Fast Bayesian Inference for Computer Simulation Inverse Problems

Matthew Taddy, Herbert K. H. Lee & Bruno Sansó
University of California, Santa Cruz,
Department of Applied Mathematics and Statistics

January 11, 2008

Abstract

Computer models that simulate phenomena regulated by complicated dependencies on unknown parameters are increasingly used in environmental applications. In this paper we present two such applications where the values of the parameters need to be inferred from scarce observations and abundant simulated output. One consists of a climate simulator and the other of a groundwater flow model. Detailed exploration of the parameter space for a computer simulator can be costly, as running the code for a given parameter configuration can be very time consuming. The common Bayesian approach is to develop a surrogate model for the simulator and to use it to find the likelihood for sampling from the posterior distribution of inputs without having to re-run the simulator. However, in many cases like the ones presented in this paper, there is a large bank of simulated values and the need for a faster algorithm that does not require predicted output at new locations. We discuss a sampling importance resampling (SIR) algorithm that works in conjunction with kernel density estimation to resample from the original computer output according to the posterior distribution of input values, which allows for a fully Bayesian analysis without the need for MCMC.

Keywords: Importance Sampling, SIR, Computer Model, Computer Experiment.

1 Introduction

Sophisticated environmental studies often require the use of mathematical models depending on physical parameters. Additionally, in some cases, the computer implementation of the models requires tuning that involves setting some parameter values. The setting or calibration of such parameters is usually done by comparing the output of the model to actual observations obtained from experimentation or historical records. The optimal value of the parameters corresponds to the solution of an inverse problem. It is often the case that, given a set of parameter values, obtaining the corresponding model output , i.e. solving the forward problem, is very time consuming. This makes it impossible to build a likelihood function and embed its evaluation within an iterative procedure. In some cases, though, a record

of model output is available for a large set of parameter combinations. Alternatively a fast statistical model may be available as a surrogate for the complicated mathematical model. In many cases relevant prior knowledge is available and samples from a prior distribution for the parameters will be easy to obtain.

The field of design and analysis of computer experiments has received considerable attention during the last two decades. See, for example (Sacks et al., 1989; Kennedy and O'Hagan, 2001; Santner et al., 2003), with most of the attention directed towards applications in engineering. Here we focus first on an application to climate modeling and second on the problem of measuring the permeability of soil with respect to groundwater flow. In the first example inference for the inputs to a climate model simulator translates into knowledge about important properties of the climate system. The problem involves a relatively small bank of computer runs over a poorly spaced input sample. In the second example, a huge bank of permeability fields was generated from a spatial prior and input to the computer model. The data correspond to a site with substantial underground pollution. Quantifying the permeability of the ground is important for the soil remediation effort.

The classic setting of a computer experiment is that there is some true process, $\zeta(x, \theta)$, that in the real world characterizes the functional relationship between sets of inputs $\{x, \theta\}$, with x known and θ unknown, and a potentially multivariate output y. With additive noise, $y(x, \theta) = \zeta(x, \theta) + \varepsilon(x)$, where ε is a zero mean random variable. If $\eta(x, \theta)$ is the computer simulation function:

$$\begin{aligned} y(x, \theta) &= \eta(x, \theta) + e(x, \theta) \\ e(x, \theta) &= \zeta(x, t) - \eta(x, \theta) + \varepsilon(x) \; . \end{aligned}$$

The error, $e(\cdot)$, includes random noise as well as the bias between simulation and reality.

The inverse problem involves solving for the values of the underlying variables that have led to an observed data set. Given a set of true response values $\mathbf{y}$, with known inputs $\mathbf{x}$, what is our uncertainty about the unknown input θ values that led to this response? An overview of such problems can be found in the book by Kaipio & Somersalo (2004). We take a Bayesian approach in order to fully account for our uncertainty. Characteristically, there is a single $\{\mathbf{x},\ \mathbf{y}\}$ data set corresponding to an unknown input θ vector, and a set of simulated response values for known inputs from our computer code. If the model is $y(x, \theta) = \eta(x, \theta) + (\zeta(x, \theta) - \eta(x, \theta)) + \epsilon(x)$ as defined above, then the goal is to estimate $x|\tilde{y}$ in the presence of the random error from ϵ and the bias in our code. If the data consist of only one set $\{\mathbf{x},\ \mathbf{y}\}$ from a single unknown input set θ, then we cannot hope to estimate the bias and must assume that $\eta(x, \cdot) = \zeta(x, \cdot)$. The model is then $y(x, \theta) = \eta(x, \theta) + \varepsilon(x)$. Note that this simplification, although necessary, should be used with caution. If we proceed as such, even though the code does not accurately characterize the relationship between inputs and outputs, then we will be merely *tuning* the computer simulator to match the observed output rather than actually *inverting*.

The posterior density (ignoring the many possible nuisance parameters) is then of the form:

$$P(\theta|\mathbf{x}, \mathbf{y}, \eta) \propto \mathcal{L}(\theta|\mathbf{x}, \mathbf{y}, \eta)\pi(\theta|\mathbf{x}) \; .$$

Because we have assumed that $\eta = \zeta$, the likelihood for θ will be based upon only $\hat{\varepsilon}(\mathbf{x}) = \mathbf{y} - \eta(\mathbf{x}, \theta)$. For example, if $\varepsilon(x) \overset{iid}{\sim} \mathrm{N}(0, \sigma^2)$ then $\mathcal{L}(\theta|\mathbf{x}, \mathbf{y}, \eta) \propto \exp\left(-\frac{1}{2\sigma^2}\sum_{x\in\mathbf{x}} \hat{\varepsilon}(x)^2\right)$. And if $\varepsilon(x) \sim \mathrm{GP}(x)$, then $\mathcal{L}(\theta|\mathbf{x}, \mathbf{y}, \eta) = \mathrm{N}(\hat{\varepsilon}(\mathbf{x})|0, \mathrm{K}(\mathbf{x}))$, where $\mathrm{K}(\cdot)$ is the covariance implied by the Gaussian Process, $\mathrm{GP}(\cdot)$. The prior $\pi(\theta|\mathbf{x})$ can also have many different forms. The paper by Cornford et al. (2004) includes a nice application where t are wind vectors over a spatial field, and they assume a GP prior. The two applications considered in this paper describe independent priors over each dimension of θ, as well as a GP prior over a two-dimensional spatial θ.

It is not often possible to find an analytic solution for the posterior. Common strategies for inference center on Markov Chain Monte Carlo methods. If η is easy to evaluate, this inference is straightforward. If η is expensive to run, then we need to calculate the likelihood with a surrogate model based upon a limited number of runs. Strategies of this sort can be found in the paper by Higdon et al. (2003). We were faced with the situation where there is a huge bank of simulation runs, but there is no appetite or ability to generate new points from η. Building a surrogate model around a huge number of runs can get very expensive. Often, one wants to perform the inversion repeatedly, over different observed $\mathbf{y}$ values. MCMC with an expensive surrogate is not feasible. We need a "sledgehammer" methodology: something that provides an estimate of the posterior very quickly, using a huge dataset of simulation runs. We thus turn to resampling methods, as described in the next two sections. Our 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 tuned for use in inverse problems.

2 Sampling Importance Resampling

SIR is an extremely fast method for sampling from a posterior distribution (Rubin, 1988). If you have a sample of points $\{\theta_i\}_{i=1,\ldots,n}$ from the probability distribution defined by $g(\theta)$, and you desire a sample of points from $f(\theta)$, then you can resample with replacement according to weights

$$w(\theta_i) = \frac{f(\theta_i)}{g(\theta_i)} \, .$$

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

There are a variety of ways to measure the performance of a SIR algorithm. The SIR weights create an empirical probability function for our initial sample. We are implicitly using $\hat{I} = \frac{1}{n}\sum w(\theta_i)$, where $\mathbf{S}$ is of size n, as an approximation to $I = \int \mathcal{L}(\theta|\mathbf{x}, \mathbf{y}, \eta)\pi(\theta|\mathbf{x})d\theta$. (This is inherent in our construction of a discrete approximation to the true posterior.) The variance of this estimator is approximately proportional to $\mathrm{var}[\mathrm{w}(\theta)]$. Thus it is essential that we monitor the variance of the weights, as it is directly related to the variance of our implicit estimator. One can see this relationship by considering the sample variance $\approx \frac{1}{n}\frac{\sum(w(\theta)-\hat{I})^2}{n-1}$ for reasonably large n, since $\mathbb{E}[w(x)] = 1, \ \forall x$.

Liu et al. (1998) propose monitoring by using the coefficient of variation, $C(w)$, which is computed

$$C^2(w) = \frac{1}{n}\sum_{j=1}^{n}\{nw(t_j) - 1\}^2 .$$

This estimator is an approximation to the variance of the w values. They also propose the use of a heuristic approximation:

$$\mathrm{ESS}^1 = \frac{n}{1 + C^2(w)} .$$

As it is for much of the research surrounding SIR, the motivating example here is from particle filtering. The idea is that these statistics are used to monitor the degeneracy of the sample, and thus give an indication of when SIR should be used to rejuvenate the particle sample. These same ideas are used by Balakrishnan & Madigan (2006) to rejuvenate the weighted sample from the posterior as the analysis moves through a massive dataset. They use an alternative approximation to the effective sample size,

$$\mathrm{ESS}^2 = \frac{\left(\sum w(\theta_i)\right)^2}{\sum w(\theta_i)^2} .$$

We can use these ideas of rejuvenation and weight degeneracy not so much to know when SIR is appropriate (with a large arbitrary sample of inputs, the inverse problem will almost always lead to a highly variable set of w's), but to monitor our algorithm and understand the properties of our estimates.

The paper by Skare, Bølviken, & Holden (2003) derives some convergence properties for SIR algorithms and proposes a new 'Improved SIR' algorithm, that has better asymptotic properties than the Rubin version. For their Improved SIR with replacement, the resampling weights are adjusted:

$$\tilde{w}(\theta_i) \propto \frac{w(\theta_i)}{\sum_{\theta_j \in \mathbf{S}, j \neq i} w(\theta_j)}$$

where w is the standard SIR weight, and S is the bank of inputs. They show that, if $\hat{f}_n$ is the SIR estimate of f from a bank of size n, then the relative point error at $\theta' \in \mathbf{S}$ for a Rubin SIR algorithm satisfies:

$$\frac{\hat{f}_n(\theta')}{f(\theta')} - 1 = \frac{1}{n}(1 - w(\theta') + \mathrm{var}\{\mathrm{w}(\theta)\}) + \mathrm{O}(\frac{1}{\mathrm{n}^2}) ,$$

while the relationship for improved SIR is simply

$$\frac{\hat{f}_n(\theta')}{f(\theta')} - 1 = O(\frac{1}{n^2}) .$$

It is important to note that the variance of the resampling weights, $\mathrm{var}\{\mathrm{w}(\theta)\}$, is in the leading terms of the Taylor expansion behind the $O(\frac{1}{n^2})$ term, such that although the effect of highly variable w values is now negligible asymptotically, 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 practice we found that the performance of a without replacement algorithm is quite poor in inverse problem settings, where we are generally dealing with a somewhat limited set of points of peaked likelihood, even when the magnitude of $\mathbf{S}$ is quite large. In our experience the without replacement algorithms led to a posterior that was more diffuse than the data would have indicated.

3 Sampling Inverse Importance Resampling

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 $g(\cdot)$. Then we can resample from the posterior after designating the sampling weights,

$$w(\theta_i) = \frac{f(\theta_i)}{g(\theta_i)} = \frac{\pi(\theta_i)\mathcal{L}(\theta|\mathbf{x},\mathbf{y},\eta)}{g(\theta_i)} \quad .$$

Thus application of the SIR algorithm is straightforward and we are able to obtain a discrete approximation to the inverse problem posterior without having to re-run the computer simulator. Alternatively, we can use these inverse importance weights in a Monte Carlo integration for any point estimation. In the presence of nuisance parameters, we can sample from the hyperprior at some subset of the resampling iterations, and couple these values with the sample from $g(\theta)$ for resampling.

In order to calculate these weights, we need to know $g(\theta_i)$ at each $\theta_i \in \mathrm{S}$. Often, with computer models where the input configuration has been decided by some previous user, we will have no knowledge (at least no usable knowledge) 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 θ_i are discrete with a manageable support, we can compute the empirical probability function to estimate the $g(\theta_i)$ marginals. When this is not possible, we use a Kernel Density Estimate (KDE). For Normal kernels, this generally describes estimates of the sort

$$\hat{g}(\theta) = \tfrac{1}{n}\textstyle\sum_j \mathrm{N}(\theta|m_j, Vh^2)$$

V is an estimate of the variance of $g(t)$, h is a smoothing parameter or bandwidth, and m_j is a location dependent upon θ_j. The version which we use below, with shrinkage for the individual means, is described in West (1993).

$$\begin{aligned} \mathbf{m}_j &= \theta_j\sqrt{1-h^2} + \bar{\theta}(1-\sqrt{1-h^2}) \\ h &= \left(\frac{4}{n(1+2p)}\right)^{\frac{1}{1+4p}}, \quad p = \dim(\theta) \end{aligned}$$

The literature on KDE methods is vast, and the best choice will be application specific. See the books by Bowman & Azzalini (1997) and Simonoff (1996) for examples. This estimation

is, however, a critical point for our methodology. Not only is the KDE often required for estimation of the initial g sampling distribution, the resultant resample is only a discrete approximation, such that some form of density estimation will be required for visualization of a continuous posterior.

4 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 the 499 times series in a way that is useful to understand possible global climate changes, three different statistics or "diagnostics" are used. See Sansó et al. (2007) 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 $\theta =$(**S**, **Kv**, **Faer**). The model output, β_θ, is the deep ocean temperature trend. The assumption of a linear trend leads to the model,

$$y_i = \alpha_\theta + x_i\beta_\theta + \delta_i \ ,$$

where the y_i correspond to the annual deep ocean temperature obtained from the Levitus climatology, $\mathbf{x} = (YEAR - 1950)$, $\alpha_\theta = \bar{y} - \beta_\theta\bar{x}$, and $\delta_i \overset{iid}{\sim} \mathrm{N}(0, \sigma^2)$, $i = 1, \ldots, 45$. The

conditional likelihood is thus

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

Using different starting values to initialize the model, it is possible to obtain ensembles of temperature simulations. Thus we are not faced with the standard situation of a single-input/single-output deterministic simulator. In our application we used four ensembles, so that we have four outputs, say $\beta_{\theta j}$, for each input θ. Assume that $\beta_{\theta j} \sim \mathrm{N}(\beta_\theta, \tau_\theta^2)$ and set:

$$\eta(x, \theta) = \hat{\alpha}_\theta + x\hat{\beta}_\theta$$

$$\hat{\beta}_\theta = \frac{1}{4}\sum_{j=1}^{4}\beta_{\theta j} \qquad \hat{\alpha}_\theta = \bar{y} - \bar{x}\hat{\beta}_\theta \ .$$

Then we have

$$\begin{aligned} \mathbb{E}[\eta(x,\theta)] &= \alpha_\theta + x\beta_\theta \\ \mathrm{var}\{\eta(\mathrm{x},\theta)\} &= \frac{\tau_\theta^2}{4}(x - \bar{x})^2 \ . \end{aligned}$$

We put flat priors on β_θ and set

$$\begin{aligned} \tau^2 &= \hat{\tau}^2 = \mathrm{var}\{\beta_{\theta \mathrm{i}}\}_{\mathrm{i}=1..4} \\ \sigma^2 &= s^2 = \frac{1}{43}\sum_{i=1}^{45}(y_i - \alpha_{mle} - x_i\beta_{mle})^2 \ . \end{aligned}$$

Then

$$\begin{aligned} y_i &= \eta(x_i, \theta) + \varepsilon(x_i) \\ \varepsilon(x_i) &\sim \mathrm{N}(0, s^2 + \frac{\hat{\tau}^2}{4}(x_i - \bar{x})^2) \\ &\Rightarrow \\ P(\theta|\mathbf{y}, \mathbf{x}, \eta) &\propto \exp\left(\sum_{i=1}^{45} -\frac{(y_i - \eta(x_i,\theta))^2}{2(s^2 + \frac{\hat{\tau}^2}{4}(x_i - \bar{x})^2)}\right)\pi(\theta) \ . \end{aligned}$$

An informative prior, $\pi(\theta)$, was elicited from the literature about climate properties. According to this, each input variable has an independent prior with

$$\begin{aligned} \frac{\mathbf{S}}{6} &\sim \mathrm{Beta}(3.5, 6) \\ \frac{\mathbf{Kv}}{15} &\sim \mathrm{Beta}(2.85, 14) \\ \frac{\mathbf{Faer} + 1.5}{2} &\sim \mathrm{Beta}(4, 4) \end{aligned}$$

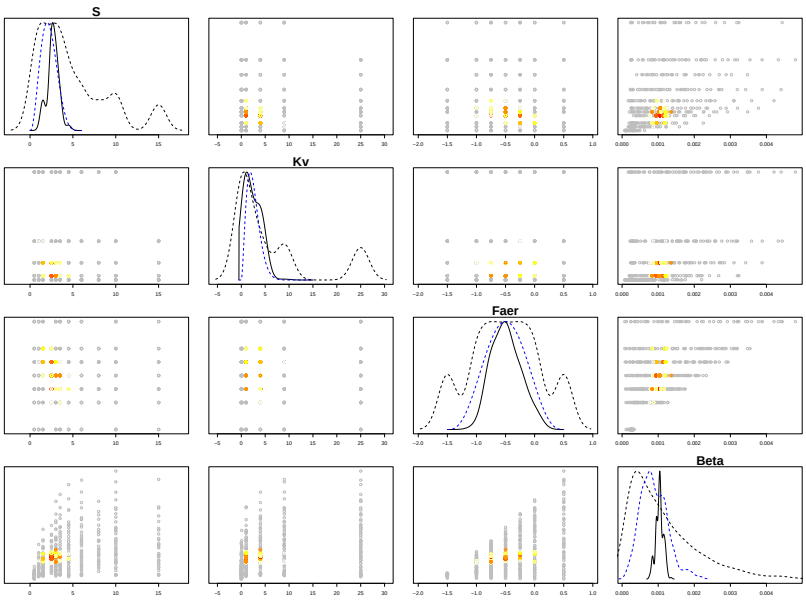

Figure 1: Resampling and Density Estimates under an informative prior. The solid line is the KDE of the posterior resample. The dotted blue line shows the prior and the dotted black line is the KDE of the original sample. The off-diagonal scatterplots show the original sample, and the colours indicate resampling frequency.

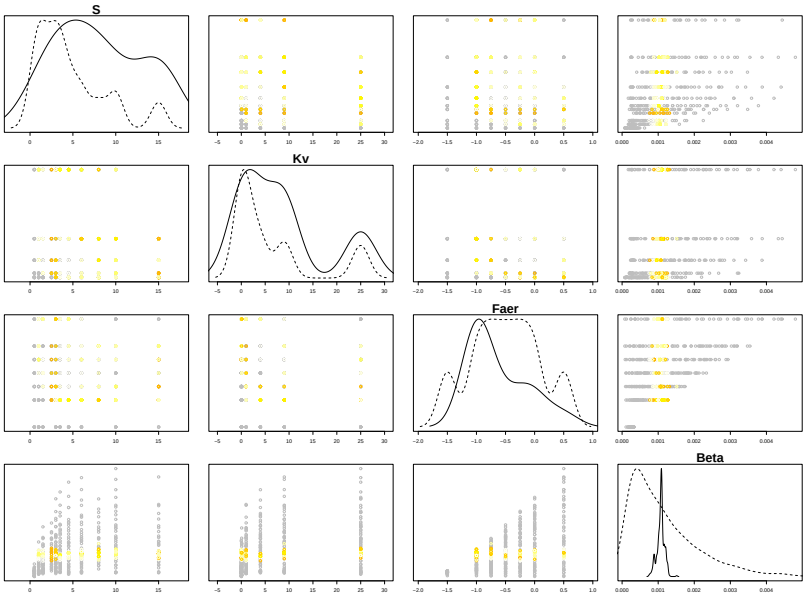

Figure 2: Resampling and Density Estimates with a non-informative prior. The solid line is the KDE of the posterior resample and the dotted black line is the KDE of the original sample. The off-diagonal scatterplots show the original sample, and the colours indicate resampling frequency.

Figure 4 illustrates the results of our analysis. Due to the discrete nature 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, $\pi(\theta_i) = \frac{1}{n}$. We see in Figure 4 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. The data would again indicate that **S** is likely near the higher end of our support, with or without prior information on the variables.

5 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 dataset 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. Here we make use of 1,000,000 simulations of permeability fields that were run through the program S3D developed by King and Datta-Gupta (1998). These fields were all generated from a Gaussian process with a Gaussian correlation structure and correlation parameters consistent with the expected geologic structure of the study site. Note that this is a pure inverse problem, in that no direct permeability measurements are available, so all inference must rely on the indirect information in the flow data.

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 dataset, 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

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

where the observed breakthrough times at the sampling wells are denoted by y_i, and the times produced by the computer simulator (when the 14x11 permeability field θ is input) are denoted by $\hat{y}_i(\theta)$. σ^2 is fixed to a value based on input from the geologists, as it is not possible to estimate it from the available data.

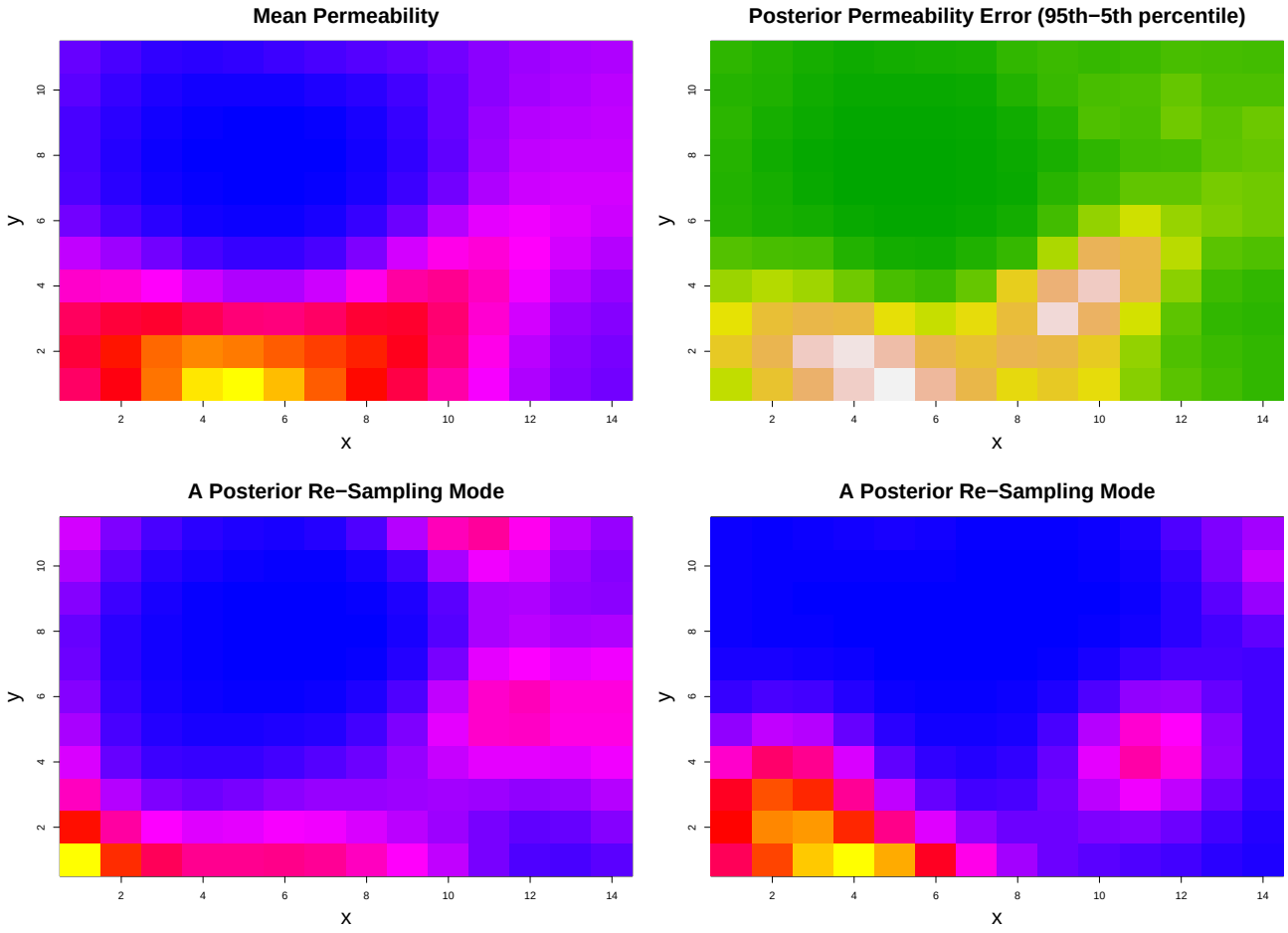

Figure 3: Posterior mean permeability field, variability estimate, and two of the highly weighted sample fields

Figure 5 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

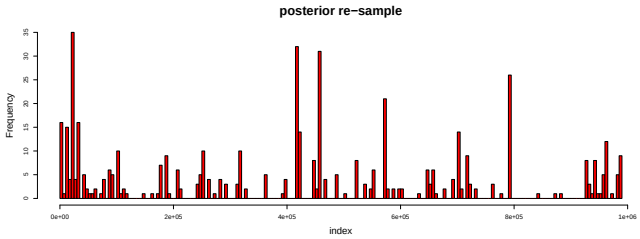

Figure 4: Resampling frequencies for the permeability fields. In a 500 point resample, 35 was the maximum resampling frequency.

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 5 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 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.

6 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, resampling for inverse problems provides a natural solution. Our method is also applicable when a very large bank of existing runs is available, such that conventional GP surrogate modeling is infeasible (the matrix inversions make using more than several thousand points difficult). Furthermore, the method can be employed retrospectively, when no additional simulator runs can be performed, preventing any use of MCMC.

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. Technical report, MIT. Submitted to GRL.

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., 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., 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.

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(1):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(3):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(452):1244–1255.

Rubin, D. (1988). Using the SIR algorithm to simulate posterior distributions by data augmentation. In Bernardo, J., DeGroot, M., and Lindley, D. amd Smith, A., editors, *Bayesian statistics 3*. 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. (2007). Statistical calibration of climate system properties. Technical Report ams2007-06, Applied Mathematics and Statistics, University of California Santa Cruz.

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*. PhD thesis, Texas A&M University, Department of Petroleum Engineering.