Fast Bayesian Inference for Computer Simulation Inverse Problems

Matthew Taddy[1], **Herbert K. H. Lee**[2] & **Bruno Sansó**[2]

1 University of Chicago Graduate School of Business
2 Department of Applied Mathematics and Statistics
University of California, Santa Cruz

E-mail: matt.taddy@chicagogsb.edu

Abstract. Computer models for the simulation of physical and environmental phenomena are often regulated by complicated dependencies on unknown parameters. When there already exists a large bank of simulated values, it may be difficult or impossible to fit a complicated statistical model to the existing parameter evaluations or to develop a Markov chain Monte Carlo (MCMC) sampling scheme, as the common Bayesian statistical approaches would require. In response to this motivation for a fast Bayesian statistical analysis which does not require model fitting or MCMC sampling, we discuss a sampling importance resampling algorithm that works in conjunction with kernel density estimation to resample from the original computer output according to the posterior distribution of input values. We present two applications where input parameter values are to be inferred from scarce observations and abundant simulated output. One consists of a climate simulator and the other of a groundwater flow model.

1. Introduction

Problems in engineering and environmental studies often require the use of mathematical forward simulation models which depend upon unknown parameters. The setting of such parameters, or calibration of the simulator, is based upon a comparison of model output against actual observations obtained through experimentation or from historical record. Thus, inference about the optimal values for unknown input parameters constitutes a statistical inverse problem. There is a rapidly growing literature on this problem, but little discussion of how to perform fully Bayesian inference when large numbers of computer runs are available, or when Markov chain Monte Carlo is not practical. This paper addresses those cases.

The field of design and analysis of computer experiments has received considerable attention during the last two decades (see, e.g., Sacks et al., 1989; Kennedy and O'Hagan, 2001; Santner et al., 2003). The classic setting of a computer experiment involves some true process, $\zeta(x, \theta)$, which in the real world characterizes a functional relationship between sets of inputs $\{x, \theta\}$, with x known and θ unknown, and an output y. Each of x, y, θ may be multivariate and it is not uncommon that θ will depend upon x. With additive noise, $y(x, \theta) = \zeta(x, \theta) + \varepsilon(x)$, where ε is a zero mean random variable with a distribution which may depend upon x. The computer simulation function, $\eta(x, \theta)$, is then incorporated into the framework such that $y(x, \theta) = \eta(x, \theta) + e(x, \theta)$, where the simulator error, $e(x, \theta) = \zeta(x, t) - \eta(x, \theta) + \varepsilon(x)$, includes random noise as well as the bias between simulation and reality.

The inverse problem involves solving for the values of underlying variables that have led to an observed data set. Given a set of true response values $\mathbf{y} = \{y_1, \ldots, y_n\}$, with known inputs $\mathbf{x} = \{x_1, \ldots, x_n\}$, what is our uncertainty about the unknown input $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_n\}$ values that led to this response? An overview of such problems can be found in the book by Kaipio and Somersalo (2004).

Characteristically, it is not possible to identify which portion of the discrepancy between data and simulation is due to simulator bias. This is the case when, for example, the real world response values at different known input locations are all assumed to

correspond to a single unknown input θ vector (i.e., $\theta_i = \theta \; \forall \; i$). This is also true when, although there is a distinct θ_i corresponding to each (x_i, y_i) pair, the prior on $\boldsymbol{\theta}$ holds that the entire parameter set is a single realization of a stochastic process indexed by $\mathbf{x}$ (e.g, when $\boldsymbol{\theta}$ forms a spatial field and $\mathbf{x}$ is a spatial grid of locations). In these situations, we cannot hope to estimate the bias and must assume that $\eta(x, \cdot) = \zeta(x, \cdot)$. The model is then

$$y(x, \theta) = \eta(x, \theta) + \varepsilon(x), \tag{1}$$

and the likelihood for $\boldsymbol{\theta}$ is based upon only the residuals $\mathbf{y} - \eta(\mathbf{x}, \boldsymbol{\theta}) = [y_1 - \eta(x_1, \theta_1), \ldots, y_n - \eta(x_n, \theta_n)]'$. For example, if the error is modeled such that $\varepsilon(x_i) \overset{iid}{\sim} \mathrm{N}(0, \sigma^2)$ for $i = 1, \ldots, n$, then $\Pr(\mathbf{y}|\mathbf{x}, \eta, \boldsymbol{\theta}) \propto \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - \eta(x_i, \theta_i)\right)^2\right)$. Note that this simplification, although necessary, should be used with caution: if we proceed in this manner in analysis of a simulator which does not accurately characterize the relationship between inputs and outputs, we will be merely *tuning* the computer simulator to match the observed output rather than actually solving for any physical interpretation of $\boldsymbol{\theta}$.

We assume a Bayesian approach in undertaking to solve for $\boldsymbol{\theta}$. Optimal decisions about unknown parameter values may then be made through minimization of a loss function depending upon the posterior distribution for $\boldsymbol{\theta}$, which is of the form,

$$\Pr(\boldsymbol{\theta}|\mathbf{x}, \mathbf{y}, \eta) \propto \Pr(\mathbf{y}|\mathbf{x}, \eta, \boldsymbol{\theta}) \pi(\boldsymbol{\theta}|\mathbf{x}). \tag{2}$$

The prior $\pi(\boldsymbol{\theta}|\mathbf{x})$ can assume a variety of different forms, and the two applications considered in this paper describe independent priors over each dimension of a single θ vector, as well as a Gaussian Process prior for $\boldsymbol{\theta}$ over a two-dimensional spatial grid defined by $\mathbf{x}$.

It is seldom possible to obtain an analytical solution for the density represented in (2). Common strategies for inference center on Markov chain Monte Carlo (MCMC) (see, for example, Gamerman and Lopes, 2006) sampling from the posterior for unknown parameters. This approach requires repeated evaluation of a likelihood, $\Pr(\mathbf{y}|\mathbf{x}, \eta, \boldsymbol{\theta})$, which depends upon simulator output at unique parameter input locations. If η is easy to evaluate, this inference is straightforward. However, it is often the case that, given a set of parameter values, solving the forward problem and obtaining the corresponding model output is very time consuming. Embedding this computation within an iterative procedure such as MCMC will be prohibitively expensive. As a result, most applications allow for a parametric statistical model $\hat{\eta}(x, \theta|\psi)$, where ψ are the model parameters, fit to a limited number of simulator runs to act as a cheap surrogate for the actual computer model. A fully Bayesian analysis requires uncertainty about the statistical surrogate model (i.e., about ψ) to be incorporated into the analysis. Thus the canonical framework for a Bayesian solution to statistical inverse problems is to alternatively draw from the conditional probability distributions for ψ given set of computer simulation runs, and for $\boldsymbol{\theta}$ conditional on real world data and the surrogate $\hat{\eta}(\cdot; \psi)$. Strategies of this sort can be found in the paper by Higdon et al. (2003).

The alternative methodology proposed herein is motivated by the common situation where a huge bank of simulation runs have already occurred and it is desirable to find

a solution based upon this data without building a surrogate model or making use of iterative MCMC algorithms. This situation arises for a variety of different reasons. Primarily, fully Bayesian fitting of a surrogate model around a huge number of runs can get very expensive and may require more sophisticated modeling techniques than the researchers are willing to entertain. In addition, if one wants to perform the inversion repeatedly or on-line, over different observed $\{\mathbf{x}, \mathbf{y}\}$ sets, multiple MCMC runs will be very expensive and seldom feasible. There is thus a need for Bayesian inverse problem methodology which provides an estimate of the posterior very quickly, using a huge data set of simulation runs, without requiring a sophisticated statistical surrogate model.

Our proposed solution is based upon a weighted resampling of existing simulator runs. The approach shares much in common with the ideas of Bayesian Melding (Poole and Raftery, 2000; Bates et al., 2002), although our work here is more directly suited for use in inverse problems. After the general methodological development in the following section, we will consider in Section 3 two data analysis examples which illustrate practical implementation of the approach and provide encouraging results. The first example is an application to climate modeling, the second involves the problem of measuring the permeability of soil with respect to groundwater flow. In the first example, results from a set of climate simulator runs are compared with actual temperature data. Inference for the inputs to a climate model simulator translates into knowledge about important properties of the climate system. In the second example, the data correspond to groundwater flow at a site with substantial underground pollution, and this is compared with huge bank of simulated flow results for permeability fields generated from a spatial prior. Quantifying the permeability of the ground is important for the soil remediation effort.

2. Methodology

2.1. Sampling Importance Resampling

Sampling Importance Resampling (SIR) is an extremely fast method for sampling from a posterior distribution (Rubin, 1988). Given a sample of points $\tilde{\boldsymbol{\theta}} = \{\tilde{\theta}_1, \ldots, \tilde{\theta}_m\}$ from the probability distribution defined by probability $g(\theta)$, a sample from the distribution characterised by a density proportional to the unnormalized $h(\theta)$ (i.e., from $f(\theta) = h(\theta) \,/\, \int h(\theta) d\theta$) is obtained by resampling $\tilde{\boldsymbol{\theta}}$ with replacement such that the probability assigned to each $\tilde{\theta}_i$ is equal to the SIR weight $w(\tilde{\theta}_i) = w^\star(\tilde{\theta}_i) \,/\, \sum_{j=1}^m w^\star(\tilde{\theta}_i)$, where

$$w^\star(\tilde{\theta}_i) = \frac{h(\tilde{\theta}_i)}{g(\tilde{\theta}_i)}. \tag{3}$$

If the support for f is contained within the support for g, the resampled points will have the desired distribution.

The SIR weights create an empirical probability function for our initial sample, and this defines a discrete approximation to the distribution corresponding to f. Hence, the sum $\hat{I} = \frac{1}{m} \sum w^\star(\tilde{\theta}_i)$ is implicitly relied upon as an approximation to $I = \int h(\theta) d\theta$. The

variance of this estimator is approximately proportional to var$[w(\theta)]$ for $\theta \sim g(\theta)$. Thus the success of an SIR scheme depends upon the variability of the resampling weights, as this is directly related to the variance of our density estimator.

The paper by Skare, Bølviken, and Holden (2003) derives convergence properties for SIR algorithms and proposes a new Improved SIR (I-SIR) algorithm. For their I-SIR with replacement, the SIR weights are adjusted to provide the I-SIR weights $v(\tilde{\theta}_i) = v^\star(\tilde{\theta}_i) \;/\; \sum_{j=1}^m v^\star(\tilde{\theta}_i)$, where

$$v^\star(\tilde{\theta}_i) \propto \frac{w(\tilde{\theta}_i)}{\sum_{\tilde{\theta}_j \in \tilde{\boldsymbol{\theta}}: j \neq i} w(\tilde{\theta}_j)}.$$

They show that the relative point error for the density estimate $\hat{f}_m(\tilde{\theta})$ at $\tilde{\theta} \in \tilde{\boldsymbol{\theta}}$ based upon a Rubin SIR algorithm with sample size m is such that

$$\frac{\hat{f}_n(\tilde{\theta})}{f(\tilde{\theta})} - 1 = \frac{1}{m}(1 - w(\tilde{\theta}) + \text{var}\{w(\boldsymbol{\theta})\}) + O(\frac{1}{m^2})$$

while the relationship for I-SIR is simply

$$\frac{\hat{f}_n(\tilde{\theta})}{f(\tilde{\theta})} - 1 = O(\frac{1}{m^2})$$

It is important to note that the variance of the resampling weights, var$\{w(\boldsymbol{\theta})\}$, appears in the leading terms of the Taylor expansion behind the $O(\frac{1}{n^2})$ term in (4), such that although the effect of highly variable w values is rendered negligible asymptotically for I-SIR, the variance of the weights should still be carefully monitored. Skare et al. also show that an SIR algorithm *without replacement* has better asymptotic convergence in total variation norm than any corresponding SIR algorithm (improved or not) with replacement. However, in our experience, for finite initial sample sizes the without replacement algorithms led to a posterior that was more diffuse than the data would have indicated.

2.2. Sampling Inverse Importance Resampling

The motivating applications consist of the real world data $\mathcal{D} = \{\mathbf{x}, \mathbf{y}\} = \{x_j, y_j : j = 1, \ldots, n\}$, accompanied by a bank of simulated response values corresponding to a sample of input vectors, $\mathcal{S} = \left\{\tilde{\boldsymbol{\theta}}, \mathbf{x}, \{\eta(x_j, \tilde{\theta}_i) : i = 1, \ldots, m; j = 1, \ldots, n\}\right\}$. Recall that the goal is to estimate the *true* θ, where $y(x, \theta) = \eta(x, \theta) + \epsilon(x)$, conditional on $\mathcal{D}$ and $\mathcal{S}$. The inverse likelihood is inexpensive to evaluate at any input location where the simulator has already been run. Suppose that the inputs for our bank of computer output were sampled independently from some distribution defined by the density $g(\theta)$. We are then able to sample from the approximate posterior for θ given $\mathcal{D}$ and $\mathcal{S}$ by resampling with replacement from $\tilde{\boldsymbol{\theta}}$ according to either the SIR or I-SIR probability function built upon the basic weights

$$w^\star(\tilde{\theta}_i) = \frac{\pi(\tilde{\theta}_i)}{g(\tilde{\theta}_i)} \Pr(\mathbf{y} | \{\eta(x_j, \tilde{\theta}_i) : j = 1, \ldots, n\}).$$

Thus application of the I-SIR algorithm is straightforward, with $v^\star(\tilde{\theta}_i)$ defined in terms of $w^\star$ as in Section 2.1, and we are able to obtain a discrete approximation to the inverse problem posterior without having to re-run the computer simulator or fit a surrogate model. Alternatively, we can use these inverse importance weights in a Monte Carlo integration for point estimation of any function of θ.

In order to calculate these weights, we need to know $g(\theta_i)$ at each $\theta_i \in \mathrm{S}$; that is, we need to know the density for the sampled simulator input locations. When a large bank of simulator runs is available, it is often the case that the input configuration has been decided by some previous modeler and there is no available information about the nature of g. In fact, it may seem odd to assume that the sampling was random at all. However, the role of g in the weights is to counter the effect of the original sampling on any posterior estimate, and this remains the case whether or not we believe that g truly describes the sampler's intent. In the case where the variables $\tilde{\theta}$ are discrete with a manageable support, we can compute the empirical probability function to estimate $g(\tilde{\theta})$. When this is not possible, we use a Kernel Density Estimate (KDE) for g. The literature on KDE methods is vast, and the best choice will be application specific. See the books by Bowman and Azzalini (1997) and Simonoff (1996) for examples. With the standard choice of Normal kernels, this generally describes estimates of the sort $\hat{g}(\tilde{\theta}) = \frac{1}{m}\sum_{i=1}^{m} \mathrm{N}(\tilde{\theta}|M_i, Vh^2)$, where V is an estimate of the variance of $\tilde{\theta}$, h is a smoothing parameter or bandwidth, and M_i is a location dependent upon $\tilde{\theta}_i$. The version which we use below, with shrinkage for the individual means, is described in West (1993) such that $M_j = \sqrt{1-h^2}\tilde{\theta}_j + (1-\sqrt{1-h^2})\frac{1}{m}\sum_{i=1}^{m}\tilde{\theta}_i$ and $h = \left(\frac{4}{m(1+2p)}\right)^{\frac{1}{1+4p}}$, where p is the dimension of $\tilde{\theta}$ and V is the sample covariance matrix.

3. Examples

3.1. Climate Model

Computer climate models contain parameterizations that allow for the exploration of climate system properties. In this application we consider the MIT 2D Climate Model described in Sokolov and Stone (1998). In the 2D models, like the one that is considered in this paper, climate system properties are controlled via simple parameterizations. Here we study the following three parameters: climate sensitivity, deep ocean temperature diffusion rate, and net anthropogenic aerosol forcings. Climate sensitivity, S, is defined as the equilibrium global mean temperature response to a doubling of CO_2. This sensitivity has been singled out as a critical parameter with extensive uncertainty. Deep ocean temperature diffusion rate, Kv, is controlled by varying a diffusion coefficient. Net anthropogenic aerosol and unmodeled forcings, key inputs to the simulator, are written here as *Faer*. These quantities can not be derived from physical principles and must be inferred by comparing computer output to historical records in a typical inverse problem setting. Estimating their values and assessing their variability is key for forward looking projections of climate change that

may be used for policy making.

The output from the MIT 2D climate model, considered in this paper, consists of temperatures over a grid of zonal bands corresponding to 46 latitudes, averaging over all longitudes in the band. It has 11 vertical layers for a grid of 506 cells for every time step. Typical output corresponds to periods of 50 years with data every 30 minutes. This information can be summarized in a variety of ways. A standard approach is to run the MIT 2D climate model for many choices of the uncertain parameters *S*, *Kv* and *Faer*, selected systematically on a non-uniform grid (Forest et al., 2000, 2001, 2002, 2005). The grid considered in this paper consists of 499 points. To summarize this data in a way that is useful for understanding possible global climate change, three different statistics or "diagnostics" are used. See Sansó et al. (2008) for further details. In this paper we focus on the *Deep ocean* temperature trend, calculated for the period [1948–1995]. The corresponding observational data are annual deep ocean temperature measurements obtained from Levitus et al. (2000), for the period [1950–1995].

Following the notation in previous sections we have that θ =(*S*, *Kv*, *Faer*). The dependence of y on θ is characterized by the deep ocean temperature gradient, β_θ. The assumption of a linear trend leads to the full model, $y_i = \eta(x_i, \theta) + \varepsilon_i = \alpha_\theta + x_i\beta_\theta + \varepsilon_i$, where the y_i correspond to the annual deep ocean temperature obtained from the Levitus climatology, $x_i = (year_i - 1950)$, $\alpha_\theta = \bar{y} - \beta_\theta\bar{x}$, and $\varepsilon_i \overset{iid}{\sim} \mathrm{N}(0, \sigma^2)$, $i = 1, \ldots, 45$. The probability model is thus

$$\Pr(\mathbf{y}|\beta_\theta, \mathbf{y}, \mathbf{x}) \propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{45}(y_i - \alpha_\theta - \beta_\theta x_i)^2\right).$$

The simulator output is an estimate of deep ocean temperature gradient conditional on an input $\tilde{\theta}$ vector of values for (*S*, *Kv*, *Faer*). Using different starting values to initialize the model, it is possible to obtain ensembles of temperature simulations. In our application we obtain four random ensembles, leading to four output temperature trends, say $\{b_{i1}, \ldots, b_{i4}\}$, for each input $\tilde{\theta}_i$. The full simulator output is defined for the purposes of the inverse problem as $\eta(x, \tilde{\theta}_i) = \alpha_{\tilde{\theta}_i} + x\beta_{\tilde{\theta}_i}$, where $\beta_{\tilde{\theta}_i} = \frac{1}{4}\sum_{j=1}^{4} b_{ij}$ and $\alpha_\theta = \bar{y} - \bar{x}\beta_{\tilde{\theta}_i}$. Assuming a variance of $\tau^2_{\tilde{\theta}_i}$ for the four b_{ij}, this implies that $\mathrm{var}\{\eta(x, \tilde{\theta}_i)\} = \tau_i^2(x - \bar{x})^2/4$ introduces an additional source of variability that needs to be incorporated into the conditional likelihood. Setting each τ_i^2 to the sample variance of $\{\beta_{\theta i}\}_{i=1..4}$, and plugging-in $s^2 = \frac{1}{43}\sum_{i=1}^{45}(y_i - \alpha_{OLS} - x_i\beta_{OLS})^2$ in place of σ^2 (α_{OLS} and β_{OLS} are the least squares temperature trend estimates), we obtain the completed conditional likelihood for each $\tilde{\theta}_i$:

$$\Pr(\mathbf{y}|\beta_{\tilde{\theta}_i}, \mathbf{y}, \mathbf{x}) \propto \exp\left(-\frac{\sum_{i=1}^{45}(y_i - \alpha_{\tilde{\theta}_i} - x_i\beta_{\tilde{\theta}_i})^2}{2\left(s^2 + \frac{\tau_i^2}{4}(x_i - \bar{x})^2\right)}\right).$$

An informative prior, $\pi(\theta)$, was elicited from the literature about climate properties. According to this, each input variable is assigned an independent prior with $S/6 \sim \mathrm{Be}(3.5, 6)$, $Kv/15 \sim \mathrm{Be}(2.85, 14)$, and $(Faer + 1.5)/2 \sim \mathrm{Be}(4, 4)$, where $\mathrm{Be}(a, b)$ denotes the Beta probability distribution with mean a/b.

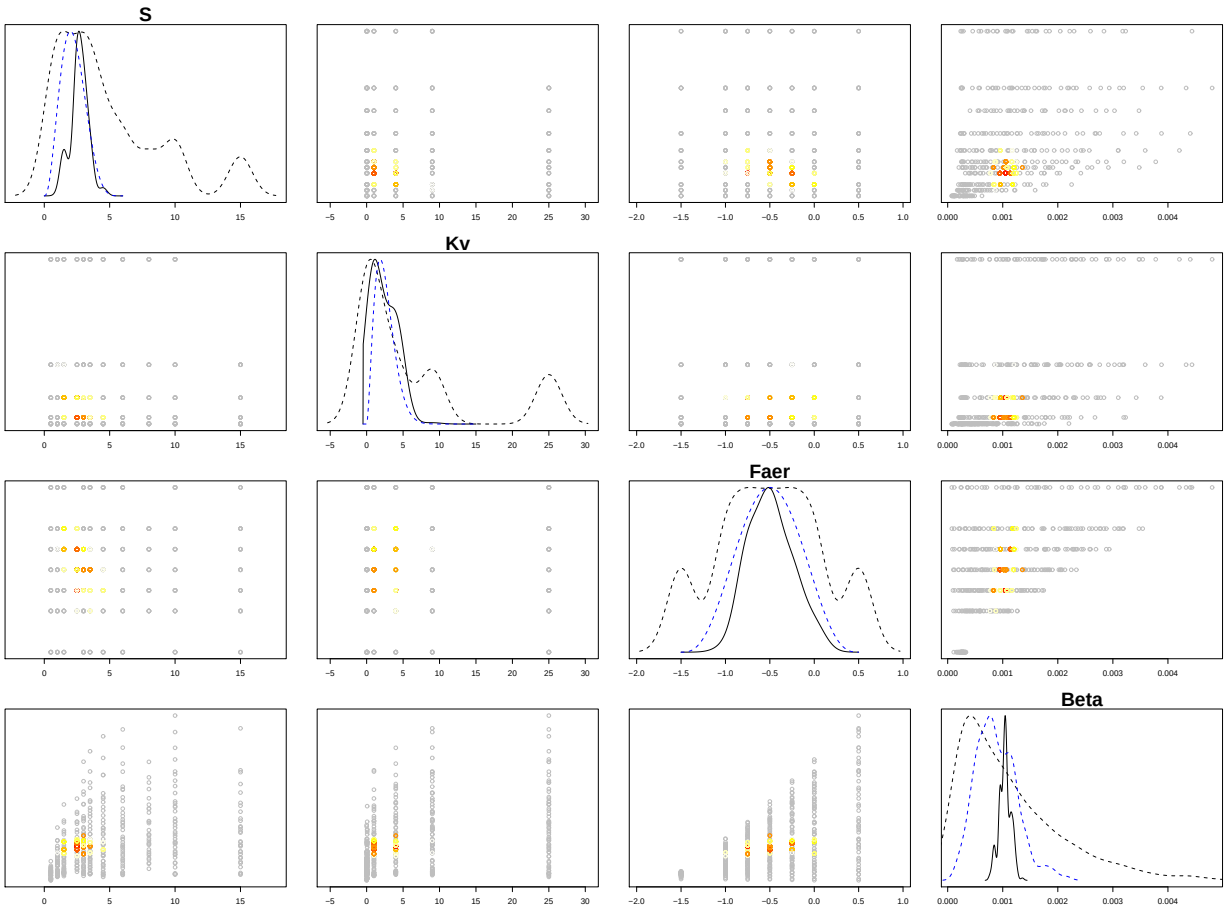

Figure 1. Resampling and Density Estimates under an informative prior. The solid line is the KDE of the posterior resample and the dotted black line is the KDE of the original sample. Scatterplots show the original sample and the colours indicate resampling frequency (rising from yellow to red; grey were not resampled).

Following the steps in Section 2.2, this prior is combined with $\Pr(\mathbf{y}|\beta_{\tilde{\theta}_i}, \mathbf{y}, \mathbf{x})$ and the KDE for $g(\tilde{\theta})$, the sampling distribution for the input vectors, to obtain I-SIR weights. Figure 3.1 illustrates the results of the resampling based upon these weights. Due to the sparsity of the original sample, the posterior resampling is concentrated on a relatively small number of particles. However, even in the presence of these limitations, it is possible through careful choice of posterior KDE to properly visualize the posterior. Certainly, we are able to see the strong influence of the prior. But it is also evident that the data are influencing our posterior, and we see that our posterior uncertainty about S has been pushed to the right. This is an important result since, as mentioned above, this is a critical parameter with extensive uncertainty.

We also obtained a posterior resample under a non-informative prior (i.e., $\pi(\tilde{\theta}_i) = \frac{1}{m}$) and we see in Figure 3.1 that the results are very different from those obtained under the informative prior, with more uniform resampling weights and a higher proportion of the original particles included in the posterior sample. As the posterior weights were more uniform, we have a larger number of sample particles included in the resample.

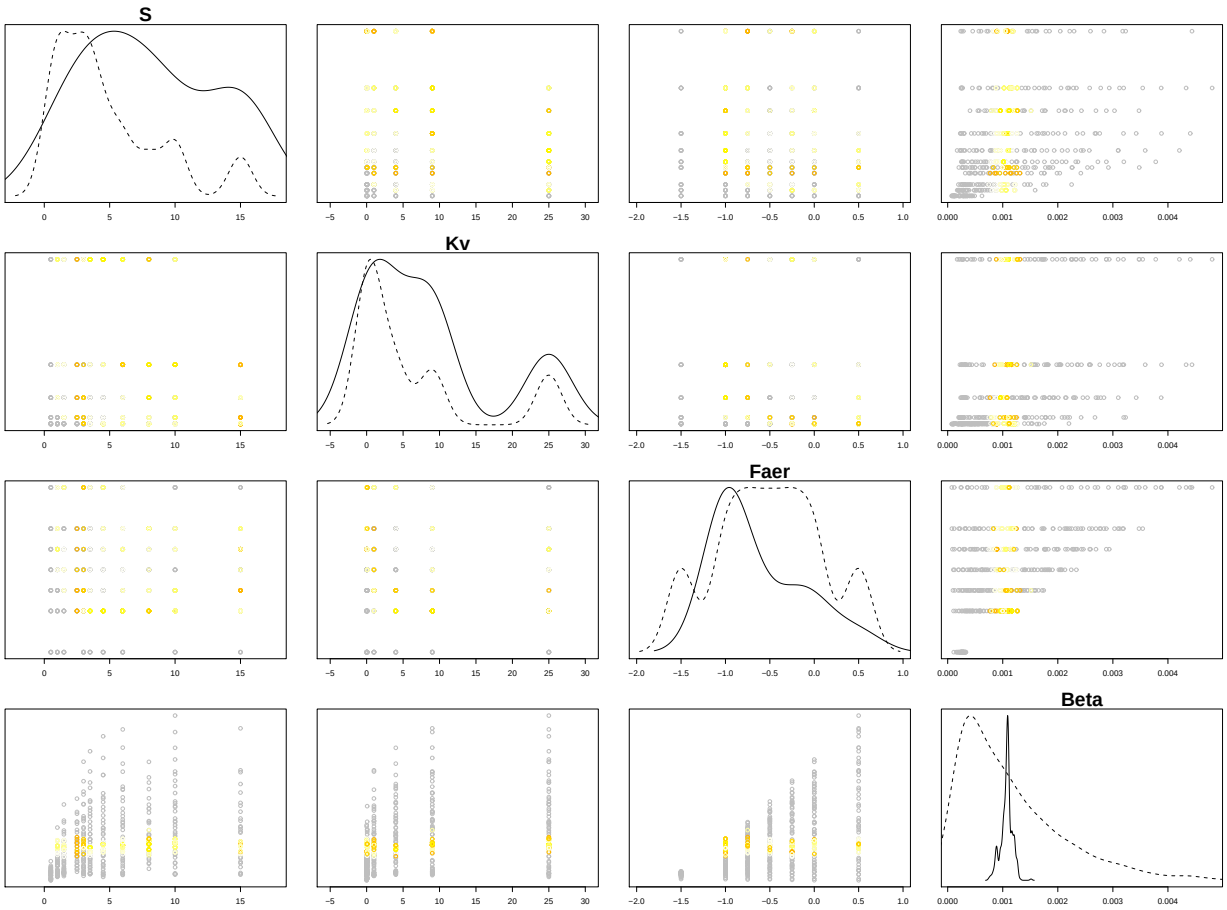

Figure 2. Resampling and Density Estimates with a non-informative prior. The solid line is the KDE of the posterior resample, the dotted black line is the KDE of the original sample, and the blue line is the prior density. Scatterplots show the original sample and the colours indicate resampling frequency (rising from yellow to red; grey were not resampled).

However, the data indicate that S is likely near the higher end of our support, with or without prior information on the variables.

3.2. Groundwater application

Our second example is one in groundwater flow. Of interest is a spatial field of permeability values, parameterized on a 14x11 grid. Thus we need to perform inference on a high-dimensional (154) but highly correlated parameter space.

This particular data set comes from part of a larger study Annable et al. (1998); Yoon (2000) of an area at the Hill Air Force Base in Utah, where flow experiments were done to learn about the soil structure as part of a project in cleaning up a polluted section of the ground. In order to perform effective soil remediation, it is necessary for the engineers to understand the soil structure, in particular the permeability field. Permeability is a measure of how well water flows through the soil at a point, and it

varies spatially. It is difficult to measure, with core samples providing expensive yet noisy estimates at point locations, so flow experiments are often conducted instead. Thus the goal is to find the distribution of permeability configurations most consistent with the flow data, which involves matching the observed flows to the output of flow experiments on the proposed permeability configuration as determined by computer simulators. Such simulators solve systems of differential equations numerically to determine flow experiment outputs under various possible aquifer configurations.

The field experiment involves four injector wells along the left side of the aquifer which force water across the site to the three production wells along the right side of the aquifer. Water is pumped in until the system reaches equilibrium. Then, a tracer is injected at the injector wells, and the amount of time taken for the tracer to reach each of the five sampling wells (situated between the injector and production wells) is recorded as the *breakthrough time*. These five breakthrough times are the available data. A previous analysis of this data set, along with additional details on the experiment, can be found in Lee, et al. (2002). Following the standard approach in the literature, we use an independent Gaussian likelihood for the breakthrough times

$$\Pr(\mathbf{y}|\theta,\sigma^2) \propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{5}(y_i-\eta(\theta))^2\right),$$

where the observed breakthrough times at the sampling wells are denoted by y_i, the θ is the theoretical 14x11 permeability field, and $\eta(\theta)$ is the mean breakthrough time corresponding to this permeability structure.

The simulated data consists of mean breakthrough times for 1,000,000 permeability fields which were run through the program S3D developed by King and Datta-Gupta (1998). These fields are all generated from a Gaussian process with a Gaussian correlation structure and correlation parameters consistent with the expected geologic structure of the study site. In other words, $\tilde{\boldsymbol{\theta}}$ was generated from our spatial prior for the true θ, such that $g(\theta) = \pi(\theta)$ in the notation of Section 2 and there is no need for a KDE. Hence, the I-SIR probabilities for these 10^6 fields are derived from the basic resampling weights, $w^\star(\tilde{\theta}_i) =\propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{5}(y_i-\eta(\tilde{\theta}_i))^2\right)$, where the expected breakthrough time produced by the computer simulator is denoted by $\eta(\tilde{\theta}_i)$ and σ^2 is fixed to a value based on input from the geologists.

We resampled the permeability weights according to the derived I-SIR probabilities. Figure 3 shows the resulting posterior mean permeability field in the upper left plot, with blue (dark) representing lower permeability values and yellow (light) for higher values. The resulting picture is consistent with the slow breakthrough time observed near the center of the site and the fast breakthrough time observed at a sampling well in the lower middle section of the site. The upper right plot shows the spread of a 90% credible interval, with more variability apparent for the larger permeability values. The lower two plots show two of the modes in the resampling of the fields.

Figure 4 shows the re-sampling frequencies for each of the million fields that were run through the simulator. There are four or five more prominent fields, but there are

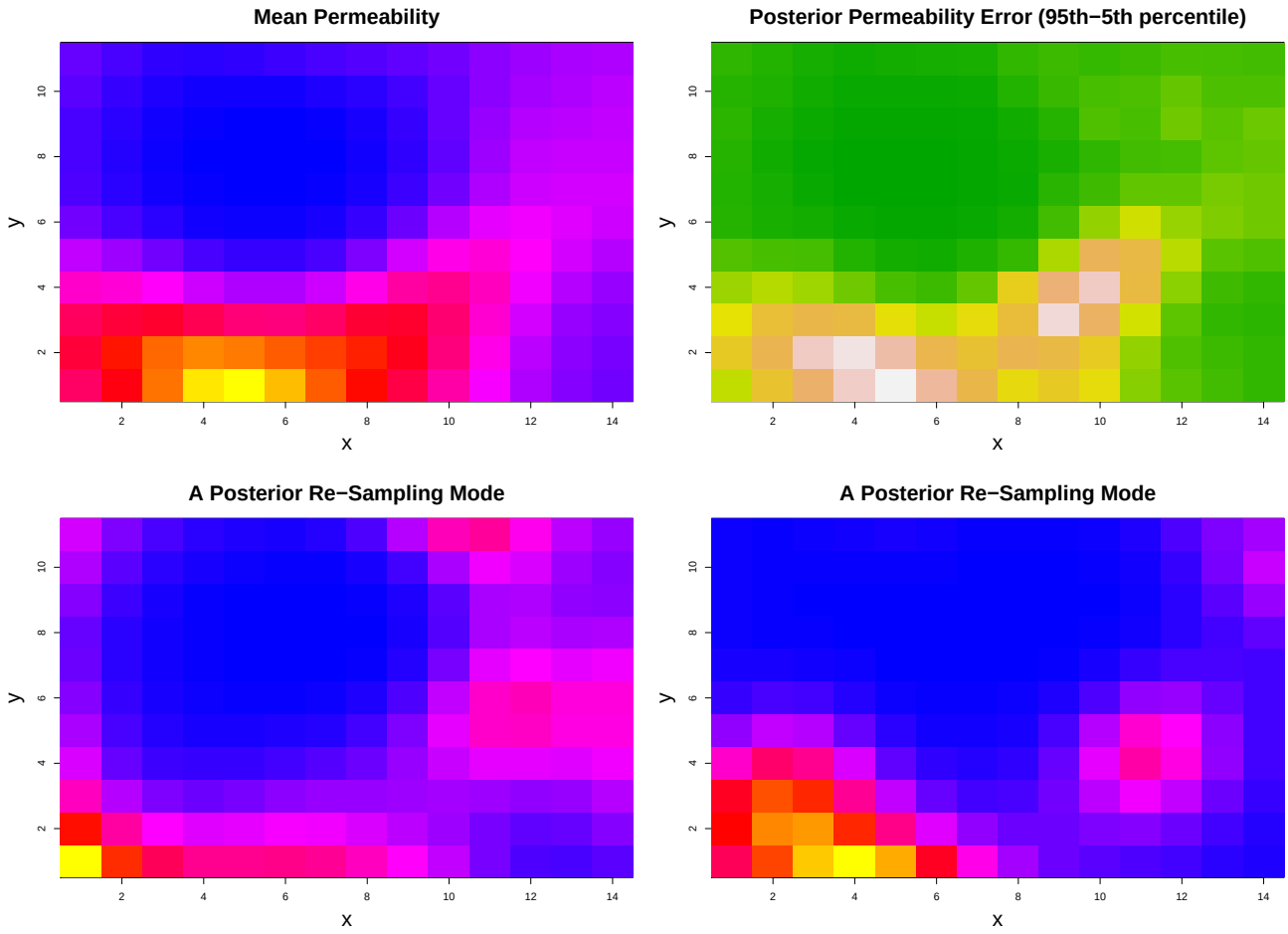

Figure 3. Posterior mean permeability field, variability estimate, and two of the highly weighted sample fields. Permeability is rising from blue to yellow, and variability is rising from green to white.

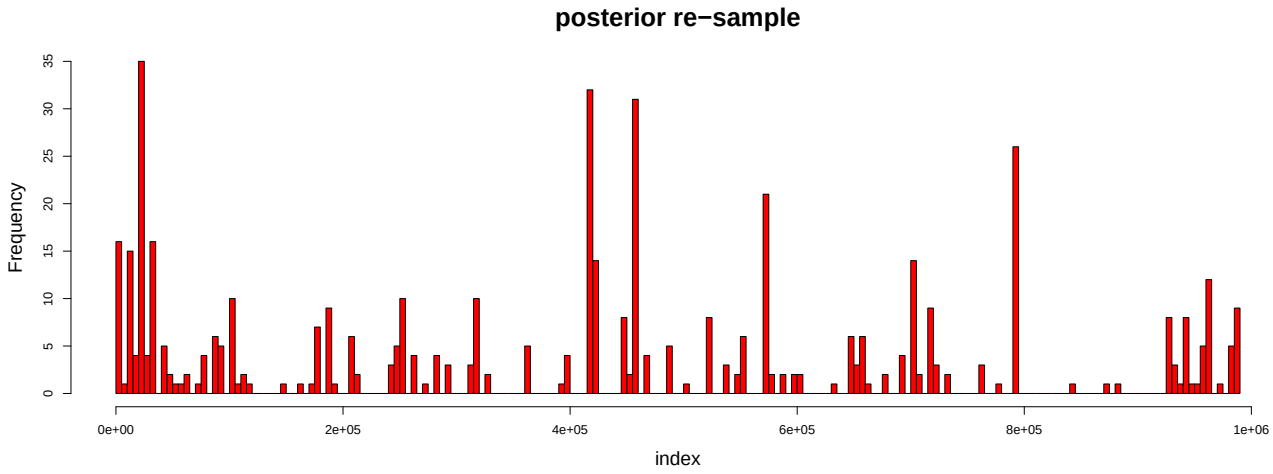

Figure 4. Resampling frequencies for the permeability fields. In a 500 point resample, 35 was the maximum resampling frequency.

a number of fields that were re-sampled and no single field dominates the sampling, so we are pleased by the lack of weight degeneracy, and thus more confident in our results. We are also pleased that our present results are comparable to those of earlier studies Lee et al. (2002), but here our resampling took only about five seconds, compared to the week-long MCMC runs that had been done previously.

4. Conclusion

In each of the two applications, our results were comparable to those obtained through alternative, more conventional, methodologies. Such methods, outlined in the already cited literature, typically involved some sort of surrogate modeling and MCMC sampling and required days of computing time. Resampling by inverse importance provided solutions that, not including the time spent running the computer model at the bank of input locations, required only 5-10 seconds of computing time. While we make no claim that our method is superior to or should replace the conventional solutions, it may be the best choice when a very fast solution is required. If, for example, the inversion needs to be performed repeatedly over a large set of observed data-values, or there already exists a massive bank of simulator output, resampling for inverse problems provides a natural solution.

Acknowledgments

The authors thank Dave Higdon of Los Alamos National Laboratory for providing the initial idea and impetus. The data for the MIT2D climate model were provided by Chris Forest. This work was partially supported by Los Alamos National Laboratory subcontract 26918-002-06 and National Science Foundation grant NSF-Geomath 0417753.

References

Annable, M. D., Rao, P. S. C., Hatfield, K., Graham, W. D., Wood, A. L., and Enfield, C. G. (1998), "Partitioning Tracers for Measuring Residual NAPL: Field-Scale Test Results," *Journal of Environmental Engineering*, 124, 498–503.

Balakrishnan, S. and Madigan, D. (2006), "A one-pass sequential Monte Carlo method for Bayesian analysis of massive datasets," *Bayesian Analysis*, 1.

Bates, S., Cullen, A., and Raftery, A. E. (2002), "Bayesian Uncertainty Assessment in Multicompartment Deterministic Simulation Models for Environmental Risk Assessment," *Environmetrics*, 13, 1–17.

Bowman, A. W. and Azzalini, A. (1997), *Applied Smoothing Techniques for Data Analysis*, Oxford University Press.

Cornford, D., Csatò, L., Evans, D., and Opper, M. (2004), "Bayesian analysis of the scatterometer wind retrieval inverse problem: some new approaches," *Journal of the Royal Statistical Society, Series B*, 66, 609–626.

Forest, C., Stone, P., and Sokolov, A. (2005), "Estimated PDFs of climate system properties including natural and antrophogenic forcings," Tech. rep., MIT, Submitted to GRL.

Forest, C. E., Allen, M. R., Stone, P. H., and Sokolov, A. P. (2000), "Constraining uncertainties in climate models using climate change detection methods," *Geophysical Research Letters*, 27, 569–572.

Forest, C. E., Allen, M. R., Sokolov, A. P., and Stone, P. H. (2001), "Constrainting climate model properties using optimal fingerprint detection methods," *Climate Dynamics*, 18, 277–295.

Forest, C. E., Stone, P. H., Sokolov, A. P., and Allen, M. R. (2002), "Quantifying uncertainties in climate system properties with the use of recent climate observations," *Science*, 295, 113–117.

Gamerman, D. and Lopes, H. F. (2006) *Markov Chain Monte Carlo - Stochastic Simulation for Bayesian Inference*. London, UK: Chapman and Hall, second edn.

Higdon, D., Lee, H., and Holloman, C. (2003), "Markov chain Monte Carlo-based approaches for inference in computationally intensive inverse problems," *Bayesian Statistics*, 7.

Kaipio, J. and Somersalo, E. (2004), *Statistical and Computational Inverse Problems*, Springer-Verlag.

Kennedy, M. and O'Hagan, A. (2001), "Bayesian Calibration of Computer Models," *Journal of the Royal Statistical Society, Series B Statistical Methodology*, 63, 425–464.

King, M. J. and Datta-Gupta, A. (1998), "Streamline Simulation: A Current Perspective," *In Situ*, 22, 91–140.

Lee, H., Higdon, D., Bi, Z., Ferreira, M., and West, M. (2002), "Markov Random Field Models for High-Dimensional Parameters in Simulations of Fluid Flow in Porous Media," *Technometrics*, 44, 230–241.

Levitus, S., Antonov, J., Boyer, T. P., and Stephens, C. (2000), "Warming of the World Ocean," *Science*, 287, 2225–2229.

Liu, J. S., Chen, R., and Wong, W. H. (1998), "Rejection control and sequential importance sampling," *Journal of the American Statistical Association*, 93, 1022–1031.

Poole, D. and Raftery, A. E. (2000), "Inference for Deterministic Simulation Models: The Bayesian Melding Approach," *Journal of the American Statistical Association*, 95, 1244–1255.

Rubin, D. (1988), "Using the SIR algorithm to simulate posterior distributions by data augmentation," in *Bayesian statistics 3*, eds. J. Bernardo, M. DeGroot, and A. Lindley, D. amd Smith, Oxford University Press Inc.

Sacks, J., Welch, W., Mitchell, T., and Wynn, H. (1989), "Design and analysis of computer experiments," *Statistical Science*, 4, 409–435.

Sansó, B., Forest, C., and Zantedeschi, D. (2008), "Inferring Climate System Properties Using a Computer Model", with discussion," *Bayesian Analysis*, 03, pp 1–62.

Santner, T., Williams, B., and Notz, W. (2003), *The Design and Analysis of Computer Experiments*, Springer-Verlag.

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

Skare, Ø., Bølviken, E., and Holden, L. (2003), "Improved sampling importance resampling and reduced bias importance sampling," *Scandinavian Journal of Statistics*, 30, 719–737.

Sokolov, A. P. and Stone, P. H. (1998), "A flexible climate model for use in integrated assessments," *Climate Dynamics*, 14, 291–303.

West, M. (1993), "Approximating posterior distributions by mixtures," *Journal of the Royal Statistical Society, Series B, Methodological*, 55, 409–422.

Yoon, S. (2000), "Dynamic Data Integration Into High Resolution Reservoir Models Using Streamline-Based Inversion," Ph.D. thesis, Texas A&M University, Department of Petroleum Engineering.