March 23, 2000

Hierarchical Modeling of Arsenic Concentrations at Entry Points in US Public Drinking Water Supplies (ADA report)

Yangang Zhang

A Bayesian hierarchical model is built to describe arsenic concentrations in treated water from sources of public drinking water systems. The model allows us to decompose the total variability in arsenic concentration into three components: between-system, between-source (within system) and within-source variabilities. Predictions about what percentage of a state's systems and sources affected by the various proposed maximum contaminant level (MCL) regulations are simulated based on the posterior predictive distribution. We investigate the potential impact of the between-source variability on this percentage by comparing predictions based on the full model and based on the reduced model which eliminates the between-source variability. Other issues addressed in the modeling are the possibility that the arsenic concentration in source water is changing over time and the possibility that changes in measurement methods and their detection limits cause a change in the precision and accuracy of the measurement methods. The analysis is conducted based on data from four states: California, Illinois, New Mexico and Utah.

KEY WORDS: Bayesian analysis; censored data; hierarchical linear model; Markov chain Monte Carlo; variance component; water quality.

1. INTRODUCTION

Maximum contaminant levels (MCLs) are set by the US Environmental Protection Agency (EPA) to help to limit the presence of dangerous chemicals and water-borne bacteria and viruses in public drinking water. The 1996 Safe Drinking Water Act (SDWA) requires EPA to re-evaluate the MCL for arsenic in drinking water. Exposure to high levels of arsenic has been correlated with the risk of bladder, lung and skin cancer. Data also suggest that arsenic may affect the vascular system in humans and may be associated with the development of diabetes (Hopenhayn-Rich et al. 1997; Pontius et al. 1994). The National Research Council Subcommittee on Arsenic in Drinking Water has concluded that the current MCL (50 $\mu g/L$)

Yangang Zhang is Doctoral candidate, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213, Tel: 412-268-2717. The report is advised by professor Mark Schervish, Department of Statistics and professor Mitchell J. Small, Department of Engineering and Public Policy and Civil and Environmental Engineering, Carnegie Mellon University, Pittsburgh, PA 15213. This paper has not undergone UA 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.

is not low enough to adequately reduce exposure to the carcinogen (Christen 1999). The revised MCL may be between 2-20 $\mu g/L$.

Lowering the MCL will require drinking water systems to invest in treatment technologies sufficient to comply with the new standard. In order to estimate compliance costs, an estimate of the distribution of arsenic concentrations in drinking water systems is needed . As a consequence of the nature of the currently available national data, preliminary national occurrence estimates have not accounted for the distribution of arsenic concentrations between different sources within a system. Most of the data from a water industry national survey consist of a single measurement from each system (Frey and Edwards 1997). The potential drawback of such analysis is that it may over- or underestimate the mitigation costs for systems obtaining water from more than one source, because it assumes that the arsenic levels in all of the sources within a system have the same concentration and would exceed the limits and require treatment at the same time.

To address the limitations of analysis based on single measurements for a system, we analyze the compliance monitoring data from four states that include multiple measurements from multiple sources in individual systems. The differences in arsenic concentrations among different sources within systems (between-source variability) and the sample to sample difference within sources (within-source variability) are estimated through use of a hierarchical model. The model is then used to estimate the potential mis-estimation of the mitigation costs that can result due to omission of the between-source and within-source variability. The arsenic samples were collected over the past 20 years and were analyzed by different measurement methods. Such a dataset suggests that we investigate the possibility that the mean arsenic levels in sources change with time and the possibility of accuracy and precision differences among different measurement methods. A Bayesian approach is used to fit the model.

In section 2, we describe the source and nature of the data used for the analysis. In section 3, we describe our hierarchical model for estimating the arsenic concentration for the sources within systems. In addition, we identify the questions and issues which are addressed by our model. We also provide the methods used for model checking and prediction. The results of the model, including decomposition of the variance of arsenic concentrations within the states into between system, between source and within source components and the implications for the prediction of the percentage of a state's systems and sources affected by various MCLs are provided in section 4. Finally, conclusions and a discussion of our findings are presented in section 5.

2. SOURCE OF DATA

The availability of data with multiple measurements of arsenic concentrations from multiple sources in individual systems is limited. The recent nation-wide arsenic surveys, like the National Arsenic Occurrence Survey (NAOS) (Frey and Edwards 1997), and data provided by the United States Geological Survey (USGS) (Lockwood et al. 1999) did not provide such data. However, because regulatory compliance and reporting are required for all water sources entering a water distribution system, compliance data reported to EPA by state regulatory agencies do provide such information. States measure the arsenic concentration

Table 1. Period and sample size for the four state datasets.

State	Sampling Date	Systems	Sources	Samples
California	1980 - 1997	1774	3418	27567
Illinois	1993 - 1999	1226	1810	3565
New Mexico	1985 - 1999	699	1782	3069
Utah	1980 - 1999	763	1841	6433

at entry points of treated source water into a distribution system in each period of time to insure compliance with EPA's MCLs. The arsenic concentration is measured in treated water. Samples at the entry points represent the water quality at individual "sources" within a system.

Data from twenty-five states is included in the Arsenic Compliance Database (ACD), most of which do not record the name of the source from which the water sample is collected. Four state datasets are selected for detailed analysis, because they provide sampling results from multiple sources (entry points) for an appreciable number of water systems and they record the source names. The states are California, Illinois, New Mexico and Utah, which are representative of the western and middle parts of the US. Unfortunately, EPA data on the eastern part of the country are not available. Arsenic is known to vary geographically, therefore, our results should not be applied to make national prediction without more data.

The datasets from the four states share some common features. The data were collected over the past 20 years. Most of samples (about 90%) are from ground water sources; only about 10% of the measurements are from surface water sources. Figure 1 shows the plots of arsenic concentration vs. time for all four states, which indicates that the arsenic measurements are censored at multiple levels. The arsenic concentrations are measured by different chemical methods, each of which has a method detection limit (MDL). That is, if the arsenic concentration is below the MDL, the method could not detect the presence of arsenic so the only information we know is that the arsenic concentration is below the MDL. Federal regulations (40 CFR Part 136.1) provide more details about MDLs. The non-detected record can be treated as left censored data at the value of MDL. The plots from Figure 1 show that, in every state, there were initially one or two predominant MDLs. Then, in the middle of 1990's, they were changed to significantly different levels. The MDL was changed from $10\mu g/L$ and $5\mu g/L$ to $2\mu g/L$ for California; $1\mu g/L$ to $2\mu g/L$ for Illinois; $5\mu g/L$ to $1\mu g/L$ for New Mexico; and $1\mu g/L$ to $5\mu g/L$ for Utah. We consider these data features in our analysis.

3. METHODS

Our goal is, first, to build a hierarchical model to describe the distribution of arsenic concentration in sources by including between-source variability – the difference in arsenic concentrations between different sources within a water system. Then, based on this hierarchical model, we investigate the potential impact of this between-source variability on

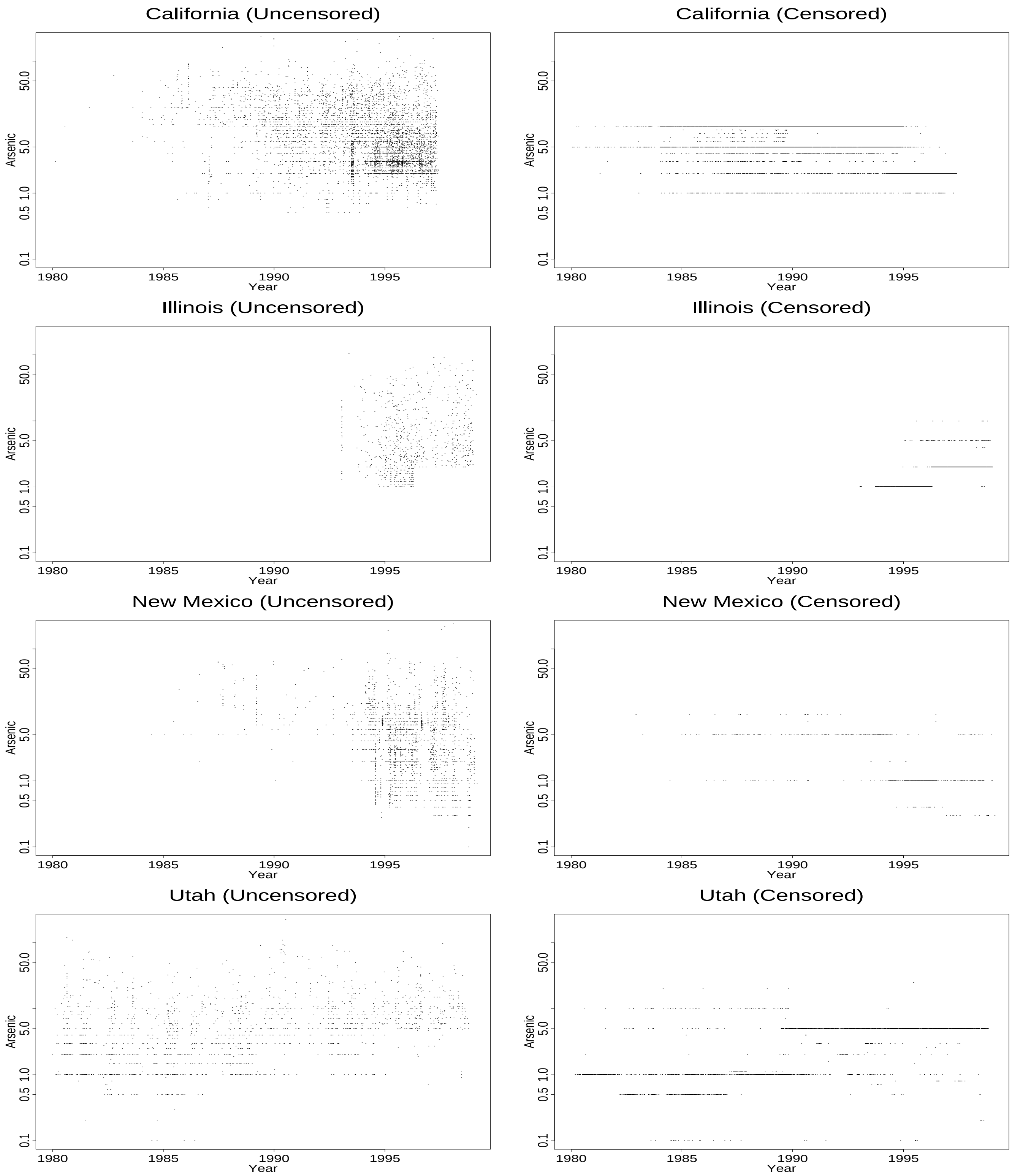


Figure 1. Plots of arsenic concentration vs. time for all four states: California, Illinois, New Mexico and Utah.

the prediction of the systems and sources that would be affected by possible changes of the arsenic MCL.

It is natural to model our problem hierarchically. Let's define η_i as the arsenic concentration in $\mu g/L$ in the ith system and ρ_{ij} as the arsenic level in the jth source within the ith system. It might be reasonable to expect that estimates of the η_i should be related to each other in the different systems, and that, in the different sources within the ith system, estimates of ρ_{ij} should be related to each other. Such a two-stage hierarchical structure can be modeled in a natural way if we use one prior distribution in which the η_i's are viewed as a sample from a common population distribution for a whole state. The common population distribution is specified by the parameter $\boldsymbol{\psi_1}$. We use another prior distribution in which the ρ_{ij}'s are viewed as a sample from a common population distribution for the ith system. The ρ_{ij} are given a probabilistic specification in terms of η_i and $\boldsymbol{\psi_2}$. Finally, we model the observed data $\boldsymbol{Y_{ij}}$ from the jth source within the ith system parametricly in terms of ρ_{ij} and $\boldsymbol{\psi_3}$. The observed data $\boldsymbol{Y_{ij}}$ are then used to estimate aspects of the population distribution of η_i and ρ_{ij}.

The computation of the features of the posterior distribution of parameters $\boldsymbol{\rho}$, $\boldsymbol{\eta}$, $\boldsymbol{\psi_1}$, $\boldsymbol{\psi_2}$ and $\boldsymbol{\psi_3}$ given the observed data $\boldsymbol{Y}$ was performed via MCMC simulation, by the following factorization:

$$P(\boldsymbol{\rho}, \boldsymbol{\eta}, \boldsymbol{\psi_1}, \boldsymbol{\psi_2}, \boldsymbol{\psi_3}|\mathbf{y}) \propto P(\mathbf{y}|\boldsymbol{\rho}, \boldsymbol{\psi_3})P(\boldsymbol{\rho}|\boldsymbol{\eta}, \boldsymbol{\psi_2})P(\boldsymbol{\eta}|\boldsymbol{\psi_3})P(\boldsymbol{\psi_1}, \boldsymbol{\psi_2}, \boldsymbol{\psi_3})$$

Prediction of future data $\mathbf{Y}'$ can be made based on the posterior predictive distribution, which is found from the posterior density of the parameters:

$$P(\mathbf{y}'|\mathbf{y}) = \int P(\mathbf{y}'|\boldsymbol{\rho}, \boldsymbol{\eta}, \boldsymbol{\psi_1}, \boldsymbol{\psi_2}, \boldsymbol{\psi_3})P(\boldsymbol{\rho}, \boldsymbol{\eta}, \boldsymbol{\psi_1}, \boldsymbol{\psi_2}, \boldsymbol{\psi_3}|\mathbf{y})d\boldsymbol{\rho}d\boldsymbol{\eta}d\boldsymbol{\psi_1}d\boldsymbol{\psi_2}d\boldsymbol{\psi_3}$$

There are a number of issues that must be addressed in the modeling of the data.

- Earlier research (Lockwood et al. 1999) and exploratory data analysis of our data indicate that the characteristics of arsenic concentration in ground and surface source water are generally different. It is necessary to model surface and ground source water separately.

- Samples have been collected over the past 20 years. The possibility that the arsenic levels in sources have been changing over time should be investigated. Recent research (Mclaren and Kim 1995) reported that one-year period seasonal variation in arsenic concentration was observed in some surface waters. Such variation may be due to seasonal conversion of arsenic in river sediments to more soluble forms. Trend effects may be observed due to the improvement of water treatment methods in treatment facilities, since our arsenic levels were measured in treated water.

- The apparent change of MDLs around the middle of the 1990's may be associated with changes in measurement methods or in laboratory conditions. If this is true, differences in precision and accuracy of the measurements before and after the MDL change are worthwhile issues to address. The accuracy and precision of the measurements are characterized by the bias and variability of measurements.

Explanatory variables $\mathbf{X}$ in the hierarchical model are thus chosen to include the effects of time, the changes in the MDLs , and the source water type.

3.1 The Hierarchical Model

In this section, we specify the two prior distributions to model the two-stage hierarchical structure and we specify the parametric model to describe the observed data Y_{ijk} with respect to ρ_{ij}. Y_{ijk} is the logarithm of the arsenic concentration for the kth measurement in the jth source of the ith system, where $i = 1, ..., I$, $j = 1, ..., m_i$, and $k = 1, ..., n_{ij}$.

- **System Level Prior:** (η_i)

$$\eta_i|\mu, \sigma^2_{sys} \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2_{sys})$$

  The variance component σ^2_{sys} describes the variability of the arsenic concentrations among the different systems within a state.

- **Source Level Prior:** (ρ_{ij})

$$\rho_{ij}|\eta_i, \mu_{sou(ij)}, \sigma^2_{sou(ij)} \overset{\text{i.i.d.}}{\sim} N(\eta_i + \mu_{sou(ij)}, \sigma^2_{sou(ij)})$$

  where $\mu_{sou(ij)}$ and $\sigma^2_{sou(ij)}$ are defined as follows. Define $W_{ij} = 1$ if a source is a ground water source, and $W_{ij} = -1$ if a surface water source. Then

$$\begin{aligned}\mu_{sou(ij)} &= W_{ij}\alpha \\ \sigma^2_{sou(ij)} &= \sigma^2_{sou.g} I_{\{W_{ij}=1\}} + \sigma^2_{sou.s} I_{\{W_{ij}=-1\}}\end{aligned}$$

  At this level, we model the ground and surface water sources separately. In $\mu_{sou(ij)}$, α represents the mean difference of log(arsenic) between ground and surface water sources. In $\sigma^2_{sou(ij)}$, we separately model the variability of arsenic concentration among ground water sources within systems as $\sigma^2_{sou.g}$ and variability among surface water sources within systems as $\sigma^2_{sou.s}$.

- **Within Source Level Model:** (Y_{ijk})

$$Y_{ijk}|\rho_{ij}, \mu_{InSou(ijk)}, \sigma^2_{InSou(ijk)} \overset{\text{i.i.d.}}{\sim} N(\rho_{ij} + \mu_{InSou(ijk)}, \sigma^2_{InSou(ijk)})$$

  To define $\mu_{InSou(ijk)}$ and $\sigma^2_{InSou(ijk)}$, first, we model the change of the true arsenic concentration in the sources as a function of time. Second, we model the measurement accuracy and precision differences before and after the apparent change in the MDL. The bias and variability of measurement methods are investigated. We continue to model ground and surface water separately.

  The observed mean arsenic concentration for each source is equal to the real mean arsenic concentration of that source plus the bias of the measurement method. We model the observed arsenic level as a joint function of time, water type and measurement methods (a function of time). The detailed format of the joint function we used is discussed in the next part. Generally, it can be expressed as

$$\mu_{InSou(ijk)} = g(t_{ijk}, W_{ijk}).$$

Similarly, the observed variability of arsenic concentrations in a single source should be a function of time, measurement and water type. We use the following simple variance function:

$$\begin{aligned}\sigma^2_{InSou(ijk)} &= V(t_{ijk}, W_{ij}) \\ &= \sigma^2_{00} I_{\{W_{ij}=1, t_{ijk} \le t_0\}} + \sigma^2_{01} I_{\{W_{ij}=1, t_{ijk} > t_0\}} \\ &\quad + \sigma^2_{10} I_{\{W_{ij}=-1, t_{ijk} \le t_0\}} + \sigma^2_{11} I_{\{W_{ij}=-1, t_{ijk} > t_0\}},\end{aligned}$$

where t_0 is the point in time at which the MDL changes. That is, we fit different variances for ground and surface water before and after t_0.

- **Hyperprior Distribution:**
  We assume that the seven variance components in the above model are independently and identically distributed as $Gamma^{-1}(\alpha_p, \beta_p)$. With the regression parameters, we assume $Normal(\mu_p, \sigma_p^2)$. The conjugate priors are chosen for easy computation. We pick $\alpha_p = \beta_p = 0.01$ and $\mu_p = 0$, $\sigma_p^2 = 10000$, so that the prior distributions are very flat.

Regression Function for Fitting Arsenic Change with Time

The general regression function for the observed mean arsenic concentration in source water $g(t_{ijk}, W_{ij})$ is specified as follows. Define $U_+ = U$ if $U \ge 0$ and $U_+ = 0$ if $U < 0$. And $U_- = U$ if $U \le 0$ and $U_- = 0$ if $U > 0$. Consider a point in time at t_0 when the MDL changes. For ground and surface source water, we fit two different regression curves. Each curve can be expressed as:

$$\begin{aligned}R(t_{ijk}) &= \beta_0(t-t_0)^0_- + \beta_{10}(t-t_0)^1_- + \beta_{20}(t-t_0)^2_- + \beta_{30}(t-t_0)^3_- \\ &- \beta_0(t-t_0)^0_+ + \beta_{11}(t-t_0)^1_+ + \beta_{21}(t-t_0)^2_+ + \beta_{31}(t-t_0)^3_+ \\ &+ \beta_4\cos(2\pi t) + \beta_5\sin(2\pi t)\end{aligned}$$

The $\beta_0(t-t_0)^0_+$ and $\beta_0(t-t_0)^0_-$ allow the function to jump at t_0. β_0 represents the bias difference in measurements before and after the MDL change. The power terms represent the trend effects. Before t_0 and after t_0 we fit two different regression curves. The seasonal effect is represented by $\beta_4\cos(2\pi t) + \beta_5\sin(2\pi t)$, with an one-year period. Since the transition of MDL is continuous, t_0 is chosen to best represent the changes.

3.2 Model Implementation

We estimate our model by Markov Chain Monte Carlo (MCMC) simulation. Gibbs Sampling (Geman and Geman 1984) algorithms written in C are used to compute the posterior distributions of the model parameters. Then, samples are drawn based on the posterior predictive distributions for use in model evaluation and prediction. Here, we focus on several non standard features of the implementations.

Efficient Re-parameterization

The Monte Carlo algorithm converges very slowly due to high correlation amongst the parameters and weak identifiability of certain parameters. An appropriate parameterization can dramatically affect the shape of the posterior surface and improve convergence and stability of the samples.

- With random effects, it is generally more efficient to use "hierarchical centering" (Gelfand et al. 1995), in which the nested random effect distribution is centered around a higher population level, which is the way we parameterize our model. The centered parameterization is $(\boldsymbol{\mu}, \boldsymbol{\eta}, \boldsymbol{\rho})$. This may be particularly true with multiple random effects, like in our case.
- With regression parameters, we standardize covariates around their mean in order to make the estimated parameters more orthogonal. In our model, only the covariate "Sample collection time" is centered by subtracting the sample mean value.

Data Augmentation - Censored Data

The treatment of censored arsenic observations is handled through data augmentation (Gelman et al. 1995). Before each iteration of the sampling algorithm, simulated values for the censored arsenic concentrations are imputed from their conditional distribution based on the parameters sampled in the previous iteration. This conditional distribution is a truncated lognormal where the truncation point is the MDL (Lockwood et al. 1999).

3.3 Model Checking & Prediction

We apply our hierarchical model to a new set of data $\tilde{\mathbf{Y}}$, which come from a set of water sources $\tilde{\rho} = (\tilde{\rho_{11}}, \tilde{\rho_{12}}, ..., \tilde{\rho_{ij}})$ with associated explanatory variables $\tilde{\mathbf{X}}$, representing the water type, sampling time, and measurement methods. Assume $\tilde{\mathbf{Y}}$ is independent of the original data $\mathbf{Y}$ given $\boldsymbol{\Theta}$ and has distribution $P(\tilde{\mathbf{y}}|\tilde{\mathbf{x}}, \theta)$. The predictive distribution of $\tilde{\mathbf{Y}}$ can then be obtained from the posterior distribution of the parameters by $P(\tilde{\mathbf{y}}|\tilde{\mathbf{x}}, \mathbf{y}, \mathbf{x}) = \int P(\tilde{\mathbf{y}}|\tilde{\mathbf{x}}, \theta)P(\theta|\mathbf{y}, \mathbf{x})d\theta$. To obtain a draw from the posterior predictive distribution of $\tilde{\mathbf{Y}}$, first obtain a draw from $P(\theta|\mathbf{y}, \mathbf{x})$ and then draw the predictive data $\tilde{\mathbf{Y}}$ from $P(\tilde{\mathbf{y}}|\tilde{\mathbf{x}}, \theta)$. Using this method, many aspects of the model can be investigated.

Checking a Model by Comparing Data to the Posterior Predictive Distribution.

Our basic technique for checking the fit of a model to data is to draw simulated values from the posterior predictive distribution of replicated data and compare the empirical CDF of the observed data to the CDF of the posterior predictive simulation. Replicated data are simulated arsenic values from the same sources with the same explanatory variables as the original data. Any systematic difference between the simulations and the data indicates a potential failing of the model. The empirical CDF of the data was estimated by the Kaplan-Meier (product-limit) estimator (Kaplan and Meier 1958).

Prediction of the Number of the Systems and Sources Exceeding the MCLs

One interesting feature of the model prediction is the prediction of the numbers of water systems and sources in which the arsenic levels are exceeding the various possible MCLs. The "Source-Based" predictions are the predictions derived by considering variability among sources within individual systems. Our objective is to compare the "Source-Based" predictions with the predictions that do not take between-source variability into account, called "System-Based" predictions. See ICF Kaiser International Consulting Group (1999) for further discussion of "System-Based" and "Source-Based" predictions. The prediction difference gives an idea of the potential biases and uncertainties in compliance cost estimates that are associated with the use of the "System-Based" estimation.

The predictions of the numbers of the systems and the sources affected by a MCL are simulated using a sample from the posterior distribution obtained via MCMC. In every iteration of the Markov chain conditional on the parameters simulated from the posterior distribution, an arsenic sample is drawn for each source. According to the recent measurement procedure, if the sample value is below the MCL, declare the source is in compliance. If the one sample is above the MCL, take three more samples and average the four. Declare the source to be out of compliance if and only if this average exceeds the MCL. Then, the number of sources affected by the MCL is estimated by counting the number of sources out of compliance in each iteration. A system affected by the MCL is defined as one with at least one source above the MCL. We repeat the simulation for the different MCLs $a_1, a_2, ..., a_s$. Thereby, at each MCL a_i for $i = 1, ..., s.$ we have a sample of the numbers of systems and sources affected at that MCL. The 90% uncertainty bands are pointwise (at each a_i) 5th and 95th percentiles of these samples. Pointwise medians are also calculated.

For "System-Based" prediction, the posterior distribution is derived from a reduced model eliminating the between-system variability, which is the only difference between the reduced model used by the "System-Based" prediction and the full model used by the "Source-Based" prediction.

4. RESULTS

First, we present comprehensive results for parameter estimation, model checking and prediction for the Utah dataset. Then, we repeat the analysis for the other three states and compare the results for the four states.

4.1 Variance Components

Variance between Sources vs. Variance between Systems

The decomposition of the arsenic concentration variance for the Utah measurements is summarized in Figure 2. Figure 2(top) compares the estimated posterior densities of the variance between systems to the variance between sources (within systems) for the Utah dataset. Source-level variances are present, but they are smaller than the system-level variance. This indicates that the model for characterizing the arsenic concentrations can be improved by including the source-level variabilities, but the improvement should not dramaticly change the estimated effects of alternative MCLs and the national distribution of arsenic concentration based on the model which only includes system-level-variability. There is very little information in the data about between-source variability for surface water, since many of the surface source water systems (in Utah and elsewhere) have only one intake point. This is the reason that the posterior distribution for the source-level variance for surface waters in Utah has a much greater spread than that for ground waters.

Variances of Random Errors

The estimated posterior densities of the variance components for random errors are shown in Figure 2(bottom). The characteristics of the posterior densities of random error variabilities are significantly different before and after the MDL change (assumed to occur at $t_0 = 1989$). For ground water, the variability becomes significantly smaller after the MDL change. If such change can be explained by the change of the measurement methods, the methods used after the MDL change have smaller variability than those used before MDL change. For surface water, the posterior density of variability becomes very flat after the MDL change. Most of the surface water data are censored, because the new methods use a higher MDL (5 $\mu g/l$; prior to 1989 a MDL of 1 $\mu g/l$ is more common) and surface waters generally have arsenic concentrations below this higher level. Therefore the data can not provide accurate estimation for the variances of surface water measurements, especially after 1989. The new measurement methods thus appear to have smaller measurement variability than the old methods when they are used to measure samples with high concentrations of arsenic. But they are unable to resolve lower concentrations. The better resolution of the estimated variances for ground water vs. surface water sources is also due to the larger sample size for the former (n: 5753) vs. the latter (n: 680).

4.2 Water Type Effects vs. Measurement Bias

Figure 3 shows the posterior densities of mean arsenic for ground and surface source water before and after the MDL change. Similar as in the variances of random errors, the characteristics of the mean arsenic concentrations are significantly different before and after the MDL change. The results suggest that the measurement methods used before and after the MDL change have different measurement bias. Before the MDL change, the ground source water has a similar mean arsenic level as the surface source water, but, mean arsenic levels are significantly different between ground and surface water after the MDL change. Therefore, we can not tell if there is a real difference in mean arsenic levels, because the observed difference could be caused by the bias of the measurement methods.

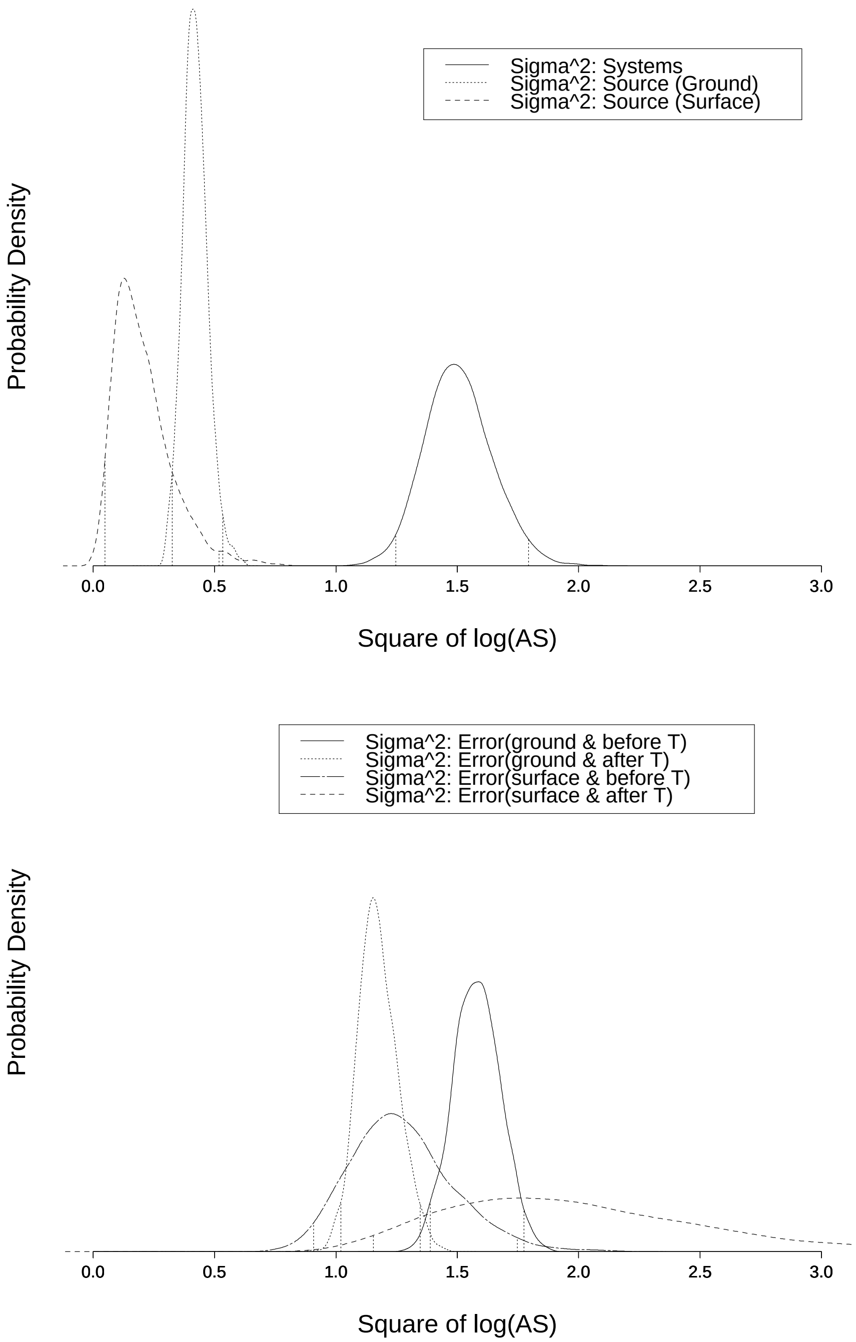


Figure 2. The estimated posterior densities of the components of the variances of arsenic concentration in Utah source waters. (Top): System-level and source-level variances. (Bottom): Within-source (random error) variances. The MDL transition point for Utah is given by $t_0 = 1989$.

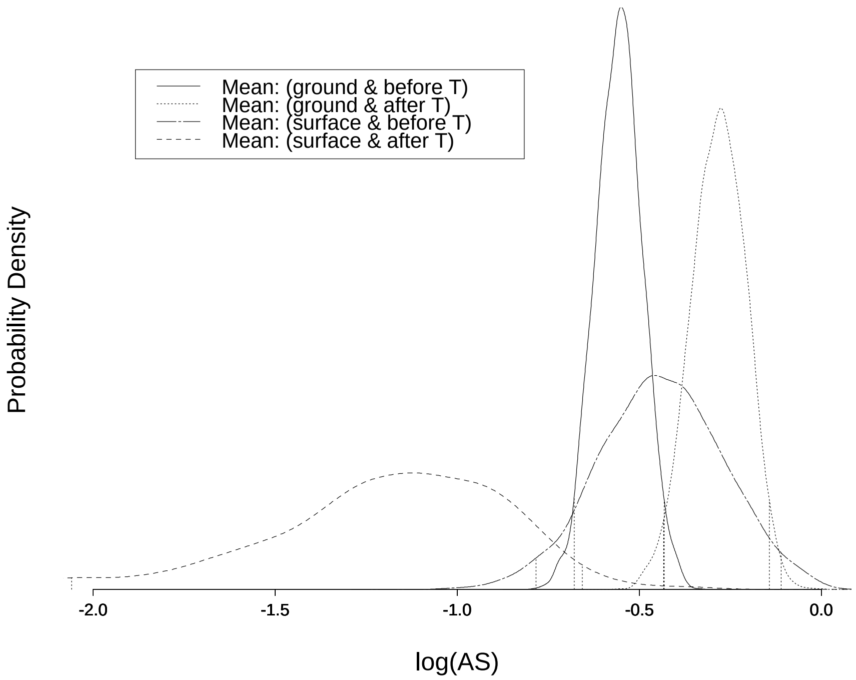


Figure 3. The posterior densities of the mean arsenic concentration of ground and surface source water before and after the MDL change for the Utah dataset.

4.3 Time Dependent Effects

Although the long-term trend effects are not significant, one-year-period seasonal effects are observed for both surface and ground water concentrations. Figure 4 shows the posterior densities of $2\sqrt{\beta_4^2 + \beta_5^2}/\sigma_{err(t>t_0)}$ for ground water (Fig.4(left)) and surface water (Fig.4(right)) , which indicates the relative magnitude of the maximum seasonal variability compared to the random error variability after the MDL change. The posterior mode of that ratio for surface water (25%) is much higher than that for the ground water (8%). The greater seasonal variability in surface vs. ground water samples is expected since the former are subject to greater variations in flow rate and temperature, while such variations tend to be dampened in ground water sources.

4.4 Goodness of Fit

Figure 5 shows the empirical CDFs of the data and the CDFs of the posterior predictive simulation of the replicated data from the full model for both ground and surface water. Generally, the fit of the model to the data is good. For most of the arsenic range, the empirical CDFs fall inside the 90% uncertainty band calculated from the CDFs of the predictive simulation. The simulated and observed CDFs differ greatly from each other only at low arsenic levels (below 1 μg/L) for ground water. At this level, the estimation of the empir-

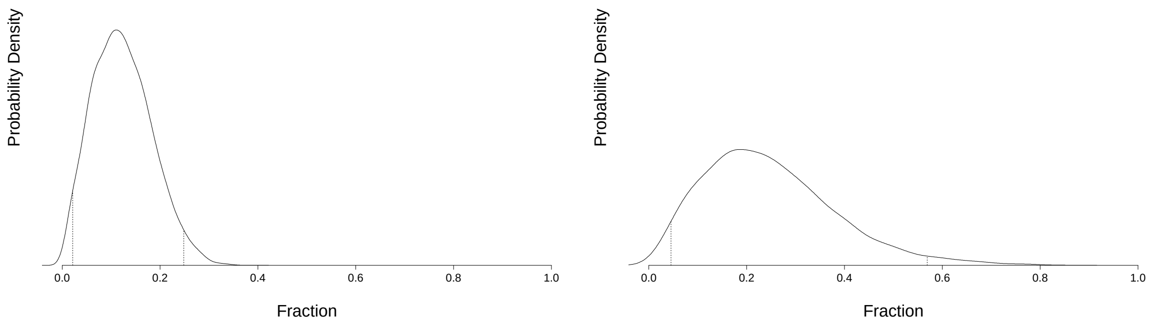


Figure 4. The posterior densities of $\frac{2\sqrt{\beta_4^2+\beta_5^2}}{\sigma_{err(t>t_0)}}$ for the Utah dataset. It indicates the relative magnitude of the maximum seasonal variability compared to the random error variability after the MDL change for ground water (left) and surface water (right).

ical CDF is not accurate because the estimation is based only on a small amount of data. Accurate prediction of the CDF is most important at high arsenic concentrations, since this is where heath concerns occur and where MCLs will have then primary impact.

The CDF of the simulated data for surface water has a wider uncertainty band than that for ground water, because the posterior densities of the parameters for surface water are more flat than those for ground water. It is again due to the fact that only about 10% of the data are from surface water and many of these measurements are censored.

4.5 'Source-Based" vs. "System-Based" Prediction

The two most important results of our analysis are: (1) the prediction of the percentage of all of a state's sources and systems in which the arsenic concentration exceeds a number of potential MCLs; (2) the comparison between the "Source-Based" prediction from the full model and the "System-Based" prediction from the reduced model. Figure 6 shows our main results for Utah. A system which is out of compliance is defined as a system with at least one source out of compliance. We analyze the potential MCLs of 2, 5, 10, 20 and 50 $\mu g/l$, which is the range currently under consideration by the EPA (Frey et al. 1997). Determining whether a system or source is out of compliance is based on the multiple measurement procedure. More details on this procedure are presented in section 3.3.

Our predictions (Figure 6) have several features in common as the prediction from the nationwide single-sample model (Lockwood et al. 1999). First, the rate of change in the proportion of exceedance for both sources and systems increases as the proposed MCL decreases. This indicates that the number of systems and sources requiring treatment and likely the corresponding cost of the regulation increase more rapidly at low values. Second,

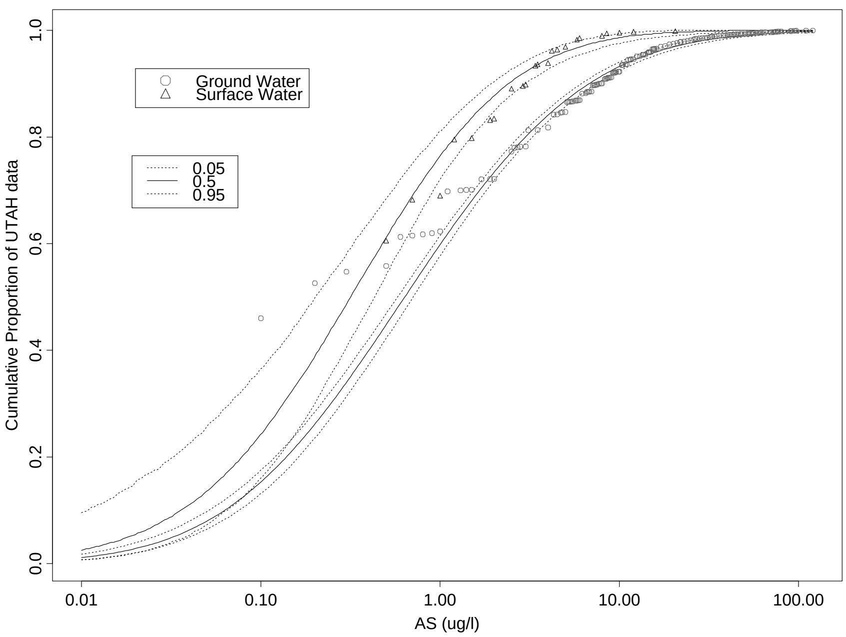


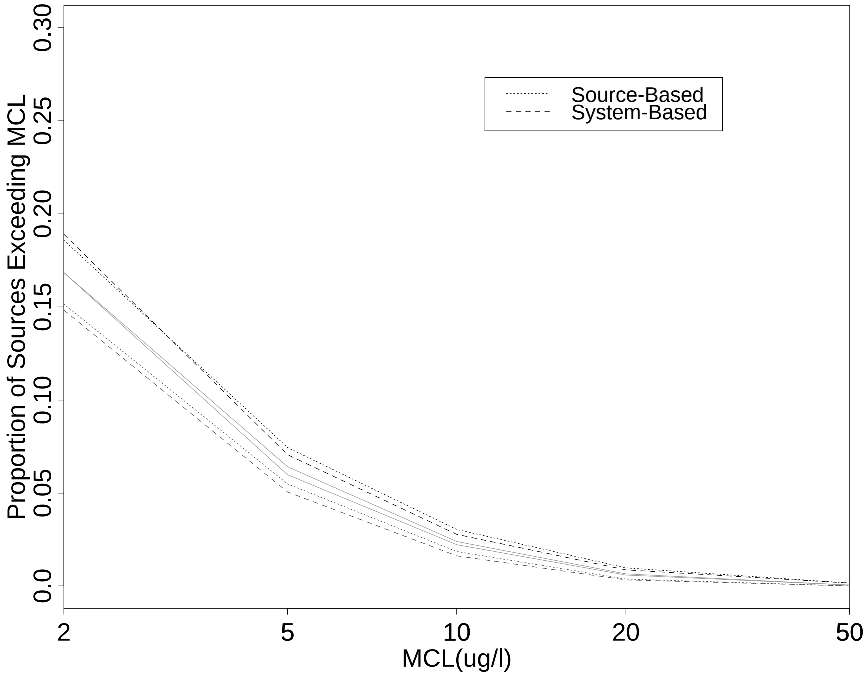
Figure 5. Empirical CDF of the data and CDF of the posterior predictive simulation of replicated data.

the uncertainty in exceedance is a decreasing function of the MCL; this makes sense due to the additional uncertainty about low arsenic concentrations fostered by the censored data.

Our multiple-sample in multiple-source model provides other features which could not be obtained from the single-sample model. First, the proportion of exceedance for systems with at least one source out of compliance is higher than the proportion of excess for sources out of compliance at the same MCL; this indicates that it is common that more than one, but not all, sources, are out of compliance for a system with multiple sources, therefore, calculating the cost of the regulation associated with sources instead of systems will provide more accurate estimation. Also, the difference between "Source-Based" and "System-Based" predictions is very small in the Utah dataset; this is consistent with the fact that the between-source variability is smaller than the between-system and within-source variabilities, two of which have similar magnitude, thus, eliminating the source level variance dose not change the prediction much. This is not true for all of the other states.

4.6 Results for Other States

We repeated the above analysis for three other states. Figures 7, 8, 9 and 10 summarize the results about the variance components, the mean arsenic concentrations, the seasonal variations and the proportion of exceedance for the four states.



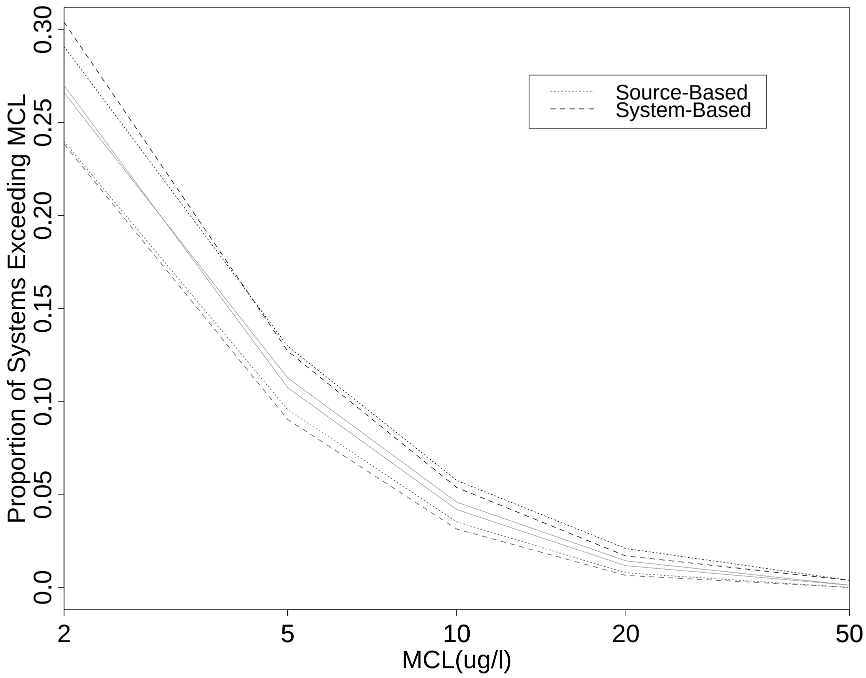


Figure 6. Predictions of the fractions of the sources (top) and the systems (bottom) in Utah affected by the various potential MCLs. The dotted lines represent pointwise 5 and 95 percentiles of the posterior distributions. The solid lines are 50 percentiles.

There are some common as well as different features among the four states.

- The between-system variability is larger than the between-source variability and the random error variabilities in all cases. The size of the between-system variability also varies among the states. Illinois has the highest value (posterior mode: 3.5) and California has the lowest value (posterior mode: 1.2).
- The between-source and the random error variabilities generally have the same size except in Utah where the random error variance is similar in magnitude to the between-system variability.
- Generally, the size of the difference between "Source-Based" and "System-Based" predictions is similar to the size of the uncertainty bands and the size is small; this is consistent with the fact that the between-source variability is similar in size to the within-source (random error) variabilities, but smaller than the between-system variability.
- The characteristics of random error variabilities are different before and after the MDL change. Generally, these variances become smaller after the MDL change. This indicates that the later measurement methods have smaller measurement variability than the former methods.
- The difference in the observed mean arsenic concentration before and after the MDL change (see Figure 8) is more apparent in California and Utah than in the other two states. This may be due to the fact that the changes of the MDL in California and Utah are more dramatic than in the other two states. The shift of the observed mean arsenic level before and after the MDL change is consistent in the ground and the surface source water. Therefore, from both measurement methods, we can conclude that mean arsenic concentration for ground water is higher than that for surface water. We could not, however, have such a conclusion for the Utah dataset. See section 4.2 for a detailed analysis of the Utah dataset.
- The arsenic concentrations in surface source water have higher seasonal variation than those for ground source water. The posterior density of the relative seasonal variation for surface source water in Illinois is much higher and flatter than that in the other states. This may be due to the fact that more of the samples of surface source water in Illinois have arsenic levels which are below the censoring level.

5. CONCLUSION

Generally, the between-source variability is smaller than the between-system variability. The random error variability has a magnitude similar to the between-source variability, especially when assessed by applying the recent measurement methods. This indicates that the model for characterizing the arsenic concentrations can be improved by including the source-level and within-source (random error) variabilities, but the improvement should not

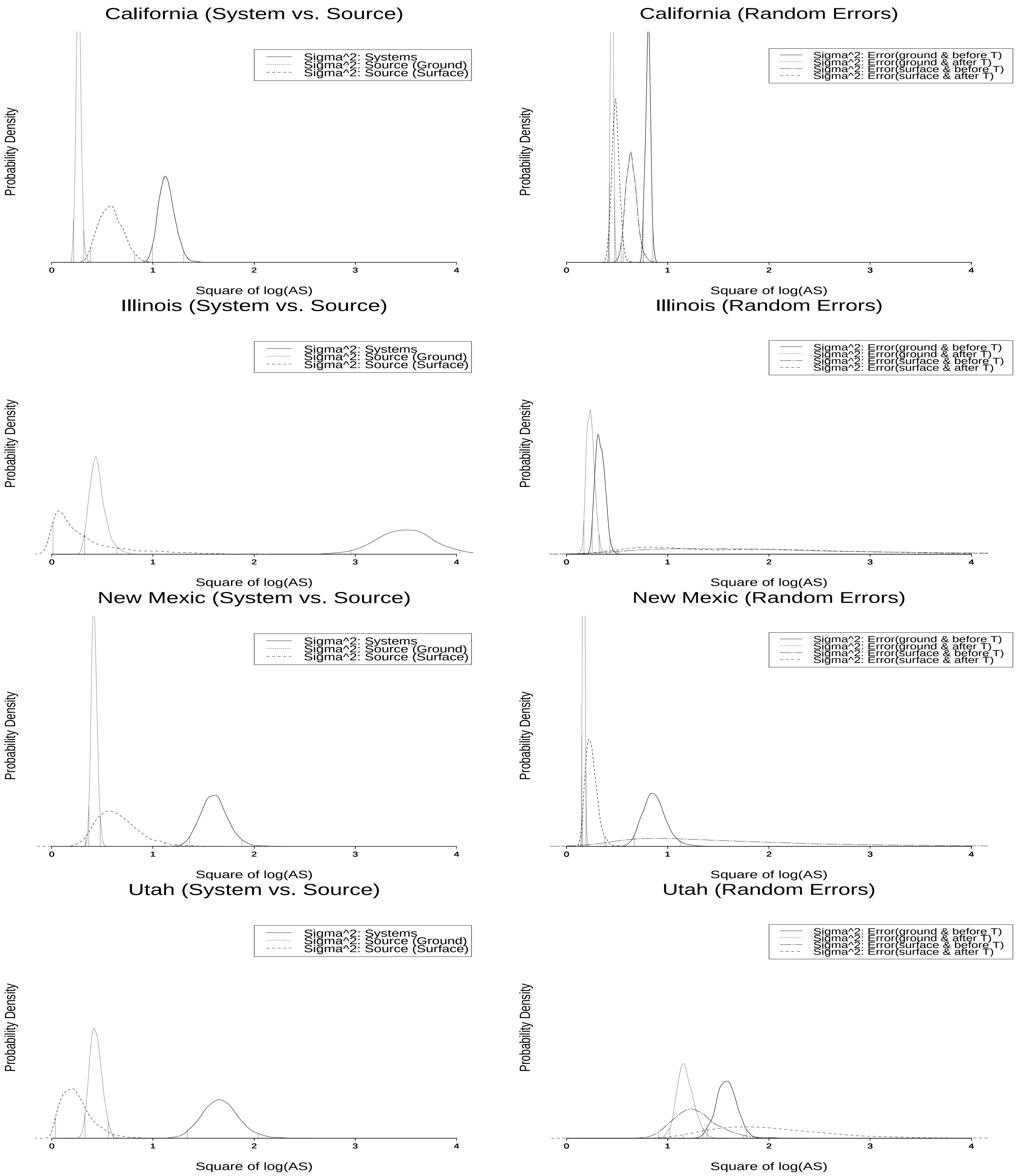


Figure 7. The posterior densities of the variance components for the four states. Refer to Figure 2 for the details about the plots.

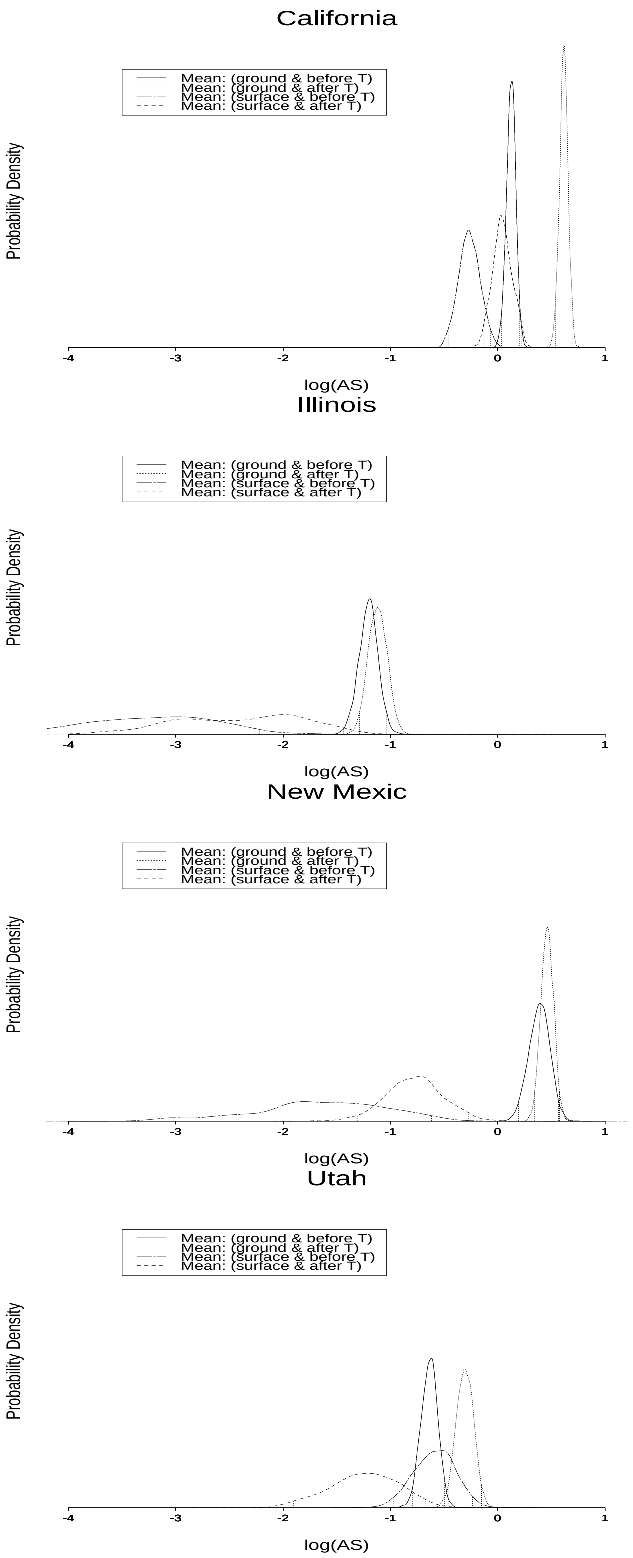


Figure 8. The posterior densities of the mean arsenic concentration for ground and surface source water of the four states. Refer to Figure 3 for the details about the plots.

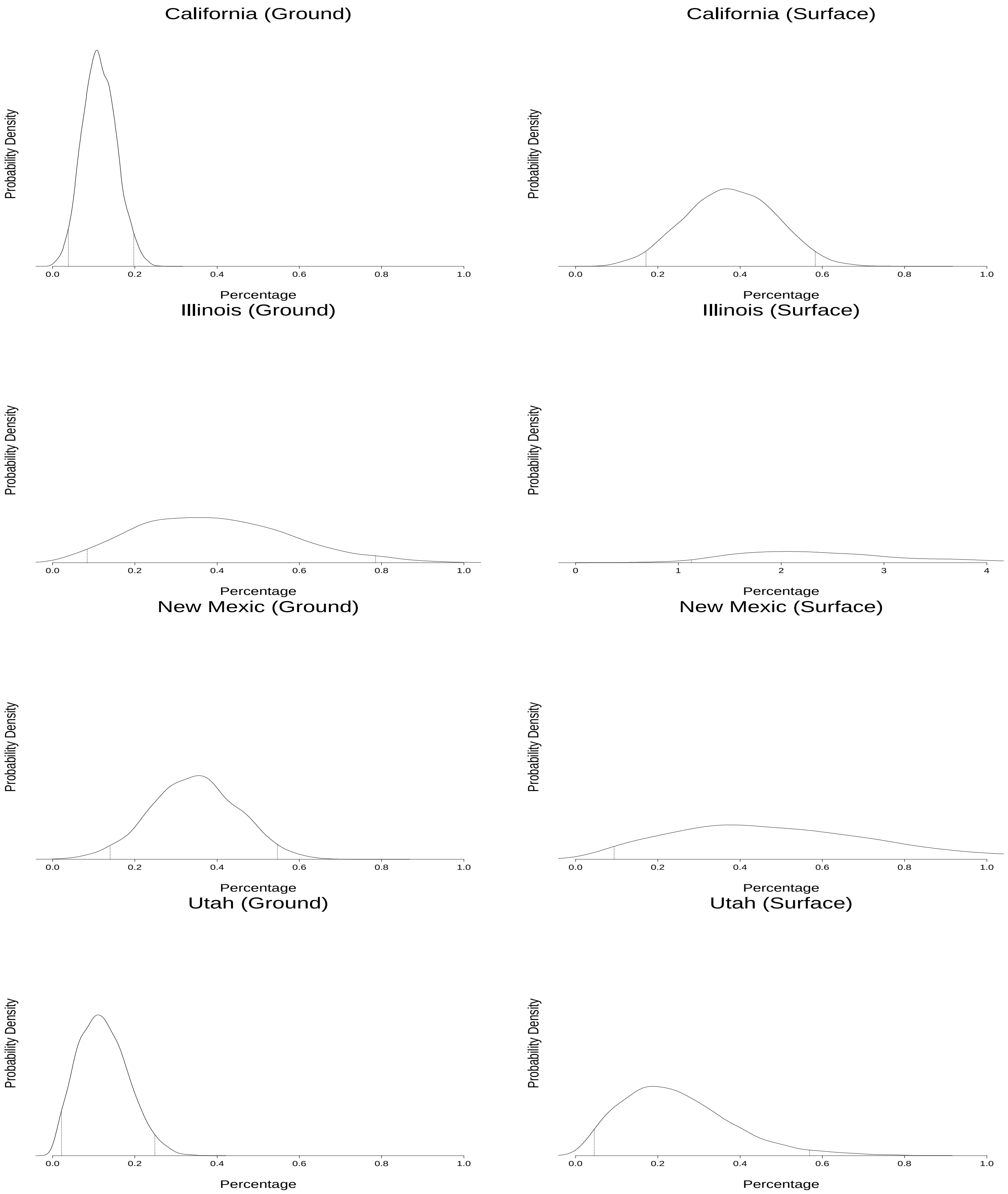


Figure 9. The posterior densities of $\frac{2\sqrt{\beta_4^2+\beta_5^2}}{\sigma_{err(t>t_0)}}$ for ground and surface source water of the four states. Refer to Figure 4 for the details about the plots.

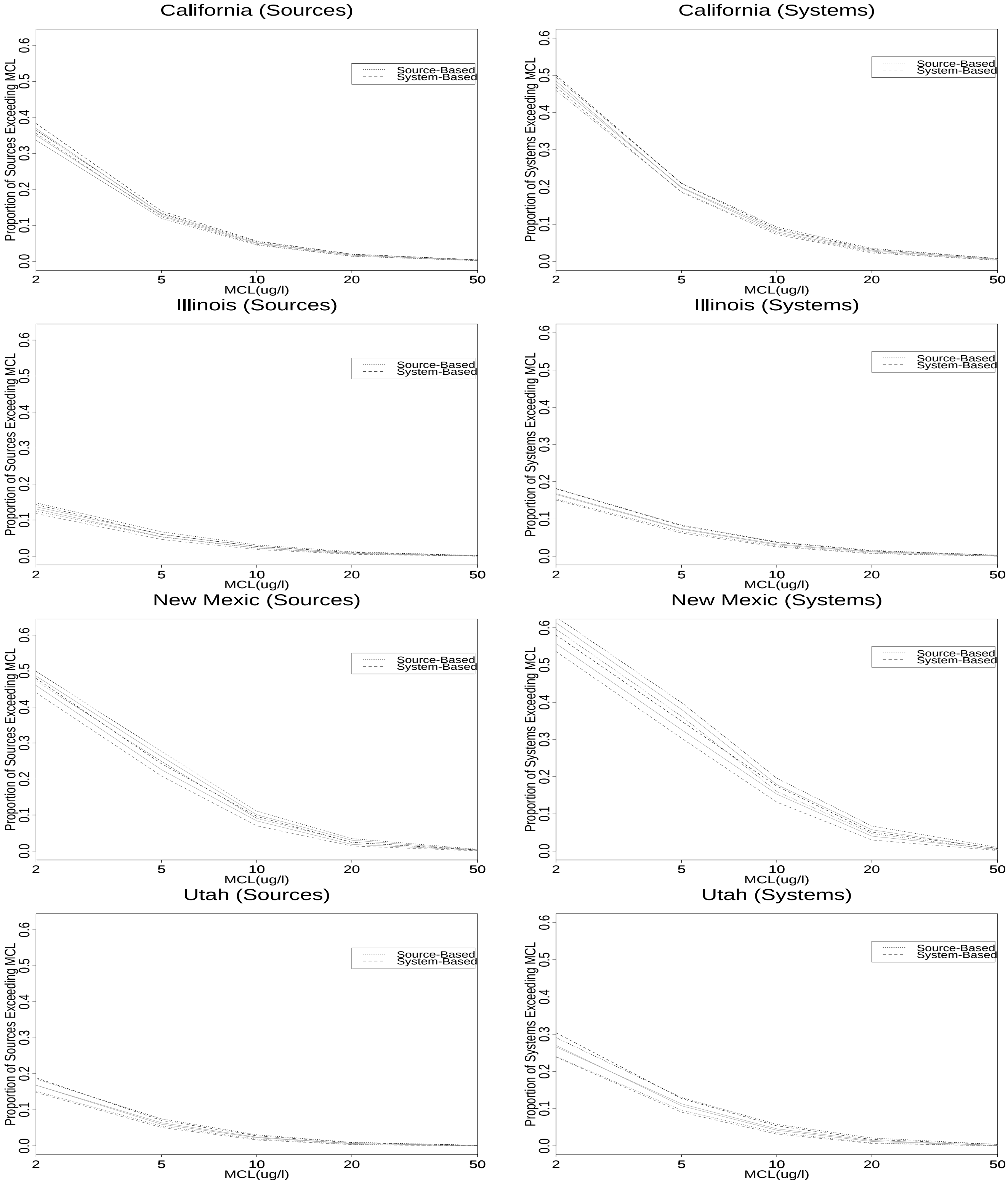


Figure 10. Predictions of the fractions of the sources and the systems affected by the various potential MCLs for the four states. Refer to Figure 6 for the details about the plots.

dramaticly change the estimated effects of alternative MCLs and the national distribution of arsenic concentration based on the model which considers only system-level-variability.

We observed differences in the distributions of the arsenic measurements before and after the apparent MDL change. Such differences indicate the existence of the difference in bias and variance between the former and the later measurement methods.

REFERENCES

Christen, K. (1999), "Arsenic standard for drinking water too high, NRC says," *Environmental Science and Technology*, 188A.

Frey, M. and Edwards, M. (1997), "Survey arsenic occurrence," *J. AWWA*, 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," Technical report, Water Industry Technical Action Fund Report.

Gelfand, A., Sahu, S., and Carlin, B. (1995), "Efficient Parametrisations for Normal Linear Mixed Models," *Biometrika*, 82, 479–488.

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

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.

Hopenhayn-Rich, C., Biggs, M. L., Fuchs, A., Bergoglio, R., Tello, E., Nicolli, H., and Smith, A. (1997), "Bladder Cancer mortality associated with arsenic in drinking water in Argentina," *Epidemiology*, 117, 7(2).

ICF Kaiser International Consulting Group (1999), "Intra-System variability in groundwater radon levels," Technical report, Office of Groundwater and Drinking Water.

Kaplan, E. L. and Meier, P. (1958), "Nonparametric Estimation From Incomplete Observations," *Journal of the American Statistical Association*, 53, 457–481.

Lockwood, J., Schervish, M., Gurian, P., and Small, M. (1999), "Characterization of arsenic occurrence in US drinking water treatment facility source waters," Technical report, Department of Statistics, Carnegie Mellon University.

Mclaren, S. and Kim, N. (1995), "Evidence for a Seasonal Fluctuation of Arsenic in New-Zealands Longest River and the Effect of Treatment on Concentrations in Drinking-Water," *Environmental Pollution*, 67–73, 90: (1).

Pontius, F., Brown, K., and Chen, C. (1994), "Health implication of arsenic in drinking water," *J. AWWA*, 86(9).