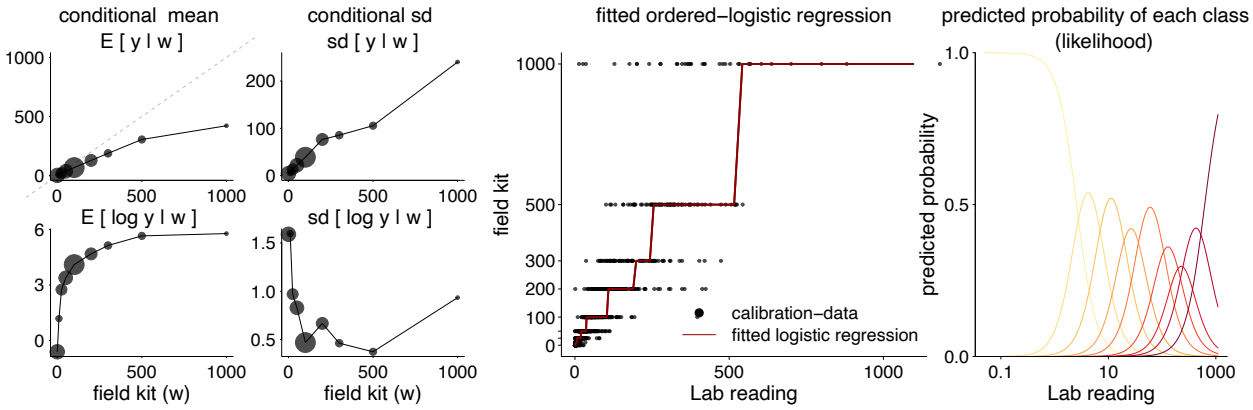

Figure 6: *First two columns: conditional mean and standard deviation of lab measurement y conditioning on the face value of eco-kit measurement w. The dot size is the sample size in that category in the calibration set. Third column: the fitted ordered-logistic regression using the the quality-control sample. The red line is the posterior prediction of w given y, while the black dot is the observation in the calibration set. Rightmost: the predicted probability of each class* $\Pr(w = k|y, \hat{\beta}^{\text{cal}}, \hat{c}^{\text{cal}})$*, as a function of lab reading y. Colors from shallow yellow to dark red represents kit-test classes from low to high.*

laboratory measurements of the same water sample. The kit reading exhibits a systemically negative bias and a considerable variance. For calibration, we treat the kit measurement w as a nominal variable that only takes integer values $1, 2, \ldots, 9$. We model the kit-measurement $p(w_i^{\text{cal}}|y_i^{\text{cal}})$ by an ordered logistic regression with a flat prior and the following likelihood:

$$\Pr(w_i^{\text{cal}} \le k \,|\, \beta^{\text{cal}}, c^{\text{cal}}, y_i^{\text{cal}}) = \text{logit}^{-1}\left(c_k^{\text{cal}} + \beta^{\text{cal}} \log y_i^{\text{cal}}\right), \quad k = 1, 2, \ldots, 9. \tag{6}$$

The ordered logistic regression does not take the face value of the kit measurement as true value, but does keep the sequential order of the nominal readings. The fitted result is shown in Figure 6. With $n = 944$ quality control samples, the joint posterior distribution of these 10 parameters is nearly a point mass.

When it comes to the kit-measurement w_{2i} in the large blanket survey that we will model in the next subsection, the lab measurement is missing, and we will impute it probabilistically from the ordered-logistic regression learned here.

3.2. Modeling the mixing process

We model the spatial arsenic distribution with a smoothing spline and superimpose a diffusion-like process to reflect changes in the distribution between 2000–2001 and 2012–2013.

Smoothing spline and its Laplacian

For the i-th well in the 2000–2001 survey at location $(x_{1i}^{\text{N}}, x_{1i}^{\text{E}})$, we decompose the log laboratory reading $\log y_{1i}$ into (a) the spatial baseline factor, modeled by a bivariate spline of $(x_{1i}^{\text{N}}, x_{1i}^{\text{E}})$, (b) the depth dependence, modeled by a linear function of depth $(d_{1i} - d_0)$, where d_0 is the average

depth of all observed wells, and (c) an independent observational noise.

$$\log y_{1i} = \beta_0 + \sum_{l=1}^{L} \beta_l B_l(x_{1i}^{\mathrm{N}}, x_{1i}^{\mathrm{E}}) + \beta_{\mathrm{depth}}(d_{1i} - d_0) + \mathrm{normal}(0, \sigma_{\mathrm{obs}}),\ 1 \le i \le n_1. \tag{7}$$

We construct the bivariate spline from tensor products of one-dimensional cubic B-splines, with L basis function $B_l(x_{2i}^{\mathrm{N}}, x_{2i}^{\mathrm{E}}) = B_l^{\mathrm{N}}(x_{2i}^{\mathrm{N}}) B_l^{\mathrm{E}}(x_{2i}^{\mathrm{E}})$. After trimming, the number of product-basis-functions is $L = 485$. We detail the spline knot choice in Appendix A.3.

From this spline model, for all sampling locations $x_{2i}^{\mathrm{N}}, x_{2i}^{\mathrm{E}}$ in the second survey, the counterfactual log baseline value is described by parameters θ_{1i}: the expected log As value if well i in the second had existed in 2000–2001 and we had tested its sample using lab reading.

$$\theta_{1i} = \beta_0 + \sum_{l=1}^{L} \beta_l B_l(x_{2i}^{\mathrm{N}}, x_{2i}^{\mathrm{E}}),\ 1 \le i \le n_2. \tag{8}$$

The spline model describes not only the value of θ_{1i}, but also the geometry of the static As value surface. In particular, the Laplace operator $\Delta = \frac{\partial^2}{\partial (x^{\mathrm{N}})^2} + \frac{\partial^2}{\partial (x^{\mathrm{E}})^2}$ describes the local curvature. The Laplacian of basis functions comes in closed form $\Delta B_l(x_{2i}^{\mathrm{N}}, x_{2i}^{\mathrm{E}}) = \Delta B_l^{\mathrm{N}}(x_{2i}^{\mathrm{N}}) B_l^{\mathrm{E}}(x_{2i}^{\mathrm{E}}) + \Delta B_l^{\mathrm{E}}(x_{2i}^{\mathrm{E}}) B_l^{\mathrm{N}}(x_{2i}^{\mathrm{N}})$, where ΔB^{E} and ΔB^{N} is the second order derivative of the one-dimensional cubic spline basis functions. Therefore, we can extract pointwise Laplacian δ_i as a linear transformation of spline coefficients:

$$\delta_i = \sum_{l=1}^{L} \beta_l \Delta B_l(x_{2i}^{\mathrm{N}}, x_{2i}^{\mathrm{E}}),\ 1 \le i \le n_2, \tag{9}$$

which reflects the As surface curvature in 2000–2001 at location $(x_{2i}^{\mathrm{N}}, x_{2i}^{\mathrm{E}})$.

From the ordered logistic regression, we use a variable η_{2i} to represent the log lab reading of well-i: the inferred log lab measurement if we had done the second survey in lab. We plug-in the posterior mean of $\hat{c}_k^{\mathrm{cal}}, \hat{\beta}^{\mathrm{cal}}$ learned from the calibration model (3.1), and the likelihood of the missing lab measurement is expressed by

$$\mathrm{logit}(\Pr(w_{2i} \le k)) = \hat{c}_k^{\mathrm{cal}} + \hat{\beta}^{\mathrm{cal}} \eta_{2i}, \quad k = 1, 2, \ldots, 9,\ 1 \le i \le n_2. \tag{10}$$

Like the observed $\log y_{1i}$ in data model (7), the hypothetical log lab reading η_{2i} also contains the noise-free baseline As surface θ_{2i}, the depth dependence, and another independent Gaussian noise.

$$\eta_{2i} = \theta_{2i} + \beta_{\mathrm{depth}}(d_{2i} - d_0) + \mathrm{normal}(0, \sigma_{\mathrm{obs}}),\ 1 \le i \le n_2. \tag{11}$$

Mixing dynamic on latent arsenic surface

So far we have modeled θ_{1i} and θ_{2i}: the static baseline log As surfaces at the same well i in 2000–2001 and 2012–2013 respectively. Next, we model the temporal dynamic.

In an ideal noise-free isotropic mixing scheme, the As concentration surface $y(x, t)$ as a function of location x and time t would follow a diffusion process, $\frac{\partial}{\partial t} y(x, t) \propto \Delta_x y(x, t)$. Based on this heuristics, we build an autoregression:

$$\theta_{2i} = \theta_{1i} + \alpha_\delta + (\beta_\delta + \gamma_i)\delta_i + \mathrm{normal}(0, \tau),\ 1 \le i \le n_2. \tag{12}$$

The temporal change between θ_{1i} and θ_{2i} contains the global shift α_δ, the product of the mixing coefficient $\beta_\delta + \gamma_i$ and the local Laplacian δ_i, and independent regression residuals.

This autoregression dynamic (12) is not equivalent to the diffusion process because it models θ, the log scale As concentration. Besides, the discretization error over $t = 12$ years cannot be ignored. On the other hand, we do not expect the flow and recharge of groundwater As being perfectly described by a molecular diffusion and hence the diffusion process itself is not realistic.

That being said, we view the Laplacian δ_i as an extracted feature that describes how the As concentration in a well is compared with its nearby neighbors. In order to learn the a more general dynamic, we allow the coefficient $\beta_\delta + \gamma_i$ to vary by the initial well As, instead of a homogeneous diffusion constant. In effect, we use (12) to learn a mixing dynamic

$$\frac{\partial}{\partial t}\theta(x,t) \propto \alpha_\delta + (\beta_\delta + \gamma(\theta))\Delta_x(\theta). \tag{13}$$

The regression residual in (12) determines the fidelity of the solution to this dynamic.

Data-dependent mixing coefficient

The mixing coefficient contains a constant term β_δ and a random term γ_i. We parameterize the latter term by a function of θ_{1i},

$$\gamma_i = \alpha_y \exp(\theta_{1i}/2) + \alpha_\theta \theta_{1i}, \tag{14}$$

The combination of linear and exponential inputs is designed to fit both the large and small end of the initial value θ_{1i}.

Priors and computation

Apart from the calibration parameter $\{c_k^{\text{cal}}\}_{k=1}^9$, β^{cal} learned in advance, the complete model contains 16950 free parameters in total: $\{\beta_l\}_{0=1}^{485}$, σ_{obs}, $\{\theta_{2i}\}_{i=1}^{8229}$, $\{\eta_{2i}\}_{i=1}^{8229}$, α_0, α_y, α_θ, α_δ, β_δ, τ, on which we place weakly informative priors:

$$\begin{gathered}\beta_0 \sim \text{normal}(4,2),\ \beta_l \sim \text{normal}(0,0.5), 1 \le l \le L = 485.\\ \alpha_y \sim \text{normal}(0,0.2),\ \alpha_\theta \sim \text{normal}(0,0.5),\ \alpha_\delta \sim \text{normal}(0,1),\ \beta_\delta \sim \text{normal}(0,1).\\ \sigma_{\text{obs}} \sim \text{InvGamma}(5,5),\ \tau \sim \text{InvGamma}(5,5).\end{gathered} \tag{15}$$

We fit the complete model (7)–(15) in Stan based on 4 chains and 2000 posterior simulation draws of all these parameters. Besides using the usual dynamic Hamiltonian Monte Carlo sampler, we employ sparse matrix algebra when computing the log joint density, since the basis functions $B_l(x_{1i}^{\text{N}}, x_{1i}^{\text{E}})$, $B_l(x_{2i}^{\text{N}}, x_{2i}^{\text{E}})$ and their Laplacians ΔB are sparse matrices.

4. Inference results for blanket surveys

4.1. Individual prediction

The posterior distribution of η_{2i} provides inference for the As value in the i-th sampled well in 2012–2013. It is both spatially smoothed owing to the spline, and calibrated against the measurement errors. Figure 7 displays the posterior mean and 10% and 90% quantile for all wells in 2012–2013. For what matters to public health, we also compute the posterior probability of each well exceeding the safety threshold (10, 50, and 100 ppb) in Figure 8.

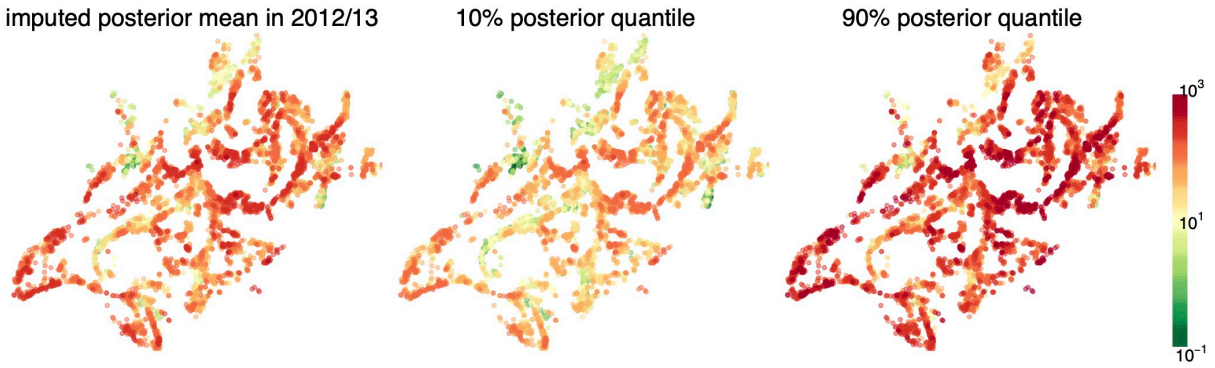

Figure 7: *The posterior mean, 10%, and 90% quantile of η_2: the calibrated As level of the well in 2012–2013.*

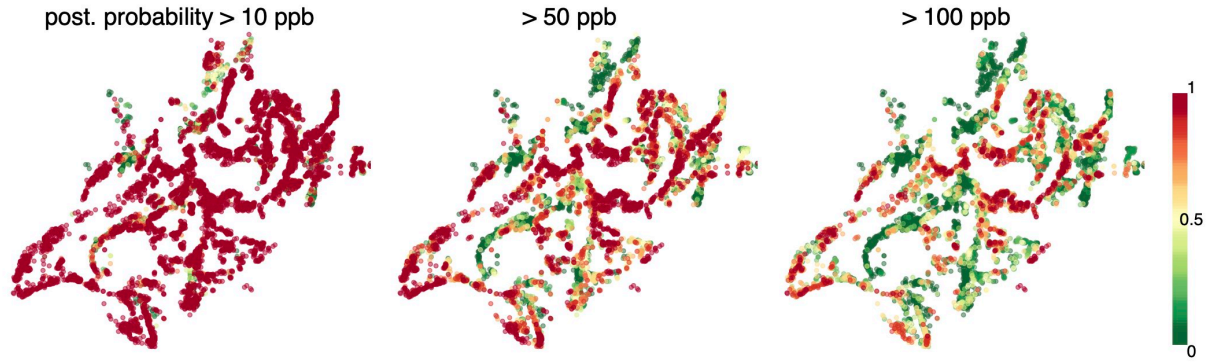

Figure 8: *The posterior probability that each sampled well was exceeding the As safety threshold (10, 50, and 100 ppb) in 2012–2013.*

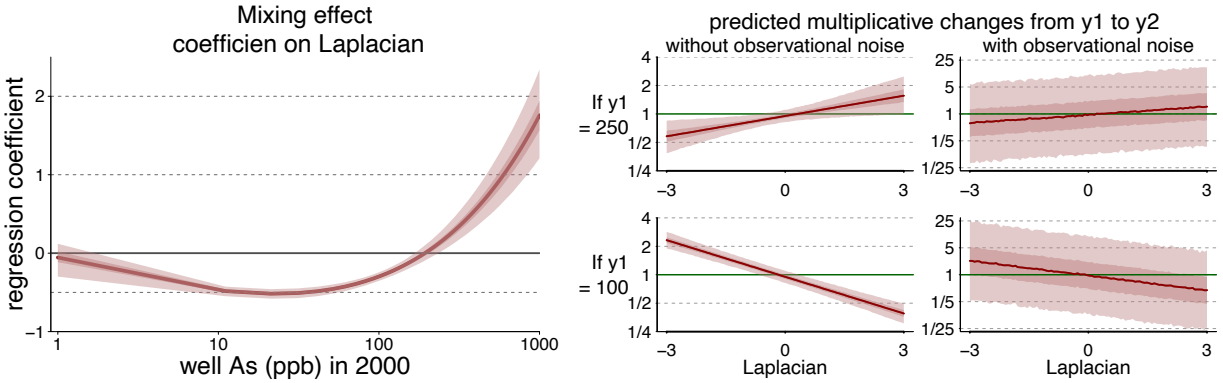

Figure 9: *Left: the posterior mean, 50% and 95% confidence interval of the regression coefficient before the Laplacian $\beta_\delta + \alpha_y \exp(\theta_{1i}/2) + \alpha_\theta \theta_{1i}$. A positive coefficient represents a diffusion effect, which is allowed to vary over θ_1. Middle column: the posterior mean, 50% and 95% confidence interval for predictive multiplicative change term $\alpha_\delta + (\beta_\delta + \alpha(\theta_1))\delta$ as a function of δ for two fixed values $\exp(\theta_{\theta_1}) = 100$ and 250 ppb. Right: posterior mean, 50%, and 95% confidence intervals for predictive multiplicative change in the observational level, with extra noise term added.*

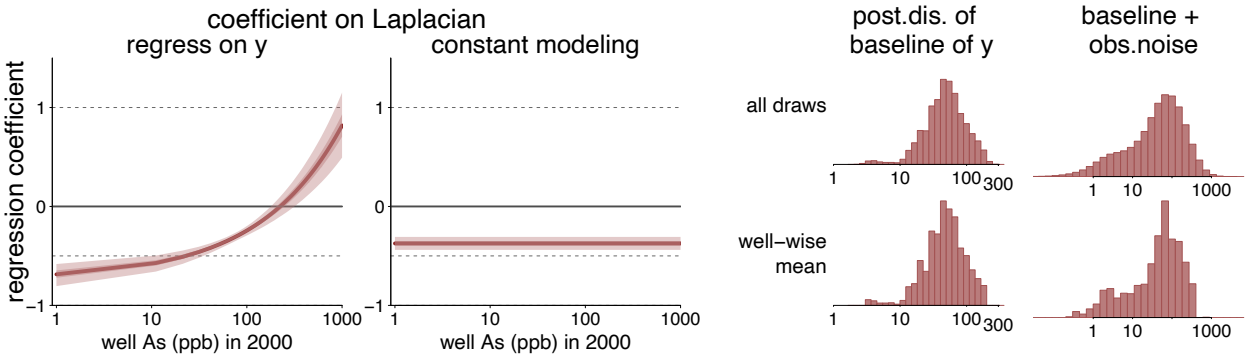

Figure 10: *To check how sensitive the model is to the regression assumption, we consider two alternative models where the mixing effect coefficient is modeled by a a linear regression on* $\exp(\theta)$ *and a constant respectively.*

4.2. Mixing effect

The log As level changes from 2000–2001 to 2012–2013 in the i-th well is modeled by $\theta_{2i} - \theta_{1i} = \alpha_\delta + (\beta_\delta + \gamma_i)\delta_i +$ noise. Figure 9 shows the posterior mean, 50% and 95% confidence interval of the mixing coefficient, $(\beta_\delta + \gamma_i)$, as a function of the initial baseline As concentration $\exp(\theta_{1i})$. When the initial value is very high (> 200 ppb), the mixing coefficient is positive, which would drive locally large As values in 2000–2001 to drop in 2012–2013. For small θ_1, the mixing coefficient is estimated to be negative. As a result, the portions of the aquifers that had As levels lower than than both their local neighbors and 200 ppb in 2000–2001 are likely to become safer in average over time. This pattern is consistent with the trend displayed for the subset of resampled wells in Figure 5.

The middle column in Figure 9 simulates the predictive multiplicative change term $\exp\left(\alpha_\delta + (\beta_\delta + \alpha(\theta_1))\delta\right)$ as a function of the Laplacian δ conditional on two fixed initial values $\exp(\theta_1) = 100$ and 250 ppb respectively. To give a sense of the additional observational noise, in the rightmost column, we simulate from the actual observable change by adding back the observational noise $\exp\left(\alpha_\delta + (\beta_\delta + \alpha(\theta_1))\delta + \text{normal}(0, \sigma_{\text{obs}}) + \text{normal}(0, \tau)\right)$.

Should we worry about the the restriction that the coefficient $\beta_\delta + \alpha_y \exp(\theta_1/2) + \alpha_\theta\theta_1$ only has two degrees of freedom? First this is already an extension from the constant-diffusion-coefficient model. We also consider two alternative models that replace replacing this functional form by a linear regression on $\exp(\theta_1$ or a constant. The fitted result is shown in Figure 10. The newly fitted linear model is close to our main model fit except for the low end of the range in As concentrations.

Notably, θ_{1i} itself is a latent variable and we do not know its range a priori. The third and fourth column in Figure 10 exhibit the histogram of all posterior simulation draws of the θ_{1i}, the histogram of well-wise posterior mean $\text{E}_{\text{post}}(\theta_{1i})$, and the same two distribution for observed quantity η_{2i}. The majority of θ_{1i} is supported on the $\log[10, 300]$ interval. Consequently, the diffusion coefficient, as a function of θ_1 will rely on extrapolation beyond this support. This explains why different parametric assumptions yield similar inference on this interval.

4.3. Overall trend

Besides mixing effect, we are also interested in the direction and magnitude of the As levels change over time. In the left panel in Figure 11, we simulate the average multiplicative well-wise change

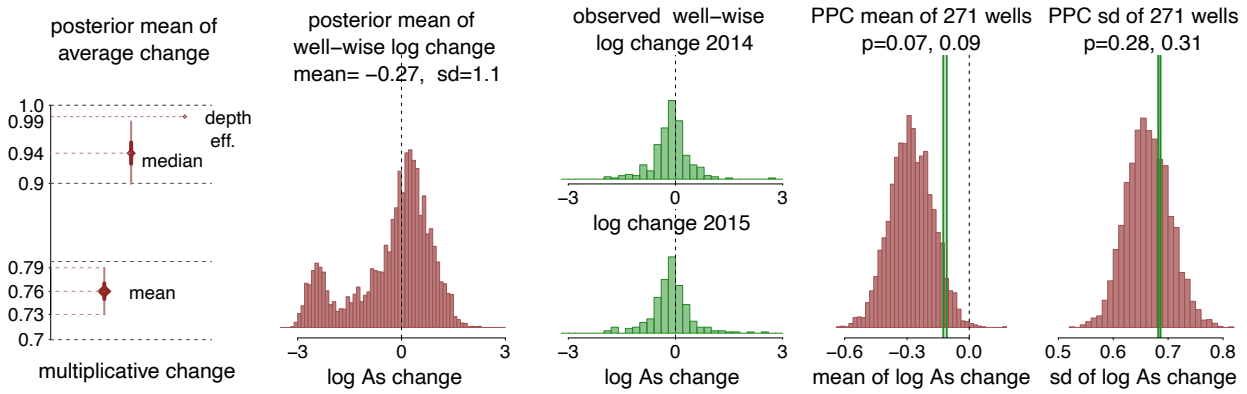

Figure 11: *(1) Averaging over n_2 wells, the mean change from baseline θ_{1i} to θ_{2i} is equivalent to a 76% multiplicative change in As concentration, with 95% confidence interval (73%, 79%). The median change is a 6% decrease. Apart from the baseline θ, the average depth of wells decreases 0.2 meter, adding to another 1% drop. (2) The distribution of posterior mean of n_2 individual well-wise log changes. (3) The observed well-wise change in the resampled 271 wells in 2014 and 2015. (4–5) The posterior predictive distributions of mean and standard deviation of log As change computed by 271 random wells each draw. They match with the observed mean and sd of log As changes in the resampled wells.*

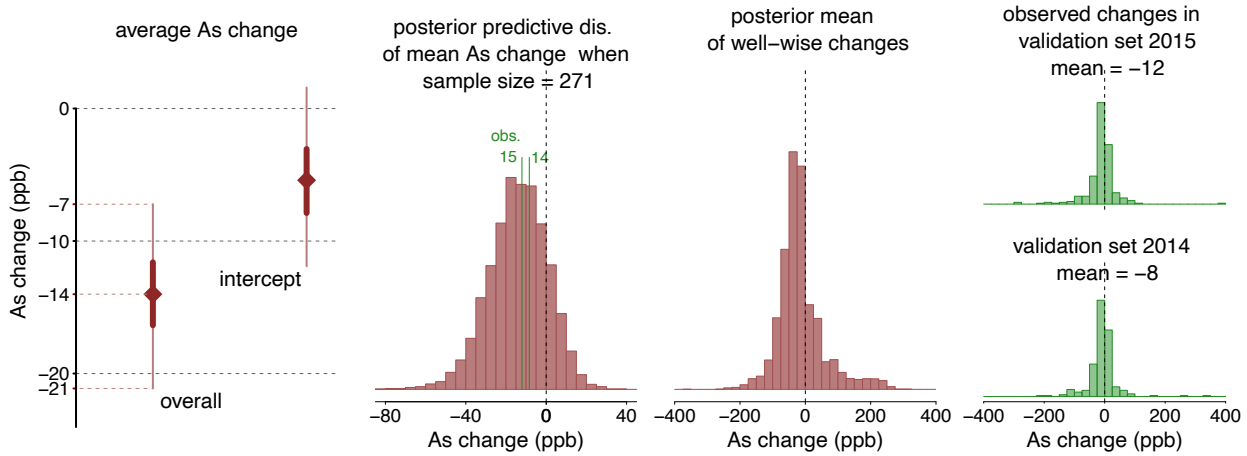

Figure 12: *(1) The posterior mean of overall As change averaged over all n_2 wells is 14 ppb decreasing, among which 5 ppb of the decrease is attributed to the intercept. (2) The posterior predictive distribution of overall As change averaged over 271 random wells. It matches with the observed mean As changes in the resampled wells. (3) The distribution of posterior mean of n_2 individual well-wise changes, among with 73% of wells are inferred to decrease. (4) The observed well-wise change in the resampled wells.*

among all wells in 2012–2013 by $\exp(\frac{1}{n_2}\sum_{i=1}^{n_2}(\theta_{2i}-\theta_{1i}))$ and compute its posterior mean and 95 % confidence interval, which equals 76% (95% CI (73%, 79%)). In other words, the average As concentration (among all n_2 sampling locations) decreased by 24% from 2000–2001 to 2012–2013. Replacing the mean by median in the previous simulation, the median decrease is 6% (95%

CI (2%,10%)). These inferred mean shifts are evaluated at the same well in the same sampling location and is net of depth dependence. The sample mean depth of 2012–2013 wells is slightly shallower (0.2 meter, or 1%) than from 2000–2001, which contributes to an additional 1% As mean decrease through the $\beta_{\text{depth}}d$ term.

However, a 24% drop in average does not imply that concentrations in all wells had to decline over a decade. The second panel in Figure 11 displays the histogram of posterior mean of well-wise log change $\theta_{2i} - \theta_{1i}$: the log baseline change in the i-th well. The mean is -0.27 (equivalently 76% in multiplicative scale) while the standard deviation of 1.1 is much larger. This large standard deviation reflects the intrinsic variation of individual level well-As changes, and will not shrink with a larger sample size.

To better interpret this temporal change, we draw posterior simulations of the the linear scale As change (in ppb) of each well. The hypothetically-lab-measured average well As in the 2012–2013 blanket survey is $\frac{1}{n_2}\sum_{i=1}^{n_2}\exp(\eta_{2i})$, which has a posterior mean of 96 ppb (95% CI: $(94, 99)$). The hypothetically-lab-measured well As at these locations in 2000–2001 is $1/n_2\sum_{i=1}^{n_2}(\exp(\theta_{1i} + \beta_{\text{depth}}(d_{2i} - d_0) + \text{normal}(0, \sigma_{\text{obs}})))$, which has a posterior mean of 110 ppb (95% CI: $(104, 117)$). As for a reference, the sample mean of the first blanket survey in 2000–2001 is 109.6 ppb. For each well in the second blanket survey, we can compute the paired change $\exp(\eta_{2i}) - \exp(\theta_{1i} + \beta_{\text{depth}}(d_{2i} - d_0) + \text{normal}(0, \sigma_{\text{obs}}))$. Averaged over $i = 1, \ldots, n_2$, the mean well As change is -14 ppb with 95% confidence interval $(-7, -21)$, as shown in the first panel in Figure 12. More precisely, the model interprets the overall log change $\theta_2 - \theta_1$ into two terms, the intercept α_δ and the diffusion effect $(\beta_\delta + \gamma_i)\delta_i$. The intercept represents the expected change if the mixing process akin to diffusion is completely blocked. We can generate the linear scale of posterior distribution of the intercept by simulation draws of $\frac{1}{n_2}\sum_{i=1}^{n_2}\exp(\eta_{2i})(1 - \exp(\alpha_\delta))$, which has a posterior mean of -5 ppb.

Again, a 14 ppb decrease in average As concentration does not mean all wells have to drop. The third panel in Figure 12 is the distribution of the well-wise posterior mean of As change, among which 27% of individual wells are inferred to increase.

4.4. Posterior predictive model check on external validation data

In order to verify the pattern we discovered in the blanket survey data, we use the resampled 271 wells as a validation set to check the model fit of the blanket survey data. As a caveat, the resampled wells are not spatially identical to the blanket survey: the former one is relative coarse in the northeast, such their sample average of lab measurements in 2000–2001 also differ slightly. Nevertheless, we still expect them to share a similar overall trend.

Because the well As is highly skewed in the blanket survey sample, the inferred mean log change (-0.27, or -24% in multiplicative scale) is not the same as the log change of mean (the average well As drops from 110 to 96 ppb, or -12%) at the same locations. We check the model fit in both scales.

We first check the log changes of individual wells. We may compare the inferred well-wise posterior mean of log well As change among the blanket survey data (the second panel in Figure 11) with the observed well-wise log As changes in the resampled dataset over a decade (the third column in Figure 11), but it is misleading as they differ in sample sizes. To make the correct model check, for each joint simulation draw, we randomly choose 271 wells from the second blanket survey and compute the mean and standard deviation of the log As change among these 271 random wells. The right two columns in Figure 11 visualize the posterior predictive distribution of these two quantities, both yielding acceptable p-values when compared with the observed validation data.

We then the check the linear scale changes. Again, we randomly draw a sample of size 271

among all n_2 wells in the second blanket survey and compute the sample average change. The second column of Figure 12 displays the histogram of 4000 posterior simulation draws of this random average, which matches well with the observed changes in both validation sets.

5. Discussion

5.1. Limitations and assumptions

Our model in section 3.2 contains three components: (a) imputing the counterfactual lab measurements from inaccurate kit data, (b) spatially smoothing the latent As surface, and (c) estimating the mixing dynamic over time.

Our approximation to the differential equation (13) is in line with Ramsay et al. (2007)'s approach to parameter estimation in general differential equations by fitting well As surface using a collection of basis functions. To this end, the spatial modeling part (8) describes the static log As surfaces by tensor products of B-spines. This spatial component is similar to Stone (1988)'s approach to approximating a thin plate spline. Instead of penalizing the roughness directly, the smoothness penalization in our model is achieved by the prior of spline coefficients (15). The B-spline approach was later criticized in spatial smoothing and interpolation for regions with irregular shapes and boundaries, and its alternatives include products of P-splines (Eilers and Marx, 1996), soap-film splines (Wood et al., 2008), and spatial spline regression (Sangalli et al., 2013). In our data (Figure 1), the outer boundary of the sampled wells is approximately convex, but there do exist several holes inside, because the wells are mainly located around the residential settlements, which are clustered and separated by open rice fields. Nevertheless, these human-settlement-holes do not define hard physical barriers for groundwater flows, hence we adopt simple Euclidean distance without extra modeling of the holes.

Our analysis relies on several assumptions. First, we modeled the log scale of the well As concentration and a constant observational error. We make this log transformation based on evidence from the residual plots in Figures 3 and 6: After the log transformation of observations, the error term has similar scales for different observations, which otherwise is highly skewed in the linear scale. That said, we do not expect the observational error to be of exact same variance after the log transformation. This heterogeneity of variance may still exist, although it s partially remedied by a flexible regression form (Gaussian process or bivaraite splines in our model). Second, we learn the differentiation equation of the log As surface with one-step time-discretization (12 years). This is only reasonable to explain the long term dynamic. We expect to better model the moderate the short term variation by through more frequent monitoring and sampling in future research. Third, the assumed dynamic (13) is isotropic, while the groundwater flow can be more complicated based on local geological structures. Fourth, the θ_1-varying mixing effect coefficient is only reasonable for the interval where θ_1 is supported in observations. The claim on extreme values ($>300\,\mu g/L$) relies on extrapolations. Lastly, we model the dependence on well depth, but do not model other potential selection mechanism of well-reinstallation, a potential confounder. Most households reinstalled their wells during the decade separating the two surveys but those who did so with the goal of reducing their arsenic exposure did so outside the depth range disturbed by irrigation pumping that was considered for this analysis (Jamil et al., 2019). The shallower well reinstallations analyzed here are therefore less likely to be biased.

5.2. Practical implications of inferred spatiotemporal trends

Our analysis shows considerable variability in arsenic concentrations, but fortunately, no indication from the two larger surveys conducted a decade apart that low-arsenic portions of the shallow aquifer became systematically contaminated over time. In particular, for wells meeting the local safety range (As $<50\,\mu g/L$), the inferred mixing dynamic suggests they are unlikely to be elevated due to neighboring influence. Consequently, the well-switching should be more systematically encouraged in Araihazar and many other parts of Bangladesh.

This mixing effect could have been a concern because groundwater flow patterns have been thoroughly altered by irrigation pumping drawing large volumes of water from shallower aquifers during each winter season throughout the region (Harvey et al., 2006).

The documented decline in mean arsenic for the study region is of considerable interest as well because it provides new evidence that accelerated groundwater pumping has been withdrawing arsenic from the shallow aquifer, albeit at the cost of redepositing this arsenic in rice fields (Dittmar et al., 2007). These withdrawals are compensated each year at the onset of the monsoon by recharge of low-arsenic surface water. The resulting flushing of the shallow aquifer has evidently only partially been compensated by a release of arsenic from arsenic sands through a previously proposed exchange process (van Geen et al., 2008; Mozumder, 2019). The much larger pool of arsenic present in aquifer sediments compared to the pool of arsenic in groundwater has therefore delayed the decline in groundwater arsenic concentrations due to flushing but has probably also anchored the distribution of arsenic to the local geology. This has prevented greater convergence of concentrations to the areal mean due to the turning on and off of irrigation pumps. Such conclusions could not have been reached with a detailed statistical analysis of the available data.

5.3. The merit of field kits

In agreement with with previous findings (van Geen et al., 2003; Mailloux et al., 2020), our model confirms a considerable variation in individual well As, both spatially (between neighboring wells or villages, explained by σ_{obs} in (7)), and temporally (despite an overall trend, individual wells are likely to drop or incline over a decade, explained by τ in (12)).

The field kit measures As concentrations with much less precision compared to laboratory measurements, exhibiting both large bias and variance. On one hand, this measurement error amplifies the already-large noise-to-signal ratio in observations, on top of the spatial and temporal variations. Takings the face value of field kit measurements increases the chance of incorrect and inconsistent labeling. That said, laboratory precision is not necessary when concentrations range across several orders of magnitude (van Geen et al., 2005). Health-based thresholds are often somewhat arbitrary. The health impacts of drinking water that contains 9 or $11\,\mu g/L$ As are to first approximation a linear function of exposure and therefore not that different, for instance. What is more important is for rural residents in Bangladesh to know if their well As is closer to 1 or $100\,\mu g/L$.

On the other hand, field kits comes with the distinct advantage that the cost of the measurement is an order of magnitude lower and, perhaps, more importantly that the result can be delivered on the spot. Bringing a sample to the laboratory and the result back to the well owner is logistically much more complicated. By incorporating information from the quality-control sample, spatial smoothing, and dynamic pattern over time, our statistical model calibrates the individual filed kit results and makes personalized probabilistic prediction for each well (Figures 7 and 8), which is helpful for residents to distinguish current safe and unsafe wells, and instructive for determining future well-installation locations.

Looking forward, either for the purpose of better understanding the dynamic pattern or monitoring individual household exposures, the high variability in groundwater As or other toxicants in the environment calls for more frequent and extensive sampling. Imprecise but widely-accessible field kit tests in companion with flexible statistical modeling that facilitates this open-ended data gathering can provide a balance between total cost and accuracy in many areas of geoscience research and policy.

Acknowledgements

Data collection in the field and in the laboratory was supported in part by NIEHS Superfund Research Program grant P42 ES010349 and NSF grant ICER 1414131. We also thank the National Science Foundation, Institute of Education Sciences, Office of Naval Research, National Institutes of Health, Sloan Foundation, and Schmidt Futures for financial support and Daniel Simpson for helpful discussions.

References

Chen, Y., van Geen, A., Graziano, J. H., Pfaff, A., Madajewicz, M., Parvez, F., Hussain, A. I., Slavkovich, V., Islam, T., and Ahsan, H. (2007). Reduction in urinary arsenic levels in response to arsenic mitigation efforts in Araihazar, Bangladesh. *Environmental Health Perspectives*, 115:917–923.

Cheng, Z., Zheng, Y., Mortlock, R., and van Geen, A. (2004). Rapid multi-element analysis of groundwater by high-resolution inductively coupled plasma mass spectrometry. *Analytical and Bioanalytical Chemistry*, 379:512–518.

Dhar, R., Zheng, Y., Stute, M., van Geen, A., Cheng, Z., Shanewaz, M., Shamsudduha, M., Hoque, M., Rahman, M., and Ahmed, K. (2008). Temporal variability of groundwater chemistry in shallow and deep aquifers of Araihazar, Bangladesh. *Journal of Contaminant Hydrology*, 99:97–111.

Dittmar, J., Voegelin, A., Roberts, L. C., Hug, S. J., Saha, G. C., Ali, M. A., Badruzzaman, A. B. M., and Kretzschmar, R. (2007). Spatial distribution and temporal variability of arsenic in irrigated rice fields in Bangladesh. *Environmental Science & Technology*, 41:5967–5972.

Eilers, P. H. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, pages 89–102.

Fendorf, S., Michael, H. A., and van Geen, A. (2010). Spatial and temporal variations of groundwater arsenic in south and southeast Asia. *Science*, 328:1123–1127.

Flanagan, S. V., Johnston, R. B., and Zheng, Y. (2012). Arsenic in tube well water in Bangladesh: Health and economic impacts and implications for arsenic mitigation. *Bulletin of the World Health Organization*, 90:839–846.

Gelman, A., Trevisani, M., Lu, H., and Van Geen, A. (2004). Direct data manipulation for local decision analysis as applied to the problem of arsenic in drinking water from tube wells in Bangladesh. *Risk Analysis*, 24:1597–1612.

George, C. M., Zheng, Y., Graziano, J. H., Rasul, S. B., Hossain, Z., Mey, J. L., and van Geen, A. (2012). Evaluation of an arsenic test kit for rapid well screening in Bangladesh. *Environmental Science & Technology*, 46:11213–11219.

Harvey, C. F., Ashfaque, K. N., Yu, W., Badruzzaman, A., Ali, M. A., Oates, P. M., Michael, H. A., Neumann, R. B., Beckie, R., and Islam, S. (2006). Groundwater dynamics and arsenic contamination in Bangladesh. *Chemical Geology*, 228:112–136.

Harvey, C. F., Swartz, C. H., Badruzzaman, A., Keon-Blute, N., Yu, W., Ali, M. A., Jay, J., Beckie, R., Niedan, V., and Brabander, D. (2002). Arsenic mobility and groundwater extraction in Bangladesh. *Science*, 298:1602–1606.

Jamil, N. B., Feng, H., Ahmed, K. M., Choudhury, I., Barnwal, P., and van Geen, A. (2019). Effectiveness of different approaches to arsenic mitigation over 18 years in Araihazar, Bangladesh: implications for national policy. *Environmental Science & Technology*, 53:5596–5604.

Korfmacher, K. S. and Dixon, S. (2007). Reliability of spot test kits for detecting lead in household dust. *Environmental Research*, 104:241–249.

Landes, F. C., Paltseva, A., Sobolewski, J. M., Cheng, Z., Ellis, T. K., Mailloux, B. J., and van Geen, A. (2019). A field procedure to screen soil for hazardous lead. *Analytical Chemistry*, 91:8192–8198.

Mailloux, B. J., Procopio, N. A., Bakker, M., Chen, T., Choudhury, I., Ahmed, K. M., Mozumder, M. R. H., Ellis, T., Chillrud, S., and van Geen, A. (2020). Recommended sampling intervals for arsenic in private wells. *Groundwater*.

McArthur, J., Banerjee, D., Sengupta, S., Ravenscroft, P., Klump, S., Sarkar, A., Disch, B., and Kipfer, R. (2010). Migration of As, and $^3H/^3He$ ages, in groundwater from West Bengal: implications for monitoring. *Water Research*, 44:4171–4185.

Mihajlov, I., Mozumder, M. R. H., Bostick, B. C., Stute, M., Mailloux, B. J., Knappett, P. S., Choudhury, I., Ahmed, K. M., Schlosser, P., and van Geen, A. (2020). Arsenic contamination of Bangladesh aquifers exacerbated by clay layers. *Nature Communications*, 11:1–9.

Mozumder, R. H. (2019). *Impacts of pumping on the distribution of arsenic in Bangladesh groundwater*. PhD thesis, Columbia University.

Ramsay, J. O., Hooker, G., Campbell, D., and Cao, J. (2007). Parameter estimation for differential equations: a generalized smoothing approach. *Journal of the Royal Statistical Society: B*, 69:741–796.

Sangalli, L. M., Ramsay, J. O., and Ramsay, T. O. (2013). Spatial spline regression models. *Journal of the Royal Statistical Society: B*, pages 681–703.

Stan Development Team (2020). *Stan Modeling Language Users Guide and Reference Manual*. Version 2.25, `http://mc-stan.org`.

Stone, G. (1988). *Bivariate splines*. PhD thesis, University of Bath.

van Geen, A., Ahmed, E. B., Pitcher, L., Mey, J. L., Ahsan, H., Graziano, J. H., and Ahmed, K. M. (2014). Comparison of two blanket surveys of arsenic in tubewells conducted 12 years apart in a 25 km^2 area of Bangladesh. *Science of the Total Environment*, 488:484–492.

van Geen, A., Ahsan, H., Horneman, A. H., Dhar, R. K., Zheng, Y., Hussain, I., Ahmed, K. M., Gelman, A., Stute, M., Simpson, H. J., et al. (2002). Promotion of well-switching to mitigate the current arsenic crisis in Bangladesh. *Bulletin of the World Health Organization*, 80:732–737.

van Geen, A., Cheng, Z., Seddique, A., Hoque, M., Gelman, A., Graziano, J., Ahsan, H., Parvez, F., and Ahmed, K. (2005). Reliability of a commercial kit to test groundwater for arsenic in bangladesh. *Environmental science & technology*, 39:299–303.

van Geen, A., Zheng, Y., Goodbred Jr, S., Horneman, A., Aziz, Z., Cheng, Z., Stute, M., Mailloux, B., Weinman, B., and Hoque, M. (2008). Flushing history as a hydrogeological control on the regional distribution of arsenic in shallow groundwater of the Bengal Basin. *Environmental Science & Technology*, 42:2283–2288.

van Geen, A., Zheng, Y.-J., Versteeg, R., Stute, M., Horneman, A., Dhar, R., Steckler, M., Gelman, A., Small, C., and Ahsan, H. (2003). Spatial variability of arsenic in 6000 tube wells in a 25 km^2 area of Bangladesh. *Water Resources Research*, 39.

Wood, S. N., Bravington, M. V., and Hedley, S. L. (2008). Soap film smoothing. *Journal of the Royal Statistical Society: B*, 70:931–955.

A. Supplementary material

A.1. Replication data and code

Available at `https://github.com/yao-yl/As-measurement-code`.

A.2. Graphical representation of models

Figure 13 and 14 summarize the model for 271 re-identified wells (section 2.2), and the model for blanket survey data (section 3.2, the main model) in graphs.

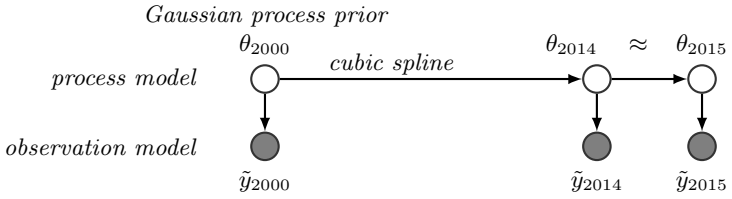

Figure 13: *Graphical summary of the model for 271 resampled wells. The observed arsenic concentration $\tilde{y}$ is modeled by a noisy realization of the baseline value θ, and θ is placed a Gaussian process prior for spatial smoothness. The change from θ_{2010} to θ_{2014} and θ_{2015} is modeled by an auto regression with cubic splines.*

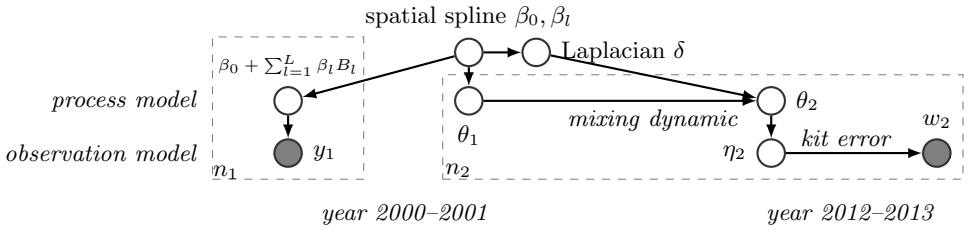

Figure 14: *Graphical summary of the model for blanket survey data (section 3.2). We model the spatial distribution by a spline, and the temporal dynamic is modeled by an autoregression on the initial baseline value and the Laplacian. The field kit is calibrated back to the hypothetical lab reading η_2, and further linked to the baseline value θ_2 via a Gaussian noise.*

A.3. Knots of bivariate B-spline

In section 3.2, the spatial modeling (7) relies on bivariate B-splines. We place the inner knots on a uniform grid spanned by equally-spaced vertical coordinates $\tilde{x}^{\mathrm{N}}_1, \ldots, \tilde{x}^{\mathrm{N}}_{N_1}$ and horizontal coordinates $\tilde{x}^{\mathrm{E}}_1, \ldots, \tilde{x}^{\mathrm{E}}_{N_2}$ such that $\tilde{x}^{\mathrm{N}}_{i+1} - \tilde{x}^{\mathrm{N}}_i = \tilde{x}^{\mathrm{E}}_{j+1} - \tilde{x}^{\mathrm{E}}_j = x_0$. In other words, the inner knots are placed on the vertexes of $x_0 \times x_0$ blocks.

The spacing is chosen such that there are 30 inner knots of longitudes, which results in $x_0 = 293$ meters, $N_1 = 30$, and $N_2 = 22$. For a comparison, the posterior mean of the length scale ρ from the Gaussian process regression (2) is estimated to be 381 meters.

On each coordinate, we construct the usual cubic spline basis functions with knots $\tilde{x}^{\mathrm{N}}_1, \ldots, \tilde{x}^{\mathrm{N}}_{N_1}$ or $\tilde{x}^{\mathrm{E}}_1, \ldots, \tilde{x}^{\mathrm{E}}_{N_2}$. The number of basis functions on each dimension is $N_1 + 3 = 33$ and $N_2 + 3 = 25$.

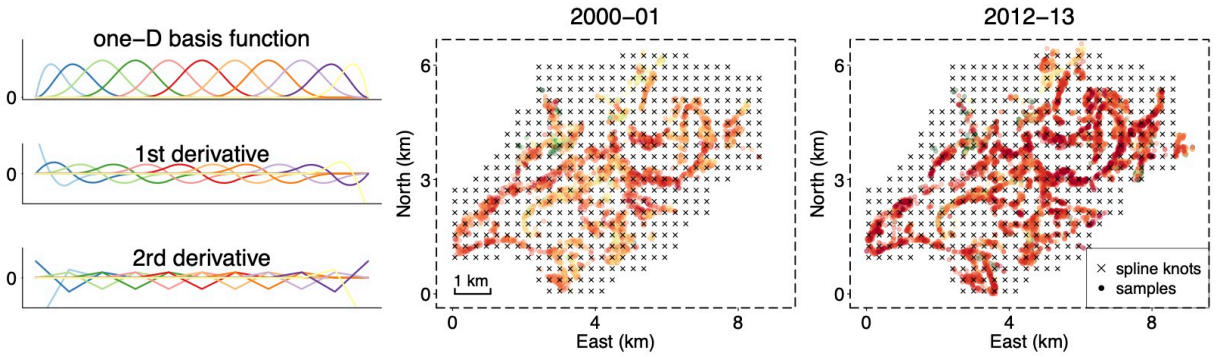

Figure 15: *Left: Illustration of one-dimensional cubic spline basis functions and their first two derivatives. Right two panels: We build the bivariate splines from the tensor product of one dimensional B-splines. The knots are represented by black crosses on the graph. They are equally-spaced, and trimmed outside sample range.*

We consider the tensor product of these two dimensions,

$$B_l(x_{2i}^{\mathrm{N}}, x_{2i}^{\mathrm{E}}) = B_l^{\mathrm{N}}(x_{2i}^{\mathrm{N}}) B_l^{\mathrm{E}}(x_{2i}^{\mathrm{E}}),$$

where l is the rearranged index of the product basis functions, ranging from 1 to $25 \times 33 = 825$.

We prune the knots by remove the blocks not covering any observed wells from blanket surveys. The pruned knots have holes inside, and we remove these holes by completing the convex hull of the pruned knots. After this procedure, the number of not-constant-zero basis function is $L = 485$. We relabel the index l from 1 to 485. We plot the trimmed knots with reference to observations in Figure 15.

The basis function is evaluated at location x_{1i} and x_{2i}, which are stored by two matrices $B(x_1)$ and $B(x_2)$ with dimension 4574×485 and 8295×485 respectively.

Finally we compute the Laplacian of basis functions. This is straightforward for tensor products for

$$\Delta B_l(x_{2i}^{\mathrm{N}}, x_{2i}^{\mathrm{E}}) = \Delta B_l^{\mathrm{N}}(x_{2i}^{\mathrm{N}}) B_l^{\mathrm{E}}(x_{2i}^{\mathrm{E}}) + \Delta B_l^{\mathrm{E}}(x_{2i}^{\mathrm{E}}) B_l^{\mathrm{N}}(x_{2i}^{\mathrm{N}}).$$

We store this product $\Delta B(x_2)$, Laplacian basis evaluated at x_2, by a 8295×485 matrix.

The matrix of basis functions $B(x_1)$, $B(x_2)$ and $\Delta B(x_2)$ only come into the log joint density via matrix multiplication, during which employ sparse matrix algebra.

A.4. Imputation result of pointwise Laplacian

Figure 16 displays the posterior mean of the Laplacian δ_i from the fitted blanket survey data. A positive Laplacian means the well has lower As value than its neighboring wells. The Laplacian is not scaling invariant. In our computation, we standardize the input matrix ΔB by a factor 1000, and standardize the location coordinate $x^{\mathrm{N}}, x^{\mathrm{E}}$ by a factor 9.1 km (so as to make them unit-scaled).To interpret the scale of the Laplacian, we use the heuristic from the finite difference approximation

$$\begin{aligned}\Delta\theta\left(x_2^{\mathrm{N}}, x_2^{\mathrm{E}}\right) \approx & \frac{1}{4}h^2\big(\theta\left(x_2^{\mathrm{N}}+h, x_2^{\mathrm{E}}\right)+\theta\left(x_2^{\mathrm{N}}-h, x_2^{\mathrm{E}}\right) \\ & +\theta\left(x_2^{\mathrm{N}}, x_2^{\mathrm{E}}+h\right)+\theta\left(x_2^{\mathrm{N}}, x_2^{\mathrm{E}}-h\right)-4\theta\left(x_2^{\mathrm{N}}, x_2^{\mathrm{E}}\right)\big).\end{aligned}$$

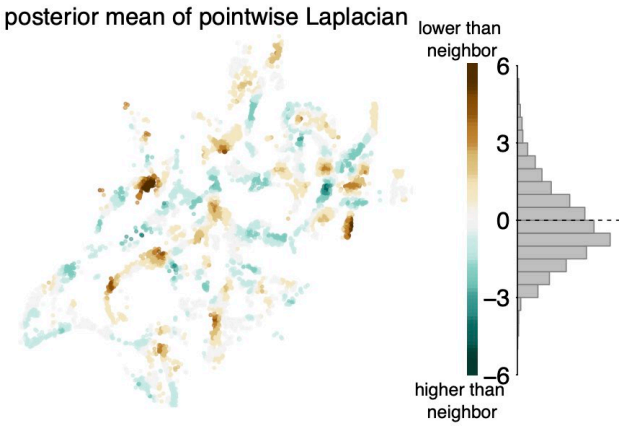

Figure 16: *The posterior mean of the Laplacian δ_i in year 2000. A positive Laplacian means the well has lower As value than its neighboring wells. On the right is the histogram of the posterior mean of well-wise Laplacian.*

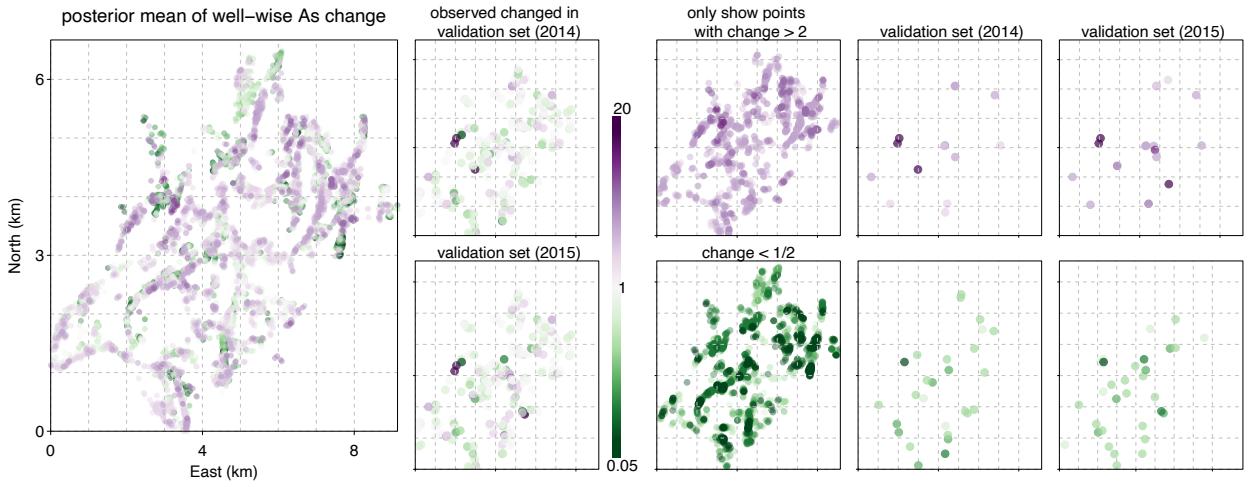

Figure 17: *Column (1) The posterior mean of inferred log As changes of n_2 individual well: $\theta_{2i}-\theta_{1i}$. (2) The observed log changes in the 271 resampled well. (3-5) The subset of wells with inferred or observed changes more than doubling (first row) or less than halving (second row). The fitted model is able to match certain local patterns in the validation set.*

Using the average closest distance in the second survey, $h = 28$ meters, a unit scale of Laplacian is approximately $\exp(1000(28/9100)^2) = 1.01$ in multiplicative factor, i.e., Laplacian$= \pm 1$ equivalents being 1% lower/higher than all closest well in the neighborhood.

One caveat is that the pointwise value of the Laplacian δ_i is unlikely to be estimated accurately, not only because the latent variable θ is not observable, but also because the spline fit implicitly penalizes the roughness. That said, we view the fitted δ_i as an extracted feature to represent the neighboring difference.

Posterior predictive check of the inferred Laplacian. How reliable is our conclusion concerning the diffusion-like process of homogenisation of arsenic concentrations? The challenge in checking the model is that all associated quantities describing the process, the baseline log As value θ_1, θ_2 and the pointwise Laplacian, are latent variables. Nevertheless, we have already seen a consistent

pattern in Figure 5 and 9. We further check the model claim by visually checking if the model is able to pick the spatially local change pattern. The leftmost panel in Figure 17 visualizes the posterior mean of inferred log As changes of n_2 individual well: $\theta_{2i} - \theta_{1i}$, and the second column is the observed log changes in the 271 resampled well. To reduce overlapping, we also draw the subset of wells that have inferred or observed changes more than doubling or less than halving. Although we cannot completely separate noise, the fitted model is able to match certain local patterns in the validation set. The local regions with drastic changes are colored in agreement in the fitted model.

A.5. Dependence on well depth

The first and third panel in Figure 18 displays the relation between well depth and log As concentration in the resampled dataset and blanket survey 2000. A direct univariate regression has coefficient 0.12, which ignores the spatial dependence on well depth. In our model, the posterior mean of β_{depth} is 0.03 and 0.064 (95% CI: (0.057, 0.07)) in the resampled and blanket-survey dataset. The latter one has smaller standard deviation for its larger sample size.

It is previously known that among shallow (<30 m) wells, the well As is positively correlated with well depth even after controlling the spatial variation (Gelman et al., 2004), although this correlation does not necessarily imply a causal relation between them.

Our model for blanket surveys contains both the spatial distribution and the depth dependence. The well-wise are therefore the controlled-comparison: the inferred well-wise change given the same well depth. Overall the depth dependence is weak compared with spatiotemporal variations. From 2000–2001 to 2012–2013, the average well depth decreased by 0.23 meters, which adds to $1 - \exp(-0.2 \times 0.06) = 1\%$ decrease in the sample mean As concentration.

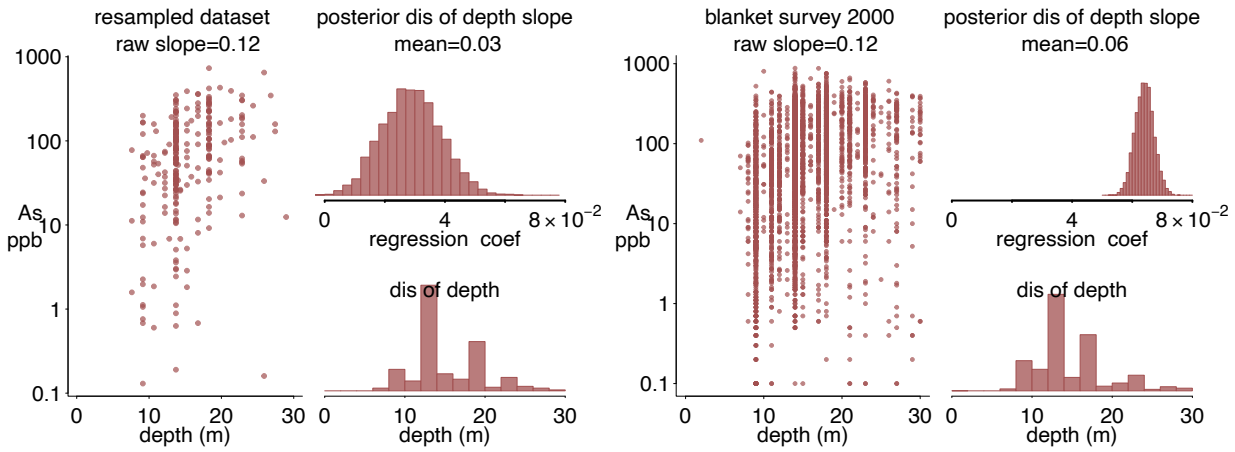

Figure 18: *Dependence of well-As on well depth. First column: the observed As concentration-well depth in the 271 resampled dataset in 2015. Second column: the posterior distribution of* β_{depth} *from the first model and resampled dataset, and the histogram of the well depth in the sample. Right two columns: the same scatterplot and histogram, inferred using the second model and larger dataset.*