Modeling the U.S. national distribution of waterborne pathogen concentrations with application to *Cryptosporidium parvum*

Ciprian M. Crainiceanu,[1] Jery R. Stedinger,[2] David Ruppert,[3] and Christopher T. Behr[4]

Received 15 August 2002; revised 27 February 2003; accepted 8 April 2003; published 4 September 2003.

[1] This paper provides a general statistical methodology for modeling environmental pathogen concentrations in natural waters. A hierarchical model of pathogen concentrations captures site and regional random effects as well as random laboratory recovery rates. Recovery rates were modeled by a generalized linear mixed model. Two classes of pathogen concentration models are differentiated according to their ultimate purpose: water quality prediction or health risk analysis. A fully Bayesian analysis using Markov chain Monte Carlo (MCMC) simulation is used for statistical inference. The applicability of this methodology is illustrated by the analysis of a national survey of *Cryptosporidium parvum* concentrations, in which 93% of the observations were zero counts. *INDEX TERMS:* 1860 Hydrology: Runoff and streamflow; 1871 Hydrology: Surface water quality; 1894 Hydrology: Instruments and techniques; *KEYWORDS:* Bayesian analysis, Markov Chain Monte Carlo, waterborne pathogens, *Cryptosporidium parvum*, generalized linear mixed model

Citation: Crainiceanu, C. M., J. R. Stedinger, D. Ruppert, and C. T. Behr, Modeling the U.S. national distribution of waterborne pathogen concentrations with application to *Cryptosporidium parvum*, *Water Resour. Res.*, *39*(9), 1235, doi:10.1029/2002WR001664, 2003.

1. Introduction

[2] Pathogenic waterborne microorganisms represent a serious threat to drinking water quality. In 1990, a Science Advisory Panel to the U.S. Environmental Protection Agency (U.S. EPA) described microorganisms as the greatest remaining challenge to health risk management for drinking water suppliers [*Science Advisory Board*, 1990]. To assess the health effects of long-term exposure to a specific waterborne pathogen and better define regulations for treatment plants, one needs first to understand the distribution of background pathogen concentrations.

[3] There are several challenges in collecting and modeling pathogen concentration data. First, the true concentration cannot be observed directly. Instead, laboratory technicians discretely count the number of pathogens in a sample of water. Moreover, the counting process of microscopic organisms in a laboratory is subject to various sources of error leading to a count of only a fraction of the organisms originally present in a water sample. This fraction varies randomly and can depend on the laboratory that analyzes the sample and on water quality attributes.

[4] The pathogen concentrations are likely to exhibit spatial and temporal correlations as well as dependence on covariates, such as turbidity, streamflow rates, land use and seasonality. The main objective of this paper is to provide a general methodology for the statistical modeling of spatially distributed pathogen concentrations based upon available count data. For this purpose we develop a model incorporating covariates that are likely to have a causal effect on concentrations. The model has a hierarchical structure for observations within sites, sites, regions and an overall national average. A fully Bayesian approach using Markov chain Monte Carlo (MCMC) simulation is used for statistical inference. As an illustration of this methodology we use the EPA's national Information Collection Rule (ICR) survey that yielded *Cryptosporidium* count data. We also use ICR *Giardia* data to illustrate some interesting phenomena not found in the *Cryptosporidium* data set.

[5] *Cryptosporidium parvum* is a microscopic waterborne pathogen that can produce gastrointestinal illness in healthy individuals and serious complications or even death in individuals with a weakened immune system [*Meinhardt et al.*, 1996]. *Cryptosporidium* and *Giardia* spread in the environment as microscopic spore-like structures called oocysts for *Cryptosporidium* and cysts for *Giardia*. These (oo)cysts are resistant to many environmental stresses. In particular, *Cryptosporidium* oocysts are resistant to chlorination causing it to be a health risk in water supplies that depend upon chlorination without filtration or other processes that would reliably remove them.

[6] In response to recent outbreaks, the US EPA conducted a national survey of *Cryptosporidium* concentrations under the ICR described by *U.S. EPA* [2001]. The ultimate goal of the investigations was to develop revised water treatment standards. The ICR survey was conducted over a period of 18 months and included 350 major water users.

[7] Figure 1 shows the oocyst counts for the ICR survey. A huge proportion of these observations are zero counts (93%). Among the remaining nonzero observations most are just one or two, but there exists several large counts (of 38, 35, 30) indicating strong overdispersion with respect to the Poisson distribution. The statistical analysis of these

[1]Department of Statistical Science, Cornell University, Ithaca, New York, USA.
[2]School of Civil and Environmental Engineering, Cornell University, Ithaca, New York, USA.
[3]School of Operational Research and Industrial Engineering, Cornell University, Ithaca, New York, USA.
[4]eDesign Dynamics LLC, West New York, New Jersey, USA.

Copyright 2003 by the American Geophysical Union.
0043-1397/03/2002WR001664$09.00

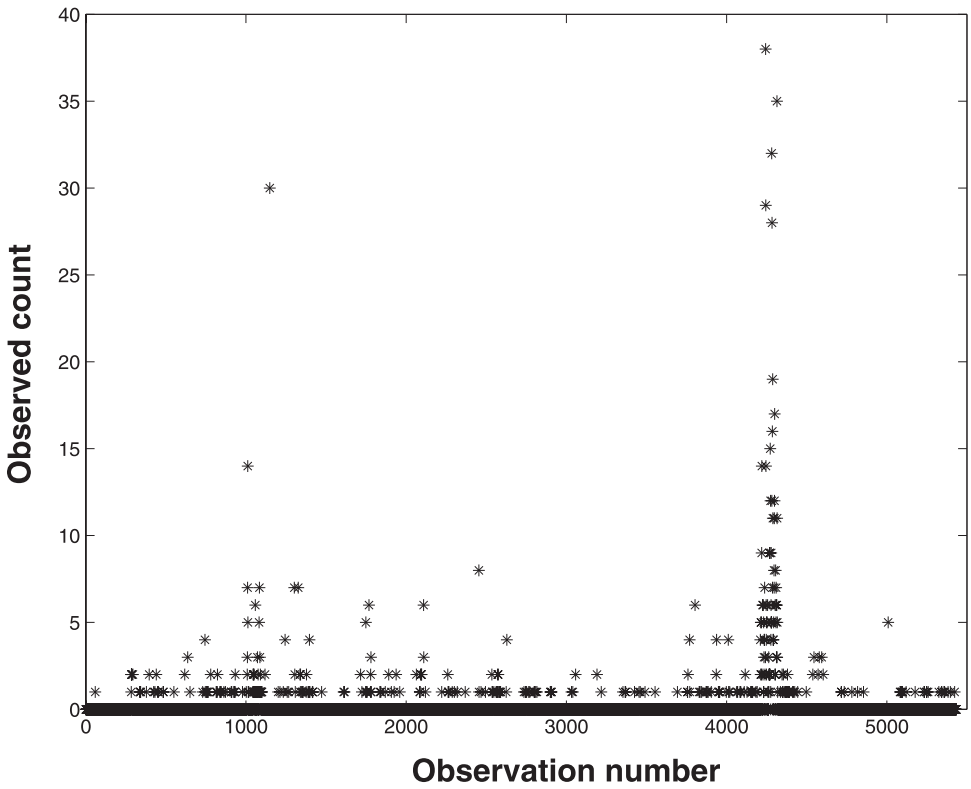

Figure 1. *Cryptosporidium* count data in the ICR survey.

data is challenging because of the discrete nature of the response variable (oocyst counts), the high frequency of zero counts, lack-of-fit for the standard Poisson distribution, spatial and temporal correlation structure, seasonality, random effects, variable recovery rates, and missing observations [*Messner and Wolpert*, 2002].

[8] The paper is organized as follows. Section 2 provides a literature review for *Cryptosporidium* risk and statistical modeling of count data. Section 3 discusses Bayesian analysis using Markov Chain Monte Carlo (MCMC) simulation and comparisons with standard likelihood analysis. Section 4 provides a general methodology for Bayesian modeling of pathogen counts including a subsection on modeling recovery rates. Inference and results for the recovery rate models are also discussed. Section 5 presents results for the ICR data set for a selected set of models. Section 6 discusses the prediction of pathogen distributions. Section 7 provides the conclusions of the paper.

2. Literature Review

2.1. *Cryptosporidium parvum* Risk

[9] Public health concerns have been high ever since *Cryptosporidium* caused large outbreaks of gastrointestinal illness in the US, most notably in Milwaukee [*Solo-Gabrielle and Neumeister*, 1996]. Additional outbreaks have been detected in the United Kingdom, Canada, Australia, and Japan [*Meinhardt et al.*, 1996; *Clancy*, 2000; *Hashimoto et al.*, 2001; *Laing*, 2002]. Studies have also shown endemic rates of such illnesses to be between 0.5 and 5 million cases in the U.S. each year [*Fayer et al.*, 1997].

[10] *Rose et al.* [1997] document the frequency of occurrence and prevalence of *Cryptosporidium* in surface waters across North America based upon a number of different reports. The studies that had been conducted failed to provide a complete picture of such a complex issue. *Cryptosporidium* is largely unaffected by chlorination levels that are acceptable for water treatment while other treatments are costly [*Juranek*, 1995; *Okun et al.*, 1997; *Brodeur*, 2001; *Dugan et al.*, 2001]. Models are needed to support health risk analyses that guide public policy and development of the EPA regulations in the search for the appropriate tradeoff among microbial risks, chlorination and other disinfection procedures, and treatment costs [*Putnam and Wiener*, 1995; *Regli et al.*, 1999; *Casman et al.*, 2000; *Messner et al.*, 2001; *U.S. EPA*, 2001].

[11] Public health concerns have led to the collection of data on *Cryptosporidium* concentrations and models of its frequency of occurrence [*Haas and Rose*, 1996; *Parkhurst and Stern*, 1998]. Count data describing the environmental concentration of *Cryptosporidium* appear to be overdispersed relative to the Poisson distribution, as would be expected if environmental concentrations vary over time [*Stedinger and MacKay*, 1998].

[12] Likely sources of *Cryptosporidium* are of great concern. Large mammals, particularly cows in dairy production, are an obvious potential source [*National Research Council*, 1999; *Ong et al*., 1996]. *LeChevallier and Norton* [1992] found water turbidity to be a significant indicator of elevated concentrations. Other possible indicators are *Giardia*, fecal coliforms, *Clostridium*, alkalinity, water hardness, pH, and river discharge [*Atherholt et al.*, 1998; *Walker*, 1999, chap. 15–19]. Several studies found relationships between concentrations and seasonal factors [*Hansen and Ongerth*, 1991]. *Walker and Stedinger* [1999] suggest that wastewater treatment plants are likely to be important

sources of oocysts. *Poulton et al.* [1992] found that oocyst concentrations were highest during peak river flows and immediately after rainfall events, suggesting sources of oocysts included sewage discharges.

2.2. Modeling of Count Data

[13] *McCullagh and Nelder* [1989] present generalized linear models (GLMs) as a flexible class of regression models that can accommodate discrete responses. Poisson regression, a standard GLM, is used by many statisticians to explain discrete count data. *Grenfell and Wilson* [1997] applied GLMs to parasite data using Poisson and negative binomial errors. *Walker* [1999] studied a *Cryptosporidium* data set for the Delaware River using several GLMs and concluded that a negative binomial model provided a better fit than a Poisson distribution.

[14] National pathogen surveys generate complex data sets with significant spatial and temporal correlation structure that is not captured by simple GLMs. Random effects can be added to represent this structure resulting in generalized linear mixed model (GLMM) [*Breslow and Clayton*, 1993; *Littell et al.*, 1996; *McCulloch and Searle*, 2001; *Maiti*, 2001]. Mixed models with random effects have been shown to be effective for count data with spatial variation [*Christensen and Waagepetersen*, 2002; *Wikle*, 2002; *Best et al.*, 2002]. To model the additional uncertainty in the pathogen counting process, the pathogen model in section 4 proceeds beyond the standard GLMMs.

[15] While models including complex random effects structures that reasonably represent the data set are desirable, until recently the practitioner was faced with insurmountable computational problems. Section 3 explains why the maximum likelihood estimation for such models requires computation of analytically intractable integrals. This could be very difficult even for problems with a modest number of random effects [*Carlin and Louis*, 2000]. GLMM's can be fitted using "PROC GLIMMIX" in SAS; however, GLIMMIX uses a Laplace approximation which is known to introduce bias in the estimation of variance components [*Lin and Breslow*, 1996]. *McCulloch* [1997] shows how to calculate the exact maximum likelihood estimate using a computationally intensive Monte Carlo EM algorithm. Bayesian analysis using MCMC simulation has become the standard statistical tool for inference in such complex models [see, e.g., *Gilks et al.*, 1996; *Carlin and Louis*, 2000] and WinBugs [*Spiegelhalter et al.*, 2000] is the standard software package for such analyses. Case studies in Bayesian statistics are provided, for example, by *Gatsonis et al.* [2002a, 2002b]. Bayesian Monte Carlo methods are very flexible and are already seeing wide application in water resources [*Kuczera*, 1999; *Bates and Campbell*, 2001; *Wang*, 2001]. *Messner and Wolpert* [2002] apply MCMC techniques to pathogen data sets.

3. Bayesian Analysis Using MCMC

[16] Bayesian analysis using MCMC simulation has become the standard for statistical inference in complex models. It is preferred in this paper over the more classical approaches, such as Maximum Likelihood (ML) inference, because it is computationally feasible and provides the entire joint posterior distribution of the parameters given the data.

[17] Assume that the response variables $y_1, \ldots, y_n$ (in our case pathogen counts) have a conditional distribution $f(y_i|\boldsymbol{u}, \beta, \sigma^2)$, where $\boldsymbol{u}$ are random effects, β are parameters for the fixed effects, and σ^2 represents the variance components. Assuming independence among the y_i given $(\boldsymbol{u}, \beta, \sigma^2)$, the conditional (given $\boldsymbol{u}$) likelihood function is

$$f(\boldsymbol{y}|\boldsymbol{u}, \beta, \sigma^2) = \prod_{i=1}^{n} f(y_i|\boldsymbol{u}, \beta, \sigma^2). \quad (1)$$

If the distribution for the random effects $\boldsymbol{u}$ has probability density function (pdf) $\rho(\boldsymbol{u}|\boldsymbol{D})$, where $\boldsymbol{D}$ are the parameters of the distribution of $\boldsymbol{u}$, then the unconditional likelihood function with this data is

$$L(\beta, \sigma^2, \boldsymbol{D}|\boldsymbol{y}) = \int f(\boldsymbol{y}|\boldsymbol{u}, \beta, \sigma^2)\rho(\boldsymbol{u}|\boldsymbol{D})d\boldsymbol{u}. \quad (2)$$

If it were available, this unconditional likelihood would be the basis of all likelihood inference. For example, the maximum likelihood estimate would be the value of (β, σ^2) maximizing (2). However, computation of this unconditional likelihood function at any point of the parameter space requires the evaluation of an integral whose dimension is equal to the dimension of the random effects vector, $\boldsymbol{u}$, which is usually very large. Although there are recently developed methods for approximating the unconditional likelihood or computing the MLE, e.g., the expectation maximization (EM) algorithm, Bayesian analysis through MCMC provides a more elegant approach. Mixed models can be viewed as a step toward Bayesian statistics. In mixed models, some parameters, called the random effects, are modeled as random variables. Taking this approach a little further, we arrive at Bayesian analysis where all parameters of the model are considered to be random variables with specified prior distributions. Let $\rho(\beta)$, $\rho(\sigma^2)$, and $\rho(\boldsymbol{D})$ be the pdf's of the prior distributions of β, σ^2, and $\boldsymbol{D}$, respectively. Then the pdf of the joint distribution of the data, random effects, and parameters parameters $(\boldsymbol{y}, \boldsymbol{u}, \psi)$, with $\psi = (\beta, \sigma^2, \boldsymbol{D})$ is

$$\rho(\boldsymbol{y}, \boldsymbol{u}, \psi) = f(\boldsymbol{y}|\boldsymbol{u}, \beta, \sigma^2)\rho(\boldsymbol{u}|\boldsymbol{D})\rho(\beta)\rho(\sigma^2)\rho(\boldsymbol{D}). \quad (3)$$

Using Bayes rule, the joint posterior density of the parameters and random effects given the data has the pdf

$$\rho(\boldsymbol{u}, \psi|\boldsymbol{y}) = \frac{\rho(\boldsymbol{y}, \boldsymbol{u}, \psi)}{\int \rho(\boldsymbol{y}, \boldsymbol{u}, \psi)d\boldsymbol{u}\, d\psi} \quad (4)$$

The numerator of this pdf has a known analytical form but the denominator, representing the normalizing constant, is a high dimensional integral, even more complicated than the integral in equation (2). Therefore, the exact computation of the posterior pdf is intractable. Markov Chain Monte Carlo allow simulation from distributions with unknown normalizing constants. The general philosophy behind MCMC is to produce a correlated sample from distributions that eventually converge to the target distribution whose normalizing constant may be unknown. The outcome of this procedure is not only the posterior mean and variance but also the entire posterior distribution of the parameter vector, or any explicit function of the parameters. A direct implication is that no asymptotic approximation is needed to describe the variability of

estimates as is often done in maximum likelihood analyses [*Gelman et al.*, 1995, p. 4].

[18] MCMC can be used to generate a sample from any probability distribution whose density is known up to a multiplicative constant. Given a full joint distribution (prior+likelihood) the goal is to sample values from the posterior distribution of the unknown parameters given the data. The standard software for Bayesian analysis through MCMC simulations is WinBUGS 1.3 [*Spiegelhalter et al.*, 2000]. WinBUGS uses Gibbs sampling and Metropolis-within-Gibbs algorithm for difficult full conditional distributions. The basic idea of Gibbs sampling is to successively sample from the conditional distribution of each random variable given all the others. Under weak conditions this process eventually provides samples from the joint posterior distribution of the unknown parameters [*Spiegelhalter et al.*, 2000]. *Crainiceanu et al.*, 2002a, provide a discussion of the application of WinBUGS to hierarchical models of count data, and recommendations for model formulation to improve numerical performance.

4. Bayesian Modeling of Pathogen Counts

[19] The basic data set for this analysis is the number of pathogens counted by a laboratory in a sampled volume of water. Therefore it is crucial to identify and understand the processes that affect the number of pathogens counted. The relatively simple process of counting microscopic pathogens is subject to complex influences, some quantifiable and systematic, and some random.

[20] Consider a given water sample of volume V taken from a natural water body at a given time. Assuming that the pathogens are homogeneously distributed in water and denoting by C the unobserved pathogen concentration, one expects to have on average $N = C \times V$ pathogens in the volume of water sampled. However, the counting process is imperfect and subject to uncontrollable limitations in the retention and identification of microorganisms. Therefore, the expected number of pathogens actually counted is only a fraction, R, of the total number N in the water. R is called the recovery rate. Thus, the expected number of organisms counted is $R \times V \times C$.

[21] While the outcome of the counting process is the number of pathogens observed, the unobserved pathogen concentration, C, is the quantity of concern for health risk analysis and new water treatment regulations. A variety of factors could have a causal relationship with pathogen concentrations, such as water quality (turbidity, temperature), basin characteristics (land usage, standardized flow or residence time, type of water body), seasonality, and spatial and temporal correlations. One can link the concentration to these factors using a standard linear regression model for the logarithm of concentration with random effects to account for correlations, as shown in section 4.1 below [see also *Best et al.*, 2002].

[22] Variations in the recovery rate, or the probability of identifying and counting one microscopic pathogen in water, introduces additional complexity. For example, a zero or other low count can be due either to a low pathogen concentration, or to a low laboratory recovery rate, or both. Because of this inherent duality in the interpretation of lab results, additional data is needed to understand the properties of recovery rates. The EPA addressed this issue with a "spiking survey" wherein laboratories were given samples of water containing a known, but undisclosed, number of oocysts. The ratio of the observed counts to the exact counts is an estimator of the recovery rate. Section 4.2 develops a model for recovery rates incorporating water quality covariates and laboratory random effects.

4.1. Bayesian Model Formulation

[23] This section considers the following model of pathogen counts:

$$\begin{aligned} Y_{ij}|\lambda_{ij} &\sim \text{POISSON}(\lambda_{ij}) \\ \lambda_{ij} &= V_{ij} R_{ij} C_{ij} \\ \log(C_{ij}) &= X'_{ij}\beta + t_{ij} \\ t_{ij}|s_i, \sigma_t &\sim N(s_i, \sigma_t^2) \\ s_i|r_{k(i)}, \sigma_s &\sim N(r_{k(i)}, \sigma_s^2) \\ r_k|\mu, \sigma_r &\sim N(\mu, \sigma_r^2) \end{aligned} \tag{5}$$

[24] Here Y_{ij} is the observed pathogen count for site i and month j. Conditional on their mean λ_{ij}, the pathogen counts Y_{ij} have Poisson distributions. The mean of the distribution of the observed pathogen count λ_{ij} is the product of the volume of water analyzed V_{ij}, the true concentration of microorganisms C_{ij}, and the recovery rate probability R_{ij}. A model of R_{ij} is developed in section 4.2.

[25] The site number i runs from 1 to M, where M is the number of sites considered. The month index j runs from 1 to N_i, where N_i is the number of observations for site i. In general $N_i = 18$, but there are sites with $N_i < 18$ due to missing observations. Site i with mean s_i belongs to geographic region $k(i)$ whose mean is r_k. In a few cases 2 or 3 sampling points (locations from which utilities drew water) were actually the same water source; these records were combined into a single super-site with up to 54 observations [see *Behr*, 2001, pp. 15–18]. As a result, the number of stream sites was reduced from 111 to 87, while the number of reservoir-lake sites was reduced from 185 to 147. This was important because the Weber River in Utah appeared to be unusual in that region as well as nationally, and records for 3 Weber-River withdrawal points were combined into one super-site.

[26] The third relationship in the model links variability in the pathogen concentrations C_{ij} to available covariates X_{ij} and a random error component. This relationship is approximated by the log linear model

$$\log(C_{ij}) = X'_{ij}\beta + t_{ij}. \tag{6}$$

A basic assumption in (6) is that the pathogen concentration in natural waters can be very small but never exactly zero. The log transformation converts the concentration $C_{ij} > 0$ to $\log(C_{ij})$, which is unbounded both above and below, and thus is consistent with a standard linear regression analysis with normal errors. Each entry of the vector X_{ij} represents a covariate which may be a nonlinear function of an observed and measured quantity. Possible covariates include turbidity, temperature, streamflow, residence time, source water type and seasonal spline functions

(see Appendix A). The vector β contains the regression model parameters.

[27] Conditional on their means, the random time-site effects t_{ij}, site effects s_i and regional effects r_k are independent and normally distributed. This is an assumption about their prior distribution. Although independence and normality may not be an entirely satisfactory assumption, there was little prior information to enable us to form an alternative formulation. The estimates of these effects should depend mostly on the data and little on the prior assumptions. The time-site effects capture longitudinal within-site variation, the site effects represent between-site variation in the site means within geographic regions, and the regional effects capture variation between the EPA regions within the United States. These effects are an important aspect of the model because they jointly capture the stochastic variation in concentrations over time at a single site, the unique character of each site within a region, and also allow for regional differences. The resulting spatial correlation structure along with R_{ij} describe the observed overdispersion in the actual counts.

[28] The spatial and temporal stochastic dependence of the data is modeled by a hierarchy of normal random effects. As formulated in equation (5), the time-site random effects t_{ij} are centered on the random site effects s_i, the s_i are centered on the regional random effects $r_k(i)$, and the r_k are centered on the national mean μ. This model formulation significantly improves the performance of the MCMC simulation which was an important consideration [*Crainiceanu et al.*, 2002a]. Covariates were centered by subtracting their sample mean and standardized by dividing by their sample standard deviation. Centering covariates is widely used in statistical computation to make them orthogonal to the intercept and thus improve MCMC mixing properties [*Crainiceanu et al.*, 2002a, *Spiegelhalteret al.*, 2000]. Standardizing covariates also makes them unitless and allows direct comparison of their effects on concentration.

[29] *Messner and Wolpert* [2002] consider a set of models similar to that in equation (6), with site effects but no regional effects, a single continuous covariate representing turbidity, and an index variable for each of the twelve months. In addition, they used a beta distribution to describe independent identically distributed recovery rates, whereas our analysis employs a logit-normal recovery rate distribution that includes a statistically significant covariate corresponding to the volume of the sample analyzed. Their analysis of the basic ICR data set and supplemental survey for both *Cryptosporidium* and *Giardia* serves as the basis of the proposed EPA regulations on water treatment [*U.S. EPA*, 2001]. Their analysis focuses on the median, 5th and 95th percentiles of the national distribution of site means. This paper focuses on the performance of the basic model, selection of a recovery rate model and of covariates, and the role of observations, site means, and regional effects.

4.2. Recovery Rates

[30] For a given laboratory, the random recovery rate R is the probability that a lab technician observes and counts a pathogen included in the sample. The precision of laboratory recovery rates for *Cryptosporidium* has long been of concern [*Gimbel and Nahrstedt*, 1996; *Walker*, 1999, chap. 18; *Bukhari et al.*, 1999; *Young and Komisar*, 1999; *Connell et al.*, 2000]. Because the mean recovery rate for *Cryptosporidium parvum* is approximately 11% [*Messner*, 2000], failure to include recovery in the model would result in severe underestimation of concentrations C_{ij} and exaggeration of their variability [*Stedinger and MacKay*, 1998]. The pathogen counts have two main sources of overdispersion: variation in true concentrations and potential variation in laboratory recovery rates. These sources should be distinguished because they have different implications for risk management.

[31] The EPA conducted a lab-spiking study to determine the distribution of recovery rates R for the ICR analysis [*Scheller et al.*, 2002]. *Messner* [2000] reports that: "The ICR Laboratory Spiking Study was designed to assess the ICR Method's performance when testing actual drinking water sources. For this purpose, a subset of ICR utilities (70) collected 100 L volumes of their source waters on two separate occasions concurrent with ICR sample collection. The 100 L samples were shipped to a central lab, spiked with a known number of *Cryptosporidium* and *Giardia* organisms, filtered according to the ICR Method, and (the filters) shipped to the utilities' selected analytical laboratories." For every experiment, collected data included the volume of water filtered (which is in general different from 100 L), number of organisms added to this volume of water (for *Cryptosporidium* and *Giardia*), standard deviation of the number of organisms spiked (variation is due to the technique used: hemacytometer enumeration), number of organisms actually counted by the laboratories, volume of water analyzed by the laboratories (which is typically smaller than the volume of water filtered), and three variables (turbidity, temperature and pH) measured at the time of filtration, and laboratory identification number. A total of 21 laboratories participated in the study.

4.2.1. Recovery Rates Model

[32] There are many imprecisely controlled processes in the ICR method that could be responsible for variability. Some organisms could pass through or be trapped in the filter; others could be lost during the flotation or staining procedure, or are hidden by other particles in the water [*Young and Komisar*, 1999]. Even for a given probability of counting an organism (recovery rate), the number of oocysts counted by the laboratory is still random. The water spiking process with a large and nondeterministic number of pathogens is another source of variability. Let j be the j-th experiment at laboratory i, N_{ij} be the number of oocysts spiked in the total volume of water T_{ij}, and Z_{ij} be the observed count. Therefore, the true concentration in the spiked sample is N_{ij}/T_{ij}. Because $E[N_{ij}]$ is generally in the thousands, a continuous Gamma distribution is more than satisfactory for modeling the count N_{ij}. Consider the following model:

$$\begin{aligned} N_{ij}|a_{ij}, b_{ij} &\sim \text{GAMMA}(a_{ij}, b_{ij}) \\ Z_{ij} &\sim \text{POISSON}(\theta_{ij}) \\ \theta_{ij} &= V_{ij}N_{ij}R_{ij}/T_{ij} \\ \text{logit}(R_{ij}) &= W'_{ij}\beta_W + u_{ij} \\ u_{ij}|L_i, \sigma_u &\sim N(L_i, \sigma_u^2) \\ L_i|\delta, \sigma_L &\sim N(\delta, \sigma_L^2) \end{aligned} \tag{7}$$

Given the nature of the spiking step, which involves an imperfect counting process of very small organisms, the exact number N_{ij} is unknown. We account for its variability by assuming that the true number of oocysts in the spike has a Gamma distribution. For the ICR spiking *Cryptosporidium* data, the reported coefficient of variation (CV) of the N_{ij} distributions is between 9.7% and 29.9% with an average CV of 15.9%. The lab-spiking data provides the expected value, $E[N_{ij}]$, and variance, $\mathrm{Var}[N_{ij}]$ of the number of organisms spiked. One important assumption is that this process is unbiased. The parameters a_{ij} and b_{ij} of the gamma distribution are chosen so that its mean and variance match $E[N_{ij}]$ and $\mathrm{Var}[N_{ij}]$ respectively:

$$\begin{aligned} E\left[N_{ij}\right] &= a_{ij}/b_{ij} \\ \mathrm{Var}\left[N_{ij}\right] &= a_{ij}/b_{ij}^2 \end{aligned} \tag{8}$$

Conditional on the true concentration N_{ij}/T_{ij}, the expected number of oocysts in the volume analyzed V_{ij} is $V_{ij}N_{ij}/T_{ij}$. However, because of lack of accuracy in the laboratory counting process, the expected number of oocysts is smaller; the correction factor R_{ij} is the recovery rate. Conditional on the recovery rate and on the true concentration in the volume of water analyzed, the number of oocysts actually counted has a Poisson distribution with mean $\theta_{ij} = V_{ij}N_{ij}R_{ij}/T_{ij}$.

[33] The third relationship in the model links the variability in the recovery rate to the value of available covariates and a random error component. This relationship is described by the logit model

$$\log\left[R_{ij}/\left(1-R_{ij}\right)\right] = W'_{ij}\beta_W + u_{ij}, \tag{9}$$

where $W'_{ij}\beta_W$ is a linear combination of the covariates using parameters β_W, and the u_{ij} are conditionally independent random errors with mean L_i representing the laboratory effect. Laboratory effects have normal distributions about the overall mean for the model δ with variance σ_L^2. The role of the logit transformation is to convert the recovery rate $0 < R < 1$ into a variable that is unbounded both above and below, and thus is consistent with a linear model including normally distributed within-lab variation and laboratory effects. In the statistics literature model (7) would be referred to as a Generalized Linear Mixed Model, because u_{ij} and L_i are random. Appendix B discusses alternative recovery rate models.

4.2.2. Recovery Rate Inference

[34] Of interest is whether a statistical analysis will find any significant covariates or laboratory effects when the observed counts were small, and the magnitude of the initial spike included uncertainty.

[35] The initial data set included 140 observations for the recovery rate of *Cryptosporidium parvum* and *Giardia lamblia*. Because for one observation no covariates were measured, and for two other observations temperature was not recorded, we discarded those three observations. For models that do not include covariates, results for the entire data set (containing 140 observations) were almost identical to results for the reduced data (containing 137 observations). Because some very large values of turbidity were observed (e.g., 1200 ntu and 394 ntu, when the other observations had mean turbidity 9.58 ntu with standard deviation 17.04 ntu), all turbidity values above a threshold of 30 ntu were set equal to that threshold, while values below were left unchanged. This transformation affected about 10% of the observations, and results were not sensitive to the selected bound.

Table 1. Results of Bayesian Analysis for the Recovery Rate Models of *Cryptosporidium*: Posterior Means and Standard Error[a]

	Cryptosporidium		Giardia	
Parameters	Full	Reduced	Full	Reduced
Overall mean	−2.47 (0.12)	−2.39 (0.10)	−1.20 (0.22)	−1.19 (0.21)
Turbidity	0.13 (0.12)		−0.15 (0.11)	
Temperature	0.20 (0.11)		0.04 (0.10)	
pH	0.05 (0.09)		0.03 (0.08)	
log-V	−0.43[b] (0.13)	−0.42[b] (0.10)	−0.35[c] (0.17)	−0.35[b] (0.12)
Log-N	0.09 (0.11)		−0.08 (0.09)	
σ_u	1.10 (0.09)	1.10 (0.09)	1.00 (0.09)	0.98 (0.08)
σ_L	0.12 (0.09)		0.69 (0.24)	0.70 (0.27)
RS	0.02 (0.03)		0.32 (0.15)	0.33 (0.16)

[a]Standard errors are given in parentheses.
[b]Posterior symmetric 95% credible interval of the fixed effect does not contain 0.
[c]Posterior symmetric 90% credible interval of the fixed effect does not contain 0.

[36] Table 1 reports the posterior means and standard deviations for the parameters of four recovery rate models. Analyses employed WinBUGS 1.3 with diffuse proper priors for all hyperparameters. The first two columns correspond to two versions of the model in equation (7) applied to *Cryptosporidium* data. The first column provides results for a model including all available covariates and laboratory random effects. For *Cryptosporidium*, the logarithm of the volume of water analyzed (log-V) is statistically significant (the 95% credible interval of its parameter does not contain zero). Turbidity, temperature, pH, log of the number of oocysts spiked (log-N) were not statistically significant.

[37] To examine the importance of laboratory effects in explaining the variability in the data we computed the posterior distribution of

$$RS = \frac{\sigma_L^2}{\sigma_L^2 + \sigma_u^2}.$$

The *RS* statistic is inspired from the standard linear regression and represents the fraction of the total variance explained by the random laboratory effects. For *Cryptosporidium RS* is smaller than 0.1 with 0.95 probability and smaller than 0.14 with 0.975 probability, showing that laboratory random effects explain only a very small fraction of the variation in the *Cryptosporidium* data.

[38] The second column provides results for a *Cryptosporidium* recovery rate model that only includes log-V as a covariate and does not contain laboratory random effects. The log-V parameter does not change and remains significant. This second recovery rate model will be used in the complete analysis of *Cryptosporidium* concentrations.

[39] Simple, nonparametric estimates of the recovery rates can be obtained by ignoring the Poisson variability in model (7) which gives the "empirical recovery rates"

$$\hat{R}_{ij} = \frac{Z_{ij}T_{ij}}{N_{ij}V_{ij}}.$$

Figure 2 displays these values for *Cryptosporidium* relative to the log of the volume analyzed. The $\widehat{R}_{ij}$ are larger

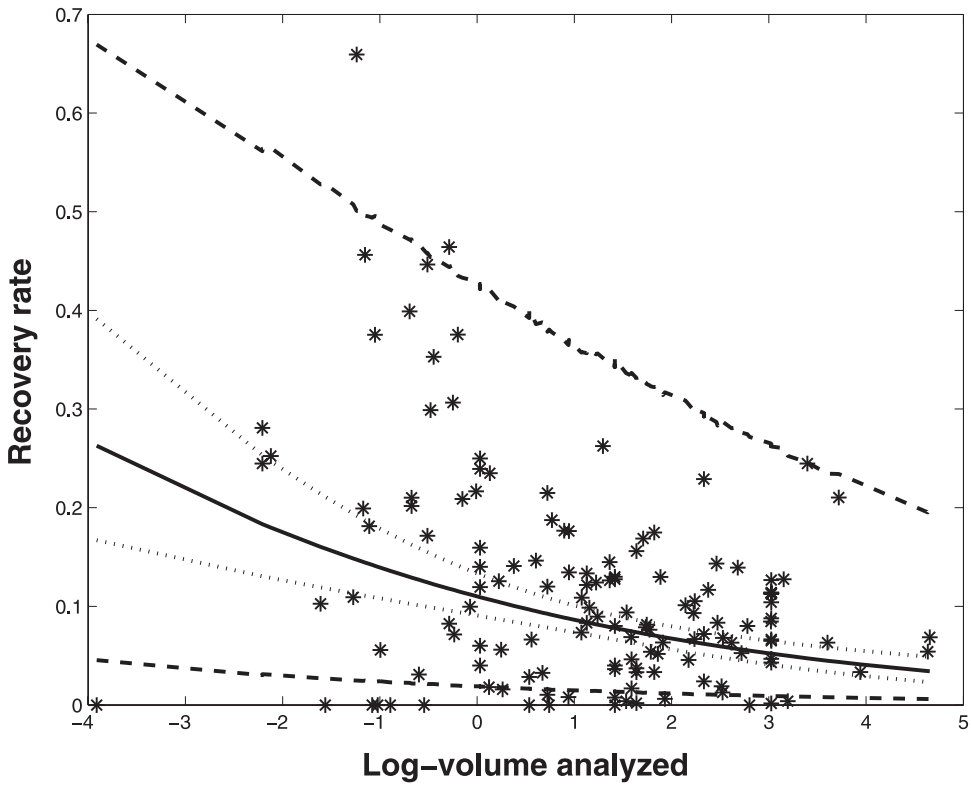

Figure 2. Stars are empirical recovery rates; solid line is median recovery rate; dotted line is 90% credible interval on the median recovery rate; short-dashed line is 90% credible interval for the unknown true recovery rate.

for smaller values of the log-V, which is consistent with our model-based findings for the recovery rates. The advantage of model (7) is that it quantifies this relationship.

[40] *Messner* [2000] analyzed the same data set to derive the MLEs and the posterior distribution of the parameters of a Beta distribution for R_{ij} using a Poisson model for the observed counts, but neglecting any covariates or laboratory effects. *Walker* [1999] employed the empirical recovery rates in a GLM Gamma regression analysis thus neglecting the sampling variability both in Z_{ij} and N_{ij}. See equation (B2). This seems inappropriate because several Z_{ij} were zero, as shown in Figure 2.

[41] The median of R_{ij} can be obtained by replacing u_{ij} by its median δ in equation (9). The solid line and the dotted lines in Figure 2 represent the posterior 5th, 50th, and 95th percentiles of the true median recovery rate. The dashed lines in Figure 2 represent the posterior 5th and 95th percentiles of the posterior distribution of possible recovery rates.

[42] The negative effect of log-V on *Cryptosporidium* recovery rates (-0.42) is shown by the decreasing median recovery rate in Figure 2. The recovery rate credible intervals are wider than the credible intervals for the median recovery rate due to the additional variability represented by the random effects u_{ij} around δ. The recovery rate credible intervals do not contain exactly 90% of the empirical recovery rates since, because of variability of the counts, the empirical rates are more variable than the actual rates.

[43] The last two columns of Table 1 present results for *Giardia* recovery rates. The overall mean on the log-scale increases from -2.47 for *Cryptosporidium* to -1.20 for *Giardia* showing larger recovery rates. This may be due to the larger size of the *Giardia* cysts relative to the *Cryptosporidium* oocysts, which makes them easier to retain and identify. As with *Cryptosporidium* the only covariate that has a sizable effect on recovery rates is log-V (-0.35). However, the fraction of the total variance explained by the random laboratory effects (*RS*) is much larger (0.32 for *Giardia* relative to 0.02 for *Cryptosporidium*) showing that for *Giardia* there are sizable differences in recovery rates among the 21 laboratories. The last column presents results for the simple model of recovery rates for *Giardia* with all nonsignificant covariates omitted.

[44] The physical relationship that causes recovery rate to drop with the volume of water analyzed is not clear. It could be related to the suspended solids in the water that results in the reduction in the volume of water run through the filter before clogging, and the collected solids from which a sample was taken for inspection. One needs to be careful using volume analyzed as a recovery rate covariate because this value could be manipulated by utilities and laboratories, and the derived relationship pertains to the historical sampling and laboratory procedures. Bias would be introduced into the analysis if utilities or laboratories change their procedures and sampling rules, and thus the relationship between real recovery rates and the volume analyzed.

5. *Cryptosporidium* Concentration Results

[45] If one is interested in real-time prediction of pathogen concentrations for day-to-day water supply management, it is reasonable to use all available covariates. Such a model

Table 2. Posterior Means and Posterior Standard Deviations of Parameters for Eight *Cryptosporidium* Models[a]

	Models							
Parameters	S-NOV	S-V	S-V-S	S-HR	R-NOV	R-V	R-V-S	R-HR
Overall mean	−1.93 (0.42)	−1.83 (0.46)	−1.69 (0.46)	−1.63 (0.47)	−4.97 (0.59)	−4.27 (0.58)	−4.01 (0.56)	−4.19 (0.56)
Turbidity	0.87[b] (0.16)	0.67[b] (0.14)	0.55[b] (0.15)		0.54[b] (0.22)	0.53[b] (0.20)	0.54[b] (0.19)	
Temperature	−0.30 (0.23)	−0.29 (0.21)			−0.32 (0.32)	−0.38 (0.32)		
T-coli	−0.41[c] (0.29)	−0.42[c] (0.29)			0.12 (0.17)	0.08 (0.16)		
Season 1	0.08 (0.13)	0.06 (0.12)	0.05 (0.12)	−0.02 (0.12)	−0.08 (0.24)	−0.08 (0.20)		
Season 2	−0.36[c] (0.19)	−0.33[c] (0.18)	−0.16 (0.11)	−0.29[c] (0.10)	−0.21 (0.26)	−0.22 (0.24)		
Season 3	−0.20 (0.17)	−0.20 (0.16)	−0.04 (0.12)	−0.03 (0.12)	−0.16 (0.25)	−0.17 (0.23)		
pH	0.37[c] (0.19)	0.27 (0.17)			−0.20 (0.29)	−0.24 (0.27)		
Log-pop.	−0.22 (0.27)	−0.29 (0.26)			0.32 (0.28)	0.31 (0.25)		
Resid. time					−0.35 (0.28)	−0.34 (0.25)		
Export-L	−0.49[c] (0.28)	−0.49[c] (0.26)			−0.19 (0.28)	−0.21 (0.27)		
σ_t	1.11 (0.16)	1.02 (0.16)	0.99 (0.17)	1.03 (0.16)	1.67 (0.31)	1.32 (0.30)	1.27 (0.30)	1.46 (0.29)
σ_s	1.58 (0.23)	1.37 (0.22)	1.34 (0.22)	1.36 (0.23)	1.62 (0.37)	1.46 (0.30)	1.32 (0.28)	1.41 (0.29)
σ_r	0.73 (0.51)	0.99 (0.48)	1.02 (0.45)	1.06 (0.49)	0.90 (0.61)	0.90 (0.56)	0.90 (0.47)	0.72 (0.48)

[a]Posterior standard deviations are given in parentheses. S-NOV: stream data, recovery rate that does not incorporate log-volume analyzed. S-V: stream data, recovery rate incorporating log-volume analyzed. S-V-S: stream data, recovery rate incorporating log-volume analyzed, statistically significant covariates only. S-HR: stream data, recovery rate incorporating log-volume analyzed, health risk model. R-NOV: reservoir data, recovery rate that does not incorporate log-volume analyzed. R-V: reservoir data, recovery rate incorporating log-volume analyzed. R-V-S: reservoir data, recovery rate incorporating log-volume analyzed, statistically significant covariates only. R-HR: reservoir data, recovery rate incorporating log-volume analyzed, health risk model.
[b]The 95% equal tail probability credible interval does not contain 0.
[c]The 90% equal tail probability credible interval does not contain 0.

could be used to quickly determine periods with a risk of higher pathogen concentrations so that utilities could take appropriate actions (e.g., intensify water quality monitoring, increase levels of water treatment or switch water sources).

[46] For health risk analysis associated with long-term exposure, one is interested in the annual average concentration at a given site and its predictive distribution for decision-making [*US EPA*, 2001]. Because future values of some covariates are not known, they are not useful for long-term risk analysis, though their site-specific historical averages could be used. For risk analysis we want to retain in the model all covariates that are available for long term prediction. Therefore, we divide the covariates into two categories.

[47] A first category includes time-site specific covariates (turbidity, temperature, water-flow, etc.), whose future values are not known at present. The second category includes perfectly predictable covariates (type of water source, seasonal effects, basin characteristics, etc.). Their exact values are known at present for every site and future period. Moreover, if there is a missing observation for a given month, then the time-site specific covariates (e.g., turbidity) are unavailable whereas the perfectly predictable covariates are known, whether or not a sample was taken for site i in month j. Grouping covariates into these two distinct categories results in two sets of models, one for water quality prediction using all covariates, and one for risk analysis using only perfectly predictable covariates. *Messner and Wolpert* [2002] took another approach which was to fit lognormal distributions to the sets of at-site averages computed at each MCMC iteration. The median of those distributions then represented their median national distribution for at-site *Cryptosporidium* means, including parameter uncertainty. The 5th and 95th percentiles were also computed.

[48] Model (5) can be applied to different pathogen data sets, including the entire ICR data or subsets that exhibit special characteristics not found in the rest of the data. The ICR data set was partitioned according to the water source type (streams versus reservoirs and lakes), because stream water sources are likely to exhibit higher *Cryptosporidium* concentrations and different causal relationships between oocyst counts and covariates. Moreover, some covariates such as residence time and flow rate are defined for one water source but not the other.

[49] Table 2 provides posterior means and standard deviations for a representative set of parameters for eight selected models for *Cryptosporidium* based upon 1444 observations for streams and 1771 observations for reservoirs/lakes. Results were obtained using 4000 burn-in simulations (not used for inference) and 20,000 simulations from the target distribution (used for inference). As convergence diagnostics we used multiple chains with different initial starting points for parameters and visual inspection of the chains corresponding to parameters of interest.

[50] Each column corresponds to one particular case of the model described in equation (5). The first column (S-NOV) corresponds to results for model (5) applied to stream data incorporating many available covariates and using a recovery rate model that does not include log-volume analyzed as a covariate. Model S-V includes the log-volume analyzed as an explanatory variable for recovery rates, but not laboratory effects and nonsignificant covariates (temperature, pH, log-number of oocysts spiked). This refinement of the recovery rate model has a large impact on the effect of turbidity on concentrations. While statistically significant in both models, its effect is reduced from 0.87 for S-NOV to 0.67 for S-V, or almost 25%. A similar impact can be observed for pH, but the value of pH in both models is very modest.

[51] Complex models such as S-NOV and S-V, that include a large number of covariates, can be used in a first phase of statistical modeling to identify those variables that affect pathogen concentrations. However, in our models many covariates were not statistically significant and could be discarded from the analysis to help focus the analysis on covariates that are important. The third column (S-V-S) of

Table 2 corresponds to a simpler case of model S-V from which most covariates with small effects were discarded. The S-V-S model retains only turbidity and seasonal functions as explanatory variables for *Cryptosporidium* concentrations. The effect of turbidity is further reduced to 0.55 but remains significant, while none of the the seasonal functions is statistically significant.

[52] To obtain a model for health risk analysis it is easiest to use only perfectly predictable covariates. An example of risk analysis model for streams is presented in the fourth column (S-HR) and is obtained by dropping turbidity from model S-V-S. One can see that the seasonal splines described in Appendix A become statistically significant, most probably by capturing the seasonal effects previously captured by turbidity.

[53] Reservoir/Lakes results for four models are presented in the last 4 columns of Table 2. Notation is the same as for streams models. The only significant covariate in the first 3 models (R-NOV, R-V, R-V-S) is turbidity whose effect (0.53) is practically unchanged by the various transformations of the model. Using a more refined model for recovery rates reduces the posterior means of the standard deviations for time-site effects σ_t (from 1.67 in R-NOV to 1.32 in R-V, or 21%) and for site effects σ_s (from 1.62 in R-NOV to 1.46 in R-V, or 10%) but left it unchanged for region effects (0.90). Eliminating the noise contained in the nonsignificant covariates further reduces σ_t. For the health risk model R-HR the inability to use the time-site specific turbidity, which was statistically significant in the other models, moderately inflates σ_t and σ_s, and decreases σ_r. As with the stream models, σ_r is the most uncertain of the variance components; this should not be a surprise because there are only 10 regions whereas there are 84 stream sites and 114 reservoir/lakes sites, most of which have 18 monthly observations which help determine σ_t, the most precisely determined variance component in the stream models. For reservoir-lakes σ_t and σ_s had the same precision most likely because so many counts were zero.

[54] The mean count for streams is much larger than for reservoirs (0.48 compared to 0.08) and the same is true for the percentage of nonzero counts (13.1% compared to 4.2%). Because the covariates are mean-centered, differences between reservoir and stream concentrations are captured by the differences in the overall mean on the natural log-scale ($\mu = -1.83$ for S-V versus $\mu = -4.27$ for R-V). This difference between *Cryptosporidium* concentrations in streams and reservoirs is very large compared to all other covariate effects and is by far the most important effect identified in the ICR *Cryptosporidium* data.

[55] Of interest is the importance of the hierarchical structure accounting for time-site, site and regional effects. The importance of each is reflected in the variance components σ_t^2, σ_s^2 and σ_r^2. For stream models, the largest variance component is the site effect, followed closely by the time-site and regional effect. Of the total variation in concentrations explained by the site mean, roughly 30% $[\sigma_r^2/(\sigma_r^2 + \sigma_s^2) \times 100\%]$ of that variation could be explained by region, both for streams and reservoirs. This illustrates the importance of including regions in the model to correctly represent the stochastic dependence among observations, as well as the potential value of knowing a region upon estimation of site means.

Table 3. Posterior Means and Posterior Standard Deviations of Parameters for Four Models of *Cryptosporidium* Concentrations in Reservoirs

Parameters	Models			
	Full	Model 1	Model 2	Model 3
Overall mean	−4.32 (0.75)	−4.00 (0.69)	−3.96 (0.61)	−3.92 (0.58)
Turbidity	0.53[b] (0.23)	0.49[b] (0.20)	0.46[b] (0.20)	0.45[b] (0.19)
Carbonate hard.	0.63[b] (0.23)	0.62[b] (0.22)	0.58[b] (0.22)	0.62[b] (0.21)
Total organic carbon	0.33[b](0.15)	0.35[b](0.13)	0.36[b] (0.13)	0.38[b] (0.13)
Log-Urban land area	−0.26 (0.28)	−0.21 (0.25)	−0.16 (0.21)	
Log-residence time	−0.12 (0.27)	−0.13 (0.24)	−0.15 (0.25)	
Sediment export	−0.15 (0.22)	−0.12 (0.20)		
Export-L	−0.18 (0.29)	−0.19 (0.27)		
Temperature	0.19 (0.18)			
Season 1	−0.05 (0.23)			
Season 2	−0.09 (0.19)			
Season 3	−0.09 (0.22)			
Log-population	0.06 (0.24)			
Soil permeability	0.02 (0.21)			
σ_t	1.52 (0.41)	1.35 (0.40)	1.33 (0.33)	1.38 (0.32)
σ_s	0.92 (0.43)	0.81 (0.39)	0.91 (0.37)	0.78 (0.36)
σ_r	1.15 (0.56)	1.05 (0.48)	1.01 (0.51)	1.06 (0.49)

[a]Posterior standard deviations are given in parentheses. Different models are obtained from the full model by dropping covariates that are not statistically significant. A smaller subset than the one used for results in Table 2, because fewer covariates were available.

[b]Posterior symmetric 95% credible interval of the fixed effect does not contain 0.

[56] Table 3 consists of posterior means and standard deviations for several models of *Cryptosporidium* at reservoir sites only. The models have been adapted from earlier analyses in which a smaller sample of the ICR data set is explored (1469 observations) to accommodate a larger set of covariates [*Behr*, 2001]. The basic structure of the models is given in equation (5), where again log-volume is included in the model for recovery rates. The results indicate that turbidity, carbonate hardness, and total organic carbon are all significant. The regional variance component σ_r^2 is now larger than the site effect variance σ_s^2. The decrease in σ_t from 1.52 to 1.33 is difficult to interpret because the overall mean increases at the same time from −4.32 to −3.96; the probability of a nonzero count depends upon the joint distribution of both parameters.

6. Model-Based Prediction of Pathogen Distributions

[57] Our hierarchical Bayesian models are of necessity complex because the data's structure is complex. A major attraction of the model is that the complexity is built by exploiting relatively simple conditional relationships, so that the model is reasonably easy to understand. To further understand the model, we now study how it predicts under different sets of available information. We will use prediction intervals to illustrate the following points: (1) Prediction intervals for log concentrations become narrower when there is more information available on actual count, site, or region. (2) The location and width of

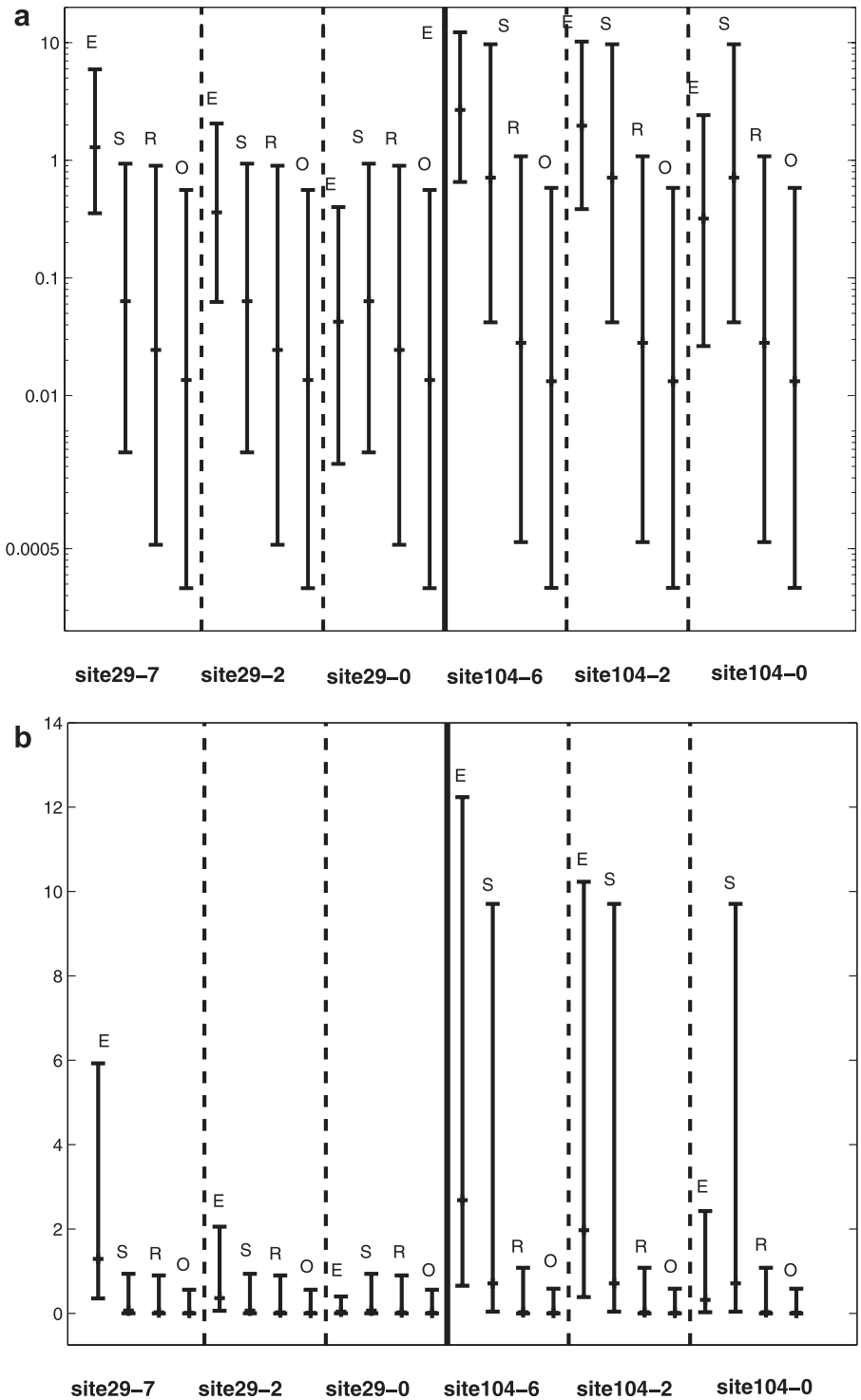

Figure 3. Posterior 5th, 50th, and 95th percentiles of *Cryptosporidium* concentrations using model R-HR, conditional upon observations at sites 29 and 104, with different amounts of information, as explained in the text. Figure 3a has log-scale vertical axis, and Figure 3b has linear vertical axis.

prediction intervals are affected by covariates, if they are available and included in the model. (3) When predicting concentrations rather than log concentrations, heteroscedasticity is severe and prediction intervals are considerably wider when the medians of the predicted values are larger. (4) The heteroscedasticity effect can dominate the effect of increased information: observing a high count, rather than having no observation, increases the prediction interval's width despite the additional information the observation provides.

[58] The information displayed in Figures 3 and 4 was produced by the MCMC algorithm as it samples the posterior distribution. At each step in the Markov chain, the algorithm generates a sample from the posterior of possible values for all parameters and random effects. These provide the corresponding concentration for a given site and

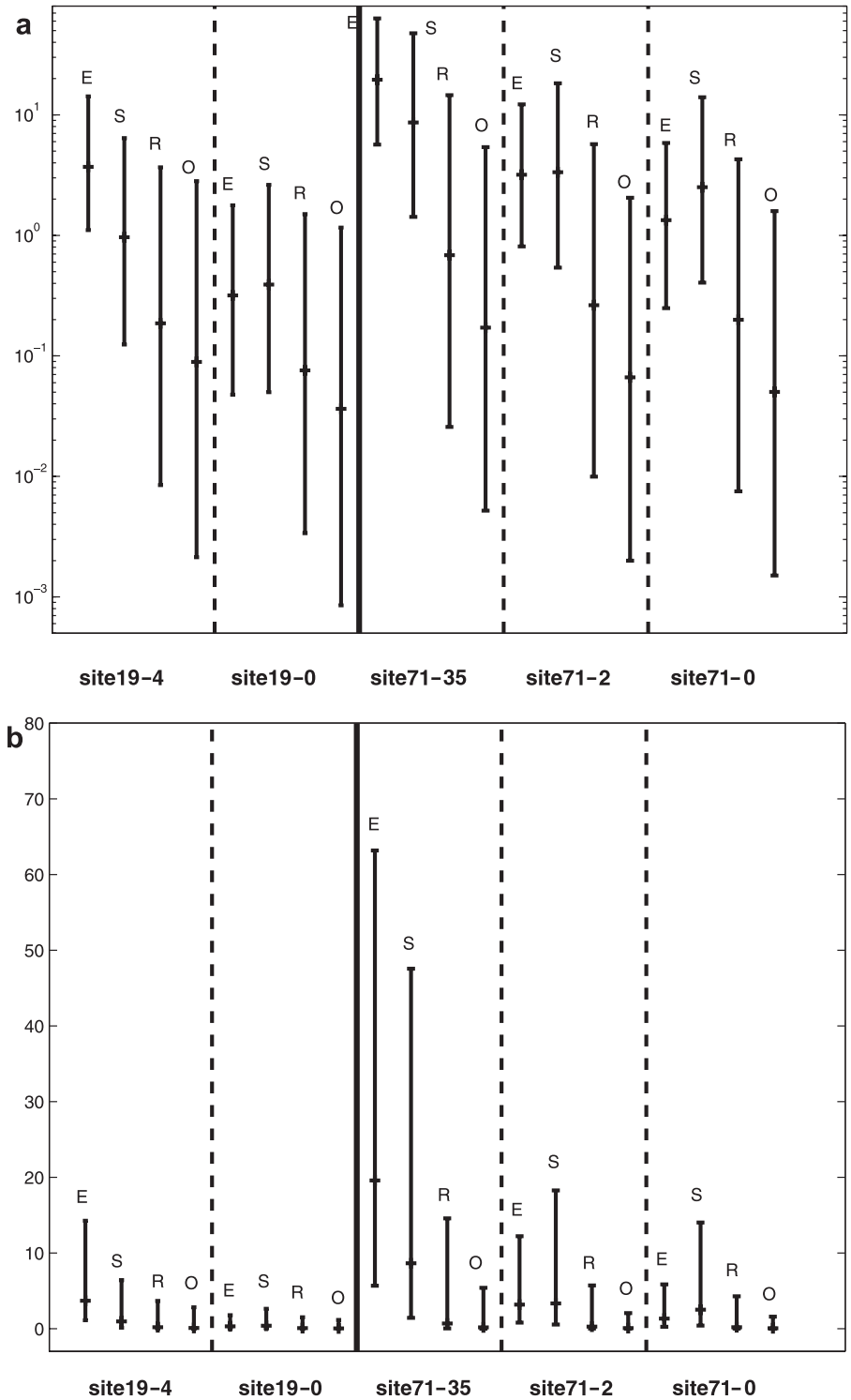

Figure 4. Posterior 5th, 50th, and 95th percentiles of *Cryptosporidium* concentrations using model S-V. Figure 4a has log-scale vertical axis, and Figure 4b has linear vertical axis.

time. That *Cryptosporidium* concentration distribution is labeled "E." The posterior E distribution depends upon the count observed at the time and site where concentration is being predicted. Figures 3 and 4 report the 5th, 50th and 95th percentiles of the generated distribution.

[59] The "S" distribution corresponds to same site and month as "E" (with the same covariates) but at a time with no observed count. To get this distribution, at each step of the chain, a t_{ij} is generated from its prior distribution $N(s_i, \sigma_t^2)$ to compute a value of the environmental concentrations for *Cryptosporidium*. The "S" distribution is the prediction of the *Cryptosporidium* concentration distribution at this site, without a concurrent observation.

[60] The "R" distribution uses random unconditional values of the time-site effect t_{ij} and the site effect s_i for a selected region. The "R" predictive distribution describes a site without count data so that the only information for predicting concentrations is knowledge of the region and covariates.

[61] Finally, the "O" distribution corresponds to the *Cryptosporidium* concentrations likely to be observed in a randomly selected region, at a randomly selected site, on the specified date of sampling, without using sample information for that site. The quantiles in Figures 3 and 4 reflect variability in the time-site effects, the site, the regional effects, and also the uncertainty in the regression coefficients β and the variance components of the model: σ_t^2, σ_s^2, and σ_r^2.

[62] The difference between the two graphs in Figure 3 is that Figure 3a uses a log-scale vertical axis while Figure 3b uses a linear scale. Thus Figure 3a shows the effects of information on log-concentrations equal to $X'_{ij}\beta + t_{ij}$, whereas Figure 3b shows the effects on the corresponding real concentrations $C_{ij} = \exp(X'_{ij}\beta + t_{ij})$.

[63] The left-hand side of Figure 3a shows the impact of observing a count of 7, of 2 or of 0 at site 29 in the EPA region 3. The count of 7 results in a posterior distribution represented by "E" that is larger than with a count of 2, and much larger than with a 0 count. However, an observation does not eliminate all uncertainty because of variation in the recovery rates and because the observed Poisson count does not determine the actual concentration.

[64] Similar patterns can be observed among the "S" distributions. Because 5th to 95th percentile ranges for "S" are smaller than for "R" and "O," the log-concentration distribution is tighter when the site is known than when only the region is known, or when the region is selected at random. The same holds when "R" is compared with "O", because that knowledge of the region reduces the uncertainty in the *Cryptosporidium* concentration, though the difference is relatively modest because the regional effect was not very large.

[65] These effects look very different in Figure 3b, where the intervals for "E" are wider than those for "S", "R", and "O" for counts of 7 and 2, but smaller when the count was zero. The uncertainty seen in Figure 3b is mostly driven by heteroscedasticity rather than the relative precision of the predictive distribution, as demonstrated by Figure 3a. Because the R-HR model has no covariates, under "S," "R," and "O" the 5th to 95th percentiles are the same for all time-site pairs.

[66] Figure 4 displays results for the stream model including all covariates (S-V). Estimates of concentrations correspond to two counts, 4 and 0, at site 19 which is in region 3 and three counts, 35, 2, and 0, at site 71 which is in region 8. The basic patterns are the same as in Figure 3 with one important difference: the "S", "R", and "O" estimates are not the same within a site because the covariates have different values which depend on the time of sampling.

7. Conclusions

[67] In natural waters, pathogen concentrations vary over time and space. This paper develops models that describe such variation as functions of water quality and hydrologic covariates, as well as time, site and regional random effects. Modeling the variation of pathogen concentration using observed count data is a significant challenge, especially in the ICR study where over 90% of the observed counts are zero, and at some sites all counts were zero; the large variation in laboratory recovery rates further complicates the analysis [*Messner and Wolpert*, 2002]. From the ICR spiking study we found that volume-analyzed did help explain *Cryptosporidium* and *Giardia* recovery rates, and for *Giardia* there were significant laboratory effects. Including volume-analyzed in the recovery rate model caused a large reduction in the coefficients for turbidity and pH in the *Cryptosporidium* stream model, but not in the reservoir-lake model. Despite the large percent of zeros, the analysis demonstrated that *Cryptosporidium* concentrations were on average larger in streams than in reservoirs and lakes, several covariates were statistically significant, and there were important differences among regions.

[68] This paper develops a fully Bayesian modeling framework for understanding the variation in environmental pathogen concentrations across sites and across time. The hierarchical model captures site and regional effects. The statistical model is applicable even when the historical pathogen counts are subject to sizable variation in recovery rates and include many small counts, zero counts, and missing data. The methodology is also applied to understand laboratory recovery rates wherein one is concerned with discrete counts whose mean is explained by covariates including log-volume analyzed, and laboratory effects; variation in the number of organisms added to the sample was also a concern.

[69] Even though both models are relatively complex, they were easily analyzed with WinBUGS, a standard package for the numerical evaluation of the posterior distribution of a Bayesian model using Markov Chain Monte Carlo simulation. Overall this study shows that hierarchical Bayesian models are an incredibly flexible and numerically feasible general statistical methodology to describe environmental concentrations of pathogen and microbiological organisms.

Appendix A: Covariates for *Cryptosporidium* Recovery Rate and Concentration Models

[70] Turbidity is a measure of cloudiness of water measured at the time of sampling and it is expressed in nephelometric turbidity units (ntu). Two different transformations were employed for turbidity. For the lab-spiking study all values larger than a given threshold

Table A1. Transformations of Covariates for Recovery Rates Model

	Transformations			
Covariates	Log	Threshold	Centering	Standardized
Turbidity	no	yes	yes	yes
Temperature	no	no	yes	yes
pH	no	no	yes	yes
Volume analyzed	yes	no	yes	yes
N-spiked	yes	no	yes	yes

Table A2. Transformations of Covariates *Cryptosporidium* Concentrations

Covariates	Transformations		
	Log	Centering	Standardized
Turbidity	yes	yes	yes
Temperature	no	yes	yes
T-coli	no	yes	yes
pH	no	yes	yes
Population	yes	yes	yes
Residence time	no	yes	yes
Export-L	no	yes	yes

(30 ntu) were set equal to the threshold. For the ICR study the turbidity was log-transformed. Temperature is the temperature of the water at the time of sampling expressed in degrees Celsius. T-coli is the total number of Coliform bacteria found in the sampled water. pH is the pH of water at the time of sampling. log-V is the logarithm of the volume analyzed by the laboratory in the lab-spiking study. The volume analyzed is only a fraction of the total volume spiked with a known, but undisclosed number of oocysts. log-N is the logarithm of the number of oocysts spiked into the total volume of water in the lab-spiking study. Population is the number of people in the cities that use water from a given site. Residence time of reservoirs or lakes is the average time it takes for water to flow from an inlet to the outlet, equal to the reservoir volume divided by the flow rate. Export-L is the proportion of nitrogen export due to livestock. Seasons are four spline functions designed to capture the intraannual variations not captured by the other covariates. We used four spline functions, one for each season starting with summer. The coefficient of one of the spline functions (end of spring beginning of summer) was set to zero to avoid linear dependencies. Like all other covariates, the seasonal spline functions were centered and standardized.

[71] All models were fit with complete data sets, that is each observation included all covariates. Several transformations of covariates from the original ICR data aimed at reducing the impact of outliers and improving numerical stability. Tables A1 and A2 provide a summary of those transformations. The log-transformation was applied before centering and standardization.

Appendix B: Recovery Rate Models

[72] The models in equations (B1) and (B2) below, were also considered for modeling the pathogen recovery rates. These models provided a statistical interpretation of the data that is consistent with that obtained using the basic model (7), both for *Cryptosporidium* and *Giardia* [*Crainiceanu et al.*, 2002b].

[73] The model in equation (B1) is similar to the model (7) with the important difference that the recovery rates R_{ij} are assumed to follow a Beta distribution with mean m_{ij} and variance $m_{ij}(1 - m_{ij})/(\psi + 1)$. Here ψ is considered constant across all Beta distributions, with a small ψ being indicative of highly variable recovery rates. For the model in (B1), the logit of the mean recovery rate, and not of the recovery rate itself, is assumed to follow a linear mixed model with exchangeable random effects. The variability in the recovery rates in model (B1) is described by the Beta distribution, whereas the random effects t_{ij} described that variability in model (7).

$$\begin{aligned} N_{ij}|a_{ij}, b_{ij} &\sim \text{GAMMA}(a_{ij}, b_{ij}) \\ Z_{ij} &\sim \text{POISSON}(\theta_{ij}) \\ \theta_{ij} &= V_{ij}N_{ij}R_{ij}/T_{ij} \\ R_{ij} &\sim \text{BETA}[\psi m_{ij}, \psi(1 - m_{ij})] \\ \text{logit}(m_{ij}) &= W'_{ij}\beta_W + L_i \\ L_i|\delta, \sigma_L &\sim N(\delta, \sigma_L^2) \end{aligned} \tag{B1}$$

[74] Yet a third model is described in equation (B2). The important difference from the basic model in (7) is in the model of recovery rates. Here $\log(-\log(1 - R_{ij}))$ is modeled as a linear mixed model with random lab-time effects $\log(g_{ij})$, where g_{ij} are assumed to have a Gamma distribution. This model was considered because the Gamma distribution in model (B2) is more flexible at zero than the lognormal distribution in equation (7), especially for small recovery rates.

$$\begin{aligned} N_{ij}|a_{ij}, b_{ij} &\sim \text{GAMMA}(a_{ij}, b_{ij}) \\ Z_{ij} &\sim \text{POISSON}(\theta_{ij}) \\ \theta_{ij} &= V_{ij}N_{ij}R_{ij}/T_{ij} \\ \log(-\log(1 - R_{ij})) &= W'_{ij}\beta_W + L_i + \log(g_{ij}) \\ g_{ij} &\sim \text{GAMMA}(c, c) \\ L_i|\delta, \sigma_L &\sim N(\delta, \sigma_L^2) \end{aligned} \tag{B2}$$

This is a generalization of the model developed by *Walker* [1999], who neglected variability in N_{ij} and Z_{ij}, and did not include laboratory effects.

[75] **Acknowledgments.** We wish to thank two reviewers for their careful reading of the original manuscript and for their comments that significantly improved the paper. We appreciate the assistance provided by Russell Walker, Susan Boutros (Environmental Associates Ltd.) and John Fox (US EPA), who were helpful in the formulation of this research project. Sarah Hattery assisted in the development of the literature review. Michael Messner (U.S. EPA) provided significant and consistent guidance along the way, and suggested the use of volume analyzed as a covariate in the recovery rate model. This research was supported by grant R-82795201 from the U.S. Environmental Protection Agency NCERQA STAR program, "Statistical modeling of waterborne pathogen concentrations."

References

Atherholt, T. B., M. W. LeChevallier, W. D. Norton, and J. S. Rosen, Effect of Rainfall on *Giardia* and *Cryptosporidium*, *J. Am. Water Works Assoc.*, *90*(9), 66–80, 1998.

Bates, B. C., and E. P. Campbell, A Markov chain Monte Carlo scheme for parameter estimation and inference in conceptual rainfall-runoff modeling, *Water Resour. Res.*, *37*(4), 937–947, 2001.

Behr, C. T., Modeling *Cryptosporidium* Concentrations: A Bayesian GLMM of Regional Discrete Count Data, M.S. thesis, Cornell Univ., Ithaca, N. Y., 2001.

Best, N. G., K. Ickstadt, R. L. Wolpert, S. Cockings, P. Elliott, J. Benett, A. Bottle, and S. Reed, Modeling the Impact of traffic-related air pollution on childhood respiratory illness, in *Case Studies in Bayesian Statistics*, vol. V, edited by C. Gatsonis et al., pp. 153–253, Springer-Verlag, New York, 2002.

Breslow, N. E., and D. G. Clayton, Approximate inference in generalized linear mixed models, *J. Am. Stat. Assoc.*, *88*(421), 9–25, 1993.

Brodeur, T. P., G. C. Cline, C. A. Cotton, and D. M. Owen, UV disinfection costs for inactivating *Cryptosporidium*, *Am. Water Works Assoc.*, *93*(6), 82–94, 2001.

Bukhari, Z., J. L. Clancy, Z. Matheson, R. M. McCuin, and C. R. Fricker, USEPA method 1622, *J. Am. Water Works Assoc.*, *91*(9), 60–98, 1999.

Carlin, B. P., and T. A. Louis, *Bayes and Empirical Bayes Methods for Data Analysis*, 2nd ed., Chapman and Hall, New York, 2000.

Casman, E. A., B. Fischhoff, C. Palmgren, M. J. Small, and F. Wu, An integrated risk model of a drinking-water-borne Cryptosporidiosis outbreak, *Risk Anal.*, *20*(4), 495–511, 2000.

Christensen, O. F., and R. Waagepetersen, Bayesian prediction of spatial count data using generalized linear mixed models, *Biometrics*, *58*(2), 280–286, 2002.

Clancy, J. L., Sydney's 1998 water quality crisis, *J. Am. Water Works Assoc.*, *92*(3), 55–66, 2000.

Connell, K., C. C. Rodgers, H. L. Shank-Givens, J. Scheller, M. L. Pope, and K. Miller, Building a better protozoan data set, *J. Am. Water Works Assoc.*, *92*(10), 30–43, 2000.

Crainiceanu, C. M., D. Ruppert, J. R. Stedinger, and C. T. Behr, Improving MCMC mixing for a GLMM describing pathogen concentrations in water supplies, in *Case Studies in Bayesian Statistics*, vol. VI, edited by C. Gatsonis et al., pp. 207–221, Springer-Verlag, New York, 2002a.

Crainiceanu, C. M., D. Ruppert, and J. R. Stedinger, Bayesian recovery rates modeling for waterborne pathogens (Applications to ICR spiking lab data), technical report, Cornell Univ. Ithaca, N. Y., 2002b. (Available at www.orie.cornell.edu/ davidr/epa-risk/)

Dugan, N. R., K. R. Fox, J. H. Owens, and R. J. Miltner, Controlling *Cryptosporidium* oocysts using conventional treatment, *J. Am. Water Works Assoc.*, *64*(12), 64–76, 2001.

Fayer, R., C. Speer, and J. Dubey, The general biology of *Cryptosporidium*, in *Cryptosporidium and Cryptosporidiosis*, edited by R. Fayer, pp. 1–41, CRC Press, Boca Raton, Fla., 1997.

Gatsonis, C., R. E. Kass, B. Carlin, A. Carriquiry, A. Gelman, I. Verdinelli, and M. West (Eds.), *Case Studies in Bayesian Statistics*, vol. V, Springer-Verlag, New York, 2002a.

Gatsonis, C., R. E. Kass, B. Carlin, A. Carriquiry, A. Gelman, I. Verdinelli, and M. West (Eds.), *Case Studies in Bayesian Statistics*, vol. VI, Springer-Verlag, New York, 2002b.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis*, Chapman and Hall, New York, 1995.

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

Gimbel, R., and A. Nahrstedt, A statistical method for determining the reliability of the analytical results in the detection of *Cryptosporidium* and *Giardia* in water, *Aqua*, *45*(3), 101–111, 1996.

Grenfell, B. T., and K. Wilson, Generalized linear modelling for parasitologists, *Parasitol. Today*, *13*, 33–37, 1997.

Haas, C. N., and J. B. Rose, Distribution of *Cryptosporidium* Oocysts in a water supply, *Water Resour.*, *30*(10), 2251–2254, 1996.

Hansen, J., and J. Ongerth, Effects of time and watershed characteristics on the concentration of *Cryptosporidium* oocysts in river water, *Appl. Environ. Microbiol.*, *57*(10), 2790–2795, 1991.

Hashimoto, A., T. Hirata, and S. Kunikane, Prevalence of *Cryptosporidium* oocysts and *Giardia* cysts in the drinking water supply in Japan, *Water Research*, *36*(3), 59–526, 2001.

Juranek, D. D., Cryptosporidiosis: Sources of infection and guidelines for prevention, *Clinical Infectious Diseases*, *21*, suppl. 1, 57–61, 1995.

Kuczera, G., Comprehensive at-site flood frequency analysis using Monte Carlo Bayesian inference, *Water Resour. Res.*, *35*(5), 1551–1557, 1999.

Laing, R. D., Lack of settling contributed to Canadian Crypto outbreak, condensation of report by Honorable Justice R. D. Laining, Commissioner, March 28, 2002, *Opflow*, *28*(6), 12–21, 2002.

LeChevallier, M. W., and W. D. Norton, Examining relationships between particle counts and *Giardia*, *Cryptosporidium*, and turbidity, *J. Am. Water Works Assoc.*, *87*(9), 54–60, 1992.

Lin, Z., and N. E. Breslow, Bias correction in generalized linear mixed models with multiple components of dispersion, *J. Am. Stat. Assoc.*, *91*(435), 1007–1016, 1996.

Littell, R. C., G. A. Milliken, W. W. Stroup, and R. D. Wolfinger, SAS system for mixed models, SAS Inst., Cary, N. C., 1996.

Maiti, T., Robust generalized linear mixed models for small area estimation, *J. Stat. Plann. Inf.*, *98*(1–2), 225–238, 2001.

McCullagh, P., and J. A. Nelder, *Generalized Linear Models*, Chapman and Hall, New York, 1989.

McCulloch, C. E., Maximum likelihood algorithms for generalized linear mixed models, *J. Am. Stat. Assoc.*, *92*(437), 162–190, 1997.

McCulloch, C. E., and S. R. Searle, *Generalized, Linear, and Mixed Models*, John Wiley, New York, 2001.

Meinhardt, P. L., D. P. Casemore, and K. B. Miller, Epidimiologic aspects of human Cryptosporidiosis and the role of waterborne transmission, *Epidimiol. Rev.*, *18*(2), 118–136, 1996.

Messner, M., Precision, accuracy, and detection capability of the ICR method for measuring *Cryptosporidium* and *Giardia* in water—Results of the ICR Laboratory Spiking Study, paper presented at the 1999 Joint Statistical Meetings, Section on Statistics and the Environment, Am. Stat. Assoc., Alexandria, Va., 2000.

Messner, M. J., and R. L. Wolpert, *Cryptosporidium* and *Giardia* occurrence in ICR drinking water sources—Statistical analysis of ICR Data, in *Information Collection Rule Data Analysis*, edited by M. J. McGuire, J. McLain, and A. Obolensky, pp. 463–481, AWWA Res. Found., Denver, Colo., 2002.

Messner, M. J., C. L. Chappell, and P. C. Okhuysen, Risk assessment for *Cryptosporidium*: A hierarchical Bayesian analysis of human dose, *Water Res.*, *35*(16), 3934–3940, 2001.

National Research Council, *Watershed Management for Potable Water Water Supply: Assessing New York City's Approach*, Natl. Acad. Press, Washington, D. C., 1999.

Okun, D. A., G. F. Craun, J. K. Edzwald, J. B. Gilbert, and J. B. Rose, New York City: To filter or not to filter, *J. Am. Water Works Assoc.*, *89*(3), 62–74, 1997.

Ong, C., W. Moorehead, A. Ross, and J. IsaacRenton, Studies of *Giardia* spp. and *Cryptosporidium* spp. in two adjacent watersheds, *Appl. Environ. Microbiol.*, *62*(8), 2798–2805, 1996.

Parkhurst, D., and D. Stern, Determining Average concentrations of *Cryptosporidium* and other pathogens in water, *Environ. Sci. Technol.*, *32*(21), 3424–3429, 1998.

Poulton, M., J. Colbourne, and J. Rose, *Cryptosporidium* monitoring in the UK and risk assessment, paper presented at 1992 Water Quality Technology Conference, Am, Water Works Assoc., Toronto, Canada, 1992.

Putnam, S. W., and J. B. Wiener, Seeking safe drinking water, in *Risk Versus Risk: Tradeoffs in Protecting Health and the Environment*, edited by J. D. Graham and J. B. Wiener, pp. 124–148, Harvard Univ. Press, Cambridge, Mass., 1995.

Regli, S., R. Odon, J. Cromwell, M. Lustic, and V. Blank, Benefits and costs of the IESWTR, *J. Am. Water Works Assoc.*, *91*(4), 148–158, 1999.

Rose, J. B., J. T. Lisle, and M. LeChavallier, Waterborne *Cryptosporidium*: Incidence, outbreaks and treatment strategies, in *Cryptosporidium and Cryptosporidiosis*, edited by R. Feyer, pp. 93–109, CRC Press, Boca Raton, Fla., 1997.

Scheller, J., K. Connell, H. Shank-Givens, and C. Rodgers, Design, implementation, and results of the EPA's 70-Utility ICR Laboratory Spiking Program, in *Information Collection Rule Data Analysis*, edited by M. J. McGuire, J. McLain, and A. Obolensky, pp. 483–500, AWWA Res. Found., Denver, Colo., 2002.

Science Advisory Board (SAB), Reducing risk: Setting priorities and strategies for environmental protection, U.S. Environ. Prot. Agency, Washington, D. C., 1990.

Solo-Gabriele, H., and S. Neumeister, U.S. outbreaks of Cryptosporidiosis, *J. Am. Water Works Assoc.*, *88*(9), 76–86, 1996.

Spiegelhalter, D., A. Thomas, and N. Best, WinBugs version 1.3 user manual, Medical Res. Counc. Biostat. Unit, Cambridge, UK, 2000.

Stedinger, J. R., and R. J. MacKay, Interpretation of *Cryptosporidium* and *Giardia* monitoring data generated by ICR program invited presentation, EPA Workshop on Statistics, U.S. Environ. Prot. Agency, Washington, D. C., 1998.

U.S. Environmental Protection Agency (U.S. EPA), National primary drinking water regulations: Long term enhanced surface water treatment

rule, Washington, D. C., 27 Nov. 2001. (Available at http://www.epa.gov/safewater/lt2/)

Walker, F. R., Jr., Statistical analysis of hydrologic and environmental data, Ph.D. thesis, Cornell Univ., Ithaca, N. Y., 1999.

Walker, F. R., Jr., and J. R. Stedinger, Fate and transport model of *Cryptosporidium*, *J. Environ. Eng.*, *125*(4), 325–333, 1999.

Wang, Q., A Bayesian joint probability approach for flood record augmentation, *Water Resour. Res.*, *37*(6), 1707–1712, 2001.

Wilke, C. K., Spatial modeling of count data: A case study in modelling breeding bird survey data on large spatial domains, in *Spatial Cluster Modelling*, edited by A. Lawson and D. Denison, pp. 199–209, Chapman and Hall, New York, 2002.

Young, P., and S. Komisar, The variability introduced by partial sample analysis to numbers of *Cryptosporidium* oocysts and *Giardia* cysts reported under the information collection rule, *Water Resources*, *33*(11), 2660–2668, 1999.

C. T. Behr, eDesign Dynamics LLC, 116 65th Street, # 2, West New York, NJ 07093, USA. (cbehr@edesigndynamics.com)

C. M. Crainiceanu, Department of Statistical Science, Cornell University, Malott Hall, NY 14853, USA. (cmc59@cornell.edu)

D. Ruppert, School of Operational Research and Industrial Engineering, Cornell University, Rhodes Hall, NY 14853, USA. (ruppert@orie.cornell.edu)

J. R. Stedinger, School of Civil and Environmental Engineering, Cornell University, Hollister Hall, NY 14853, USA. (jrs5@cornell.edu)