

Bayesian Calibration of Computationally Expensive Models Using Optimization and Radial Basis Function Approximation ^{*}

Nikolay Bliznyuk [†] David Ruppert [‡] Christine Shoemaker [§]
Rommel Regis [¶] Stefan Wild ^{||} Pradeep Mugunthan ^{**}

August 4, 2007

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 a data vector \mathbf{Y} corresponding to the regression model $\mathbf{f}(\boldsymbol{\beta})$, find plausible values of $\boldsymbol{\beta}$. As an intermediate step, \mathbf{Y} and \mathbf{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 \mathbf{Y} using MCMC. To reduce computational cost, we limit evaluation of \mathbf{f} to a small number of points chosen on a high posterior density region found by optimization. Then, we approximate the logarithm of the posterior density 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 the frequentist coverage properties of credible intervals. Our experiments

^{*}This paper should be cited as: Bliznyuk, N., Ruppert, D., Shoemaker, C.A., Regis, R., Wild, S., and Mugunthan, P. (2007), “Bayesian Calibration of Computationally Expensive Models Using Optimization and Radial Basis Function Approximation,” accepted to the *Journal of Computational & Graphical Statistics*.

[†]Nikolay Bliznyuk is a graduate student, School of Operations Research and Information 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 Information 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 Information 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 Information 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.

indicate that our method can produce results similar to those when the true “expensive” posterior density is sampled by MCMC while reducing computational costs by well over an order of magnitude.

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 (e.g., Benaman, Shoemaker and Haith (2005), Tolson and Shoemaker (2007), Tolson and Shoemaker (2007, in press), Shoemaker, Regis and Fleming (2007, in press)). Calibration means estimation of unknown parameters, for example, initial conditions or reaction and diffusion rates in a system modeled by partial differential equations. In this paper, we propose a Monte Carlo-based strategy for Bayesian calibration when the models are specified by computationally expensive computer codes, also referred to as simulators.

Our focus is on computer codes that, in a single run, produce deterministic d -dimensional output vectors $f(X, \beta)$ for all vector “indices” X in some specified set and a given parameter vector β . 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 β . We assume that one has a sample Y_1, \dots, Y_n of observation vectors in \mathbb{R}^d that correspond to the model values $f(X_1, \beta), \dots, f(X_n, \beta)$, and the goal is to make inferences about β . The vector X_i , which may contain covariates for the statistical model, is assumed to be known to the experimenter and can thus be regarded as a label for the model $f(X_i; \beta)$ 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 include the temporal instants and spatial locations where the concentrations were measured, although our methodology is applicable to a wider range of problems. Evaluating $f(X_1, \beta), \dots, f(X_n, \beta)$ for a single value of β 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). Thus accurate calibration of such models is infeasible without special methods such as those introduced here.

Given that Y_i is $f(X_i, \beta^{(0)})$ plus noise for value $\beta^{(0)}$ of β , calibration is seen to be a nonlinear regression problem. However, ordinary (nonlinear) least squares is not recommended

since practitioners often find that the variation of Y_i about $f(X_i, \boldsymbol{\beta})$ is non-normally distributed with nonconstant variance and correlated across time and space. We accommodate the non-normality 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 dependencies in the noise, we use a parametric space-time covariance model. The statistical model will be stated precisely in Section 2.

Specifying the likelihood of the data Y_1, \dots, Y_n and prior densities for parameters, we obtain the expression for the unnormalized posterior density. Even though our interest is in the models with the likelihood specified in Section 2, any alternative form of the likelihood can be used. We assume that the posterior density has a single mode in the interior of the parameter space and is differentiable twice; however, derivatives of the simulator with respect to $\boldsymbol{\beta}$ are not assumed to be given.

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) interpolant of the logarithm of the posterior density; and (4) draw a sample from the approximate posterior density using a Markov Chain Monte Carlo (MCMC) algorithm. As a result, the computational burden is reduced considerably since step (4) does not require expensive function evaluations. Using the sample from the approximate posterior distribution allows us to solve the problems of Bayesian calibration and of prediction for $F(\boldsymbol{\beta})$ by estimating moments and quantiles of the posterior distributions 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 density. However, our methodology requires far fewer evaluations of the simulator than are needed if the exact posterior density were sampled, e.g., in our application approximately 150 expensive function evaluations are used but the RBF approximation is evaluated 10,000 times.

To the best of our knowledge, this is the first investigation that uses a nonparametric approximation to the posterior density on a region of high posterior probability found by

derivative-free optimization. In Section 4.1 we introduce a new transformation family, which is more attractive from the Bayesian perspective than the usual power family and allows a systematic treatment of data transformation, which is typically carried out in an *ad hoc* fashion.

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 density 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 of computationally intensive models.

2 DESCRIPTION OF THE 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 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 j th components of Y_i and $f(X_i, \boldsymbol{\beta})$, respectively, where $j = 1, \dots, d$. Let $\{h(\cdot, \lambda) : \lambda \in \Lambda\}$ be a parametric family of differentiable increasing transformations that are indexed by λ and whose range is the real line for every λ . We assume that, for some λ_j , $h(Y_{i,j}, \lambda_j)$ is distributed $N[h\{f_j(X_i, \boldsymbol{\beta}), \lambda_j\}, \sigma_j^2]$, where σ_j is constant as a function of X_i ; later in this section we discuss a possible extension to account for simulator inadequacy. 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 4.1 we describe a new transformation family that we have used in our application.

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

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

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

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

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

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

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

where $J_h(\mathbf{Y}, \boldsymbol{\lambda})$ is the Jacobian of the transformation from \mathbf{Y} to $h\{\mathbf{Y}, \boldsymbol{\lambda}\}$ and $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ belongs to a family of covariance matrices parameterized by $\boldsymbol{\theta}$. 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, $\|x\|_{\mathbf{A}}^2 = x^\top \mathbf{A}x$.

Define the noise vectors $\epsilon_i = (\epsilon_{i,1}, \dots, \epsilon_{i,d})^\top = h\{Y_i, \boldsymbol{\lambda}\} - h\{f(X_i, \boldsymbol{\beta}), \boldsymbol{\lambda}\}$ for $i = 1, \dots, n$, $\epsilon_{\bullet,j} = (\epsilon_{1,j}, \dots, \epsilon_{n,j})^\top$ for $j = 1, \dots, d$, and $\boldsymbol{\epsilon} = (\epsilon_1^\top, \dots, \epsilon_n^\top)^\top$. The covariance between $\epsilon_{i,j}$ and $\epsilon_{i',j'}$ is modeled parsimoniously using a separable covariance function of the form $\mathbf{C}_{j,j'} \cdot \rho_{ST}(X_i, X_{i'}; \boldsymbol{\gamma})$, where \mathbf{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 $\boldsymbol{\gamma}$. Let $\mathbf{S}(\boldsymbol{\gamma})$ be the $n \times n$ space-time correlation matrix with $\mathbf{S}_{i,i'}(\boldsymbol{\gamma}) = \rho_{ST}(X_i, X_{i'}; \boldsymbol{\gamma})$. Then $\text{Var}\{\epsilon_{\bullet,j}\} = \mathbf{C}_{j,j} \cdot \mathbf{S}(\boldsymbol{\gamma})$ and, more generally, $\text{Var}\{\boldsymbol{\epsilon}\} = \boldsymbol{\Sigma}(\boldsymbol{\theta}) = \mathbf{S}(\boldsymbol{\gamma}) \otimes \mathbf{C}$, where $\boldsymbol{\theta} = (\boldsymbol{\gamma}, \mathbf{C})$ and \otimes denotes the Kronecker product of two matrices.

In equation (1) we assume that the Gaussian noise term $\epsilon_{i,j}$ is the sum of a model misspecification error (model inadequacy function) and the observation error, which is in the spirit of Higdon, Lee and Holloman (2003) and Craig, Goldstein, Rougier and Seheult (2001). In general, it is impossible to separate these two types of errors, and only their sum is identified. Only with additional assumptions can the individual errors be identified. For example, Kennedy and O'Hagan (2001) assume that the observation errors are independent.

In this case, if one assumes that the model misspecification errors are a continuous Gaussian process, then the observation error is a nugget effect and can be identified. Whether the observation errors are independent will, of course, be application-specific. In some cases, it will be not clear whether a specific type of error should be considered “observation error” or model inadequacy. For example, in our current work with a stream runoff model, we have observed that often, after a large rainfall event, the residuals are consistently either positive or negative. This pattern is not surprising, since there are large sampling errors when rainfall is estimated from a few gauges. Sampling error in rainfall could be called observation error or model inadequacy, depending on one’s viewpoint. In fact, we would rather consider it a third type of error, measurement error in a covariate (rainfall). (In our notation, covariates are included in X .) Because of the difficulties in identifying the different sources of error, we only model their sum. Hence, $\Sigma(\boldsymbol{\theta})$ is the sum of the various covariance matrices.

If there is evidence, a priori or from intermediate diagnostics (see, for example, Bates and Watts (1988)), that the simulator is a deficient representation of the underlying physical process that generates observations, equation (2) can be generalized as in Kennedy and O’Hagan (2001), by replacing $h\{\mathbf{f}(\boldsymbol{\beta}), \boldsymbol{\lambda}\}$ with $[h\{\mathbf{f}(\boldsymbol{\beta}), \boldsymbol{\lambda}\} + \mathbf{g}(\boldsymbol{\beta}, \boldsymbol{\eta})]$ or with $h\{\mathbf{f}(\boldsymbol{\beta}) + \mathbf{g}(\boldsymbol{\beta}, \boldsymbol{\eta}), \boldsymbol{\lambda}\}$. The vector-valued function $\mathbf{g}(\boldsymbol{\beta}, \boldsymbol{\eta})$ is the (statistical) model for the mean of the model inadequacy function that may involve X_i ’s as well as additional predictors. However, Beven (2001) cautions that “. . . experience with Monte Carlo simulations for complex environmental models used in the generalized likelihood uncertainty estimate (GLUE) . . . suggest that it may be very difficult to formulate an inadequacy function.”

Unless n is very large, computation of \mathbf{f} usually presents the main computational challenge in evaluation of the likelihood. In our algorithm of Section 3, we take advantage of this to evaluate the likelihood for multiple values of non-simulator parameters for each run of the simulator.

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 joint posterior density for $\boldsymbol{\beta}$, it may be necessary to evaluate the expensive function for additional space-time indices X_{n+1}, \dots, 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

$$\mathbf{f}^*(\boldsymbol{\beta}) = (\mathbf{f}\{\boldsymbol{\beta}\}^\top, f\{X_{n+1}, \boldsymbol{\beta}\}^\top, \dots, f\{X_{n^*}, \boldsymbol{\beta}\}^\top)^\top. \quad (3)$$

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 $\mathbf{f}(\boldsymbol{\beta})$ and then for $f(X_i, \boldsymbol{\beta})$ for all $i > n$).

3 METHODOLOGY

3.1 An Approximation to the Posterior Density

Since our procedure approximates the posterior density by an interpolant, it is subject to the curse of dimensionality. In this subsection we briefly review ways to lower the dimension of the argument of the posterior density and introduce some new notation.

Given a prior density $[\boldsymbol{\beta}, \boldsymbol{\zeta}]$, one has a posterior density

$$[\boldsymbol{\beta}, \boldsymbol{\zeta} | \mathbf{Y}] = \frac{[\mathbf{Y} | \boldsymbol{\beta}, \boldsymbol{\zeta}] \cdot [\boldsymbol{\beta}, \boldsymbol{\zeta}]}{\int [\mathbf{Y} | \boldsymbol{\beta}, \boldsymbol{\zeta}] \cdot [\boldsymbol{\beta}, \boldsymbol{\zeta}] d\boldsymbol{\beta} d\boldsymbol{\zeta}}. \quad (4)$$

As before, $\boldsymbol{\beta}$ is the argument of the simulator and $\boldsymbol{\zeta}$ is the vector of non-simulator parameters. We associate $\boldsymbol{\zeta}$ with nuisance parameters $\{\boldsymbol{\lambda}, \boldsymbol{\theta}\}$ from the previous section. The case with model inadequacy function parameters $\boldsymbol{\eta}$ can be treated similarly and is not considered.

In applications, $\boldsymbol{\beta}$ is the primary parameter and interest centers on its marginal posterior density $[\boldsymbol{\beta} | \mathbf{Y}] = \int [\boldsymbol{\beta}, \boldsymbol{\zeta} | \mathbf{Y}] d\boldsymbol{\zeta}$. It is often possible to integrate out a sub-block of parameters in $\boldsymbol{\zeta}$ either analytically, by using a conjugate family of prior densities as shown in Appendix A.4, or numerically. In what follows, let $\boldsymbol{\zeta}$ be the subvector of the remaining non-simulator parameters after the integration. Also, \mathbf{Y} is always regarded as fixed and $[\boldsymbol{\beta}, \mathbf{Y}]$ and $[\boldsymbol{\beta}, \boldsymbol{\zeta}, \mathbf{Y}]$ refer, respectively, to arbitrary unnormalized marginal and joint posterior densities.

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

Our goal is to obtain an accurate and cheap-to-evaluate nonparametric approximation to $[\boldsymbol{\beta}, \mathbf{Y}]$ or $[\boldsymbol{\beta}, \boldsymbol{\zeta}, \mathbf{Y}]$ based on a relatively small number of evaluations of \mathbf{f} . One can use

the resulting surface as a surrogate for the respective unnormalized posterior density in a M-H algorithm, as the sampler does not require specification of normalizing constants. (In Section 5, we contrast the proposed procedure with similar approaches in the literature.) When the expression for $[\boldsymbol{\beta}, \mathbf{Y}]$ is not available, we first approximate $[\boldsymbol{\beta}, \mathbf{Y}]$ heuristically. For a fixed value of $\boldsymbol{\beta}$, let $\widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta})$ be the maximizer of $[\boldsymbol{\beta}, \boldsymbol{\zeta}, \mathbf{Y}]$ with respect to $\boldsymbol{\zeta}$. One possible heuristic approximation is the *profile* posterior density

$$\pi_{\max}(\boldsymbol{\beta}, \mathbf{Y}) = \sup_{\boldsymbol{\zeta}} [\boldsymbol{\beta}, \boldsymbol{\zeta}, \mathbf{Y}] = [\boldsymbol{\beta}, \widehat{\boldsymbol{\zeta}}(\boldsymbol{\beta}), \mathbf{Y}]. \quad (5)$$

A more sophisticated *Laplace* approximation of Tierney and Kadane (1986) multiplies (5) by a correction factor. A simplification to (5), referred to as *pseudoposterior* density, 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 density $[\boldsymbol{\beta}, \boldsymbol{\zeta} | \mathbf{Y}]$. Each of these approximations to $[\boldsymbol{\beta}, \mathbf{Y}]$ attempt to avoid the more difficult task of integrating out nuisance parameters by maximization. It should be kept in mind, though, that neither the integration nor the maximization requires extra evaluations of \mathbf{f} . In the sequel, the notation $\pi(\cdot, \mathbf{Y})$ will be used to refer to any of these heuristic approximations to $[\boldsymbol{\beta}, \mathbf{Y}]$.

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

3.2 The Algorithm

In our algorithm, \mathbf{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 density 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 density has a single mode located in the interior of the parameter space and that \mathbf{f} is twice differentiable in a neighborhood $\widehat{\boldsymbol{\beta}}$, but we are currently generalizing the approach to multimodal densities. While selecting “design points” (the values of $\boldsymbol{\beta}$ at which to evaluate \mathbf{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 \mathbf{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 $\log\{[\boldsymbol{\beta}^{(0)}, \boldsymbol{\zeta}, \mathbf{Y}]\}$ 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 $\log\{\pi_{\max}(\boldsymbol{\beta}^{(0)}, \mathbf{Y})\}$ from (5). We perform this maximization using a constrained minimization routine (sequential quadratic programming) with analytical gradients and Hessians, implemented in MATLAB’s `fmincon`. Consequently, we maximize $\log\{\pi_{\max}(\boldsymbol{\beta}, \mathbf{Y})\}$ with respect to $\boldsymbol{\beta}$ to find $\widehat{\boldsymbol{\beta}}$ and then $\log\{[\widehat{\boldsymbol{\beta}}, \boldsymbol{\zeta}, \mathbf{Y}]\}$ 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 \mathbf{f} comes at low cost along with \mathbf{f} because finite differencing to estimate derivatives produces clusters of “uninformative” design points. We maximize $\log\{[\widehat{\boldsymbol{\beta}}, \boldsymbol{\zeta}, \mathbf{Y}]\}$ using publicly available software `CONDOR` described by Vanden Berghen and Bersini (2005), which implements a derivative-free trust-region algorithm `UOBYQA` of Powell (2000). Other derivative-free optimization methods could also be used in this step including those applied (without accompanying uncertainty analysis) to environmental calibration problems as discussed in the papers by Shoemaker et al. (2007, in press) and Tolson and Shoemaker (2007).

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 $[\boldsymbol{\beta}|\mathbf{Y}]$, defined as $C_R(\alpha) = \{\boldsymbol{\beta} : [\boldsymbol{\beta}, \mathbf{Y}] > \kappa(\alpha)\}$, where $\kappa(\alpha)$ is chosen so that the credible region $C_R(\alpha)$ contains the fraction $1 - \alpha$ of the mass of $[\boldsymbol{\beta}, \mathbf{Y}]$. Here α is a tuning parameter, for example, 0.05 or 0.01.

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

partitioning $\widehat{\mathbf{I}}^{-1}$ into blocks corresponding to $\boldsymbol{\beta}$ and $\boldsymbol{\zeta}$ one gets

$$\begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\zeta} \end{bmatrix} \underset{\text{approx.}}{\sim} MVN \left(\begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{\zeta}} \end{bmatrix}, \begin{bmatrix} \widehat{\mathbf{I}}^{\beta\beta} & \widehat{\mathbf{I}}^{\beta\zeta} \\ \widehat{\mathbf{I}}^{\zeta\beta} & \widehat{\mathbf{I}}^{\zeta\zeta} \end{bmatrix} \right), \text{ where} \quad (6)$$

$$\widehat{\mathbf{I}}^{-1} = \begin{bmatrix} \widehat{\mathbf{I}}_{\beta\beta} & \widehat{\mathbf{I}}_{\beta\zeta} \\ \widehat{\mathbf{I}}_{\zeta\beta} & \widehat{\mathbf{I}}_{\zeta\zeta} \end{bmatrix}^{-1} = \begin{bmatrix} \widehat{\mathbf{I}}^{\beta\beta} & \widehat{\mathbf{I}}^{\beta\zeta} \\ \widehat{\mathbf{I}}^{\zeta\beta} & \widehat{\mathbf{I}}^{\zeta\zeta} \end{bmatrix}. \quad (7)$$

Estimation of $\widehat{\mathbf{I}}$ by finite differences is wasteful as it does not produce new informative design points. We estimate $\widehat{\mathbf{I}}$ by fitting a quadratic surface to $\log\{[\boldsymbol{\beta}, \boldsymbol{\zeta}, \mathbf{Y}]\}$ 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 $\widehat{\mathbf{I}}$ and its blocks from (7) to refer solely to the estimated Hessian and its blocks.

We define

$$\widehat{C}_R(\alpha) = \left\{ \boldsymbol{\beta} : (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})^\top \left[\widehat{\mathbf{I}}^{\beta\beta} \right]^{-1} (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) \leq \chi_{p,1-\alpha}^2 \right\}, \quad (8)$$

where $\widehat{\mathbf{I}}^{\beta\beta} = \left[\widehat{\mathbf{I}}_{\beta\beta} - \widehat{\mathbf{I}}_{\beta\zeta} \cdot \widehat{\mathbf{I}}_{\zeta\zeta}^{-1} \cdot \widehat{\mathbf{I}}_{\zeta\beta} \right]^{-1}$ and $\chi_{p,1-\alpha}^2$ is the $(1 - \alpha)$ th quantile of the χ_p^2 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 $[\boldsymbol{\beta}|\mathbf{Y}]$ based on equation (6). We will use evaluations of \mathbf{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 **Step 2A**.

We remark that $\widehat{\mathbf{I}}$ is crucial for subsequent analysis. If \mathbf{H} is any square full-rank matrix such that $\widehat{\mathbf{I}}^{\beta\beta} = \mathbf{H}\mathbf{H}^\top$, for example, a Cholesky factor of $\widehat{\mathbf{I}}^{\beta\beta}$, then we apply the linear transformation \mathbf{H}^{-1} to $\boldsymbol{\beta}$ to ensure the same scale and to reduce correlation in parameters. Extra design points are chosen with respect to the 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, $\widehat{\mathbf{I}}$ is also used in the MCMC stage to define one of the scale parameters of the proposal density.

Let \mathcal{B}_O and \mathcal{B}_H be sets of values of $\boldsymbol{\beta}$ at which \mathbf{f}^* is evaluated during optimization in **Step 1** and during estimation of $\widehat{\mathbf{I}}$ 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 an

approximate *maximin* 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. (When working with $[\boldsymbol{\beta}, \boldsymbol{\zeta}, \mathbf{Y}]$, we evaluate the joint posterior density for multiple values of non-simulator parameters $\boldsymbol{\zeta}$ for each given evaluation of the simulator.) Further details and motivation 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. We typically use $\alpha \leq 0.1$ and $\alpha' = 0.01$ or 0.005 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 $l(\cdot) = \log\{\pi(\cdot, \mathbf{Y})\}$ at the design points $\mathcal{B}_D = \{\boldsymbol{\beta}^{(1)}, \dots, \boldsymbol{\beta}^{(N)}\}$ of the form

$$\widetilde{l}(\boldsymbol{\beta}) = \sum_{i=1}^N a_i \phi(\|\boldsymbol{\beta} - \boldsymbol{\beta}^{(i)}\|_2) + q(\boldsymbol{\beta}), \quad (9)$$

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 $q(\boldsymbol{\beta}) = (1, \boldsymbol{\beta}^\top) \cdot \mathbf{c}$.

Our choice of RBF approximation over alternatives was influenced by success with application of this method to related problems reported in Regis and Shoemaker (2007a and 2007b, both in press). However, other interpolation methods could be used. The closely related technique of kriging assumes that l is a realization of a Gaussian process (GP), determined by mean and covariance functions, and uses best linear prediction as a means of interpolation. The RBF interpolation model is a form of universal kriging with a generalized

(not necessarily positive definite) covariance function (Cressie 1991, sec. 4.4.5). Unlike a general RBF model, kriging allows one to use the covariance function of the GP (conditional on the process values at design points) to assess prediction uncertainty, which may be used to sequentially select design points for additional simulator runs.

In our case study with synthetic data in Section 4, we found that the RBF interpolant gave results that were virtually indistinguishable from the exact results. Nonetheless, the ability to assess the uncertainty of prediction is useful, especially in higher dimensional problems, or under other circumstances, where the RBF interpolant is likely to be less accurate. It may be possible to devise a similar measure in the case of RBF interpolation, and we intend to investigate this in the future.

Selection of extra design points \mathcal{B}_E in our implementation of RBF model is guided by the convergence results of \tilde{l} to l (Buhmann 2003, chap. 5) that suggest that the rate of convergence is governed by the maximum (over all points in $\widehat{C}_R(\alpha)$) distance from any point in $\widehat{C}_R(\alpha)$ to the closest design point in \mathcal{B}_D (coverage radius). We are not aware of similar convergence results for a kriging model, in part because the covariance function is typically chosen for other reasons. (However, see Appendix A.2 for design optimality considerations for GP interpolation.)

3.2.4 MCMC Sampling (Step 4)

In **Step 4** of the algorithm we draw an MCMC sample from the density proportional to $\tilde{\pi}(\cdot, \mathbf{Y}) = \exp\{\tilde{l}(\cdot)\}$ restricted to the approximate HPD region $\widehat{C}_R(\alpha')$, see the end of Section 3.2.2 and equation (9). This is done to prevent sampling of $\tilde{\pi}(\cdot, \mathbf{Y})$ in the low-probability regions of $[\boldsymbol{\beta}|\mathbf{Y}]$ where $\pi(\cdot, \mathbf{Y})$ is not approximated well enough. Sampling can be carried out using any MCMC algorithm that does not require the normalizing constant of $\tilde{\pi}(\cdot, \mathbf{Y})$ to be known. We work with the autoregressive Metropolis-Hastings algorithm (Tierney (1994)) that 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.*, $\boldsymbol{\beta}^c = \boldsymbol{\mu} + \boldsymbol{\rho}(\boldsymbol{\beta}^{(t)} - \boldsymbol{\mu}) + \mathbf{e}_t$, where $\boldsymbol{\mu}$ is the location parameter, $\boldsymbol{\rho}$ is the autoregressive parameter (matrix), and \mathbf{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 $\widehat{\boldsymbol{\Gamma}}^{\boldsymbol{\beta}\boldsymbol{\beta}}$. The

location parameter $\boldsymbol{\mu}$ is set to the MAP $\hat{\boldsymbol{\beta}}$. 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 evaluation of \boldsymbol{f} .

3.3 Bayesian Inference

Once the MCMC sample \mathcal{B}_M from the approximate posterior density is obtained as discussed in Section 3.2.4, inference about $\boldsymbol{\beta}$ can proceed using 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 distribution 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 distribution 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. The 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 density.

Examples of methods designed for calibration and uncertainty analysis of computationally expensive environmental models are Mugunthan et al. (2005) and Mugunthan and Shoemaker (2006), respectively. The methods in both of these papers are applied to a remediation problem at an 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.

Before starting with the details of the environmental application, it is necessary to define a new transformation for positive data that are common in science.

4.1 A New Transformation Family

Since we are modeling concentrations, we assume in our application that both the vector of observed concentrations \mathbf{Y} and the simulator $f(X, \boldsymbol{\beta})$ are positive. The usual transformation family used with the transform-both-sides method for such data is the Box-Cox family where $h_{BC}(y, \lambda)$ is $(y^\lambda - 1)/\lambda$ if $\lambda \neq 0$ and is $\log(y) = \lim_{\lambda \rightarrow 0} (y^\lambda - 1)/\lambda$ if $\lambda = 0$. In typical applications, λ takes values between 0 and 1. Lower values of λ define more concave transformations.

Notice that the requirement of Section 2 that the range of $h(\cdot, \lambda)$ is the real line does not hold for $h_{BC}(\cdot, \lambda)$ except when $\lambda = 0$. Then one needs to “truncate” the normal distribution of $\epsilon_{i,j}$ in equation (1) to the set where the inverse of $h_{BC}(\cdot, \lambda)$ is defined. Consequently, to make the expression in equation (2) a valid density, one must multiply it by a normalizing constant, whose computation is feasible only for the simplest models.

To avoid this difficulty, we propose the *CO*nvex combination of *I*ntity and *L*og (*COIL*) family defined as

$$h_C(y, \lambda) = \lambda y + (1 - \lambda) \log(y), \quad 0 < \lambda \leq 1. \quad (10)$$

As in the Box-Cox family, λ similarly controls the degree of concavity.

Our simulation experiments with the transform-both-sides method show that the entire Box-Cox family for $\lambda \in [0, 1)$ can be approximated well by our family. The empirical study of the *COIL* family, including its generalizations to more concave transformations, will be reported in a separate paper.

4.2 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 one point 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 $s \geq 0$):

$$C(s, t; M, D, L, \tau) = \frac{M}{\sqrt{4\pi Dt}} \exp\left[\frac{-s^2}{4Dt}\right] + \frac{M}{\sqrt{4\pi D(t-\tau)}} \exp\left[\frac{-(s-L)^2}{4D(t-\tau)}\right] \cdot \mathbb{I}(\tau < t), \quad (11)$$

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 $s_j = 0, 0.5, 1, 1.5, 2.5$ record 200 concentration readings at times $t_k = 0.3, 0.6, \dots, 50.7, 60$. The corresponding expensive model function is $\mathbf{f}(\boldsymbol{\beta}) = \{f(X_1, \boldsymbol{\beta}), \dots, f(X_{1000}, \boldsymbol{\beta})\}^\top$, 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 $\{(3, 40), (3, 45), \dots, (3, 140)\}$. As dis-

cussed in Section 2, the expensive model of equation (3) that we evaluate is

$$\mathbf{f}^*(\boldsymbol{\beta}) = (\mathbf{f}\{\boldsymbol{\beta}\}^\top, f\{(3, 40), \boldsymbol{\beta}\}, \dots, f\{(3, 140), \boldsymbol{\beta}\})^\top.$$

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 \mathbf{Y} of observed concentrations is generated from models of equations (2) and (11) with $\lambda = 0.333$ for the COIL family given by (10) and the values of $\boldsymbol{\beta}_i$'s and respective parameter spaces (domains) given in Table 1. The likelihood from equation (2) is discontinuous since $C(s_i, t_j; \boldsymbol{\beta})$ in equation (11) 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 \mathbf{Y} are independent, with variance of $h(Y_i, \lambda)$ for every i equal to the sample variance of $h\{f(X_1, \boldsymbol{\beta}), \lambda\}, \dots, h\{f(X_{1000}, \boldsymbol{\beta}), \lambda\}$, 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 density on $(\boldsymbol{\beta}, \lambda)$ over the parameter spaces mentioned earlier and an overdispersed inverse-gamma prior density 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 $[\boldsymbol{\beta}, \lambda, \mathbf{Y}]$ is the unnormalized posterior density from which we derive $\pi(\cdot, \mathbf{Y})$ as discussed in Section 3.1.

4.3 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 $(\hat{\boldsymbol{\beta}}, \hat{\lambda})$ of $[\boldsymbol{\beta}, \lambda, \mathbf{Y}]$ was found by CONDOR via maximization of $\pi_{\max}(\cdot, \mathbf{Y})$ given by equation (5), and $\hat{\mathbf{I}}$ 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 $\widehat{\mathbf{I}}$ 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 $\widehat{\mathbf{I}}^{-1}$, indicate that $[\boldsymbol{\beta}|\mathbf{Y}]$ is likely to be close to the conditional density $[\boldsymbol{\beta}|\lambda = \widehat{\lambda}, \mathbf{Y}]$.

Experiments were run for elliptical design regions $\widehat{C}_R(\alpha)$ given by equation (8), with α in $\{0.2, 0.1, 0.05, 0.01\}$ and numbers of extra experimental design points ($|\mathcal{B}_E|$) in $\{0, 10, 20, \dots, 100\}$. It was observed that the estimates of the posterior densities based on MCMC samples from approximate surfaces are not very sensitive to the volume 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 $\widehat{C}_R(0.1)$ region with 30 extra design points for $\boldsymbol{\beta}$, the same for every surface. In the case of RBF interpolation of $\log\{[\boldsymbol{\beta}, \lambda, \mathbf{Y}]\}$, 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 $\{(\boldsymbol{\beta}^{(i)}, \lambda_{ij}) \text{ for } i = 1, \dots, N \text{ and } j = 1, \dots, 10\}$, chosen as outlined in Appendix A.2. All tabular summaries pertain to this RBF approximation to $[\boldsymbol{\beta}, \lambda, \mathbf{Y}]$ and the true surface $[\boldsymbol{\beta}, \lambda, \mathbf{Y}]$. All results are reported for the autoregressive M-H sampler with $\rho = -0.25$, 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.

Samples from the approximate posterior distribution 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 (9) and then evaluating the resulting interpolant at the MCMC samples from the RBF approximations to the pseudoposterior, to the profile posterior (with and without Laplace correction) and to the joint posterior densities; see Section 3.3 for a discussion. Likewise, the sample from the true posterior distribution of $F(\boldsymbol{\beta})$ was obtained by evaluating F at the sample from $[\boldsymbol{\beta}|\mathbf{Y}]$.

Figure 1 presents plots of the *differences between sample quantiles* of the components of $\boldsymbol{\beta}$ based on MCMC samples from the approximate posterior surfaces and the corresponding

sample quantiles based on an MCMC run using the exact joint posterior surface (ordinate) *against the sample quantiles* of the components of $\boldsymbol{\beta}$ based on an MCMC run using the exact joint posterior surface (abscissa). Figure 2 overlays similar plots based on sample quantiles for the exact and approximate posterior distributions of $F(\boldsymbol{\beta})$. (MCMC samples of length 30,000, rather than 10,000, were used for all plots to reduce the variability in the estimates of tail quantiles.) Comparing the magnitudes of the differences between the sample quantiles to the respective interquartile range or to some other measure of dispersion, one can appreciate the accuracy of these RBF approximations. The plots of the differences between the sample quantiles appeared the most informative to us because the q-q plots of the MCMC sample quantiles from the RBF approximation against those from the exact surface looked like a straight line with slope 1.

Overlaid plots of kernel density estimates (not reported in this paper) for the marginal densities of the components of $\boldsymbol{\beta}$ (and similar plots for $F(\boldsymbol{\beta})$) using MCMC samples from the exact joint posterior density and the approximate posterior densities showed close agreement between the estimates of the exact and approximate densities. Striking similarities between the exact and RBF results in Table 1, Figure 1 and 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.

For each of the 1000 replications of \mathbf{Y} under the same model and parameter values but a different realization of noise, we found the MAP and $\hat{\mathbf{I}}$. We then took an MCMC sample of size 10,000 using the true surface $[\boldsymbol{\beta}, \lambda, \mathbf{Y}]$ and the respective RBF surface with approximate HPD region $\hat{C}_R(0.1)$ and the number of extra experimental design points $|\mathcal{B}_E| = 30$. Table 2 reports the observed coverage proportions of components of $\boldsymbol{\beta}$ by symmetric credible intervals of sizes 0.9, 0.95 and 0.99, along with the 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.

These results show that the coverage properties and lengths of the credible intervals based on exact and approximate surfaces are similar. The tables also suggest that the approximate credible intervals can serve as frequentist confidence intervals, as the observed coverage proportions are close to the nominal confidence coefficients.

5 DISCUSSION

5.1 Survey of Literature

Most of the literature dealing with Bayesian calibration of complex computer models focuses on reducing the number of expensive function evaluations via approximation of \mathbf{f} .

Papers can be roughly divided into two groups. In the first group, 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 and accuracy. (In Kennedy and O’Hagan (2001), the unobservable physical model, approximated by a complex code, plays the role of the top-level code.) Craig et al. (2001) model the unobserved physical process that generates measurements \mathbf{Y} as a sum of simulator and inadequacy functions, the latter assumed to have mean zero and a covariance matrix determined by an expert. Goldstein and Rougier (2004, 2006) develop a logical framework for inference about the physical system using multiple simulators (some of them hypothetical) of different quality. In each of these three papers, the simulator \mathbf{f} is approximated component-wise. We remark that so long as the likelihood of the data \mathbf{Y} can be evaluated, our algorithm of Section 3.2 can be applied to any of the models from these papers.

In the second group of papers, Higdon, Lee and Holloman (2003) 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. Christen and Fox (2005) use a cheap-to-evaluate approximation to the unnormalized posterior density, to evaluate the expensive posterior density only for the MCMC moves that are likely to be accepted. The emphasis, however, is on the models for which approximation to the posterior density is obtained by replacing \mathbf{f} by an approximation, for example, by linearization. The approach of Rasmussen (2003) uses a GP interpolant of the logarithm of the posterior density and of its first derivatives to generate proposal states for a Hybrid Monte Carlo algorithm. Each of these three papers requires at least one evaluation of the expensive posterior density for each accepted state.

5.2 Differences from Earlier Approaches

The route we take is significantly different from those mentioned. First, we do not use coarse and inexpensive versions of the expensive code, because in our experience these often

do not exist. Second, given our interest in the sample from the posterior density for β , we approximate the (scalar-valued) posterior density *directly*, and not through approximation of the high-dimensional model output f . Third, we realize that in many problems sampling thousands of times from the exact posterior surface is not computationally feasible and thus work solely with the approximation, unlike Higdon et al. (2003) and Christen and Fox (2005).

By virtue of working with GP interpolants (see our Section 3.2.3), the paper of Rasmussen (2003) has a number of similarities with ours, although our attention is not restricted to Hybrid Monte Carlo. The main requirement that the two approaches share is that the logarithm of the posterior density must be approximated well on a HPD region. However, in order to make a GP or RBF approximation strategy practical, one needs to resolve several issues, which are not addressed in Rasmussen’s work. First, since interpolation suffers from the curse of dimensionality, it makes sense to separate explicitly the argument of the posterior distribution into simulator (β) and non-simulator (ζ) parameters. When evaluation of f is the main computational bottleneck, it is beneficial to evaluate the posterior density for multiple values of ζ for the same β , which can increase the number of design points by orders of magnitude. Second, the sequential design procedure of Rasmussen does not guarantee that the whole HPD region is covered when the number of allowed runs of f is fixed and small. We avoid this problem by defining the design region explicitly. Third, any sampler drawing from an approximate posterior density must be restricted to the region where the approximation is good (the approximate HPD region); otherwise, a large mass of the proposal density may be in a region of low probability under the exact posterior distribution. Fourth, if one attempts to generate variates from the exact posterior distribution, the computational budget needs to be split in advance into parts for approximation and sampling. If the approximation is accurate, there is not much extra benefit from evaluating the exact posterior density in the MCMC run, as Rasmussen’s first example shows; otherwise, the sampler will be wasting simulator runs for the states rejected by M-H. Since accurate approximation is possible only at the expense of the MCMC run length, we allocate the whole budget to interpolation and sample only the approximate density. With our approach, optimality of placement of design points can (potentially) be enforced, whereas, even if design points from the sampling stage are re-used to improve the approximation (which Rasmussen does not consider), one has no control of their placement. Our work addresses all of these concerns.

5.3 Other Considerations (Limitations and Extensions)

A potential weakness of our algorithm is its reliance on the quadratic approximation of the logarithm of the posterior density used to define the design region. If the MAP happens to lie on the boundary of the parameter space, it will not be possible to obtain an approximate HPD region using equation (8); however, encountering this situation is likely to be a sign of a misspecified model or parameter space. If there is extreme skewness in the posterior density, then an asymmetric approximation to a HPD region is expected to be superior to our elliptical region.

In order to ensure that the design region $\widehat{C}_R(\alpha)$ covers the true HPD region adequately, the parameter α that controls the size of the region should be tuned in practice. Starting with a “smaller” $\widehat{C}_R(\alpha_0)$ and design points on it, let $\alpha_1 < \alpha_0$ and choose additional design points in the region $\widehat{C}_R(\alpha_1)$ that contains $\widehat{C}_R(\alpha_0)$. Then MCMC samples from the two approximate densities restricted to the respective design regions can be obtained and compared by means of a distribution test, controlling for dependence. Ideally, one would continue to “grow” the design region until a “discrepancy” between consecutive MCMC runs becomes small. To allow more general region shapes, one would start with an initial (e.g., elliptical) region and “grow” it outwards in the directions where the RBF surface is highest using the feedback from preceding MCMC runs. This is one of the lines of our current work.

Using an MCMC sample from the approximate posterior surface restricted to a HPD region is likely to produce accurate estimates of non-extreme quantiles, but estimated moments may be misleading. We are not aware of work on approximation of expensive models that resolves this issue.

In applications, the component-wise output of \mathbf{f} may be discontinuous on a very fine scale due to discretization, e.g., when an appropriate system of differential equations is solved numerically. Theoretical convergence results for the corresponding exact posterior densities are not applicable. Nevertheless, our approach still captures the shape of the true surface – by maintaining separation of design points we are essentially interpolating a smooth version of the exact posterior density.

Consider two methods for approximating the posterior density. The *direct method*, which we use, approximates the logarithm of the posterior density itself. The *indirect method* approximates each component of \mathbf{f} and plugs these into the logarithm of the posterior

density. One advantage of the direct method is that it approximates the scalar-valued log-posterior surface, whereas the indirect method must approximate a surface whose dimension can be quite high, e.g., 1,000 in the example in Section 4—although this is a synthetic example, it is typical of many actual applications. With the indirect method, there is an interpolation error for each component of \mathbf{f} , and the cumulative effect of component-wise approximation errors is unclear but could be large. If the indirect method is applied by modeling \mathbf{f} as a multivariate GP, then specification of the cross-covariance matrix requires substantial subject-matter expertise, whereas interpolation of the log-posterior surface by the direct method can be automated.

A referee asked whether, because the direct method does not interpolate \mathbf{f} , one can check the goodness-of-fit of the model. For diagnostics, one would normally use the simulator output only at a single point estimate of $\boldsymbol{\beta}$, say the MAP $\hat{\boldsymbol{\beta}}$, to compute residuals. This value $\mathbf{f}(\hat{\boldsymbol{\beta}})$ is known from **Step 1** of the algorithm, so no extra computation is required.

5.4 Summary and Conclusions

This paper presented a Bayesian calibration method suitable when the allowed number of evaluations of the computationally expensive simulator \mathbf{f}^* is relatively small and no inexpensive approximation to it is available. Sampling the exact posterior distribution many thousands of times during an MCMC run is questionable, if feasible, under such restrictions.

The main contribution is the algorithm of Section 3.2, which re-uses a subset of well-separated design points from a derivative-free optimization search (**Step 1**), augmented with additional design points (**Step 2**), to build an RBF approximation for the posterior density on the region of high posterior probability (**Step 3**). This allows one to draw arbitrarily long samples from the cheap proxy to the true expensive posterior density in **Step 4**. Derivative-based optimization routines that use finite differences are undesirable for **Step 1** since they produce clusters of nearby design points that carry little new information about the log-posterior surface once the surface value at any one of these points is known. Furthermore, all points in each cluster cannot be re-used in the RBF interpolation without creating numerical instability in the linear system (12) of Appendix A.3.

In our experiments presented in Section 4, a very accurate approximation to the exact posterior density was obtained, on average, using 150 runs of the simulator (**Step 1** and **Step 2**). The computational effort of our approach is well over an order of magnitude

below that required to carry out several thousands of steps in an MCMC run using the exact posterior density. 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 \mathbf{f} under less restrictive smoothness assumptions, posterior densities 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. Under the GP model, we have devised and are currently studying the properties of algorithms to sample from the densities determined by the individual realizations of the conditional GP and to integrate out the uncertainty due to approximation by MCMC. 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 Bliznyuk 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 Mungunthan 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 $\hat{\mathbf{I}}$

Let $p = \dim(\boldsymbol{\beta})$ and $u = \dim(\boldsymbol{\zeta})$. To approximate the Hessian $\hat{\mathbf{I}}$ of $-\log\{[\boldsymbol{\beta}, \boldsymbol{\zeta}, \mathbf{Y}]\}$ at $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\zeta}})$ by forward differences generally requires (around) $(p + u + 1)(p + u + 2)/2$ function evaluations. Partition $\hat{\mathbf{I}}$ as in equation (7). In our problem, $\hat{\mathbf{I}}_{\boldsymbol{\zeta}\boldsymbol{\zeta}}$ can be found analytically, and the off-diagonal blocks of $\hat{\mathbf{I}}$ can be computed entirely from the evaluations of \mathbf{f} used to

estimate the diagonal of $\widehat{\mathbf{I}}_{\beta\beta}$.

Unfortunately, the $(p+1)(p+2)/2$ points for evaluation of \mathbf{f} by finite differences to compute $\widehat{\mathbf{I}}_{\beta\beta}$ are very close to $\widehat{\beta}$ 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 the approximation $\log\{[\beta, \widehat{\zeta}(\widehat{\beta}), \mathbf{Y}]\} \approx \text{const} - \frac{1}{2}\|\beta - \widehat{\beta}\|_{\widehat{\mathbf{I}}_{\beta\beta}}^2$ on the ellipsoid $\mathcal{E}(c) = \{\beta : \|\beta - \widehat{\beta}\|_{\widehat{\mathbf{I}}_{\beta\beta}}^2 \leq c^2\}$ 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 $\widehat{\mathbf{I}}_{\beta\beta}$ 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 \mathbf{e}_i the i th standard basis vector for \mathbb{R}^p and notice that the boundary of $\mathcal{E}(c)$ passes through points $\widehat{\beta} \pm \mathbf{b}_i$, where $\mathbf{b}_i = \mathbf{e}_i \cdot c / \sqrt{\widehat{\mathbf{I}}_{\beta\beta}(i, i)}$ and $\widehat{\mathbf{I}}_{\beta\beta}(i, i)$ is the i th diagonal entry of $\widehat{\mathbf{I}}_{\beta\beta}$. The convex hull of these points

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

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 $\widehat{\mathbf{I}}_{\beta\beta}$ by forward differences using at most $2 \cdot p$ extra function evaluations, half of which are reused in computation of the off-diagonal blocks $\widehat{\mathbf{I}}_{\beta\zeta}$. (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 $[\beta | \zeta = \widehat{\zeta}, \mathbf{Y}]$. 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 $\widehat{\beta}$. We had some success with the following procedure to estimate $\widehat{\mathbf{I}}$ and to produce the set \mathcal{B}_H from Section 3.2.2:

1. Initialization:

- (a) Choose a moderate initial value of c , say, $\sqrt{\chi_{p,0.1}^2}$.
- (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 β for which \mathbf{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 \mathbf{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 $\hat{\mathbf{I}}$. If the resulting estimate of $\hat{\mathbf{I}}$ (not only that of $\hat{\mathbf{I}}_{\beta\beta}$) is not positive definite, reduce c and return to the previous step. Otherwise terminate; return $\hat{\mathbf{I}}$ 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 \mathbf{f}^* has been evaluated previously (“fixed points”). This objective is related to the *maximin* criterion that attempts to *maximize* the *minimum* 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).

As we remarked in Section 3.2.3, ideally we would like a design that minimizes the coverage radius of design points (minimax design). Johnson, Moore and Ylvisaker (1990) argue that minimax design minimizes maximum prediction variance in GP interpolation when intersite correlation is low, thereby linking optimality of designs for GP and RBF interpolation. As a heuristic, we bound the minimax distance by the intersite distance of the optimal maximin design, and then find an approximate maximin design.

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 j th 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 $\log\{[\boldsymbol{\beta}, \boldsymbol{\zeta}, \mathbf{Y}]\}$, 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 $[\boldsymbol{\zeta}|\boldsymbol{\beta} = \boldsymbol{\beta}^{(i)}, \mathbf{Y}]$, derived from equation (6). As a consequence, the increase in the dimension of the argument of the posterior density (going from $\pi(\cdot, \mathbf{Y})$ to $[\boldsymbol{\beta}, \boldsymbol{\zeta}, \mathbf{Y}]$) in this problem need not translate into the increase in the number of evaluations of 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 $q(\boldsymbol{\beta}) = (1, \boldsymbol{\beta}^\top) \cdot \mathbf{c}$. 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 $i, j = 1, \dots, N$. Let $\mathbf{P} \in \mathbb{R}^{N \times (p+1)}$ be the matrix with $(1, \{\boldsymbol{\beta}^{(i)}\}^\top)$ as the i th row for $i = 1, \dots, N$. The coefficients for the RBF surface that interpolates $l(\cdot) = \log(\pi(\cdot, \mathbf{Y}))$ at the points $\boldsymbol{\beta}^{(1)}, \dots, \boldsymbol{\beta}^{(N)}$ are obtained by solving the system

$$\begin{pmatrix} \boldsymbol{\Phi} & \mathbf{P} \\ \mathbf{P}^\top & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{a} \\ \mathbf{c} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mathcal{L}} \\ \mathbf{0} \end{pmatrix}, \quad (12)$$

where $\boldsymbol{\mathcal{L}} = [l(\boldsymbol{\beta}^{(1)}), \dots, l(\boldsymbol{\beta}^{(N)})]^\top$, $\mathbf{a} \in \mathbb{R}^N$ and $\mathbf{c} \in \mathbb{R}^{p+1}$.

The coefficient matrix in equation (12) is invertible if and only if the rank of \mathbf{P} is $p + 1$ (Powell 1992). For the case of a cubic RBF with a linear tail, this holds if and only if the set of (distinct) design points contains $p + 1$ points that are *affinely independent*. For stability purposes, we solve equation (12) by means of matrix factorizations, as described in Powell (1996).

A.4 Details on Integrating \mathbf{C} out

Let $\mathbf{Z} = \mathbf{Z}(\boldsymbol{\beta}, \boldsymbol{\lambda})$ be the matrix with the i th row $[h\{Y_i, \boldsymbol{\lambda}\} - h\{f(X_i, \boldsymbol{\beta}), \boldsymbol{\lambda}\}]^\top$ for $i = 1, \dots, n$, \mathbf{R} be the upper-triangular Cholesky factor of $\mathbf{S}(\boldsymbol{\gamma})$ and $\widetilde{\mathbf{Z}} = \mathbf{R}^{-\top} \mathbf{Z}$. Notice that, under the separable covariance model $\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \mathbf{S}(\boldsymbol{\gamma}) \otimes \mathbf{C}$ of Section 2, the likelihood equation (2) implies

that the rows $\tilde{\mathbf{Z}}_{1,\bullet}, \dots, \tilde{\mathbf{Z}}_{n,\bullet}$ of $\tilde{\mathbf{Z}}$ are *i.i.d.* $MVN(\mathbf{0}, \mathbf{C})$. Notice that

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

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

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

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

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

References

- [1] Bates, D.M., and Watts, D.G. (1988), *Nonlinear Regression Analysis and Its Applications*. New York: Wiley.
- [2] Benaman, J., Shoemaker, C. A., and Haith, D. A. (2005), “Calibration and Validation of Soil and Water Assessment Tool on an Agricultural Watershed in Upstate New York,” *ASCE Journal of Hydrologic Engineering*, 10, 363–374.
- [3] Beven, K. (2001), discussion of Kennedy and O’Hagan (2001), “Bayesian Calibration of Computer Models,” *Journal of the Royal Statistical Society, Series B*, 63, 456.
- [4] Box, G. E. P., and Cox, D. R. (1964), “An Analysis of Transformations,” *Journal of the Royal Statistical Society, Series B*, 26, 211–246.

- [5] Buhmann, M. D. (2003), *Radial Basis Functions*, New York: Cambridge University Press.
- [6] Carroll, R. J., and Ruppert, D. (1984), “Power Transformation When Fitting Theoretical Models to Data,” *Journal of the American Statistical Association*, 79, 321–328.
- [7] ————— (1988), *Transformation and Weighting in Regression*, New York: Chapman & Hall.
- [8] Christen, J. A., and Fox, C. (2005), “Markov Chain Monte Carlo Using an Approximation,” *Journal of Computational & Graphical Statistics*, 14, 795–810.
- [9] Craig, P. S., Goldstein, M., Rougier, J., and Seheult, A. H. (2001), “Bayesian Forecasting for Complex Systems Using Computer Simulators,” *Journal of the American Statistical Association*, 96, 717–729.
- [10] Cressie, N. (1991), *Statistics for Spatial Data*, New York: Wiley.
- [11] Devroye, L. (1986), *Non-Uniform Random Variate Generation*, New York: Springer-Verlag.
- [12] Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004), *Bayesian Data Analysis* (2nd ed.), Boca Raton: Chapman & Hall/CRC.
- [13] Goldstein, M., and Rougier, J. C. (2004), “Probabilistic Formulations for Transferring Inferences from Mathematical Models to Physical Systems,” *SIAM Journal on Scientific Computing*, 26, 467–487.
- [14] Goldstein, M. and Rougier, J. (2006), “Reified Bayesian Modelling and Inference for Physical Systems”. Submitted to the *Journal of Statistical Planning and Inference*, available at <http://www.maths.dur.ac.uk/stats/people/jcr/Reify.pdf>
- [15] Higdon, D., Lee H., and Holloman, C. (2003), “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.

- [16] Johnson, M., Moore, L., and Ylvisaker, D. (1990), “Minimax and Maximin Distance Designs,” *Journal of Statistical Planning and Inference*, 26, 131–148.
- [17] 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.
- [18] ——— (2001), “Bayesian Calibration of Computer Models,” *Journal of the Royal Statistical Society, Series B*, 63, 425–464.
- [19] Morris, M., Mitchell, T., and Ylvisaker, D. (1993), “Bayesian Design and Analysis of Computer Experiments: Use of Derivatives in Surface Prediction,” *Technometrics*, 35, 243–255.
- [20] Mugunthan, P. and Shoemaker, C. A. (2006), “Assessing the Impacts of Parameter Uncertainty for Computationally Expensive Groundwater Models,” *Water Resources Research*, 42, W10428, doi: 10.1029/2005WR004640.
- [21] 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.
- [22] 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.
- [23] 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.
- [24] ——— (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.
- [25] ——— (2002), “UOBYQA: Unconstrained Optimization by Quadratic Approximation,” *Mathematical Programming*, 92, 555–582.

- [26] Rasmussen, C.E. (2003), “Gaussian Processes to Speed up Hybrid Monte Carlo for Expensive Bayesian Integrals,” in *Bayesian Statistics 7*, eds. J. M. Bernardo, J. O. Berger, A. P. Berger and A. F. M. Smith, pp. 651–659.
- [27] Regis, R. G., and Shoemaker, C. A. (2007a, in press), “A Stochastic Radial Basis Function Method for the Global Optimization of Expensive Functions,” *INFORMS Journal of Computing*.
- [28] Regis, R. G., and Shoemaker, C. A. (2007b, in press), “Parallel Radial Basis Function Methods for the Global Optimization of Computationally Expensive Functions,” *European Journal of Operations Research*
- [29] Santner, T.J., Williams, B.J. and Notz, W. (2003), *The Design and Analysis of Computer Experiments*. New York: Springer–Verlag.
- [30] Shoemaker, C., Regis, R., and Fleming, R. (2007, in press), “Watershed Calibration Using Multistart Local Optimization and Evolutionary Optimization with Radial Basis Function Approximation,” *Journal Of Hydrologic Science*.
- [31] Tierney, L. (1994), “Markov Chains for Exploring Posterior Distributions,” *The Annals of Statistics*, 22, 1701–1786.
- [32] Tierney, L., and Kadane, J. B. (1986), “Accurate Approximations for Posterior Moments and Marginal Densities,” *Journal of the American Statistical Association*, 81, 82–86.
- [33] Tolson, B., and Shoemaker, C. A. (2007), “The Dynamically Dimensioned Search Algorithm for Computationally Efficient Automatic Calibration of Environmental Simulation Models,” *Water Resources Research*, 43, W01413, doi: 10.1029/2005WR004723.
- [34] Tolson, B., and Shoemaker, C. A. (2007, in press), “Cannonsville Reservoir Watershed SWAT2000 Model Development, Calibration and Validation,” *Journal of Hydrology*.
- [35] Trosset, M. W. (1999), “Approximate Maximin Distance Designs,” in *American Statistical Association Proceedings of the Physical and Engineering Sciences Section*, pp. 223–227.

- [36] 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 $[\beta, \lambda, \mathbf{Y}]$ as the surface. The RBF approximations use, on average, 150 expensive function evaluations compared to 10,000 for the exact results.

	domain	true	MC mean		ratio of lengths of cred. int.'s		
			exact	RBF	size 0.9	size 0.95	size 0.99
β_1	[7, 13]	10	10.0057 (0.0866)	10.0061 (0.0893)	0.9969 (0.0602)	0.9961 (0.0624)	0.9844 (0.0738)
β_2	[0.02, 0.12]	0.07	0.07008 (0.00097)	0.07008 (0.00101)	0.9910 (0.0592)	0.9888 (0.0612)	0.9687 (0.0673)
β_3	[.01, 3]	1	1.0005 (0.0136)	1.0005 (0.0134)	0.9671 (0.0785)	0.9662 (0.0765)	0.9604 (0.0750)
β_4	[30.01, 30.295]	30.16	30.1610 (0.0096)	30.1610 (0.0096)	0.9786 (0.0779)	0.9709 (0.0818)	0.9403 (0.0835)
$F(\beta)$	–	128.998	129.063 (1.087)	129.067 (1.100)	0.9959 (0.062)	0.9937 (0.0628)	0.9841 (0.0695)

Table 2: Observed probabilities of coverage with (standard errors) of symmetric credible intervals based on 1000 dataset replications and joint posterior density as the surface for MCMC. The RBF approximations use, on average, 150 expensive function evaluations compared to 10,000 for the exact results.

	size 0.9 cred. int.		size 0.95 cred. int.		size 0.99 cred. int.	
	exact	RBF	exact	RBF	exact	RBF
β_1	0.905 (0.009)	0.904 (0.009)	0.950 (0.007)	0.944 (0.007)	0.986 (0.004)	0.990 (0.003)
β_2	0.908 (0.009)	0.903 (0.009)	0.954 (0.007)	0.951 (0.007)	0.991 (0.003)	0.987 (0.004)
β_3	0.916 (0.009)	0.899 (0.010)	0.953 (0.007)	0.954 (0.007)	0.989 (0.003)	0.988 (0.003)
β_4	0.904 (0.009)	0.909 (0.009)	0.947 (0.007)	0.945 (0.007)	0.988 (0.003)	0.987 (0.004)
$F(\beta)$	0.904 (0.009)	0.902 (0.009)	0.947 (0.007)	0.937 (0.008)	0.994 (0.002)	0.980 (0.004)

List of Figures

- 1 Interpolated *pairwise differences between sample quantiles* of β_i based on MCMC samples from each approximate posterior surface and the respective sample quantiles of β_i based on an MCMC run using the exact joint posterior surface (ordinate) *against the sample quantiles* of β_i based on an MCMC run using the exact joint posterior surface (abscissa). All plots are of the form (approximate minus exact) vs exact quantiles for the RBF approximations to the joint posterior (*), profile posterior with (+) and without (o) the Laplace correction and pseudoposterior (Δ) densities. Markers are placed at the $(-0.05 + 0.1 \cdot j)$ th sample quantiles for $j = 1, 2, \dots, 10$. These RBF approximations for a single representative dataset use $\hat{C}_R(0.1)$ and $|\mathcal{B}_E| = 30$. MCMC run length is 30,000. 35
- 2 Interpolated *pairwise differences between sample quantiles* of $F(\beta)$ based on MCMC samples from each approximate posterior surface and the respective sample quantiles of $F(\beta)$ based on an MCMC run using the exact joint posterior surface (ordinate) *against the sample quantiles* of $F(\beta)$ based on an MCMC run using the exact joint posterior surface (abscissa). Markers are placed at the $(-0.025 + 0.05 \cdot j)$ th sample quantiles for $j = 1, 2, \dots, 20$. The dataset, exact and RBF surfaces and samples \mathcal{B}_M , as well as the plot identifiers, are the same as in Figure 1. 36

Figure 1: Interpolated *pairwise differences between sample quantiles* of β_i based on MCMC samples from each approximate posterior surface and the respective sample quantiles of β_i based on an MCMC run using the exact joint posterior surface (ordinate) *against the sample quantiles* of β_i based on an MCMC run using the exact joint posterior surface (abscissa). All plots are of the form (approximate minus exact) vs exact quantiles for the RBF approximations to the joint posterior (*), profile posterior with (+) and without (o) the Laplace correction and pseudoposterior (Δ) densities. Markers are placed at the $(-0.05 + 0.1 \cdot j)$ th sample quantiles for $j = 1, 2, \dots, 10$. These RBF approximations for a single representative dataset use $\hat{C}_R(0.1)$ and $|\mathcal{B}_E| = 30$. MCMC run length is 30,000.

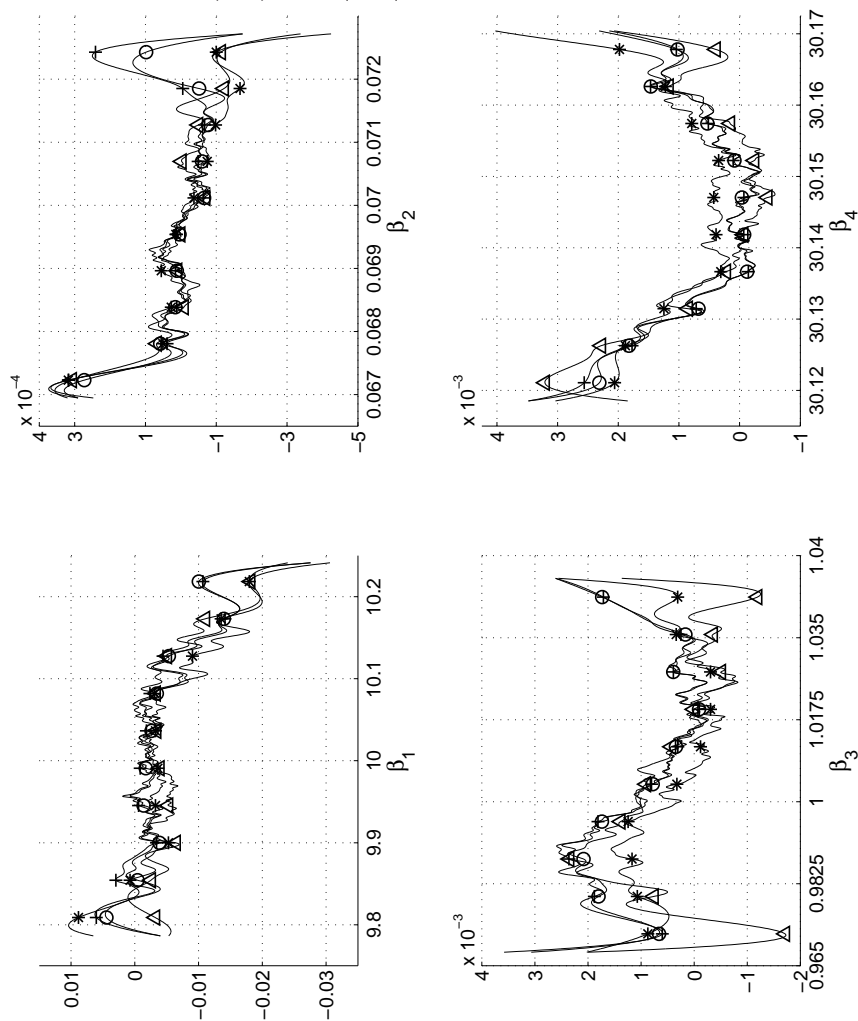


Figure 2: Interpolated *pairwise differences between sample quantiles* of $F(\beta)$ based on MCMC samples from each approximate posterior surface and the respective sample quantiles of $F(\beta)$ based on an MCMC run using the exact joint posterior surface (ordinate) *against the sample quantiles* of $F(\beta)$ based on an MCMC run using the exact joint posterior surface (abscissa). Markers are placed at the $(-0.025 + 0.05 \cdot j)$ th sample quantiles for $j = 1, 2, \dots, 20$. The dataset, exact and RBF surfaces and samples \mathcal{B}_M , as well as the plot identifiers, are the same as in Figure 1.

