January 2, 2002

Bayesian Prediction of National Multi–Contaminant Trends in Community Water System Sources

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

The current framework for U.S. Environmental Protection Agency regulation of water quality in community drinking water supplies consists of sequential rules for either single contaminants or small groups of similar contaminants. For both substantive and pragmatic reasons, promulgating less frequent rules for larger contaminant classes may be desirable. Such a change would require the expansion of existing regulatory evaluation technologies to account for joint occurrence distributions of the contaminants. This paper extends existing methods for modeling the distributions of a single contaminant in community water system source waters to the simultaneous consideration of multiple contaminants. It considers alternatives for addressing the implementation difficulties inherent in the multivariate setting, providing solutions of general methodological interest. Through case studies involving arsenic, sulfate, magnesium and calcium, it shows how jointly modeling contaminants provides better fit and predictive power than marginal models, and emphasizes how inferences about critical regulatory quantities can be improved through joint modeling. The methods presented in this paper make significant progress in redressing several shortcomings of existing analyses.

KEY WORDS: water quality regulation; regulatory impact assessment; Markov Chain Monte Carlo; multivariate spatial data; data augmentation.

J.R. Lockwood is Associate Statistician, RAND Statistics Group, Pittsburgh, PA 15213. Mark J. Schervish is Professor, Department of Statistics and Mitchell J. Small is the H. John Heinz III Professor of Environmental Engineering, Departments of Engineering and Public Policy and Civil and Environmental Engineering, Carnegie Mellon University, Pittsburgh, PA 15213. Patrick Gurian is Assistant Professor, Department of Civil Engineering, University of Texas at El Paso, El Paso, TX 79968. The support and guidance provided by Judy Calem, Andrew Schulman, Ben Smith, Lew Summers and Jennifer Wu 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 or RAND peer review, and no official endorsement should be inferred. This research was supported by US EPA award No. R826890-01-0, the Cooperative Agreement CR825188-01-3 between US EPA Office of Policy, Planning and Evaluation (OPPE) and Carnegie Mellon University, and RAND.

1. INTRODUCTION

Regulations on the quality of water distributed by community water systems are set by the U.S. Environmental Protection Agency (EPA). The agency mandates uniform national upper bounds, or *maximum contaminant levels* (MCLs), on the concentrations of various organic and inorganic contaminants in community drinking water supplies, including microbial pathogens, radionuclides and heavy metals. The regulatory process begins by establishing a *maximum contaminant level goal* (MCLG), a non–enforceable standard which is the maximum concentration of the contaminant believed to result in no adverse health effects. Prior to the 1996 Safe Drinking Water Act Amendments (PL 104-182), the EPA was required to set the MCL as close to the MCLG as "feasible." One focus of the amendments was the establishment of a more formal and flexible decision framework for the regulatory process. For each proposed rule, the EPA must publish a "regulatory impact assessment" (RIA) comparing the estimated costs associated with upgrading treatment facilities and the estimated health benefits of reduced exposure. At its discretion, EPA may set the final MCL to a value higher than the feasible level if the costs of regulating to the feasible level would not be justified by the anticipated health benefits.

Community water suppliers must comply simultaneously with the MCLs for all regulated substances, the number of which is currently in excess of 90 (EPA 2001b) and is slated to grow continuously (at a rate of up to one per year) according to the regulatory plan detailed in the 1996 amendments. The initial post-amendment regulatory activity has focused on separate rules for single contaminants such as arsenic and radon, as well as rules for groups of similar contaminants such as radionuclides, pathogens and disinfection byproducts. Each of these contaminants or groups of contaminants are treated in isolation from a regulatory perspective, with independent RIAs either completed or underway for each rule. Water systems will be forced to comply successively with each rule as it is promulgated. However, because the RIAs are performed sequentially, the net realized costs and benefits for the collection of rules may be vastly different from those implied by the individual analyses.

The are numerous sources of potential discrepancy. First, because many treatments remove

a broad array of substances, a treatment technology which would be prohibitively expensive for treating a given contaminant class may be more attractive if it helps a system comply with multiple standards simultaneously. On the other hand, the presence of substance A may interfere substantially with the performance of a particular treatment in removing substance B. In this case, a more expensive treatment option may be necessary to adequately remove substance B, but a cost–benefit analysis that ignores the presence of substance A would not reflect these additional costs. In addition, health impacts can depend heavily on the joint occurrence of contaminants because certain substances such as pesticides may exhibit "additive or synergistic" toxic effects (Kolpin, Barbash, and Gilliom 1998; Gennings, Schwartz, Carter Jr., and Simmons 1997). Risk analyses performed one contaminant at a time thus may seriously misstate the actual risks caused by the simultaneous exposure to multiple substances. Finally, the failure to consider the joint behavior of contaminants neglects both the natural and anthropogenic activities which are known to induce relationships in contaminant occurrence. Not accounting for covariations in the presence of a high degree of contaminant co-occurrence is likely to produce estimated national occurrence distributions that are unrealistic, which in turn biases independently estimated costs and benefits.

In addition to these substantive issues, the sequential regulatory approach has been criticized on pragmatic grounds. Community water systems generally have only limited ability to augment treatment technologies. Most systems are quite small, serving less than a few thousand people, and thus financial constraints dictate that sweeping treatment changes cannot be made often. However, the current regulatory protocol produces new standards sequentially and frequently, which in the worst case may impose an unfeasible sequence of treatment upgrades (Roberson and Power 2000). This is at odds with a more holistic approach to regulation, whereby systems could optimize treatment upgrades to meet long-term water quality goals (Neukrug 2000). These factors suggest that a regulatory framework in which rules are made less frequently but for larger groups of contaminants might be more scientifically sound and practical. This paradigm shift would necessitate the development of a flexible, integrated RIA process capable of addressing a multitude of potentially diverse contaminants. Two of the stumbling blocks to the development of such a "multi-contaminant RIA" are the lack of adaptable methods for simultaneously analyzing groups of contaminants with re-

spect to their tendencies to occur and to be removed together. Both topics are in need of additional research, a fact noted by Guttorp (2000) with regard to air quality regulation.

Moreover, as the recent debates over the reduction of the arsenic MCL (EPA 2001a) have highlighted, systematically quantifying uncertainties is of paramount importance to an effective RIA process. Uncertainty analysis can be invaluable to navigating and arbitrating the often disparate estimates of key regulatory decision quantities made by different stakeholders. Previous work on arsenic regulation treats uncertainty in an integrated Bayesian framework by providing a raw (i.e. untreated) water occurrence model (Lockwood, Schervish, Gurian, and Small 2001) and the use of this model in a RIA (Gurian, Small, Lockwood, and Schervish 2001c,a). The cost estimates and associated uncertainties resulting from this work provided insights into the nature of the disagreements among other published analyses. It is thus imperative that the component models of a multi-contaminant RIA retain, if not augment, these capabilities for addressing uncertainty.

The current study develops and illustrates a method for calculating one of the key inputs to a multi-contaminant RIA in a manner that coherently quantifies uncertainty. In particular, it extends the work of Lockwood et al. (2001) on estimating national raw water occurrence for a single contaminant by deriving a flexible framework for estimating joint raw water occurrence distributions of multiple contaminants. Raw water concentrations, while typically not the focus of regulatory impact assessments, are an integral component of the proper quantification of treatment upgrade costs and their uncertainties. Their characterization is especially interesting from a statistical perspective because compliance monitoring is based only on finished (i.e. treated) water measurements, leading to generally sparse measurement and reporting of raw water data. Hence, as discussed in Lockwood et al. (2001), inferences about national raw water contaminant levels often must be based on only a single sample from a small number of community water systems, making the problem particularly well-suited to statistical analysis.

The model developed here is used successfully by Gurian et al. (2001b) in a multi-contaminant RIA evaluating standards for arsenic, uranium and nitrate, with consideration of calcium, magnesium, manganese and sulfate for their aesthetic and treatment performance implications. In the absence of methods for estimating the joint occurrence distributions of these contaminants, a

multi-contaminant RIA would rely on marginal occurrence distributions, which could compromise its validity on several fronts when the contaminants are correlated. The analyses presented here exhibit precisely how the multivariate modeling approach enhances the fit to existing data, the predictive power for future data, and ultimately the inferences for important regulatory quantities. Moreover, the model structure offers solutions to a number of shortcomings in existing water quality modeling strategies, including the lack of rigorous analyses of data collected over a large spatial scale, the tendency to ignore either multivariate or spatial aspects of data, and suboptimal treatment of censored observations.

The remainder of the paper is organized as follows. Section 2 establishes a statistical framework for estimating joint distributions of contaminants based on limited data from community water system raw waters. Section 3 briefly reviews an existing methodology for modeling a single contaminant, and provides an extension of this model to the simultaneous consideration of an arbitrary number of contaminants. Section 4 discusses technical issues regarding the implementation of the multi-contaminant model. Section 5 presents some key applications of the multi-contaminant model, highlighting how jointly modeling contaminants can improve model fit, predictive power, and inferences about key regulatory quantities. Finally, Section 6 discusses how the model addresses some shortcomings of existing approaches, and provides possible extensions.

2. STATISTICAL FRAMEWORK

Suppose that for each of n_0 community water systems is available an observation vector providing raw water concentrations for some number of contaminants. The general goal is to model the joint distribution of p contaminants as a function of system characteristics represented by the covariates $\boldsymbol{x}_0 = (\boldsymbol{x}_{0,1}, \ldots, \boldsymbol{x}_{0,n_0})$, which include auxiliary system information such as location, source-water type and population served. For the i^{th} sampled system, $i = 1, \ldots, n_0$, let $\boldsymbol{Y}_i$ be a p-dimensional vector of the natural logarithms of the concentrations for the p contaminants of interest. Components of these vectors may be left-censored (Section 4.1), and may also be missing, because not all of the p contaminants of interest will be represented by each of the sampled systems. Let $\boldsymbol{Y}_0$ be the entire collection of contaminant vectors, with observed value $\boldsymbol{y}_0$. Let $\boldsymbol{\Theta}$

be a vector of unknown parameters quantifying the dependence of contaminant concentrations on system characteristics $\boldsymbol{x}_0$. Denote the conditional distribution of the contaminant concentrations given $\boldsymbol{\Theta} = \boldsymbol{\theta}$ by $f_{\boldsymbol{Y}_0|\boldsymbol{\Theta},\boldsymbol{X}_0}(\cdot|\boldsymbol{\theta}, \boldsymbol{x}_0)$. To quantify prior uncertainty about the parameter, suppose that $\boldsymbol{\Theta}$ has a prior distribution with density $f_{\boldsymbol{\Theta}|\boldsymbol{\Lambda}}(\cdot|\boldsymbol{\lambda})$, where $\boldsymbol{\Lambda}$ is a vector of hyperparameters which is conditionally independent of the data given $\boldsymbol{\Theta}$, having a prior distribution $f_{\boldsymbol{\Lambda}}(\cdot)$ depending only on fixed hyperparameters. Bayes Rule then allows calculation of the posterior distribution of the model parameters:

$$f_{\boldsymbol{\Theta},\boldsymbol{\Lambda}|\boldsymbol{Y}_0,\boldsymbol{X}_0}(\boldsymbol{\theta}, \boldsymbol{\lambda}|\boldsymbol{y}_0, \boldsymbol{x}_0) \propto f_{\boldsymbol{Y}_0|\boldsymbol{\Theta},\boldsymbol{X}_0}(\boldsymbol{y}_0|\boldsymbol{\theta}, \boldsymbol{x}_0) f_{\boldsymbol{\Theta}|\boldsymbol{\Lambda}}(\boldsymbol{\theta}|\boldsymbol{\lambda}) f_{\boldsymbol{\Lambda}}(\boldsymbol{\lambda}) \tag{1}$$

From this the marginal posterior density of $\boldsymbol{\Theta}$, $f_{\boldsymbol{\Theta}|\boldsymbol{Y}_0,\boldsymbol{X}_0}(\cdot|\boldsymbol{y}_0, \boldsymbol{x}_0)$, can be obtained either by integration or, more typically, with a sample from the joint posterior distribution of all unobserved quantities. The posterior distribution of any function of the parameters, as well as the predictive distributions of any future data, can be calculated stochastically via Markov Chain Monte Carlo (MCMC) methods (Carlin and Louis 2000; Gilks, Richardson, and Spiegelhalter 1996; Gelman, Carlin, Stern, and Rubin 1995) with a sample $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m$ from $f_{\boldsymbol{\Theta}|\boldsymbol{Y}_0,\boldsymbol{X}_0}(\cdot|\boldsymbol{y}_0, \boldsymbol{x}_0)$.

The posterior inferences that are potentially of interest are diverse and numerous. Of obvious importance are the relationships between system characteristics and contaminant levels as well as the covariances among the contaminants. An issue of paramount concern is the use of the model to predict unobserved contaminant concentrations $\boldsymbol{Y}_1$ in a collection of n_1 community water systems with associated covariates $\boldsymbol{x}_1 = (\boldsymbol{x}_{1,1}, \ldots, \boldsymbol{x}_{1,n_1})$. If this set of systems is a subset of the observed data, the predictions can be used for model validation with posterior predictive checking techniques (Gelman, Meng, and Stern 1996; Gilks et al. 1996). The more important case from a regulatory perspective is when the set of systems is the totality of community water systems in the country. These predictions, as well as their associated uncertainties, would provide a valuable input to either single contaminant RIA methods (Frey, Chwirka, Kommineni, Chowdhury, and Narasimhan 2000; Frey, Owen, Chowdhury, Raucher, and Edwards 1998) or multi-contaminant RIAs (Gurian et al. 2001b).

The prediction of contaminant levels in the Bayesian framework is straightforward. Assume that

$\boldsymbol{Y}_1$ is conditionally independent of the observed data $\boldsymbol{Y}_0$ given $\boldsymbol{\Theta}$ with density $f_{\boldsymbol{Y}_1|\boldsymbol{\Theta},\boldsymbol{X}_1}(\cdot|\boldsymbol{\theta},\boldsymbol{x}_1)$. The fundamental distribution of interest is the posterior predictive distribution (Schervish 1995)

$$f_{\boldsymbol{Y}_1|\boldsymbol{X}_1,\boldsymbol{Y}_0,\boldsymbol{X}_0}(\boldsymbol{y}_1|\boldsymbol{x}_1,\boldsymbol{y}_0,\boldsymbol{x}_0) = \int f_{\boldsymbol{Y}_1|\boldsymbol{\Theta},\boldsymbol{X}_1}(\boldsymbol{y}_1|\boldsymbol{\theta},\boldsymbol{x}_1) f_{\boldsymbol{\Theta}|\boldsymbol{Y}_0,\boldsymbol{X}_0}(\boldsymbol{\theta}|\boldsymbol{y}_0,\boldsymbol{x}_0) d\boldsymbol{\theta}. \tag{2}$$

The most practically feasible way to examine the features of this distribution is by sampling from it. This can be achieved by first obtaining a sample $\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_m$ from $f_{\boldsymbol{\Theta}|\boldsymbol{Y}_0,\boldsymbol{X}_0}(\cdot|\boldsymbol{y}_0,\boldsymbol{x}_0)$, and for each i, sampling $\boldsymbol{y}_{1,i}$ from $f_{\boldsymbol{Y}_1|\boldsymbol{\Theta},\boldsymbol{X}_1}(\cdot|\boldsymbol{\theta}_i,\boldsymbol{x}_1)$. Then the sampled concentrations $\boldsymbol{y}_{1,1},\ldots,\boldsymbol{y}_{1,m}$ follow the posterior predictive distribution in Equation (2). At this point any feature of the distribution, such as the number of systems exhibiting concentrations above possible MCLs for some contaminants, can be calculated. Moreover, because the predictive distribution represents integration with respect to the posterior distribution of $\boldsymbol{\Theta}$, it encompasses the uncertainties in contaminant occurrence that result from not knowing the value of $\boldsymbol{\Theta}$. It is precisely these uncertainties that must be recognized and quantified in a RIA, both in the univariate and multivariate case.

3. MODEL DEVELOPMENT

3.1 A Model for a Single Contaminant

A logical point of departure for the process of modeling the joint behavior of contaminants is the development of an effective model for a single contaminant. Such a model for arsenic is presented in Lockwood et al. (2001). In that study, a collection of nineteen models were compared in a cross-validation analysis. One model provided the best available compromise between fit and predictive power given the resolution of the raw water data. This model, summarized in the remainder of this section, is thus taken as the starting point for the development of the multi-contaminant model. The covariates of the model include the source-water type used by the community water system, and the U.S. state in which the system is located. Justifications for these covariates as well as other features of the model structure are given in Lockwood et al. (2001).

The data under consideration are, for $i = 1,\ldots,n_0$ observed community water systems, a single (possibly censored) log contaminant concentration denoted Y_i. Also, let $X(i)$ indicate the U.S. state in which system i is located, and let $W_i = -1$ if the system i is classified as surface water

and $W_i = 1$ if it is classified as ground water. The model is a Bayesian hierarchical model with the following stages:

Stage 0: $Y_i|\mu_i, \sigma_i \sim N(\mu_i, \sigma_i^2)$ independently for $i = 1, \ldots, n_0$,

Stage 1: For $i = 1, \ldots, n_0$,

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

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

The superscripts "(m)" and "(v)" are mnemonic devices for "mean" and "variance". Let $\boldsymbol{\alpha}^{(m)}$ refer to the vector $(\alpha^{(m)}_1, \ldots, \alpha^{(m)}_{50})$, and make a similar definition for each of the other three vectors $\boldsymbol{\alpha}^{(v)}$, $\boldsymbol{\beta}^{(m)}$, $\boldsymbol{\beta}^{(v)}$.

Stage 2: The four vectors in Stage 1 all have a similar prior structure, only one of which is described here with the others being analogous. Conditional on a vector of prior means $\boldsymbol{\alpha}^{(m)}_0$, a scalar prior precision $\tau^{(m)}_\alpha$ and a positive scalar parameter ρ, suppose that $\boldsymbol{\alpha}^{(m)}$ has a 50-dimensional multivariate normal distribution with mean vector $\boldsymbol{\alpha}^{(m)}_0$ and covariance matrix $\boldsymbol{C}(\rho)/\tau^{(m)}_\alpha$. For each value of ρ, the matrix $\boldsymbol{C}(\rho)$ is a non-singular 50-dimensional correlation matrix with entries $c_{p,q} = \exp[-\rho d^2(p, q)]$, 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. 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. 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*.

3.2 Extension to Multiple Contaminants

This section extends the structure of the univariate model in Section 3.1 to modeling the joint raw water occurrence distributions for an arbitrary number p of contaminants. The model attempts to maintain as much of the structure of the univariate model as is practically feasible, while introducing dependence among the contaminants and the parameters governing their distributions. The fundamental problem that must be addressed is that the scalar measurements Y_i from each water system are replaced by p-dimensional vectors $\boldsymbol{Y}_i = (Y_{i,1}, \ldots, Y_{i,p})'$ denoting the vector of natural logarithms of the contaminant concentrations (e.g., in $\mu g/L$), as discussed in Section 2. The covariates under consideration are again the location and source water type of each system. To retain generality, suppose that the systems under consideration are divided into k indexed locations, and let $X(i)$ be the location to which system i is allocated. The location basis is entirely arbitrary and may be based on political boundaries (e.g. U.S. states as in Section 3.1 or counties), geological entities such as watersheds, or in the most extreme case, the systems themselves. Finally, as in the univariate model, let $W_i = -1$ if the system i is classified as a surface-water system and $W_i = 1$ if it classified as a ground-water system.

It is easiest to deal first with the likelihood function by generalizing Stage 0 of the univariate model as $\boldsymbol{Y}_i|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i \sim N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ independently for $i = 1, \ldots, n_0$. Here $\boldsymbol{\mu}_i = (\mu_{i,1}, \ldots, \mu_{i,p})'$ is a vector of means and $\boldsymbol{\Sigma}_i$ is a $(p \times p)$ positive definite symmetric matrix. This will force a fundamental divergence from the structure used in the univariate model, in which it was possible to place the same ANOVA structure on the data means and log precisions. Now that covariation between contaminants is allowed through the matrices $\boldsymbol{\Sigma}_i$, these sets of parameters must be treated differently. We next consider possible prior structures for each of these sets of parameters.

Prior structure of data means: Generalizing Equation (3) in Stage 1 of the univariate model to handle the vector mean parameters $\boldsymbol{\mu}_i$ is straightforward. Suppose that for $i = 1, \ldots, n_0$, $\boldsymbol{\mu}_i = \boldsymbol{\alpha}^{(m)}_{X(i)} + \boldsymbol{\beta}^{(m)}_{X(i)} W_i$. Here, unlike in the univariate model, the terms $\boldsymbol{\alpha}^{(m)}_{X(i)}$ and $\boldsymbol{\beta}^{(m)}_{X(i)}$ are not scalars, but vectors of length p giving main effects and source-type adjustments for each contaminant in location $X(i)$. Some difficulty arises in the specification of the joint prior distribution for each

of these sets of k vectors. Focus now on the collection $\{\boldsymbol{\alpha}_j^{(m)}\}$ for $j = 1, \ldots, k$, the discussion for $\{\boldsymbol{\beta}_j^{(m)}\}$ being analogous. Define $\boldsymbol{\alpha}^{(m)}$ to be the $(k \times p)$ matrix with rows $\boldsymbol{\alpha}_j^{(m)}$ for $j = 1, \ldots, k$, and thus $\text{vec}(\boldsymbol{\alpha}^{(m)})$ refers to the kp-dimensional vector with effects organized in p contaminant blocks. The fundamental challenge is in the specification of a positive-definite $(k \times p) \times (k \times p)$ covariance structure for $\boldsymbol{\alpha}^{(m)}$. Independent replicates of the univariate model would have a block diagonal covariance matrix, the blocks being the spatial correlation matrices $\boldsymbol{C}(\rho_i)$ for $i = 1, \ldots, p$. This structure forces prior independence among parameters for different contaminants. Prior dependence among parameters for different contaminants, both within a location and across locations, can be accomplished by replacing zero entries in this covariance matrix by non-zero values. This task, however, is not easy because the resulting matrix must be ensured to be positive semi-definite. In the most general case, where different contaminants are allowed to have different values of ρ, the problem approaches intractability because of the complex constraints this places on all of the elements off of the block diagonal. This is clearly a place where additional research would prove valuable, because this sort of problem arises quite naturally in most multivariate analyses and has been identified as one of the more challenging aspects of modeling complex environmental data (Cressie 2000; Berliner 2000). For example, Woodard et al. (2000) develop a more flexible structure for modeling the joint spatial covariances of two random quantities, but the problem for higher dimensions persists. This paper does not address this issue further because the solution to be discussed (which constrains the value ρ to be the same for all p contaminants) places only modest limitations on the complexity of the models that can be considered.

Assuming that the contaminants share a common value of ρ, which is tantamount to specifying that the correlation structure of effects across locations within each contaminant is the same for all contaminants, progress can be made via a multivariate Gaussian process prior distribution with Kronecker covariance structure. Let $\boldsymbol{C}(\rho)$ be a $(k \times k)$ correlation matrix with entries based on location distances as before, and let $\boldsymbol{\gamma}$ be a $(p \times p)$ positive definite symmetric matrix. It is then possible to model the covariance structure of $\boldsymbol{\alpha}^{(m)}$ as $\text{Cov}(\text{vec}(\boldsymbol{\alpha}^{(m)})) = \boldsymbol{C}(\rho) \otimes \boldsymbol{\gamma}$. $\boldsymbol{\gamma}$ gives the prior covariance across contaminants of all of the effects within the same location, and this is assumed to be the same for all locations. $\boldsymbol{C}(\rho)$ gives the correlation structure of the effects across locations for

all the effects for the same contaminant, and this is assumed to be the same for all contaminants. As long as the two component matrices are positive semi-definite, their Kronecker product is also positive semi-definite, a guarantee not afforded by more general structures. Note that if $\boldsymbol{\gamma}$ is a diagonal matrix, the parameters for different contaminants are independent and the structure used in the univariate model is recovered (with the additional constraint that a common ρ is shared by all contaminants, a constraint unnecessary if $\boldsymbol{\gamma}$ is diagonal). Hence, while forcing all contaminants to share a common value of ρ results in some loss of generality, it does allow prior dependence between contaminant parameters both within locations and across locations in a way that results in a valid positive definite covariance matrix. Because $\boldsymbol{\rho}$ is a hyperparameter which impacts only the form of prior distributions of the parameters of interest and not the parameters themselves, the loss of generality does not seriously limit the resulting model for the joint distributions of multiple contaminants.

Prior structure of data covariances: As mentioned previously, modeling the data covariance structure is more complicated than the extension possible in the mean structures because it is necessary to place distributions on the positive definite matrices $\boldsymbol{\Sigma}_i$ as opposed to just scalar variance terms. In order to organize the discussion, this section classifies the approaches to the problem as either "direct" or "decomposition". In the direct approach, the matrices $\boldsymbol{\Sigma}_i$ (or, to mimic the previous models, $\boldsymbol{\Sigma}_i^{-1}$) are modeled directly with probability distributions over the space of positive definite matrices. A convenient choice for this distribution that has been employed often in Bayesian multivariate analysis problems is the Wishart distribution (Gelman et al. 1995). If one is willing to sacrifice some generality and assume, for example, that the covariance structure depends only on source-water type and not location, then one possible approach would be to model the surface water and ground water covariance matrices with independent Wishart prior distributions.

On the other hand, if one wishes to allow each location and source-water type to have a different covariance structure, a feature necessary to attain the most effective model in the univariate arsenic analysis, then directly modeling the covariance matrices becomes far more complicated. The fundamental difficulty is that this approach results in $2k$ matrices representing a total of $kp(p+1)$

parameters. This number of nominal parameters could be justified if enough data were available from each location and source-type combination to estimate them adequately or if the effective number of parameters could be reduced by forcing relationships between these parameters beyond whatever restrictions are imposed by the positive definiteness constraint. The former solution of relying on sufficient data to inform the large number of parameters is not viable because even with the relatively coarse location delineations based on U.S. states, some states have little or no raw water data available. This entails the need for some mechanism for reducing the effective number of parameters by inducing relationships among the covariance matrices for different locations and source types. For example, analogous to the role of the correlation matrices $\boldsymbol{C}(\rho)$ in the univariate case, geographical proximity could serve as both a pragmatic and physically justifiable mechanism for forming relationships among covariance matrices of neighboring locations. Unfortunately this problem does not lend itself to a ready solution in the multivariate setting. For example, it is not clear what kind of prior distribution for the matrices $\boldsymbol{\Sigma}_i$ would have the property that the covariance matrices for neighboring locations are more similar to one another than those for more separated locations. Moreover, there is no clear multivariate analog to the ANOVA structure of correlated parameters used for the scalar data precision terms in the univariate model because of the restriction of positive definiteness. It is not possible to subtract matrix "effects" and still ensure positive definiteness of the resulting matrix, eliminating from consideration structures such as $\boldsymbol{\Sigma}_i = \boldsymbol{\Theta}_{X(i)} + \boldsymbol{\Lambda}_{X(i)} W_i$. If one could somehow coerce the various matrix "effects" such that only positive definite matrices were added together, there would still be the substantive concern that matrices would continue to get "larger" as each effect is added. In summary, if one wants to allow each location and source water type to have its own covariance matrix, and if these matrices are to be related to one another geographically, then methods for directly modeling the matrices $\boldsymbol{\Sigma}_i$ are not apparent. Of course, that is not to say that no solution exists, and the problem raises a potentially fruitful avenue for future research.

If the focus is switched to modeling some decomposition of the covariance matrices $\boldsymbol{\Sigma}_i$ rather than the matrices themselves, more progress is possible. The particular decomposition addressed here separates the treatment of inter-contaminant correlation and intra-contaminant variation and

is discussed by Barnard et al. (2000). This decomposes the covariance matrices as

$$\boldsymbol{\Sigma}_i = \boldsymbol{D}_i \boldsymbol{R}_i \boldsymbol{D}_i \tag{5}$$

where $\boldsymbol{D}_i = \text{diag}(\sigma_{i,1}, \ldots, \sigma_{i,p})$, a diagonal matrix of contaminant standard deviations for the i^{th} system and $\boldsymbol{R}_i$ is a correlation matrix. Such a structure is particularly useful in a Bayesian context where the prior information available about the coordinate standard deviations is greater than that available for correlation matrices (Barnard et al. 2000), which is arguably the case with the problem at hand. Relatively informative prior information about individual contaminant variation is known, while the paucity of published analyses of joint contaminant occurrence (especially after controlling for location effects) makes prior information about co-occurrence more elusive. Moreover, the structure is practically feasible and has been used successfully in applications (Barnard et al. 2000; Brav 2000; Boatwright, McCulloch, and Rossi 1999).

The generality of allowing $\boldsymbol{R}_i$ to depend on both source type and location results in a problem no simpler than what was faced while attempting to model the matrices $\boldsymbol{\Sigma}_i$ directly. However, under the assumption that the correlation structure depends only on source type, the number of parameters under consideration changes from $kp(p+1)$ to $2kp + p(p+1)$. This results in substantial parsimony, especially for large k and p. Moreover, because the decomposition separates the scalar standard deviations from the correlations, the log precisions $-2\log(\sigma_{i,j})$ can be modeled in precisely the same manner by which the mean parameters are modeled (i.e., with the matrix normal distribution). This introduces both spatial and source-type dependence in the data variance structure, extending the structure successfully implemented in the univariate case. On the other hand, the correlation matrices $\boldsymbol{R}_{-1}$ for surface water systems and $\boldsymbol{R}_1$ for ground water systems are assumed to be the same for all locations. While this assumption is probably more restrictive than is justified by the complexity of the underlying phenomenon, the decision is practically sound. The separation of the treatment of the contaminant variances from their covariances balances the desire to make use of geographical proximity in modeling the data covariance structure with the need to limit the number of parameters to a tractable amount. Of course, a more flexible model would allow the correlation matrices to depend on location as well as source type, but

have geographical dependence among matrices in different locations. This would keep the effective number of parameters manageable while maintaining the generality of the model. Unfortunately, as is the case with the covariance matrices $\boldsymbol{\Sigma}_i$, it is not clear what form of prior distribution would be appropriate for introducing such spatial relationships between the matrices.

Other decompositions of the data covariance matrices are available. For example, one could use the orthogonal diagonalization decomposition $\boldsymbol{\Sigma}_i = \boldsymbol{Q}_i \boldsymbol{T}_i \boldsymbol{Q}_i'$ where $\boldsymbol{Q}_i$ is orthogonal with columns equal to the eigenvectors of $\boldsymbol{\Sigma}_i$ and $\boldsymbol{T}_i$ is a diagonal matrix of the eigenvalues of $\boldsymbol{\Sigma}_i$. This decomposition has mathematical appeal but is not nearly as intuitive as that in Equation (5). The eigenvalues of $\boldsymbol{\Sigma}_i$ are variances of linear combinations of the components of the data vector rather than variances of the components themselves, and this would be difficult to reconcile with the ANOVA structure for the data component variances used in the arsenic analysis. Similar concerns pertain to the Cholesky decomposition $\boldsymbol{\Sigma}_i = \boldsymbol{L}_i \boldsymbol{L}_i'$ for $\boldsymbol{L}_i$ a lower triangular matrix (Pinheiro and Bates 1996) and the other decompositions summarized in Barnard et al. (2000). The remainder of this paper considers only the standard deviation and correlation decomposition in Equation (5).

Distributions over the space of correlation matrices: Probably the most difficult aspect of using the decomposition in Equation (5) is the formulation of a prior distribution for the correlation matrices. The space $\mathcal{R}$ of positive definite correlation matrices is as complicated as the space of general positive definite matrices, with the additional constraint that the diagonal elements of the matrices are 1. This section discusses a number of options for distributions over this space and considers the relative merits of each.

Barnard et al. (2000) deal only with the case where prior information about the correlation matrices under consideration is non-informative, and they suggest two different prior distributions that could reasonably be considered as reference priors. The first is derived from examination of the marginal distribution of the correlation matrix from a matrix which has a Wishart distribution, and has the property that the marginal distribution of each correlation coefficient is uniform on $(-1, 1)$. The disadvantage of this approach is that the joint distribution of subsets of the elements of the matrix are not nearly as intuitive. The other is a uniform distribution over the space of

correlation matrices, with density $p(\boldsymbol{R}) = cI_{\mathcal{R}}(\boldsymbol{R})$ for some normalizing constant c. Because of the compactness of the space, such a distribution is proper. Unlike the marginal distribution of the Wishart matrix, the marginal distributions of the individual matrix elements are not uniform. Instead, because of the positive definiteness constraint, the marginal distributions are concentrated near zero, with more concentration as the dimension increases. Barnard et al. (2000) examine this phenomenon in detail. For the dimension $p = 3$, the degree of concentration is minimal, while for $p = 10$, the concentration places very low density on only nearly singular matrices. In a simulation study designed to examine the influence of these prior distributions, Barnard et al. (2000) conclude that the "informativeness may have a tolerable impact on the marginal posteriors, as long as [the dimensionality] is not too large relative to the amount of data." Of course, this depends on the data and likelihood of whatever model is under consideration. In the current application, if the modeler is content to allow the correlation matrices to be in large part determined by the data, this goal will be realized for small to moderate numbers of contaminants ($p \leq 10$). For p very large, $p(\boldsymbol{R}) = cI_{\mathcal{R}}(\boldsymbol{R})$ will probably not provide the requisite flexibility, and less restrictive prior structures will be necessary.

Neither of these possibilities for modeling the correlation matrices is entirely satisfactory because it is not possible to manifest any substantive prior information about the correlation matrices. However, they each have some intuitive appeal from the perspective of a researcher desiring a non-informative prior. From a practical standpoint, the uniform distribution $p(\boldsymbol{R}) = cI_{\mathcal{R}}(\boldsymbol{R})$ is the more attractive of the two because the density is bounded even near the boundary of the parameter space. This precludes the kinds of MCMC convergence problems that arise from the use of unbounded density functions. In order to make progress with the underlying problem, the remainder of this paper uses $p(\boldsymbol{R}) = cI_{\mathcal{R}}(\boldsymbol{R})$ in all analyses. However, future applications may benefit from a more flexible prior structure that could be informed more effectively by expert opinion, and such work is underway (Garthwaite and Al-Awadhi 2001).

3.3 Model formulation

The discussions of the previous section culminate in the formulation of a model which incorporates the bulk of the flexibility necessary to model inter-contaminant dependence while maintaining mathematical tractability and conciseness. The stages are as follows.

Stage 0: Independently for $i = 1, \ldots, n_0$,

$$\boldsymbol{Y}_i | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i \sim N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{6}$$

Stage 1: For $i = 1, \ldots, n_0$,

- *Mean structure:*

$$\boldsymbol{\mu}_i = \boldsymbol{\alpha}^{(m)}_{X(i)} + \boldsymbol{\beta}^{(m)}_{X(i)} W_i \tag{7}$$

- *Variance structure:* From Equation (5), write

$$\boldsymbol{\Sigma}_i = \boldsymbol{D}_i \boldsymbol{R}_{W_i} \boldsymbol{D}_i \tag{8}$$

  where $\boldsymbol{R}_{W_i}$ is a source-type specific correlation matrix and $\boldsymbol{D}_i$ is a diagonal matrix whose elements are the square roots of the diagonal elements of $\boldsymbol{\Sigma}_i$. Let $\boldsymbol{\sigma}_i$ be the vector of these diagonal elements, and for a vector $\boldsymbol{x}$, let $f(\boldsymbol{x}) = (f(x_1), \ldots, f(x_p))$. Then suppose

$$-2 \log \boldsymbol{\sigma}_i = \boldsymbol{\alpha}^{(v)}_{X(i)} + \boldsymbol{\beta}^{(v)}_{X(i)} W_i \tag{9}$$

Now let

$$\boldsymbol{\alpha}^{(m)} = \begin{bmatrix} \alpha^{(m)}_{1,1} & \cdots & \alpha^{(m)}_{1,p} \\ \vdots & \ddots & \vdots \\ \alpha^{(m)}_{k,1} & \cdots & \alpha^{(m)}_{k,p} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\alpha}^{(m)\prime}_1 \\ \vdots \\ \boldsymbol{\alpha}^{(m)\prime}_k \end{bmatrix} \tag{10}$$

Also organize the other three sets of parameters analogously. Note that this stage introduces no additional randomness; it merely re-expresses the mean and covariance structures of the data as deterministic functions of parameters which can be modeled more easily.

Stage 2a: Each of the four $(k \times p)$ matrices $\boldsymbol{\alpha}^{(m)}$, $\boldsymbol{\alpha}^{(v)}$, $\boldsymbol{\beta}^{(m)}$ and $\boldsymbol{\beta}^{(v)}$ have a similar Gaussian process prior structure based on the matrix normal distribution, one of which is described here. Conditional on the scalar parameter $\rho_{\alpha^{(m)}}$, a $(p \times p)$ positive definite matrix $\boldsymbol{\gamma}_{\alpha^{(m)}}$, and a fixed $(k \times p)$ prior mean matrix $\boldsymbol{\alpha}_0^{(m)}$, suppose that the $(k \times p)$ matrix $\boldsymbol{\alpha}^{(m)}$ has a matrix normal distribution:

$$\boldsymbol{\alpha}^{(m)} | \boldsymbol{\alpha}_0^{(m)}, \rho_{\alpha^{(m)}}, \boldsymbol{\gamma}_{\alpha^{(m)}}, \sim N_{(k \times p)} \left(\boldsymbol{\alpha}_0^{(m)}, \boldsymbol{C}(\rho_{\alpha^{(m)}}) \otimes \boldsymbol{\gamma}_{\alpha^{(m)}} \right) \tag{11}$$

Remark: Analogously to the univariate model, it is possible replace the parameters $\rho_{\alpha^{(m)}}$, $\rho_{\alpha^{(v)}}$, $\rho_{\beta^{(m)}}$, and $\rho_{\beta^{(v)}}$ with an omnibus parameter ρ which is the same for all 4 matrices. However, the structure is written here in terms of the more general case, where such a reduction would be a special case of the general model.

Stage 2b: Assume that the two source-specific correlation matrices $\boldsymbol{R}_1$ and $\boldsymbol{R}_{-1}$ are independent with a uniform prior distribution over the space of positive definite correlation matrices.

Stage 3: Suppose the logarithms of the four parameters $\rho_{\alpha^{(m)}}$, $\rho_{\alpha^{(v)}}$, $\rho_{\beta^{(m)}}$, $\rho_{\beta^{(v)}}$ are iid univariate normal conditional on scalar parameters μ_ρ (mean) and ξ_ρ (standard deviation). Finally, suppose that the four matrix parameters $\boldsymbol{\gamma}_{\alpha^{(m)}}$, $\boldsymbol{\gamma}_{\alpha^{(v)}}$, $\boldsymbol{\gamma}_{\beta^{(m)}}$, $\boldsymbol{\gamma}_{\beta^{(v)}}$ are iid Wishart conditional on a scalar degrees of freedom parameter a_γ and a fixed scale matrix parameter $\boldsymbol{b}_\gamma$.

This formulation leads to the following joint distribution of the data and parameters:

$$\begin{aligned}
& \left[\prod_{i=1}^{n_0} p_0(\boldsymbol{y}_i | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \right] \qquad \text{(likelihood function)} \\
\times \quad & p_1 \left(\boldsymbol{\alpha}^{(m)} | \boldsymbol{\alpha}_0^{(m)}, \rho_{\alpha^{(m)}}, \boldsymbol{\gamma}_{\alpha^{(m)}} \right) p_1 \left(\boldsymbol{\alpha}^{(v)} | \boldsymbol{\alpha}_0^{(v)}, \rho_{\alpha^{(v)}}, \boldsymbol{\gamma}_{\alpha^{(v)}} \right) \\
\times \quad & p_1 \left(\boldsymbol{\beta}^{(m)} | \boldsymbol{\beta}_0^{(m)}, \rho_{\beta^{(m)}}, \boldsymbol{\gamma}_{\beta^{(m)}} \right) p_1 \left(\boldsymbol{\beta}^{(v)} | \boldsymbol{\beta}_0^{(v)}, \rho_{\beta^{(v)}}, \boldsymbol{\gamma}_{\beta^{(v)}} \right) \\
\times \quad & p_2(\boldsymbol{R}_1) p_2(\boldsymbol{R}_{-1}) \\
\times \quad & p_3(\log \rho_{\alpha^{(m)}} | \mu_\rho, \xi_\rho) p_3(\log \rho_{\alpha^{(v)}} | \mu_\rho, \xi_\rho) p_3(\log \rho_{\beta^{(m)}} | \mu_\rho, \xi_\rho) p_3(\log \rho_{\beta^{(v)}} | \mu_\rho, \xi_\rho) \\
\times \quad & p_4(\boldsymbol{\gamma}_{\alpha^{(m)}} | a_\gamma, \boldsymbol{b}_\gamma) p_4(\boldsymbol{\gamma}_{\alpha^{(v)}} | a_\gamma, \boldsymbol{b}_\gamma) p_4(\boldsymbol{\gamma}_{\beta^{(m)}} | a_\gamma, \boldsymbol{b}_\gamma) p_4(\boldsymbol{\gamma}_{\beta^{(v)}} | a_\gamma, \boldsymbol{b}_\gamma)
\end{aligned} \tag{12}$$

where p_1 is the matrix normal density, p_2 is the uniform density over the space of positive definite correlation matrices, p_3 is the univariate normal density, and p_4 is the Wishart density.

4. IMPLEMENTATION ISSUES

As is the case in all Bayesian models, the posterior distribution of the model parameters is, up to a constant, the same as the joint distribution of the data and parameters viewing the data as fixed. In order to make inferences about the parameters from this joint distribution, MCMC methods can be used to generate a large sample with which to learn about expected values of functions of the parameters. The complexity of the data and the model in the current study necessitate some more specialized considerations, which are discussed in the subsections that follow.

4.1 Censoring and data augmentation

A common problem in quantifying levels of trace substances is that contaminant concentrations cannot be quantified unless they are sufficiently high, resulting in left-censoring of some contaminant measurements. A nice summary of this issue is given in Piegorsch et al. (1998). The net result is that most observations $\boldsymbol{y}_i$ will not be a p-dimensional vector of observed concentrations, but a mixture of observed concentrations and missing components. In the case of censored observations, the missing components will contain the information that the corresponding measurement was less than some censoring limit which in general may be a function of both the contaminant and the observation vector. Moreover, the algorithm must allow for components that are missing due to an unmeasured contaminant. An unmeasured coordinate is equivalent to a coordinate censored at positive infinity, so that an appropriate analytical technique for handling the censored observations suffices for the unmeasured ones as well. As such, the remainder of the discussion refers to both censored and unmeasured coordinates as "missing."

In the arsenic analysis (Lockwood et al. 2001), missing observations are handled with a data augmentation technique (Dyk and Meng 2001; Gelman et al. 1995; Tanner and Wong 1987), which imputes missing observations at each iteration of the MCMC algorithm. In particular, before updating any of the parameters for the next iteration, the algorithm simulates observations for the missing values conditional on the parameters from the last stage and possibly on the fact that the observations were known to be less than some concentration. Then, based on the imputed complete data, it generates the next set of parameters.

For the multiple contaminant case, the same general data augmentation approach is possible, but the technique requires sampling from a truncated multivariate conditional distribution. The proper conditional distribution can be formed in two steps, the first isolating the missing components and the second accounting for the censoring of these components where applicable. Step one uses the formula for the conditional distribution of the missing coordinates given the observed coordinates, available in closed form because the entire vector has a multivariate normal distribution (Morrison 1990). The additional knowledge about censoring, where applicable, must now be incorporated into the conditional distribution of the missing components given the others. This results in a multivariate distribution which is truncated in some of the dimensions. There are a number of methods available for sampling from this distribution, including rejection sampling, introduction of latent variables (Damien and Walker 2001), the method of Geweke (1991), and the "one for one" algorithm of Gelfand et al. (1992). This final method, a multivariate generalization of the typical inverse CDF method, has proven successful in applications (Boatwright et al. 1999; McCulloch and Rossi 1994) and was used for all results presented here. The method reduces the problem of imputation in the multivariate setting to a series of univariate imputations, where a clear solution to sampling from the distribution exists whether the missing value is censored or not. It proceeds as follows. First, find the conditional distribution of the first missing component given all of the observed components and then sample from this (possibly truncated) univariate distribution. This step is not difficult because sampling from a univariate truncated distribution is easily accomplished with the inverse CDF method. Then treat this imputed value as observed and sample from the conditional distribution of the next missing component given the new set of observed components. Repeating in this manner until all missing components are imputed results in a realization that has the desired truncated conditional distribution.

4.2 Reparameterization

As discussed in the arsenic analysis in Lockwood et al. (2001), the spatial correlation $\boldsymbol{C}(\rho)$ in the prior distribution for $\boldsymbol{\alpha}^{(m)}$ (a k-dimensional vector in the univariate case) necessitates special consideration when attempting to update elements of the vector $\boldsymbol{\alpha}^{(m)}$. The primary difficulty is that

small values of ρ imply high correlations among the elements, causing the chain to move inefficiently through the parameter space. A solution to this problem, discussed in detail in Lockwood et al. (2001), 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

$$\boldsymbol{\alpha}^{(m)} = \boldsymbol{\alpha}_0^{(m)} + \frac{1}{\sqrt{\tau_\alpha^{(m)}}} \boldsymbol{L}(\rho)\boldsymbol{\theta}_{\alpha^{(m)}}, \tag{13}$$

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. In addition, if ρ is small enough that the matrix is numerically singular, an appropriate modification to the matrix $\boldsymbol{L}(\rho)$ is possible. This is also discussed in detail in Lockwood et al. (2001).

Clearly an analogous and potentially more severe situation occurs with the random matrices of location and contaminant effects when $p > 1$. For concreteness, focus on the $(k \times p)$ matrix $\boldsymbol{\alpha}^{(m)}$ of the multi-contaminant model. Let $\boldsymbol{\theta}_{\alpha^{(m)}}$ be a $(k \times p)$ random matrix whose elements are iid standard normal. Anticipating the transformation of this matrix to another with the same distribution as $\boldsymbol{\alpha}^{(m)}$, note that the distribution of $\boldsymbol{\theta}_{\alpha^{(m)}}$ is a matrix normal distribution with mean matrix $\mathbf{0}$, a $(k \times p)$ matrix of zeros, and with covariance matrix $\boldsymbol{I}_k \otimes \boldsymbol{I}_p$. Now perform the lower-triangular factorization $\boldsymbol{C}(\rho_{\alpha^{(m)}}) = \boldsymbol{L}(\rho_{\alpha^{(m)}})\boldsymbol{L}'(\rho_{\alpha^{(m)}})$ and $\boldsymbol{\gamma_{\alpha^{(m)}}} = \boldsymbol{L}(\boldsymbol{\gamma}_{\alpha^{(m)}})\boldsymbol{L}'(\boldsymbol{\gamma}_{\alpha^{(m)}})$, and appeal to a result concerning linear functions of normal random matrices. Let $\boldsymbol{X}$ be a $(k \times p)$ normal random matrix with mean matrix $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\sigma} \otimes \boldsymbol{\nu}$. Then for fixed real matrices $\boldsymbol{a}$ and $\boldsymbol{b}$ of dimensions $(a \times k)$ and $(b \times p)$ respectively, it can be shown that $\boldsymbol{aXb'}$ has a matrix normal distribution (of dimension $(a \times b)$) with mean matrix $\boldsymbol{a\mu b'}$ and covariance matrix $(\boldsymbol{a\sigma a'}) \otimes (\boldsymbol{b\nu b'})$. Now suppose that $\boldsymbol{\theta}_{\alpha^{(m)}}$ is related to $\boldsymbol{\alpha}^{(m)}$ by

$$\boldsymbol{\alpha}^{(m)} = \boldsymbol{\alpha}_0^{(m)} + \boldsymbol{L}(\rho_{\alpha^{(m)}})\boldsymbol{\theta}_{\alpha^{(m)}}\boldsymbol{L}'(\boldsymbol{\gamma}_{\alpha^{(m)}}) \tag{14}$$

It is immediate from the result about linear transformations of normal random matrices that the conditional distribution of $\boldsymbol{\alpha}^{(m)}$ given $\boldsymbol{\alpha}_0^{(m)}$, $\rho_{\alpha^{(m)}}$ and $\boldsymbol{\gamma}_{\alpha^{(m)}}$ is the same as that posited in Equation (11). This is the multivariate analog to the transformation in Equation (13) and achieves identical benefits by treating $\boldsymbol{\theta}_{\alpha^{(m)}}$ as the parameter. Finally, introduce parameters $\boldsymbol{\theta}_{\alpha^{(v)}}, \boldsymbol{\theta}_{\beta^{(m)}}$ and $\boldsymbol{\theta}_{\beta^{(v)}}$ which are related to $\boldsymbol{\alpha}^{(v)}, \boldsymbol{\beta}^{(m)}$ and $\boldsymbol{\beta}^{(v)}$ in a manner given by Equation (14). These reparameterizations, similar to those developed by Rue (2001) for sampling Markov random fields, greatly improve the performance of the algorithm used to sample from the posterior distribution. It is important to note, however, that the joint distribution of the data and parameters specified in Equation (12) changes in light of the reparameterization. In the original formulation, the data are conditionally independent of the ρ and $\boldsymbol{\gamma}$ hyperparameters given the $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{R}$ parameters. Hence when updating the ρ and $\boldsymbol{\gamma}$ parameters, it is not necessary to evaluate the likelihood function because it is constant with respect to the conditional posterior distribution of these parameters given the others. However, when treating the $\boldsymbol{\theta}$ quantities as the parameters, the likelihood function is implicitly a function of all model parameters via the relationship in Equation (14) and its analogs for the other three matrices of parameters. That is, the data are not conditionally independent of the ρ and $\boldsymbol{\gamma}$ parameters given the $\boldsymbol{\theta}$ and $\boldsymbol{R}$ parameters. The net result is that the Metropolis-Hastings algorithm applied to updating any of the model parameters must take the likelihood function into account, forcing a greater number of likelihood evaluations per iteration of the Markov chain. Hence, while the reparameterization is necessary to produce a viable MCMC algorithm, each iteration of the modified chain will be somewhat more computationally intensive.

4.3 Updating positive definite matrix parameters

The $\boldsymbol{\gamma}$ matrices and the data correlation matrices $\boldsymbol{R}_{-1}$ and $\boldsymbol{R}_1$ are constrained to be positive definite, and thus warrant special consideration in the MCMC setting. The simpler case regards the four $\boldsymbol{\gamma}$ matrices, for which updating was performed on the individual elements of the Cholesky decomposition $\boldsymbol{L} = ((\ell_{ij}))$ of each $\boldsymbol{\gamma}$. We used lognormal proposal distributions for the diagonal elements ℓ_{ii}, and normal proposal distributions for all subdiagonal elements. Because the transformation from $\boldsymbol{\gamma}$ to $\boldsymbol{L}$ has a relatively simple Jacobian of $2^p \prod_{i=1}^{p} \ell_{ii}^{p-i+1}$ (Olkin 1953), the density

for $\boldsymbol{L}$ is available in closed form. This updating scheme provided superb mixing properties in all applications.

Developing an effective updating strategy for the data correlation matrices $\boldsymbol{R}_{-1}$ and $\boldsymbol{R}_1$ presents somewhat more difficulty. The discussion that follows uses "$\boldsymbol{R}$" to refer to either of these matrices. The primary problem is that because the diagonal elements of $\boldsymbol{R}$ must be one, the subspace of positive definite matrices that must be navigated, $\mathcal{C}$, is more complicated than the space of all positive definite matrices. For example, the Cholesky decomposition strategy used for the $\boldsymbol{\gamma}$ matrices cannot be used in the same way because the Cholesky decomposition for a correlation matrix has rows which are constrained to have Euclidean length less than (or, in case of positive semi-definiteness, equal to) one. This not only disqualifies the use of a normal proposal distribution for the subdiagonal elements, but also makes it impossible to consider the Cholesky elements in isolation of one another. The method used here, a modification of the algorithm first presented by by Barnard et al. (2000), directly updates $\boldsymbol{R}$ one element at a time. The method is based on the fact that conditional on all of the other elements of $\boldsymbol{R}$, the range of allowable values for a particular element r_{ij} based on the constraint of positive definiteness is a subinterval of $(-1, 1)$ whose endpoints are available in closed form. The precise equations of the interval boundaries are based on determinants of a matrix $\boldsymbol{R}^*$ obtained by varying only r_{ij} and are detailed in their paper. The net result is that updating $\boldsymbol{R}$ proceeds one element at a time, with proposed values being drawn from some distribution over the allowable range for that particular matrix element.

In their implementation, Barnard et al. (2000) used the so-called "griddy" Gibbs sampler (Ritter and Tanner 1992), which essentially uses a discrete approximation to the marginal posterior distribution of each $r_{i,j}$ as a proposal distribution. Because the models presented here were part of an extensive cross-validation study involving multiple data sets, the model was fit to a relatively large collection of different observation sets. Thus, a method not requiring approximations to the various marginal posterior distributions is preferred. An easy alternative is a uniform distribution over the allowable interval for the element at hand. However, this proposal distribution is not centered at the current value, cannot be changed in any way to optimize the mixing of the chain, and places non-negligible probability of values near the boundary of positive definiteness. This

results in numerically singular proposed correlation matrices, which presents additional computational difficulties. A more flexible solution is to alter the proposal distribution to one centered in some manner at the current value and which has vanishingly small probability near the boundaries. The strategy pursued here is to use a normal proposal distribution on a transformation of the correlation coefficient which produces the desired result. Let $(a, b) \subset (-1, 1)$, $a < b$, be the allowable range of a particular correlation element r conditional on the rest of the matrix. We construct a proposal distribution $f_R^*(\cdot|r_0)$ over (a, b) which is centered at the current value r_0 as follows. Consider the transformation $y = g(r) = \tan(\pi((r-a)/(b-a) - 0.5))$ which maps (a, b) to $(-\infty, +\infty)$ and which has inverse transformation $r = g^{-1}(y) = a + (b-a)\left(\tan^{-1}(y)/\pi + 0.5\right)$. Then, if the current value of a correlation R is r_0, then the suggested distribution of the proposed value is that of $g^{-1}(Y)$ where $Y \sim N(g(r_0), \sigma^2)$. The density of this distribution is available by the usual change-of-variables formula. Because g is strictly monotonic and continuous on (a, b), g^{-1} is as well. Hence quantiles are preserved and the median of the proposal distribution is r_0, providing the desired centering at the current value. This method proved successful in all applications, examples of which are discussed next.

5. APPLICATION

The primary motivation of jointly modeling contaminants is to exploit relationships among the contaminants to achieve a better descriptor of the underlying phenomena generating the data. This section highlights some applications of the model presented in Section 3, showing how in the presence of contaminant covariations, the model provides better fit and predictive power than modeling contaminants individually. In addition, this section discusses how inferences about important regulatory quantities can depend on which modeling approach is used. The purpose of this section is not to provide a comprehensive analysis of the available data, but rather to present key results highlighting the benefits of joint contaminant models to the underlying policy objective of developing a multi-contaminant RIA.

The data used in these applications originate from the same two data sets discussed in Lockwood et al. (2001), to which a lengthier discussion is deferred. The first, derived from the National Arsenic

Occurrence Survey (NAOS) (Frey, Edwards, and Amy 1997; Frey and Edwards 1997), provides raw water concentrations for approximately 25 substances from nearly 500 community water systems from across the U.S. The other data were provided by the United States Geological Survey (USGS) as part of the National Water Information System (Focazio, Welch, Watkins, Helsel, and Horn 2000). These data provide analytical results for approximately 15 substances from nearly 2000 community water systems using groundwater.

The applications presented here focus on two pairs of contaminants, the first being arsenic (As) and sulfate (SO_4). Unlike arsenic, sulfate is neither carcinogenic nor acutely toxic. However, water with high sulfate concentrations is aesthetically bothersome because of a salty taste, pungent odor and mild laxative effect (Backer 2000; EPA 1999). The joint occurrence of arsenic and sulfate is most interesting from a water treatment perspective. Both arsenic and sulfate are negatively charged particles, and hence are removed by any treatment whose mode action derives from particle charge (e.g. anion exchange). This is a benefit to systems designed to remove both contaminants, but can be a detriment to a system concerned with only one or the other of the substances. In either case, properly accounting for occurrence covariation between arsenic and sulfate would be an integral part of the estimation of treatment costs and their uncertainties in a multi-contaminant RIA. The other pair of contaminants under consideration is magnesium (Mg) and calcium (Ca), which together comprise what is commonly termed "water hardness." Their typically high degree of covariation, along with their inherently joint interest from a treatment perspective, motivates an isolated study of these contaminants. All four constituents are reported at most sites in both the NAOS and USGS data sets.

This section presents comparisons of two different models. The first, called "Model I" for "Independence Model," is the model in Section 3.3 fit separately (e.g., with $p = 1$) to the contaminants under consideration, taking $k = 50$ corresponding to the U.S. states. This is essentially equivalent to fitting the arsenic model in Lockwood et al. (2001) independently to each contaminant. The other model, called "Model J" for "Joint Model," is the full multi-contaminant model of Section 3.3, again with $k = 50$. Specific details regarding prior distributions and convergence diagnostics can be found in Lockwood (2001). In short, the results of this section were quite robust to different sensible

prior specifications, which were based on substantive information about the marginal distributions of the contaminants under consideration. In all cases, prior distributions for correlation matrices are uniform as discussed in Section 3.2, and a common value of ρ was used for $\rho_{\alpha^{(m)}}$, $\rho_{\beta^{(m)}}$, $\rho_{\alpha^{(v)}}$ and $\rho_{\beta^{(v)}}$ because allowing differences in these parameters added little to the model.

5.1 Assessing Model Fit

A natural first item to consider is how the model fits differ for the two models, which can be evaluated effectively with posterior predictive checks (Gelman et al. 1996; Gilks et al. 1996). Figure 1 summarizes these posterior predictive distributions for the two pairs of contaminants in the NAOS data, separately for surface and ground water systems but aggregated across states. Each frame was constructed by first obtaining, for each of Model I and Model J, a sample from the posterior predictive distribution of hypothetical data with the same sampling design as the observed data, as described in Section 2. In all cases, 10,000 such simulated hypothetical data sets were used to characterize the predictive distribution, based on sampling a new data set conditional on every 100th parameter vector of a MCMC sample of size one million from the relevant posterior distribution conditional on the NAOS data. The net result is, for each pair of contaminants, a large sample from two different posterior predictive distributions: one based on Model I and the other based on Model J. The samples were then used to estimate both marginal and joint posterior predictive densities via univariate and bivariate kernel density estimators (Simonoff 1996). The estimated marginal predictive densities, along with marginal histograms of the observed data, form the margins of each frame. The image in each frame represents the log of the ratio of the estimated posterior predictive density for Model J to that of Model I. Hence, a value of 0 represents a region assigned approximately equal density under the two models, while positive values correspond to regions where Model J provides higher density than Model I. The images are restricted to regions where both densities are at least 1% of their respective maxima to avoid unstable ratios.

For each pair of contaminants, it is clear that the two models provide nearly identical marginal fits to the data, and these fits are acceptable. This is not surprising since the marginal structure of the bivariate model is nearly identical to that of the univariate model. However, the models

provide substantially different joint fits to the contaminants. The posterior distributions of the within–location correlation coefficients of Model J for each pair of contaminants within each source type is approximately unimodal and symmetric, with posterior means of 0.86 (surface) and 0.83 (ground) for Mg and Ca and 0.52 (surface) and 0.18 (ground) for As and SO_4. These degrees of positive correlation are evident in the data even after aggregation across locations, and it is clear that Model J captures these correlations in a way not possible with Model I. For the most part, the data are concentrated in regions of the observation space to which Model J assigns higher posterior predictive density than Model I. The case in which the two are most similar is As and SO_4 in ground water, for which the contaminants exhibit only a mild positive association. Thus, in the presence of contaminant covariation, explicitly accounting for these correlations provides a better fit to the data. Plots similar to those depicted in Figure 1 formed within locations display similar patterns, and in no case was there strong evidence to refute the assumption of a common correlation structure across locations.

5.2 Comparison of Predictive Ability

Because the analyses presented here are motivated by prediction, it is necessary to examine whether the additional structure of Model J results in better predictions of new data in addition to its superior fit relative to Model I. This issue for the particular case of arsenic and sulfate was the subject of an extensive cross-validation study, provided in Lockwood (2001). After controlling for substantial Monte Carlo variability in the cross-validation comparison criterion (predictive density of the testing data), there was evidence that Model J provided a moderate predictive advantage relative to Model I. The degree of the advantage was commensurate with the relatively modest degree of covariation between these contaminants. For the more highly correlated substances Mg and Ca, Model J provides a stark predictive advantage relative to Model I. Figure 2 provides an example. The left frame compares the joint posterior predictive distributions of the two contaminants in New Jersey ground water systems based on the fit of Model I and Model J to the NAOS data. The overlaid data in this case are the corresponding observations from the USGS database, for which the data from New Jersey are most numerous. Hence, unlike Figure 1, the overlaid data were not

used to fit the model; thus the plot exhibits the predictive abilities of the models rather than their fit. Both models do an adequate job of predicting the marginal distributions of the substances, but Model J capitalizes on the high correlations to predict more accurately their joint distribution. Plots for other states represented in the USGS data display similar advantages for Model J, and the net result is summarized in the right frame of Figure 2. This provides the log joint predictive density in Equation (2) evaluated at the entire collection $\boldsymbol{y}_1$ of USGS Mg and Ca observations, estimated by 1,000,000 posterior samples from the fit of Models I and J to the NAOS data. The predictive density was estimated using three different methods under Model J (Lockwood 2001). With each aggregate estimate marked on the horizontal axis is associated a density estimate providing some indication of the Monte Carlo variability based on blocks of 10,000 parameter vectors. Although this variability is considerable, the plot provides convincing evidence that the predictive density based on Model J is substantially larger than that based on Model I, complementing the results provided for New Jersey.

5.3 Implications for Regulatory Inferences

As discussed in Section 1, the current study is motivated by the long-term goal of improving the sequential regulatory process by one in which rules for larger, more diverse groups of contaminants are considered. In the existing structure, the EPA examines available data on contaminant occurrence during the RIA process in order to form national estimates of contaminant occurrence. What is meant by a "national estimate" is some indication of the fraction or number of all community water systems in the nation expected to have either raw or finished water contaminant concentrations in excess of various concentrations of interest. Some measure of uncertainty about these estimates also should be an integral part of the decision process.

A multi-contaminant RIA would require analogous quantities for the collection of contaminants under consideration. For example, in the simplest case of two substances, it would be useful to estimate the fraction of all community water systems expected to have a raw water concentration of substance 1 in excess of concentration c_1 *and* and a raw water concentration of substance 2 in excess of c_2 for selected values of c_1 and c_2. Moreover, for optimal utility to the RIA, it is imperative

that such estimates be coupled with quantitative assessments of their uncertainties. Inferences of this nature are available from estimated joint distributions of the contaminants, and estimated uncertainties about these joint distributions, as a function of community water system attributes.

Without a formal structure in which to address contaminant covariations, estimated joint distributions would be based on either products of marginal distributions, or possibly some informal mechanism for accounting for occurrence relationships among the contaminants. One would expect that for contaminants that occur essentially independently, rigorous estimation of their joint distributions based on multivariate modeling would provide little to no benefits relative to, for example, concatenation of marginal models. On the other hand, for contaminants that covary strongly, the differences could be pronounced, and proper estimation of contaminant joint distributions would be required to maintain the integrity of the multi-contaminant RIA. The methods developed in this paper provide a flexible structure with which to estimate contaminant joint distributions in a manner that explicitly accounts for contaminant covariations. The purpose of this section is to exhibit the power of these methods in making inferences about quantities of critical importance to the regulatory process, and moreover, to demonstrate to what extent these inferences differ from those based on simpler models.

The fundamental quantities of interest are percentages (or numbers) of systems expected to have raw water concentrations in excess of various concentrations. In general, we assume that these quantities can be expressed as functions of some model parameters $\boldsymbol{\theta}$; for example, those from Model I, Model J, or some other parametric statistical model. For a given contaminant m, a concentration c_m, a fixed value of all model parameters $\boldsymbol{\theta}$, and a given water system s, let $P_s^{(m)}(\boldsymbol{\theta}, c_m)$ be the probability that the system has raw water concentration for contaminant m in excess of c_m. Because $P_s^{(m)}(\boldsymbol{\theta}, c_m)$ is a function of $\boldsymbol{\theta}$, it has a posterior distribution conditional on the raw water data, and features of this distribution are readily obtained via MCMC. The posterior distributions used in this section are conditional on both the NAOS and USGS data.

In the current modeling framework, with differentiation only by source type and location (in particular by U.S. state), systems in the same location using the same source water type have identical values of this probability. We make this explicit by replacing the index s with ij, so that

$P_{ij}^{(m)}(\boldsymbol{\theta}, c_m)$ is the probability that a system using source water type i in location (state) j has raw water concentration for contaminant m in excess of c_m. Under assumed lognormality of the concentrations, as in Models I and J,

$$P_{ij}^{(m)}(\boldsymbol{\theta}, c_m) = 1 - \Phi\left(\frac{\log c_m - \mu_{ij}(\boldsymbol{\theta})}{\sigma_{ij}(\boldsymbol{\theta})}\right), \tag{15}$$

where $\mu_{ij}(\boldsymbol{\theta})$ and $\sigma_{ij}(\boldsymbol{\theta})$ are the mean and standard deviation for the log contaminant distribution source water type i and state j as a function of $\boldsymbol{\theta}$.

Inference at the national level is based on the EPA Safe Drinking Water Information System (SDWIS) (EPA 2000), which provides comprehensive information about water system characteristics for all systems in the country. The national estimates discussed in this section are based on a 1998 SDWIS query which provided information for 10,637 surface water and 44,087 ground water systems. Let n_{ij} be the number of the 54,724 SDWIS systems in source type i ($i = 1, 2$) and state j ($j = 1, \ldots, 50$), and let $n_{i\cdot}$ denote the source water type marginal totals. These values are summarized in Table 1. By the assumed conditional independence of the systems, $P_{ij}^{(m)}(\boldsymbol{\theta}, c_m)$ in Equation (15) is the expected proportion of the n_{ij} systems having concentrations in excess of c_m. While location-specific estimates are of some interest for understanding local variations in contaminant occurrence, inferences aggregated across locations are more relevant due to the national scope of regulations [1]. Thus, we consider the marginal probabilities

$$P_{i\cdot}^{(m)}(\boldsymbol{\theta}, c_m) = \sum_{j=1}^{50} \frac{n_{ij}}{n_{i\cdot}} P_{ij}^{(m)}(\boldsymbol{\theta}, c_m), \tag{16}$$

which for $i = 1, 2$ is the fraction of the $n_{i\cdot}$ national systems using source water type i expected to have raw water concentrations in excess of c_m. These marginal probabilities are available from either Model I or Model J, and similar to the marginal comparisons of model fit, inferences about them are virtually identical for the two models. Figure 3 presents some summary results for arsenic and sulfate. The frames provide posterior means and ± 2 times posterior standard deviations of $P_{i\cdot}^{(m)}(\boldsymbol{\theta}, c_m)$ for arsenic (top row) and sulfate (bottom row) in surface water (left column) and

[1] In fact, there is considerable regional variability in contaminant levels for all of the contaminants discussed in this paper. The most pervasive trend is that states in the eastern part of the U.S. have lower concentrations than most states in the west.

ground water (right column). The selected concentrations c_m for each contaminant are reflective of those considered during current regulatory activity. The historical MCL for arsenic was 50 $\mu g/L$, and during the recent revision process, MCLs as low as 2 $\mu g/L$ were suggested based on minimizing health risks. For sulfate, a 500 mg/L MCL was proposed but was never adopted, and there is an existing "secondary" standard of 250 mg/L. (Secondary standards are not motivated by health impacts, but rather aesthetic concerns.) Estimates such as those presented in Figure 3, along with their associated uncertainties, would be considered during the RIA process. It is interesting to note the substantially higher uncertainties regarding the surface water marginal fractions, which derives from the fact that only about 7% of the systems in the combined NAOS and USGS are surface water systems.

In a multi-contaminant RIA, it would be necessary to estimate percentages of systems expected to have raw water concentrations of two (or more) contaminants simultaneously exceeding two (or more) given concentrations, in addition to the marginal percentages. For clarity, we focus on the two contaminant case of arsenic and sulfate. Let $P_{ij}^{(*)}(\boldsymbol{\theta}, c_1, c_2)$ denote the probability that a system using source water type i in state j simultaneously has a raw water arsenic concentration greater than c_1 and a raw water sulfate concentration greater than c_2. Analogous to the discussion for a single contaminant, a quantity that would be of particular interest in a multi-contaminant RIA would be the estimated fraction of all national systems of a given source type that would simultaneously exceed given concentrations for both contaminants. This is the marginal probability

$$P_{i\cdot}^{(*)}(\boldsymbol{\theta}, c_1, c_2) = \sum_{j=1}^{50} \frac{n_{ij}}{n_{i\cdot}} P_{ij}^{(*)}(\boldsymbol{\theta}, c_1, c_2). \tag{17}$$

Credible estimation of $P_{i\cdot}^{(*)}(\boldsymbol{\theta}, c_1, c_2)$ requires that the $P_{ij}^{(*)}(\boldsymbol{\theta}, c_1, c_2)$ reflect any covariations in the occurrence of the contaminants. For example, under Model J and for a given value of $\boldsymbol{\theta}$, $P_{ij}^{(*)}(\boldsymbol{\theta}, c_1, c_2)$ is a bivariate normal orthant probability based on the joint distribution of the contaminants in source type i and state j. On the other hand, when the contaminants are modeled independently as in Model I, $P_{ij}^{(*)}(\boldsymbol{\theta}, c_1, c_2) = P_{ij}^{(As)}(\boldsymbol{\theta}, c_1) P_{ij}^{(SO4)}(\boldsymbol{\theta}, c_2)$. One would expect

$$P_{ij}^{(*)}(\boldsymbol{\theta}, c_1, c_2) > P_{ij}^{(As)}(\boldsymbol{\theta}, c_1) P_{ij}^{(SO4)}(\boldsymbol{\theta}, c_2)$$

when the contaminants are positively correlated. Thus, the posterior distributions of $P_{i\cdot}^{(*)}(\boldsymbol{\theta}, c_1, c_2)$ could be quite different for the two models, with greater differences for more highly correlated contaminants. As before, these distributions are readily obtained with MCMC samples. The calculations for Model J are based on samples from a single chain, while the results for Model I derive from appropriate manipulation of samples from separate chains for the contaminants under consideration.

Figure 4 shows the degree of disparity in these posterior distributions. The top row of the figure compares the posterior distribution of $100P_{i\cdot}^{(*)}(\boldsymbol{\theta}, c_1, c_2)$, i.e. the percentage of national systems of a given source type that would exhibit raw water concentrations simultaneously in excess of concentrations of $c_1 = 5\ \mu g/L$ for arsenic and $c_2 = 100\ mg/L$ for sulfate. The bottom row provides analogous results for magnesium at 25 mg/L and calcium at 100 mg/L. In each frame, the histogram depicts the posterior distribution obtained from Model J, while the dotted density estimate gives that obtained from Model I. In all cases, the posterior distributions based on Model I are shifted toward lower percentages than those based on Model J, with more striking differences for the more highly correlated magnesium and calcium. For arsenic and sulfate, the difference is somewhat more pronounced for surface water, which is consistent with the fact that the contaminants are more highly correlated in surface water. Although the raw percentages are small, these are precisely the sorts of percentages often faced in regulatory decisions because in most cases, only a small fraction of systems are affected by a contaminant MCL.

The differences exhibited in Figure 4 have important implications for national cost estimates performed during regulatory investigation. As a simple but instructive example, consider the situation where there is only a single treatment, and this treatment can effectively remove both arsenic and sulfate. Suppose that it is of interest to estimate the total cost, across all community water systems, of implementing this treatment to bring finished water concentrations of the two contaminants below pre-specified concentrations. The total cost estimate is roughly the sum of the marginal costs for each contaminant, minus the costs for the systems requiring treatment for both contaminants. When the contaminants are positively correlated, the marginal occurrence models underestimate the latter quantity because they do not account for this correlation. Thus their

estimated costs would be inflated artificially. On the other hand, the joint occurrence models would produce more realistic cost estimates because they account for the contaminant covariations. The primary conclusion is that the estimates of important regulatory quantities can be quite different depending on whether the contaminants are considered jointly or independently, underscoring the value of the methods presented in the current study.

6. DISCUSSION

Through extension of existing methods for modeling a single contaminant, this paper develops and demonstrates an effective method for modeling joint distributions of contaminants in community water system source waters. Especially for highly correlated contaminants, the necessity of a joint modeling approach for integration into a multi-contaminant RIA process is clear. Modeling contaminants individually, while convenient, results in a potentially severe loss of co-occurrence information that can negatively impact predictive power and inferences about key regulatory quantities. The framework presented here provides the requisite formal structure which is flexible enough to handle any groups of contaminants, as well as covariances among contaminants and other quantities that may affect cost estimates (e.g. pH).

Moreover, this study has identified issues with water quality modeling and available data sets that limit the development of more effective statistical characterizations. Most water quality analyses focus on data collected over a limited spatial scale, where intricate hydrodynamics can be successfully modeled. However, because these models are unrealistic for data collected at more expansive spatial scales, national data are typically not subjected to statistical modeling. The lack of rigorous analyses at large spatial scales is detrimental to the process of setting drinking water quality standards for two reasons. The first is that because EPA MCLs apply to all community water suppliers, national occurrence distributions, as opposed to analyses of local chemistry, are of paramount importance. The other is that an integral part of the RIA process is an analysis of uncertainty, which is not properly obtainable without statistical modeling. The Bayesian hierarchical methods presented here provide a way to make national predictions and statements about uncertainty based on spatially sparse data. In addition, for cross-sectional, spatial and multivariate

data, most water quality analyses consist of either a sequence of univariate spatial analyses (e.g. kriging) or a multivariate analysis that ignores spatial relationships (e.g. principal components). The methods presented here simultaneously account for both aspects of the data. With appropriate (albeit difficult) extensions, the methods could be used to examine temporal trends as well. Finally, censored data are ubiquitous in water quality analyses, especially among trace substances such as arsenic. Despite the existence of numerous methods for properly handling them, many analyses revert to *ad hoc* treatment. Often censored observations are ignored or replaced with an arbitrary constant, neither of which properly account for the actual information provided by the censored observations. The data augmentation techniques used in the models presented here are both easy to implement and theoretically justifiable, as they treat the missing data as any other unobserved quantity in the Bayesian framework. In all, Bayesian hierarchical models provide a useful paradigm from which to approach water quality analyses at the national level.

The methods presented here lend themselves to numerous extensions, some of which we are currently pursuing. Although this paper presents results for only bivariate cases, the multi-contaminant model has been implemented successfully in estimating joint distributions for seven contaminants, as mentioned previously. The model was used in conjunction with a statistical model of contaminant removal by water treatment processes to model finished water concentrations for multiple contaminants, which in turn were used to conduct a multi-contaminant RIA (Gurian et al. 2001b). Moreover, we are exploring modifications of Stage 2a in Equation (11) in which the fixed matrices $\boldsymbol{\alpha}_0^{(m)}$, $\boldsymbol{\alpha}_0^{(v)}$, $\boldsymbol{\beta}_0^{(m)}$, and $\boldsymbol{\beta}_0^{(v)}$ are replaced with more structured matrices that have some unknown parameters. This additional hierarchical level provides a more adaptive modeling structure by allowing the data to inform the location parameters, and can be used to account for possible database heterogeneity in contaminant measurements. Preliminary results indicate that the extra adaptivity may be beneficial, and some degree of systematic differences among the measurements for the NAOS and USGS do exist. Finally, using the U.S. states as the location basis is acceptable for generally sparse raw water data, where borrowing of strength across locations is of paramount concern. However, the model is sufficiently general to handle the allocation of the systems to an arbitrary collection of k locations, e.g. counties, watersheds, or the systems themselves, although such richer location

bases entail additional complexity and possible difficulties. The simple distance–based correlation structure is sufficient to capture large–scale spatial trends across states, but finer location bases may warrant more careful consideration of the prior correlation structures. In addition, a greater number of locations leads to larger matrices, introducing computational difficulties. As these types of Gaussian process structures have wide applicability in statistical modeling, the development of methods to deal efficiently with their computational aspects is of great importance.

REFERENCES

Backer, L. (2000), "Assessing the acute gastrointestinal effects of ingesting naturally occurring, high levels of sulfate in drinking water," *Critical Reviews in Clinical Laboratory Sciences*, 37(4), 389–400.

Barnard, J., McCulloch, R., and Meng, X.-L. (2000), "Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage," *Statistica Sinica*, 10, 1281–1311.

Berliner, L. M. (2000), "Hierarchical Bayesian modeling in the environmental sciences," *Journal of the German Statistical Society*, 84, 141–153.

Boatwright, P., McCulloch, R., and Rossi, P. (1999), "Account-level modeling for trade promotion: An application of a constrained parameter hierarchical model," *Journal of the American Statistical Association*, 94(448), 1063–1073.

Brav, A. (2000), "Inference in long-horizon event studies: A Bayesian approach with application to initial public offerings," *Journal of Finance*, 55(5), 1979–2016.

Carlin, B. and Louis, T. (2000), *Bayes and Empirical Bayes Methods for Data Analysis* (Second ed.), Boca Raton, FL: Chapman and Hall/CRC Press.

Cressie, N. (2000), "Position Paper from The Ohio State University Workshop on Hierarchical Modeling in Environmental Statistics,", submitted to the U.S. Environmental Protection Agency.

Damien, P. and Walker, S. (2001), "Sampling truncated normal, beta and gamma densities," *Journal of Computational and Graphical Statistics*, 10(2), 206–215.

Dyk, D. v. and Meng, X. (2001), "The art of data augmentation (with discussion)," *Journal of Computational and Graphical Statistics*, 10(1), 1–111.

EPA (1999), "Health Effects from Exposure to High Levels of Sulfate in Drinking Water Study and Sulfate Workshop; Notice," *Federal Register*, 64(28), 7027–7030.

——— (2000), "Safe Drinking Water Information System,", Part of EPA Envirofacts Data Warehouse.

——— (2001a), "National Primary Drinking Water Regulations; Arsenic and Clarifications to Compliance and New Source Contaminants Monitoring; Delay of Effective Date," *Federal Register*, 66(99), 28342–28350.

——— (2001b), "National Primary Drinking Water Standards,", EPA Publication 816-F-01-007.

Focazio, M., Welch, A., Watkins, S., Helsel, D., and Horn, M. (2000), "A Retrospective Analysis of the Occurrence of Arsenic in Ground Water Resources in the United States and Limitations in Drinking Water Supply Characterizations,", U.S. Geological Survey Water Resources Investigations Report 99-4279, Reston, VA.

Frey, M., Chwirka, J., Kommineni, S., Chowdhury, Z., and Narasimhan, R. (2000), "Cost Implications of a Lower Arsenic MCL Final Report,", American Water Works Research Foundation, Denver, CO.

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.

Frey, M., Owen, D., Chowdhury, Z., Raucher, R., and Edwards, M. (1998), "Costs to utilities for a lower MCL for arsenic," *Journal of the American Water Works Assocation*, 90(3), 89–102.

Garthwaite, P. and Al-Awadhi, S. (2001), "Non-conjugate prior distribution assessment for multivariate normal sampling," *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 63(1), 95–110.

Gelfand, A., Smith, A., and Lee, T. (1992), "Bayesian analysis of constrained parameter and trun-

cated data problems using Gibbs sampling," *Journal of the American Statistical Association*, 87, 523–532.

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.

Gennings, C., Schwartz, P., Carter Jr., W. H., and Simmons, J. E. (1997), "Detection of departures from additivity in mixtures of many chemicals with a threshold model," *Journal of Agricultural, Biological, and Environmental Statistics*, 2(2), 198–211.

Geweke, J. (1991), "Efficient simulation from the multivariate normal and Student-t distributions subject to linear constraints," in *Computing science and statistics: Proceedings of the twenty-third symposium on the interface*, *American Statistical Association*, Alexandria, VA, pp. 571–578.

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

Gurian, P., Small, M., Lockwood, J., and Schervish, M. (2001a), "Addressing uncertainty and conflicting cost estimates in revising the arsenic MCL," *Environmental Science & Technology*, 35(22), 4414–4420.

——— (2001b), "Assessing nationwide cost-benefit implications of multi-contaminant drinking water standards,", In preparation.

——— (2001c), "Benefit-cost estimation for alternative drinking water MCLs," *Water Resources Research*, 37(8), 2213–2226.

Guttorp, P. (2000), "Statistics in the year 2000: Vignette on environmental statistics," *Journal of the American Statistical Association*, 95(449), 289–292.

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.

Lockwood, J. (2001), *Estimating Joint Distributions of Contaminants in U.S. Community Water System Sources*, unpublished Ph.D. dissertation, Department of Statistics, Carnegie Mellon University.

Lockwood, J., Schervish, M., Gurian, P., and Small, M. (2001), "Characterization of arsenic occurrence in source waters of US community water systems," *Journal of the American Statistical Association*, in press.

McCulloch, R. and Rossi, P. (1994), "An exact likelihood approach to analysis of the MNP model," *Journal of Econometrics*, 64, 207–240.

Morrison, D. F. (1990), *Multivariate Statistical Methods* (Third ed.), New York: McGraw-Hill.

Neukrug, H. (2000), "Deemphasizing contaminant-by-contaminant regulation," *Journal of the American Water Works Association*, 92(3), 24–30.

Olkin, I. (1953), "Note on "The Jacobians of certain matrix transformations useful in multivariate analysis"," *Biometrika*, 40, 43–46.

Piegorsch, W., Smith, E., Edwards, D., and Smith, R. (1998), "Statistical advances in environmental science," *Statistical Science*, 13(2), 186–208.

Pinheiro, J. and Bates, D. (1996), "Unconstrained parameterizations for variance-covariance matrices," *Statistics and Computing*, 6, 289–296.

Ritter, C. and Tanner, M. (1992), "Facilitating the Gibbs sampler: The Gibbs stopper and the griddy Gibbs sampler," *Journal of the American Statistical Association*, 87, 861–868.

Roberson, J. and Power, J. (2000), "The groundwater train wreck," *Journal of the American Water Works Association*, 92(3), 8,103.

Rue, H. (2001), "Fast sampling of Gaussian Markov random fields," *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 63, 325–338.

Schervish, M. (1995), *Theory of Statistics* (Second ed.), New York: Springer-Verlag.

Simonoff, J. S. (1996), *Smoothing Methods in Statistics*, New York: Springer-Verlag.

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.

Woodard, R., Sun, D., and He, Z. (2000), "Bayesian hierarchical models for spatially correlated data from survey sampling," Technical report, University of Missouri at Columbia.

Table 1. Cross classification of 54,724 SDWIS community water systems by state and source water type (n_{ij}).

FIPS Code	Surface	Ground	FIPS Code	Surface	Ground
AK	154	553	MT	89	581
AL	239	346	NC	340	1937
AR	246	482	ND	62	276
AZ	73	1074	NE	11	621
CA	776	2748	NH	51	615
CO	264	554	NJ	94	519
CT	59	537	NM	38	635
DE	3	229	NV	49	294
FL	418	2235	NY	750	1947
GA	195	1476	OH	297	1146
HI	29	155	OK	638	560
IA	124	1035	OR	209	685
ID	58	686	PA	437	1808
IL	551	1255	RI	24	62
IN	107	804	SC	178	548
KS	327	605	SD	115	385
KY	364	131	TN	295	257
LA	72	1211	TX	923	3620
MA	170	361	UT	96	363
MD	59	455	VA	265	1262
ME	97	317	VT	91	354
MI	281	1188	WA	201	2112
MN	38	964	WI	43	1152
MO	241	1180	WV	322	305
MS	9	1256	WY	65	206

Figure 1. Comparison of posterior predictive distributions for NAOS data for Model I and Model J, separately for surface water (left) and ground water (right) and for As/SO_4 (top) and Mg/Ca (bottom). Imaged values are the estimated log posterior predictive density ratio of Model J to Model I, with the associated NAOS data overlaid. Estimate maxima for the joint predictive densities are provided with colored circles. Censored observations fall along dotted lines.

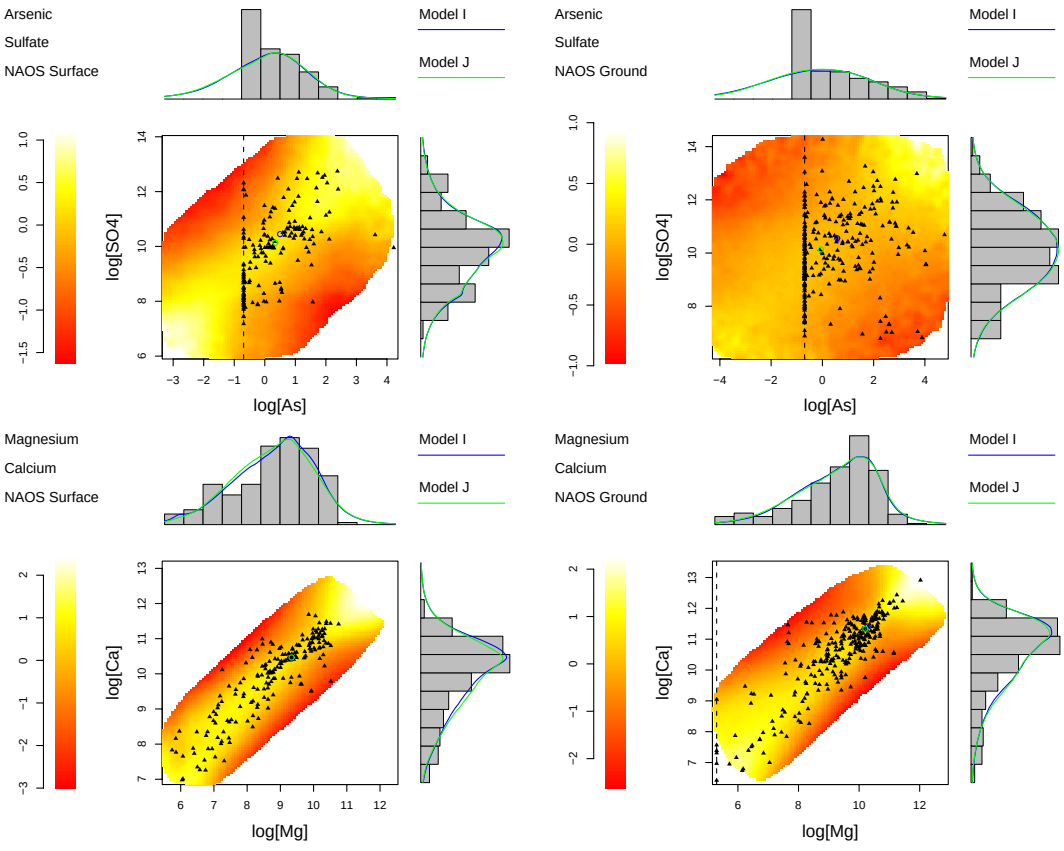

Figure 2. (Left frame) Comparison of posterior predictive distributions for Mg and Ca in New Jersey groundwater based on fit to NAOS, with corresponding data from USGS overlaid. Imaged values are the estimated log posterior predictive density ratios of Model J to Model I. Estimate maxima for the joint predictive densities are provided with colored circles. (Right frame) Comparison of Model I and Model J predictive densities based on one million MCMC samples. Three different methods for the calculation of the Model J log predictive density are marked with "x"; the one for Model I is marked with "o". The associated histograms indicate Monte Carlo variability of the estimate based on parameter blocks of size 10,000.

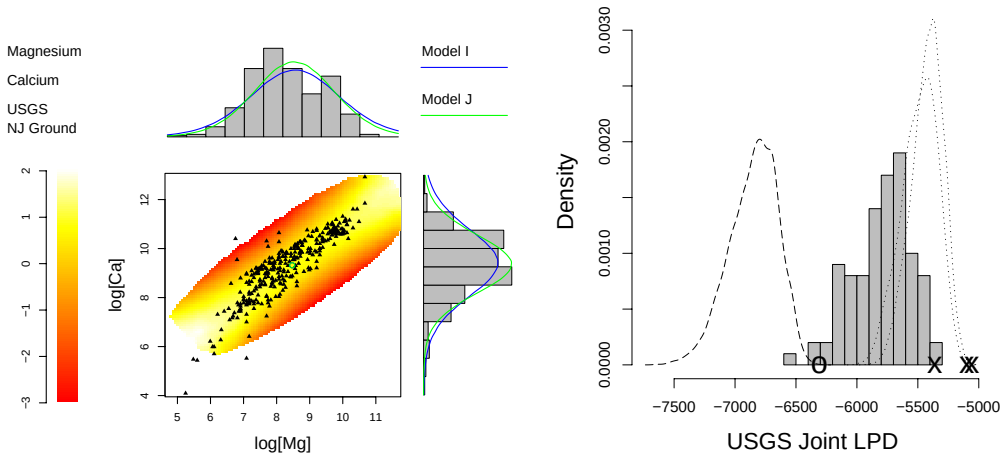

Figure 3. Posterior distributions of $P_{i\cdot}^{(As)}(\boldsymbol{\theta}, c_m)$ (top row) and $P_{i\cdot}^{(SO4)}(\boldsymbol{\theta}, c_m)$ (bottom row) for selected concentrations c_m. Posterior means indicated by "o" and bars delimit ± 2 posterior standard deviations from the posterior means.

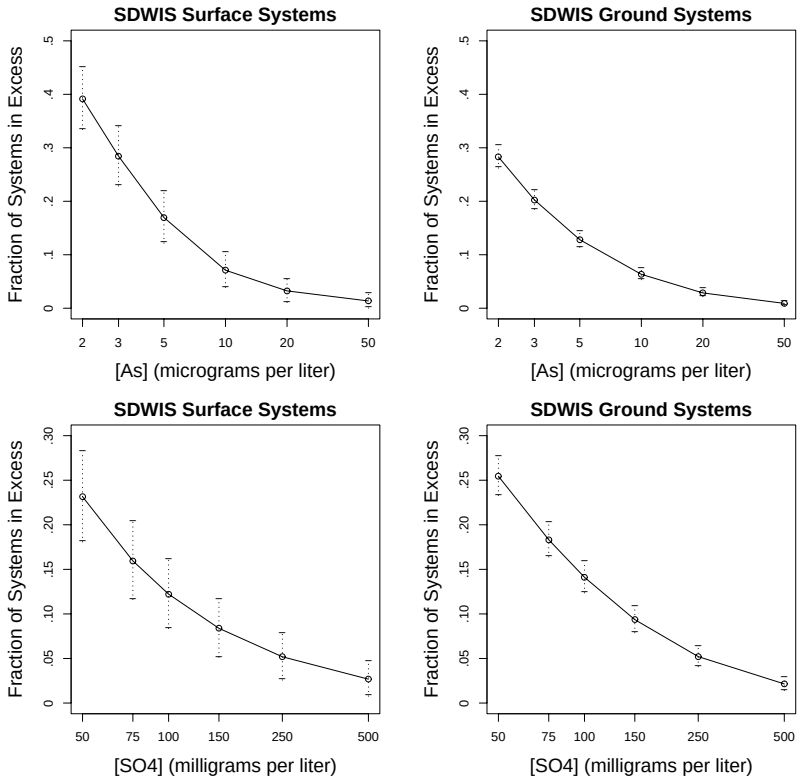

Figure 4. Estimated posterior distributions of $100P_{i\cdot}^{(*)}(\boldsymbol{\theta}, c_1, c_2)$ *based on Model I (density estimates) and Model J (histograms), conditional on both the NAOS and USGS data. The top row considers concentrations of* $c_1 = 5\ \mu g/L$ *for arsenic and* $c_2 = 100\ mg/L$ *for sulfate, and the bottom row considers magnesium at 25* mg/L *and calcium at 100* mg/L*. Posterior means for Model J are marked with "x", and those for Model I are marked with "o".*