Spring 2014

A comparison of Bayesian Monte Carlo methods for deterministic models

Amanda Marie Miller
James Madison University

Follow this and additional works at: https://commons.lib.jmu.edu/honors201019

Recommended Citation
https://commons.lib.jmu.edu/honors201019/450

This Thesis is brought to you for free and open access by the Honors College at JMU Scholarly Commons. It has been accepted for inclusion in Senior Honors Projects, 2010-current by an authorized administrator of JMU Scholarly Commons. For more information, please contact dc_admin@jmu.edu.
A Comparison of Bayesian Monte Carlo Methods for Deterministic Models

A Project Presented to
the Faculty of the Undergraduate
College of Science and Mathematics
James Madison University

in Partial Fulfillment of the Requirements
for the Degree of Bachelor of Science

by Amanda Marie Miller
May 2014

Accepted by the faculty of the Department of Mathematics and Statistics, James Madison University, in partial fulfillment of the requirements for the Degree of Bachelor of Science.

FACULTY COMMITTEE:

Project Advisor: Samantha Prins, Ph.D.,
Associate Professor, Mathematics & Statistics

Reader: Ling Xu, Ph.D.,
Associate Professor, Mathematics & Statistics

Reader: Lihua Chen, Ph.D.,
Assistant Professor, Mathematics & Statistics

HONORS PROGRAM APPROVAL:

Barry Falk, Ph.D.,
Director, Honors Program
Contents

1 Introduction 5

1.1 The Deterministic Model Framework . . . . . . . . . . . . . . . . . . . . . . . 8

2 Methods 10

2.1 Bates et al. (2003) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10

2.2 Sohn and Small (2000) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 15

3 Results 17

3.1 Bates et al. (2003) Posterior Distributions . . . . . . . . . . . . . . . . . . . 17

3.2 Comparison of Bates et al. (2003) and Sohn and Small (2000) Results . . . 17

3.3 Sensitivity Analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 25

4 Discussion 30
List of Figures

1  Dot plot of the natural log of PCB concentration for both air and soil. . . . . 13
2  Normal Quantile-Quantile plots of the observed values of natural log of PCB concentration in both air and soil. . . . . . . . . . . . . . . . . . . . . . . 14
3  Graphs of $b$, $\rho$, and $D$ . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 18
4  Graphs of $s_e$, $s_f$, $s_v$, $Vd$ . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 19
5  Graphs of $\mu_g$, $\mu_d$, $\sigma_g$, $\sigma_d$ . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 20
6  Graphs of $\mu_{ln(C_a)}$, $\sigma_{ln(C_a)}^2$, $C_a$, $C_s$ . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 21
7  Graphs of $ln(C_a)$ for Sohn and Small (2000) and Bates et al. (2003) Methods 23
8  Graphs of $ln(C_s)$ for Sohn and Small (2000) and Bates et al. (2003) Methods 24
9  Sohn and Small (2000) and Bates et al. (2003) Likelihoods . . . . . . . . . 26
Acknowledgements

I cannot express enough thanks to my committee for their continued support and encouragement. Special thanks to Dr. Samantha Prins for sticking with me for three long semesters of this research, being patient when I was struggling, and helping me whenever I needed her. Also sincere thank you to my readers Dr. Ling Xu and Dr. Lihua Chen for supporting me in this research and being such great help in the end when I was rushing to finalize everything.

Thank you to all of my friends and family for being here for me throughout this whole process; your love and support has carried me through these semesters of research. Extra special thanks to my amazing boyfriend for enduring all of the tears and stress that this project has brought on and sticking with me through it all no matter what, I don’t know what I would do without you.

I am very grateful for the opportunity to work with Dr. Samantha Prins for the past year and a half, it has been the best learning experience of my college career. I cannot thank you enough for everything I learned throughout this time and I will never forget it.
Abstract

We discuss a comparison of the Bayesian approaches to uncertainty assessment in deterministic models developed in Sohn and Small (2000) and Bates et al. (2003). The methods were compared within the context of the environmental risk assessment model discussed in Bates et al. (2003). Each approach was run with the same data and priors, their specific likelihood forms, and a sample from the posterior distributions obtained using the same algorithm, namely, sampling importance resampling. To determine similarities and differences between the two approaches we compared the general shape, location and spread of the posterior distributions as well as the analytic form of the likelihoods each used. The comparison showed that there was a difference in the likelihoods of each model and that this resulted in differences in some of the posterior distributions. Bates et al. (2003) used the mean and standard deviation of the observed data in their likelihood while Sohn and Small (2000) used each individual data point in their likelihood. For this reason, we believe that Sohn and Small (2000) seemed to represent the data better.

1 Introduction

Statistics is a science that is used to gather, summarize, analyze, and present data in a way that answers specific questions that one is interested in (Bolstad, 2007). In almost all cases the data we collect reflects uncertainty or variability in the underlying process. Methods for statistical inference try to account for some or all of this uncertainty and statistical inference problems are generally framed in terms of parameters, $\theta$. These parameters describe particular characteristics of the population being looked at. The Bayesian
approach considers the parameters to be random variables due to the uncertainty about their true value. Bolstad (2007) describes the Bayesian framework in more detail; we summarize it here.

In the Bayesian approach we use prior distributions to represent our uncertainty in our parameters. These priors are based on information that is known about the parameter values. A (joint) prior distribution for parameters \( \theta = (\theta_1, \ldots, \theta_k) \), which we will denote as, \( p(\theta) \), represents the relative weights that the individual performing the Bayesian analysis gives to the particular parameter values \( \theta = (\theta_1, \ldots, \theta_k) \). This prior knowledge (in the form of weights) could come from expert knowledge or previous studies and different individuals may select different prior distributions. If the \( k \) parameters are considered apriori to be independent from each other then an appropriate (joint) prior distribution would be the product of the \( k \) marginal prior distributions i.e.

\[
p(\theta) = \prod_{i=1}^{k} p_i(\theta_i)
\]  

where \( p_i(\theta_i) \) is the relative weight of the particular value \( \theta_i \) assigned by the prior for the \( i^{th} \) parameter.

In addition to apriori information on the parameters, we might also have data, \( y \). The likelihood distribution represents how likely this data is given the parameter values \( \theta \), and will be denoted by \( L(y|\theta) \).

Bayesians see the complete inference on the parameters, \( \theta \), as the posterior distribution of the parameter given the data that occurred. The posterior distribution gives the relative weights given to each parameter value after the information contained in the data
Bayes’ rule states that:

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_{j=1}^{n} P(A|B_j)P(B_j)}$$  (2)

where $B_i, i = 1, \ldots, n$ are unobservable events that partition the universe and $A$ is an event that has occurred. In Bayesian inference, the unobservable events are the values of $\theta$ and the data $y$ are considered the events that have occurred. Then, in terms of prior and likelihood distributions, the following can be used to find the posterior distribution, $g(\theta|y)$, for the parameters $\theta$ given the data $y$:

$$g(\theta|y) = \frac{L(y|\theta)p(\theta)}{\int L(y|\theta)p(\theta)d\theta}$$  (3)

But, because the denominator is an integral over all values of $\theta$, $g(\theta|y)$ is proportional to $p(\theta)L(y|\theta)$, the prior times the likelihood.

Appropriate priors can be selected through various means including: from the conjugate family of the posterior, by interpolation over discrete prior weights placed on several values of the parameter(s) within the apriori plausible range of those parameter(s), or as a uniform prior when no prior knowledge is available. The prior must assign reasonable probability to all values that are realistically possible. Although the choice of the prior distribution is left up to the investigator, if there is a reasonable amount of data, $y$, quite disparate prior distributions should still result in similar posterior distributions. In fact, investigators should investigate the sensitivity of the posterior distribution that results from an analysis, to the prior distribution that was specified.

Bayes’ theorem is the only consistent way to change our beliefs about the parameters given data that has actually occurred (Bolstad, 2007). Thus one aspect in which the
Bayesian approach is better than the frequentist approach is because it uses data that has already occurred and not just possible samples that may occur like the ones used in the frequentist approach.

1.1 The Deterministic Model Framework

Deterministic models have the form

\[ \phi = M(\theta) \]  

(4)

where \( M \) is some non-stochastic function/model of \( k \) inputs, \( \theta = (\theta_1, \ldots, \theta_k) \) that results in the \( l \) outputs, \( \phi = (\phi_1, \ldots, \phi_l) \). Note that the model, \( M \), is deterministic as opposed to stochastic; repeated "runs" of the model \( M \) with a given set of inputs \( \theta \), result in the exact same values of the outputs, \( \phi \). Uncertainty analysis can be used in order to determine the effect of uncertainty in the inputs, \( \theta \), on the uncertainty in the outputs, \( \phi \), of deterministic models. Two Bayesian approaches to uncertainty analysis for deterministic models were Sohn and Small (2000) and Bates et al. (2003). Both were based on Monte Carlo Studies, which are used to perform sample space averaging. The Monte Carlo studies take a large number of random samples and calculate the Bayesian statistics for each sample. The sampling distribution is approximated using the empirical distribution of these calculated statistics. In the end the statistics of the Monte Carlo sample are used to approximate the statistics of the sampling distribution (Bolstad, 2007).

Sohn and Small (2000) used Bayesian Monte Carlo methods in their research on reducing uncertainty in ground-water flow and chemical transport predictions. They com-
pared model results and existing observations using Bayesian updating. Instead of using a simple Monte Carlo study, Sohn and Small (2000) used a Bayes Monte Carlo approach in which updates of the prior distribution were done using Bayes’ rule to relate the model predictions to the observed data. As more information was obtained, Sohn and Small (2000) used the computed posterior from one round of updating as the prior in the next update round. These sequential rounds of updating were possible because their initial prior distribution gave non-zero weights to a sufficiently large range of values for the inputs, $\theta$.

Bates et al. (2003) used a Bayesian approach to carry out research on environmental risk assessment. In particular, they used a special case of Bayesian melding to work with multicompartment deterministic models. This approach allowed them to account for uncertainty in the inputs of the deterministic model while allowing information on one model (perhaps representing a single compartment or source) to inform the results of another model (perhaps representing the same compartment or source in addition to others).

Neither the Sohn and Small (2000) or Bates et al. (2003) approaches accounted for model uncertainty because both sets of authors assumed the underlying model, $M$, was valid.

This project will compare and contrast the methods described in Bates et al. (2003) and Sohn and Small (2000) within the context of a simple one-compartment deterministic model for the concentration of polychlorinated biphenyls (PCBs) in soil (abbreviated as $C_s$). This model is explained fully in Bates et al. (2003) and Taylor (1992) and expresses the single output $\phi = C_s$ as a deterministic function of several inputs.
2 Methods

2.1 Bates et al. (2003)

PCB concentration in soil was modeled using a deterministic function of five inputs in Bates et al. (2003). The inputs used were PCB concentration in the air \((C_a)\), deposition velocity \((Vd)\), decay constant in soil \((b)\), density of soil \((\rho)\), and mixing depth of soil \((D)\). The inputs can be used to calculate \(C_s\) deterministically using equation 5.

\[
C_s = \frac{C_a Vd}{b \rho D}
\] (5)

The prior distributions we used for these inputs (Table 1) were chosen based on the priors given in Bates et al. (2003). Deposition velocity can depend on the size of the particle, and to model this we introduce the vector, \(s = (s_c, s_f, s_v)\), representing the proportion of coarse and fine particles, and vapor phase. A dirichlet hyperprior is placed on \(s\) to represent our uncertainty in particle class. Deposition velocity is then represented as a mixture distribution based on \(s\) with different log-uniform prior distributions for \(Vd\) depending on the class of particle.

The prior for \(C_a\) is a random log-normal based on parameters \(\mu_{\ln(C_a)}\) and \(\sigma^2_{\ln(C_a)}\). To better represent the fact that \(C_a\) is an annual average of PCB concentration in air and how PCB concentration in the air varies over time, in particular during the growing season for produce versus the dormant season, Bates et al. (2003) introduced four inputs representing the mean and standard deviation of the 24 hour average air PCB concentration in the growing season \((\mu_g, \sigma_g)\) and the corresponding values in the dormant season \((\mu_d, \sigma_d)\). The inputs, \(\mu_g, \sigma_g, \mu_d, \sigma_d\), were then weighted according to the proportion of the year in
Table 1: Priors (and hyperpriors) for inputs to the deterministic model of PCB concentration in soil, Equation 5.

<table>
<thead>
<tr>
<th>Input</th>
<th>Prior</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C_a$</td>
<td>$\text{Log}<em>e\text{Normal}\left(\mu</em>{\ln(C_a)}, \sigma^2_{\ln(C_a)}\right)$</td>
</tr>
<tr>
<td>$b$</td>
<td>Uniform(0.0001, 0.0002)</td>
</tr>
<tr>
<td>$\rho$</td>
<td>$\text{Log}_e\text{Normal}(1.4 \times 10^6, 1.15)$ $^a$</td>
</tr>
<tr>
<td>$D$</td>
<td>Uniform(0.15, 0.25)</td>
</tr>
<tr>
<td>$V_d$ for:</td>
<td></td>
</tr>
<tr>
<td>Coarse particles</td>
<td>$\text{Log}_{10}\text{Uniform}(10, 1000)$ $^b$</td>
</tr>
<tr>
<td>Fine particles</td>
<td>$\text{Log}_{10}\text{Uniform}(1, 100)$</td>
</tr>
<tr>
<td>Vapor phase</td>
<td>$\text{Log}_{10}\text{Uniform}(10, 1000)$</td>
</tr>
<tr>
<td>$s=\left(s_c, s_f, s_v\right)$</td>
<td>$\text{Dirichlet}(20 \times (0.06, 0.01, 0.9)$</td>
</tr>
<tr>
<td>$\mu_g$</td>
<td>$\text{Normal}(-13.67, 1.54)$</td>
</tr>
<tr>
<td>$\sigma^2_g$</td>
<td>$\text{Inv-\chi}^2(1, 0.19)$ $^c$</td>
</tr>
<tr>
<td>$\mu_d$</td>
<td>$\text{Normal}(-15.49, 1.54)$</td>
</tr>
<tr>
<td>$\sigma^2_d$</td>
<td>$\text{Inv-\chi}^2(1, 1.11)$</td>
</tr>
</tbody>
</table>

$^a$If $X \sim \text{Log}_e\text{Normal}(a, b)$ then $\ln(X) \sim \text{Normal}(\ln(a), b^2)$.

$^b$If $X \sim \text{Log}_{10}\text{Uniform}(a, b)$ then $\log_{10}(X) \sim \text{Uniform}(\log_{10}(a), \log_{10}(b))$.

$^c$If $X \sim \text{Inv-\chi}^2(a,b^2)$ then $X$ is a scaled inverse $\chi^2$ with $a$ df and scale $b$.

Dartmouth, MA that each season lasts, to calculate the mean and variance of year-round 24 hour average air concentration $(\mu_a, \sigma^2_a)$ using:

$$
\mu_a = \frac{1}{365} (122\mu_g + 243\mu_d) \\
\sigma^2_a = \frac{1}{365^2} \left( (122\mu_g)^2 + (243\mu_d)^2 \right)
$$

(6)
The quantities $\mu_g$, $\mu_d$ and $\sigma_g$, $\sigma_d$ were given independent diffuse unit information priors estimated using the observed data. The forms of these priors represent the information that a single observation would provide. All of these priors are detailed in Table 1.

Weather is believed to occur in five-day cycles in the region where data was collected and there are 73 cycles per year. We can assume the expected annual average air concentration is equal to the expected 24 hour average air concentration, but the variability in annual average concentration is smaller due to the factor of 73 (Cullen and Frey, 1999). The 24 hour average values can be converted to annual average parameters $(\mu_{\ln(C_a)}, \sigma^2_{\ln(C_a)})$ using these assumptions and the properties of the lognormal distribution using equation (6).

$$\mu_{\ln(C_a)} = \mu_a + 0.5\sigma_a^2 - 0.5 \ln \left( 1 + \frac{1}{73} (\exp \{ \sigma_a^2 \} - 1) \right)$$  \hspace{1cm} (7)

$$\sigma^2_{\ln(C_a)} = \ln \left( 1 + \frac{1}{73} (\exp \{ \sigma_a^2 \} - 1) \right)$$  \hspace{1cm} (8)

Sampling for all of these priors and hyperpriors was done in such a way as to ensure finite, positive values of $\sigma^2_{\ln(C_a)}$ and $C_a$.

In order to obtain a sample from the posterior distribution for inputs, $\theta$, and output, $\phi = C_s$ we used sampling importance resampling (SIR). SIR takes an initial sample from some known distribution and then resamples from this initial sample according to calculated SIR weights. If the weights are calculated correctly, the resample is then a sample from your target distribution. In the Bayesian case, we initially sample values of $\theta$ from the prior distribution, $p(\theta)$, and then resample using an SIR weight that is proportional to the likelihood, $L(y|\theta)$. This is because the SIR weight is the ratio of what you want, the
Figure 1: Dot plot of the natural log of polychlorinated biphenyl (PCB) concentration for both air and soil.

posterior density, over what you initially sampled from, the prior density:

\[
\text{SIR weight} = \frac{g(\theta|y)}{p(\theta)} \propto \frac{L(y|\theta)p(\theta)}{p(\theta)} = L(y|\theta)
\]  

where \(\theta\) are the inputs, \(y\) is the observed data, \(p(\theta)\) is the prior distribution on the inputs, \(g(\theta|y)\) is the posterior of the inputs given the data, and \(L(y|\theta)\) is the likelihood for the inputs.

Data from a site in New Bedford Harbor, MA is available on which to base independent normal likelihood distributions for the natural logarithm of PCB concentrations in soil and in air. Specifically, we have 19 measurements of PCB concentration in air and 18 of PCB concentration in soil taken from samples obtained from Dartmouth, MA. Figure 1 summarizes the observed data and Figure 2 suggests the use of a normal likelihood is appropriate. In addition, information is also available Bates et al. (2003) for prior distributions for each of the remaining seven model inputs listed in Table 1.

In our case, following the approach of Bates et al. (2003), \(L(y|\theta)\) is the product (assur-
Figure 2: Normal Quantile-Quantile plots of the observed values of natural log of poly-
chlorinated biphenyl concentration in both air and soil.

The likelihood for air data is independent of that for soil data) of a likelihood based on
the independent observations in air and a likelihood based on the independent observa-
tions in soil. The likelihood piece for air is based on the data collected on contamination
in the air, given the values of air contamination in $\theta$. The likelihood piece for soil is based
on the data collected for contamination in soil, given the induced values of $C_s = M(\theta)$.

The likelihoods for air and soil are both log normal distributions with parameters equal
to the means and standard deviations for the natural log of observed soil concentrations
($T_s = 18$ data points) and observed air concentrations ($T_a = 19$ data points). We must
assume independence of the air and soil likelihoods because we have no information on
how to model the dependence between them. The breakdown of this product is,
\[ L(\theta_k | y) = \prod_{t=1}^{T_a} L(\theta_k | y_{a_t}) \prod_{t=1}^{T_s} L(\theta_k | y_{s_t}) \]  

\[ = \prod_{t=1}^{T_a} \frac{1}{\theta_k s_a \sqrt{2\pi}} \exp \left\{ -\frac{1}{2} \left[ \ln \theta_k - \bar{y}_a \right]^2 \right\} \times \prod_{t=1}^{T_s} \frac{1}{M(\theta_k) s_s \sqrt{2\pi}} \exp \left\{ -\frac{1}{2} \left[ \ln M(\theta_k) - \bar{y}_s \right]^2 \right\} \]  

(10)

where \( \theta_k \) is the \( k^{th} \) value of \( C_a \), \( M(\theta_k) \) is the \( k^{th} \) induced value of \( C_s \), \( y_{a_t} \) is the \( t^{th} \) observed air data, \( y_{s_t} \) is the \( t^{th} \) observed soil data, \( \bar{y}_a \) and \( s_a \) are the mean and standard deviation of the natural logarithm of observed air data, and \( \bar{y}_s \) and \( s_s \) are the mean and standard deviation of the natural logarithm of observed soil data. Note that \( C_a \) is the only input used in the first product. One thing to note is that the roles of \( y \) and \( \theta_k \) are switched to reflect how the likelihood uses the observed data.

The resulting SIR resample would thus be from the posterior distribution. Our SIR weights are then based on the likelihoods of the observed data, \( y \), given the input values sampled from the prior and the induced output values, \( M(\theta) \), calculated using equation (5), and after resampling 10% of the values using these weights, we have a sample from the posterior distribution.

2.2 Sohn and Small (2000)

In order to better understand the Bayesian Monte Carlo approach of Sohn and Small (2000) we analyzed the PCB contamination concentration in soil using his approach. The first step involved identifying the conceptual model, such as the model of \( C_s \) given in equation (5), and then the distributions of the model parameters (inputs and outputs) as given
in Table 1. Sohn and Small (2000) then compared the model results i.e. the induced values of $C_s$, with the observed data $y_i$, using Bayesian updating. In a general context, the likelihood is the product of the likelihoods for each individual observed data value with parameters of this distribution based on the inputs, $\theta$, and induced values, $M(\theta)$. That is,

$$ L(y|\theta_k) = \prod_{t=1}^{T} L(y_t|\theta_k) = \prod_{t=1}^{T} \frac{1}{\sqrt{2\pi} y_t \sigma_{e}} \exp\left\{-\frac{1}{2} \left[ \frac{\ln(y_t) - \ln(\theta_k)}{\sigma_{e}} \right]^2 \right\} $$

where $T$ is the number of independent observations and $\sigma_{e}^2$ is the observation error variance.

In the context of the PCB concentration problem from Bates et al. (2003), the Sohn and Small (2000) likelihood is the product of a likelihood based on independent observations in air and a likelihood of independent observations in soil. The likelihood is then:

$$ L(y|\theta_k) = \prod_{t=1}^{T_a} L(y_{at}|\theta_k) \prod_{t=1}^{T_s} L(y_{st}|\theta_k) $$

$$ = \prod_{t=1}^{T_a} \frac{1}{y_{at} s_a \sqrt{2\pi}} \exp\left\{-\frac{1}{2} \left[ \frac{\ln y_{at} - \ln \theta_k}{s_a} \right]^2 \right\} \times $$

$$ \prod_{t=1}^{T_s} \frac{1}{y_{st} s_s \sqrt{2\pi}} \exp\left\{-\frac{1}{2} \left[ \frac{\ln y_{st} - \ln M(\theta_k)}{s_s} \right]^2 \right\} $$

where $T_a$ is the number of independent observations of PCB concentration in air, $T_s$ is the number of independent observations of PCB concentration in soil, $\theta_k$ is the $k^{th}$ value of $C_a$, $M(\theta_k)$ is the $k^{th}$ induced value of $C_s$, and $y_{at}$, $y_{st}$, $s_a$, and $s_s$ are as defined previously. Not that $C_a$ is the only input used in the first product.

Sohn and Small (2000) uses the likelihood to compute SIR weights and these weights are used to determine the posterior distribution, just as in Bates et al. (2003).
3 Results

3.1 Bates et al. (2003) Posterior Distributions

The posterior distributions for each input and the outputs are approximated by the resample obtained through SIR. To view these posteriors we use density estimates. Figures 3-6 show the posterior distributions for inputs and outputs obtained using only the soil data in the likelihood, using only the air data in the likelihood, and then using both the air and soil data in the likelihood. For example, in the case of the air data only equation (11) would be:

$$L(\theta_k|y) = \prod_{t=1}^{T_o} L(\theta_k|y_{a_t})$$  \hspace{1cm} (15)

Each figure has a solid line representing the sample from the prior, a dashed line for the posterior based on soil alone, a dotted line for the posterior based on air alone, and a dashed/dotted line for the posterior based on both air and soil. We can use these plots to determine whether the likelihoods of the soil data, air data, or both combined give us information about any of the inputs or outputs. For instance the soil data appears to be driving the updates in deposition velocity and PCB concentration in soil, but the air data appears to be driving the updates in PCB concentration in air.

3.2 Comparison of Bates et al. (2003) and Sohn and Small (2000) Results

Since comparison of Bates et al. (2003) and Sohn and Small (2000) is our goal we provide the Sohn and Small (2000) posteriors only relative to the Bates et al. (2003) posteriors,
Figure 3: Graphs of $b$, $\rho$, and $D$. The solid black line is the sample from the prior. The red dashed line is the posterior based on soil alone. The blue dotted line is the posterior based on air alone. The green dashed and dotted line is the posterior based on both air and soil together.
Figure 4: Graphs of $s_c$, $s_f$, $s_v$, $Vd$. The solid black line is the sample from the prior. The red dashed line is the posterior based on soil alone. The blue dotted line is the posterior based on air alone. The green dashed and dotted line is the posterior based on both air and soil together.
Figure 5: Graphs of $\mu_g$, $\mu_d$, $\sigma_g$, $\sigma_d$. The solid black line is the sample from the prior. The red dashed line is the posterior based on soil alone. The blue dotted line is the posterior based on air alone. The green dashed and dotted line is the posterior based on both air and soil together.
Figure 6: Graphs of $\mu_{\ln(C_a)}$, $\sigma^2_{\ln(C_a)}$, $C_a$, $C_s$. The solid black line is the sample from the prior. The red dashed line is the posterior based on soil alone. The blue dotted line is the posterior based on air alone. The green dashed and dotted line is the posterior based on both air and soil together.
rather than summarizing them by themselves. At first glance the Sohn and Small (2000) method looks extremely similar to the Bates et al. (2003) method of Bayesian analysis. The likelihoods are the same in that both methods assume a log normal likelihood for air and soil, and both use the standard deviation of the observed PCB concentrations ($s_a$ and $s_s$) in the likelihoods. Before examining the form of the likelihoods any further, we show a comparison of the posteriors that the two methods obtained. The comparison graphs of priors and posteriors for $\ln(C_a)$ and $\ln(C_s)$ values are shown in Figures 7 and 8. These graphs are based on an initial sample of 50,000 and a resample of size of 5,000. They show the distinct differences in the Sohn and Small (2000) and Bates et al. (2003) method. We can see that in Figure 8 of $\ln(C_s)$ the Sohn and Small (2000) and Bates et al. (2003) posterior based on air alone (green dashed and dotted respectively) are similar. However, when soil data is incorporated, or used alone, the posteriors of the two approaches are quite different, favoring different values of $C_s$. In the graph of $\ln(C_a)$ we can see the Sohn and Small (2000) and Bates et al. (2003) posterior based on soil alone (red dashed and dotted respectively) is quite similar, however when air data is incorporated, or used alone, the posteriors are very different.

To explain why there was such a big difference in the two methods we looked into how each approach treated the data, $\theta_k$ and $M(\theta_k)$ values in the analytic form of the likelihoods (equations (11) and (14)), and what influence the data had. Comparing the Bates et al. (2003) likelihood in equation (11) and the Sohn and Small (2000) likelihood in equation (14) we can see that the two likelihoods are different. The differences come in the mean of the log normal distribution. The mean of the likelihood in Sohn and Small (2000) is based on the prior or induced values, $\theta_k$ and $M(\theta_k)$. The mean and standard deviation of
the likelihood in Bates et al. (2003) is based on the observed soil and air data, \(\bar{y}_a, \bar{y}_s, s_a,\) and \(s_s.\)

To examine this further we can look at Figure 9 which plots the form of the likelihoods using soil data alone (second product in equations (11) and (14)) over the range of induced \(\ln(C_s)\) values of -15 to -5 (which roughly matches our results in Figure 5).

Each log normal density is of the form:

\[
f(x, \mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2} \left[ \frac{\ln x - \mu}{\sigma} \right]^2 \right\}
\]

(16)

where \(\mu\) is the mean of \(\ln(x)\) and \(\sigma\) is the standard deviation of \(\ln(x)\). In Bates et al. (2003), \(\ln x = \ln(M(\theta_k))\) ie. the natural logarithm of the induced \(C_s\) value and \(\mu\) is the mean of the observed \(\ln(C_s)\) values. In Sohn and Small (2000), \(\ln x\) is the natural logarithm of one observed \(C_s\) value, and \(\mu\) is \(\ln(M(\theta_k))\), a single induced value. The Sohn and Small (2000) likelihood gives higher likelihood to induced values of \(C_s\) closer to the observed data than the Bates et al. (2003) likelihood does.

### 3.3 Sensitivity Analysis

The sensitivity to priors and random seed number were tested in order to make sure that the original seed number produced typical results and that the posterior distributions were not underly sensitive to the chosen priors. All origirinal results were done using an original seed of 1281992, and to quantify the Monte Carlo error in the results three extra seeds were used. The prior and posterior distributions (using all available data and both the Bates et al. (2003) and Sohn and Small (2000) methods) were found using each of these seeds. Table 2 gives the 2.5\(^{th}\), 50\(^{th}\), and 97.5\(^{th}\) percentiles of the prior and
Figure 9: Sohn and Small (2000) and Bates et al. (2003) Likelihoods. Each likelihood is calculated for a sequence of values between -15 and -5, representing potential induced values of $\ln(C_s)$. The black dashed line represents the likelihood using the Sohn and Small (2000) approach (equation (14)). The solid black line represents likelihood using the Bates et al. (2003) approach (equation (11)). The points represent the observed $C_s$ data.
posterior distributions for both $C_s$ and $C_a$ using the original seed, and the largest change (as percentages) observed in each of the percentiles using the new seeds.

Both the prior for $C_a$ and induced distribution for $C_s$ show large percentage differences in the $2.5^{th}$ and $97.5^{th}$ percentiles. This is expected because the prior distributions for $C_a$ and induced for $C_s$ are diffuse and the values sampled in the tails of these priors can vary widely due to different seed number. The thing to note is that the median values for both $C_a$ and $C_s$ were less than five percent different.

All differences in the posterior distributions for the percentiles were less than 10 percent, which shows evidence that the Monte Carlo error in the results was not too high.

The sensitivity to increasing and decreasing the width of the priors by 10 percent was also tested using the original seed of 1281992 for all simulations. When the priors were decreased (Table 3) you can obviously see that the percent differences from original priors was very large in the $2.5^{th}$ percentile but very small everywhere else. In the posteriors there was almost no differences made due to decreasing the width of the priors, all of the differences were less than one percent.

When the width of the priors were increased similar results occurred (Table 4). There as an obvious difference in the $97.5^{th}$ percentile for both prior distributions, but this was expected because we were increasing the priors. The $50^{th}$ and $2.5^{th}$ percentile differences were very small. The posteriors also had very minimal differences due to increasing the priors, all of them were less than a one percent difference from when the original priors were used.

When testing the sensitivity to increasing and decreasing the width of the priors we would like to see all of the percentages in Tables 3 and 4 be within the corresponding
Table 2: Percentage differences from original seed due to seed change. The percentage differences were tested for the 2.5\textsuperscript{th}, 50\textsuperscript{th}, and 97.5\textsuperscript{th} percentiles. Each row represents the actual percentile value observed with the original seed of 1281992, and in parentheses the largest absolute change in that percentile resulting from using 3 other seeds. The change is expressed as a percentage of the original value. Results are included for the prior and induced distributions (Row 1), posterior based on Bates et al. (2003) method (Row 2), and posterior based on Sohn and Small (2000) method (Row 3).

**PCB Concentration in Soil ($C_s$)**

<table>
<thead>
<tr>
<th>Distribution</th>
<th>2.5\textsuperscript{th} Percentile</th>
<th>50\textsuperscript{th} Percentile</th>
<th>97.5\textsuperscript{th} Percentile</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prior</td>
<td>$9.23 \times 10^{-91}$ (68577.93%)</td>
<td>$1.68 \times 10^{-6}$ (2.06%)</td>
<td>$9.93 \times 10^{+66}$ (2266.52%)</td>
</tr>
<tr>
<td>Bates</td>
<td>$5.62 \times 10^{-7}$ (2.91%)</td>
<td>$4.46 \times 10^{-6}$ (2.18%)</td>
<td>$3.12 \times 10^{-5}$ (7.64%)</td>
</tr>
<tr>
<td>Sohn</td>
<td>$2.35 \times 10^{-5}$ (4.28%)</td>
<td>$4.28 \times 10^{-5}$ (4.62%)</td>
<td>$7.25 \times 10^{-5}$ (2.97%)</td>
</tr>
</tbody>
</table>

**PCB Concentration in Air ($C_a$)**

<table>
<thead>
<tr>
<th>Distribution</th>
<th>2.5\textsuperscript{th} Percentile</th>
<th>50\textsuperscript{th} Percentile</th>
<th>97.5\textsuperscript{th} Percentile</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prior</td>
<td>$1.99 \times 10^{-91}$ (137330.99%)</td>
<td>$6.42 \times 10^{-7}$ (1.34%)</td>
<td>$7.51 \times 10^{+66}$ (872.54%)</td>
</tr>
<tr>
<td>Bates</td>
<td>$6.60 \times 10^{-8}$ (7.88%)</td>
<td>$3.24 \times 10^{-7}$ (0.54%)</td>
<td>$1.62 \times 10^{-6}$ (0.85%)</td>
</tr>
<tr>
<td>Sohn</td>
<td>$2.69 \times 10^{-7}$ (4.90%)</td>
<td>$4.71 \times 10^{-7}$ (6.40%)</td>
<td>$7.74 \times 10^{-7}$ (3.14%)</td>
</tr>
</tbody>
</table>
Table 3: The sensitivity to priors was tested by comparing the actual values of each percentile with the original seed to the percentiles after the priors were decreased by 10 percent. The rows labeled Bates and Sohn represent the posteriors using the Bates et al. (2003) and Sohn and Small (2000) methods.

### PCB Concentration in Soil \( (C_s) \)

<table>
<thead>
<tr>
<th>Distribution</th>
<th>2.5(^{th}) Percentile</th>
<th>50(^{th}) Percentile</th>
<th>97.5(^{th}) Percentile</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prior</td>
<td>855674.37%</td>
<td>0.22%</td>
<td>1.00%</td>
</tr>
<tr>
<td>Bates</td>
<td>0.19%</td>
<td>0.22%</td>
<td>0.33%</td>
</tr>
<tr>
<td>Sohn</td>
<td>0.03%</td>
<td>0.10%</td>
<td>0.04%</td>
</tr>
</tbody>
</table>

### PCB Concentration in Air \( (C_a) \)

<table>
<thead>
<tr>
<th>Distribution</th>
<th>2.5(^{th}) Percentile</th>
<th>50(^{th}) Percentile</th>
<th>97.5(^{th}) Percentile</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prior</td>
<td>1636343.57%</td>
<td>0.04%</td>
<td>1.00%</td>
</tr>
<tr>
<td>Bates</td>
<td>0.19%</td>
<td>0.14%</td>
<td>0.09%</td>
</tr>
<tr>
<td>Sohn</td>
<td>0.03%</td>
<td>0.10%</td>
<td>0.03%</td>
</tr>
</tbody>
</table>

percentages in Table 2. This would allow the sensitivity to prior changes to be within Monte Carlo error. All of the percentages in Table 3 and 4 did fall within the Monte Carlo error in Table 2 except for the priors when the width of the priors were decreased. They were slightly higher than the Monte Carlo error.
Table 4: The sensitivity to priors was also tested by comparing the actual values of each percentile with the original seed to the percentiles after the priors were increased by 10 percent. The rows labeled Bates and Sohn represent the posteriors using the Bates et al. (2003) and Sohn and Small (2000) methods.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>2.5&lt;sup&gt;th&lt;/sup&gt; Percentile</th>
<th>50&lt;sup&gt;th&lt;/sup&gt; Percentile</th>
<th>97.5&lt;sup&gt;th&lt;/sup&gt; Percentile</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prior</td>
<td>0.63%</td>
<td>0.25%</td>
<td>266.68%</td>
</tr>
<tr>
<td>Bates</td>
<td>0.00%</td>
<td>0.23%</td>
<td>0.52%</td>
</tr>
<tr>
<td>Sohn</td>
<td>0.07%</td>
<td>0.01%</td>
<td>0.03%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Distribution</th>
<th>2.5&lt;sup&gt;th&lt;/sup&gt; Percentile</th>
<th>50&lt;sup&gt;th&lt;/sup&gt; Percentile</th>
<th>97.5&lt;sup&gt;th&lt;/sup&gt; Percentile</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prior</td>
<td>0.62%</td>
<td>0.04%</td>
<td>123.40%</td>
</tr>
<tr>
<td>Bates</td>
<td>0.27%</td>
<td>0.17%</td>
<td>0.05%</td>
</tr>
<tr>
<td>Sohn</td>
<td>0.05%</td>
<td>0.07%</td>
<td>0.04%</td>
</tr>
</tbody>
</table>

4 Discussion

Our comparison of the Sohn and Small (2000) and the Bates et al. (2003) methods in the context of a deterministic model for PCB concentration in soil found differences in the posterior distributions for some inputs and outputs to the model in terms of location and scale. These differences appear to be due to the differing treatment of the observed air and soil data in the likelihoods. The Sohn and Small (2000) approach based its likelihood
on each individual observed data value, while the Bates et al. (2003) approach based its likelihood on the mean of the observed data. This led to the Sohn and Small (2000) approach having their likelihood being centered around the observed data, while the Bates et al. (2003) approach had their likelihood centered at the low end of the observed data. Thinking back to what a likelihood actually is we know the likelihood is the probability of the data given the prior and induced output values. Thus the Sohn and Small (2000) likelihood chooses the input and induced values that give high probability to individual values of the observed data while the Bates et al. (2003) likelihood chooses the prior and induced values that give high probability for only the average of the observed data. Therefore the likelihood from Sohn and Small (2000) makes the best use of the observed data.

Other than the likelihood differences, we did not see any differences in the methods. Both can be used in a sequential manner as more data becomes available.

References


