June 20, 2001

Characterization of Arsenic Occurrence in US Drinking Water Treatment Facility Source Waters (Revised Version)

J.R. Lockwood III, Mark J. Schervish, Patrick Gurian, and Mitchell J. Small

The 1996 amendments to the US Safe Drinking Water Act (SDWA) mandate revision of current maximum contaminant levels (MCLs) for various harmful substances in public drinking water supplies. The determination of a revised MCL for any contaminant must reflect a judicious compromise between the potential benefits of lowered exposure and the feasibility of obtaining such levels. This evaluation is made as part of a regulatory impact assessment (RIA) requiring detailed information about the occurrence of the contaminant and the costs and efficiencies of the available treatment technologies. Our work focuses on the first step of this process, using a collection of data sources to model arsenic occurrence in treatment facility source waters as a function of system characteristics such as source water type, location and size. We fit Bayesian hierarchical models to account for the spatial aspects of arsenic occurrence as well as to characterize uncertainty in our estimates. After model selection based on cross-validation predictive densities, we use a national census of treatment systems and their associated covariates to predict the national distribution of raw water arsenic concentrations. We then

J.R. Lockwood III is Graduate student and Mark J. Schervish is Professor, Department of Statistics, Patrick Gurian is Graduate student, Department of Engineering and Public Policy, and Mitchell J. Small is Professor, Departments of Engineering and Public Policy and Civil and Environmental Engineering, Carnegie Mellon University, Pittsburgh, PA 15213. The support and guidance provided by Judy Calem and Ben Smith of the US EPA Office of Ground Water and Drinking Water, Standards and Risk Management Division are gratefully acknowledged. This paper has not undergone US EPA peer review, and no official endorsement should be inferred This research was supported by US EPA award No. R826890-01-0 and Cooperative Agreement CR825188-01-3 between US EPA Office of Policy, Planning and Evaluation (OPPE) and Carnegie Mellon University.

examine the relationship between proposed MCLs and the number of systems requiring treatment and identify classes of systems which are most likely to be problematic. The posterior distribution of the model parameters, obtained via Markov Chain Monte Carlo, allows us to quantify the uncertainty in our predictions.

KEY WORDS: Bayesian analysis; censored data; cross-validation; hierarchical models; Markov chain Monte Carlo; predictive density.

---

1. INTRODUCTION

The overwhelming majority of U.S. residents receive their drinking water from one of approximately 55,000 community water suppliers. These systems collect raw water from both surface and ground water sources, treat it to remove various chemical and microbial contaminants, and distribute the finished water to consumers. The Safe Drinking Water Act of 1974 empowers the U.S. Environmental Protection Agency (USEPA) to mandate uniform national upper bounds, or maximum contaminant levels (MCLs), on the concentrations of contaminants in the waters distributed by community water systems. The 1996 amendments to the Safe Drinking Water Act establish more rigorous guidelines for identifying and regulating problematic substances by requiring USEPA to publish a cost–benefit analysis, or regulatory impact assessment, comparing the merits of proposed MCLs for each potentially regulated contaminant (USEPA 1998). Along with establishing this protocol, the amendments also require re-evaluation of the MCLs for all currently regulated substances.

The MCL for arsenic was one of the first to be revisited because of growing concerns about its associated health risks. Historically, high arsenic concentrations in drinking water have been linked to various types of cancer, especially that of the skin, stomach, and bladder

(Chen, Chen, Wu, and Kuo 1992; Pontius, Brown, and Chen 1994; Hopenhayn-Rich et al. 1996). Data suggested that the previous arsenic MCL of 50 $\mu g/L$ was not low enough to adequately reduce exposure to the carcinogen (National Research Council 1999). In June 2000, USEPA proposed lowering the existing MCL to 5 $\mu g/L$, and accepted public comment on possible MCLs of 3 $\mu g/L$, 10 $\mu g/L$ and 20 $\mu g/L$ for ninety days following (USEPA 2000a). The original statutory deadline of January 1, 2001 for finalizing the standard was extended until June 2001 (PL 106-377), giving USEPA additional time to consider the diverse opinions of the stakeholders. On January 16, 2001, a standard of 10 $\mu g/L$ was adopted.

The determination of a revised MCL for any contaminant must reflect a judicious compromise between the potential benefits of lowered exposure and the feasibility of obtaining such levels. Of particular concern is the cost of augmenting existing treatment technology in order to ensure compliance with a lower MCL. While examination of concentrations in finished water contributes substantially to the assessment of these costs, an analysis of raw water is equally meaningful. For example, given that the finished water of a system fails to meet a specified standard, determining the class of treatments that will redress the problem requires specific information about both contaminant occurrence in raw water and the removal efficiencies of various treatment options. This paper focuses on the characterization of raw water arsenic occurrence as a first step in the decision process.

A complete assessment of raw water arsenic occurrence in U.S. community water supplies ideally would be based on repeated samples from the source waters of each system. However, such data are generally unavailable because all federal water quality standards are based on finished water. Hence, the limited resources available to systems for analysis are focused on systematically sampling finished water in order to monitor compliance with USEPA standards. The practical import is that a quantification of raw water arsenic concentrations must be based on limited observational data, typically consisting of a single sample from only a small subset of water systems. A critical component of the estimation of compliance costs is thus one which is also difficult to characterize.

Fortunately, Bayesian methods can help to bridge the gap between sparse observational data and a comprehensive summary of national raw water arsenic levels by allowing various sources of variability and uncertainty in occurrence to be explicitly quantified. This type of analysis allows us to assess how much information is lost by not having more complete data. In this way, we can identify the components of uncertainty which are most important to resolve, and we can make suggestions about how to improve future data collection. The purpose of our analysis is thus twofold: first, we will model arsenic occurrence as a necessary step in evaluating alternative MCLs; second, we hope to exhibit methodology which is valuable for similar types of analyses and is able to inform the data collection process.

In Section 2 we outline our methods for estimating a national distribution of arsenic from a model which combines the available occurrence and system data. We then detail the application of our methods in Section 3, and provide selected results and predictions in Section 4. A discussion of our findings as well as implications for future research are presented in Section 5.

2. METHODOLOGY

Our general approach is to use several available data sets (to be described in Section 3) to model arsenic concentrations in the raw waters of community water systems as a function of other available information about the systems. The value of a model based on system covariates is that it allows prediction of concentrations for systems with a given covariate structure. In particular, if characteristics of every water system are known, then prediction based on this set of covariates can be used as a surrogate for a comprehensive survey of raw water arsenic occurrence. This estimate of the national arsenic occurrence distribution, coupled with the uncertainties surrounding it, is an informative prerequisite for an assessment of current exposure and the potential costs of reaching compliance with a revised MCL.

2.1 Deriving a Model to Describe Raw Water Arsenic in Sampled Systems

Consider a collection of comparable data sets which, when combined, represent n_0 water systems with covariates $\boldsymbol{x_0} = (\boldsymbol{x_{0,1}}, \ldots, \boldsymbol{x_{0,n_0}})$. With each system is also associated a measurement $Y_{0,i}$ of arsenic with observed value $y_{0,i}$. Let $\boldsymbol{Y_0} = (Y_{0,1}, \ldots, Y_{0,n_0})$ and let the observed values be $\boldsymbol{y_0} = (y_{0,1}, \ldots, y_{0,n_0})$. We suppose that $\boldsymbol{Y_0}$ has a density $f_0(\boldsymbol{y_0}|\boldsymbol{x_0}, \theta)$ where θ is a vector of unknown parameters quantifying various model features. For example, θ could include parameters which describe the relationship between arsenic concentration and system location or the type of source water used by the system. To quantify prior uncertainty about the parameter, suppose that θ has a prior distribution with density $p(\theta|\eta)$, where η is a vector of hyperparameters which is conditionally independent of the data given θ and which has prior distribution $p(\eta)$. Bayes Rule allows calculation of the posterior distribution of the model parameters by $p(\theta, \eta|\boldsymbol{y_0}, \boldsymbol{x_0}) \propto f_0(\boldsymbol{y_0}|\boldsymbol{x_0}, \theta)p(\theta|\eta)p(\eta)$, from which Markov chain Monte Carlo (MCMC) methods (Gilks, Richardson, and Spiegelhalter 1996) can be used to generate a sample $(\theta_1, \eta_1), \ldots, (\theta_m, \eta_m)$. Inferences about any function of the joint posterior distribution, or the marginal posterior distribution $p(\theta|\boldsymbol{y_0}, \boldsymbol{x_0})$, can be made with this sample.

2.2 Extending the Model to Estimate Raw Water Arsenic Concentrations in All Systems

Now consider an additional collection of n_1 systems with associated covariates $\boldsymbol{x_1} = (\boldsymbol{x_{1,1}}, \ldots, \boldsymbol{x_{1,n_1}})$ representing systems for which we wish to predict some, as yet unobserved, arsenic concentrations $\boldsymbol{Y_1} = (Y_{1,1}, \ldots, Y_{1,n_1})$. Although $\boldsymbol{x_1}$ in general could represent any collection of systems, we focus on the case in which $\boldsymbol{x_1}$ consists of the system characteristics for all systems under federal regulation. This case is most important to us as we attempt to simulate the results of a national assessment of raw water concentrations. Assume that $\boldsymbol{Y_1}$ is conditionally independent of the observed data given θ and that $\boldsymbol{Y_1}$ has a density $f_1(\boldsymbol{y_1}|\boldsymbol{x_1}, \theta)$, where θ is the same parameter describing the distribution of $\boldsymbol{Y_0}$. After observing $\boldsymbol{Y_0} = \boldsymbol{y_0}$ and $\boldsymbol{x_0}$, θ has density $p(\theta|\boldsymbol{y_0}, \boldsymbol{x_0})$. Because the density f_1 is a function of the parameter θ, features of its posterior distribution can be determined from a sample $\theta_1, \ldots, \theta_m$ from $p(\theta|\boldsymbol{y_0}, \boldsymbol{x_0})$. The underlying approach is to allow the sampled systems to update prior beliefs about the parameter θ, then to use the revised beliefs to quantify uncertainty about features of the model for all systems.

A feature of the posterior distribution of $f_1(\boldsymbol{y_1}|\boldsymbol{x_1}, \theta)$ that is of particular concern is what would be a likely result of a single-sample assessment of raw water arsenic from every community water system in the country. A survey of this kind, combined with additional information, could form the basis for determining arsenic occurrence levels in the exploration of alternative MCLs. One can easily simulate the results of such a study with the current approach, as well as quantify the uncertainty surrounding the outcome. To simulate the posterior distribution of the results of a single-sample assessment of all community water systems, one could first obtain a sample $\theta_1, \ldots, \theta_m$ from $p(\theta|\boldsymbol{y_0}, \boldsymbol{x_0})$ via MCMC. For each sampled parameter θ_i, one can simulate observed arsenic concentrations $\boldsymbol{y_{1i}}$ for the systems by sampling from $f_1(\boldsymbol{y_1}|\boldsymbol{x_1}, \theta_i)$. At this point any calculation based on observed data which would be of interest, such as the number of systems exhibiting a concentration above a certain MCL, can be performed. In addition, the distribution of the simulated results over the sample $\theta_1, \ldots, \theta_m$ can be used to quantify the uncertainty in the outcome of the experiment fostered by not knowing the precise value of the parameter.

3. APPLICATION

Our underlying objective is to build a model describing arsenic occurrence using observed concentration data and then to use the conditional distribution of the parameters given the data together with the covariates of all national utilities to estimate features of the national raw water arsenic concentration distribution.

3.1 Data Description

Several data sets addressing raw water arsenic occurrence have been collected by USEPA, independent consulting firms, and groups of community water suppliers (Wade Miller Associates, Inc. 1992; USEPA 1994; Frey, Edwards, and Amy 1997.) The primary data set we used to build the occurrence model resulted from the National Arsenic Occurrence Survey (NAOS) (Frey et al. 1997; Frey and Edwards 1997), conducted in the mid-1990s. The NAOS researchers selected a stratified random sample of 809 U.S. community water systems based on system characteristics such as location, type of water used (community systems are classified as either surface water or ground water systems depending on the type of source water used) and number of people served. These systems were sent sample bottles and instructions, and the 517 responding systems resulted in 457 valid arsenic measurements. The NAOS data set provided an attractive basis for a raw water arsenic model for a number of reasons. First, because of the stratified study design, the NAOS achieved nearly comprehensive coverage of the major classes of water suppliers. Moreover, the study is comparatively recent and samples were analyzed at the same laboratory with a highly sensitive measurement technique. Despite the voluntary response survey design, the designers believe the integrity of the sample was maintained, and they present empirical evidence supporting this conclusion (Frey and Edwards 1997).

The other occurrence data that we used were provided by the United States Geological Survey (USGS) as part of the National Water Information System (United States Geological Survey 1997). The USGS data set, collected from 1976 to 1996, provides a single raw water arsenic measurement for each of nearly 2000 drinking water wells serving community water systems. Hence, while providing more observations than the NAOS, the USGS data set is less representative because it consists entirely of measurements from ground water systems.

Both the NAOS and USGS data are subject to left censoring. For each sampled system we learn either the observed concentration or the fact that the concentration is believed to be less than some number which we call the "censoring point." For both of our data sets, this number represents the lowest concentration that could be reliably quantified. Thus we have reasonable confidence that a reported concentration above the censoring point is an accurate representation of the underlying arsenic concentration. In the absence of any other information, we treated all censored observations as being known to be less than the associated censoring point. The censoring point for the NAOS observations is 0.5 $\mu g/L$, which is among the lowest of those used in existing arsenic occurrence assessments (Frey and Edwards 1997). For reference, among the 457 NAOS systems, 163 have censored concentrations. For the 1945 USGS sampled systems, the censoring point is 1 $\mu g/L$ and there are 789 censored observations. Additional details about censoring are provided in Appendix A.

Finally, the data set providing the water system characteristics needed to make inferences about national arsenic occurrence is the Safe Drinking Water Information System (SDWIS) (USEPA 2000b). SDWIS offers no observed arsenic measurements, but instead provides a comprehensive census of U.S. community water suppliers including information on source water type, size and location by state of all suppliers in the nation. Initial modeling is based on the NAOS data, while the USGS data are used first to validate, then to update the parameter distributions. SDWIS is then used as the source of the population distribution of covariates for predictions and simulations.

3.2 Model Description

Exploratory analysis of the NAOS data set as well as input from several experts in water quality analysis (Wade Miller Associates, Inc. 1992; Frey and Edwards 1997; Frey et al. 1997; Cadmus Group Inc. 1997) indicated that the following factors possibly covary with source water arsenic concentrations:

- Source water type: Most utilities use primarily surface water or ground water. Standard beliefs indicate that these types intrinsically differ with respect to concentrations of contaminants, including arsenic. For example, it may be the case that concentrations of chemicals in surface water tend to be less variable than those in ground water due to the inherent averaging of chemical concentrations over surface water drainage basins (Small 1989).

- Location: Ambient arsenic levels are in large part a function of underlying geological characteristics which vary geographically.

- System size: While probably not directly influential, system size (in terms of the number of people served) is a proxy for other features which may directly impact arsenic levels, such as degree of industrialization and the selectivity and sophistication in choosing source waters.

Aside from information on other contaminants, this essentially exhausts the covariate information available in either the NAOS or USGS data sets. Our approach was to posit a series of increasingly complex models for arsenic occurrence which included these covariates in various forms. We then chose what appeared to be the most appropriate model based on a cross-validation comparison of the models, described in Section 3.5. We next explain the structure of what proved to be the most effective model we considered, leaving the description of the other models, including all of those in which system size was considered, for section Section 3.5.

To establish model notation, let Y_i denote the natural logarithm of arsenic concentration in $\mu g/L$ measured at community water system i for $i = 1, \ldots, n$. We let $X(i)$ indicate the U.S. state in which system i is located and we let $W_i = -1$ if the system i is classified as surface water and $W_i = 1$ if it is classified as ground water. We propose a hierarchical model with the following stages:

Stage 0: For $i = 1, \ldots, n$, $Y_i \sim N(\mu_i, \sigma_i^2)$ independently given μ_i and σ_i. The assumption of lognormality of concentrations is standard in environmetrics and can be justified on a physical basis (Ott 1995). Note that in general, for some systems Y_i may be left censored at some censoring limit c_i. As described in Section 3.1, all systems in the NAOS data set had a common censoring limit of 0.5 $\mu g/L$, while all systems in the USGS data had a common censoring limit of 1 $\mu g/L$.

Stage 1: For $i = 1, \ldots, n$,

$$\mu_i = \alpha^{(m)}_{X(i)} + \beta^{(m)}_{X(i)} W_i \tag{1}$$

$$-2 \log \sigma_i = \alpha^{(v)}_{X(i)} + \beta^{(v)}_{X(i)} W_i \tag{2}$$

(The superscripts "(m)" and "(v)" are mnemonic devices for "mean" and "variance".) This is essentially assuming a one-way ANOVA structure within each state for the means and log-precisions of the arsenic distribution where the factor is the source water type. The design is unbalanced because there are different numbers of ground and surface water systems in each state, both in the data and the population. We think of the four adjustment terms on the right-hand sides of Equation (1) and Equation (2) as being the ith coordinates of vectors with the number of components equal to the number of states. We let $\boldsymbol{\alpha}^{(m)}$ refer to the entire vector $(\alpha_1^{(m)}, \ldots, \alpha_{50}^{(m)})$, and similarly for each of the other three vectors.

Stage 2: The four vectors in Stage 1 all have a similar prior structure. We describe one of them here. Conditional on a vector of prior means $\boldsymbol{\alpha}_0^{(m)}$, a scalar prior precision $\tau_\alpha^{(m)}$ and a positive scalar parameter ρ, we suppose that $\boldsymbol{\alpha}^{(m)}$ has a 50-dimensional multivariate normal distribution with mean vector $\boldsymbol{\alpha}_0^{(m)}$ and covariance matrix $\boldsymbol{C}(\rho)/\tau_\alpha^{(m)}$. For each value of ρ, the matrix $\boldsymbol{C}(\rho)$ is a non-singular 50-dimensional correlation matrix, the structure of which is described in Section 3.3. The parameter ρ and the matrix $\boldsymbol{C}(\rho)$ are common to all four vectors while each vector has its own prior mean and precision. The prior means are fixed hyperparameters but the individual precision parameters are updated by the model and are addressed in Stage 3. Conditional on ρ and all of the individual hyperparameters, the four vectors are independent.

Stage 3: The logarithms of the four precision terms from Stage 2 are assumed to be independent and identically distributed as $N(\mu_\tau, \xi_\tau)$ for fixed hyperparameters μ_τ and ξ_τ. We also suppose that $\log(\rho)$ is distributed $N(\mu_\rho, \xi_\rho)$ for fixed hyperparameters μ_ρ and ξ_ρ. These five parameters are assumed independent *a priori*.

3.3 Description and Rationale of Structure

It is well-known that arsenic levels in water sources vary geographically, suggesting the state-by-state differentiation of arsenic distributions in Stage 1. The hierarchical structure specified in Stage 2 was motivated by considerations of parsimony as well as a desire to account for likely geographic correlation between the values of, for example, $\alpha_{j_1}^{(m)}$ and $\alpha_{j_2}^{(m)}$ for states j_1 and j_2 in close proximity. This led us to impose a distance-based covariance structure through the elements of $\boldsymbol{C}(\rho)$. In particular, the (p, q) entry of the matrix $\boldsymbol{C}(\rho)$ is given by $c_{p,q} = \exp[-\rho d^2(p, q)]$ where $d^2(p, q)$ is the squared distance between the p^{th} and q^{th} states based on estimates of the geographical centroids of the respective states. This parameterization is akin to formulations used in other applications (Cook and Pocock 1983; Conlon and Waller 1998.) Approximate centroids in longitude and latitude were determined from an internal *S-plus* data set (MathSoft 1991) as well as from various mapping programs available on the World Wide Web (e.g., `http://tiger.census.gov` and `http://MapsOnUs.switchboard.com`). Distances in miles based on coordinates were calculated using the Haversine formula for great circle distance (Sinnott 1984). When $p = q$, corresponding to the diagonal of the correlation matrix, the distance is zero and the correlation one. Bochner's theorem (Feller 1971, Section XIX.2) guarantees that the resulting correlation matrix is positive definite.

There are several advantages to a prior covariance structure of this form. First, it constrains the elements of the vectors in such a way that the effective number of parameters can be considerably less than the nominal number. In this way, the prior information reduces the burden of the data in estimating the parameters, which should help to prevent overfitting the data with the richly-parameterized model. Another advantage of the structure is that it allows for "borrowing of strength" in estimation between neighboring states. Intuitively, the extent to which the data from one state should influence the parameter distributions of another depends on both the distance between the states as well the relative amounts of observed data in the states (and surrounding states). In particular, when a state has either no or very few observations, its parameter estimates are shrunk toward the nearest states that have data. The disparity in state-by-state representation in the NAOS data set, a byproduct of the disparity in the actual number of systems in each state and the stratification used in the NAOS, made such a structure attractive. For example, five states are not represented (Hawaii, New Hampshire, Rhode Island, Vermont, West Virginia), while some states (e.g., California and Pennsylvania) have in excess of forty observations.

3.4 Model Implementation

For this model, as well as for all of the more complicated models described in Section 3.5, samples from the posterior distribution of the model parameters were obtained using standard Markov Chain Monte Carlo (MCMC) methods (Gilks et al. 1996) programmed in C. There were several non-standard features of the implementation. The first was the treatment of the censored arsenic observations, handled through data augmentation techniques (Gelman, Carlin, Stern, and Rubin 1995; Tanner and Wong 1987). For each censored observation at each iteration of the Markov chain, we simulated an observation from the conditional distribution of log arsenic given the current values of the model parameters and the fact that the observation was censored. For example, if a ground water observation i was censored at c, we simulated a value y_i^* from the distribution with cumulative distribution function

$$F(y) = \frac{\Phi([y - \alpha^{(m)}_{X(i)} - \beta^{(m)}_{X(i)}] \exp[(\alpha^{(v)}_{X(i)} + \beta^{(v)}_{X(i)})/2])}{\Phi([c - \alpha^{(m)}_{X(i)} - \beta^{(m)}_{X(i)}] \exp[(\alpha^{(v)}_{X(i)} + \beta^{(v)}_{X(i)})/2])},$$

for $-\infty < y < c$. The simulated value y_i^* was used, until the next iteration of the Markov chain, as if it had been the observed value of Y_i. This method of dealing with the censoring considerably simplified some aspects of the MCMC, such as the calculation of the likelihood function.

The other implementation difficulty resulted from the form of the correlation matrix $\boldsymbol{C}(\rho)$. For small values of ρ, the matrix implies very high correlations between certain parameters. For even smaller values of ρ, the matrix is actually numerically singular. Both of these cause a traditional application of the Metropolis-Hastings algorithm (Hastings 1970; Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller 1953) to lead to unacceptably slow mixing of the chain. Our solution to this problem was to reparameterize to vectors of uncorrelated parameters obtained via a linear transformation with a matrix which acts like a Cholesky factor for a singular matrix. Theoretical and implementational details are provided in Appendix B.

3.5 Model Validation

The model detailed in Section 3.2 was introduced as the most effective model we considered. This determination was based on an extensive cross–validation study comparing the predictive performances of a number of models. We had two primary motivations for carrying out such an extensive model comparison analysis. First, the dependence of arsenic concentrations on system size (e.g., population served) is only tenuously justifiable from a scientific standpoint, and it is important to evaluate quantitatively whether or not this covariate provides enough additional predictive power to warrant inclusion in the model. Moreover, one may have noticed that the model detailed in Section 3.2 has a large number of parameters. When dealing with richly parameterized models, it is necessary to justify the additional complexity by demonstrating that the more complicated models actually provide better predictive power than simpler models. In our case, the NAOS data set consists of 457 points, while the model in Section 3.2 has 200 nominal parameters. We will consider models with more than double this number of nominal parameters later in this section. Although the effective number of parameters is considerably restricted by the prior covariance structure, it is still conceivable that we could vastly overfit the data and lose generality. To these ends, we carried out a somewhat elaborate cross and external validation scheme to assess the predictive performance of a battery of models. In this section we discuss this validation study and we succeed in settling on a model which provides an excellent compromise between fit and predictive power.

Our validation study consisted of comparing a sequence of increasingly complex models in seven validation contexts:

1. four sets of cross validation within the NAOS data;
2. an external validation fitting the models with the entire NAOS data set and predicting the USGS data; and
3. two sets of cross validations fitting the models with some of NAOS and some of USGS and then predicting the rest.

The validations involving the USGS data were somewhat limited by the facts that the USGS data contain only ground water arsenic measurements, and that the number of people served by particular wells are not available. The former shortcoming simply limits the ability to validate predictions with the USGS data to ground water systems only. The latter limitation was addressed by treating the missing populations as parameters, sampling from their conditional distributions given other model quantities (Appendix C). For the NAOS cross validations, we split the NAOS data into $NAOS_1$ and $NAOS_2$ by a stratified random sample in which eight strata were used. The strata were formed by considering three binary categories: source type (surface/ground), location (east/west) and system size (large/small). The determination of location was based on the Mississippi River, where Minnesota and Louisiana are classified as west. The size determination of a system was based on whether the population served was greater or less than 25000 people, corresponding to the middle of twelve established EPA size classes (Raucher et al. 1995). We then split each of $NAOS_1$ and $NAOS_2$ roughly in half along the same lines and recombined half of each into $NAOS_3$ and the other half of each into $NAOS_4$. We also split the USGS data into $USGS_1$ and $USGS_2$ roughly along geographical lines. The validations consisted of fitting the sequence of models to one set of data ("training data") and then predicting another ("testing data"). We thought that it was not sensible to use USGS as the training data since there were no surface water systems in USGS. The seven pairs of training and testing data are summarized in Table 1.

Validation	Training Data	Test Data
V_1	$NAOS_1$	$NAOS_2$
V_2	$NAOS_2$	$NAOS_1$
V_3	$NAOS_3$	$NAOS_4$
V_4	$NAOS_4$	$NAOS_3$
V_5	$NAOS$	$USGS$
V_6	$NAOS_3 + USGS_1$	$NAOS_4 + USGS_2$
V_7	$NAOS_4 + USGS_2$	$NAOS_3 + USGS_1$

Table 1. Outline of the validation training/testing combinations.

There were nineteen different models included in the validation study, including the model described in Section 3.2. We classify these models into three different groups in order to organize the discussion. The first group consists of models substantially simpler in structure than that in Section 3.2. These models are included primarily as so-called "strawman" models because the expectation (which is indeed borne out) is that they are not sufficiently rich to capture the complexity of the data, making them relatively poor predictors of external data. The next class consists of four hierarchical models which include the system size covariate in a discrete form, and the final class consists of nine hierarchical models which include the system size covariate in a quantitative, linear form. Both of these classes include the model in Section 3.2 as special cases. In the following sections we present the details of the various models in these three classes. Each section begins with an explanation of the class of models, followed by an itemized list of the specific details of each model in the class, including the nominal number of parameters that directly impact the likelihood function. The final section relays the results and conclusions of the validation efforts.

3.5.1 Simple Models

As mentioned previously, a primary concern with the use of complex hierarchical models is they may too closely fit the data, thereby reducing their predictive power. It is thus necessary to consider the predictive performance of simpler models in order to justify the use of more complicated models. The latter should only be used if simpler models are not able to adequately capture subtle features of the data necessary for more accurate predictions of new data. The models considered in this section have many fewer parameters than the models in the other classes. This is primarily due to the fact that they use geographic distinctions far more coarse that the U.S. states considered in the other models. For example, some of the models divide the country into an "east" and "west" region; this delineation was referenced briefly in the discussion of the stratification used to create the validation data sets. The

Mississippi River was used as the dividing line between the two regions, where Minnesota and Louisiana are classified as west. The Mississippi River is not only a natural dividing line for the U.S., but has been found to roughly divide the country into regions of different arsenic profiles (Pontius et al. 1994).

The other feature of these models warranting further comment is the way in which system size was included. As mentioned previously, it is believed that, although the population served by a system may not be directly relevant to predicting arsenic concentrations, it may be a proxy for other factors which are. This being the case, we believed that including population as a quantitative variable could overstate its usefulness as a covariate despite the fact that the NAOS data provide actual numbers of people served by each system. Moreover, exploratory data analysis indicated that there appeared to be slight disparity in the arsenic profiles between systems which served a large number of people and those which did not (Appendix F), but no quantitative relationship was immediately evident. These factors led us to consider a class of models in which the population served by a system only indicates whether that system was classified as "large" or "small". Models which include system size in a quantitative manner are discussed in Section 3.5.3. Also, the distinction between "large" and "small" in both models in this class that include system size depends on source water type. The actual impact that the size class distinction has on the distribution of the data depends not only on source water type, but location. This is primarily due to the fact that the observed dependence on system size was almost exclusively isolated to large ground water systems in the western U.S.

Model Definitions:

- (Model 1) Common lognormal distribution for all arsenic observations (2 parameters)

- (Model 2) Separate lognormal distributions for each source type (4 parameters)
- (Model 3) Separate lognormal distributions for each source type and east/west combination (8 parameters).
- (Model 4) Separate lognormal distributions for each source type and NAOS region combination. Although we would not choose this as our final model because we do not want to use the NAOS regions as a location basis (see Section 3.3), we wanted to fit it using maximum likelihood to see if it was possible to achieve a good fit using some level of geographical information short of the states themselves. (28 parameters)
- (Model 5) Separate lognormal distributions for each source type, east/west designation, and size classification, where each of the 8 categories has a separate mean, but the variances are common for a particular source and location combination across both sizes. Here size refers to large or small, and separate parameters for ground and surface were fit to distinguish these categories. (14 parameters)
- (Model 6) Separate lognormal distributions for each source type, east/west and size category. This is just like Model 5, but no restrictions on the data variances are made. Also like Model 5, two size cutoff parameters were fit (18 parameters).

3.5.2 Hierarchical Models I: Population Thresholds

In this section, we consider two models that, like Models 5 and 6, include system size in the model only to classify each system as small or large. However, we build upon the rich structure of the model in Section 3.2 so that the effect that this distinction has on the distribution of the data can depend on the state of each system (as opposed to the east/west regional classification) in addition to the source water type. To describe the model, let Z_i

denote the logarithm of population served by the system i. Define $U_i(\delta_{W_i}) = -1$ if $Z_i < \delta_{W_i}$ and $U_i(\delta_{W_i}) = 1$ otherwise. That is, the values δ_{-1} and δ_1 are source-type specific cutoffs on the scale of log-population which determine whether a particular system is "large" or "small;" these are the same parameters that were used in Models 5 and 6.

We then generalize Stage 1 of the previous model (Equations 1 and 2) as for $i = 1, \ldots, n$,

$$\mu_i = \alpha^{(m)}_{X(i)} + \beta^{(m)}_{X(i)} W_i + \gamma^{(m)}_{X(i)} U_i(\delta_{W_i}) + \eta^{(m)}_{X(i)} W_i U_i(\delta_{W_i}) \tag{3}$$

$$-2\log\sigma_i = \alpha^{(v)}_{X(i)} + \beta^{(v)}_{X(i)} W_i + \gamma^{(v)}_{X(i)} U_i(\delta_{W_i}) + \eta^{(v)}_{X(i)} W_i U_i(\delta_{W_i}) \tag{4}$$

That is, we replace the one-way ANOVA structure of the parameters used in the previous model with a two-way ANOVA structure, where the additional factor is the size classification of the system. The main effect of size is essentially characterized by the γ parameters, while the η parameters quantify the interactions of size with source water type. The implications of this structure become more clear with an example. Suppose that a ground water system in state j serves a (log) population which is less than δ_1. That system is then classified as being "small", which adjusts the mean of the distribution for log arsenic by $(-\gamma^{(m)}_j - \eta^{(m)}_j)$ and the log precision by $(-\gamma^{(v)}_j - \eta^{(v)}_j)$. Other contrasts of these parameters are used to adjust the distributions for other combinations of the source water type and the size classification.

Stages 2 and 3 of the original model are suitably generalized to accommodate the additional parameters. As before, we concatenate analogous parameters for different states into vectors of length 50, labeling them $\boldsymbol{\gamma}^{(m)}$,$\boldsymbol{\gamma}^{(v)}$,$\boldsymbol{\eta}^{(m)}$ and $\boldsymbol{\eta}^{(v)}$. These vectors have prior distributions just like the $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ source type adjustment vectors, with fixed prior mean vectors $\boldsymbol{\gamma}^{(m)}_0$,$\boldsymbol{\gamma}^{(v)}_0$,$\boldsymbol{\eta}^{(m)}_0$ and $\boldsymbol{\eta}^{(v)}_0$, unknown prior precision parameters $\tau^{(m)}_\gamma$, $\tau^{(v)}_\gamma$,$\tau^{(m)}_\eta$ and $\tau^{(v)}_\eta$, and the common distance-based correlation matrix $C(\rho)$. The value of ρ is shared with the $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ source type adjustment vectors. Finally, the prior distribution for the log prior precision terms is iid $N(\mu_\tau, \xi_\tau)$, and the two size class cutoff parameters δ_{-1}, δ_{+1} are modeled as iid $N(\mu_\delta, \xi_\delta)$ for fixed hyperparameters μ_δ and ξ_δ.

Model Definitions:

- (Model 7) Force the 4 vectors $\boldsymbol{\gamma}^{(m)}$,$\boldsymbol{\eta}^{(m)}$,$\boldsymbol{\gamma}^{(v)}$ and $\boldsymbol{\eta}^{(v)}$ to be vectors of zeroes, so that there is no dependence on population. This is precisely the model described in detail in Section 3.2, and from this point forth, we refer to the model described therein as "Model 7." (200 parameters)

- (Model 8) Force the 2 vectors $\boldsymbol{\gamma}^{(m)}$ and $\boldsymbol{\eta}^{(m)}$ to be vectors of zeroes, so that the dependence on system size is only reflected in the data precisions. (302 parameters)

- (Model 9) Force the 2 vectors $\boldsymbol{\gamma}^{(v)}$ and $\boldsymbol{\eta}^{(v)}$ to be vectors of zeroes, so that the dependence on system size is only reflected in the data means. (302 parameters)

- (Model 10) Place no restrictions on the on the four vectors $\boldsymbol{\gamma}^{(m)}$,$\boldsymbol{\eta}^{(m)}$,$\boldsymbol{\gamma}^{(v)}$ and $\boldsymbol{\eta}^{(v)}$, so that the dependence on system size is manifested in both the means and precisions of the data. (402 parameters)

3.5.3 Hierarchical Models II: Population Linear

In addition to the threshold models, we also fit a battery of models which included system size in a quantitative, linear fashion. Our motivation was twofold. First, the threshold models performed relatively poorly at predicting out of sample, indicating that they were too richly parameterized. Moreover, these models presented a number of implementational difficulties stemming from the fact that the likelihood function is not a continuous function of the population threshold parameters δ_1 and δ_{-1}. Both of these will be expounded upon in Section 3.5.4. The net result is that we found it desirable to try a more traditional regression-like approach to the inclusion of the system size covariate, and we wanted to reduce the number of parameters used to quantify this dependence. In order to describe the models, we define E_i as an indicator of whether system i is in the eastern or western region of the country. The basis for this family of models is again Model 7; we generalize Stage 1 (Equations 1 and 2), for $i = 1, \ldots, n$, as

$$\mu_i = \alpha^{(m)}_{X(i)} + \beta^{(m)}_{X(i)} W_i + \lambda^{(m)}_{W_i,E_i} Z_i \quad (5)$$

$$-2 \log \sigma_i = \alpha^{(v)}_{X(i)} + \beta^{(v)}_{X(i)} W_i + \lambda^{(v)}_{W_i,E_i} Z_i \quad (6)$$

That is, the means and log precisions of the data distributions now depend in a linear fashion on the log of population served by each system via the coefficients $\lambda^{(m)}_{W_i,E_i}$ and $\lambda^{(v)}_{W_i,E_i}$. The indices indicate that these coefficients are allowed to depend on whether the system is surface water or ground water, and whether the system is eastern or western, comprising a maximum of eight additional parameters. The various models in this class are distinguished by which of these eight parameters are introduced into the model. In particular, for each of the means and variances, we allowed three levels of complexity for the parameters λ_{W_i,E_i}. The first sets all of them to zero, thereby removing the associated dependence on population. The next removes the dependence on location, forcing $\lambda_{W_i,0} = \lambda_{W_i,1}$ for $W_i = -1, 1$. The last allows each of the four parameters to be non-degenerate. We did not consider any models in which the population effect did not depend on source water type because such models did not make sense based on examination of the data. Because each of the three levels of complexity could be allocated independently to each of $\lambda^{(m)}_{W_i,E_i}$ and $\lambda^{(v)}_{W_i,E_i}$, the net result is nine possible models. Of course, the one in which all the coefficients are zero for both the mean and log precisions is equivalent to Model 7. The remaining eight models are as follows:

Model Definitions:

- (Model 11) $\lambda^{(m)}_{W_i,E_i} \equiv 0$ $\lambda^{(v)}_{W_i,0} = \lambda^{(v)}_{W_i,1}$ (202 parameters)
- (Model 12) $\lambda^{(m)}_{W_i,E_i} \equiv 0$ $\lambda^{(v)}_{W_i,E_i}$ has no restrictions (204 parameters)
- (Model 13) $\lambda^{(m)}_{W_i,0} = \lambda^{(m)}_{W_i,1}$ $\lambda^{(v)}_{W_i,E_i} \equiv 0$ (202 parameters)
- (Model 14) $\lambda^{(m)}_{W_i,0} = \lambda^{(m)}_{W_i,1}$ $\lambda^{(v)}_{W_i,0} = \lambda^{(v)}_{W_i,1}$ (204 parameters)
- (Model 15) $\lambda^{(m)}_{W_i,0} = \lambda^{(m)}_{W_i,1}$ $\lambda^{(v)}_{W_i,E_i}$ has no restrictions (206 parameters)
- (Model 16) $\lambda^{(m)}_{W_i,E_i}$ has no restrictions $\lambda^{(v)}_{W_i,E_i} \equiv 0$ (204 parameters)
- (Model 17) $\lambda^{(m)}_{W_i,E_i}$ has no restrictions $\lambda^{(v)}_{W_i,0} = \lambda^{(v)}_{W_i,1}$ (206 parameters)
- (Model 18) $\lambda^{(m)}_{W_i,E_i}$ has no restrictions $\lambda^{(v)}_{W_i,E_i}$ has no restrictions (208 parameters)

3.5.4 Validation Results

In this section we describe how we carried out the validation study of the previously discussed models, and we present the results. The simple Models 1 through 6 were fit by maximum likelihood, while the remaining models were assessed via large samples obtained from the posterior distribution of their respective parameters using Markov Chain Monte Carlo methods. A Bayesian version of Model 3 (labeled "$3B$") was also fit using MCMC. For each of these models, we used a burn-in period of 100,000 and a sampling period of 1,000,000. The only exception is Model 7, which was run for these durations twice because it was the degenerate case in both classes of the population models. A complete description of the prior distributions used for each of the models, as well as a sensitivity analysis to these prior distributions, can be found in Appendix D. In summary, the results were reasonably robust to even large changes in the specification of the prior hyperparameters. Finally, to the extent possible, we ensured that the chains were converging properly by confirming the consistency of the results after starting from several configurations of widely dispersed starting values.

Special considerations regarding population values were necessary in all of the models, stemming from both missing populations (which occurred for 16 of the 457 NAOS observations and all of the USGS observations) and the form in which population was used in the threshold models. For the simple models fit by maximum likelihood, missing populations were explicitly integrated over when calculating the likelihood functions. Moreover, we restricted the set of values for the cutoffs δ_{-1} and δ_1 in Models 5 and 6 to lie in the interval between 8.11 and 11.95, which are the 0.15 and 0.85 quantiles of the empirical distribution of log population in the NAOS data. The reason for this was that, if the population cutoff got too extreme in Model 6, we could end up with only one uncensored observation in a source type, size class, and location cell, and this would allow the likelihood to be unbounded as the variance for that cell approached 0.

For the Bayesian models, missing populations were treated as any other parameter. Within each state and source type, the SDWIS data set was used to determine a distribution for populations served by all public water systems in that state and source water type classification. These served as prior distributions for all missing population figures. At each iteration of the Markov Chain, missing populations were sampled from their conditional distributions given all other model quantities. For the threshold models, this distribution was available in closed form; for the quantitative models, the Metropolis algorithm was necessary. More specific details about how the population full conditional distributions were obtained and sampled, as well as the estimated prior distributions for each state and source water type, are given in Appendix C. Finally, we restricted the range of the population cutoff parameters in exactly the same way as we did in the simple models, for the same reason.

The validation criterion used in all seven cases $V_1, \ldots, V_7$ was the predictive density of the associated test data. The predictive density of testing data $\boldsymbol{y_1}$ (with covariates $\boldsymbol{x_1}$) given training data $\boldsymbol{y_0}$ (with covariates $\boldsymbol{x_0}$) is $\int f(\boldsymbol{y_1}|\boldsymbol{x_1}, \theta)p(\theta|\boldsymbol{y_0}, \boldsymbol{x_0})d\theta$. The ratio of two such predictive densities, first considered by Lempers (1971), is known as a "partial" Bayes factor because of its relationship to the traditional Bayes factor (O'Hagan 1995). If the predictive density for $\boldsymbol{y}_1$ under model A is higher than that for model B, $\boldsymbol{y}_1$ is more likely under model A than model B, helping to make the case that model A is a better descriptor of the phenomena generating the data. Other reasons for this choice of criterion and more detail on its theoretical and implementational properties are given in Appendix E. For the models fit by maximum likelihood, the predictive density was taken as the density of the testing data assuming that the MLE is the parameter. For the Bayesian models, the predictive density was obtained by integrating the likelihood function of the test data with respect to the posterior distribution of the model parameters. These integrals were estimated by Monte Carlo integration using a sample $\theta_1, \ldots, \theta_m$ from the relevant posterior distribution of the parameters given the associated training data.

A summary of our calculations is given in Table 2, which gives the estimated log predictive density of the testing data for each of the models in each of the seven validation settings. We first consider the "Simple" Models 1 through 6. The first predominant feature to note is the large predictive densities of Models 3 and 3B compared to Model 1 in all of the validations. This is substantial evidence for predictable covariation of arsenic concentrations

Model	V_1	V_2	V_3	V_4	V_5	V_6	V_7
1	−351.3	−361.1	−358.9	−353.9	−2664.3	−1672.4	−1680.3
2	−346.5	−353.3	−354.3	−347.2	−2729.5	−1673.6	−1677.7
3	−343.8	−328.8	−345.8	−325.0	−2620.0	−1597.6	−1584.4
3B	−339.9	−327.4	−342.1	−324.3	−2544.1	−1596.7	−1585.5
4	−493.5	−332.5	−382.4	−317.1	−2601.8	−1541.2	−1523.7
5	−347.4	−334.7	−368.4	−331.7	−2618.8	−1621.6	−1588.8
6	−422.0	−335.0	−455.5	NA	−2725.8	−3023.4	−1568.3
7	−314.4	−301.0	−316.9	−297.3	−2378.4	−1393.4	−1429.3
8	−315.5	−304.9	−319.7	−300.7	−2418.2	NA	NA
9	−321.3	−310.6	−317.1	−299.7	−2453.7	−1418.3	−1434.0
10	−319.9	−310.7	−315.5	−297.3	−2501.7	NA	NA
11	−317.2	−301.5	−316.3	−298.0	−2402.7	−1397.4	−1428.2
12	−315.7	−301.7	−315.7	−297.0	−2415.7	−1399.7	−1428.6
13	−315.8	−302.3	−319.5	−297.5	−2384.8	−1393.1	−1430.4
14	−313.7	−300.8	−318.8	−296.6	−2403.8	−1396.7	−1427.9
15	−315.0	−302.5	−316.5	−298.5	−2401.1	−1397.7	−1430.8
16	−315.9	−302.3	−317.0	−296.1	−2402.0	−1398.5	−1428.6
17	−314.5	−303.1	−315.4	−298.2	−2412.2	−1397.0	−1427.3
18	−316.5	−301.6	−313.2	−298.9	−2403.1	−1400.9	−1427.6

Table 2. Results from the validations. The tabled values are the logarithms of the predictive density of the appropriate test data ("LPD"). The row for Model 7 corresponds to the combined inferences from the two different realizations of the Model 7 predictions.

with source water type and location. Model 4, which subdivides the east/west designation into the seven NAOS regions, is also noteworthy. For validations V1 through V3 and V5, this model predicts more poorly than Model 3B, which seems to indicate that the data are being overfit by the more finely divided regions. However, its performance in V4, V6 and V7 is superior to Model 3B, the latter two being most telling. Because the validations sets V6 and V7 were made by combining half of each of the 457 NAOS observations and the 1945 USGS observations, the training data for these validations have approximately 1200 observations, far more than the other five validations. This suggests that we ought to produce substantially more precise estimates of the parameters for ground water systems based on the training data for these validations than we do for the other validations. With these less variable estimates, we are able to achieve a better characterization of the testing data with the richer spatial structure of Model 4. This indicates that the poorer performance of Model 4 compared to Model 3B in four of the other five validations is more a function of highly-variable estimates than it is of a fundamental overparameterization of the model. That is, at least among ground water systems, the dependence of arsenic profiles on location is apparently more structured than can be characterized by the dichotomous east/west classification.

Models 5 and 6, which include the size class distinction, perform considerably more poorly than the best previous models in all of the validations. Of course, because the size class distinction subdivides the data even further, and because the USGS data do not have population figures, this could be purely a function of overly variable estimates rather than an indication of the irrelevance of system size. The trade-off of partitioning the data into more specifically defined classes versus the variability of the resulting estimates is epitomized in the performance of Model 6 in validation V4, where the maximum likelihood estimate does not exist due to sparseness.

We now turn to a comparison of Models 7 through 10, which include the three Bayesian population threshold models as well as the fundamental model presented in Section 3.2. The first obvious feature is that the estimated predictive densities are substantially larger than any of the simple models. The only exceptions were Models 8 and 10 in validations V6 and V7, where no predictive density could be calculated because the model fits were hampered by convergence problems. We believe this was due to not sufficiently restricting the range of the population cutoffs, giving rise to an unbounded likelihood function. We did not perform a more rigorous examination of the problem because it was clear from the rest of the analysis that these models were dominated by other models and hence warranted no further attention.

Despite the fact that the point estimates of the predictive densities from Models 7 through 10 are considerably larger than that for the best simple model in each validation, it is impossible to determine whether these differences are significant without some examination of the variability of the Monte Carlo estimates. This is because the predictive densities from the simple models are known exactly, as they are calculated by evaluating the likelihood of the testing data at a single point. The only exception is Model 3B, in which the predictive density was estimated by Monte Carlo integration with respect to the posterior distribution. However, the Monte Carlo variability of the estimates for Model 3B was essentially negligible because the posterior distribution was very concentrated. For all intents and purposes, the predictive densities of all of the simple models can be treated as known quantities.

Quite the opposite is true for the more richly parameterized Models 7 through 10. The predictive density estimates for these models are not nearly as well behaved because the posterior distributions are orders of magnitude more sparse than those for Model 3B. Hence, the predictive density point estimates based on even long MCMC runs can be quite variable because the regions of the parameter space where the likelihood of the external data is high can have very low probability under the posterior distribution given the training data. Our extensive empirical experience indicated that the point estimates are often driven by one or two parameter vectors in MCMC samples of size 100,000 or more. Because the estimates are driven by such a small number of sampled parameter vectors, assessing the standard error of the predictive density estimator is rather difficult. Appendix E presents the results of both a traditional and a Bayesian bootstrap analysis to assess whether the estimated predictive superiority of Model 7, as compared to the best simple model in each of the validations, is great enough to conclude that the actual unknown predictive densities assigned by this model exceed those of the other models. The results of the analysis indicated that in each of the seven validations, the probability is at least 0.99995 that the predictive density for Model 7 is at least tens of thousands of times as large as the largest predictive density observed for any of the simple models. Hence, we have strong evidence that for any one of the validations, Model 7 offers greater predictive power for the external data than even the best simple model. The fact that this holds for all seven validation contexts solidifies the conclusion that the richer structure of Model 7 makes it more effective than any of the simple models considered. The bootstrap results in Appendix E are based on a preliminary analysis that used a slightly different prior distribution than that used to calculate the numbers presented in Table 2, as discussed in Appendix D. However, the log predictive density estimates were quite similar for the two prior specifications, so that we can be confident that the predictive superiority of Model 7 holds in the current case as well.

That leaves the question of how Models 7 through 10 perform relative to one another. The point estimates somewhat suggest that the population structure in Models 8 through 10 lead to slightly noisier, less effective predictions than Model 7. In light of the previous discussion about Monte Carlo variability, this conclusion is at best tenuous without a more rigorous examination of the variability of these estimates. Figure 1 attempts to characterize this variability in a way that allows a more definitive assessment of their relative performances. For each model in each validation, we divided the 1,000,000 posterior samples into 100 contiguous blocks of 10,000 parameter vectors. The only exception to this was Model 7, for which we had 2,000,000 posterior samples to divide into 200 blocks of 10,000. We then estimated the log predictive density of the testing data, via Monte Carlo integration, separately for each of the posterior sample blocks. This presents us with an indication of

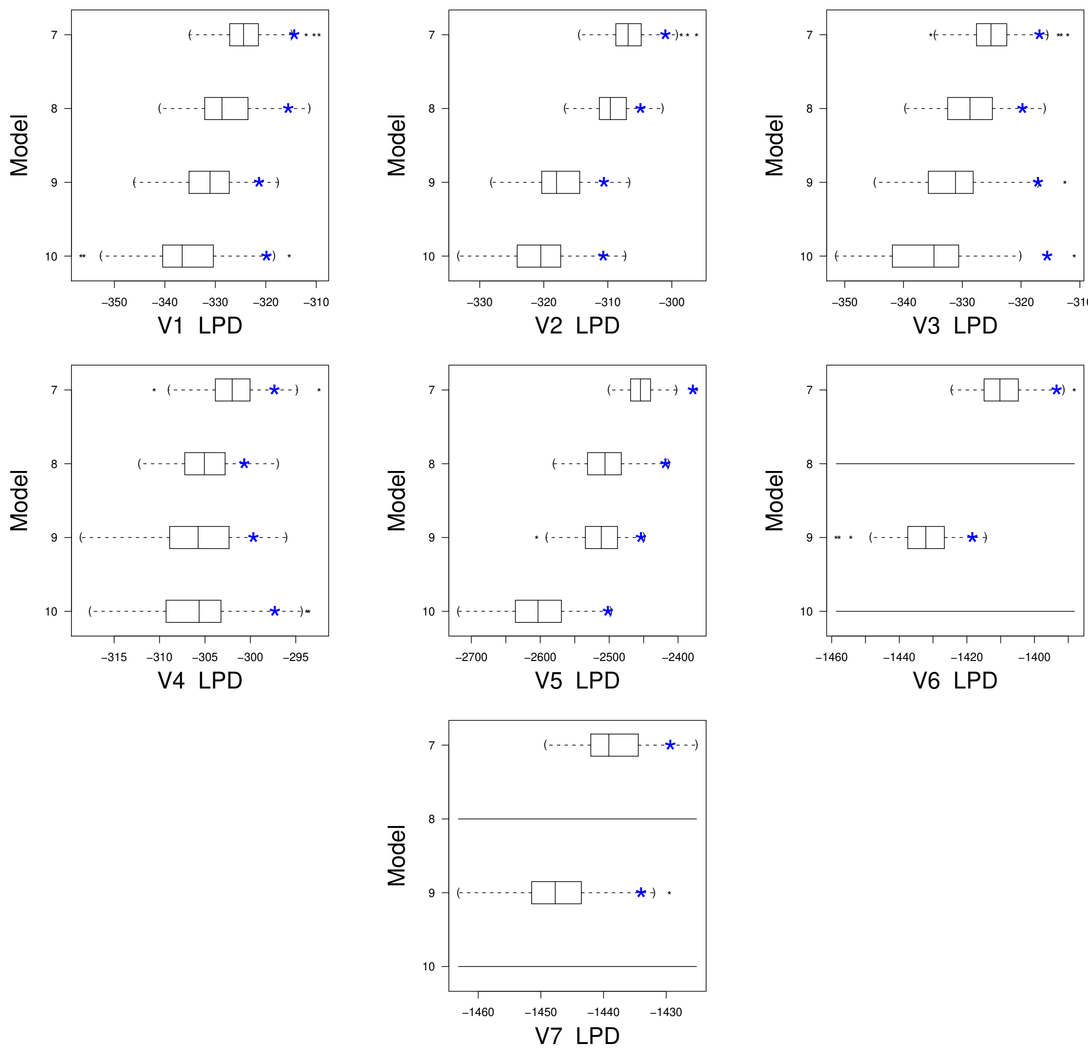


Figure 1. Comparison of log predictive density estimates for Models 7 through 10 based on blocks of 10,000 parameter vectors. Each boxplot represents the distribution of the block estimates of log predictive density of the testing data for a particular model and validation. Aggregate estimates based on all posterior samples are indicated by the asterisks. Boxplots for Model 7 based on 200 blocks of 10,000 while those for the other models are based on 100 blocks of 10,000. Predictive densities for Models 8 and 10 in validations V6 and V7 were not available.

Monte Carlo variability of the predictive density estimates based on posterior sample sizes of 10,000, allowing us to assess (at least subjectively) our belief in the accuracy of the estimate based on the full sample of one million parameter vectors. The figure contains one frame for each validation, and in each frame is a boxplot of the log predictive density block estimates for each of the four models. The large asterisk indicates the aggregate estimate based on the entire sample. These plots clearly indicate that Model 7 consistently results in higher predictive density estimates, even based on only 10,000 parameter vectors. Even in the cases where the aggregate estimate for Model 7 is not the best among the four models (e.g. V3 and V4), the sampling distributions of the Model 7 block estimates are shifted toward higher values. The conclusion is that Model 7, which does not include the system size covariate, appears to offer more predictive power than either of the population threshold models. This does not definitively conclude that system size is not relevant to predicting arsenic concentrations, but it does indicate that the threshold structure (at least in its current formulation) is not the proper quantification of the relationship.

The final class of models, based on the inclusion of system size in a linear manner, seeks a better quantification. However, an examination of Table 2 for Models 11 through 18, as well as the other degenerate form of Model 7, provides even more evidence that system size is not particularly effective at enhancing predictive power. In each of the validation contexts, the log predictive density estimates are remarkably similar across the different models. [1]. The fundamental similarity of the predictive performances of the different models is highlighted in Figure 2, where we plot log predictive density estimates based on contiguous blocks of 10,000 parameters, analogous to what is depicted in Figure 1. Unlike what was observed in

[1]The only possible exception to this is in V5, where some additional variability is evident. However, this variability is almost certainly due to the intra-model variability of the estimate rather than inter-model variability in predictive power. In all of our work, predictive densities for V5 were subject to a high degree of Monte Carlo variance. This issue is discussed more fully in Appendix E.

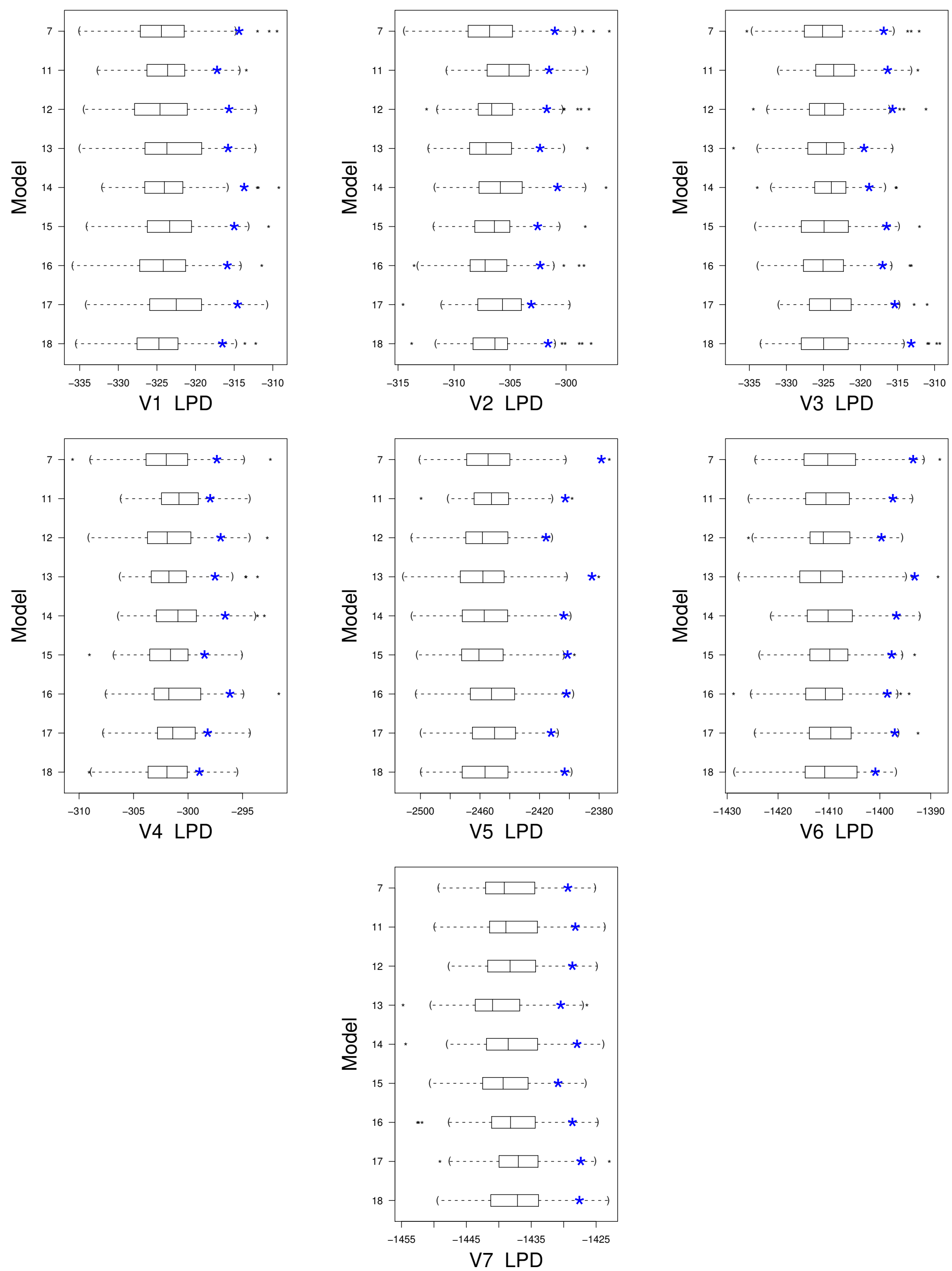


Figure 2. Comparison of log predictive density estimates for Models 7 and 11 through 18 based on blocks of 10,000 parameter vectors. Each boxplot represents the distribution of the block estimates of log predictive density of the testing data for a particular model and validation. Aggregate estimates based on all posterior samples are indicated by the asterisks. Boxplots for Model 7 based on 200 blocks of 10,000 while those for the other models are based on 100 blocks of 10,000.

Figure 1, the block estimate distributions in Figure 2 are nearly identical across different models within each validation context. Essentially no clear predictive differences among the various linear population models, including the degenerate one, is evident. In defense of the relevance of system size, there is slight evidence that the distributions for Models 11 and 14 are shifted toward somewhat larger values than those for Model 7 among validations V1 through V4. It is only in these validations where the numbers of people served was known for nearly all the systems. It is thus conceivable that, if the system size were available for the systems in the USGS data, the slight effect might become more evident. The results thus indicate that there may be some (probably small) systematic covariation of arsenic concentrations with system size, but that this effect is essentially impossible to quantify given the currently available data. Some additional discussion of this matter is given in Appendix F.

The fundamental conclusion of the validation study is that Model 7 is clearly more efficacious predictor of arsenic trends than any of the simpler models, as well as the more complicated population threshold models. Moreover, it predicts as well as any of the linear population models, whereby the principle of parsimony suggests that it dominates the more complicated models. It is these facts, along with its good fit (Section 4.1), that led us to conclude that Model 7 was the "best" among those we considered. We stress again that we would be remiss in claiming that the results of the validation study vanquish the relevance of system size in quantifying patterns in arsenic occurrence. Without more convincing evidence of its usefulness, however, we decided not to let it hamper our progress in making national predictions of arsenic levels.

4. RESULTS

4.1 Model Fit

The predictive power of Model 7 is complemented by the fact that it fits the data quite well. Posterior predictive checks (Gelman, Meng, and Stern 1996; Gilks et al. 1996; Datta, Lahiri, Maiti, and Lu 1999) were used to ensure that Model 7 is able to capture effectively the trends in arsenic occurrence observed in the data. Figure 3 is an example

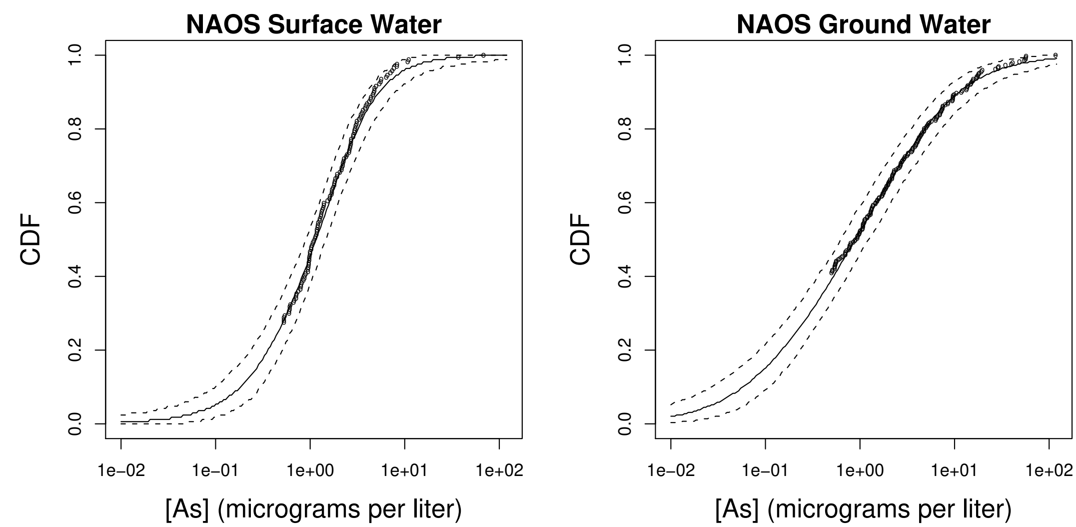


Figure 3. Empirical and posterior predictive cumulative distribution functions of NAOS arsenic concentration for surface water (left) and ground water (right) based on fit of Model 7 to the NAOS data. Dotted lines represent pointwise 0.025 and 0.975 quantiles of the predictive CDFs. Plotted points are observed data above the censoring point of 0.5 $\mu g/L$.

of the kind of posterior predictive check that was used throughout the modeling process. The horizontal axes of Figure 3 represent logarithmically spaced concentrations. The dots represent the empirical CDFs of the relevant data, in this case, separate empirical CDFs for the NAOS ground and surface water arsenic observations. The censored observations are not plotted; however they contribute to the empirical CDF since they all correspond to observations less than the censoring point of 0.5 $\mu g/L$. That is why the empirical CDFs conspicuously truncate at the censoring point. The curves are intended as predictions of the empirical distributions of future samples of systems with the characteristics that correspond to those systems whose empirical CDF is shown in the plot. Imagine that a future arsenic concentration is measured at each of these systems. What does the model predict the empirical CDF of this future sample to be? This is the question that the plot answers, and it is calculated as follows. For each parameter in a posterior sample from the MCMC analysis based on the fit of Model 7 to the NAOS data, we simulate one observation at every system in the relevant set of systems. We then compute the empirical CDF of these simulated values at each point $a_1, \ldots, a_s$ along a grid on the horizontal axis. Repeating this procedure for each parameter vector results in a sample of empirical CDF values at each a_i for $i = 1, \ldots, s$. The curves in each plot are pointwise (at each a_i) 0.025, 0.5 and 0.975 quantiles of these samples. While the pointwise approach is not equivalent, or even as informative, as a more complicated function-wise prediction interval, it nevertheless provides a simple yet useful tool for the assessment of the variability of model predictions.

The figure suggests that arsenic in ground water is more variable than in surface water, a trend verified by comparing the distributions from the two source water types within each state. From the figure it is evident that the model rather successfully captures the differences between the NAOS surface water and ground water observations. In order to assess more fully the degree to which the model explains the patterns observed in the data, plots such as these were made for all nonempty cells in the (50×2) design of state and source type as well as aggregate classes in addition to those depicted here. Virtually all of the plots reflect the kind of agreement between model predictions and observed data displayed in Figure 3, indicating that the model fits the data well.

4.2 Geographic Trends

We highlight the spatial patterns in arsenic occurrence in Figure 4. For each state and

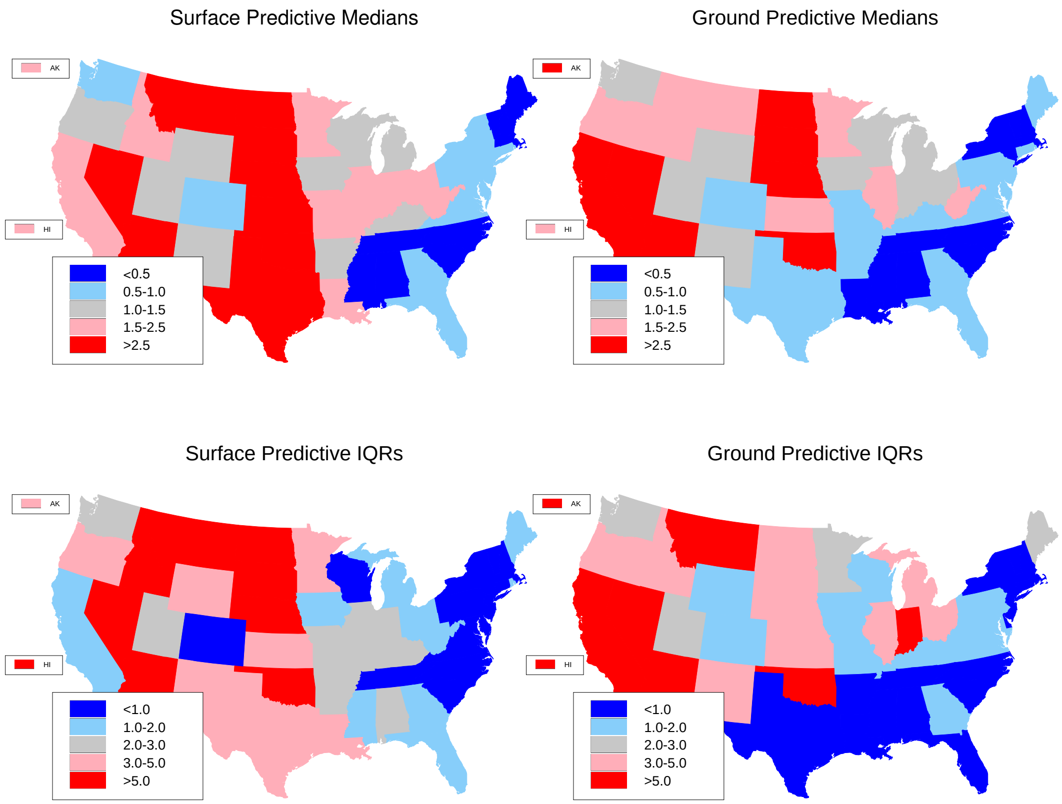


Figure 4. Median (top row) and IQR (bottom row) of the predictive distribution for a single arsenic concentration in each state and source water type, based on the fit of Model 7 to the combined NAOS and USGS data.

source water type, we plot the median (top row) and IQR (bottom row) of the predictive distribution of a single arsenic concentration for systems in that state using that source water type. The predictive distributions are based on the fit of the model to the combined NAOS and USGS data. We see a marked difference in the predicted median level of arsenic from eastern to western states; in particular, the west generally appears to have substantially higher values. These predictions are reflective of the corresponding patterns observed in the data (for example, the USGS measurements depicted in Figure 5). Both the Missis-

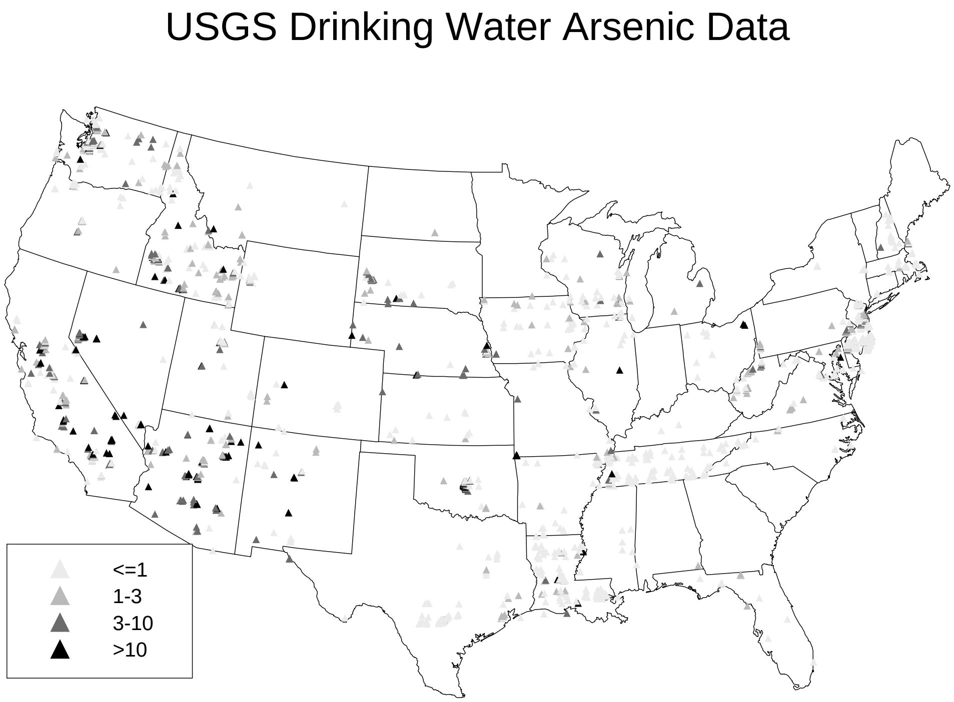


Figure 5. USGS community drinking water system arsenic measurements ($\mu g/L$).

sippi and Missouri Rivers approximately divide the country into regions of different arsenic profiles, which is consistent with previous findings (Pontius et al. 1994). The figures also indicate a high degree of correlation between ground and surface water concentrations, though differences are apparent in selected cases (e.g., Washington and Oregon, where relatively high concentrations of arsenic are found in ground water, along with low concentrations in surface waters). Colorado is unusual in the west with low arsenic levels for both ground and surface waters. The great disparity between ground water and surface water in Texas is possibly due to the fact that the surface water is heavily influenced by influx from the west, while the ground water is more representative of the low arsenic levels of the eastern coastal region. Finally, the spatial trends in the inter-quartile ranges of the predictive distributions (bottom row of Figure 4) roughly match those for the predictive medians. This makes sense given that each predictive distribution for arsenic is a mixture of lognormal distributions, for which the variability increases as the central location increases.

4.3 Effect of Updating with USGS

It is instructive to consider how updating with the USGS data changes the posterior distribution of the model parameters. Figure 6 shows the posterior predictive CDFs of

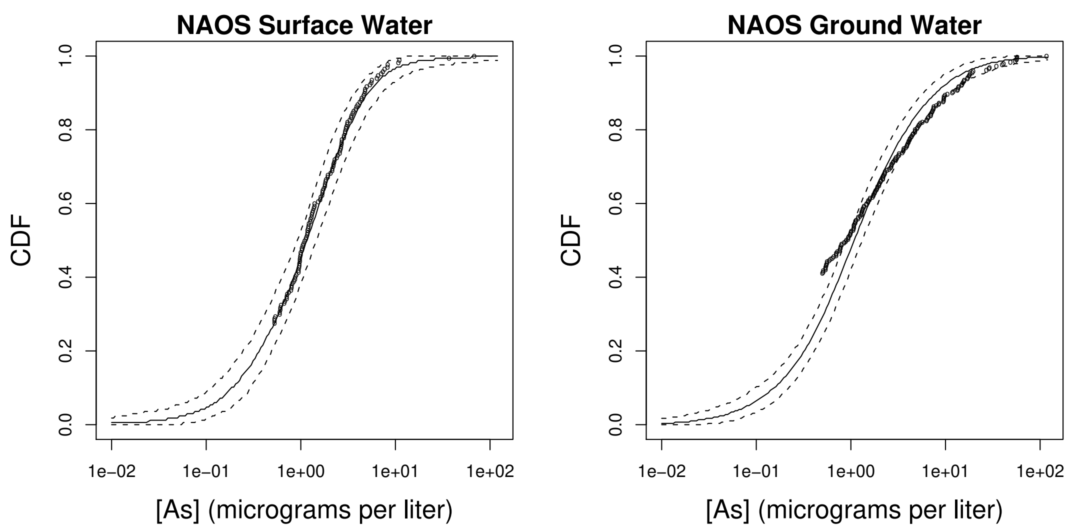


Figure 6. Empirical and posterior predictive cumulative distribution functions of NAOS arsenic concentration for surface water (left) and ground water (right) based on fit of Model 7 to the combined NAOS and USGS data. Dotted lines represent pointwise 0.025 and 0.975 quantiles of the predictive CDFs. Plotted points are observed data above the censoring point of 0.5 $\mu g/L$.

arsenic concentrations for the NAOS systems based on the fit of the model to the combined NAOS and USGS data. These figures are analogous to those depicted in Figure 3, but are based on parameter vectors from the posterior distribution conditional on both the NAOS and USGS data rather than the NAOS data alone. As expected, because the USGS data contain no observations from surface water sources, the impact of the USGS on predictions for surface water sites is negligible. This is clear from the fact that the left frames of Figures 3 and 6 are virtually identical. However, ground water predictions are impacted by using the USGS data to update the posterior distribution of the parameters. In particular, the predictive CDF reflects a change in the locations of the distributions of some of the parameters induced by the USGS data. This is evident from the fact that the predictive CDF based on all of the data does not match the empirical CDF of the NAOS ground water observations as closely as that based only on the NAOS data; the model estimated with both data sets suggests a ground water arsenic distribution with somewhat lower variance than that based on the NAOS data alone. Moreover, the consideration of the USGS data also reduces parameter uncertainty; this can be seen in the tighter 95% credible bounds around the predictive CDF computed from the posterior conditioned on all of the data. These impacts of the USGS data are revisited in Section 5.

4.4 National Predictions

The most fundamental result from our analysis is the prediction of the outcome of a national assessment of raw water arsenic concentrations and the uncertainty in this prediction. As noted previously, we use the SDWIS database to provide a census of all systems under USEPA jurisdiction. Considering only community water suppliers within the fifty states, the SDWIS database includes surface water 10,637 systems and 44,087 ground water systems. Figure 7 summarizes our results. In this figure, the posterior distribution of the model parameters given both the NAOS and USGS data is used to predict the fraction of all national systems which would exhibit raw water arsenic concentrations in excess of a number of alternative MCLs in a single-sample assessment of all systems. In particular,

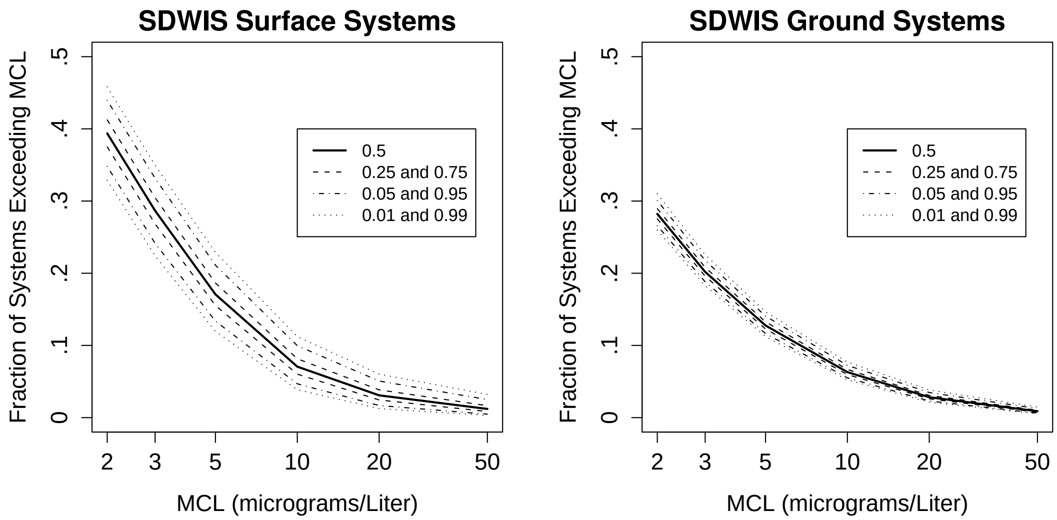


Figure 7. Predictions of fractions of national utilities exhibiting raw water arsenic concentrations in excess of alternative MCLs. Left figure represents predictions for 10,637 SDWIS surface systems and right figure for 44,087 SDWIS ground systems. Different lines represent indicated quantiles of posterior distribution. Predictions are based on a single-sample assessment from each system.

we consider MCLs of 2, 3, 5, 10, 20 and 50 $\mu g/L$, which includes both the new standard and others considered during the public comment period (USEPA 2000a). The posterior distributions of the fractions of systems with raw water arsenic concentrations exceeding the alternative MCLs are summarized by quantiles from their respective distributions.

There are several features of our predictions warranting comment. First, the uncertainty in the fraction of systems exceeding a certain MCL is greater for surface systems than for ground systems. For example, the model predicts that the percentage of surface water systems with raw water arsenic concentration in excess of 10 $\mu g/L$ is between 4% and 11%, while the corresponding range for ground water systems, between 5% and 8%, is considerably smaller. This is consistent with the fact that the inclusion of the USGS data reduces uncertainty for those functions of the parameters which describe the distribution of arsenic in ground water sources, but does so much less for those functions governing surface water arsenic concentrations. We also see that the uncertainty in the fraction of systems exceeding a certain MCL is a decreasing function of the MCL. This makes sense because the number of systems exceeding a given MCL, conditional on a particular value of the parameters, has a mixture of binomial distributions. As the MCL increases, the probabilities of exceeding this MCL for virtually all of those binomial distributions become closer to zero where the variance is smaller. Another interesting feature is that the median predicted behavior for the two types of systems is approximately the same for MCLs greater than or equal to 10 $\mu g/L$, but surface water systems are proportionally more problematic for MCLs less than 10 $\mu g/L$.

Finally, we examined how the predictions differed by state in addition to source type. Figure 8 shows the posterior median percentage of SDWIS systems expected to have raw

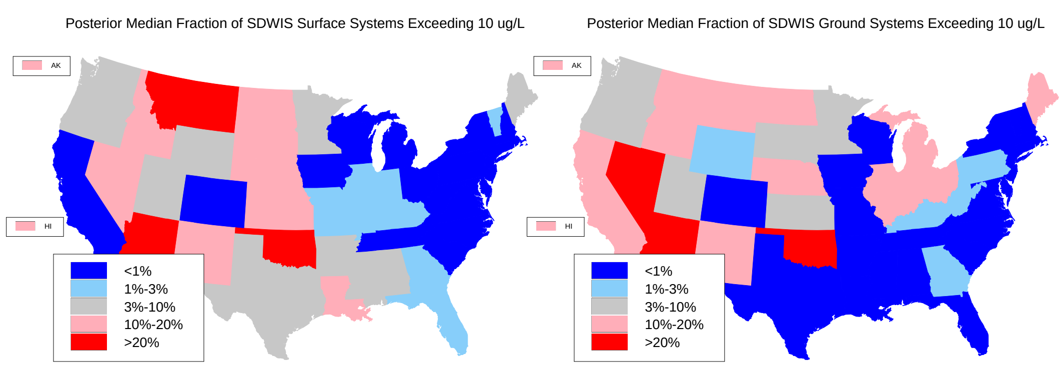


Figure 8. Posterior median percentage of SDWIS systems expected to have raw water arsenic concentrations exceeding 10 $\mu g/L$ separately for each state and source water type. Left frame is for surface water and right is for ground water. Results based on the fit of Model 7 to the combined NAOS and USGS data.

water arsenic concentrations exceeding 10 $\mu g/L$ separately for each state and source water type. The patterns observed are quite similar to those depicted in Figure 4. For both source types, the raw waters of systems in western states are generally more likely to exceed a given MCL than those in eastern states.

4.5 Examination of Uncertainty

Additionally, it is natural to consider which classes of water systems most substantially contribute to the nationwide uncertainty in the number of systems with raw water arsenic concentration in excess of a given MCL. It is these systems that should be the focus of future raw water arsenic data collection in order to most effectively inform any further regulatory decisions. Let n_{ij} be the number of SDWIS systems in state i ($i = 1, \ldots, 50$) using source water type j ($j = 1, 2$). Under the model, the number X_{ij} of these systems that would exhibit a raw water arsenic concentration in excess of, for example, 10 $\mu g/L$ in a single-sample assessment is distributed Binomial(n_{ij}, P_{ij}). Because P_{ij} is unknown, $\mathrm{Var}(X_{ij}) = \mathrm{Var}\left(\mathrm{E}(X_{ij}|P_{ij})\right) + \mathrm{E}\left(\mathrm{Var}(X_{ij}|P_{ij})\right)$, where the outer variance and expectation operations on the right hand side are with respect to the posterior distribution of P_{ij}. The term $\mathrm{E}\left(\mathrm{Var}(X_{ij}|P_{ij})\right)$ essentially quantifies only the variability in X_{ij} conditional on a particular value of P_{ij}, and it is positive even if P_{ij} is known precisely. This is in contrast to $\mathrm{Var}\left(\mathrm{E}(X_{ij}|P_{ij})\right) = n_{ij}^2 \mathrm{Var}(P_{ij})$. In order for a class of systems to have a large value of $n_{ij}^2 \mathrm{Var}(P_{ij})$, it must have both a substantial number of systems and significant uncertainty about P_{ij}, and the term is zero when P_{ij} is known. This matches what intuitively characterizes a class of systems that would benefit most significantly from future data collection, and thus we use this quantity to rank order the 100 classes of systems delineated by our model. Figure 9 plots the square root of $n_{ij}^2 \mathrm{Var}(P_{ij})$ for all classes of systems. Generally, ground water systems exhibit higher uncertainty than surface water systems. This indicates that

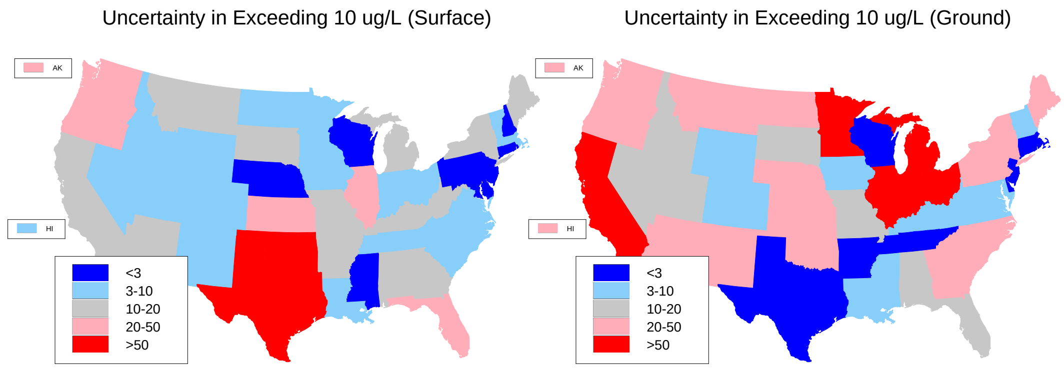


Figure 9. Uncertainty estimates of the number of systems with raw water arsenic concentrations in excess of 10 μg/L for each state and source water type. Plotted quantity is $n_{ij}\,Var^{1/2}(P_{ij})$.

although ground water systems already represent over 93% of our data, this disparity is not sufficient to overcome the fact that they comprise over 80% of the community water systems. However, because $n_{ij}\mathrm{Var}^{1/2}(P_{ij})$ balances numbers of systems against uncertainty in arsenic occurrence, it is not simply the classes with the largest numbers of systems that dominate the uncertainty. For example, in the Midwest, the ground water systems of Ohio, Michigan, Indiana and Minnesota contribute most heavily to uncertainty, while none of these states ranks in the upper quartile with respect to the number of ground water systems.

5. DISCUSSION

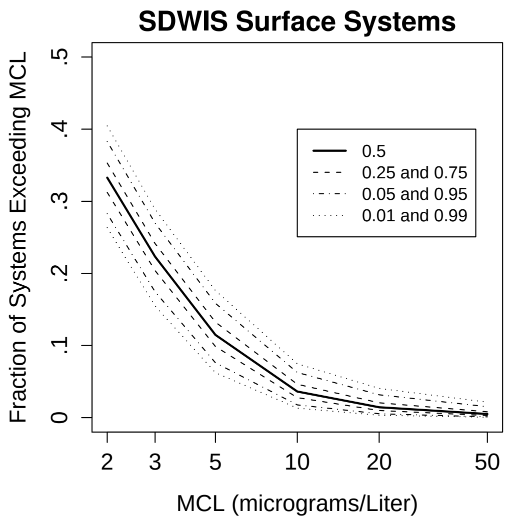
5.1 Summary of Analysis

Through simulation, our model lends insight into national raw water arsenic levels in community water systems. The ability to make statements about uncertainty highlights a primary advantage of our Bayesian hierarchical model over simpler, more traditional procedures. A common technique for estimating fractions (and eventually numbers) of water systems that would exceed a given MCL based on observations of only a subset of systems is the following: the population of community water suppliers is cross–classified by such characteristics as source type, location and number of people served, and the collection of sampled systems is subjected to the same stratification. The fraction of total systems in each class that exceed the MCL is then assumed to be equal to the corresponding fraction observed in the sample. While this procedure can offer a reasonable point estimate, it cannot be used to estimate the uncertainty in the estimate without an underlying statistical model. In contrast, the Bayesian approach to the problem coherently quantifies this uncertainty.

5.2 Shortcomings and Suggestions for Future Data Collection

A major drawback to the present analysis, which is a consequence of the nature of the currently available data, is the inability to distinguish between various sources of variability and uncertainty concerning arsenic occurrence. For example, because the data consist of a single assessment from each system, we have no way to distinguish sampling variability within a system from the variability among utilities with the same covariates. A much stronger and more valid model would separate these sources of variability, which would require more detailed data than those generally available now. There are some state data on arsenic occurrence which have multiple observations within a single water system, but the value of multiple measurements has generally been ignored by those reporting and analyzing data. For example, one analysis of the state occurrence data began by replacing all multiple observations with their average and then treating the average as a single measurement (USEPA 1994).

The issue of within–system variability warrants more sophisticated consideration. As a complement to the numerous existing single-sample arsenic data sets, it would be quite useful to have a systematic, multiple-sample assessment to address the issue of intra-system variability on a national basis. It is unlikely that the cost of assessing a small number of nationally representative utilities would be any more prohibitive than a single-sample study of a large number of utilities, while the potential to aid our understanding of arsenic occurrence is high. To see how much the availability of multiple measurements could affect our predictions, we return to the prediction of the number of systems whose raw waters would need treatment in order to meet various MCLs. One procedure recently used for determining compliance with the current MCL was the following. Take one sample, and if it is below the MCL, declare the system to be in compliance. If the one sample is above the MCL, take three more samples and average the four. Declare the system to be out of compliance if and only if this result exceeds the MCL. We simulated this procedure by assuming that the variation in arsenic concentrations within a system is equal to that between systems with the same covariates and produced Figure 10. The comparison of Figures 7 and 10 illustrates two extreme cases that could arise from data with multiple measurements. For example, if the variability within a system were 0 and all variability were between systems, then Figure 7 would not change if the more complicated assessment procedure described here were used because all four observations from a given system would be the same. At the other extreme, if there were no variability between system means and all variability between systems were the same as the variability within systems, then Figure 10 would be appropriate for predicting because we used the same mean and the same variance to simulate all samples from systems with the same covariates. Clearly, neither of these extremes is realistic, and the truth should be somewhere in between. Figure 11 shows the medians from all of the plots in Figures 7 and 10. It allows comparison of the two extremes



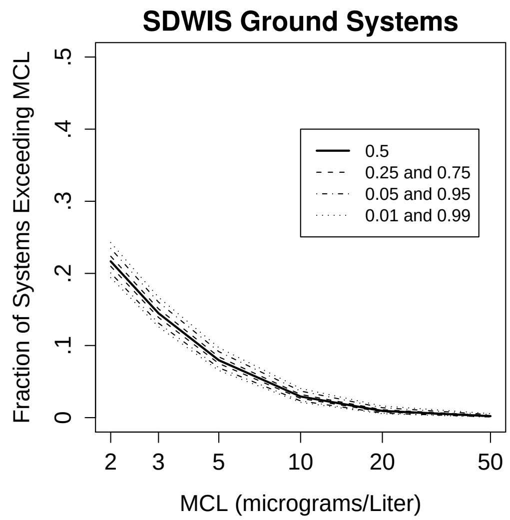


Figure 10. Predictions of fractions of national utilities exhibiting raw water arsenic concentrations in excess of various potential MCLs. Left figure represents predictions for 10,637 SDWIS surface systems and right figure for 44,087 SDWIS ground systems. Different lines represent indicated quantiles of posterior distribution. Predictions are based on a procedure of drawing a single sample and, if it is too large, averaging it with three additional samples.

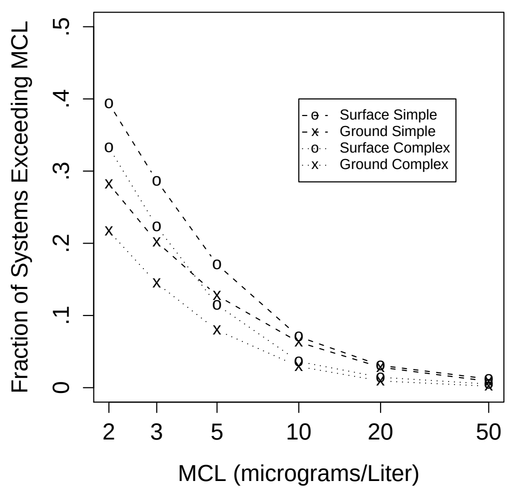


Figure 11. Comparison of simple and more complex compliance assessment methods. The plotted lines are the same as the median lines from Figures 7 and 10.

together with direct comparison of ground and surface waters. The differences between the two extreme cases is on the same order of magnitude as the difference between ground and surface for small MCLs and it is even larger than the ground/surface difference at larger MCLs. This suggests that it would be very valuable to be able to separate within system variability from between system variability. Of course, compliance with an MCL will be based on finished water rather than raw water, and we will revisit this issue when we have a model for finished water.

Another potential shortcoming is a lack of comparability between the measurements in the NAOS and USGS data sets. There are a variety of measurement techniques for arsenic which are specifically designed to give consistent results, but it is likely that measurements produced by different laboratories are systematically different to some extent. Because the two data sets were collected by independent agencies, it is possible that we should not be treating the measurements as if they came from a common population. In fact, the results presented in Figure 6 may indicate that the two data sets actually represent two distinct populations of measurements. If the data sets were two large samples from the same population, we would expect the effect of updating the model with the USGS data to result in reduced uncertainty but not a major change in location or scale in the predictive distribution of the NAOS data. A part of the shift we see in the right frame of Figure 6 is a result of the differences between the two data sets which result from sampling variability, but it is possible that the shift is also in part due to underlying differences between the two populations of measurements. We discuss this further in Section 5.3.

Finally, the spatial resolution of our analysis was not ideal. We aggregated systems at the state level for our analysis because this was the most specific location information offered by the NAOS and initially available SDWIS water system data. While this provides some degree of spatial differentiation sufficient to capture gross trends in arsenic occurrence, a more valid approach would use more refined geographical data able to capture occurrence trends at smaller spatial scales. In particular, more specific spatial information about the systems would make possible the consideration of the local influences that probably dominate covariation in arsenic levels between neighboring systems.

5.3 Future Work

We are currently pursuing a number of extensions to this analysis. First, we are in the process of assessing the need to model different arsenic data sets differently, in light of what we saw in Figure 6. The most straightforward approach, in the spirit of meta-analysis, would be to include a random effect in the two equations (1) and (2) in Stage 1 of the model. The random effect would have different values for each data set, allowing each data set to have slightly different mean and variance, but otherwise the same structure across states, size classes, and source types. We have not included such an analysis in this paper for two reasons. First, the current model fits surprisingly well to both data sets as it is, and second, we need to make sure that a model that distinguishes data sets is still able to make out-of-sample predictions for the entire nation. There is an issue as to whether the national predictions using the SDWIS database should be considered as a third sample, an extension to one of the original two samples, or some collection of more than one new sample.

Next, we are gathering information on existing treatment technologies and their removal efficiencies to build models for simulating arsenic treatment. We intend to couple the treatment models and the occurrence models to produce estimates of national finished water arsenic concentrations which will be validated and updated with observed finished water data. In addition to an assessment of existing treatments, we are also building a cost-based model for the choice of additional treatment technologies by systems requiring them. The net result of our models will be an estimate of the national cost of compliance with various MCLs. Uncertainties in the costs and efficiencies of the treatment technologies will be combined with the uncertainties in the national occurrence distribution in order to characterize the uncertainties in our cost estimate.

Finally, we intend to generalize the entire analysis in the previous paragraph to the simultaneous consideration of multiple contaminants. Current EPA RIAs generally consider the costs associated with lowering concentrations of contaminants individually (e.g., Odom, Regli, Messner, Cromwell, and Javdan 1999; Regli, Odom, Cromwell, Lustic, and Blank 1999). The failure to consider contaminants jointly ignores the overlap in effectiveness of treatments, possible interference of different treatment types, and the potentially "additive or synergistic" toxic effects displayed by some contaminants (Kolpin, Barbash, and Gilliom 1998). Also, the joint occurrence of contaminants must be explicitly modeled because underlying physical processes as well as human activity tend to cause certain contaminants to be found with others (Kolpin et al. 1998; Grady and Weaver 1989; Page 1981). We have made substantial progress in generalizing the models discussed in this paper to the analysis of multiple contaminant occurrence.

APPENDIX A: CENSORING

A common problem in quantifying levels of trace substances is that concentrations can be so low that even sensitive equipment can have difficulty detecting the presence of the chemical, making measurements at low concentrations unreliable. The extent to which this occurs depends on the contaminant and the measurement technique. There are several distinct notions of the lower bound on the ability of a technique to measure trace elements. The first, called the *method detection limit* (MDL), quantifies the ability of a technique to detect the presence of a substance. More specifically, MDL is defined as "the minimum concentration of an analyte that can be measured and reported with 99% confidence that the analyte concentration is greater than zero" (40 CFR Part 136.1). Put another way, the MDL is that concentration which has probability 0.01 of being exceeded when measured on a blank (zero concentration) sample (Berthouex and Hau 1991).

While the MDL addresses the ability of a measurement technique to detect the presence of a substance, it does not reflect the ability of the technique to quantify accurately the concentration of that substance. In addition to the MDL, researchers generally provide a *reliable quatitation limit* (RQL) that represents the lowest concentration of a substance that is not only likely to be detected, but also likely to be accurately measured. This is necessary because although a measurement technique may be able to detect a very low concentration of a substance, it may not be able to produce a reliable quantitative measurement of the amount unless the concentration is higher. Like the MDL, the RQL is specific to a particular measurement technique and substance, and the RQL is obviously bounded below by the MDL. The RQL is related to a third lower bound used in measuring concentrations called the *practical quantitation limit* (PQL), which is a standard set for each contaminant by the USEPA. It is defined as "the lowest concentration of an analyte than can be reliably measured within the specified limits of precision and accuracy during routine laboratory operating conditions." (52 FR 25699). The idea of behind the PQL is that it is intended to be a "meta-RQL", providing a concentration above which most laboratories can expect to be able to make reliable measurements. The USEPA determines PQLs by examining the measurement techniques used by different laboratories (USEPA 2000a).

By the definition of the RQL (or PQL), if a measured concentration is above the RQL, it is reasonably certain that the measurement is an accurate reflection of the underlying concentration. Moreover, in the absence of additional information to the contrary, it is usually assumed that measured concentrations below the RQL represent a true concentration below the RQL. Thus, measured concentrations below the RQL can properly be treated as left censored (Gibbons 1994).

Both the NAOS and USGS arsenic data contain measurements reported to be less than some particular concentration. Careful reading of the available documentation of these studies leads to the conclusion that in both cases, these concentrations represent a particular RQL. Hence, all measurements of this type are treated as left censored in all analyses in this report, and as such, this report generally refers to the RQL as a "censoring point." As noted previously, the censoring point for all NAOS arsenic observations is 0.5 $\mu g/L$, and among the 457 NAOS systems, 163 have censored arsenic concentrations. For the 1945 USGS community water system arsenic measurements, the censoring point is 1 $\mu g/L$ and there are 789 censored observations. The PQL for arsenic is 3 $\mu g/L$ (USEPA 2000a), indicating that both the NAOS and the USGS used more sensitive techniques than those used by other laboratories. In fact, the NAOS censoring point of 0.5 $\mu g/L$ is among the lowest of those used in existing arsenic occurrence assessments (Frey and Edwards 1997).

APPENDIX B: DIFFICULTIES IN CARRYING OUT MCMC WITH THE MATRIX $\boldsymbol{C}(\rho)$:

In order to carry out Hasting's algorithm in fitting the hierarchical models, it is necessary, for example, to sample from the conditional posterior distribution of $\boldsymbol{\alpha}^{(\boldsymbol{m})}$ given the other parameters, including ρ. Examining the structure of the matrix $\boldsymbol{C}(\rho)$, it is evident that as $\rho \to 0$, this matrix approaches the rank 1 matrix with all entries equal to 1. In large samples from the posterior with a prior for ρ which has mass near zero, it is inevitable that some sampled values of ρ will be very small, causing $\boldsymbol{C}(\rho)$ to be numerically singular. This presents two problems. The first is that the Metropolis algorithm in this parameterization requires the inverse of this matrix to calculate the prior density of a proposed value $\boldsymbol{\alpha}^{(m)*}$ of $\boldsymbol{\alpha}^{(m)}$ conditional on ρ. This is easily handled by calculating a generalized inverse of the matrix (Graybill 1976). The more serious problem, however, cannot be solved in this manner. The difficulty arises in accepting the proposed value $\boldsymbol{\alpha}^{(m)*}$. When ρ is very small, there will be at least one linear function $\boldsymbol{\lambda}'\boldsymbol{\alpha}^{(m)}$ which has very high probability of being almost constant, that constant being $\boldsymbol{\lambda}'\boldsymbol{\alpha}_0^{(m)}$. That is, $\boldsymbol{\alpha}^{(m)}$ has essentially a singular prior distribution conditional on small values of ρ. Because the proposed value is sampled by perturbing the elements of $\boldsymbol{\alpha}^{(m)}$ according to some proposal distribution, it is extremely unlikely that the proposed vector will satisfy all of the necessary (near) linear constraints. Hence, the prior density at the proposed value will be nearly zero, making it virtually impossible to accept this value. The net result is unacceptably slow mixing of the Markov chain. A similar problem occurs when trying to update ρ. If $\boldsymbol{C}(\rho)$ is nearly singular, then a small change to a proposed value ρ^* will produce another nearly singular matrix, but with different linear functions that need to be nearly constant. The current value of $\boldsymbol{\alpha}^{(m)}$ will most likely not satisfy the new linear constraints, making it unlikely that the proposed value will be accepted.

A solution to this problem is to reparameterize to parameters that are uncorrelated *a priori*, resulting in a parameter space which is more efficiently navigated. Instead of directly considering the parameter $\boldsymbol{\alpha}^{(m)}$, introduce a parameter $\boldsymbol{\theta}_{\alpha^{(m)}}$ which has a k-dimensional standard normal prior distribution. $\boldsymbol{\theta}_{\alpha^{(m)}}$ is related to $\boldsymbol{\alpha}^{(m)}$ by

$$\boldsymbol{\alpha}^{(m)} = \boldsymbol{\alpha}_0^{(m)} + \frac{1}{\sqrt{\tau_\alpha^{(m)}}} \boldsymbol{L}(\rho) \boldsymbol{\theta}_{\alpha^{(m)}}, \tag{A.1}$$

where $\boldsymbol{L}(\rho)$ is a lower-triangular Cholesky factor of the matrix $\boldsymbol{C}(\rho)$ satisfying $\boldsymbol{L}(\rho)\boldsymbol{L}'(\rho) = \boldsymbol{C}(\rho)$. Rudimentary results about transformations of normal random variables implies that conditional on ρ and $\tau_\alpha^{(m)}$, the prior distribution for $\boldsymbol{\alpha}^{(m)}$ is $N(\boldsymbol{\alpha}_0^{(m)}, \boldsymbol{C}(\rho)/\tau_\alpha^{(m)})$. Thus the prior distribution of the parameter of interest in unchanged. However, the new parameterization greatly alleviates the difficulty in accepting proposed values of $\boldsymbol{\alpha}^{(m)}$ when ρ is small. Suppose that ρ is small enough such that there exists $\boldsymbol{\lambda}$ with the property that $\boldsymbol{\lambda}'\boldsymbol{\alpha}^{(m)}$ is nearly degenerate at $\boldsymbol{\lambda}'\boldsymbol{\alpha}_0^{(m)}$. Then as described, a proposed value $\boldsymbol{\alpha}^{(m)*}$ of $\boldsymbol{\alpha}^{(m)}$ obtained by directly perturbing $\boldsymbol{\alpha}^{(m)}$ will almost certainly be rejected. However, when treating $\boldsymbol{\theta}_{\alpha^{(m)}}$ as the parameter, a proposed value $\boldsymbol{\theta}^*_{\alpha^{(m)}}$ is obtained by perturbing $\boldsymbol{\theta}_{\alpha^{(m)}}$ and is substituted for it in Equation (A.1) in order to obtain $\boldsymbol{\alpha}^{(m)*}$. Because $\boldsymbol{\lambda}'\boldsymbol{L}(\rho)$ is proportional to the standard deviation of $\boldsymbol{\lambda}'\boldsymbol{\alpha}^{(m)}$, it is very nearly a zero vector. Hence the value of $\boldsymbol{\lambda}'\boldsymbol{\alpha}^{(m)*}$ will be very near $\boldsymbol{\lambda}'\boldsymbol{\alpha}_0^{(m)}$ as required. Similar benefits are realized when attempting to move from ρ to a proposed value ρ^*.

The above approach is most beneficial anytime $\boldsymbol{C}(\rho)$ has at least one small eigenvalue, but in fact can be used to improve mixing in any situation where some parameter vector has a multivariate normal prior distribution with non-zero correlations. As written, it requires that the Cholesky factor $\boldsymbol{L}(\rho)$ be defined, which in turn requires that $\boldsymbol{C}(\rho)$ be nonsingular. However, in the course of the MCMC routine used to sample from the posterior distribution in the current problem, one will eventually obtain a value of ρ that is small enough that $\boldsymbol{C}(\rho)$ is numerically singular. Fortunately, a method exists for calculating a lower rank matrix $\boldsymbol{L}(\rho)$ with the property that $\boldsymbol{L}(\rho)\boldsymbol{L'}(\rho)$ is still approximately equal to $\boldsymbol{C}(\rho)$, allowing the method to be used no matter how small ρ becomes.

In order to describe the modification, suppose the goal is to factor $C = LL'$ where $C = ((c_{i,j}))$ is a $k \times k$ positive definite symmetric matrix. The usual Cholesky factorization sets $L = ((\ell_{i,j}))$ where $\ell_{1,1} = \sqrt{c_{1,1}}$, $\ell_{i,1} = c_{i,1}/\ell_{1,1}$ for $i = 2, \ldots, k$, and for $j > 1$, $\ell_{j,j} = \sqrt{c_{j,j} - \sum_{i=1}^{j-1} \ell_{j,i}^2}$ and for $i = j+1, \ldots, k$, $\ell_{i,j} = (c_{i,j} - \sum_{k=1}^{j-1} \ell_{j,k}\ell_{i,k})/\ell_{j,j}$ (Thisted 1988). Numerical problems arise when $\sum_{i=1}^{j-1} \ell_{j,i}^2 \geq c_{j,j}$, which cannot happen in theory, but which will often occur in practice when dealing with matrices that are nearly singular. When $c_{j,j} - \sum_{i=1}^{j-1} \ell_{j,i}^2$ becomes less than some value ϵ, one can set $\ell_{i,j} = 0$ for $i = j, \ldots, n$. (A value that worked well in the applications of this report is $\epsilon = 10^{-6}$.) The resulting matrix has nice properties for the purpose for which it is needed. If C as viewed as the covariance matrix of a vector $X = LZ$, where the coordinates of Z are uncorrelated, then the first $j-1$ entries in row j of L equal the coefficients of linear regression of X_j on $(Z_1, \ldots, Z_{j-1})$ and $\ell_{j,j}$ is the residual standard deviation of this regression. Setting $\ell_{j,j} = 0$ together with the rest of column j is tantamount to declaring that X_j is perfectly predicted by $(Z_1, \ldots, Z_{j-1})$, equivalently that it is perfectly predicted by $(X_1, \ldots, X_{j-1})$, and that there is no residual variation. Also, it is not necessary use Z_j for predicting X_i for $i > j$. These facts allow the "inversion" Equation (A.1) even though $\boldsymbol{L}(\rho)$ might be singular.

APPENDIX C: ESTIMATING MISSING POPULATIONS

Section 3.5 considers a number of models in which population served by each system is included as a covariate in either a threshold form (Models 8, 9 and 10) or a quantitative linear form (Models 11 through 18). Unfortunately, some of the NAOS observations and all of the USGS observations are missing population figures. This appendix details how missing populations are treated in the two types of models, both in terms of fitting the models and in terms of the calculation of predictive densities. As will be shown, how to treat missing population figures correctly depends on whether the missing information is part of the training data or the testing data.

In order to illustrate the difference, denote all arsenic measurements by $\boldsymbol{Y} = (\boldsymbol{Y}_0, \boldsymbol{Y}_1)$ where $\boldsymbol{Y}_0$ is the training data and $\boldsymbol{Y}_1$ is the testing data. Analogously partition all observed covariates $\boldsymbol{X} = (\boldsymbol{X}_0, \boldsymbol{X}_1)$, which includes the location (state), source water type and the observed population figures. Let $\boldsymbol{Z} = (\boldsymbol{Z}_0, \boldsymbol{Z}_1)$ denote the collection of unobserved population figures for the training and testing data. Finally, let $\boldsymbol{\theta}$ denote all parameters for the model under consideration, which is arbitrary for the purpose of this discussion. For ease of notation, $f(\cdot|\cdot)$ is used to denote all density functions; exactly which density function is indicated by its arguments.

The joint distribution of all model quantities, both observable and unobservable, is

$$\begin{aligned} f(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\theta}) &= f(\boldsymbol{\theta}) f(\boldsymbol{x}|\boldsymbol{\theta}) f(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\theta}) f(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{x}, \boldsymbol{\theta}) \\ &= f(\boldsymbol{\theta}) f(\boldsymbol{z}|\boldsymbol{x}) f(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{x}, \boldsymbol{\theta}) \\ &= f(\boldsymbol{\theta}) f(\boldsymbol{z}_0|\boldsymbol{x}_0) f(\boldsymbol{z}_1|\boldsymbol{x}_1) f(\boldsymbol{y}_0|\boldsymbol{z}_0, \boldsymbol{x}_0, \boldsymbol{\theta}) f(\boldsymbol{y}_1|\boldsymbol{z}_1, \boldsymbol{x}_1, \boldsymbol{\theta}) \end{aligned} \tag{A.2}$$

The first equation is just a standard factorization, the second follows from the fact that the observed covariates are not considered random and that all covariates are assumed to be independent of the parameters, and the third from the fact that all aspects of the testing data are conditionally independent of the training data given the parameters. It is from this joint distribution that the two quantities of interest are calculated.

C.1 Model Fit

Fitting the model refers to finding the posterior distribution $f(\boldsymbol{\theta}|\boldsymbol{y}_0, \boldsymbol{x}_0)$ of $\boldsymbol{\theta}$ given the observable training data. Implicitly this is integrated over the missing population figures, which can essentially be viewed as nuisance parameters. That is,

$$f(\boldsymbol{\theta}|\boldsymbol{y}_0, \boldsymbol{x}_0) = \int f(\boldsymbol{\theta}, \boldsymbol{z}_0|\boldsymbol{y}_0, \boldsymbol{x}_0) f(\boldsymbol{z}_0|\boldsymbol{y}_0, \boldsymbol{x}_0) d\boldsymbol{z}_0 \tag{A.3}$$

In a Monte Carlo framework, this integral is performed by obtaining a sample from the joint posterior distribution $f(\boldsymbol{\theta}, \boldsymbol{z}_0|\boldsymbol{y}_0, \boldsymbol{x}_0)$ and ignoring the sampled values of $\boldsymbol{z}_0$. This joint posterior distribution is derived only from the factors $f(\boldsymbol{\theta}) f(\boldsymbol{z}_0|\boldsymbol{x}_0) f(\boldsymbol{y}_0|\boldsymbol{z}_0, \boldsymbol{x}_0, \boldsymbol{\theta})$ from Equation (A.2). The term $f(\boldsymbol{z}_0|\boldsymbol{x}_0)$ can be viewed as the prior distribution for the missing population figures and was derived from the SDWIS census. However, while the models in Section 3.5 make use of actual numbers of people served, SDWIS contains population information only in the form of USEPA size category, consisting of twelve size classes (Raucher et al. 1995). For each system, SDWIS provides the USEPA size category, and for each size category, SDWIS provides the total number of people served by *all* systems in that size category. Based on this information, populations were assigned to each system equal to the total number of people served by its size category divided by the number of systems in that size category. The systems were then classified by state and source water type, and the maximum likelihood normal distribution was fit to the logarithms of the estimated population values in order to characterize the distribution of population within that location and source category. The means and standard deviations of the resulting distributions for each state and source water type are given in Table C.1. These distributions are used as

FIPS	Surface		Ground		FIPS	Surface		Ground	
Code	μ	σ	μ	σ	Code	μ	σ	μ	σ
AK	5.721	1.377	5.049	1.148	MT	6.325	1.870	5.078	1.158
AL	8.396	1.422	7.586	1.138	NC	7.831	1.736	5.123	1.253
AR	7.307	1.456	6.640	1.358	ND	6.681	1.752	5.917	1.177
AZ	7.567	2.626	5.749	1.587	NE	6.983	2.419	5.859	1.335
CA	7.776	2.636	5.580	1.799	NH	7.888	1.695	4.944	1.127
CO	6.676	2.492	5.267	1.334	NJ	9.963	1.398	6.697	1.989
CT	9.06	2.049	5.101	1.188	NM	7.179	1.911	5.587	1.494
DE	11.782	0.256	5.365	1.287	NV	7.555	2.307	5.378	1.426
FL	8.683	1.978	6.275	1.937	NY	7.085	1.935	5.386	1.529
GA	8.341	1.958	5.340	1.399	OH	8.115	1.949	6.219	1.730
HI	7.057	1.676	7.066	1.913	OK	6.426	1.646	6.034	1.316
IA	7.128	1.629	5.987	1.410	OR	7.196	1.938	5.264	1.346
ID	6.222	1.512	5.121	1.398	PA	8.119	1.877	5.531	1.404
IL	7.899	1.812	6.391	1.476	RI	9.324	1.851	5.436	1.858
IN	8.206	1.746	6.242	1.732	SC	8.312	1.629	5.394	1.576
KS	6.518	1.584	6.059	1.301	SD	6.417	1.488	5.543	1.268
KY	8.247	1.272	6.240	1.714	TN	8.171	1.688	6.963	1.681
LA	8.626	1.713	6.243	1.586	TX	7.585	1.886	5.982	1.513
MA	9.349	1.816	6.402	2.169	UT	8.390	2.073	6.015	1.535
MD	8.078	2.332	5.594	1.462	VA	7.691	2.041	5.080	1.162
ME	7.440	1.477	4.814	1.179	VT	6.865	1.732	5.190	1.114
MI	8.544	1.770	5.766	1.454	WA	7.050	2.377	5.215	1.402
MN	9.006	1.652	6.282	1.581	WI	9.539	1.362	5.793	1.691
MO	7.345	1.630	5.774	1.555	WV	6.992	1.414	5.687	1.442
MS	8.237	1.819	6.754	1.223	WY	7.267	1.748	5.221	1.164

Table 3. Estimated distributions for log population served for each state and source water type.

prior distributions for any missing population values. The assumption of lognormality is reasonable based on examination of the empirical population distributions in the various state and source classes.

Because the missing population figures are treated as parameters, they are sampled from their full conditional distributions at each iteration of the Markov Chain. Exactly how the missing populations are sampled depend on whether the model in question is of the threshold or regression type. For the regression models, the Metropolis algorithm (Metropolis et al. 1953) with an appropriately chosen normal proposal distribution is used to sample from the full conditional distribution of each missing population figure. For the threshold models, it is possible to identify and sample from the full conditional distribution of each missing population figure using traditional Gibbs sampling (Geman and Geman 1984). As a first simplification, because the full conditional distributions condition on all current parameter values as well as the data (both observed and imputed) and other covariates, it is possible to treat each missing population value as conditionally independent of the others and sample them one at a time. For now, focus on the logarithm Z of any one of these missing population figures and its associated location and source water type, which are always known. The only pieces of the joint distribution of all model quantities which depend on this missing population are its univariate normal prior distribution with mean μ and standard deviation σ depending on state and source water type as shown in Table C.1) and in the contribution to the likelihood function of its associated log arsenic concentration Y. This piece of the likelihood function, in turn, only depends on whether or not $Z \leq \delta$, where δ is the source-type specific cutoff between large and small systems. If $Z \leq \delta$, then Y has some normal distribution with parameters θ_s; otherwise, Y has some other normal distribution with parameters θ_b. Both the cutoff δ and the parameters θ_s and θ_b can be treated as known in the full conditional distribution. Hence, the conditional distribution of Z given all other quantities (labeled "$rest$") satisfies

$$f(z|rest) = \frac{f(z|\mu,\sigma)}{c}\left(f(y|\theta_s)I_{z\leq\delta} + f(y|\theta_b)I_{z>\delta}\right)$$

where c is some normalizing constant. Notice that Z does not show up in the terms $f(y|\theta_s)$ and $f(y|\theta_b)$ because the model specifies that the distribution of arsenic does not depend on the population itself, but rather only on whether the system is large or small. This makes it relatively easy to determine c by direct integration:

$$
\begin{aligned}
c &= \int f(z|rest)dz \\
&= f(y|\theta_s)\int_{z\leq\delta} f(z|\mu,\sigma)dz + f(y|\theta_b)\int_{z>\delta} f(z|\mu,\sigma)dz \\
&= f(y|\theta_s)\Phi\left(\frac{\delta-\mu}{\sigma}\right) + f(y|\theta_b)\left(1-\Phi\left(\frac{\delta-\mu}{\sigma}\right)\right)
\end{aligned}
$$

Hence

$$
f(z|rest) = \frac{\left[f(y|\theta_s)\Phi\left(\frac{\delta-\mu}{\sigma}\right)\right]\frac{f(z|\mu,\sigma)I_{z<\delta}}{\Phi\left(\frac{\delta-\mu}{\sigma}\right)} + \left[f(y|\theta_b)\left(1-\Phi\left(\frac{\delta-\mu}{\sigma}\right)\right)\right]\frac{f(z|\mu,\sigma)I_{z>\delta}}{1-\Phi\left(\frac{\delta-\mu}{\sigma}\right)}}{f(y|\theta_s)\Phi\left(\frac{\delta-\mu}{\sigma}\right) + f(y|\theta_b)\left(1-\Phi\left(\frac{\delta-\mu}{\sigma}\right)\right)}
$$

That is, the full conditional distribution for Z is a mixture of two distributions – one a normal distribution constrained to $Z \leq \delta$ and the other that same normal distribution constrained to $Z > \delta$. The mixing weights are the ratios of each of the summands of c to c itself and reflect the assimilation of the prior distribution of Z and the contribution of Z to the likelihood for its associated arsenic observation. Sampling from this distribution is straightforward.

Incidentally, for the threshold models, it would be possible to perform data augmentation on the random variable $I_{Z>\delta}$ rather than on Z itself. Because the actual value of Z is never needed, the only relevant quantities are the prior probability that $Z > \delta$ and the two densities for the data when Z is greater than or less than δ.

C.2 Calculations of Predictive Densities

The predictive density of the testing data given the training data is defined to be $f(\boldsymbol{y}_1|\boldsymbol{y}_0, \boldsymbol{x})$. Returning to the joint distribution in Equation (A.2), this is obtained by integrating the conditional distribution of the testing data given everything else over all unobservable quantities:

$$\begin{aligned} f(\boldsymbol{y}_1|\boldsymbol{y}_0,\boldsymbol{x}) &= \int f(\boldsymbol{y}_1|\boldsymbol{y}_0,\boldsymbol{x},\boldsymbol{z},\boldsymbol{\theta}) f(\boldsymbol{z},\boldsymbol{\theta}|\boldsymbol{y}_0,\boldsymbol{x}) d\boldsymbol{\theta} d\boldsymbol{z} \\ &= \int f(\boldsymbol{y}_1|\boldsymbol{y}_0,\boldsymbol{x},\boldsymbol{z},\boldsymbol{\theta}) f(\boldsymbol{\theta}|\boldsymbol{y}_0,\boldsymbol{x}) f(\boldsymbol{z}_0|\boldsymbol{\theta},\boldsymbol{y}_0,\boldsymbol{x}) f(\boldsymbol{z}_1|\boldsymbol{z}_0,\boldsymbol{\theta},\boldsymbol{y}_0,\boldsymbol{x}) d\boldsymbol{\theta} d\boldsymbol{z}_0 d\boldsymbol{z}_1 \\ &= \int f(\boldsymbol{y}_1|\boldsymbol{x}_1,\boldsymbol{z}_1,\boldsymbol{\theta}) f(\boldsymbol{\theta}|\boldsymbol{y}_0,\boldsymbol{x}_0) f(\boldsymbol{z}_0|\boldsymbol{\theta},\boldsymbol{y}_0,\boldsymbol{x}_0) f(\boldsymbol{z}_1|\boldsymbol{x}_1) d\boldsymbol{\theta} d\boldsymbol{z}_0 d\boldsymbol{z}_1 \end{aligned}$$

The first equation is just a definition, the second results from standard factorization of the distribution of the unobservables, and the third follows directly the from the conditional independence assumptions presented in the development of Equation (A.2).

For both the regression models and the threshold models, Monte Carlo integration is used to calculate this predictive density. The integration in both cases is of $f(\boldsymbol{y}_1|\boldsymbol{x}_1, \boldsymbol{z}_1, \boldsymbol{\theta})$ with respect to $f(\boldsymbol{z}, \boldsymbol{\theta}|\boldsymbol{y}_0, \boldsymbol{x})$, or equivalently with respect to $f(\boldsymbol{\theta}, \boldsymbol{z}_0|\boldsymbol{y}_0, \boldsymbol{x}_0) f(\boldsymbol{z}_1|\boldsymbol{x}_1)$. This latter expression is more enlightening because the first factor is simply the posterior distribution given the training data, a sample from which is already obtained in order to fit the model. Letting $(\boldsymbol{\theta}_i, \boldsymbol{z}_{0,i})$ for $i = 1, \ldots, M$ denote this sample, then the predictive density can be estimated by

$$\widehat{f}(\boldsymbol{y}_1|\boldsymbol{y}_0,\boldsymbol{x}) = \frac{1}{M} \sum_{i=1}^{M} \int f(\boldsymbol{y}_1|\boldsymbol{x}_1,\boldsymbol{z}_1,\boldsymbol{\theta}_i) f(\boldsymbol{z}_1|\boldsymbol{x}_1) d\boldsymbol{z}_1 \tag{A.4}$$

Note that the sampled values $\boldsymbol{z}_{0,i}$ are not needed for this calculation. The more important point is how the final level of integration, over the missing population values of the testing data, is performed. Similar to what is experienced when fitting the models, this is where the treatment of the regression and threshold models diverges. For the regression models, the integrand depends on the missing population values in a continuous fashion, and the integral cannot generally be calculated in closed form. [2] Hence, another level of Monte Carlo integration is used, where for each sampled value of $(\boldsymbol{\theta}_i, \boldsymbol{z}_{0,i})$ is also sampled a single collection $\boldsymbol{z}_{1,i}$ of missing populations from $f(\boldsymbol{z}_1|\boldsymbol{x}_1)$, and

$$\frac{1}{M}\sum_{i=1}^{M}\int f(\boldsymbol{y}_1|\boldsymbol{x}_1, \boldsymbol{z}_{1,i}, \boldsymbol{\theta}_i) \tag{A.5}$$

is taken as the estimated predictive density. Nothing precludes the use of more than one set of sampled values of $\boldsymbol{z}_1$ at each iteration; such a procedure would have lower Monte Carlo variance. However, because the MCMC analyses in this report are typically performed for a large number of iterations (10^6 or greater), it was not prudent to introduce any additional computational burden at each iteration.

For the threshold models, in which the conditional distribution of each observation given everything else depends only on whether the associated population is greater or smaller than some threshold δ, the integration over $f(\boldsymbol{z}_1|\boldsymbol{x}_1)$ can be done explicitly for each sampled $(\boldsymbol{\theta}_i, \boldsymbol{z}_{0,i})$. When factoring the integral into a product of pieces representing the contribution of each individual observation in the testing data, the integrand of each constituent integral takes on only one of two values depending on whether the system is "large" or "small". This is analogous to the derivation of the full conditional distributions of the missing populations given previously, and in fact is easier because the integration is over $f(\boldsymbol{z}_1|\boldsymbol{x}_1)$ rather than some more complicated distribution. Let Y denote an observation in the testing data for

[2] It is actually possible to do the calculation in closed form for the models in which only the data means depend on population. However, for programming reasons, it was more efficient to deal with all regression models in the same fashion. This necessitated the Monte Carlo rather than the analytical approach.

which the population is missing, $\boldsymbol{\theta}_{s,i}$ be the parameters of the normal distribution assumed for Y if the system is small, and $\boldsymbol{\theta}_{b,i}$ be the parameters of the normal distribution assumed for Y if the system is large. Finally, let μ and σ be the mean and standard deviation of the normal distribution for the missing population. Then the contribution to the predictive density of this observation is

$$f(y|\boldsymbol{\theta}_{s,i})\Phi\left(\frac{\delta-\mu}{\sigma}\right)+f(y|\boldsymbol{\theta}_{b,i})\left(1-\Phi\left(\frac{\delta-\mu}{\sigma}\right)\right) \tag{A.6}$$

APPENDIX D: PRIOR DISTRIBUTIONS AND FITTING

This appendix contains the prior distribution specifications for all models described in this report. It also discusses some implementational details and sensitivity analyses not given elsewhere. Models 7 through 18 as well as Model 3B in Section 3.5 all required the specification of prior distributions. First, for Models 7 through 18, $\log(\rho) \sim N(\mu_\rho, \xi_\rho)$, and μ_ρ was set to -12.5 and ξ_ρ to $156.25 = 12.5^2$. This prior is relatively non-informative, placing non-negligible prior density in the vicinity of $\log(\rho) = 5$ (corresponding to the states being essentially independent) and $\log(\rho) = -30$ (corresponding to the states being nearly perfectly correlated). In fact, for practical purposes, the range of values of $\log(\rho)$ that reflect these extreme correlation structures is about $(-20, -5)$, over which the log prior density ranges only 0.18. To check the sensitivity of various model inferences to this specification, prior distributions for $\log(\rho)$ centered at other locations with smaller standard deviations were tried. Almost all priors led to posterior distributions for ρ that were located in about the same place, and thus the model fits and predictive results were largely unaffected. The only exceptions were priors that had very small standard deviations and had means that were very far from -12.5 (for example, mean of 0 and standard deviation of 1). In addition, periodically throughout the modeling process, models with a different ρ parameter for each of the four fundamental vectors were fit. This also had virtually no impact on either the model fits or predictions.

The prior precision parameters $\tau_\alpha^{(m)}$, $\tau_\alpha^{(v)}$, $\tau_\beta^{(m)}$ and $\tau_\beta^{(v)}$ are modeled as iid $N(\mu_\tau, \xi_\tau)$, where $\mu_\tau = 0$ and $\xi_\tau = 1$. This centers the prior for the τ near 1, representing one observation-worth of data in the prior, but still provides enough spread to allow values both near 0 and as far away as 10. An alternate prior with $\xi_\tau = 5^2$ was also used in some analyses with Model 7 to check sensitivity to that aspect of the prior, but it had virtually no impact.

All fifty components of each of the vectors $\boldsymbol{\beta}_0^{(m)}$ and $\boldsymbol{\beta}_0^{(v)}$ were set to zero, reflecting no distinction between the states and source water types *a priori*. While it was expected that these covariate groups would exhibit somewhat different arsenic profiles, there was no clear prior knowledge about the magnitude of these differences. Hence, we decided to let the data drive the inferences about source water distinctions.

The situation with the parameters $\boldsymbol{\alpha}_0^{(m)}$ and $\boldsymbol{\alpha}_0^{(v)}$ is somewhat more complicated. This report has results based on two different specifications for the parameters. Some initial analyses used a default prior where all fifty components of each of the vectors were set to zero. This reflects a marginal arsenic distribution which is the same for all states and source water types, and this distribution is consistent with a median arsenic value of 1 $\mu g/L$ and values of 50 $\mu g/L$ being rare. The results presented in Appendices E, F, and part of G all are based on this prior specification. The primary results presented in the main part of the document (including the validation study in Section 3.5), all are based on a different specification which was based on expert elicitation. This distribution sets the fifty components of $\boldsymbol{\alpha}_0^{(m)}$ to $\log(2)$ and the fifty components of $\boldsymbol{\alpha}_0^{(v)}$ to -0.95177. Like the default specification, the effect is that marginal distribution of arsenic is *a priori* the same for all states and source types. In this case, however, the specification is consistent with a median arsenic value of 2 $\mu g/L$ and values of 250 $\mu g/L$ being rare. These priors were also used for the parameters of Model 3B. In practice, the differences in model fit and predictions for the models based on the two different specifications were relatively minimal.

Both the population threshold and population linear models have additional parameters that quantify the dependence of arsenic concentrations on the numbers of people served by the systems. The population threshold models introduce four additional sets of random effects labeled $\boldsymbol{\gamma}^{(m)}$,$\boldsymbol{\gamma}^{(v)}$,$\boldsymbol{\eta}^{(m)}$ and $\boldsymbol{\eta}^{(v)}$. Their prior means, $\boldsymbol{\gamma}_0^{(m)}$,$\boldsymbol{\gamma}_0^{(v)}$,$\boldsymbol{\eta}_0^{(m)}$ and $\boldsymbol{\eta}_0^{(v)}$, were all set to vectors of zeros, corresponding to a prior belief that the net effect of the population covariate is zero. Their prior precisions $\tau_\gamma^{(m)}$, $\tau_\gamma^{(v)}$,$\tau_\eta^{(m)}$ and $\tau_\eta^{(v)}$ are estimated by the models and had the same standard normal prior distribution as that used for the corresponding parameters of the distributions of the source type main effects and adjustments. For both the size adjustments and their prior precisions, there was little sensitivity of the results to the prior distributions. The two size class cutoff parameters δ_{-1}, δ_{+1} are modeled as iid $N(\mu_\delta, \xi_\delta)$ with a prior mean $\mu_\delta = \log(25000)$ and prior variance $\xi_\delta = 5$. This is a relatively diffuse prior distribution centered at the end of the sixth of twelve USEPA system size categories (Raucher et al. 1995). Also, as discussed in Section 3.5, this prior distribution was restricted to the interval $(8.11, 11.95)$ to prevent convergence problems.

The linear population models introduce only the parameters $\lambda^{(m)}$ and $\lambda^{(v)}$ which, in the most general case, take on different values for each source type and east/west designation. For all models, the subset of these parameters allowed to vary freely were modeled as iid $N(\mu_\lambda, \xi_\lambda)$. By the same reasoning used in the population threshold models, μ_λ was set to 0. The determination of the variance ξ_λ resulted from the following reasoning. Take $\log(25000)$ as a typical log population served. Also, assume that on the natural measurement scale of arsenic, an extreme impact of population would be to multiply the mean or the precision by a factor of 2 or 1/2. Forcing these impacts to correspond to 3 prior standard deviations away from zero implies a value of λ of approximately ± 0.06845. This implies a value of $\xi_\lambda = 0.023^2$.

For all parameters in all Bayesian models discussed in this report, the Metropolis algorithm (Metropolis et al. 1953) or Hasting's (1970) generalization of it was used to sample from the posterior distributions. Moreover, in all cases, a Gaussian proposal distribution on either the parameters or some transformation of them was employed. For scalar parameters, proposal standard deviations in these algorithms were chosen to achieve acceptance rates of $40\% - 50\%$ per parameter (Gelman, Roberts, and Gilks 1996). Proposal standard deviations for parameters updated in blocks were chosen to achieve block acceptance rates of $30\% - 35\%$, in accordance with the recommendations of Gelman, Carlin, Stern and Rubin (1995). Block updating was only used for the 50-dimensional vectors $\boldsymbol{\theta}$ of uncorrelated parameters discussed in Appendix B. These parameters were updated together, which reduced the number of evaluations of $\boldsymbol{L}(\rho)\boldsymbol{\theta}$ by a factor of 50 in Models 7 through 18. This greatly improved the efficiency of the algorithm without negatively impacting mixing.

Chains were typically burned in for 50,000 to 200,000 iterations. Periodically throughout the analyses, multiple chains were started from widely-dispersed starting values produced by sampling from the proposal distributions with standard deviations inflated by factors of five or more. Convergence was assessed graphically and numerically. The posterior distributions of the parameters, summarized by the marginal posterior means and standard deviations, were compared from the dispersed chains for consistency and were nearly identical in all cases. The same is true of graphical posterior predictive checks (e.g. Figure 3 *et al.*) of a large number of subclasses of the data.

Finally, for completeness, we briefly review some alternative models that we considered, but that did not figure in the analyses presented elsewhere in this report. We fit a version of the linear population model (Section 3.5) in which we allowed fifty different regressions, one for each state, but the regression coefficients had a correlation structure just like the the $\boldsymbol{\alpha}^{(m)}$ vector. This model fit the data about as well as other models and made predictions very similar to Model 9.

We also fit a version of Model 7 in which there were four different ρ parameters, one for each of $\boldsymbol{\alpha}^{(m)}$, $\boldsymbol{\alpha}^{(v)}$, $\boldsymbol{\beta}^{(m)}$, and $\boldsymbol{\beta}^{(v)}$. While model fit was essentially unaffected, this model predicted out of sample slightly worse than Model 7. The different ρ parameters did not all have similar posterior distributions, however. This suggests that the role of the ρ parameter in our model needs to be studied more closely in the future. We also fit a version of Model 7 in which the $\alpha^{(v)}_{X(i)}$ were all equal and the $\beta^{(v)}_{X(i)}$ were all equal. That is, there is a separate variance for ground water and surface water, but the variances are the same from state to state. This model predicted almost as well as Model 7 in some validations, but not all. It did a little better than Model 7 in the prediction of the USGS data (validation V_5). This may be due to the fact that the dependence of variance on location is slightly less pronounced for ground water than it is for surface water.

APPENDIX E: DETAILS FOR MODEL VALIDATION

We chose predictive density as our cross-validation criterion for two reasons. First, it contains a part that is similar to one of the traditional cross-validation criteria, namely sum of squared errors. Second, the ratio of two predictive densities for two different models gives an approximation to a Bayes factor for choosing between the two models.

The predictive density of test data $\boldsymbol{y_1}$ (with covariates $\boldsymbol{x_1}$) given training data $\boldsymbol{y_0}$ (with covariates $\boldsymbol{x_0}$ is

$$\int f(\boldsymbol{y_1}|\boldsymbol{x_1}, \theta) p(\theta|\boldsymbol{y_0}, \boldsymbol{x_0}) d\theta. \tag{A.7}$$

Since our MCMC analysis gives us a sample $\theta_1, \ldots, \theta_m$ from the posterior distribution of θ, we can approximate (A.7) by

$$\frac{1}{m} \sum_{j=1}^{m} f(\boldsymbol{y_1}|\boldsymbol{x_1}, \theta_j).$$

To see how this contains a part like sum of squared deviations, note that each term $f(\boldsymbol{y_1}|\boldsymbol{x_1}, \theta_j)$ is a product of one factor for each observation in the test data set. For each uncensored observation y_i, the factor is

$$\frac{1}{\sqrt{2\pi}\sigma_i(\theta_j)} \exp\left(-0.5 \frac{[y_i - \mu_i(\theta_j)]^2}{\sigma_i(\theta_j)^2}\right), \tag{A.8}$$

where $\mu_i(\theta_j)$ is that function of θ_j that specifies the mean of log-arsenic at site i and $\sigma_i(\theta_j)$ is the function that specifies the standard deviation. The negative of the logarithm of (A.8) is a weighted squared deviation plus a penalty for understating precision. The penalty for overstating precision is the fact that the squared deviation is divided by $\sigma_i(\theta_j)^2$. For each censored observation i, the factor is the probability that log-arsenic at site i is below the detection limit. This provides a penalty for predicting a large value when the observation was censored. So, large values of predictive density arise from models that assign high probabilities of being below the censoring limit for the censored observations and low values of the penalized weighted sum of squared deviations.

To see how (A.7) is like part of a Bayes factor, suppose that the training data $(\boldsymbol{y_0}, \boldsymbol{x_0})$ were known *a priori*, and that we used it to create the prior distribution with density $p(\theta|\boldsymbol{y_0}, \boldsymbol{x_0})$. Then (A.7) would be the marginal density of the observed data $(\boldsymbol{y_1}, \boldsymbol{x_1})$ under the model. The Bayes factor for comparing two models based on observed data is the ratio of the marginal densities of the observed data under the two models. To be more precise, suppose that k models are under consideration and that they have prior probabilities $q_1, \ldots, q_k$. Suppose that model ℓ says that the marginal density of the observed data is $g_\ell(\boldsymbol{y_1}, \boldsymbol{x_1})$ for $\ell = 1, \ldots, k$. Then the posterior probability of model ℓ is

$$\frac{q_\ell g_\ell(\boldsymbol{y_1}, \boldsymbol{x_1})}{\sum_{t=1}^{k} q_t g_t(\boldsymbol{y_1}, \boldsymbol{x_1})}. \tag{A.9}$$

The ratio of (A.9) with $\ell = \ell_1$ to (A.9) with $\ell = \ell_2$ is

$$\frac{q_{\ell_1}}{q_{\ell_2}} \frac{g_{\ell_1}(\boldsymbol{y_1}, \boldsymbol{x_1})}{g_{\ell_2}(\boldsymbol{y_1}, \boldsymbol{x_1})}.$$

The second factor in this expression is the Bayes factor for comparing model ℓ_1 to model ℓ_2. If the Bayes factor is large, the ratio of the posterior probability of model ℓ_1 to the posterior of model ℓ_2 will be higher than the ratio of the prior probabilities. If the Bayes factor is very large and if the ratio of prior probabilities is not very small, then model ℓ_1 will have much larger posterior probability than model ℓ_2. Using the ratio of cross-validation predictive densities as if it were a Bayes factor is like pretending that the training data were known *a priori* and the test data were the observations being modeled.

The interpretation of predictive density as part of a Bayes factor highlights the major problem that we encountered in calculating predictive densities. The formula (A.7) is the same sort of integral as that required to obtain the marginal density of all of the observed data, namely the integral of the conditional density of the data given the parameters times the density of the parameters. (This integral is often called the *normalizing constant* because it is the denominator that appears in Bayes' theorem.) In (A.7), the density of the parameters is the posterior given the training data. When finding the marginal density of the data, the density of the parameters is the prior density. The posterior density used in (A.7) should be more concentrated than the prior density, but it is still surprisingly flat, at least in the vicinity of where the likelihood (the other factor in the integral in (A.7)) is large. This has both positive and negative consequences. On the down side, it makes accurate calculation of predictive densities very difficult. Extremely large posterior samples are required in order to get approximations with small simulation standard deviation. It is similar to, but not quite as bad as, trying to approximate the marginal density of the data by averaging the likelihood using an importance sample from the prior distribution. Indeed, we are averaging the likelihood of the test data using a dependent importance sample from the posterior based on the training data.

As an example, we present the results of calculations based on the initial default prior discussed in Appendix D. Table 4 provides numbers of MCMC cycles used to calculate log predictive densities akin to those shown in Table 2 along with a measure of Monte Carlo standard error estimates. The standard errors were calculated as follows. Suppose that a

	Model							
	3B		7		9		10	
Validation	Size	s.e.	Size	s.e.	Size	s.e.	Size	s.e.
V_1	100,000	0.041	3,844,052	1.492	3,000,000	0.794	1,000,000	0.921
V_2	100,000	0.045	2,500,000	0.758	1,500,000	0.884	1,500,000	1.495
V_3	100,000	0.041	1,000,000	0.739	1,000,000	1.222	1,000,000	0.976
V_4	100,000	0.038	1,000,000	0.672	1,000,000	0.641	1,000,000	1.202
V_5	100,000	1.558	1,800,000	1.580	1,000,000	1.644	1,000,000	1.533
V_6	1,000,000	0.006	1,000,000	0.719	1,000,000	1.090	1,000,000	0.636
V_7	1,000,000	0.006	1,000,000	0.611	1,000,000	0.962	1,000,000	0.889

Table 4. Sizes of validation samples and estimated Monte Carlo standard errors.

log predictive density estimate was the average over m MCMC cycles. Denote the individual LPD values (one from each cycle) v_i for $i = 1, \ldots, m$. We first find $v^M = \max_i v_i$, the largest value, and then compute

$$v^* = \frac{1}{m}\sum_{i=1}^{m}\exp(v_i - v^M).$$

The overall log predictive density estimate (like those presented in Table 2) is $\log(v^*) + v^M$. Next, we compute

$$v^+ = \frac{1}{m}\sum_{i=1}^{m}[\exp(v_i - v^M) - v^*]^2,$$

as an estimate of the variance of each individual $\exp(v_i - v^M)$ values. We use the delta method to approximate the variance of the logarithm of the average as $v^+/[wv^{*2}]$, were $w = 0.37m$. The factor 0.37 in w is intended to adjust for possible serial correlation between the individual predictive density values. If the lag-one serial correlation is r, and if the sequence of $\exp(v_i - v^M)$ values were a stationary autoregressive process of order 1, then the variance of the average would be less than $\sigma^2/[(1-r)m]$, where σ^2 is the variance of each term. Although there was a great deal of serial correlation between the v_i values, there was very little among the $\exp(v_i - v^M)$ values, except in some of the Model 3B validations. In those cases, the maximum serial correlation was 0.63, so we assumed that $r = 0.63$ in all cases, in order to get conservative standard errors.

The simulation standard errors for Model 3B in the first four validations are all reasonable and they suggest that we have a good approximation to the log-predictive density. The same cannot be said of the other standard errors in Table 4. Several values are very close to $1/\sqrt{0.37} = 1.644$ or half of that number because the averages to which they correspond are driven by one or two values that are much larger than the others. This makes the delta method inappropriate for approximating the simulation variance of a function of the average. Indeed it draws into question the suitability of the variance as a measure of the uncertainty that remains after the simulation. As an alternative, we used the Bayesian bootstrap to assess the degree to which Model 7 has larger predictive density than the simpler models. In particular, we tried to put a lower bound q on the probability that the LPD for Model 7 is at least a large amount c plus the LPD h for the best of the simpler models. That is, we will try to make a claim that the probability is at least q (very close to 1) that the predictive density of Model 7 is $\exp(c)$ times as large as the predictive density of each of the simpler models, where $\exp(c)$ is a very large number. Since Bayesian bootstrap probabilities depend only on the sample sizes and not on the observed values, it will prove convenient to first choose a large q, and then calculate the appropriate value of c for each validation. We used $q = 0.99995$, (1 to four decimal places.)

Although similar in nature to the bootstrap, the Bayesian bootstrap is more flexible in terms of the types of analysis one can perform, and it is more conservative in estimating how much better Model 7 is than the other models. The Bayesian bootstrap uses a setup similar to that of the bootstrap. As before, let $v_1, \ldots, v_m$ be the observed values of log-predictive density associated with the m simulated values of θ from the MCMC analysis. Let V stand for a generic random variable having the same distribution as the log-predictive density from one MCMC cycle. For convenience, suppose that the v_i are ordered from smallest to largest, $v_1 \leq v_2 \leq \cdots \leq v_m$. Both the bootstrap and the Bayesian bootstrap approximate the distribution P of V by a discrete distribution having support equal to the values $\{v_1, \ldots, v_m\}$. Let p_i stand for $P(V = v_i)$. The bootstrap assumes that $p_i = 1/m$ for all i, while the Bayesian bootstrap assumes that the vector $\boldsymbol{p} = (p_1, \ldots, p_m)$ has a Dirichlet distribution with parameter $(1, \ldots, 1)$. Both bootstraps are based on the assumption that the observed values $v_1, \ldots, v_m$ came from a random sample. Although this was not true, the θ_i values are supposed to be representative of the posterior distribution, hence if we ignore the order in which they arrived, they should look like a random sample.

The Bayesian bootstrap allows us to think of (A.7) as a random variable Z whose distribution is determined by the distribution of $\boldsymbol{p}$ using the fact that

$$Z = E(\exp(V)) = \sum_{i=1}^{m} p_i \exp(v_i).$$

We can put a lower bound on $\Pr(\log(Z) > c + h)$ by noting that $Z \geq \exp(v_{m-k}) \sum_{i=m-k}^{m} p_i$ for $k = 0, 1, \ldots$. It follows that

$$\Pr(\log(Z) > c + h) \geq \Pr\left(\sum_{i=m-k}^{m} p_i > \exp(c + h - v_{m-k}) \right). \tag{A.10}$$

Suppose that we choose $c = c_k$ where $\exp(c_k + h - v_{m-k}) = 1/m$ for some k. Since $\boldsymbol{p}$ has a Dirichlet distribution, $\sum_{i=m-k}^{m} p_i$ has beta distribution with parameters $k + 1$ and $m - k$. The probability that such a beta random variable is greater than $1/m$ is the probability that a binomial random variable with parameters m and $1/m$ is at most k. But a binomial with parameters m and $1/m$ very nearly has a Poisson distribution with parameter 1. So, the probability on the right-hand side of (A.10) is essentially the probability $Poi(k)$ that a Poisson random variable with parameter 1 is at most k. For $k \geq 7$, $Poi(k) > q = 0.99995$. So, if $k \geq 7$ and $c = c_k = v_{m-k} - h - \log(m)$, the probability is at least q that the predictive density for Model 7 is at least $\exp(c_k)$ as large as the predictive density of the best of the simpler models. Table 5 contains the values of $\exp(c_7)$ for each of the seven validations from the preliminary analysis, together with the corresponding numbers for an analysis of the sampling distribution of the LPD, which we discuss shortly. The fourth column of Table 5 indicates that, for each validation, the probability is at least $q = 0.99995$ that the predictive density for Model 7 is at least tens of thousands of times as large as the largest observed predictive density for any of the simpler models.

Validation	h	v_{m-7}	$\exp(c_7)$	v_{m-14}	$\exp(c_{14})$
V_1	-333.4	-306.8	9.12×10^4	-308.7	1.43×10^4
V_2	-320.8	-290.3	7.24×10^6	-290.7	5.00×10^6
V_3	-336.0	-307.8	1.68×10^6	-310.7	9.83×10^4
V_4	-317.1	-288.6	2.43×10^6	-289.0	1.56×10^6
V_5	-2532.7	-2374.8	2.09×10^{62}	-2381.5	2.66×10^{59}
V_6	-1541.1	-1395.9	1.10×10^{57}	-1396.6	5.80×10^{56}
V_7	-1523.7	-1471.4	1.22×10^{41}	-1482.6	9.87×10^{39}

Table 5. Bayesian bootstrap analyses of validations V_1–V_7 with $k = 7$ and $k = 14$. The value of h is taken as the largest observed LPD from Model 3B for V_1, V_2, V_3, and V_5 and the LPD of Model 4 for V_4, V_6, and V_7.

More in line with a traditional bootstrap analysis, we also let $\overline{Y}$ be the average of a random sample of m values with the distribution of $\exp(V)$. We can put a lower bound on the probability that $\log(\overline{Y}) > c + h$ by noting that $\overline{Y} \geq \exp(v_{m-k})/m$, where v_{m-k} is the largest v_i that appears in the Bayesian bootstrap sample. So, if we let $\exp(c_k + h - v_{m-k}) = 1/m$ again, we have that

$$\begin{aligned}
&\Pr(\log(\overline{Y}) > c_k + h) \\
&\quad \geq \quad \Pr(\text{at least one of } v_m, \ldots, v_{m-k} \text{ appears in the Bayesian bootstrap sample}).
\end{aligned}$$

The probability that at least one of $k + 1$ specific values appears in a Bayesian bootstrap sample is known to equal

$$1 - \prod_{i=1}^{m} \left(1 - \frac{k+1}{m+i}\right) \approx 1 - 2^{-k-1}. \tag{A.11}$$

If $k \geq 14$, then $1 - 2^{-k-1} > q = 0.99995$. From the last column of Table 5, we see that, for each of the validations, the probability is at least q that the observed LPD $\overline{Y}$ is at least ten thousand times larger than the largest observed LPD for any of the simpler models. Despite the large variance of LPD for Model 7, we believe that the Bayesian bootstrap analyses demonstrate that Model 7 is very likely far superior to the simpler models in terms of predictive density in all of our validation settings. For the interested reader, a bootstrap analysis of $\overline{Y}$ would have replaced the approximation in (A.11) with $1-(1-[k+1]/m)^m \approx 1-\exp(-k-1)$, which is somewhat larger than $1-2^{-k-1}$, so that $k=9$ would have sufficed to achieve $1-\exp(-k-1)>q$. A bootstrap analysis of Z is not available.

The positive effect of the flatness of the posterior distribution is that it gives the model the flexibility to predict many different data sets well. (This is as opposed to a model with a very peaked posterior in the wrong place.) As an illustration of this, we split the USGS data into five subsets two different ways. One way (called *split 1*) was to take every fifth observation from the data file for each subset. Since the data were arranged roughly by geography, this produced five very similar subsets. The other way of dividing (called *split 2*) was to take the first 20% of the data for the first subset, the second 20% for the second subset, etc. This produced five very different subsets. We then fit Model 7 using the NAOS data set and produced a sample of 500,000 vectors from the posterior distribution and predicted each of the ten subsets of USGS. We looked at which θ vectors did a good job of predicting each subset and checked whether those same θ vectors did a good job of predicting the other four subsets in the same split. For this purpose we defined "did a good job of predicting a subset" to mean that the log-predictive density was within log(0.00001) of the largest of the 500,000 log-predictive densities for that subset. Table 6 gives the largest log-predictive density for each of the ten subsets together with the overall LPD and the number of vectors that did a good job of predicting that subset. First, it is easy to see how much more closely the subsets in split 1 were predicted than were the subsets in split

Split 1	Subset 1	Subset 2	Subset 3	Subset 4	Subset 5
Maximum	−469.5	−468.4	−472.8	−467.4	−472.4
LPD	−481.3	−481.5	−485.8	−480.4	−485.0
Number	343	27	74	144	161
Split 2	Subset 6	Subset 7	Subset 8	Subset 9	Subset 10
Maximum	−362.9	−521.5	−397.1	−420.7	−512.4
LPD	−376.0	−533.5	−410.2	−433.6	−524.8
Number	12	376	18	35	100

Table 6. Maximum log-predictive densities, LPDs, and numbers of vectors that did a good job of predicting for each subset of USGS.

2. Note also that less than one-tenth of one-percent of the θ vectors did a good job of predicting any one subset. However, for split 1, we found that the intersection of the sets of vectors that did good jobs of predicting pairs of subsets contained at least 5% of each of the two good predicting subsets. This means that, for the split into similar subsets, those few vectors that predict one subset well were more likely to predict at least one other subset well than was the typical vector. Indeed there were 2 vectors that predicted all five subsets well, which is surprising since there were only 27 vectors that predicted subset 2 well. For split 2, where the subsets are quite different, not a single vector did a good job of predicting even two different subsets.

APPENDIX F: EVIDENCE FOR COVARIATION OF ARSENIC WITH SYSTEM SIZE

Here we describe our investigation of the value of system size as a covariate. The four frames of Figure 12 are an example of the kinds of differences we observed between models which did and did not include population. In particular, we initially divided the NAOS data into four major classes, splitting by source water type and location (east or west). We then subdivided each of these four classes into collections of “large” and “small” systems, where

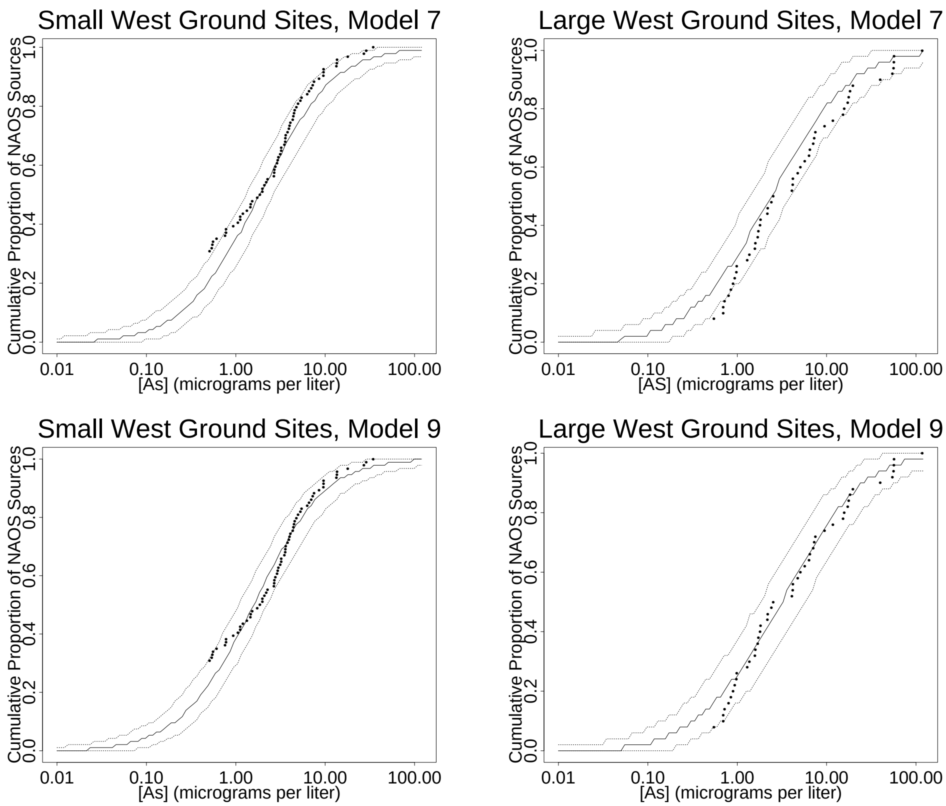


Figure 12. Empirical and posterior predictive CDFs for various classes of NAOS sites using a model which ignores population (Model 7) and one which uses it (Model 9). Both models were fit only to the NAOS data. The cutoff between "Large" and "Small" was 25,000 people; observations with missing populations not plotted. The dots are the uncensored NAOS data values.

the determination between the two size classes was based on whether the population served by each of the systems was greater than or equal to 25000 people. As noted previously, this corresponds to the division between the sixth and seventh of twelve EPA size classifications (Raucher et al. 1995). We then fit both Models 7 and 9 to the NAOS data (with the default prior discussed in Appendix D) and compared how well the models predicted the observed differences between large and small systems. In Figure 12 we present the results for the division of NAOS western ground systems. In Model 7, where we do not include system size as a covariate, there seems to be a slight underestimation of the mean arsenic level for the large systems and a corresponding overestimation for small systems. This is somewhat mitigated with Model 9, where the differences between large and small sites are captured more effectively.

We did this sort of analysis for all four divisions of the NAOS data, and the problems were most severe with western ground sites. The differences between large and small surface systems were much less distinctive, as were the differences between large and small systems in the east. This led us to explicitly introduce dependence on system size which was allowed to vary as both a function of location and source water type. In addition, although the results presented in this appendix are based on the initial prior distribution, virtually identical results are obtained when the elicited prior is used.

It was of interest to look more carefully at the validation results to see if the differences between Model 7 and Model 9 might be due to a few observations being fit much worse by one model, but the rest of the data being fit about the same by both. We reran parts of validation V_2 and found the vectors θ that had the highest log-predictive densities for Models 7 and 9 (and the MLE for Model 3). We compared the log of the predictive density for every observation from Model 7 to the same quantity for Model 9. We made this comparison for

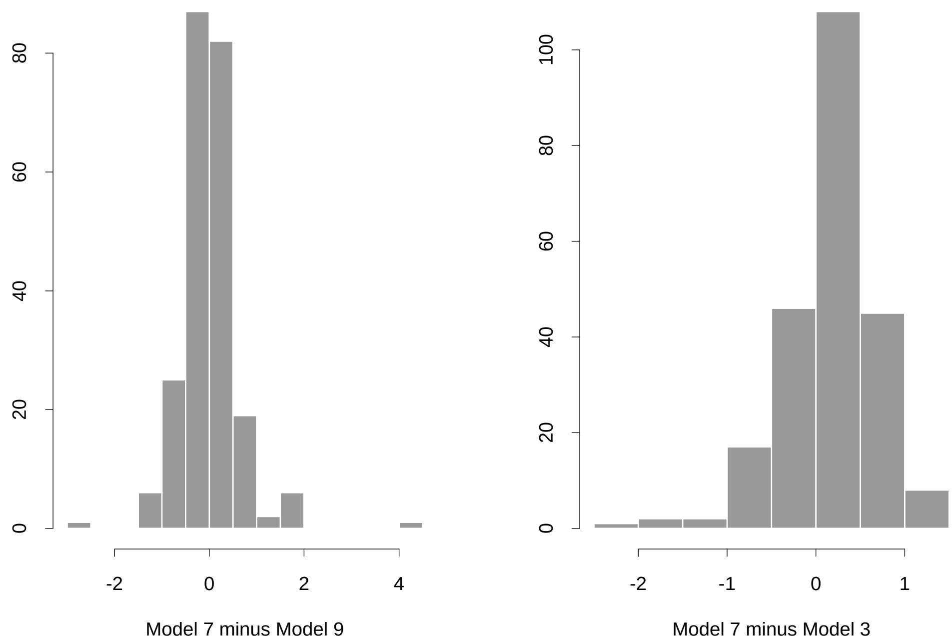


Figure 13. Histograms of differences in log-predictive densities for individual observations using the best predicting parameter vectors from Models 3, 7, and 9.

several pairs of top vectors from each model. We also made the same comparison between Models 7 and 3. There was a group of about 15 observations that neither Model 7 nor Model 9 was able to predict very well. Typically, Model 9 would predict one of these observations much more poorly than did Model 7, and this would account for almost half of the difference in log-predictive density. Indeed, we typically found that for most observations both models did equally well at predicting. Figure 13 contains histograms of the differences between the log-predictive densities from pairs of models for the 229 observations in $NAOS_1$ based on the fit to $NAOS_2$. Models 7 and 9 appear nearly identical except for one observation, while Model 7 is, on balance, somewhat better than Model 3.

Although including system size appears to lead to a better fit to the observed data as seen

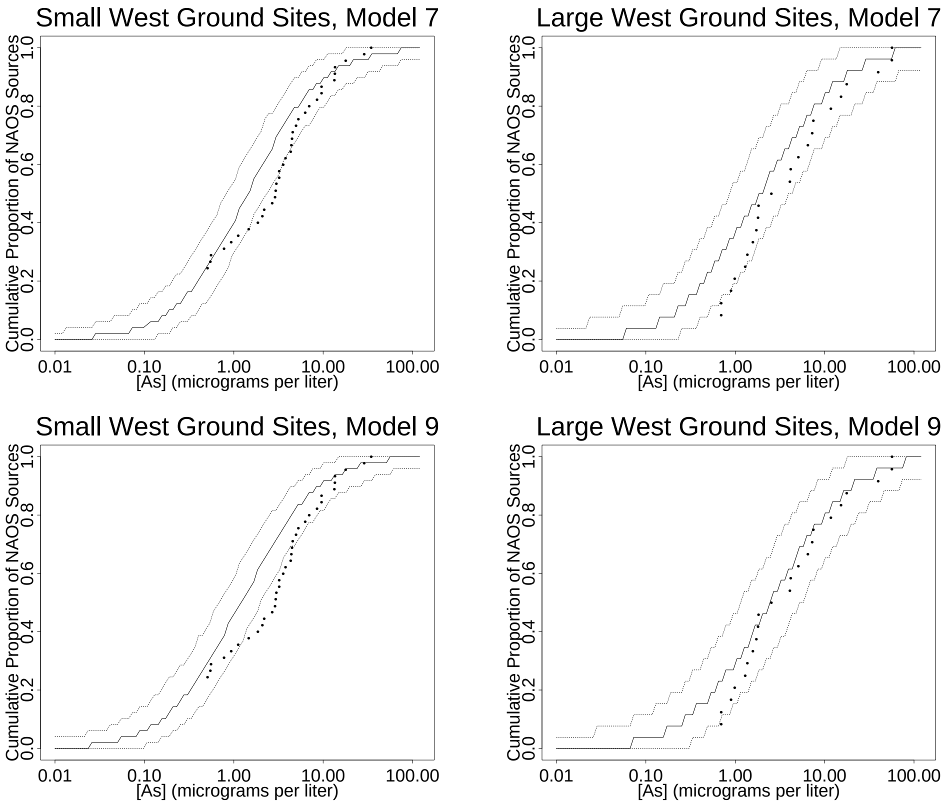


Figure 14. Empirical and posterior predictive CDFs for various classes of $NAOS_1$ sites using a model which ignores population (Model 7) and one which uses it (Model 9). Both models were fit to only the $NAOS_2$ data. The cutoff between "Large" and "Small" was 25,000 people; observations with missing populations not plotted. The dots are the uncensored NAOS data values.

in Figure 12, Models 8 through 10 do not fare as well as Model 7 in terms of out-of-sample prediction, as is evident from Table 2. This is not surprising for validation V_5 because the USGS data do not have system size information. However, based on the information in Figure 12, one would expect at least Model 9, if not Models 8 or 10, to do a better job in validations V_1–V_4. This expectation, by and large, is not realized. The case of validation V_2 is illustrated in Figure 14 , where no strong predictive advantage is clear for Model 9.

APPENDIX G: EXAMINATION OF POSTERIOR DISTRIBUTIONS

We examined the posterior distributions of some of the parameters in detail. First, we looked at ρ to see what degree of correlation between the states was being fit. Figure 15 has a plot of correlation as a function of distance based on the posterior distribution of ρ (based

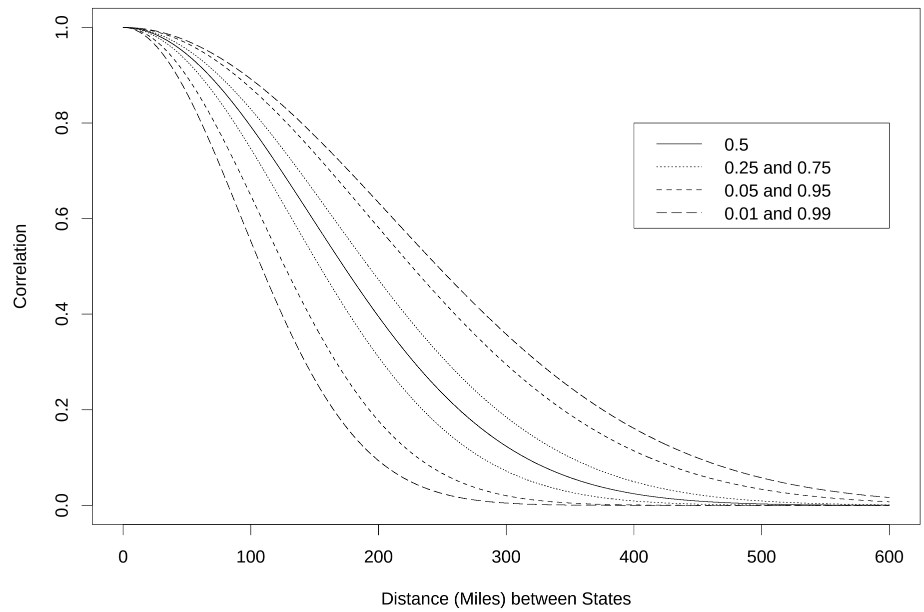


Figure 15. Plot of 0.01, 0.05, 0.25, .05, 0.75, 0.95, and 0.99 quantiles of correlation as a function of distance. The distribution of ρ comes from Model 7 fit to both NAOS and USGS.

on the initial prior distribution discussed in Appendix D). The plot shows the uncertainty that remains surrounding the correlation. The correlation has essentially dropped to 0 by 600 miles, roughly the distance from Arkansas to Nebraska, Nevada to Washington, or Maine to Maryland.

Figure 16 has boxplots of 10,000 samples (every 100th from an MCMC run of 1,000,000) from the posterior distributions of $\log(\rho)$ after fitting Model 7 to the NAOS data and after

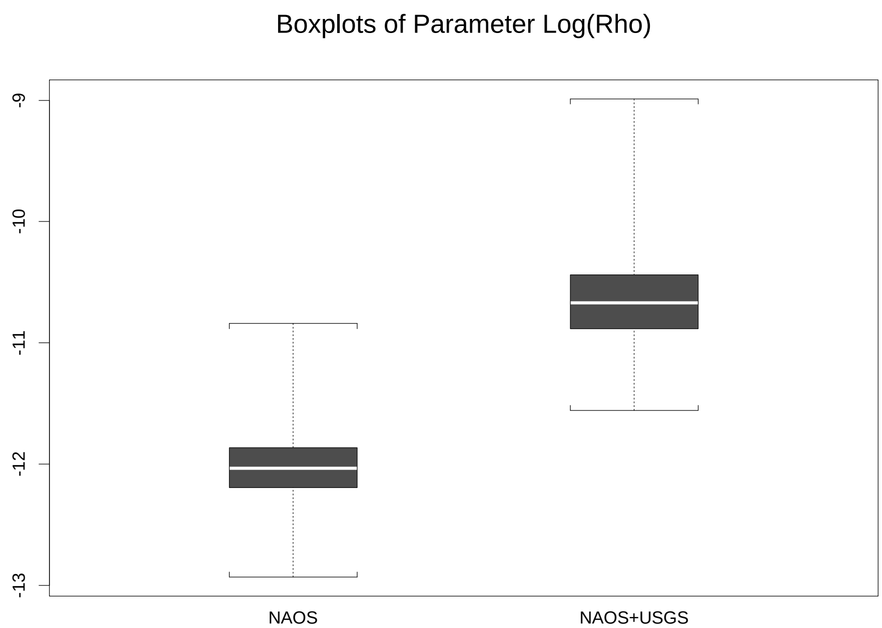


Figure 16. Boxplots of 10,000 posterior samples of $\log(\rho)$ *from Model 7 fit to only NAOS data and fit to both NAOS and USGS data. The whiskers extend to the extreme observations. There were no apparent outliers in either sample.*

fitting Model 7 to both NAOS and USGS based on the initial prior discussed in Appendix D. We see that the distribution of $\log(\rho)$ changes dramatically when the USGS data are used. This could be more evidence that the data in the two data sets should be modeled differently. (See Section 4.3 for more discussion on this point.) It could also be evidence that ground and surface waters should have separate values of ρ. (Recall that USGS is all ground water, and it is a much larger data set than NAOS.) Another possibility is that the difference reflects the more precise information that we have about at least some functions of the parameters that appear in (1) and (2). We only learn about ρ indirectly through what we learn about the parameters in (1) and (2). In fact ρ only measures a certain aspect of our uncertainty about the parameters in (1) and (2). As we learn more about those parameters our uncertainty changes, and it seems reasonable that ρ should change to reflect the change in uncertainty.

Nevertheless, we will examine the idea of using two values of ρ when we begin to model multiple contaminants. Using separate values of ρ for ground and surface waters will require us to abandon the ANOVA structure of (1) because the parameters $\alpha_{X(i)}^{(m)}$ are common to both ground and surface observations. It will be a particular challenge to construct a covariance structure for the ground and surface parameters that satisfies all three of the following:

- the marginal covariances of the parameters for the different source types are different,
- there is positive correlation between the different source types within the same state, and
- the joint covariance matrix is positive definite.

One possibility to guarantee positive definiteness would be to let the covariance matrix between the two source types be a function of the two marginal covariance matrices. For example, if C_0 and C_1 are the marginal covariance matrices for surface and ground water, we could let the covariance between surface and ground water be $(v_0 C_0^{-1} + v_1 C_1^{-1})^{-1}$, where $v_0, v_1 \geq 1$ are scalars.

Finally, the posterior distributions of the $log(\tau)$ parameters were all very nearly normal with different means and standard deviations. The four posterior distributions are quite different from the common prior which was standard normal. This indicates that the likelihood for these parameters was not flat and that we actually learned something about them from the data. Posterior means and standard deviations of all model parameters for Model 7 fit to the combined NAOS and USGS data with the informed prior are given in Tables 7 and 8.

	$\alpha^{(m)}$		$\beta^{(m)}$		$\alpha^{(v)}$		$\beta^{(v)}$	
State	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
AK	0.84	0.48	0.19	0.35	-0.79	0.57	-0.02	0.48
AL	-1.22	0.54	-0.30	0.32	-1.42	0.47	0.41	0.37
AR	-0.06	0.27	-0.30	0.24	0.35	0.49	0.61	0.41
AZ	1.55	0.38	0.15	0.36	-0.44	0.49	0.05	0.48
CA	0.90	0.09	0.19	0.09	0.08	0.19	-0.69	0.18
CO	-0.56	0.22	0.02	0.20	-0.19	0.39	-0.12	0.34
CT	-0.55	0.29	-0.10	0.24	0.33	0.47	0.17	0.40
DE	-0.54	0.26	-0.09	0.19	0.13	0.42	-0.13	0.33
FL	-0.34	0.42	-0.14	0.38	-0.13	0.60	0.30	0.52
GA	-0.54	0.39	0.04	0.29	-0.62	0.57	0.17	0.43
HI	0.72	0.97	0.01	0.40	-0.96	1.03	-0.02	0.57
IA	0.25	0.22	-0.11	0.21	0.34	0.35	-0.02	0.34
ID	0.69	0.37	0.00	0.36	-0.66	0.48	0.13	0.47
IL	0.66	0.20	-0.16	0.17	-0.21	0.28	-0.65	0.25
IN	0.50	0.26	-0.20	0.20	-0.4	0.35	-0.97	0.29
KS	0.75	0.33	-0.19	0.27	-0.23	0.51	-0.12	0.45
KY	-0.02	0.34	-0.37	0.24	-0.30	0.40	-0.43	0.32
LA	-0.15	0.34	-0.56	0.32	-0.55	0.43	0.23	0.40
MA	-0.82	0.34	-0.08	0.29	-0.12	0.51	0.14	0.44
MD	-0.45	0.18	0.07	0.14	0.01	0.31	-0.26	0.26
ME	-0.68	0.73	0.14	0.39	-1.49	0.78	-0.14	0.58
MI	0.17	0.33	0.01	0.25	-0.33	0.49	-0.87	0.43
MN	0.68	0.35	-0.18	0.27	0.18	0.61	-0.22	0.48
MO	0.26	0.25	-0.28	0.22	0.08	0.42	-0.07	0.37
MS	-1.20	0.51	-0.47	0.32	-0.59	0.59	0.63	0.45
MT	0.76	0.57	-0.24	0.38	-1.39	0.52	0.12	0.43
NC	-1.15	0.33	0.10	0.24	0.21	0.67	-0.10	0.43
ND	1.33	0.39	-0.22	0.27	0.15	0.70	-0.17	0.48
NE	1.45	0.27	-0.03	0.25	0.35	0.50	0.03	0.45
NH	-1.04	0.40	-0.05	0.31	-0.61	0.54	0.06	0.46
NJ	-0.34	0.16	-0.09	0.15	0.50	0.31	0.00	0.30
NM	0.33	0.45	-0.04	0.35	-1.04	0.56	-0.14	0.51
NV	1.37	0.29	0.16	0.27	-0.35	0.43	-0.41	0.41
NY	-0.64	0.30	-0.10	0.22	-0.04	0.52	0.00	0.37
OH	0.30	0.19	-0.13	0.17	-0.14	0.28	-0.88	0.26
OK	1.44	0.32	-0.17	0.28	-0.88	0.40	-0.28	0.38
OR	0.47	0.43	0.30	0.37	-0.53	0.58	0.37	0.50
PA	-0.54	0.15	0.06	0.14	-0.18	0.26	-0.38	0.24
RI	-0.55	0.39	-0.08	0.29	0.05	0.57	0.18	0.45
SC	-0.92	0.48	0.12	0.30	0.25	0.83	-0.01	0.50
SD	1.16	0.24	-0.21	0.22	0.19	0.42	-0.11	0.39
TN	-1.39	0.29	-0.46	0.24	-0.59	0.39	-0.09	0.33
TX	0.72	0.13	-0.85	0.13	1.35	0.34	0.17	0.33
UT	0.22	0.29	-0.08	0.26	-0.50	0.41	-0.21	0.36
VA	-0.34	0.20	0.13	0.18	0.24	0.36	-0.15	0.31
VT	-0.93	0.45	-0.06	0.31	-0.67	0.61	0.03	0.46
WA	0.14	0.37	0.21	0.36	-0.75	0.51	0.30	0.49
WI	0.23	0.15	-0.03	0.15	1.17	0.50	-0.23	0.47
WV	0.56	0.23	0.05	0.21	0.75	0.37	-0.19	0.33
WY	0.16	0.47	-0.17	0.37	-0.34	0.65	0.12	0.49

Table 7. Posterior means and standard deviations of likelihood parameters from fit of Model 7 to the combined NAOS and USGS data. Estimates based on 2,000,000 samples from the posterior distribution.

$\log(\rho)$		$\log(\tau_{\alpha^{(m)}})$		$\log(\tau_{\beta^{(m)}})$		$\log(\tau_{\alpha^{(v)}})$		$\log(\tau_{\beta^{(v)}})$	
Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
-11.16	0.20	0.17	0.29	1.89	0.43	-0.06	0.30	1.20	0.45

Table 8. Posterior means and standard deviations of hyperparameters from fit of Model 7 to the combined NAOS and USGS data. Estimates based on 2,000,000 samples from the posterior distribution.

REFERENCES

Berthouex, P. and Hau, I. (1991), "A simple rule for judging compliance using highly censored samples," *Research Journal of the Water Pollution Control Federation*, 63(6), 880–886.

Cadmus Group Inc. (1997), "The SafeWater Suite,".

Chen, C., Chen, C., Wu, M., and Kuo, T. (1992), "Cancer potential in liver, lung, bladder, and kidney due to ingested inorganic arsenic in drinking water," *British Journal of Cancer*, 66(5), 888–892.

Conlon, E. and Waller, L. (1998), "Flexible neighborhood structures in hierarchical models for disease mapping," Technical Report 98-018, Division of Biostatistics, University of Minnesota.

Cook, D. and Pocock, S. (1983), "Multiple regression in geographic mortality studies, with the allowance for spatially correlated errors," *Biometrics*, 39, 361–371.

Datta, G., Lahiri, P., Maiti, T., and Lu, K. (1999), "Hierarchical Bayes estimation of unemployment rates for the states of the US," *Journal of the American Statistical Association*, 94(448), 1074–1082.

Feller, W. (1971), *An Introduction to Probability Theory and Its Applications*, volume 2, New York: Wiley.

Frey, M. and Edwards, M. (1997), "Surveying arsenic occurrence," *Journal of the American Water Works Association*, 89(3), 105–117.

Frey, M., Edwards, M., and Amy, G. (1997), "National compliance assessment and costs for the regulation of arsenic in drinking water,", Water Industry Technical Action Fund Report, prepared in association with D. M. Owen and Z. K. Chowdhury.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (1995), *Bayesian Data Analysis*, London: Chapman & Hall.

Gelman, A., Meng, X.-L., and Stern, H. (1996), "Posterior predictive assessment of model fitness via realized discrepancies (with discussion)," *Statistica Sinica*, 6, 733–807.

Gelman, A., Roberts, G., and Gilks, W. (1996), "Efficient Metropolis jumping rules," in *Bayesian Statistics 5*, eds. J. Bernardo, J. Berger, A. Dawid, and A. Smith, Oxford: Oxford University Press, pp. 599–607.

Geman, S. and Geman, D. (1984), "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721–741.

Gibbons, R. (1994), *Statistical Methods for Groundwater Monitoring*, New York: John Wiley & Sons, Inc.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1996), *Markov Chain Monte Carlo in Practice*, London: Chapman and Hall.

Grady, S. and Weaver, M. (1989), "Evaluation of groundwater quality in relation to land use for stratified–drift aquifers in Connecticut," in *Regional Characterization of Water Quality (Proceedings of the Baltimore Symposium, May 1989) International Association of Hydrologic Sciences Publication No. 182*, Wallingford, Oxfordshire, UK: IAHS Press.

Graybill, F. (1976), *Theory and Application of the Linear Model*, Pacific Grove, CA: Wadsworth.

Hastings, W. (1970), "Monte Carlo sampling methods using Markov chains and their applications," *Biometrika*, 57, 97–109.

Hopenhayn-Rich, C., Biggs, M. L., Fuchs, A., Bergoglio, R., Tello, E., Nicolli, H., and Smith, A. (1996), "Bladder cancer mortality associated with arsenic in drinking water in Argentina," *Epidemiology*, 7(2), 117–124.

Kolpin, D., Barbash, J., and Gilliom, R. (1998), "Occurrence of pesticides in shallow groundwater of the United States: Initial results from the National Water-Quality Assessment Program," *Environmental Science & Technology*, 32, 558–566.

Lempers, F. (1971), *Posterior Probabilities of Alternative Linear Models*, Rotterdam: University Press.

MathSoft (1991), "S-plus programming environment,", MathSoft, Inc., Seattle, WA.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. (1953), "Equation of state calculations by fast computing machines," *Journal of Chemical Physics*, 21, 1087–1092.

National Research Council (1999), *Arsenic in Drinking Water*, Washington, D.C.: National Academy Press.

Odom, R., Regli, S., Messner, M., Cromwell, J., and Javdan, M. (1999), "Benefit–cost analysis of the Stage 1 D/DBP Rule," *Journal of the American Water Works Association*, 91(4), 137–147.

O'Hagan, A. (1995), "Fractional Bayes factors for model comparison (with discussion)," *Journal of the Royal Statistical Society, Ser. B*, 57, 99–138.

Ott, W. (1995), *Environmental Statistics and Data Analysis*, New York: CRC Press, Inc., Lewis Publishers.

Page, G. (1981), "Comparison of groundwater and surface water for patterns and levels of contamination by toxic substances," *Environmental Science & Technology*, 15, 1475–1481.

Pontius, F., Brown, K., and Chen, C. (1994), "Health implications of arsenic in drinking water," *Journal of the American Water Works Association*, 86(9), 52–63.

Raucher, R., Castillo, E., Dixon, A., Breffle, W., Waldman, D., and Drago, J. (1995), *Estimating the Cost of Compliance: A User's Guide*: AWWA Research Foundation.

Regli, S., Odom, R., Cromwell, J., Lustic, M., and Blank, V. (1999), "Benefits and costs of the IESWTR," *Journal of the American Water Works Association*, 91(4), 148–158.

Sinnott, R. (1984), "Virtues of the haversine," *Sky and Telescope*, 68(2), 159.

Small, M. (1989), "Regional distributions of water quality derived as averages of spatial random processes," in *Regional Characterization of Water Quality (Proceedings of the Baltimore Symposium, May 1989) International Association of Hydrologic Sciences Publication No. 182*, Wallingford, Oxfordshire, UK: IAHS Press.

Tanner, M. and Wong, W. (1987), "The calculation of posterior distributions by data augmentation (with discussion)," *Journal of the American Statistical Association*, 82(398), 528–550.

Thisted, R. (1988), *Elements of Statistical Computing: Numerical Computation*, London: Chapman & Hall.

User's Manual for the National Water Information System of the U.S. Geological Survey: Chapter 1, Ground-water Site-inventory System (1997), Reston, VA: United States Geological Survey, Open-File Report 97-633, S.B. Mathey, editor.

USEPA (1994), "Distributional assessments of state data on arsenic in drinking water and evaluation of statistical estimation methods,", Prepared by Science Applications International Corporation.

——— (1998), "Setting Standards for Safe Drinking Water,", Available at `http://www.epa.gov/safewater/standard/setting.html`.

——— (2000a), "National Primary Drinking Water Regulations; Arsenic and Clarifications to Compliance and New Source Contaminants Monitoring; Proposed Rule," *Federal Register*, 65(121), 38887–38983.

——— (2000b), "Safe Drinking Water Information System,", Part of Envirofacts Data Warehouse, `http://www.epa.gov/envirofw/index_java.html`.

Wade Miller Associates, Inc. (1992), "Occurrence assessment for arsenic in public drinking water supplies,", Prepared for USEPA under Contract 68-C0-0069.