

# Nonlinear regression modeling of nutrient loads in streams: A Bayesian approach

Song S. Qian and Kenneth H. Reckhow

Nicholas School of the Environment and Earth Science, Duke University, Durham, North Carolina, USA

Jun Zhai

Institute for Genome Sciences and Policy, Duke University, Durham, North Carolina, USA

Gerard McMahon

U.S. Geological Survey, Raleigh, North Carolina, USA

Received 24 January 2005; revised 30 March 2005; accepted 28 April 2005; published 13 July 2005.

[1] A Bayesian nonlinear regression modeling method is introduced and compared with the least squares method for modeling nutrient loads in stream networks. The objective of the study is to better model spatial correlation in river basin hydrology and land use for improving the model as a forecasting tool. The Bayesian modeling approach is introduced in three steps, each with a more complicated model and data error structure. The approach is illustrated using a data set from three large river basins in eastern North Carolina. Results indicate that the Bayesian model better accounts for model and data uncertainties than does the conventional least squares approach. Applications of the Bayesian models for ambient water quality standards compliance and TMDL assessment are discussed.

**Citation:** Qian, S. S., K. H. Reckhow, J. Zhai, and G. McMahon (2005), Nonlinear regression modeling of nutrient loads in streams: A Bayesian approach, *Water Resour. Res.*, 41, W07012, doi:10.1029/2005WR003986.

## 1. Introduction

[2] A nonlinear regression modeling approach for modeling nutrient loading in streams using simple mechanistic equations to describe watershed nutrient generation and retention processes was proposed by *Smith et al.* [1997]. Because it is identifiable and fitted from observational data, a regression model provides an estimate of prediction error and can be used to assess the value from information of additional monitoring. Information on model prediction error also is necessary for identifying impaired water bodies and evaluating the effectiveness of a total maximum daily load (TMDL) program to eliminate the impairment.

[3] Nonlinear regression assumes a strict model error structure. (Model residuals are independent random variate from the same normal distribution.) Because nutrient or other pollutant flux within a watershed is potentially highly correlated and a simple nonlinear regression model cannot be expected to explain all nutrient or pollutant generation and attenuation processes, correlation among model residuals is to be expected. Our work builds on recent advances in computational statistics to explicitly model the spatial correlation structure in river basin hydrology and in watershed characteristics, such as soil and land use.

[4] Using nonlinear regression approach for watershed modeling has two potential weaknesses, which are similar to those suggested by *McMahon et al.* [2003], who applied the spatially referenced regressions on watershed (SPARROW) model to three river basins in eastern North Carolina. First,

a nonlinear regression model does not allow different coefficient values (e.g., the nutrient delivery rates) for different subwatersheds. Second, the least squares method for estimating model coefficients does not incorporate spatial autocorrelation in the data except for the implicit component in the stream network. As a result of these weaknesses and because a simple empirical model cannot include all sources and sinks of nutrients in the watershed, model residuals are correlated and often display a spatial pattern. When residuals are correlated the least squares method is not the most efficient model coefficient estimator; that is, estimated coefficients likely are subject to large estimation errors, and the resulting uncertainty analysis may be misleading.

[5] We present a Bayesian analysis for nonlinear regression models of stream nutrient loads. The main focus of the method is to explicitly model the correlations in the model residuals, including a state space modeling strategy to model nutrient transport between subwatersheds and a conditional autoregressive model to account for additional spatial correlation. The state space modeling component separates model and observation errors. Using a least squares method, the observed upstream loading likely will be treated as input from a fixed point source; hence this method fails to model the dynamic nature of nutrient transport from upstream to downstream subwatersheds.

[6] With a Bayesian approach, the estimated model coefficients are presented in terms of a posterior joint distribution. Although the Bayesian approach using noninformative priors often produces model coefficient estimates similar to those from the maximum likelihood method, the Bayesian emphasis on the posterior distribution will provide a full

description of the nature of uncertainty, leading to better quantification of model prediction error. Consequently, the Bayesian approach is better suited for predicting the frequency of water quality standard violations, which is important for TMDL development and assessment. In addition, the Bayesian approach allows objective model comparisons through the Bayes Factor and the deviance information criteria. We illustrate our methods using the SPARROW attributes [Smith *et al.*, 1997] and the data set developed for the three river basins in eastern North Carolina by *McMahon et al.* [2003] as an example.

## 2. Method

[7] The objective of our study is to better model the spatial autocorrelation. As pointed out by *McMahon et al.* [2003], the residuals of the initial application of the SPARROW model contain a systematic (spatial) pattern. Statistical theory indicates that the least squares based model coefficient estimator is inefficient when model residuals are not independent. This inefficiency indicates potentially large coefficient estimation errors and misleading model uncertainty assessment. The proposed Bayesian analysis will address the spatial autocorrelation in three steps, with Markov chain Monte Carlo (MCMC) simulations [Gilks *et al.*, 1996] employed as the computational tool in the following procedures.

[8] 1. The MCMC simulation method is used for parameter estimation to replace the least squares method used by most nonlinear regression models including the SPARROW model. With MCMC, uncertainty about model coefficients is summarized using the joint posterior distribution of the model coefficients, and avoiding the use of the bootstrap method currently used in SPARROW applications.

[9] 2. A state space (STSP) modeling approach is used to simulate the transportation of nutrient loads through the watershed. A STSP model can better address the serial correlation resulting from the stream network structure. Using a Bayesian approach resulted in a straightforward way to account for both the model and data uncertainty.

[10] 3. A conditional autoregressive (CAR) term is added to the STSP model to account for arbitrary spatial correlation.

### 2.1. SPARROW and Its Initial Application on North Carolina Rivers

[11] In a recent application, a SPARROW total nitrogen (TN) model was calibrated using data from three large river basins in eastern North Carolina: Cape Fear, Neuse, and Tar-Pamlico. The development and use of digital spatial and tabular data sets needed to support the model and results of the model calibration were presented and discussed by *McMahon et al.* [2003].

[12] The SPARROW model expresses mean annual nitrogen load at each monitoring station, located at the downstream end of stream reach  $i$ , as a nonlinear function of monitored and unmonitored nitrogen sources and the attenuation associated with landscape and aquatic processes:

$$\log(\text{Load}_i) = \log\left(\sum_{j \in J(i)} \sum_{n=1}^N \beta_n S_{n,j} e^{(-\alpha Z_j)} H_{i,j}^S H_{i,j}^R\right) + \epsilon_i, \quad (1)$$

where

- $\text{Load}_i$  the nitrogen load or flux in reach  $i$ , measured in metric tons for the year 1992;
- $n, N$  source index, where  $N$  is the total number of individual  $n$  sources;
- $J(i)$  the set of all reaches upstream and including reach  $i$ , except reaches at or above monitoring stations upstream from reach  $i$ ;
- $\beta_n$  the estimated source coefficient for source  $n$ ;
- $\alpha$  the estimated vector of land-to-water delivery coefficients;
- $S_{n,j}$  the nitrogen mass from source  $n$  in drainage to reach  $j$ ;
- $Z_j$  land surface characteristics data associated with drainage to reach  $j$ ;
- $H_{i,j}^S$  fraction of nutrient mass present in water body  $j$  transported to water body  $i$  as a function of first-order loss processes associated with stream channels ( $H_{i,j}^S = \prod_m \exp(-k_{sm} L_{i,j,m})$ , where  $k_{sm}$  is a first-order loss coefficient,  $m$  is the number of discrete flow classes, and  $L_{i,j,m}$  is the length of the stream channel between water bodies  $j$  and  $i$  in flow class  $m$ );
- $H_{i,j}^R$  the fraction of nutrient mass present in water body  $j$  transported to water body  $i$  as a function of first-order loss processes associated with lakes and reservoirs ( $H_{i,j}^R = \prod_l \exp(-k_l q_l^{-1})$ , where  $k_l$  is an estimated first-order loss rate or “settling velocity”,  $q_l^{-1}$  is the ratio of water surface area to outflow discharge, and  $l$  is the lakes and reservoirs located between water bodies  $j$  and  $i$ );
- $\epsilon_i$  the error term assumed to be independent and identically distributed across separate subbasins defined by intervening drainage areas between monitoring stations ( $\epsilon_i \sim N(0, \sigma^2)$ ).

[13] There are eight unknown parameters (seven model coefficients plus the unknown variance  $\sigma^2$ ; Table 1). The TN loading is generated from three sources (agricultural land, nonagricultural land, and point sources). Diffuse source nitrogen is attenuated as it moves across the landscape to the stream’s edge (modeled using one land delivery variable). Additional attenuation occurs as nitrogen moves through the reservoir and stream network to a downstream monitoring station using three aquatic loss rate coefficients. Detailed information about this model form, its assumptions, and applications is available elsewhere [Smith *et al.*, 1997; Preston and Brakebill, 1999; Alexander *et al.*, 2000, 2002].

[14] To illustrate the SPARROW model specification we present an example watershed (Figure 1) with nutrient loading monitored at three reaches (reaches 3, 6, and 9; hence the  $i$  in equation (1) takes values 3, 6, and 9). This example will be used again in section 2.3 to clarify the separation of model and observational errors. The nutrient load from reach 1 is

$$\left(\beta_1 S_{1,1} e^{(-\alpha Z_1)} + \beta_2 S_{2,1} e^{(-\alpha Z_1)} + \beta_3 S_{3,1}\right) H_{3,1}^S H_{3,1}^R,$$

where  $S_{1,1}$ ,  $S_{2,1}$ ,  $S_{3,1}$  are the agricultural land area, nonagricultural land area, and point source nutrient loads, respec-

**Table 1.** Definitions of Model Coefficients

Parameters	Definitions
$\alpha$	land delivery variable: soil hydrologic group
$\beta_1$	TN sources: agricultural land areal delivery rate (MT/km <sup>2</sup> )
$\beta_2$	TN sources: nonagricultural land areal delivery rate (MT/km <sup>2</sup> )
$\beta_3$	TN sources: point sources coefficient
$k_r$	aquatic loss: reservoir settling rate (m/yr)
$k_{s1}$	aquatic loss: small-stream rate coefficient (1/km)
$k_{s2}$	aquatic loss: large-stream rate coefficient (1/km)
$\sigma$	model error standard deviation

tively, in reach 1. The estimated nutrient loads for reaches 3, 6, and 9, respectively, are as follows:

$$\log(\text{Load}_3) = \log \left\{ \sum_{j=1}^3 \left( [\beta_1 S_{1,j} + \beta_2 S_{2,j}] e^{(-\alpha Z_j)} + \beta_3 S_{3,j} \right) H_{3,j}^S H_{3,j}^R \right\} + \epsilon_3,$$

$$\log(\text{Load}_6) = \log \left\{ \sum_{j=4}^6 \left( [\beta_1 S_{1,j} + \beta_2 S_{2,j}] e^{(-\alpha Z_j)} + \beta_3 S_{3,j} \right) H_{6,j}^S H_{6,j}^R \right\} + \epsilon_6,$$

and

$$\log(\text{Load}_9) = \log \left\{ \sum_{j=7}^9 \left( [\beta_1 S_{1,j} + \beta_2 S_{2,j}] e^{(-\alpha Z_j)} + \beta_3 S_{3,j} \right) H_{9,j}^S H_{9,j}^R \right\} + \epsilon_9.$$

For reaches 6 and 9, the point source terms  $S_{3,4}$  and  $S_{3,7}$  include input from upstream nutrient load, which are  $\text{Load}_3$  and  $\text{Load}_6$ , respectively, under the current SPARROW setting.

[15] In this paper, the upstream loads are separated from the within watershed point sources, and the regression coefficient  $\beta_3$  applies only to the point source loads. That is, upstream loads are not included in the point source term  $S_{3,j}$  and the term  $\beta_3 S_{3,j}$  is replaced by

$$\beta_3 S_{3,j} + \text{Load}_{\text{upstream of } j}. \quad (2)$$

As pointed out by an anonymous reviewer, treating upstream loads as point source input may result in a systematic bias because the reported point source loads are often biased upward. We separated the upstream loads from the point source to implement the models presented in sections 2.3 and 2.4.

[16] A nonlinear least squares algorithm to equation (1) was used for parameter estimation for the SPARROW model by *McMahon et al.* [2003], using a data set containing reach-specific values for all sources and landscape terms ( $S_{n,j}$  and  $Z_j$  in equation (1)), along with stream ( $L_{i,j,m}$ ) and reservoir ( $q_l$ ) data for each reach in the stream networks of three North Carolina river basins. The uncertainty in the model parameters was estimated using 200 bootstrap samples of the set of mean annual nutrient fluxes at 44 monitoring stations in three North Carolina river

basins. Each bootstrap sample is a random collection of the same 44 monitoring stations with replacement. In other words, some stations will be included more than once and some will not be included in the same bootstrap sample. A set of model coefficients was estimated for each of the 200 bootstrap samples. Ninety-percent confidence intervals were determined from the empirical coefficient distributions for the 200 bootstrap samples by computing the minimum range of coefficient values such that the fraction of values inside the range equaled the confidence levels [Smith et al., 1997].

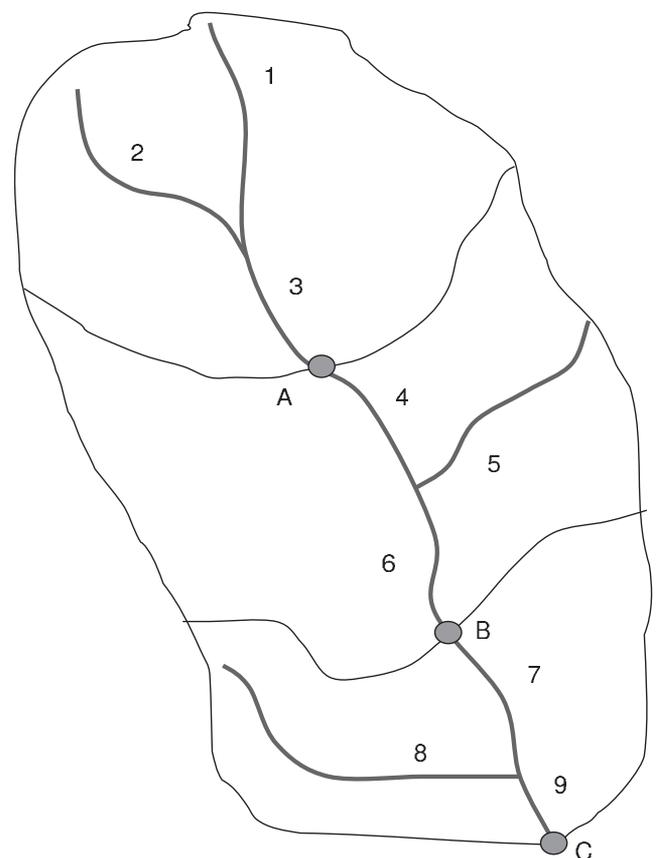
## 2.2. MCMC Model

[17] Reexpressing the nonlinear regression model (equation (1)) in terms of the probability distribution, the log nutrient loading follows a normal distribution with a mean defined by the model and a constant variance:

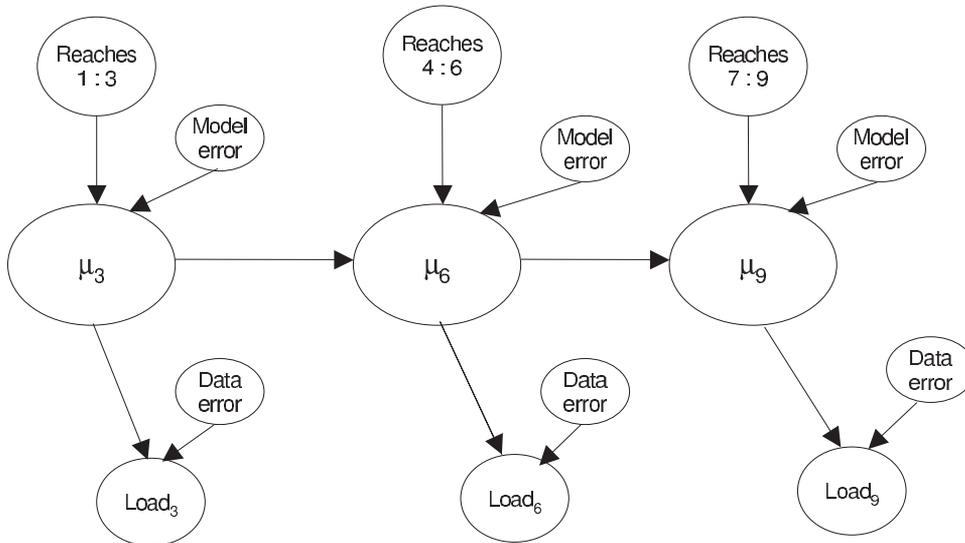
$$Y_i \sim N(\mu_i, \sigma^2) \quad (3)$$

$$\mu_i = \log \left( \sum_{n=1}^N \sum_{j \in J(i)} \beta_n S_{n,j} e^{(-\alpha Z_j)} H_{i,j}^S H_{i,j}^R \right),$$

where  $Y_i$  is the observed log nutrient load and  $\mu_i$  is the SPARROW model estimated expected log nutrient load for subwatershed  $i$ . The normal distribution provides the basis for the likelihood function. Using the Bayes' theorem, the likelihood function and prior distributions of all unknown coefficients are combined to yield the joint posterior



**Figure 1.** An example river basin shown having three subwatersheds each with three reaches, with nutrient loadings monitored at stations A, B, and C.



**Figure 2.** Graphical representation of the dynamic links between subwatersheds being modeled.

distribution. Using MCMC as the parameter estimation method, we obtain the posterior joint distribution of the model coefficients and the model error variance. Prior distributions for unknown model parameters are described in section 2.6.

### 2.3. STSP Model

[18] Statistically, the difficulty of modeling nutrient transport in streams using a regression model lies in the treatment of nutrient loads entering a subwatershed from upstream watersheds. In other words, the regression model can be summarized as

$$\log(\text{Load}_i) = f(\theta, \text{Load}_j),$$

where  $\text{Load}_j$  represents observed nutrient loads (with error) from upstream watershed(s). By using the observed  $\text{Load}_j$  as a fixed input, we use the upstream load as a predictor of the downstream load, which leads to the error in variable problem in a regression setting.

[19] Given the connection between upstream and downstream subwatersheds, the total error can be separated into process (or model) error and observation (or data) error. Process error occurs because a given parameterized model does not describe accurately the actual nutrient transport processes, whereas observation error represents all sources of error related to data gathering. The state space model framework provides a structure to include both observation and process error by separating the model-estimated nutrient loading and the observed nutrient loading. For example, model and observation errors for the example watershed in Figure 1 can be graphically expressed to illustrate the underlying statistical processes (Figure 2). In Figure 2 the actual mean nutrient loadings ( $\mu_3, \mu_6, \mu_9$ ) are estimated from a model plus a term representing the model error. The observed nutrient loadings ( $\text{Load}_3, \text{Load}_6, \text{Load}_9$ ) can be seen as samples representing the actual loadings. Mathematically, the state space model is presented in equation (4):

$$\begin{aligned} Y_i &\sim N(\mu_i, \sigma^2) \\ \mu_i &\sim N\left(f_i(\mu_{\text{upstream}}), \tau^2\right). \end{aligned} \quad (4)$$

Using this setup,  $\sigma^2$  represents the observation error variance,  $\tau^2$  represents process error variance, and  $f_i(\mu_{\text{upstream}})$  is the expected nutrient loading of subwatershed  $i$  estimated by using, for example, the SPARROW model with estimated upstream nutrient loading  $\mu_{\text{upstream}}$  as a point source input, i.e.,

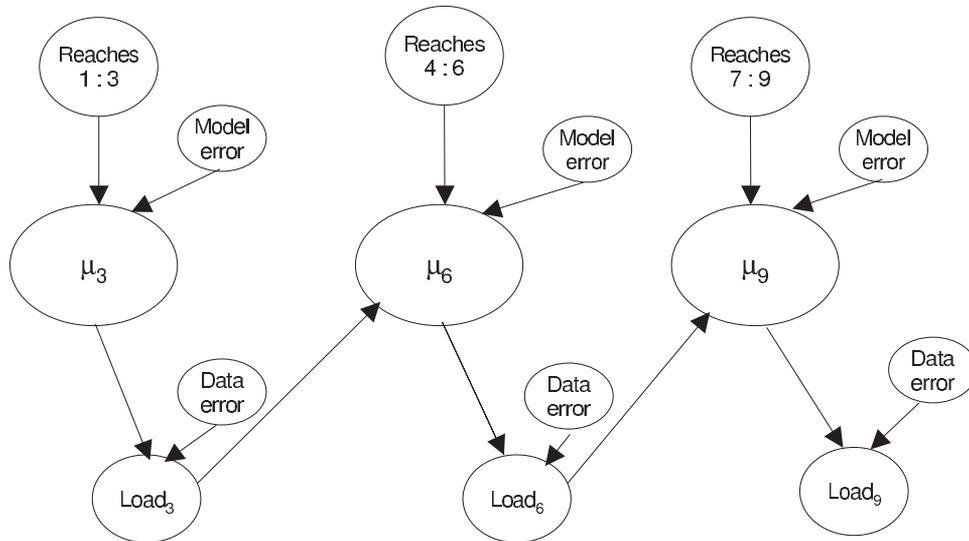
$$f_i(\mu_{\text{upstream}}) = \log\left(\sum_{n=1}^N \sum_{j \in J(i)} \beta_n S_{n,j} e^{-\alpha Z_i} H_{i,j}^S H_{i,j}^R\right), \quad (5)$$

and the model estimated upstream loading  $\mu_{\text{upstream}}$  is treated as the load from upstream (equation (2)). The STSP model adds one more unknown parameter ( $\tau^2$ ) to the list in Table 1.

[20] By entering the observed nutrient loading as input from upstream (as done in all previous SPARROW applications and the MCMC model in this paper), we equate the observed upstream log loading value  $Y_{\text{upstream}}$  and the underlying actual loading  $\mu_{\text{upstream}}$  (equation (4)) (Figure 3). Mathematically, using  $Y_{\text{upstream}}$  as upstream input is equivalent to setting the data error variance ( $\sigma^2$ ) to 0. Practically any data error will be propagated to downstream watersheds, resulting in serial correlation in the residuals.

### 2.4. CAR Model

[21] The STSP model describes spatial correlation through stream networks, which explains only part of the nutrient loading spatial correlation in the watershed. A mathematical model, mechanistic or empirical, is an approximation. It cannot include all relevant processes. The model we used includes only a subset of the parameters that are included in the national SPARROW models of *Smith et al.* [1997]. Many excluded parameters are spatially correlated. In addition, parameters included in the model may vary spatially. For example, a common nutrient delivery rate for agricultural land may not be accurate for the study area because of geographical differences in agriculture intensity. It is therefore reasonable to assume that the process error for the model will be spatially correlated.



**Figure 3.** Graphical representation of the dynamic links between subwatersheds being modeled under the current SPARROW structure.

[22] Because the nature of such spatial correlation is impossible to characterize, we opt for the Bayesian conditional autoregressive (CAR) modeling approach [Besag and Kooperberg, 1995]. Using CAR, a regional random effect term is introduced to each subwatershed. This random effect term is modeled as a Gaussian intrinsic autoregression, a treatment used in the conditional autoregressive model of Besag [1974]. Loosely speaking, the random effect term for all subwatersheds is an error term modeled by a multivariate normal distribution with mean 0 and an unknown covariance matrix:

$$\begin{aligned} Y_i &\sim N(\mu_i^{\text{CAR}}, \sigma^2), \\ \mu_i^{\text{CAR}} &= \mu_i + b_i, \\ \mu_i &\sim N\left(f_i(\mu_{\text{upstream}}^{\text{CAR}}), \tau^2\right), \end{aligned} \quad (6)$$

where  $f_i(\mu_{\text{upstream}}^{\text{CAR}})$  is the expected nutrient loading as defined in equation (5) and the estimated upstream loading  $\mu_{\text{upstream}}^{\text{CAR}}$  is used as the upstream loading (equation (2)). The distribution of the random effects  $b_i$  is proportional to

$$\exp\left[-\sum_{i \sim j} (b_i - b_j)^2 / (2s^2)\right],$$

where  $i \sim j$  denotes adjacent subwatersheds of  $i$ . This results in the conditional mean of  $b_i$  to be the mean of the adjacent subwatersheds random effects and the conditional variance of  $b_i$  to be  $s^2$  divided by the number of adjacent subwatersheds ( $n_i$ ). The conditional distribution of each term  $b_i$  is determined by the neighboring regions in the network:

$$\begin{aligned} b_i &\sim N(\bar{b}_i, s_i^2), \\ \bar{b}_i &= \frac{1}{n_i} \sum_{\text{neighbor}(i)} b_i, \\ s_i^2 &= s^2 / n_i. \end{aligned} \quad (7)$$

Intuitively, the model in equation (6) can be represented as a sum of three terms:

$$Y_i = \mu_i + b_i + e_i,$$

where  $e_i$  is the traditional sense error term distributed as normal with mean 0 and a constant variance ( $\sigma^2$ ), and the term  $b_i$  is a second error term modeled by a multivariate normal distribution. Equation (7) is the computational strategy for estimating the covariance matrix for  $b_i$ . The CAR model adds considerably more unknown parameters ( $\tau^2$ ,  $b_i$ , and  $s^2$ ) to the list in Table 1.

## 2.5. Model Comparison

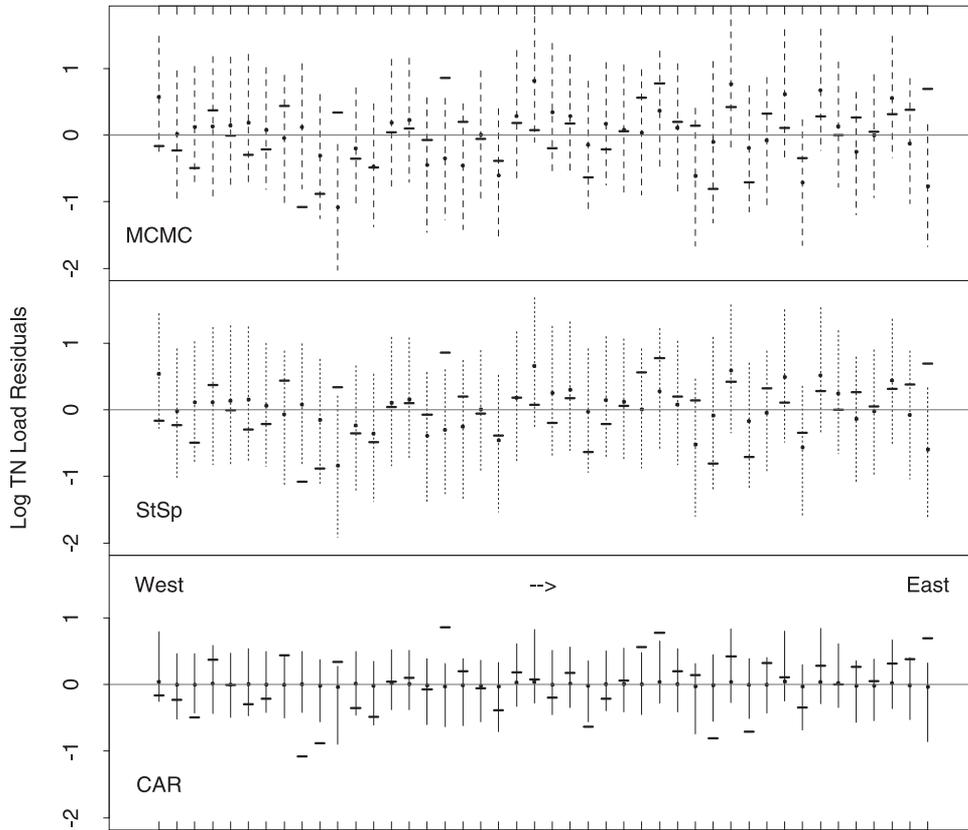
[23] The three Bayesian nonlinear models were compared using the Bayes factor (BF) [Kass and Raftery, 1995] and the Deviance Information Criterion (DIC) [Spiegelhalter et al., 2002]. BF was proposed initially by Jeffreys [1935, 1961] as a quantitative measure of the evidence in favor of a scientific theory. When comparing two alternative models, the BF is the posterior odds of one model over the other (assuming the prior probability on either model is 0.5). If the two alternative models are  $M_1$  and  $M_2$ , the BF is

$$B_{12} = \frac{\Pr(Y|M_1)}{\Pr(Y|M_2)}. \quad (8)$$

For model comparison purposes, the model likelihood ( $\Pr(Y|M_k)$ ,  $k = 1, 2$ ) is obtained by integrating over the parameter space:

$$\Pr(Y|M_k) = \int_{\Theta} \Pr(Y|\theta_k, M_k) \pi(\theta_k|M_k) d\theta_k, \quad (9)$$

where  $\theta_k$  is the parameter vector under model  $M_k$  and  $\pi(\theta_k|M_k)$  is the prior density of  $\theta_k$ . Using the MCMC method, we can estimate  $\Pr(Y|M_k)$  from posterior samples of



**Figure 4.** Ninety-five percent credible intervals (vertical lines) from the posterior predictive distributions of log TN loads estimated by the three models: (top) MCMC SPARROW, (middle) STSP SPARROW, and (bottom) CAR SPARROW. The predictive distributions are centered around the respective observed log TN loads and compared to the residuals from *McMahon et al.* [2003] (short horizontal lines).

$\theta_k$ . Letting  $\theta_k^{(i)}$  be samples from the posterior density  $\Pr(\theta_k|Y)$ , the estimated  $\Pr(Y|M_k)$  is:

$$\hat{\Pr}(Y|M_k) = \left\{ \frac{1}{m} \sum_{i=1}^m \Pr\left(Y|\theta_k^{(i)}, M_k\right)^{-1} \right\}^{-1}, \quad (10)$$

the harmonic mean of the likelihood values [Kass and Raftery, 1995].

[24] While BF compares the level of support of each model based on the fit to the data, DIC is a Bayesian measure of model complexity and fit. DIC is a sum of the posterior mean deviance  $\overline{D(\theta)}$ , a Bayesian measure of fit or “adequacy,” and a complexity measure  $p_D$ , which corresponds to the trace of the product of Fisher’s information and the posterior covariance [Spiegelhalter et al., 2002]. The Bayesian deviance  $\overline{D}$  is based on  $-2\log(\text{likelihood})$ , a measure of residual information in data conditional on parameter  $\theta$  or  $-2\log[p(Y|\theta)]$ . In our case, the data ( $Y = \log(\text{Load})$ ) are assumed from a normal distribution. The Bayesian deviance is

$$D(\theta) = -2 \log \frac{p(Y|\theta)}{f(Y)},$$

where  $f(Y)$  is the mathematical upper limit of the likelihood function (it is reached when the estimated  $\mu_i$  or  $\mu_i^{CAR}$  equal

to  $Y_i$ ). The smaller the  $D(\theta)$ , the closer the actual likelihood ( $p(Y|\theta)$ ) is to the maximum (hence a better model). The complexity measure is the the mean deviance minus the deviance evaluated at the posterior parameter means:

$$p_D = \overline{D(\theta)} - D(\hat{\theta}).$$

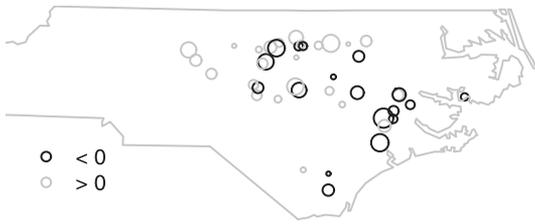
The DIC is defined as

$$\text{DIC} = \overline{D(\theta)} + p_D,$$

a Bayesian measure of model fit penalized by an additional complexity term. A smaller DIC indicates a “better” model.

### 2.6. Computation

[25] The MCMC simulation [Gilks et al., 1996] is used as the computation tool implemented with the software Bayesian inference Using Gibbs Sampler for Windows (WinBUGS) (D. Spiegelhalter et al., WinBUGS user manual, version 1.4, 2003, available at <http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml>). Qian et al. [2003] described MCMC in the context of environmental and ecological modeling. MCMC uses the fact that a joint multivariate distribution is fully defined by the full conditional distribution. Let  $\mathbf{V}$  be the collection of all unknown parameters and  $v$  be a member of  $\mathbf{V}$ . The full conditional distribution



**Figure 5.** Log residuals from least squares SPARROW model of *McMahon et al.* [2003] plotted over the three watersheds in eastern North Carolina. Sizes of the circles are proportional to the absolute value of the respective residuals.

is defined as the conditional distribution of  $v$  given  $\mathbf{V}_{-v}$  (all except  $v$ ):

$$p(v|\mathbf{V}_{-v}) \propto p(v, \mathbf{V}_{-v}) \propto \text{terms in } p(\mathbf{V}) \text{ containing } v.$$

Starting from a set of arbitrary initial values, the Gibbs sampler draws  $v$  from the full conditional distributions successively. These samples converge in distribution to the joint distribution of  $\mathbf{V}$ . An example of Gibbs sampler is given by *Qian and Richardson* [1997]. Inference about  $v$  is made based on samples from the joint distribution. These samples often are referred to as MCMC samples. In our case, the joint distribution of interest is the posterior joint distribution of unknown parameters. It is proportional to the product of the likelihood function defined by the normal distributions in equations (3), (4), or (6) and the densities of the prior distributions. Convergence is checked by using the Raftery and Lewis procedure [*Raftery and Lewis*, 1992, 1996].

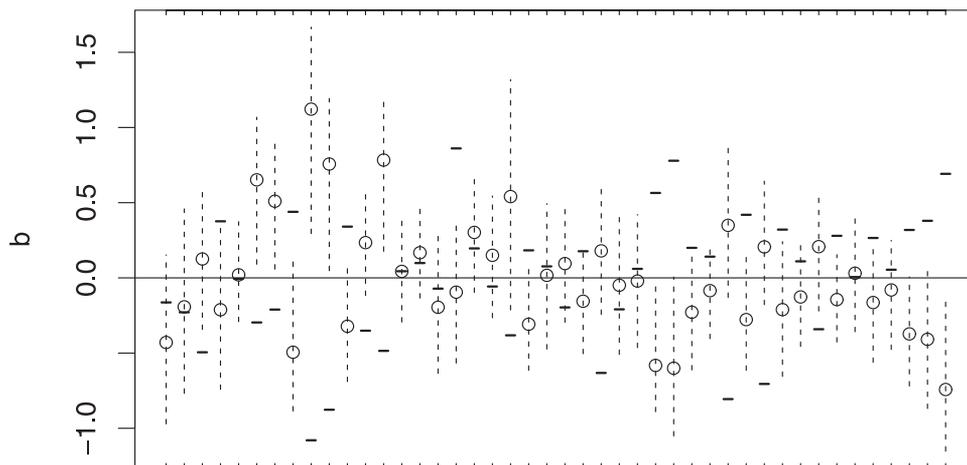
[26] We used a normal distribution with mean 0 and variance 10,000,  $N(0, 10,000)$ , as the prior distribution for model coefficients ( $\theta = \{\alpha, \beta_n, k_{sm}, k_r\}$ ) and a gamma distribution  $Ga(0.001, 0.001)$  as prior for precision parameters (e.g.,  $1/\sigma^2$ ). Both  $N(0, 10, 000)$  and  $Ga(0.001, 0.001)$  are considered “vague.” The distribution  $Ga(0.001, 0.001)$  (a proper distribution) is very close to the commonly used

Jeffrey’s prior for precision parameters, an improper distribution that does not integrate to one. Using this “barely” proper distribution avoids possible improper posteriors under the Jeffrey’s prior (see WinBUGS user manual for details). The distribution  $N(0, 10,000)$ , with a mean 0 and standard deviation of 100, is practically flat. Vague prior distributions were used because we had no information about these model parameters. When data from subsequent years are available, posterior distributions from this study can be used to develop specific priors.

### 3. Results

[27] Because of their Bayesian nature, our models produce posterior predictive distributions for TN loading at monitoring stations rather than point estimates. We present these posterior distributions using only their 95% (equal tail) credible intervals (Figure 4). The predictive distribution is presented in terms of the difference between the predicted log loads (a distribution) and the observed log loads. The CAR model consistently predicts narrower credible intervals than these from the MCMC and STSP models. If we use the distance between posterior predictive distribution medians and the observed values as a measure of model fit similar to the residuals of a regression model, the CAR model produced much smaller residuals than the nonlinear regression SPARROW of *McMahon et al.* [2003], as well as the other two Bayesian models. Although the STSP model resulted in 95% credible intervals similar to the MCMC model, the residuals from the STSP model are almost always smaller than residuals from the MCMC model. The differences in residuals between the two models are small when compared pairwise. However, the fact that the STSP model residuals are consistently smaller than the the MCMC model residuals indicates a systematic improvement and makes the addition of the state space links among subwatersheds worthwhile.

[28] *McMahon et al.* [2003] noted that the least squares fitted SPARROW tends to overpredict nutrient loads in the piedmont region and underpredict in the Coastal Plain (Figure 5). This spatial pattern is less apparent in the MCMC and STSP models, and no longer apparent in the



**Figure 6.** Ninety-five percent credible intervals of the spatial random effect term  $b_i$ , which are shown by the vertical dashed lines. The open circles are the means of  $b_i$ , and the short horizontal dashes are the log residuals from the least squares SPARROW model of *McMahon et al.* [2003].

**Table 2.** Comparisons of Estimated Model Coefficients<sup>a</sup>

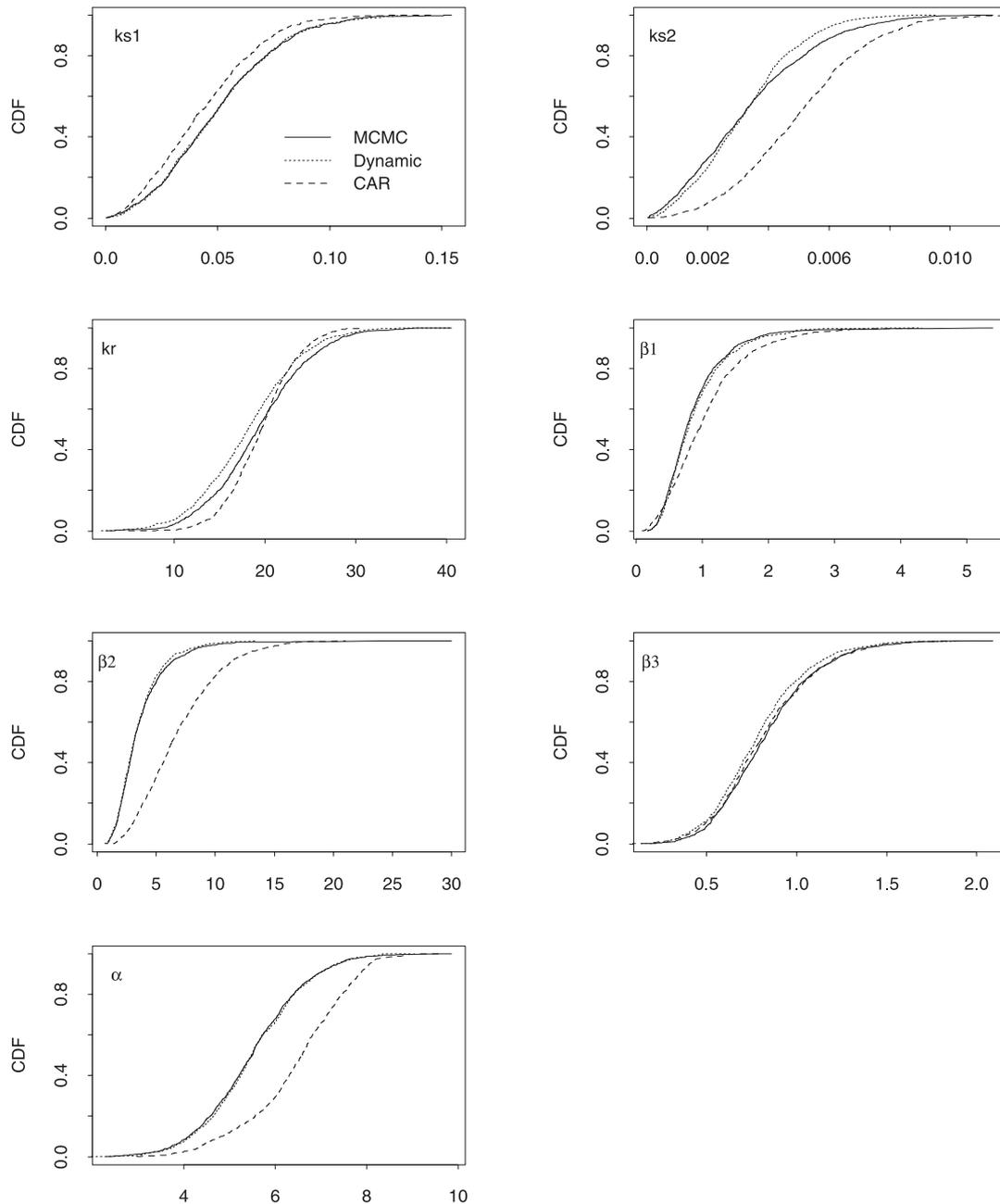
Model	$\alpha$	$\beta_1$	$\beta_2$	$\beta_3$	$k_r$	$k_{s1}$	$k_{s2}$
LS	4.1	5.9	1.8	0.85	16.4	0.08	0.002
MCMC	5.7(5.5)	2.8(3.7)	0.7(0.9)	0.77(0.83)	17.4(19.3)	0.04(0.05)	0.003(0.003)
STSP	5.3(5.5)	2.8(3.5)	0.7(0.9)	0.70(0.79)	18.4(18.2)	0.05(0.05)	0.003(0.003)
CAR	6.6(6.5)	5.5(6.9)	0.8(1.1)	0.73(0.82)	20.1(19.5)	0.03(0.04)	0.005(0.005)

<sup>a</sup>The estimated values for the three Bayesian models are the posterior mode and mean (in parenthesis).

CAR model. The random effect term  $b_i$  in the CAR model serves as a correction for the mismatch between the model predictions and the observations (Figure 6).

[29] The estimated model coefficients are different among the three Bayesian models and those estimated

by *McMahon et al.* [2003] (Table 2). The three Bayesian models coefficient estimates are the marginal posterior distribution medians, while the SPARROW estimates are coefficient values that minimize the residual sum of squares. The least squares results are close to the maximum likeli-



**Figure 7.** Empirical cumulative density functions of marginal posterior parameter distributions from three Bayesian models.

hood estimator (MLE) estimates in most situations. Theoretically, the MCMC model estimates would be close to the MLE estimate because flat priors lead to posteriors that are proportional to the likelihood function. The two sets of estimates are somewhat different.

[30] In addition to the spatial correlation in the model process error that may lead to a potentially erroneous result, a strong correlation among model coefficients can be another reason. The marginal posterior distribution of three coefficients ( $\alpha$ ,  $\beta_1$ ,  $\beta_2$ ) are highly correlated and concentrated along a narrow banana-shaped region. This type of joint distribution is common in water quality models (see *Qian et al.* [2003] for an example). The banana-shaped posterior distribution makes the search of the maximum likelihood (or the mode) very difficult using the conventional numerical optimization algorithms, because the likelihood or density value may vary only slightly along the center of this region. Consequently, a numerical search algorithm with a preset error tolerance level may stop at a different suboptimal values when different initial values are used. Using the MCMC approach, we are able to sample from the joint posterior distribution over the entire banana-shaped region, thus avoiding this difficulty in optimization.

[31] To compare the three Bayesian models, we first present the cumulative distribution function (CDF) of the marginal posteriors for each parameter (Figure 7). The CDFs for  $\sigma^2$  were not shown in Figure 7 because the three models have different error structures. The single model error variance  $\sigma^2$  in the MCMC model is divided into two parts in the STSP model ( $\sigma^2$ ,  $\tau^2$ ), and three parts in the CAR model ( $\sigma^2$ ,  $\tau^2$ , and  $s^2$ ). The estimated error standard deviations are listed in Table 3. We note that in the CAR model, each subregion has its own spatial random effect error variance  $s_i^2 = s^2/n_i$  (equation (7)). There are usually two to three neighboring regions for each region. The comparison indicates that the three Bayesian models resulted in very different posterior parameter distributions and different error structures. Among the three Bayesian models, the CAR model is overwhelmingly supported by the data based on BF and DIC (Table 4).

#### 4. Discussion

[32] Like others [e.g., *National Research Council*, 2001; *McMahon et al.*, 2003], we believe that empirical models, such as SPARROW, are a valuable modeling tool for assisting scientists and managers in TMDL pollutant allocation and listing impaired waters under section 303(d) of the Clean Water Act (303(d) listing). At the same time, we recognize that decisions regarding standard violations and TMDLs have significant consequences. Thus refinements in a conventional nonlinear regression model can help reduce costly errors associated with incorrect listings of violations of standards.

**Table 3.** Model Error Standard Deviations

Model	$\sigma$	$\tau$	$s$
MCMC	0.434	NA	NA
STSP	0.428	0.062	NA
CAR	0.063	0.061	0.647

**Table 4.** Bayes Factors and Deviance Information Criterion

	MCMC	STSP	CAR
MCMC	1	3.7	28,748 <sup>a</sup>
STSP	-	1	7,751
CAR	-	-	1
DIC	67	64	-44

<sup>a</sup>The BF values are comparing models on the top over models to the left.

[33] In our study, the Bayesian approach illustrated the improvements possible in a nonlinear regression model when spatial variability is accounted for in the model. The state space modeling approach characterizes predictive applications (i.e., load cannot be observed, because prediction represents a future, unrealized state) when upstream loading must be estimated using a regression model. The conditional autoregressive model further improves the nutrient model by capturing the spatial correlation among nearby river basins.

[34] As empirical watershed models, such as SPARROW, become more widely implemented, we envision yet another Bayesian application. At present, 303(d) listing of impaired waters is based on in situ monitoring. However, it is now recommended that hypothesis testing serve as the basis for the listing/delisting decision [*National Research Council*, 2001]; with monthly monitoring as the typical sampling frequency, a number of years of sampling may be necessary to achieve a high level of assurance on the listing decision. One way to reduce the number of samples (and hence the time required for sound listing/delisting decisions) is to use a model to augment the in situ monitoring.

[35] For example, if a commonly used lognormal distribution is assumed for the nutrient concentration, the limited monitoring concentration data can be augmented with Bayesian updating using the SPARROW model results to provide prior distributions of the nutrient concentration distribution parameters (log mean and log variance). Specifically, let the log concentration variable  $\log(C)$  follow a normal distribution with unknown mean  $\mu_c$  and variance  $\sigma_c^2$ :  $\log(C) \sim N(\mu_c, \sigma_c^2)$ . The most commonly used conjugate joint prior distribution for  $\mu_c$  and  $\sigma_c^2$  is the normal gamma distribution (with four parameters) and the posterior predictive distribution of  $\log(C)$  is a Student  $t$  distribution [*Bernardo and Smith*, 1994]. For example, using the CAR model for a specific subwatershed  $i$ , we have a series of MCMC samples of the mean log nutrient loading  $\mu_i^{\text{CAR}}$  and MCMC samples of variances  $\sigma^2$ ,  $\tau^2$  and  $s_i^2$ . The total variance of the log loading is  $\sigma_L^2 = \sigma^2 + \tau^2 + s_i^2$ . Considering a typical flow condition with log mean of  $q$  and log variance of  $\sigma_q^2$ , we can convert the MCMC samples of log loading to samples of mean nutrient concentrations ( $\mu_i^{\text{CAR}} - q$ , with variance  $\sigma_L^2 + \sigma_q^2$ ). These samples can be used to estimate the four parameters in the prior distribution (the normal gamma distribution). The resulting posterior predictive distribution (Student  $t$ ) can be used to estimate the probability of violation of standards. In addition, the resulting posterior predictive distribution can be used to assess whether there is a need for more monitoring samples based on the estimated posterior variance in the log concentration.

[36] In summary, we believe that a modeling approach combining simple mechanistic descriptions with identifi-

ability and parameter estimation represents an important direction for applied water quality modeling. As we have demonstrated, spatial correlation among watersheds can be modeled to improve a nonlinear model's predictions. Finally, when implemented in a Bayesian framework, results from a nonlinear regression model can be updated with monitoring data to assess the effectiveness of implemented TMDLs and to support the listing decision.

[37] **Acknowledgments.** Funding for this work was provided by the Water Resources Research Institute of the University of North Carolina and U.S. EPA (STAR grant R830883). We thank C.A. Stow, E. Conrad Lamon, Alix Gitelman, L.J. Steinberg, and R.J. Deckard for their constructive comments and suggestions on an earlier version of the manuscript. Discussions with Richard Smith, Greg Schwarz, and Rich Alexander helped us understand the details of SPARROW and clarify ideas. Comments and suggestions from Tim Cohn, Greg Schwarz, Graham McBride, and two anonymous reviewers are greatly appreciated.

## References

- Alexander, R. B., R. A. Smith, and G. E. Schwarz (2000), Effect of stream channel size on the delivery of nitrogen to the Gulf of Mexico, *Nature*, 403, 758–761.
- Alexander, R. B., A. H. Elliott, U. Shankar, and G. B. McBride (2002), Estimating the sources and transport of nutrients in the Waikato River Basin, New Zealand, *Water Resour. Res.*, 38(12), 1268, doi:10.1029/2001WR000878.
- Bernardo, J. M., and A. F. M. Smith (1994), *Bayesian Theory*, 586 pp., John Wiley, Hoboken, N. J.
- Besag, J. (1974), Spatial interaction and the statistical analysis of lattice systems, *J. R. Stat. Soc., Ser. B*, 36, 192–236.
- Besag, J., and C. L. Kooperberg (1995), On conditional and intrinsic autoregressions, *Biometrika*, 82, 733–746.
- Gilks, W. R., S. Richardson, and D. J. Spiegelhalter (Eds.) (1996), *Markov Chain Monte Carlo in Practice*, Chapman and Hall, New York.
- Jeffreys, H. (1935), Some tests of significance, treated by theory of probability, *Proc. Cambridge Philos. Soc.*, 31, 203–222.
- Jeffreys, H. (1961), *Theory of Probability*, 3rd ed., Oxford Univ. Press, New York.
- Kass, R. E., and A. E. Raftery (1995), Bayes factors, *J. Am. Stat. Assoc.*, 90, 773–795.
- McMahon, G., R. B. Alexander, and S. S. Qian (2003), Support of TMDL programs using spatially referenced regression models, *J. Water Resour. Plann. Manage.*, 129(4), 315–329.
- National Research Council (2001), *Assessing the TMDL Approach to Water Quality Management*, Natl. Acad. Press, Washington, D. C.
- Preston, S. D., and J. W. Brakebill (1999), Application of spatially referenced regression modeling for the evaluation of total nitrogen loading in the Chesapeake Bay watershed, *U.S. Geol. Surv. Water Resour. Invest. Rep.*, 99-4054, Va.
- Qian, S. S., and C. J. Richardson (1997), Estimating the long-term phosphorus accretion rate in the Everglades: A Bayesian approach with risk assessment, *Water Resour. Res.*, 33(7), 1681–1688.
- Qian, S. S., C. A. Stow, and M. Borsuk (2003), On Monte Carlo simulation for Bayesian analysis, *Ecol. Modell.*, 159, 269–277.
- Raftery, A. E., and S. M. Lewis (1992), One long run with diagnostics: Implementation strategies for Markov chain Monte Carlo, *Stat. Sci.*, 7, 493–497.
- Raftery, A. E., and S. M. Lewis (1996), Implementing MCMC, in *Markov Chain Monte Carlo in Practice*, edited by W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, pp. 115–130, CRC Press, Boca Raton, Fla.
- Smith, R. A., G. E. Schwarz, and R. B. Alexander (1997), Regional interpretation of water-quality monitoring data, *Water Resour. Res.*, 33, 2781–2798.
- Spiegelhalter, D., N. G. Best, B. P. Carlin, and A. van der Linde (2002), Bayesian measures of model complexity and fit, *J. R. Stat. Soc., Ser. B*, 64, 583–639.

G. McMahon, U.S. Geological Survey, 3916 Sunset Ridge Road, Raleigh, NC 27607, USA.

S. S. Qian and K. H. Reckhow, Nicholas School of the Environment and Earth Science, Duke University, Box 90328, Durham, NC 27708-0328, USA. (song@duke.edu)

J. Zhai, Institute for Genome Sciences and Policy, Duke University, Durham, NC 27708, USA.