Bayesian Calibration of Computationally Expensive Models Using Optimization and Radial Basis Function Approximation

Nikolai Bliznyouk [*] David Ruppert [†] Christine Shoemaker [‡]
Rommel Regis [§] Stefan Wild [¶] Pradeep Mugunthan [‖]

September 29, 2006

Abstract

We present a Bayesian approach to model calibration when evaluation of the model is computationally expensive. Here, calibration is a nonlinear regression problem: given data vector $\boldsymbol{Y}$ corresponding to the regression model $\boldsymbol{f}(\boldsymbol{\beta})$, find plausible values of $\boldsymbol{\beta}$. As an intermediate step, $\boldsymbol{Y}$ and $\boldsymbol{f}$ are embedded into a statistical model allowing transformation and dependence. Typically, this problem is solved by sampling from the posterior distribution of $\boldsymbol{\beta}$ given $\boldsymbol{Y}$ using MCMC. To reduce computational cost, we limit evaluation of $\boldsymbol{f}$ to small number of points chosen on a high posterior density region found by optimization. Then, we approximate the log-posterior using radial basis functions and use the resulting cheap-to-evaluate surface in MCMC. We illustrate our approach on simulated data for a pollutant diffusion problem and study frequentist coverage properties of credible intervals. Our experiments indicate that our method can produce results similar to those when the true "expensive" posterior is sampled by MCMC while reducing computational costs by well over an order of magnitude.

[*] Nikolai Bliznyouk is a graduate student, School of Operations Research and Industrial Engineering, Cornell University, Rhodes Hall, Ithaca, NY 14853, USA. (E-mail: nab36@cornell.edu.)

[†] David Ruppert is Andrew Schultz Jr. Professor of Engineering and Professor of Statistical Science, School of Operations Research and Industrial Engineering, Cornell University, Rhodes Hall, Ithaca, NY 14853, USA. (E-mail: dr24@cornell.edu.)

[‡] Christine Shoemaker is Joseph P. Ripley Professor of Engineering, School of Civil and Environmental Engineering and School of Operations Research and Industrial Engineering, Cornell University, Hollister Hall, Ithaca, NY, 14853. (Email: cas12@cornell.edu.)

[§] Rommel Regis is a Postdoctoral Associate, Cornell Theory Center, Cornell University, Rhodes Hall, Ithaca, NY, 14853. (Email: rgr6@cornell.edu.)

[¶] Stefan Wild is a graduate student, School of Operations Research and Industrial Engineering, Cornell University, Rhodes Hall, Ithaca, NY 14853, USA. (E-mail: smw58@cornell.edu.)

[‖] Pradeep Mugunthan is a Ph.D. graduate from Civil and Environmental Engineering, Cornell University, Ithaca, NY, 14853.

Keywords: computer experiments; design of experiments; interpolation; inverse problems; Markov Chain Monte Carlo; RBF; transformation.

1 INTRODUCTION

A common problem throughout science and engineering is the calibration of scientific models. Calibration means estimation of unknown parameters, for example, initial conditions or reaction or diffusion rates in a system modeled by partial differential equations. In this paper, we implement a Bayesian strategy for calibration when the models are specified by computationally expensive computer codes.

Our focus is on computer codes that, in a single run, produce deterministic d-dimensional output vectors $f(X, \boldsymbol{\beta})$ for all vector "indices" X in some specified set and a given parameter vector $\boldsymbol{\beta}$. For example, the numerical solution of a partial differential equation produces output at all points on a space-time grid for a fixed vector of coefficients $\boldsymbol{\beta}$. We assume that one has observed a sample $Y_1, \ldots, Y_n$ of stochastic response vectors in $\mathbb{R}^d$ that correspond to the model values $f(X_1, \boldsymbol{\beta}), \ldots, f(X_n, \boldsymbol{\beta})$, and the goal is to make inference about $\boldsymbol{\beta}$. The vector X_i is assumed to be known to the experimenter and can thus be regarded as a label for the model $f(X_i; \boldsymbol{\beta})$ for Y_i. We are motivated by environmental engineering problems where Y_i's are vectors of observed concentrations of chemical species and X_i's are the temporal instants and spatial locations where the concentrations were measured, although our methodology is applicable to a wider range of problems. Evaluating $f(X_1, \boldsymbol{\beta}), \ldots, f(X_n, \boldsymbol{\beta})$ for a single value of $\boldsymbol{\beta}$ can be computationally expensive taking, for example, 2.5 hours of CPU time in a groundwater bioremediation problem studied by Mugunthan, Shoemaker and Regis (2005) and Mugunthan and Shoemaker (2006, in press). Thus, accurate calibration of such models is infeasible without special methods such as those introduced here.

Given that Y_i is $f(X_i, \boldsymbol{\beta}^{(0)})$ plus noise for value $\boldsymbol{\beta}^{(0)}$ of $\boldsymbol{\beta}$, calibration is seen to be a nonlinear regression problem. However, ordinary (nonlinear) least squares (OLS) is not recommended, since practitioners often find that the variation of Y_i about $f(X_i, \boldsymbol{\beta})$ is nonnormally distributed with nonconstant variance and correlated across time and space. We accommodate the nonnormality and heteroscedasticity by the transform-both-sides methodology of Carroll and Ruppert (1984) – we assume that, after a suitable transformation, Y_i's are normally distributed and homoscedastic. To model dependence of Y_i's, we use a parametric space-time covariance model.

Specifying the likelihood of the data $Y_1, \ldots, Y_n$ and putting priors on parameters, we obtain the expression for the unnormalized posterior. Our algorithm has four main steps: (1) use numerical optimization to locate the region of the parameter space having high posterior probability; (2) evaluate the model on a suitable set of parameter values in the region of high posterior probability; (3) use the evaluations in steps (1) and (2) to construct a radial basis function (RBF) approximation to the log-posterior; and (4) draw a sample from the approximate posterior using a Markov Chain Monte Carlo (MCMC) algorithm. As a result, computational burden is reduced considerably since step (4) does not require expensive function evaluations. Using the sample from the approximate posterior allows us to solve the problems of Bayesian calibration and of prediction for $F(\boldsymbol{\beta})$ by estimating moments and quantiles of the posteriors of $\boldsymbol{\beta}$ and $F(\boldsymbol{\beta})$. Here, F is a function whose computation, for a given $\boldsymbol{\beta}$, involves evaluation of $f(\cdot, \boldsymbol{\beta})$ for multiple values of X; more precisely, $F(\boldsymbol{\beta})$ is the value of a functional of $f(\cdot, \boldsymbol{\beta})$.

Empirical studies show that our algorithm can produce estimates of posterior densities for $\boldsymbol{\beta}$ and $F(\boldsymbol{\beta})$ that are nearly the same as when sampling from the exact posterior. However, our methodology requires far fewer evaluations of the expensive function than would be needed if the exact posterior were sampled.

This appears to be the first investigation that uses a direct nonparametric approximation to the posterior surface on the region of high posterior probability found by optimization. The introduction in Section 2.2 of a new transformation family, which is more attractive from the Bayesian perspective than the usual Box-Cox power family, allows a systematic treatment of data transformation, which is typically carried out in *ad hoc* fashion. As far as we are aware, this is also the first use of the transform-both-sides model with space-time correlation and should be of independent interest.

The outline above reflects the organization of the paper: Section 2 specifies the statistical model for the data, Section 3 deals with the approximation to the posterior and contains details of the algorithm, Section 4 reports the results of a simulation study of a synthetic diffusion problem, and Section 5 discusses alternative approaches to calibration.

2 DESCRIPTION OF THE MODEL

2.1 A General Statistical Model

We assume that Y_i is $f(X_i, \boldsymbol{\beta})$ perturbed by noise, which could include model misspecification and measurement error. In many applications, the components of Y_i show right-skewed variation about $f(X_i, \boldsymbol{\beta})$ with variability that increases with $f(X_i, \boldsymbol{\beta})$. The transform-both-sides (TBS) methodology of Carroll and Ruppert (1984, 1988) is particularly well suited for such data.

Denote by $Y_{i,j}$ and $f_j(X_i, \boldsymbol{\beta})$ the jth components of Y_i and $f(X_i, \boldsymbol{\beta})$, respectively. Let $\{h(\cdot, \lambda) : \lambda \in \Lambda\}$ be a parametric family of differentiable increasing transformations indexed by λ. We assume that, for some λ_j, $h(Y_{i,j}, \lambda_j)$ is distributed $N\left[h\{f_j(X_i, \boldsymbol{\beta}), \lambda_j\}, \sigma_j^2\right]$, where σ_j is constant as a function of X_i. Stated differently, $h(\cdot, \lambda_j)$ is both a normalizing and variance-stabilizing transformation for $Y_{i,j}$. In addition, we require that the Y_i's can be transformed to have a joint multivariate normal (MVN) distribution. In Section 2.2 we describe a new transformation family that we have used in our work.

It is important to notice that both $Y_{i,j}$ and $f_j(X_i, \boldsymbol{\beta})$ are transformed in the same way. This implies that $f_j(X_i, \boldsymbol{\beta})$ is the conditional median of $Y_{i,j}$ given X_i, so, unlike when $Y_{i,j}$ alone is transformed as in Box and Cox (1964), $f_j(X_i, \boldsymbol{\beta})$ continues to be a model for $Y_{i,j}$. In fact, the model for $Y_{i,j}$ is

$$Y_{i,j} = h^{-1}\left[h\left\{f_j(X_i, \boldsymbol{\beta}), \lambda_j\right\} + \epsilon_{i,j}, \lambda_j\right],$$

where, for a fixed λ, $h^{-1}(\cdot, \lambda)$ is the inverse of $h(\cdot, \lambda)$, and $\epsilon_{i,j} \sim N(0, \sigma_j^2)$. For example, if $h(\cdot, \lambda_j)$ is the log transformation, then

$$Y_{i,j} = \exp\left[\log\{f_j(X_i, \boldsymbol{\beta})\} + \epsilon_{i,j}\right] = f_j(X_i, \boldsymbol{\beta}) \exp(\epsilon_{i,j}),$$

so the model has multiplicative, lognormal variation about the conditional median, $f_j(X_i, \boldsymbol{\beta})$.

Let $\boldsymbol{Y} = \left[Y_1^{\mathsf{T}}, \ldots, Y_n^{\mathsf{T}}\right]^{\mathsf{T}}$ be the nd-dimensional column vector of observed responses and $\boldsymbol{f}(\boldsymbol{\beta}) = \left[f(X_1, \boldsymbol{\beta})^{\mathsf{T}}, \ldots, f(X_n, \boldsymbol{\beta})^{\mathsf{T}}\right]^{\mathsf{T}}$ be the corresponding value of the regression function. Define $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_d)^{\mathsf{T}}$, and denote by $h\{\boldsymbol{Y}, \boldsymbol{\lambda}\}$ and $h\{\boldsymbol{f}(\boldsymbol{\beta}), \boldsymbol{\lambda}\}$ the coordinate-wise transformations of $\boldsymbol{Y}$ and $\boldsymbol{f}(\boldsymbol{\beta})$, where every coordinate corresponding to the jth outcome is transformed by $h(\cdot, \lambda_j)$. Our statistical model is then $h\{\boldsymbol{Y}, \boldsymbol{\lambda}\} \sim MVN\left[h\{\boldsymbol{f}(\boldsymbol{\beta}), \boldsymbol{\lambda}\}, \boldsymbol{\Sigma}(\boldsymbol{\theta})\right]$

with the corresponding likelihood function

$$[\boldsymbol{Y}|\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\theta}] = \frac{\exp\left(-0.5\|h(\boldsymbol{Y},\boldsymbol{\lambda}) - h\{\boldsymbol{f}(\boldsymbol{\beta}),\boldsymbol{\lambda}\}\|^2_{\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}}\right)}{(2\pi)^{nd/2}|\boldsymbol{\Sigma}(\boldsymbol{\theta})|^{1/2}} \cdot |J_h(\boldsymbol{Y},\boldsymbol{\lambda})|, \tag{1}$$

where $J_h(\boldsymbol{Y},\boldsymbol{\lambda})$ is the Jacobian of the transformation from $\boldsymbol{Y}$ to $h\{\boldsymbol{Y},\boldsymbol{\lambda}\}$ and $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ belongs to a family of covariance matrices parameterized by $\boldsymbol{\theta}$, as described in Section 2.3. Here we use the now standard notation that [*list*] is the joint density of the random variables in *list* and [*list 1*|*list 2*] is the conditional density of the random variables in *list 1* given those in *list 2*. We also use the conventional notation for the generalized norm, $\|x\|^2_{\boldsymbol{A}} = x^\top \boldsymbol{A} x$.

When the goal of the study is the posterior distribution of the value $F(\boldsymbol{\beta})$ of some functional of $f(\cdot,\boldsymbol{\beta})$, as well as the posterior for $\boldsymbol{\beta}$, it may be necessary to evaluate the expensive function for additional space-time indices $X_{n+1},\ldots,X_{n^*}$, for which no response Y_i was observed, in order to compute or approximate $F(\boldsymbol{\beta})$; further details are found in Section 3.3. We assume that, for a single value of $\boldsymbol{\beta}$, a run of the expensive model produces the entire vector $\boldsymbol{f}^*(\boldsymbol{\beta}) = \left[\boldsymbol{f}(\boldsymbol{\beta})^\top, f(X_{n+1},\boldsymbol{\beta})^\top,\ldots,f(X_{n^*},\boldsymbol{\beta})^\top\right]^\top$. This leads to computational savings, for example, in models relying on numerical meshes or grids, for which it is often more beneficial to obtain values of $f(X_i,\boldsymbol{\beta})$ for all values X_i of interest in a single step rather than to compute them in two stages (i.e., first for $\boldsymbol{f}(\boldsymbol{\beta})$ and then for $f(X_i,\boldsymbol{\beta})$ for all $i > n$). Once $\boldsymbol{f}$ has been evaluated at $\boldsymbol{\beta}^{(0)}$, $[\boldsymbol{Y}|\boldsymbol{\beta}^{(0)},\boldsymbol{\lambda},\boldsymbol{\theta}]$ is usually a "cheap" function of $(\boldsymbol{\lambda},\boldsymbol{\theta})$, unless n is very large.

2.2 A New Transformation Family

The usual transformation family used with the TBS method is the Box-Cox family where $h_{BC}(y,\lambda)$ is $(y^\lambda - 1)/\lambda$ if $\lambda \neq 0$ and is $\log(y) = \lim_{\lambda\to 0}(y^\lambda - 1)/\lambda$ if $\lambda = 0$. We assume that both the Y_i's and $f(X,\boldsymbol{\beta})$ are positive so the transformation should map the positive real number to the entire real line; otherwise, the transformation is to a truncated normal distribution. The only transformation in the Box-Cox family mapping $(0,\infty)$ to $(-\infty,\infty)$ is the log transformation. Consequently, if one uses the Box-Cox family, one is assuming implicitly that the transformation is to a truncated normal distribution. (For discussion and examples, see Gelman, Carlin, Stern and Rubin (2004) and Thyer, Kuczera and Wang (2002).) Typically, practitioners ignore this fine point and approximate the truncated normal by a normal distribution. This results in an invalid statistical model since with positive probability the noise term $\epsilon_{i,j}$ added to $h_{BC}\{f_j(X_i, \boldsymbol{\beta}), \lambda_j\}$ produces values for which the inverse transformation $h_{BC}^{-1}(\cdot, \lambda_j)$ is not defined. If one works with the truncated normal distribution, then the normalizing constant must be found by computing an nd-dimensional integral, which is likely to be difficult, if feasible, for all but the simplest models. Perhaps more importantly, the interpretation of $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ as a covariance matrix is lost when working with truncated distributions.

It seems better to find a simple transformation family such that every transformation in it maps $(0, \infty)$ to $(-\infty, \infty)$. We propose the *COnvex combination of Identity and Log (COIL)* family

$$h_C(y, \lambda) = \lambda y + (1 - \lambda) \log(y), \quad 0 \leq \lambda \leq 1. \tag{2}$$

When λ is 0 or 1, then $h_C(\cdot, \lambda) = h_{BC}(\cdot, \lambda)$. We will restrict λ to $[0, 1)$, since $h_C(\cdot, 1)$ does not map $(0, \infty)$ to $(-\infty, \infty)$. Our simulation experiments show that the entire Box-Cox family for $\lambda \in [0, 1)$ can be approximated well by our family; specifically, for each $\lambda \in [0, 1)$ there are constants $\lambda' \in [0, 1)$ and $a, b \in \mathbb{R}$ such that $h_{BC}(y, \lambda) \approx a + b h_C(y, \lambda')$ for a wide range of y values. The constants a and b do not play a role when the COIL transformation is combined with TBS since a is canceled out and b is absorbed into the variance parameter for identifiability.

The inverse transformation $h_C^{-1}(\cdot, \lambda)$ does not have a closed form but can be evaluated easily by interpolation.

Generalization of the COIL family to include more severe concave transformations by taking convex combinations of, for example, the log transformation, the identity transformation, and $h_{BC}(\cdot, -1)$, is possible but is not pursued here.

2.3 A Model for the Covariance

Define the noise vectors $\epsilon_i = (\epsilon_{i,1}, \ldots, \epsilon_{i,d})^\mathsf{T} = h\{Y_i, \boldsymbol{\lambda}\} - h\{f(X_i, \boldsymbol{\beta}), \boldsymbol{\lambda}\}$ for $i = 1, \ldots, n$, $\epsilon_{\bullet,j} = (\epsilon_{1,j}, \ldots, \epsilon_{n,j})^\mathsf{T}$, and $\boldsymbol{\epsilon} = (\epsilon_1^\mathsf{T}, \ldots, \epsilon_n^\mathsf{T})^\mathsf{T}$. The covariance between $\epsilon_{i,j}$ and $\epsilon_{i',j'}$ is modeled parsimoniously using a separable covariance function of the form $\boldsymbol{C}_{j,j'} \cdot \rho_{ST}(X_i, X_{i'}; \boldsymbol{\gamma})$, where $\boldsymbol{C}$ is a $d \times d$ covariance matrix for ϵ_i and $\rho_{ST}(X_i, X_{i'}; \boldsymbol{\gamma})$ is a space-time correlation function parameterized by $\boldsymbol{\gamma}$. One could consider nonseparable models if there were enough data, but in many applications a parsimonious model is highly desirable. Under some regularity conditions, misspecification of the covariance model will cause some inefficiency, but not bias, in the estimation of $\boldsymbol{\beta}$. For this reason we favor a parsimonious model over a more complex but more realistic model. In many if not most cases, ρ_{ST} will also be separable, that is, the product of spatial and temporal correlation functions.

Let $\boldsymbol{S}(\boldsymbol{\gamma})$ be the $n \times n$ space-time correlation matrix with $\boldsymbol{S}_{i,i'}(\boldsymbol{\gamma}) = \rho_{ST}(X_i, X_{i'}; \boldsymbol{\gamma})$. Then $\text{Var}\{\boldsymbol{\epsilon}_{\bullet,j}\} = \boldsymbol{C}_{j,j} \cdot \boldsymbol{S}(\boldsymbol{\gamma})$ and, more generally, $\text{Var}\{\boldsymbol{\epsilon}\} = \boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{S}(\boldsymbol{\gamma}) \otimes \boldsymbol{C}$, where $\boldsymbol{\theta} = (\boldsymbol{\gamma}, \boldsymbol{C})$ and $\otimes$ denotes the Kronecker product of two matrices.

3 METHODOLOGY

3.1 An Approximation to the Posterior

Given a prior $[\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\theta}]$, one has a posterior

$$[\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\theta} | \boldsymbol{Y}] = \frac{[\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\theta}, \boldsymbol{Y}]}{\int [\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\theta}, \boldsymbol{Y}]\, d\boldsymbol{\beta}\, d\boldsymbol{\lambda}\, d\boldsymbol{\theta}}, \tag{3}$$

where $[\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\theta}, \boldsymbol{Y}] = [\boldsymbol{Y} | \boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\theta}] \cdot [\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\theta}]$ is the joint density of the data and the parameters. (Recall the bracket notation for densities that was introduced in Section 2.1 following Equation (1).)

In applications, $\boldsymbol{\beta}$ is the primary parameter and interest centers on its marginal posterior $[\boldsymbol{\beta} | \boldsymbol{Y}] = \int [\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\theta} | \boldsymbol{Y}]\, d\boldsymbol{\lambda}\, d\boldsymbol{\theta}$. Even though it may be infeasible to integrate out all of the nuisance parameters $(\boldsymbol{\lambda}, \boldsymbol{\theta})$, it is often possible to integrate out some of them either analytically, by using conjugate priors as shown in Appendix A.4, or numerically. Let $\boldsymbol{\zeta}$ be the subvector of $(\boldsymbol{\lambda}, \boldsymbol{\theta})$ of remaining nuisance parameters after the integration. (This notation allows $\boldsymbol{\zeta}$ to be $(\boldsymbol{\lambda}, \boldsymbol{\theta})$ or be empty.) In what follows, $\boldsymbol{Y}$ is always regarded as fixed and $[\boldsymbol{\beta}, \boldsymbol{Y}]$ and $[\boldsymbol{\beta}, \boldsymbol{\zeta}, \boldsymbol{Y}]$ refer, respectively, to arbitrary unnormalized marginal and joint posteriors.

If $\boldsymbol{f}$ were inexpensive to evaluate, then we could sample from $[\boldsymbol{\beta}, \boldsymbol{\zeta} | \boldsymbol{Y}]$ using $[\boldsymbol{\beta}, \boldsymbol{\zeta}, \boldsymbol{Y}]$ with a Metropolis-Hastings (M-H) algorithm, and then the sample of $\boldsymbol{\beta}$ would be a sample from $[\boldsymbol{\beta} | \boldsymbol{Y}]$. However, drawing large samples is computationally prohibitive in our setting.

Our goal is to obtain an accurate and cheap-to-evaluate nonparametric approximation to $[\boldsymbol{\beta}, \boldsymbol{Y}]$ or $[\boldsymbol{\beta}, \boldsymbol{\zeta}, \boldsymbol{Y}]$ based on a relatively small number of evaluations of $\boldsymbol{f}$ in order to use the resulting surface as a surrogate for the respective unnormalized posterior in a M-H algorithm, as the sampler does not require specification of normalizing constants. When the expression for $[\boldsymbol{\beta}, \boldsymbol{Y}]$ is not available, we approximate $[\boldsymbol{\beta}, \boldsymbol{Y}]$ heuristically first. For a fixed value of $\boldsymbol{\beta}$, let $\widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta})$ be the maximizer of $[\boldsymbol{\beta}, \boldsymbol{\zeta}, \boldsymbol{Y}]$ with respect to $\boldsymbol{\zeta}$. One possible approximation is the *profile* posterior

$$\pi_{\max}(\boldsymbol{\beta}, \boldsymbol{Y}) = \sup_{\boldsymbol{\zeta}} [\boldsymbol{\beta}, \boldsymbol{\zeta}, \boldsymbol{Y}] = [\boldsymbol{\beta}, \widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}), \boldsymbol{Y}]. \tag{4}$$

A more sophisticated *Laplace* approximation of Tierney and Kadane (1986) multiplies (4) by a correction factor. A simplification to (4), referred to as *pseudoposterior*, is obtained by replacing $\widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta})$ by $\widehat{\boldsymbol{\zeta}}(\widehat{\boldsymbol{\beta}})$, where $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\zeta}}(\widehat{\boldsymbol{\beta}}))$ is the maximum a posteriori (MAP) estimator, the mode of the joint posterior $[\boldsymbol{\beta}, \boldsymbol{\zeta}|\boldsymbol{Y}]$. All these approximations to $[\boldsymbol{\beta}, \boldsymbol{Y}]$ attempt to avoid the more difficult task of integrating out nuisance parameters by maximization with respect to them. It should be kept in mind, though, that neither the integration nor the maximization requires extra evaluations of $\boldsymbol{f}$. In the sequel, the notation $\pi(\cdot, \boldsymbol{Y})$ will be used to refer to any of these heuristic approximations to $[\boldsymbol{\beta}, \boldsymbol{Y}]$.

As a nonparametric approximation, we use interpolation of the logarithms of $\pi(\cdot, \boldsymbol{Y})$ or $[\boldsymbol{\beta}, \boldsymbol{\zeta}, \boldsymbol{Y}]$ by radial basis functions (RBFs). We state our algorithm for $\pi(\cdot, \boldsymbol{Y})$ as the surface of interest. The treatment of $[\boldsymbol{\beta}, \boldsymbol{\zeta}, \boldsymbol{Y}]$ is similar.

3.2 The Algorithm

In our algorithm, $\boldsymbol{f}^*$ is evaluated only during the optimization stage in order to find the MAP estimate (**Step 1**) and for values of $\boldsymbol{\beta}$ in a high posterior density region (**Step 2**) in order to approximate the logarithm of the posterior surface accurately by an RBF surface (**Step 3**). The approximate posterior surface is subsequently sampled using MCMC in **Step 4**.

For ease of exposition, we assume that the posterior has a single mode located in the interior of the parameter space and that $\boldsymbol{f}$ is twice differentiable in a neighborhood $\widehat{\boldsymbol{\beta}}$, but we are currently generalizing the approach to multimodal posteriors. While selecting "design points" (the values of $\boldsymbol{\beta}$ at which to evaluate $\boldsymbol{f}^*$) we try to keep as small as possible the number of "uninformative" points – those very close to some point, at which the value of $\boldsymbol{f}^*$ is known, or far away from the mode.

3.2.1 Finding the MAP (Step 1)

For a given value $\boldsymbol{\beta}^{(0)}$ of $\boldsymbol{\beta}$, the gradient and Hessian of $\log\{[\boldsymbol{\beta}^{(0)}, \boldsymbol{\zeta}, \boldsymbol{Y}]\}$ with respect to $\boldsymbol{\zeta}$ are available analytically, and so this function can be maximized efficiently to produce $\widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}^{(0)})$ and thus to compute $\log\{\pi_{\max}(\boldsymbol{\beta}^{(0)}, \boldsymbol{Y})\}$ from Equation (4). We perform this maximization using a constrained minimization routine (sequential quadratic programming) with analytical gradients and Hessians, implemented in MATLAB in `fmincon`. Consequently, we maximize $\log\{\pi_{\max}(\boldsymbol{\beta}, \boldsymbol{Y})\}$ with respect to $\boldsymbol{\beta}$ to find $\widehat{\boldsymbol{\beta}}$ and then $\log\{[\widehat{\boldsymbol{\beta}}, \boldsymbol{\zeta}, \boldsymbol{Y}]\}$ with respect to $\boldsymbol{\zeta}$ to determine $\widehat{\boldsymbol{\zeta}}(\widehat{\boldsymbol{\beta}})$ and hence the MAP. The use of a gradient-based algorithm for maximization with respect to $\boldsymbol{\beta}$ is not recommended unless the Jacobian of $\boldsymbol{f}$ comes at low cost along with $\boldsymbol{f}$ because finite differencing to estimate derivatives produces clusters of "uninformative" design points. We maximize $\log\{[\widehat{\boldsymbol{\beta}}, \boldsymbol{\zeta}, \boldsymbol{Y}]\}$ using publicly available software `CONDOR` described by Vanden Berghen and Bersini (2005), which implements a derivative-free trust-region algorithm UOBYQA of Powell (2000).

3.2.2 The Experimental Design (Step 2)

Ideally, we would like to fit the RBF surface over a highest posterior density (HPD) region of $[\boldsymbol{\beta}|\boldsymbol{Y}]$, defined as $C_R(\alpha) = \{\boldsymbol{\beta} : [\boldsymbol{\beta}, \boldsymbol{Y}] > \kappa(\alpha)\}$, where $\kappa(\alpha)$ is chosen so that the credible region $C_R(\alpha)$ contains the fraction $1-\alpha$ of the mass of $[\boldsymbol{\beta}, \boldsymbol{Y}]$. Here α is a tuning parameter, for example, .05 or .01.

The size $(1-\alpha)$ HPD region cannot be computed accurately – not only for $[\boldsymbol{\beta}, \boldsymbol{Y}]$, but also for $[\boldsymbol{\beta}, \boldsymbol{\zeta}, \boldsymbol{Y}]$ and for any of the heuristic approximations to $[\boldsymbol{\beta}, \boldsymbol{Y}]$ from the previous section – without a prohibitive number of evaluations of $\boldsymbol{f}$. We obtain an approximate HPD region, $\widehat{C}_R(\alpha)$, using a Taylor expansion of $\log\{[\boldsymbol{\beta}, \boldsymbol{\zeta}, \boldsymbol{Y}]\}$ near the MAP $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\zeta}})$, which corresponds to the approximation to $[\boldsymbol{\beta}, \boldsymbol{\zeta}|\boldsymbol{Y}]$ by a multivariate normal density. Specifically, let $\widehat{\boldsymbol{I}}$ be negative of the Hessian of $\log\{[\boldsymbol{\beta}, \boldsymbol{\zeta}, \boldsymbol{Y}]\}$ with respect to $(\boldsymbol{\beta}, \boldsymbol{\zeta})$ evaluated at $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\zeta}})$. By partitioning $\widehat{\boldsymbol{I}}^{-1}$ into blocks corresponding to $\boldsymbol{\beta}$ and $\boldsymbol{\zeta}$ one gets

$$\begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\zeta} \end{bmatrix} \overset{approx.}{\sim} MVN\left(\begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{\zeta}} \end{bmatrix}, \begin{bmatrix} \widehat{\boldsymbol{I}}^{\beta\beta} & \widehat{\boldsymbol{I}}^{\beta\zeta} \\ \widehat{\boldsymbol{I}}^{\zeta\beta} & \widehat{\boldsymbol{I}}^{\zeta\zeta} \end{bmatrix} \right), \text{ where} \tag{5}$$

$$\widehat{\boldsymbol{I}}^{-1} = \begin{bmatrix} \widehat{\boldsymbol{I}}_{\beta\beta} & \widehat{\boldsymbol{I}}_{\beta\zeta} \\ \widehat{\boldsymbol{I}}_{\zeta\beta} & \widehat{\boldsymbol{I}}_{\zeta\zeta} \end{bmatrix}^{-1} = \begin{bmatrix} \widehat{\boldsymbol{I}}^{\beta\beta} & \widehat{\boldsymbol{I}}^{\beta\zeta} \\ \widehat{\boldsymbol{I}}^{\zeta\beta} & \widehat{\boldsymbol{I}}^{\zeta\zeta} \end{bmatrix}. \tag{6}$$

Estimation of $\widehat{\boldsymbol{I}}$ by finite differences is wasteful as it does not produce new informative design points. We estimate $\widehat{\boldsymbol{I}}$ by fitting a quadratic surface to $\log\{[\boldsymbol{\beta}, \boldsymbol{\zeta}, \boldsymbol{Y}]\}$ in a neighborhood of $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\zeta}})$. This procedure, stated in detail in Appendix A.1, allows one to reduce the number of wasteful design points and to reuse the points from the optimization trajectory from **Step 1**. To avoid new notation, from now on we use the old notation for the true $\widehat{\boldsymbol{I}}$ and its blocks from Equation (6) to refer solely to the estimated Hessian and its blocks.

We define

$$\widehat{C}_R(\alpha) = \left\{ \boldsymbol{\beta} : (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})^\mathsf{T} \left[\widehat{\boldsymbol{I}}^{\boldsymbol{\beta}\boldsymbol{\beta}} \right]^{-1} (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) \leq \chi^2_{p,1-\alpha} \right\}, \tag{7}$$

where $\widehat{\boldsymbol{I}}^{\boldsymbol{\beta}\boldsymbol{\beta}} = \left[\widehat{\boldsymbol{I}}_{\boldsymbol{\beta}\boldsymbol{\beta}} - \widehat{\boldsymbol{I}}_{\boldsymbol{\beta}\boldsymbol{\zeta}} \cdot \widehat{\boldsymbol{I}}^{-1}_{\boldsymbol{\zeta}\boldsymbol{\zeta}} \cdot \widehat{\boldsymbol{I}}_{\boldsymbol{\zeta}\boldsymbol{\beta}} \right]^{-1}$ and $\chi^2_{p,1-\alpha}$ is the $(1-\alpha)$th quantile of the χ^2_p distribution, with p being the dimension of $\boldsymbol{\beta}$. This approximate HPD region is the size-$(1-\alpha)$ minimum volume confidence ellipsoid for the (marginal) normal approximation to $[\boldsymbol{\beta}|\boldsymbol{Y}]$ based on Equation (5). We will use evaluations of $\boldsymbol{f}$ on this region at the same set of values of $\boldsymbol{\beta}$ to fit RBF surfaces to any of the posterior surfaces from Section 3.1, possibly all of them. This substep of determining an approximate design region is referred to as as **Step 2A**.

We remark that $\widehat{\boldsymbol{I}}$ is crucial for subsequent analysis. If $\boldsymbol{H}$ is any square full-rank matrix such that $\widehat{\boldsymbol{I}}^{\boldsymbol{\beta}\boldsymbol{\beta}} = \boldsymbol{H}\boldsymbol{H}^\mathsf{T}$, for example, a Cholesky factor of $\widehat{\boldsymbol{I}}^{\boldsymbol{\beta}\boldsymbol{\beta}}$, then we apply the linear transformation $\boldsymbol{H}^{-1}$ to $\boldsymbol{\beta}$ to ensure the same scale and to reduce correlation in parameters. Extra design points are chosen with respect to maximum separation criteria on this transformed space. We fit our RBF surface on the transformed space as well, but choose not to introduce new notation to emphasize this. Finally, $\widehat{\boldsymbol{I}}$ is also used in the MCMC stage to define one of the scale parameters of the proposal distribution.

Let $\mathcal{B}_O$ and $\mathcal{B}_H$ be sets of values of $\boldsymbol{\beta}$ at which $\boldsymbol{f}^*$ is evaluated during optimization in **Step 1** and during estimation of $\widehat{\boldsymbol{I}}$ in **Step 2A**, respectively. In general, points in $\mathcal{B}_O \cup \mathcal{B}_H$ do not cover $\widehat{C}_R(\alpha)$ adequately to enable us to approximate the chosen posterior surface accurately over the whole approximate HPD region. We augment these points with a *space-filling* experimental design $\mathcal{B}_E$. Specifically, we require that points in $\mathcal{B}_E$ be well-separated and do not lie close to those in $\mathcal{B}_O \cup \mathcal{B}_H$, with between-point distances measured after the mentioned linear transformation. Details are provided in Appendix A.2. We refer to the step of choosing $\mathcal{B}_E$ by **Step 2B**.

Finally, we let $\mathcal{B}_D = (\mathcal{B}_O \cup \mathcal{B}_H \cup \mathcal{B}_E) \cap \widehat{C}_R(\alpha')$ for $\alpha' \leq \alpha$ and define $N = |\mathcal{B}_D|$, the size of $\mathcal{B}_D$. The points in $\mathcal{B}_D$ will be used to build the RBF approximation. The motivation is that the optimization trajectory points $\mathcal{B}_O$ lying far outside of $\widehat{C}_R(\alpha)$ rarely improve the quality of approximation, but can be a cause of numerical problems. We typically use $\alpha \leq .1$ and $\alpha' = .01$ or $.005$ in practice.

3.2.3 The RBF Approximation (Step 3)

We use radial basis functions (Buhmann 2003, Powell 1992) to approximate the logarithm of the posterior surface by an interpolant of $l(\cdot) = \log\{\pi(\cdot, \boldsymbol{Y})\}$ at the design points $\boldsymbol{\mathcal{B}}_D = \{\boldsymbol{\beta}^{(1)}, \dots, \boldsymbol{\beta}^{(N)}\}$ of the form

$$\widetilde{l}(\boldsymbol{\beta}) = \sum_{i=1}^{N} a_i \phi(\|\boldsymbol{\beta} - \boldsymbol{\beta}^{(i)}\|_2) + q(\boldsymbol{\beta}), \tag{8}$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$, $a_1, \dots, a_N \in \mathbb{R}$, $q \in \Pi_m^p$ (the space of polynomials in $\mathbb{R}^p$ of degree less than or equal to m), $\|\cdot\|_2$ denotes the Euclidean norm, and the basis function ϕ has one of the following forms: (1) *surface spline*: $\phi(r) = r^\kappa$, $\kappa \in \mathbb{N}$, κ odd, or $\phi(r) = r^\kappa \log r$, $\kappa \in \mathbb{N}$, κ even; (2) *multiquadric*: $\phi(r) = (r^2 + \gamma^2)^\kappa$, $\kappa > 0$, $\kappa \notin \mathbb{N}$; (3) *inverse multiquadric*: $\phi(r) = (r^2 + \gamma^2)^\kappa$, $\kappa < 0$; (4) *Gaussian*: $\phi(r) = \exp(-\gamma r^2)$; where $r \geq 0$ and γ is a positive constant. The purpose of the polynomial tail is to ensure that the interpolation matrix is invertible. In the numerical experiments, we use the cubic form $\phi(r) = r^3$ with a linear tail $q(\boldsymbol{\beta}) = (1, \boldsymbol{\beta}^\mathsf{T}) \cdot \boldsymbol{c}$. Our choice of RBF interpolation over alternatives is motivated by the weakness of assumptions on the geometry of the interpolation points as compared, for instance, with those for multivariate polynomial interpolation (see Buhmann (2003, chap. 3)). This allows a greater re-use of the scattered design points from the optimization run. Another attractive feature is the ease of implementation of the RBF model; details are provided in Appendix A.3. Regis and Shoemaker (2006, in press) use RBF interpolation in global optimization of deterministic functions for the same reasons.

3.2.4 MCMC Sampling (Step 4)

In **Step 4** of the algorithm we draw an MCMC sample from the density proportional to $\widetilde{\pi}(\cdot, \boldsymbol{Y}) = \exp\{\widetilde{l}(\cdot)\}$ restricted to the approximate HPD region $\widehat{C}_R(\alpha')$, see the end of Section 3.2.2 and Equation (8). This is done to prevent sampling $\widetilde{\pi}(\cdot, \boldsymbol{Y})$ in the low-probability regions of $[\boldsymbol{\beta}|\boldsymbol{Y}]$ where $\pi(\cdot, \boldsymbol{Y})$ is not approximated sufficiently well. We use the autoregressive Metropolis-Hastings (ARMH) algorithm of Tierney (1994) which does not require the normalizing constant of $\widetilde{\pi}(\cdot, \boldsymbol{Y})$ to be known. The ARMH sampler uses a (vector) $AR(1)$ process to generate candidate points $\boldsymbol{\beta}^c$ given the current state $\boldsymbol{\beta}^{(t)}$ of the chain, *i.e.*, $\boldsymbol{\beta}^c = \boldsymbol{\mu} + \boldsymbol{\rho}(\boldsymbol{\beta}^{(t)} - \boldsymbol{\mu}) + \boldsymbol{e}_t$, where $\boldsymbol{\mu}$ is the location parameter, $\boldsymbol{\rho}$ is the autoregressive parameter (matrix), and $\boldsymbol{e}_t$'s are *i.i.d.* noise vectors from a density g. The algorithm allows much freedom in tuning its performance and includes the popular random walk M-H (when $\boldsymbol{\rho} = 1$) and the independence M-H (when $\boldsymbol{\rho} = 0$) algorithms as special cases. In our experiments, g is taken to be a finite mixture of multivariate normal and Student's t-densities centered at zero with dispersion matrices proportional to $\widehat{\boldsymbol{I}}^{\boldsymbol{\beta\beta}}$. The location parameter $\boldsymbol{\mu}$ is set to the MAP $\widehat{\boldsymbol{\beta}}$. We observed that negative values of $\boldsymbol{\rho}$ help to reduce serial correlation in the Markov chain. To improve mixing, we recommend that the tuning parameters for the sampler be calibrated to a particular application individually by conventional methods reviewed, for example, in Gelman et al. (2004), as at this stage MCMC does not require evaluations of $\boldsymbol{f}$.

3.3 Bayesian Inference

Once the MCMC sample $\mathcal{B}_M$ from the approximate posterior is obtained as discussed in Section 3.2.4, inference about $\boldsymbol{\beta}$ can proceed by standard methods. A problem of particular concern in environmental engineering is estimation of the value $F(\boldsymbol{\beta})$ of some functional of $f(\cdot, \boldsymbol{\beta})$, for example, $f(X, \boldsymbol{\beta})$ itself at values of X whose time coordinate is in the future. In this case, the set $\{F(\boldsymbol{\beta}) : \boldsymbol{\beta} \in \mathcal{B}_M\}$ is a sample from the approximate posterior of $F(\boldsymbol{\beta})$.

Since $F(\boldsymbol{\beta})$ is determined by $f(\cdot, \boldsymbol{\beta})$, it is also computationally expensive, and hence approximation is necessary to evaluate it at the points from the MCMC run. However, assuming as in Section 2 that $F(\boldsymbol{\beta})$ is a function (or can be approximated by a function) of components of $\boldsymbol{f}^*(\boldsymbol{\beta})$, it may be sufficient to compute its values only on the approximate HPD region for $\boldsymbol{\beta}$. Since we have already evaluated $\boldsymbol{f}^*$ at the design points in $\mathcal{B}_D$, it is cheap computationally to interpolate F (or an approximation to it) at the points in $\mathcal{B}_D$ and to evaluate the resulting interpolant at the points in $\mathcal{B}_M$. This approximate sample from the posterior of $F(\boldsymbol{\beta})$ can be subsequently used to estimate functionals of the posterior of $F(\boldsymbol{\beta})$.

4 AN ENVIRONMENTAL APPLICATION

In this section, we consider calibration of an environmental model for the concentrations of pollutants and illustrate our methodology on a synthetic test problem. A test problem was chosen such that $f(X, \boldsymbol{\beta})$ is given in closed form and can be evaluated inexpensively. Unlike with the expensive model functions used in many applications, this allows us to carry out an extensive Monte Carlo study comparing the coverage properties of the approximate Bayesian credible intervals based on RBF surfaces that require a relatively small number of evaluations of $\boldsymbol{f}^*$ with those of the exact credible intervals that require thousands of evaluations of the expensive exact posterior.

Examples of methods designed for calibration and uncertainty analysis of computationally expensive environmental models are Mugunthan et al. (2005) and Mugunthan and Shoemaker (2006, in press), respectively. The methods in both of these papers are applied to a remediation problem at US-DOD site that has been contaminated with chlorinated ethenes in the soil and groundwater. The simulation model there takes 2.5 hours to run. Neither of these earlier methods base analysis on the joint posterior density of the parameters as is done in this paper.

4.1 Environmental Assessment of a Chemical Spill: Formulation

Consider a chemical accident that has caused a pollutant to spill at two locations into a long and narrow holding channel. Assume it is known that the same mass M was spilled at each location (0 and L) and that the vector of the location and time of the first spill is $(0, 0)$. However, the location L and time τ of the second spill are unknown as is the value of M and the diffusion rate D in the channel. We want to estimate the average concentration of the pollutant at the end of the channel and assess the uncertainty associated with this value since harmful effects to the environment are usually estimated from pollutant concentrations. We want to know the joint posterior distribution of all the parameters, but the parameter L is of special interest because L locates the as yet unidentified industry that will need to pay for its share of the clean-up costs.

A first-order approach to modeling the concentration of substances in such channels is to assume that the channel can be approximated by an infinitely long one-dimensional system in which diffusion is the only transport device. We assume that the spills are each of mass M and occur instantaneously at space-time points $(s, t) = (0, 0)$ and $(s, t) = (L, \tau)$ and that the diffusion coefficient D is constant in both time and space. This leads to the concentration representation (for $t > 0$ and $s \geq 0$):

$$C(s, t; M, D, L, \tau) = \frac{M}{\sqrt{4\pi Dt}} \exp\left[\frac{-s^2}{4Dt}\right] + \frac{M}{\sqrt{4\pi D(t-\tau)}} \exp\left[\frac{-(s-L)^2}{4D(t-\tau)}\right] \cdot \mathbb{I}(\tau < t), \quad (9)$$

where $\mathbb{I}$ is the indicator function. We take $\boldsymbol{\beta}$ to be the vector of the four unknown environmental parameters (M, D, L, τ) and consider the scaled concentration $f\{(s, t), \boldsymbol{\beta}\} =$

$\sqrt{4\pi}C(s,t;\boldsymbol{\beta})$.

We assume that each of the five monitoring stations fixed at spatial locations $s_j = 0, .5, 1, 1.5, 2.5$ record 200 concentration readings at times $t_k = .3, .6, \ldots, 50.7, 60$. The corresponding expensive model function is $\boldsymbol{f}(\boldsymbol{\beta}) = \{f(X_1,\boldsymbol{\beta}), \ldots, f(X_{1000},\boldsymbol{\beta})\}^{\mathsf{T}}$, where $X_i = (s_j, t_k)$ if $i = (j-1)\cdot 200 + k$. In this example $f(X_i,\boldsymbol{\beta})$ is scalar because there is only one pollutant.

The ultimate goal of the study is to assess the space-time prediction uncertainty associated with the average concentration at the end of the channel, corresponding to $s = 3$, over the time interval $[40, 140]$. To this end we consider the function $F(\boldsymbol{\beta}) = \sum_{i=0}^{20} f\{(3, 40+5i), \boldsymbol{\beta}\}$ which requires evaluation of f at the additional points $\{(3,40), (3,45), \ldots, (3,140)\}$. As discussed in Section 2, the expensive model that we evaluate is

$$\boldsymbol{f}^{*}(\boldsymbol{\beta}) = \left(\boldsymbol{f}\{\boldsymbol{\beta}\}^{\mathsf{T}}, f\{(3,40),\boldsymbol{\beta}\}, \ldots, f\{(3,140),\boldsymbol{\beta}\}\right)^{\mathsf{T}}.$$

An intermediate goal is estimation of the posterior density of $\boldsymbol{\beta}$, which is partially captured by the marginal densities of its components.

The vector $\boldsymbol{Y}$ of observed concentrations is generated according to the model given by Equations (1) and (9) with the value of $\lambda = .333$ for the COIL family given by Equation (2) and the values of $\boldsymbol{\beta}_i$'s and respective parameter spaces (domains) given in Table 1. Notice that the likelihood from Equation (1) is discontinuous since $C(s_i, t_j; \boldsymbol{\beta})$ in Equation (9) explodes when $\boldsymbol{\beta}_3 \equiv L = s_i$ and $\boldsymbol{\beta}_4 \equiv \tau$ approaches t_j from below. To avoid discontinuities, the parameter space for $\boldsymbol{\beta}_4$ was restricted to the interval containing the true value of the parameter, given in Table 1. Components of $\boldsymbol{Y}$ are independent, with variance of $h(Y_i, \lambda)$ for every i equal to the sample variance of $h\{f(X_1,\boldsymbol{\beta}),\lambda\}, \ldots, h\{f(X_{1000},\boldsymbol{\beta}),\lambda\}$, computed for the true (fixed) values of $\boldsymbol{\beta}$ and λ and multiplied by a scaling constant c^2. Here, c controls the amount of noise in the transformed observed data relative to the variability of the corresponding transformed model values. For illustrative purposes – to ensure that the likelihood has a single dominant mode – c is set to .3. We put a uniform prior on $(\boldsymbol{\beta}, \lambda)$ over the parameter spaces mentioned earlier and an overdispersed inverse-gamma prior on σ^2. This is a special case of the earlier model for multiple chemical species, and, as shown in Appendix A.4, σ^2 can be integrated out analytically. Thus $[\boldsymbol{\beta}, \lambda, \boldsymbol{Y}]$ is the unnormalized posterior from which we derive $\pi(\cdot, \boldsymbol{Y})$ as discussed in Section 3.1.

4.2 Analysis

We applied our algorithm to a large number of dataset replications with the same statistical model and parameter values but a different realization of the noise. The maximizer $(\widehat{\boldsymbol{\beta}}, \widehat{\lambda})$ of $[\boldsymbol{\beta}, \lambda, \boldsymbol{Y}]$ was found by `CONDOR` via maximization of $\pi_{\max}(\cdot, \boldsymbol{Y})$ given by Equation (4), and $\widehat{\boldsymbol{I}}$ was estimated by fitting a quadratic, as explained in Sections 3.2.1 and 3.2.2 and Appendix A.1. The mean and standard deviation of the number of function evaluations to find the MAP when started at a random point from the uniform distribution on the parameter space were around 100 and 20, respectively. Nearly all of the points in $\mathcal{B}_O$ produced in **Step 1** were sufficiently separated to be used as part of the experimental design. However, usually just over one third of them were actually valuable for Hessian estimation or surface approximation while the rest were outside of $\widehat{C}_R(\alpha')$ and hence too far away from the mode. Based on 1000 dataset replications, the mean and median numbers of new design points to estimate $\widehat{\boldsymbol{I}}$ by fitting a quadratic, including the 8 points corresponding to forward differencing to obtain an estimate of the diagonal, were 20 and 19, respectively. A small correlation between $\boldsymbol{\beta}$ and λ and small variance of λ, as estimated by the entries of $\widehat{\boldsymbol{I}}^{-1}$, indicate that $[\boldsymbol{\beta}|\boldsymbol{Y}]$ is likely to be close to the conditional density $[\boldsymbol{\beta}|\lambda = \widehat{\lambda}, \boldsymbol{Y}]$.

Experiments were run for elliptical design regions $\widehat{C}_R(\alpha)$ given by Equation (7), with α in $\{.2, .1, .05, .01\}$ and numbers of extra experimental design points ($|\mathcal{B}_E|$) in $\{0, 10, 20, \ldots, 100\}$. It was observed that the estimates of the posterior densities based on MCMC samples from approximate surfaces are not very sensitive to the size of $\widehat{C}_R(\alpha)$ provided there are enough extra design points, with larger regions requiring greater numbers of extra design points. Also, for a fixed α, there is usually little (visual) improvement in density estimates when the size of $\mathcal{B}_E$ grows above 50, and the quality of approximation is often unsatisfactory for the sizes of $\mathcal{B}_E$ below 20.

All graphical summaries of posterior densities for $\boldsymbol{\beta}$ and $F(\boldsymbol{\beta})$ that we report for a single representative dataset correspond to the $\widehat{C}_R(.1)$ region with 30 extra design points for $\boldsymbol{\beta}$, the same for every surface. In the case of RBF interpolation of $\log\{[\boldsymbol{\beta}, \lambda, \boldsymbol{Y}]\}$, for each $\boldsymbol{\beta}^{(i)} \in \mathcal{B}_D$, we choose 10 design points for λ and fit the RBF surface at the design points $\{(\boldsymbol{\beta}^{(i)}, \lambda_{ij})$ for $i = 1, \ldots, N$ and $j = 1, \ldots, 10\}$, chosen as outlined in Appendix A.2. All tabular summaries pertain to this RBF approximation to $[\boldsymbol{\beta}, \lambda, \boldsymbol{Y}]$ and the true surface $[\boldsymbol{\beta}, \lambda, \boldsymbol{Y}]$. All results are reported for the ARMH sampler with $\boldsymbol{\rho} = -0.25$, the density g

being the equal-weight mixture of a multivariate normal and Cauchy distributions and with other parameters chosen as discussed in Section 3.2.4.

Figure 1 overlays plots of kernel density estimates for the components of $\boldsymbol{\beta}$ using MCMC samples of size 30,000 from the RBF approximations to the pseudoposterior, the profile posterior (with and without Laplace correction) and to the joint posterior. The close agreement among the curves supports the conjecture that the approximation of $[\boldsymbol{\beta}|\boldsymbol{Y}]$ by $[\boldsymbol{\beta}|\lambda = \widehat{\lambda}, \boldsymbol{Y}]$ is meaningful for this test problem. The plots of the estimates of the marginal densities of $\boldsymbol{\beta}_i$'s based on a M-H sample of the same length from the exact unnormalized joint posterior $[\boldsymbol{\beta}, \lambda, \boldsymbol{Y}]$, also on the same figure, confirm the validity of findings. Of course, in practice we would not be able to visualize the true surface in the way we do it here due to computational limitations.

Samples from the approximate posterior of $F(\boldsymbol{\beta})$ were obtained by first interpolating F at $\boldsymbol{\beta} \in \mathcal{B}_D$ by the (cubic) RBF surface of the form given by the right-hand side of Equation (8) and then evaluating the resulting interpolant at the MCMC samples from the approximate surfaces stated in the previous paragraph; see Section 3.3 for discussion. Likewise, the sample from the true posterior of $F(\boldsymbol{\beta})$ was obtained by evaluating F at the sample from $[\boldsymbol{\beta}|\boldsymbol{Y}]$. Striking similarity between the exact and RBF results in Table 1 and in Figure 2 suggest that our method is capable of achieving nearly the full accuracy of estimation at the expense of only a small fraction of the computational cost required to carry out MCMC sampling using the exact posterior surface.

Another important criterion for the assessment of the methodology concerns examination of the coverage properties of the Bayesian credible intervals for the components of $\boldsymbol{\beta}$ and for $F(\boldsymbol{\beta})$ based on our approximations. We believe that if the final goal is a credible interval for $F(\boldsymbol{\beta})$, then a very accurate approximation to the posterior of $\boldsymbol{\beta}$ is unnecessary, and a credible interval based on a less accurate approximation (requiring a smaller number of extra design points) is likely to have similar coverage properties to those of the credible interval based on the corresponding true posterior.

Below we report findings of the following study: for each of the 1000 replications of $\boldsymbol{Y}$ under the same model and parameter values but a different realization of noise, we find the MAP and estimate $\widehat{\boldsymbol{I}}$ as before. We then take an MCMC sample of size 10,000 using the true surface $[\boldsymbol{\beta}, \lambda, \boldsymbol{Y}]$ and the respective RBF surface with approximate HPD region $\widehat{C}_R(.1)$ and the number of extra experimental design points $|\mathcal{B}_E| = 30$. Reported in Table 2 are the observed coverage proportions of the components $\boldsymbol{\beta}_i^*$ of the true value of $\boldsymbol{\beta}$ by symmetric credible intervals of sizes .9, .95 and .99 obtained from the MCMC samples, along with the corresponding standard errors. The last three columns of Table 1 give means and standard deviations for the ratios of the lengths of RBF and exact credible intervals over all datasets.

The outcomes of the experiment allow one to conclude that the coverage properties of the credible intervals based on exact and approximate surfaces are similar, and that the respective interval lengths are close. The tables also suggest that the approximate credible intervals act as reasonable substitutes for frequentist confidence intervals, as the observed coverage proportions are close to the nominal sizes of the confidence intervals.

5 DISCUSSION

5.1 Survey of Literature

Literature dealing with Bayesian calibration and prediction of output of complex computer models focuses on reduction of the number of expensive function evaluations via approximation of $\boldsymbol{f}$ rather than of the posterior. Papers by O'Hagan, Kennedy and Oakley (1998) and Kennedy and O'Hagan (2000) assume that the model can be run at different levels of complexity, which they use to build a Markov-type model for dependence between consecutive levels of the code and to utilize the cheaper lower levels of the code to help predict the expensive top-level code output. (In Kennedy and O'Hagan (2001), the unobservable physical model, approximated by a complex code, plays the role of the top-level code.) Under a similar assumption that the computer code can be "coarsened", in particular, to lower the dimension of input ($\boldsymbol{\beta}$ in our notation), Higdon, Lee and Holloman (2002) run coarse and fine (corresponding to the original expensive model) Markov chains in tandem and use information from the faster-mixing coarse chain to improve mixing of the fine chain. A recent paper of Christen and Fox (2005) proposes to use a cheap-to-evaluate approximation to the unnormalized posterior density, which is assumed to be available, to evaluate the expensive posterior only for the MCMC moves that are likely to be accepted. The emphasis, however, is on the models for which approximation to the posterior is obtained by replacing $\boldsymbol{f}$ by an approximation, for example, by linearization.

5.2 Differences from Earlier Approaches

The route we take is significantly different from those mentioned. First, we are not assuming that the "black-box" $\boldsymbol{f}$ can be opened to produce an inexpensive approximation to the expensive code. Second, given our interest in the sample from the posterior for $\boldsymbol{\beta}$, we approximate the (scalar-valued) posterior *directly*, and not through approximation of the high-dimensional model output $\boldsymbol{f}$. Third, we realize that in many problems sampling thousands of times from the exact posterior is not computationally feasible and thus work solely with the approximation, unlike Higdon et al. (2002) and Christen and Fox (2005). For problems in which the approach of the latter two authors is feasible, our RBF approximation can be modified (by adding more mass to the tails of $\widetilde{\pi}(\cdot, \boldsymbol{Y})$) to provide an initial approximation to the exact posterior that can be gradually refined as more expensive code evaluations are made in the course of sampling.

5.3 Computational Issues

Over a number of dataset replications, the results indicate that our method can produce results similar to those that require MCMC sampling of the exact posterior surface, but with far fewer computationally expensive functions evaluations. Using an average number of 150 function evaluations, our method produces estimates of posterior densities for $\boldsymbol{\beta}$ and $F(\boldsymbol{\beta})$ that are nearly the same as those obtained at the cost of 30,000 evaluations of $\boldsymbol{f}^*$ incurred when sampling from the exact posterior.

Since $\boldsymbol{f}^*$ is a computationally expensive function, the total computation times for solving the problem are based almost entirely on the number of function evaluations. The number of function evaluations is $|\mathcal{B}_O| + |\mathcal{B}_H| + |\mathcal{B}_E|$; recall definitions in Sections 3.2.1 and 3.2.2. These evaluations of $\boldsymbol{f}^*$ take place in **Step 1** when optimization is used to find the MAP and in **Step 2** when an improvement in the accuracy of the approximation of the logarithm of posterior surface is obtained by doing additional function evaluations in a high-probability neighborhood of the maximizer. The time for other computational steps including fitting the RBF, setting up the optimization algorithm to find the next design point, and performing all the calculations in **Steps 3** and **4** take a tiny fraction of the computational time required for the evaluations of $\boldsymbol{f}^*$.

In contrast to typical optimization searches, here we want to use the function evaluations generated during the optimization search not only to find the maximizer, but also to help characterize the function around the maximizer. We then "re-use" the function evaluations from the optimization search to build an RBF surface. We use derivative-free optimization methods that generate design points, at which $\boldsymbol{f}^*$ is evaluated, that are a reasonable distance apart. Our experience with derivative-based optimization for our example was that many of the design points were so close to each other that they were not carrying much new information about the surface values (in addition to that provided by their neighbors) and that they could not all be used in RBF interpolation without creating numerical instabilities.

5.4 Conclusions

This paper presented a Bayesian solution to calibration problem when the feasible number of evaluations of the computationally expensive model is relatively small and no cheap-to-evaluate approximation to the computer model is available. Sampling from the exact expensive posterior using MCMC is questionable under such restrictions. Indeed, the computational budget may not be sufficient to ensure both that the Markov chain gets close to its limiting distribution and that the sample is "informative" about $\boldsymbol{\beta}$ if the Markov chain mixes slowly. The main contribution is the algorithm of Section 3.2, which re-uses a subset of design points from a derivative-free optimization search, augmented with additional design points, to build an RBF approximation for the posterior on the region of high posterior probability. This allows one to draw arbitrarily long samples from the cheap proxy to the true expensive posterior.

Our results required, on average, 150 evaluations of $\boldsymbol{f}^*$, which is less than 1/60 of the 10,000 evaluations used by the MCMC procedure that samples the exact posterior. (Of course, fewer iterations could be used, but this would noticeably increase Monte Carlo error.) This difference is very significant for computationally expensive functions for which many thousands of evaluations are infeasible. Our method hence shows promise as a means for doing a rigorous Bayesian uncertainty analysis on some functions (including simulation models) for which there currently does not exist a numerically feasible alternative method.

5.5 Further Developments

Our current work focuses on extending the approach to deal with $\boldsymbol{f}$ under less restrictive smoothness assumptions, posteriors with multiple important modes and pronounced skewness, and on developing a sequential procedure for determining an approximate HPD region and an experimental design on it. A systematic study of the effect of space-time dependence is also needed. After more insight into these issues is gained, the methodology will be applied to a truly expensive model.

ACKNOWLEDGMENTS

The research of Bliznyouk was supported by the National Science Foundation under grant DMS-04-538 (PI's Ruppert and Shoemaker). Research of Wild was supported by a Department of Energy Computational Science Graduate Fellowship, grant number DE-FG02-97ER25308. Regis was supported by NSF grant CCF-0305583 (PI Shoemaker) and Mugunthan on NSF grant BES-0229176 (PI Shoemaker). Ruppert and Shoemaker were also partially supported on the grants on which they are PI's.

A APPENDIX

A.1 Estimation of $\widehat{\boldsymbol{I}}$

Let $p = \dim(\boldsymbol{\beta})$ and $u = \dim(\boldsymbol{\zeta})$. To approximate the Hessian $\widehat{\boldsymbol{I}}$ of $-\log\{[\boldsymbol{\beta}, \boldsymbol{\zeta}, \boldsymbol{Y}]\}$ at $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\zeta}})$ by forward differences generally requires (around) $(p + u + 1)(p + u + 2)/2$ function evaluations. Partition $\widehat{\boldsymbol{I}}$ as in Equation (6). In our problem, $\widehat{\boldsymbol{I}}_{\boldsymbol{\zeta\zeta}}$ can be found analytically, and the off-diagonal blocks of $\widehat{\boldsymbol{I}}$ can be computed entirely from the evaluations of $\boldsymbol{f}$ used to estimate the diagonal of $\widehat{\boldsymbol{I}}_{\boldsymbol{\beta\beta}}$.

Unfortunately, the $(p + 1)(p + 2)/2$ points for evaluation of $\boldsymbol{f}$ by finite differences to compute $\widehat{\boldsymbol{I}}_{\boldsymbol{\beta\beta}}$ are very close to $\widehat{\boldsymbol{\beta}}$ and are not valuable for surface approximation. However, one can lower the number of uninformative design points using the approach below.

Taylor's theorem suggests an approximation $\log\{[\boldsymbol{\beta}, \widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}), \boldsymbol{Y}]\} \approx \text{const} - \frac{1}{2}\|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\|^2_{\widehat{\boldsymbol{I}}_{\boldsymbol{\beta\beta}}}$ on the ellipsoid $\mathcal{E}(c) = \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\|^2_{\widehat{\boldsymbol{I}}_{\boldsymbol{\beta\beta}}} \leq c^2\}$ for some c. We propose to choose $(p + 1)(p + 2)/2$ design points that are well-separated inside $\mathcal{E}(c)$, fit a quadratic surface through them and estimate $\widehat{\boldsymbol{I}}_{\boldsymbol{\beta\beta}}$ by the Hessian of the quadratic. The task is to ensure that these points lie inside $\mathcal{E}(c)$ without knowing the shape and orientation of the ellipsoid. Denote by $\boldsymbol{e}_i$ the ith standard basis vector for $\mathbb{R}^p$ and notice that the boundary of $\mathcal{E}(c)$ passes through points $\widehat{\boldsymbol{\beta}} \pm \boldsymbol{b}_i$, where $\boldsymbol{b}_i = \boldsymbol{e}_i \cdot c/\sqrt{\widehat{\boldsymbol{I}}_{\boldsymbol{\beta\beta}}(i, i)}$ and $\widehat{\boldsymbol{I}}_{\boldsymbol{\beta\beta}}(i, i)$ is the ith diagonal entry of $\widehat{\boldsymbol{I}}_{\boldsymbol{\beta\beta}}$. The convex hull of these points

$$\mathcal{H}(c) = \left\{ \begin{array}{c} \boldsymbol{\beta} : \quad \boldsymbol{\beta} = \widehat{\boldsymbol{\beta}} + \sum_{i=1}^{p} (\psi_{i,1} - \psi_{i,2}) \boldsymbol{b}_i \text{ such that } \sum_{j=1}^{2} \sum_{i=1}^{p} \psi_{i,j} = 1 \\ \text{and } \psi_{i,j} \geq 0 \text{ for } i = 1, \ldots, p \text{ and } j = 1, 2 \end{array} \right\}$$

is a subset of $\mathcal{E}(c)$, and so it is guaranteed that any experimental design on $\mathcal{H}(c)$ also lies in $\mathcal{E}(c)$. Hence one only needs to estimate the diagonal of $\widehat{\boldsymbol{I}}_{\boldsymbol{\beta}\boldsymbol{\beta}}$ by forward differences using at most $2 \cdot p$ extra function evaluations, half of which are reused in computation of the off-diagonal blocks $\widehat{\boldsymbol{I}}_{\boldsymbol{\beta}\boldsymbol{\zeta}}$. (Thus we have reduced the number of "uninformative" design points roughly by $(p+1)(p-2)/2$.)

The argument c in the definition of the ellipsoid is to be chosen by the experimenters in accord with their beliefs. It is helpful to think of c^2 as a quantile of the χ_p^2 distribution that defines a confidence ellipsoid for the multivariate normal approximation to $[\boldsymbol{\beta}|\boldsymbol{\zeta} = \widehat{\boldsymbol{\zeta}}, \boldsymbol{Y}]$. Large values of c often yield Hessians that are inaccurate or not positive definite, and very small values result in new design points close to $\widehat{\boldsymbol{\beta}}$. We had some success with the following procedure to estimate $\widehat{\boldsymbol{I}}$ and to produce the set $\mathcal{B}_H$ from Section 3.2.2:

1. Initialization:

   (a) Choose a moderate initial value of c, say, $\sqrt{\chi^2_{p,0.1}}$.

   (b) Set $\mathcal{B}_H = \mathcal{B}_O$ and remove from $\mathcal{B}_H$ points very close to each other. The set $\mathcal{B}_H$ will contain values of $\boldsymbol{\beta}$ for which $\boldsymbol{f}^*$ has been computed. Assume for now that $|\mathcal{B}_H| \leq (p+1)(p+2)/2$.

2. Augment $\mathcal{B}_H$ with new well-separated points so that $|\mathcal{B}_H \cap \mathcal{H}(c)| = (p+1)(p+2)/2$ and evaluate $\boldsymbol{f}^*$ at the new points.

3. Fit a quadratic surface through the points in $\mathcal{B}_H \cap \mathcal{H}(c)$ and plug its Hessian into the expression for $\widehat{\boldsymbol{I}}$. If the resulting estimate of $\widehat{\boldsymbol{I}}$ (not only that of $\widehat{\boldsymbol{I}}_{\boldsymbol{\beta}\boldsymbol{\beta}}$) is not positive definite, reduce c and return to the previous step. Otherwise terminate; return $\widehat{\boldsymbol{I}}$ and (the set difference) $\mathcal{B}_H = \mathcal{B}_H - \mathcal{B}_O$.

If $|\mathcal{B}_H| > (p+1)(p+2)/2$ in step 1(b) above, then no new design points are necessary and one starts by working with subsets of $\mathcal{B}_H$. Once they are exhausted, one moves to Step 2.

A.2 Choice of Design Points

While forming $\mathcal{B}_H$ and $\mathcal{B}_E$, we require that the new design points lie far from each other and from the points where $\boldsymbol{f}^*$ has been evaluated previously ("fixed points"). This objective is related to the *maximin* criterion that attempts to *max*imize the *min*imum between-point distance over all pairs of design points (Santner, Williams and Notz (2003, sec. 5.3)). Generating such designs exactly is computationally difficult, and usually one is happy to obtain a good approximate maximin design (Trosset 1999).

We devised a simple "greedy" algorithm to update the set of fixed points with N_E extra design points: (i) choose $\kappa > 1$ and draw $\lceil \kappa \cdot N_E \rceil$ *candidate* points uniformly at random on the design region; (ii) at the jth iteration, find the pair of points closest to each other that has at least one candidate point; if it has a single candidate point, delete it, otherwise delete the one that is closest to the remaining (fixed and candidate) points, until only N_E candidate points remain. This algorithm is applied to update $\mathcal{B}_O$ to produce $\mathcal{B}_H$ and then to augment $\mathcal{B}_O \cup \mathcal{B}_H$ with $\mathcal{B}_E$. As an intermediate step, one has to sample uniformly inside polytopes and spheres; for discussion, see Devroye (1986, chap. 5).

To obtain the (joint) experimental design for $\boldsymbol{\beta}$ and $\boldsymbol{\zeta}$ for fitting an RBF surface to $\log\{[\boldsymbol{\beta}, \boldsymbol{\zeta}, \boldsymbol{Y}]\}$, we start with a (marginal) design for $\boldsymbol{\beta}$ on $\widehat{C}_R(\alpha)$ as in Section 3.2.2 and augment each design point $\boldsymbol{\beta}^{(i)} \in \mathcal{B}_D$ with a (conditional) design for $\boldsymbol{\zeta}$ based on the multivariate normal approximation to $[\boldsymbol{\zeta} | \boldsymbol{\beta} = \boldsymbol{\beta}^{(i)}, \boldsymbol{Y}]$, derived from Equation (5). As a consequence, the increase in the dimension of the argument of the posterior (going from $\pi(\cdot, \boldsymbol{Y})$ to $[\boldsymbol{\beta}, \boldsymbol{\zeta}, \boldsymbol{Y}]$) in this problem need not translate into the increase in the number of evaluations of $\boldsymbol{f}$.

A.3 Details for Fitting the RBF Surface

We now describe the procedure for fitting the RBF interpolation model of Section 3.2.3 with the cubic basis function and a linear polynomial tail $q(\boldsymbol{\beta}) = (1, \boldsymbol{\beta}^\mathsf{T}) \cdot \boldsymbol{c}$. Discussion of fitting for other choices of basis functions is in Powell (1996).

Define the matrix $\boldsymbol{\Phi} \in \mathbb{R}^{N \times N}$ by: $\boldsymbol{\Phi}_{i,j} = \phi(\|\boldsymbol{\beta}^{(i)} - \boldsymbol{\beta}^{(j)}\|_2)$, for $i, j = 1, \ldots, N$. Let $\boldsymbol{P} \in \mathbb{R}^{N \times (p+1)}$ be the matrix with $(1, \{\boldsymbol{\beta}^{(i)}\}^\mathsf{T})$ as the ith row for $i = 1, \ldots, N$. The coefficients for the RBF surface that interpolates $l(\cdot) = \log(\pi(\cdot, \boldsymbol{Y}))$ at the points $\boldsymbol{\beta}^{(1)}, \ldots, \boldsymbol{\beta}^{(N)}$ are obtained by solving the system

$$\begin{pmatrix} \boldsymbol{\Phi} & \boldsymbol{P} \\ \boldsymbol{P}^\mathsf{T} & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{a} \\ \boldsymbol{c} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mathcal{L}} \\ \boldsymbol{0} \end{pmatrix}, \tag{10}$$

where $\boldsymbol{\mathcal{L}} = \left[l(\boldsymbol{\beta}^{(1)}), \ldots, l(\boldsymbol{\beta}^{(N)})\right]^{\mathsf{T}}$, $\boldsymbol{a} \in \mathbb{R}^N$ and $\boldsymbol{c} \in \mathbb{R}^{p+1}$. Our RBF interpolation model is a form of universal kriging with generalized covariances, as the above RBF interpolation equations are identical to the dual-kriging equations (Cressie 1991, sec. 4.4.5).

The coefficient matrix in Equation (10) is invertible if and only if the rank of $\boldsymbol{P}$ is $p+1$ (Powell 1992). For stability purposes, we solve Equation (10) by means of the matrix factorizations, as described in Powell (1996).

A.4 Details on Integrating $\boldsymbol{C}$ out

Let $\boldsymbol{Z} = \boldsymbol{Z}(\boldsymbol{\beta}, \boldsymbol{\lambda})$ be the matrix with the ith row $[h\{Y_i, \boldsymbol{\lambda}\} - h\{f(X_i, \boldsymbol{\beta}), \boldsymbol{\lambda}\}]^{\mathsf{T}}$ for $i = 1, \ldots, n$, $\boldsymbol{R}$ be the upper-triangular Cholesky factor of $\boldsymbol{S}(\boldsymbol{\gamma})$ and $\widetilde{\boldsymbol{Z}} = \boldsymbol{R}^{-\mathsf{T}}\boldsymbol{Z}$. Notice that, under the separable covariance model $\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{S}(\boldsymbol{\gamma}) \otimes \boldsymbol{C}$ of Section 2.3, the likelihood equation (1) implies that the rows $\widetilde{\boldsymbol{Z}}_{1,\bullet}, \ldots, \widetilde{\boldsymbol{Z}}_{n,\bullet}$ of $\widetilde{\boldsymbol{Z}}$ are *i.i.d.* $MVN(\mathbf{0}, \boldsymbol{C})$. Notice that

$$\begin{aligned}
[\boldsymbol{Y}|\boldsymbol{\beta}, \boldsymbol{\lambda}, (\boldsymbol{\gamma}, \boldsymbol{C})] &\propto |J_h(\boldsymbol{Y}; \boldsymbol{\lambda})| \cdot |\boldsymbol{S}(\boldsymbol{\gamma})|^{-d/2} |\boldsymbol{C}|^{-n/2} \prod_{j=1}^{n} \exp\left(-0.5 \cdot \|\widetilde{\boldsymbol{Z}}_{j,\bullet}\|^2_{\boldsymbol{C}^{-1}}\right) \\
&= |J_h(\boldsymbol{Y}; \boldsymbol{\lambda})| \cdot |\boldsymbol{S}(\boldsymbol{\gamma})|^{-d/2} |\boldsymbol{C}|^{-n/2} \exp\left(-0.5 \cdot tr\{\boldsymbol{C}^{-1}\widetilde{\boldsymbol{Z}}^{\mathsf{T}}\widetilde{\boldsymbol{Z}}\}\right).
\end{aligned}$$

We put a Wishart prior on $\boldsymbol{C}^{-1}$ and assume that, a priori, $\boldsymbol{C}$ and $(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\gamma})$ are independent, so that

$$[\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{C}^{-1}] \propto |\Delta|^a |\boldsymbol{C}^{-1}|^{a-(d+1)/2} \exp\left(-tr\{\Delta \boldsymbol{C}^{-1}\}\right) \cdot [\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\gamma}],$$

where $a > (d-1)/2$, $\Delta \in \mathcal{M}_d$, the space of $d \times d$ symmetric positive definite matrices, and $tr(\cdot)$ is the trace operator. This allows us to integrate $\boldsymbol{C}$ out of the joint posterior analytically:

$$\begin{aligned}
[\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{Y}] &= \int_{\mathcal{M}_d} [\boldsymbol{Y}|\boldsymbol{\beta}, \boldsymbol{\lambda}, (\boldsymbol{\gamma}, \boldsymbol{C})] \cdot [\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{C}^{-1}] d\boldsymbol{C}^{-1} \\
&\propto c(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\gamma}) \int_{\mathcal{M}_d} \exp\left(-tr\{\boldsymbol{C}^{-1}(\widetilde{\boldsymbol{Z}}^{\mathsf{T}}\widetilde{\boldsymbol{Z}}/2 + \Delta)\}\right) |\boldsymbol{C}^{-1}|^{a+(n-d-1)/2} d\boldsymbol{C}^{-1} \\
&\propto c(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\gamma}) \cdot |\widetilde{\boldsymbol{Z}}^{\mathsf{T}}\widetilde{\boldsymbol{Z}}/2 + \Delta|^{-(a+n/2)} \\
&= [\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\gamma}] \cdot |J_h(\boldsymbol{Y}; \boldsymbol{\lambda})| \cdot |\boldsymbol{S}(\boldsymbol{\gamma})|^{-d/2} \cdot |\boldsymbol{Z}^{\mathsf{T}}[\boldsymbol{S}(\boldsymbol{\gamma})]^{-1}\boldsymbol{Z}/2 + \Delta|^{-(a+n/2)}.
\end{aligned}$$

References

[1] Box, G. E. P., and Cox, D. R. (1964), "An Analysis of Transformations", *Journal of the Royal Statistical Society, Series B*, 26, 211–246.

[2] Buhmann, M. D. (2003), *Radial Basis Functions*, New York: Cambridge University Press.

[3] Carroll, R. J., and Ruppert, D. (1984), "Power Transformation When Fitting Theoretical Models to Data", *Journal of the American Statistical Association*, 79, 321–328.

[4] ———— (1988), *Transformation and Weighting in Regression*, New York: Chapman & Hall.

[5] Christen, J. A., and Fox, C. (2005), "Markov Chain Monte Carlo Using an Approximation", *Journal of Computational & Graphical Statistics*, 14, 795–810.

[6] Cressie, N. (1991), *Statistics for Spatial Data*, New York: Wiley.

[7] Devroye, L. (1986), *Non-Uniform Random Variate Generation*, New York: Springer-Verlag.

[8] Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004), *Bayesian Data Analysis* (2nd ed.), Boca Raton: Chapman & Hall/CRC.

[9] Higdon, D., Lee H., and Holloman, C. (2002), "Markov Chain Monte Carlo-Based Approaches for Inference in Computationally Intensive Inverse Problems", in *Bayesian Statistics 7*, eds. J. M. Bernardo, J. O. Berger, A. P. Berger and A. F. M. Smith, pp. 181–197.

[10] Kennedy, M. C., and O'Hagan, A. (2000), "Predicting the Output From a Complex Computer Code When Fast Approximations are Available", *Biometrika*, 87, 1–13.

[11] ———— (2001), "Bayesian Calibration of Computer Models", *Journal of the Royal Statistical Society, Series B*, 63, 425–464.

[12] Mugunthan, P., Shoemaker, C. A., and Regis, R. G. (2005), "Comparison of Function Approximation, Heuristic and Derivative-Based Methods for Automatic Calibration of

Computationally Expensive Groundwater Bioremediation Models", *Water Resources Research*, 41, W11427, doi:10.1029/2005WR004134.

[13] Mugunthan, P. and Shoemaker, C. A. (2006, in press), "Assessing the Impacts of Parameter Uncertainty for Computationally Expensive Groundwater Models", *Water Resources Research.*

[14] O'Hagan, A., Kennedy, M. C., and Oakley, J. E. (1998), "Uncertainty Analysis and Other Inference Tools for Complex Codes", in *Bayesian Statistics 6*, eds. J. M. Bernardo, J. O. Berger, A. P. Berger and A. F. M. Smith, pp. 503–524.

[15] Powell, M. J. D. (1992), "The Theory of Radial Basis Function Approximation in 1990", in *Advances in Numerical Analysis, Volume 2: Wavelets, Subdivision Algorithms and Radial Basis Functions*, ed. W. Light, New York: Oxford University Press, pp. 105–210.

[16] ———— (1996), "A Review of Algorithms for Thin Plate Spline Interpolation in Two Dimensions", in *Advanced Topics in Multivariate Approximation*, eds. F. Fontanella, K. Jetter and P. J. Laurent, River Edge, NJ: World Scientific Publishing, pp. 303–322.

[17] ———— (2002), "UOBYQA: Unconstrained Optimization by Quadratic Approximation", *Mathematical Programming*, 92, 555–582.

[18] Santner, T.J., Williams, B.J. and Notz, W. (2003), *The Design and Analysis of Computer Experiments.* New York: Springer–Verlag.

[19] Regis, R.G., and Shoemaker, C.A. (2006, in press), "A Stochastic Radial Basis Function Method for the Global Optimization of Expensive Functions," *INFORMS Journal of Computing.*

[20] Tierney, L. (1994), "Markov Chains for Exploring Posterior Distributions," *The Annals of Statistics*, 22, 1701–1786.

[21] Tierney, L., and Kadane, J. B. (1986), "Accurate Approximations for Posterior Moments and Marginal Densities," *Journal of the American Statistical Association*, 81, 82–86.

[22] Thyer, M. A., Kuczera, G., and Wang, Q. J. (2002), "Quantifying Parameter Uncertainty in Stochastic Models Using the Box-Cox Transformation," *The Journal of Hydrology*, 265, 246–257.

[23] Trosset, M. W. (1999), “Approximate Maximin Distance Designs”, in *American Statistical Association Proceedings of the Physical and Engineering Sciences Section*, pp. 223–227.

[24] Vanden Berghen, F., and Bersini, H. (2005), “CONDOR, a New Parallel, Constrained Extension of Powell’s UOBYQA Algorithm: Experimental Results and Comparison With the DFO Algorithm”, *Journal of Computational and Applied Mathematics*, 181, 157–175.

Table 1: Parameter spaces and true parameter values, mean and (standard deviation) of Monte Carlo mean, mean and (standard deviation) of ratios of lengths of RBF to exact credible intervals, based on 1000 dataset replications and $[\boldsymbol{\beta}, \lambda, \boldsymbol{Y}]$ as the surface.

			MC mean		ratio of lengths of cred. int.'s		
	domain	true	exact	RBF	size .9	size .95	size .99
$\boldsymbol{\beta}_1$	[7, 13]	10	10.0057 (.0866)	10.0061 (.0893)	.9969 (.0602)	.9961 (.0624)	.9844 (.0738)
$\boldsymbol{\beta}_2$	[.02, .12]	.07	.07008 (.00097)	.07008 (.00101)	.9910 (.0592)	.9888 (.0612)	.9687 (.0673)
$\boldsymbol{\beta}_3$	[.01, 3]	1	1.0005 (.0136)	1.0005 (.0134)	.9671 (.0785)	.9662 (.0765)	.9604 (.0750)
$\boldsymbol{\beta}_4$	[30.01, 30.295]	30.16	30.1610 (.0096)	30.1610 (.0096)	.9786 (.0779)	.9709 (.0818)	.9403 (.0835)
$F(\boldsymbol{\beta})$	–	128.998	129.063 (1.087)	129.067 (1.100)	.9959 (.062)	.9937 (.0628)	.9841 (.0695)

Table 2: Observed probabilities of coverage with (standard errors) of symmetric credible intervals based on 1000 dataset replications and joint posterior as the surface for MCMC.

	size .9 cred. int.			size .95 cred. int.			size .99 cred. int.	
	exact	RBF		exact	RBF		exact	RBF
$\boldsymbol{\beta}_1$	.905 (.009)	.904 (.009)		.950 (.007)	.944 (.007)		.986 (.004)	.990 (.003)
$\boldsymbol{\beta}_2$	.908 (.009)	.903 (.009)		.954 (.007)	.951 (.007)		.991 (.003)	.987 (.004)
$\boldsymbol{\beta}_3$	.916 (.009)	.899 (.010)		.953 (.007)	.954 (.007)		.989 (.003)	.988 (.003)
$\boldsymbol{\beta}_4$	.904 (.009)	.909 (.009)		.947 (.007)	.945 (.007)		.988 (.003)	.987 (.004)
$F(\boldsymbol{\beta})$	.904 (.009)	.902 (.009)		.947 (.007)	.937 (.008)		.994 (.002)	.980 (.004)

List of Figures

Figure 1: Kernel smoothed estimates of the marginal posterior densities of $\boldsymbol{\beta}_i$'s from MCMC with the exact joint posterior (solid line) and RBF approximations to joint posterior (dashed line), pseudoposterior (dashed-dotted line), profile posterior with and without Laplace correction (dotted and large dotted lines, respectively) for $\widehat{C}_R(.1)$ and $|\mathcal{B}_E| = 30$, for a single representative dataset.

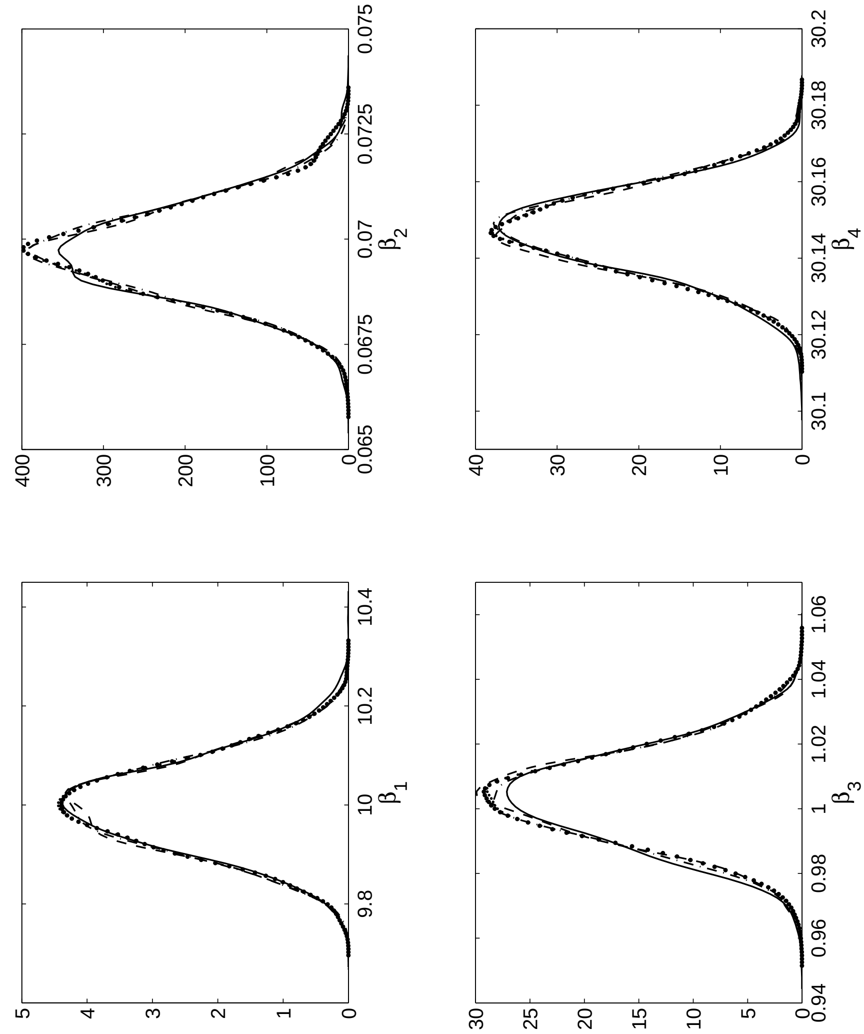
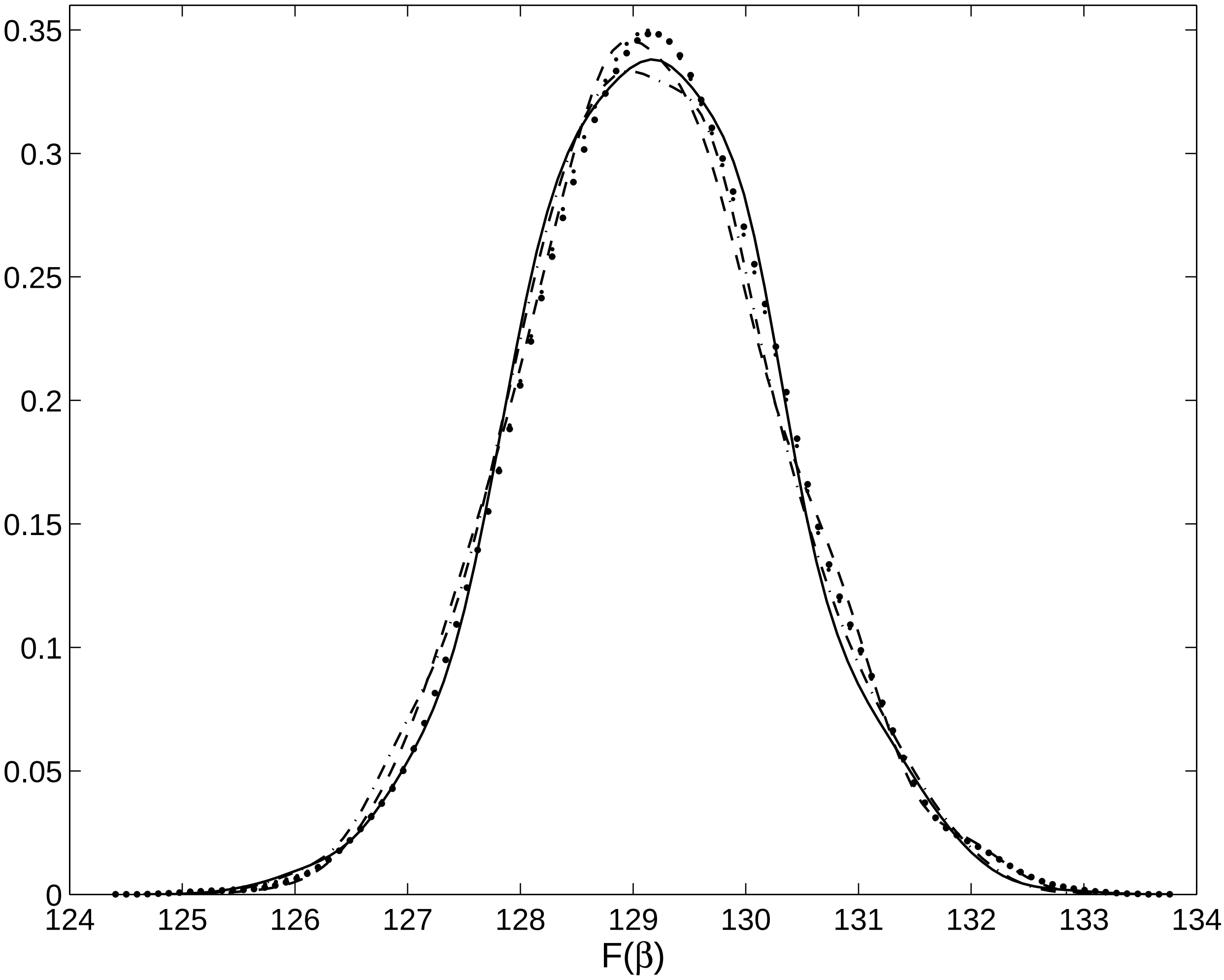


Figure 2: Kernel smoothed density estimates for the posterior of $F(\boldsymbol{\beta})$ based on the same dataset, exact and RBF surfaces and samples $\mathcal{B}_M$ as in Figure 1.


0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
124
125
126
127
128
129
130
131
132
133
134
F(β)