A Preposterior Analysis to Predict Identifiability in Experimental Calibration of Computer Models

Paul D. Arendt
Northwestern University, Department of Mechanical Engineering
2145 Sheridan Road Room B214
Evanston, IL, 60208
Phone: 847-491-5066, Fax: 847-491-3915, Email: paularendt2012@u.northwestern.edu

Daniel W. Apley*
***Corresponding Author**
Northwestern University, Department of Industrial Engineering and Management Sciences
2145 Sheridan Road Room C150
Evanston, IL, 60208
Phone: 847-491-2397, Fax: 847-491-8005, Email: apley@northwestern.edu

Wei Chen
Northwestern University, Department of Mechanical Engineering
2145 Sheridan Road Room A216
Evanston, IL, 60208
Phone: 847-491-7019, Fax: 847-491-3915, Email:

Abstract

When using physical experimental data to adjust, or calibrate, computer simulation models, two general sources of uncertainty that must be accounted for are calibration parameter uncertainty and model discrepancy. This is complicated by the well-known fact that systems to be calibrated are often subject to identifiability problems, in the sense that it is difficult to estimate the parameters precisely and to distinguish between the effects of parameter uncertainty versus model discrepancy. We develop a form of preposterior analysis that can be used, prior to conducting physical experiments but after conducting the computer simulations, to predict the degree of identifiability that will result after conducting the physical experiments for a given experimental design. Specifically, we calculate the preposterior covariance matrix of the calibration parameters and demonstrate that, in the examples that we consider, it provides a reasonable prediction of the actual posterior covariance that is calculated after the experimental data are collected. Consequently, the preposterior covariance can be used as a criterion for designing physical experiments to help achieve better identifiability in calibration problems. Supplementary materials are available for this article at the publisher's online edition of *IIE Transaction*, datasets, additional tables, detailed proofs, etc.

Key Words: Gaussian Process Models, Kriging, Bayesian Calibration, Computer Experiments

1 Introduction

In many engineering and scientific applications, computer simulations have become an important tool for design, optimization, or simply to better understand physical phenomena (Hoffman, et al. 2003, Higdon, et al. 2004, Malhotra, et al. 2012). However, computer models never perfectly represent reality. Two general sources of uncertainty that account for differences between the computer model and physical reality are parameter uncertainty and model discrepancy (Kennedy and O'Hagan 2001). The former results from imperfect knowledge of underlying physical parameters, called calibration parameters (e.g., material properties, friction coefficients, etc.), while the latter results from approximations, missing physics, and other inaccuracies of the computer simulation and is represented by a discrepancy function [see Eq. (1), introduced in Section 2]. So-called calibration methods for learning these uncertainties via combining (usually quite abundant) simulation data with (usually more limited) physical experimental data have been developed (Kennedy and O'Hagan 2001, Higdon, et al. 2004, Reese, et al. 2004, Bayarri, et al. 2007, Higdon, et al. 2008). It should be noted that if an uncertain parameter is a distributional parameter of a random input variable, one might be able to reduce its uncertainty by directly observing a sample of the input variable realizations, as in Song and Nelson (2014). We treat the very different situation (calibration) in which the uncertain parameters are fixed, deterministic variables that cannot be observed directly, and all inference on them must be made indirectly via observations of the simulated and physical experimental response variable.

When there is no discrepancy function, calibration is typically a straightforward, statistically identifiable problem (Kumar 2008, Drignei 2009, Akkaram, et al. 2010, Huan and Marzouk 2011). However, it is well known (Loeppky, et al. 2006, Han, et al. 2009, Arendt, et al. 2012a, 2012b), that when a discrepancy function is present the estimation problem is often poorly identifiable, in the sense that it is difficult to distinguish the effects of calibration parameter uncertainty from the discrepancy function or to estimate these quantities individually. Much of the prior calibration work that has considered a discrepancy function, including work on experimental design for calibrating computer models (Kennedy

and O'Hagan 2001, Higdon, et al. 2004, Bayarri, et al. 2007, Ranjan, et al. 2011, Williams, et al. 2011), focused on the more easily attainable objective of good response prediction with the calibrated computer model, even if one is unable to accurately identify the calibration parameters and distinguish their effects from the discrepancy function.

Higdon, et al. (2004), Loeppky, et al. (2006) and Han, et al. (2009) distinguished between two different types of parameters – tuning parameters and calibration parameters – that can serve as input variables in a computer model. Tuning parameters have no meaning in the physical experiment, whereas calibration parameters do have concrete physical meaning but are unknown in reality. Examples of tuning parameters are mesh density in a finite element simulation or some constant in an empirically-postulated material flow law (e.g., Maheshwari, et al. (2010)). Han, et al. (2009) and the references therein provide a number of examples of calibration parameters that have concrete physical meaning, such as the friction between bone and prothesis in a prothestic knee simulation. They discuss the importance of identifying the true values of the calibration parameters and, in light of the identifiability challenges, the need for further research on improving identifiability.

In this paper we consider the situation in which there are unknown calibration parameters that have physical meaning, and it is desired to estimate their true values and distinguish their effects from the effects of the discrepancy function. We take the viewpoint of Han, et al. (2009) and Arendt et al. (2012b) that good calibration identifiability may be important for a number of reasons. First, learning the calibration parameters may itself be the primary goal of the calibration, e.g., for scientific discovery purposes when the parameters cannot be measured directly. Second, learning the model discrepancy function may provide insight into the deficiencies of the computer model for improving future generations of the simulation code. Third, if the calibration parameters and discrepancy function cannot be identified individually, they may still allow reasonably accurate correction of the computer model in regions of the input space near which physical experimental runs have been conducted; however, better knowledge of

the calibration parameters would likely allow a less myopic correction to the computer model that provides better prediction over a broader range of input settings.

As in much of the prior work that assessed calibration identifiability, we quantify identifiability via the posterior covariance matrix of the set of calibration parameters (posterior to observing the physical experimental data, in addition to the simulation data). One might also consider using the posterior covariance function of the discrepancy function in the measure of identifiability. However, this is infinite dimensional and cumbersome. Moreover, Arendt et al. (2012b) demonstrated that the posterior covariance of the discrepancy function is closely related to the posterior covariance of the calibration parameters, in the sense that precisely estimated calibration parameters generally imply a precisely estimated discrepancy function [which is somewhat obvious from Eq. (1) in Section 2]. Consequently, we work with the more manageable posterior covariance matrix of the calibration parameters. Here, we are using the term "identifiability" rather informally to refer to whether one can expect to achieve reasonable and useful estimation of the calibration parameters with typical finite sample sizes (which is sometimes possible), as opposed to whether estimators are consistent and guaranteed to asymptotically converge to the true values when sample sizes approach infinity in some sense (which is perhaps never possible under realistic assumptions). In Section 5.4, we discuss why the latter may not be particularly meaningful for most computer model calibration problems.

It should be noted that one view within the calibration community is that identifiability of the calibration parameters in the Kennedy and O'Hagan model is impossible, or nearly impossible. We argue in Section 5.4 that good identifiability, although sometimes impossible and usually difficult, is sometimes achievable. Moreover, after collecting the simulation and experimental data and conducting the Kennedy and O'Hagan analysis, if identifiability is poor, the posterior standard deviations produced in the analysis will indicate that this is the case. The main issue that we address in this paper is how to assess the level of identifiability *prior to conducting the physical experiment*.

Others have also demonstrated that identifiability is sometimes possible and depends on the specifics of the example (e.g., Higdon, et al. (2004), Han, et al. (2009)). Arendt et al. (2012b) showed that the degree of identifiability depends strongly on the nature of the computer model response as a function of the input variables and the calibration parameters, as well as on the prior assumptions regarding the discrepancy function (e.g., the smoothness of the discrepancy function). Computer simulations are inherently multi-response, as there are many intermediate and final response variables at many different spatial and temporal locations that are automatically calculated during the course of the simulation. Arendt, et al. (2012a) showed that for systems with poor identifiability based on a single measured physical experimental response, identifiability may be improved by measuring multiple physical responses. In all situations, identifiability obviously depends on the design of the physical experiment, i.e., the number of experimental runs and the input settings for each run.

The same can be said when estimating the parameters of any parametric model based on physical experimental data, and standard criteria for designing such physical experiments are often based on measures related to parameter identifiability. However, calibration of computer simulation models involves a fundamental distinction: Prior to designing the physical experiment, one can learn via simulation a great deal about the nature of the functional dependence of the response(s) on the input variables and on the calibration parameters, and this knowledge can be exploited when designing the physical experiment to provide better identifiability.

In order to accomplish this, one needs to predict or approximate the degree of identifiability prior to conducting the physical experiments, but after conducting the computer simulations. The primary purpose of this paper is to propose and investigate the use of the preposterior covariance matrix (Berger 1985, Carlin and Louis 2000) of the calibration parameters for this purpose and to demonstrate that it provides a reasonable prediction of the actual posterior covariance, at least for the examples that we consider. Consequently, the preposterior covariance matrix can serve as a reasonable criterion for

designing physical experiments in order to achieve good identifiability when calibrating computer simulation models.

We note that a preposterior analysis cannot improve one's knowledge of the calibration parameters, because it is conducted prior to observing any physical experimental data. However, it can provide a reasonable quantification of the uncertainty of the calibration parameters that one expects will result after observing the experimental data. This is analogous to what occurs in standard optimal design of experiments (Wu and Hamada 2000, Montgomery 2005) for fitting a response surface via linear regression. Given the design matrix and the standard deviation of the random observation error, one can calculate the covariance matrix of the estimated parameters prior to conducting the experiment and use this as the experimental design criterion. Of course, one cannot estimate the parameters until the experimental data are collected.

The format of the remainder of the paper is as follows. In Section 2 we provide a brief background on the standard formulation of the computer model calibration problem and how computer simulation and physical experimental data are combined to implement the calibration. In Section 3 we present our approach to calculate the preposterior covariance matrix of the calibration parameters. In Section 4 we discuss a modification of the preposterior analysis that provides insight into the behavior of the preposterior covariance as a means of predicting the posterior covariance. In Section 5, we use a number of examples to illustrate the effectiveness of the preposterior analysis in predicting identifiability prior to conducting physical experiments and discuss the issue of identifiability of the calibration parameters. Section 6 concludes the paper.

2 Background on Combining Computer Simulation and Physical Experimental Data for Computer Model Calibration

We use the standard computer model calibration formulation introduced in Kennedy and O'Hagan (2001) and used in many subsequent works (Higdon, et al. 2004, Loeppky, et al. 2006, Williams, et al. 2006, Bayarri, et al. 2007, Han, et al. 2009). For ease of exposition, we consider only a single response variable. The standard formulation is (Kennedy and O'Hagan 2001)

$$y^e(\mathbf{x}) = y^m(\mathbf{x}, \boldsymbol{\theta}^*) + \delta(\mathbf{x}) + \varepsilon, \quad (1)$$

where $y^e(\mathbf{x})$ denotes the physical experimental response, $y^m(\mathbf{x},\boldsymbol{\theta})$ denotes the computer model response, $\mathbf{x}$ denotes a $d \times 1$ vector of controllable (or design) variables that can be controlled in the physical experiments, and $\boldsymbol{\theta}$ denotes an $r \times 1$ vector of calibration parameters. The calibration parameters are unknown constants in reality, but in the simulations their values can be specified as inputs in the same way the values of $\mathbf{x}$ are specified. Consequently, if we let $\boldsymbol{\theta}^*$ denote the vector of true calibration parameters, then Eq. (1) expresses $y^e(\mathbf{x})$ as the summation of $y^m(\mathbf{x},\boldsymbol{\theta}^*)$, a discrepancy function $\delta(\mathbf{x})$, and a random experimental observation/replication error ε. One can take the definition of $\delta(\mathbf{x})$ to be the difference between $y^e(\mathbf{x})$ (with no random error) and $y^m(\mathbf{x},\boldsymbol{\theta}^*)$, which we write as a function of only $\mathbf{x}$. The random error ε is assumed to be zero-mean, Gaussian, and independent for different experimental observations.

To estimate $\boldsymbol{\theta}^*$ and $\delta(\mathbf{x})$ and quantify their uncertainties, we use the modular Bayesian approach of (Kennedy and O'Hagan 2001, Bayarri, et al. 2007), which is comprised of the four modules illustrated in the left panel of Figure 1, the notation in which will be explained shortly. Next we briefly describe each module. Additional discussion can be found in Kennedy and O'Hagan (2001) and Arendt et al. (2012b).

In the modular Bayesian approach, both $y^m(\mathbf{x},\boldsymbol{\theta})$ and $\delta(\mathbf{x})$ are modeled as Gaussian processes (GPs) (O'Hagan 1978, Cressie 1993, Handcock and Stein 1993). In Module 1, the hyperparameters (i.e., the prior correlation and variance parameters) of the GP model for $y^m(\mathbf{x},\boldsymbol{\theta})$ are estimated using maximum likelihood estimation (MLE), based on only the observed computer simulation data. The physical experimental data are not used in Module 1 for computational reasons and because the physical experimental data contain little further information about the hyperparameters of the GP model for $y^m(\mathbf{x},\boldsymbol{\theta})$. Next, in Module 2, the hyperparameters for the GP model for $\delta(\mathbf{x})$ are estimated using MLE based on all of the observed data, treating the MLEs from Module 1 as fixed parameters and marginalizing the likelihood over the assumed prior distribution for the calibration parameters. Subsequently, in order to quantify the calibration parameter uncertainty, Module 3 calculates the posterior

distribution of the calibration parameters based on all of the observed data and on the prior for the calibration parameters, with the MLEs from Modules 1 and 2 treated as fixed values. Lastly, Module 4 calculates a posterior distribution of the experimental response and the discrepancy function using the results of Modules 1 through 3. Thus, the modular Bayesian approach provides posterior distributions that quantify the uncertainty of the calibration parameters, the discrepancy function, and the experimental response. In the next section, we will use the first three modules of the modular Bayesian approach when calculating the preposterior covariance matrix of the calibration parameters *prior* to observing the physical experimental data, which will provide a prediction of the posterior covariance matrix that will result *after* observing the physical experimental data.

3 Calculating the Preposterior Covariance Matrix of the Calibration Parameters

We next provide an overview of our algorithm for calculating the preposterior covariance matrix, illustrated in Figure 1. Following this, we give detailed descriptions of each step in the algorithm. The algorithm begins by fitting the Gaussian process model for $y^m(\mathbf{x},\boldsymbol{\theta})$ based on the computer simulation data (Step 0a). The user then specifies the physical experimental input settings and assigns priors for the GP model of $\delta(\mathbf{x})$, for the calibration parameters, and for ε (Steps 0b and 0c). Because the preposterior analysis is conducted prior to observing actual experimental data, the main body of the algorithm is a Monte Carlo (MC) simulation (Steps 1—7) in which, on each MC replicate, we generate (Steps 1—5) a hypothetical set of physical experimental response observations at the specified input settings based on the knowledge of $y^m(\mathbf{x},\boldsymbol{\theta})$ from the computer simulation data and the relevant prior distributions for $\delta(\mathbf{x})$, $\boldsymbol{\theta}$, and ε. Then we calculate (Steps 6—7) a posterior covariance matrix for $\boldsymbol{\theta}$ for the experimental response observations generated in that MC replicate. The preposterior covariance matrix for $\boldsymbol{\theta}$ is taken to be (Step 8) the average of the posterior covariance matrices over all MC replicates. Within Steps 1—7, Modules 2 and 3 of the standard modular Bayesian approach are used. Next we describe each step of the preposterior analysis in greater detail.

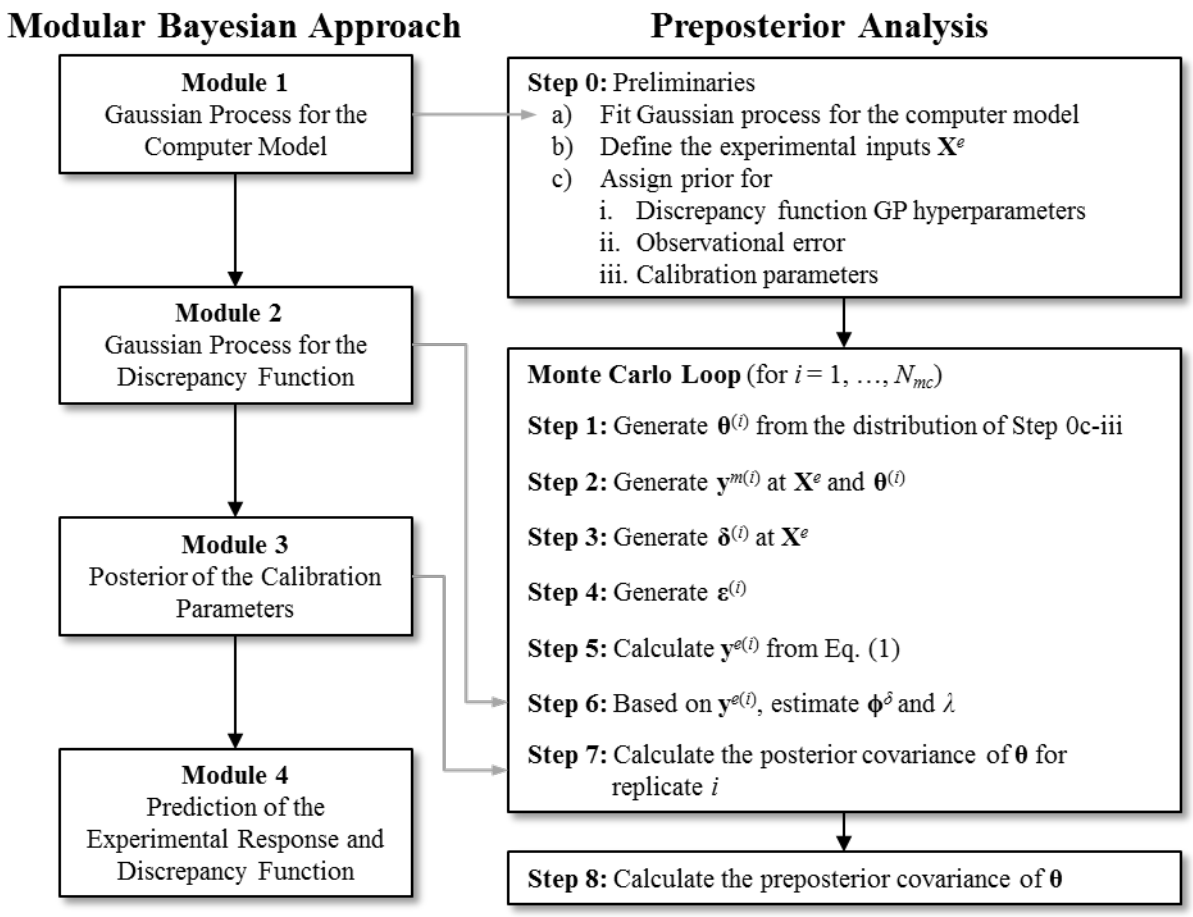

Figure 1. Flowchart for the preposterior analysis.

Step 0: Preliminaries

In this preliminary step, the GP model for $y^m(\mathbf{x},\boldsymbol{\theta})$ is fitted, and the user specifies several prior distributional quantities related to the physical experimental responses. By "fitting" the GP model for $y^m(\mathbf{x},\boldsymbol{\theta})$, we mean finding the MLEs (Schabenberger and Gotway 2005) of the hyperparameters $\boldsymbol{\phi}^m = \{\boldsymbol{\beta}^m, \sigma_m^2, \boldsymbol{\omega}^m\}$ (defined below) of the GP model based on the simulation response observations $\mathbf{y}^m = [y^m(\mathbf{x}_1^m, \boldsymbol{\theta}_1^m), \ldots, y^m(\mathbf{x}_{N_m}^m, \boldsymbol{\theta}_{N_m}^m)]^T$ at the N_m input sites $\mathbf{X}^m = [\mathbf{x}_1^m, \ldots, \mathbf{x}_{N_m}^m]^T$ and $\boldsymbol{\Theta}^m = [\boldsymbol{\theta}_1^m, \ldots, \boldsymbol{\theta}_{N_m}^m]^T$ used in the computer simulations, from which we obtain expressions for the posterior distribution of $y^m(\mathbf{x},\boldsymbol{\theta})$. This takes place in Step 0a using Module 1 of the modular Bayesian approach. In the preceding, $\mathbf{x}_i^m$

denotes the $d \times 1$ vector of inputs, and $\boldsymbol{\theta}_i^m$ denotes the $r \times 1$ vector of calibration parameters, used on the ith run of the computer simulation. In the GP model, $\mathbf{y}^m$ has mean vector $\mathbf{H}^m\boldsymbol{\beta}^m$, where $\mathbf{H}^m = [\mathbf{h}^m(\mathbf{x}_1^m)^T, \ldots, \mathbf{h}^m(\mathbf{x}_{N_m}^m)^T]^T$ with each $\mathbf{h}^m(\mathbf{x})$ a specified regression basis function and $\boldsymbol{\beta}^m$ a vector of coefficients; and $\mathbf{y}^m$ has covariance matrix $\sigma_m^2\mathbf{R}^m$, where $\mathbf{R}^m$ is the $N_m \times N_m$ matrix with kth-row, jth-column entry equal to the prior correlation $R^m(\{\mathbf{x}_j^m, \boldsymbol{\theta}_j^m\}, \{\mathbf{x}_k^m, \boldsymbol{\theta}_k^m\})$ between $y^m(\mathbf{x}_j^m, \boldsymbol{\theta}_j^m)$ and $y^m(\mathbf{x}_k^m, \boldsymbol{\theta}_k^m)$, σ_m^2 denotes the prior variance of the GP model, and $\boldsymbol{\omega}^m$ denotes a $(d + r) \times 1$ vector of prior correlation (aka, scale or roughness) parameters. For all examples in this paper, we used a constant prior mean function of unknown magnitude (i.e., $\mathbf{h}^m(\mathbf{x}) = 1$ and β^m a scalar hyperparameter to be estimated) and a Gaussian correlation function of the form $R^m(\mathbf{z}, \mathbf{z}') = exp\left(-\sum_l \omega_l(z_l - z'_l)^2\right)$, where the summation is over the elements of the vectors $\mathbf{z}$ and $\mathbf{z}'$. For numerical stability of the MLE algorithm, $\mathbf{X}^m$ and $\boldsymbol{\Theta}^m$ are transformed to the range [0, 1], and $\mathbf{y}^m$ is standardized to have a sample mean of 0 and a sample standard deviation of 1 (Rasmussen 1996). Let $\hat{\boldsymbol{\phi}}^m$ denote the MLE of $\boldsymbol{\phi}^m$.

Later, within the MC Steps 1—7, we generate a realization of $y^m(\mathbf{x},\boldsymbol{\theta})$ at various sets of $\{\mathbf{x}, \boldsymbol{\theta}\}$ values using a multivariate normal random number generator with mean and covariance taken to be the posterior mean and covariance of the GP model of $y^m(\mathbf{x},\boldsymbol{\theta})$ from Step 0a. The posterior mean and covariance are calculated using standard results from the literature (O'Hagan 1978, Rasmussen and Williams 2006). Here we use plug-in MLEs for σ_m^2 and $\boldsymbol{\omega}^m$ with a non-informative prior for $\boldsymbol{\beta}^m$. Over the remainder of the preposterior analysis, the GP model of $y^m(\mathbf{x},\boldsymbol{\theta})$ is not updated from that of Step 0a.

In Step 0b the user specifies the N_e experimental input locations $\mathbf{X}^e = [\mathbf{x}_1^e, \ldots, \mathbf{x}_{N_e}^e]^T$ at which one intends to conduct the N_e physical experimental runs (and at which the hypothetical experimental response observations will be generated within the MC loop). As described in Section 2 in the context of Module 2 of the modular Bayesian approach, we represent the discrepancy function as a GP model with hyperparameters $\boldsymbol{\phi}^\delta = \{\boldsymbol{\beta}^\delta, \sigma_\delta^2, \boldsymbol{\omega}^\delta\}$ (analogous to the hyperparameters $\boldsymbol{\phi}^m$ for the GP model of $y^m(\mathbf{x},\boldsymbol{\theta})$), to

which the user assigns prior distributions in Step 0c. This prior for $\boldsymbol{\phi}^{\delta}$ is used only in Step 3 of the MC loop, in which values of $\boldsymbol{\phi}^{\delta}$ are sampled from the specified prior before generating a hypothetical discrepancy function. Because the modular Bayesian approach calculates MLEs of $\boldsymbol{\phi}^{\delta}$ in Step 6, we do not incorporate the prior for $\boldsymbol{\phi}^{\delta}$ via maximum a posteriori estimators, although one could modify the approach to do this if desired.

Additionally in Step 0c, the user specifies the prior distribution for λ, the variance of the experimental observational error (which is assumed to be i.i.d. normal with mean 0). We treat λ in the same manner as $\boldsymbol{\phi}^{\delta}$. Its prior is used in Step 4 when generating a random value for ε, but this prior is not used when calculating the MLE of λ in Step 6.

Lastly in Step 0c, the user assigns a prior distribution $p(\boldsymbol{\theta})$ to the calibration parameters. Since we are directly interested in the uncertainty in the calibration parameters, Module 3 of the modular Bayesian approach in Step 7 calculates a full posterior distribution for $\boldsymbol{\theta}$ based on the specified prior, in contrast to the MLE-only estimation for $\boldsymbol{\phi}^{m}$ and $\boldsymbol{\phi}^{\delta}$ in Modules 1 and 2. Based on the posterior of $\boldsymbol{\theta}$, Module 3 also calculates the posterior covariance of $\boldsymbol{\theta}$ for each MC replicate.

Monte Carlo Loop (for $i = 1,\ldots, N_{mc}$)

Steps 1—7 are the steps within the MC loop for generating N_{mc} hypothetical sets of physical experimental response observations and calculating a hypothetical posterior covariance matrix for $\boldsymbol{\theta}$ for each set. We add a superscript (i) to denote that the quantity was generated in the ith MC replicate.

Step 1: Generate $\boldsymbol{\theta}^{(i)}$ from the Prior for $\boldsymbol{\theta}$ Specified in Step 0c-iii

The generated $\boldsymbol{\theta}^{(i)}$ will be treated as a "true" value for $\boldsymbol{\theta}$ when generating the hypothetical physical experimental response observations on the ith MC replicate.

Step 2: Generate $\mathbf{y}^{m(i)}$ at $\{\mathbf{X}^{e}, \boldsymbol{\theta}^{(i)}\}$

The $N_e \times 1$ vector $\mathbf{y}^{m(i)}$ denotes a random realization of values for the computer model response at the experimental input settings $\mathbf{X}^{e}$ (from Step 0b) and the calibration parameter values $\boldsymbol{\theta}^{(i)}$ (from Step 1).

The vector $\mathbf{y}^{m(i)}$ is generated from the multivariate normal posterior distribution of $y^m(\mathbf{x},\boldsymbol{\theta})$ computed in Step 0a, and it will be used in Step 5 to calculate a realization of the experimental data $\mathbf{y}^{e(i)}$.

Step 3: Generate $\boldsymbol{\delta}^{(i)}$ at $\mathbf{X}^e$

The $N_e \times 1$ vector $\boldsymbol{\delta}^{(i)}$ denotes a random realization of values for the discrepancy function at input settings $\mathbf{X}^e$. We first generate a random realization of $\boldsymbol{\phi}^\delta$, denoted by $\boldsymbol{\phi}^{\delta(i)}$, from their priors specified in Step 0c. Then $\boldsymbol{\delta}^{(i)}$ is generated from a multivariate normal distribution with mean and covariance based on the prior GP model for $\delta(\mathbf{x})$ from Step 0c using parameters $\boldsymbol{\phi}^{\delta(i)}$. Specifically, the mean is $\mathbf{H}^\delta\boldsymbol{\beta}^{\delta(i)}$, where $\mathbf{H}^\delta = [\mathbf{h}^\delta(\mathbf{x}_1^e)^T,\ldots,\mathbf{h}^\delta(\mathbf{x}_{N_e}^e)^T]^T$ with $\mathbf{h}^\delta(\mathbf{x})$ some specified regression basis functions (in our examples, we use $\mathbf{h}^\delta(\mathbf{x}) = 1$), and the covariance is the $\boldsymbol{\sigma}_\delta^{2(i)}\mathbf{R}^\delta$, where $\mathbf{R}^\delta$ is the $N_e \times N_e$ matrix with kth-row, jth-column entry equal to the prior correlation between $\delta(\mathbf{x}_k^e)$ and $\delta(\mathbf{x}_j^e)$ using parameters $\boldsymbol{\omega}^{\delta(i)}$.

Step 4: Generate $\boldsymbol{\varepsilon}^{(i)}$

The $N_e \times 1$ vector $\boldsymbol{\varepsilon}^{(i)}$ denotes a random realization of values for the experimental random error, with each element generated as a normal random variable with mean 0 and variance $\lambda^{(i)}$. Here, $\lambda^{(i)}$ is first generated from its prior distribution specified in Step 0c-i.

Step 5: Calculate $\mathbf{y}^{e(i)}$

Based on the model of Eq. (1), the $N_e \times 1$ vector of experimental response observations at $\mathbf{X}^e$ for the ith MC replicate is calculated via $\mathbf{y}^{e(i)} = \mathbf{y}^{m(i)} + \boldsymbol{\delta}^{(i)} + \boldsymbol{\varepsilon}^{(i)}$.

Step 6: Based on $\mathbf{y}^{e(i)}$, Estimate $\boldsymbol{\phi}^\delta$ and λ

For each MC replicate, we calculate the MLEs of the hyperparameters $\boldsymbol{\phi}^\delta$ of the GP model for $\delta(\mathbf{x})$ and the MLE of λ using Module 2 of the modular Bayesian approach with the hypothetical experimental data $\mathbf{y}^{e(i)}$ from Step 5. We denote the MLEs for the ith MC replicate by $\hat{\boldsymbol{\phi}}^{\delta(i)}$ and $\hat{\lambda}^{(i)}$.

Step 7: Calculate the Posterior Covariance of $\boldsymbol{\theta}$ for Replicate *i*

We calculate the posterior covariance $Cov^{(i)}[\boldsymbol{\theta}|\, \mathbf{y}^{e(i)}, \mathbf{y}^m]$ using Module 3 of the modular Bayesian approach, which treats $\boldsymbol{\phi}^m$, $\boldsymbol{\phi}^\delta$, and λ as fixed at their MLEs from Steps 0 and 6. $Cov^{(i)}[\boldsymbol{\theta}|\, \mathbf{y}^{e(i)}, \mathbf{y}^m]$ can be viewed as the actual posterior covariance (as calculated via Module 3) that would result if the actual physical experimental response observations $\mathbf{y}^e$ were equal to $\mathbf{y}^{e(i)}$. Details on calculating the posterior of the calibration parameter are discussed in Arendt et al. (2012b) and Kennedy and O'Hagan (2001). Since $r = 1$ or $r = 2$ in the examples in this paper, we used numerical integration (with Gauss-Legendre quadrature) to calculate $Cov^{(i)}[\boldsymbol{\theta}|\, \mathbf{y}^{e(i)}, \mathbf{y}^m]$. For higher dimensional problems, if one uses Markov Chain Monte Carlo (MCMC) (Robert and Casella 1999) methods to calculate the posterior distribution of $\boldsymbol{\theta}$, $Cov^{(i)}[\boldsymbol{\theta}|\, \mathbf{y}^{e(i)}, \mathbf{y}^m]$ can be calculated directly as the sample covariance matrix of the MCMC samples.

Step 8: Calculate the Preposterior Covariance of $\boldsymbol{\theta}$

After replicating Steps 1 through 7 in the MC loop a total of N_{mc} times, we calculate the preposterior covariance of the calibration parameters as the average of the N_{mc} posterior covariance matrices. That is, if we denote the preposterior covariance by $\boldsymbol{\Sigma}_{\boldsymbol{\theta},\text{pp}} = E[Cov(\boldsymbol{\theta}|\mathbf{Y}^e,\mathbf{y}^m)|\mathbf{y}^m]$, where $\mathbf{Y}^e$ denotes the (as yet) unobserved actual experimental data, then the estimate is

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta},\text{pp}} = \frac{1}{N_{mc}} \sum_{i=1}^{N_{mc}} Cov^{(i)}\left[\boldsymbol{\theta} \,|\, \mathbf{y}^{e(i)}, \mathbf{y}^m\right] . \tag{2}$$

4 A Modified Algorithm for Investigating the Behavior of the Preposterior Analysis

Consider the following modification to the algorithm of Figure 1: In Step 1, instead of generating a different $\boldsymbol{\theta}^{(i)}$ randomly from its prior distribution on each MC replicate, we use the same fixed value (denoted by $\boldsymbol{\theta}^t$, a value specified by the user outside the MC loop) for all replicates. We refer to this algorithm as the fixed-$\boldsymbol{\theta}$ preposterior analysis and denote the resulting estimate of the preposterior covariance matrix by $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta},\text{pp}}(\boldsymbol{\theta}^t)$. Recall that $\boldsymbol{\theta}^{(i)}$ in the Figure 1 algorithm is only used when generating the hypothetical experimental observations $\mathbf{y}^{e(i)}$, for which $\boldsymbol{\theta}^{(i)}$ is treated as the true values of the

calibration parameters. In reality there is only a single true value $\boldsymbol{\theta}^*$ that the laws of nature will dictate when generating the actual experimental data $\mathbf{y}^e$, based on which the actual posterior covariance of $\boldsymbol{\theta}$ will be calculated. In light of this, one might wonder how well the preposterior analysis of Figure 1 [which has no knowledge of the true $\boldsymbol{\theta}^*$ when generating $\mathbf{y}^{e(i)}$ and must, therefore, average the results over values of $\boldsymbol{\theta}^{(i)}$ drawn from the prior $p(\boldsymbol{\theta})$] can predict the actual posterior covariance matrix. In the ideal situation that $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta},\mathrm{pp}}(\boldsymbol{\theta}^t)$ does not depend on $\boldsymbol{\theta}^t$, we might expect the prediction to be good. In Section 5 we use the fixed-$\boldsymbol{\theta}$ preposterior analysis to investigate the extent to which $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta},\mathrm{pp}}(\boldsymbol{\theta}^t)$ depends on $\boldsymbol{\theta}^t$ in the examples considered.

Although it may seem counterintuitive that $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta},\mathrm{pp}}(\boldsymbol{\theta}^t)$ might not depend (strongly) on $\boldsymbol{\theta}^t$, this is precisely the situation in the standard linear regression formulation $\mathbf{y} = \mathbf{X}\boldsymbol{\theta}$ + error, where $\mathbf{X}$ is the design matrix, and $\mathbf{y}$ is the observation vector. In the linear regression formulation, under mild assumptions the covariance matrix of the regression estimate of $\boldsymbol{\theta}$ is $[\mathbf{X}^T\mathbf{X}]^{-1}$ multiplied by the error variance, which is independent of the true $\boldsymbol{\theta}$. This characteristic allows one to design experiments, prior to observing $\mathbf{y}$, that are "optimal" regardless of the true $\boldsymbol{\theta}$. The situation is more complex for the calibration problem, in part because of the black-box, potentially nonlinear dependence of $y^m(\mathbf{x},\boldsymbol{\theta})$ on $\boldsymbol{\theta}$. However, in the next section we demonstrate that the relative (when comparing different experimental designs) dependence of $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta},\mathrm{pp}}(\boldsymbol{\theta}^t)$ on $\boldsymbol{\theta}^t$ in the examples is mild enough that the preposterior covariance still allows one to distinguish good experimental designs from poor ones.

Notice that the fixed-$\boldsymbol{\theta}$ preposterior covariance $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta},\mathrm{pp}}(\boldsymbol{\theta}^t)$ is related to the preposterior covariance $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta},\mathrm{pp}}$ via

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta},\mathrm{pp}} = \int \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta},\mathrm{pp}}\left(\boldsymbol{\theta}^t\right) p\left(\boldsymbol{\theta}^t\right) d\boldsymbol{\theta}^t \tag{3}$$

where $p(\bullet)$ is the prior distribution of $\boldsymbol{\theta}$ specified in Step 0c-i. In other words, the preposterior covariance is the expected value of the fixed-$\boldsymbol{\theta}$ preposterior covariance with respect to the prior of $\boldsymbol{\theta}$.

5 Examples and Discussion

In this section, we consider various examples that illustrate how the preposterior analysis can be used, after conducting the computer simulations but before conducting the physical experiments, to predict the identifiability that will result from a proposed experimental design. The system in the first example (Section 5.1) is inherently difficult to identify even with a large amount of observed physical experiments, whereas the system in the second example (Section 5.2) is more identifiable with an appropriate experimental design.

5.1 Beam Example

5.1.1 Computer Model, Physical Experiments, and Preliminaries

We investigate the simply supported beam example considered in Arendt et al. (2012b). Briefly, the simply supported beam is comprised of a 2 m long beam with a rectangular cross-section (height of 52.5 mm and width of 20 mm). One end of the beam is rigidly fixed to a support, while the other end is supported by a roller. The design variable x is the magnitude of a static force, in N, applied to the midpoint of the beam, and the response y is the strain, in m, measured at the midpoint of the beam (at the top of the cross-section). The computer simulation is a finite element analysis (FEA) model, implemented in Abaqus 6.9, with a simplified material law that depends on Young's modulus (the calibration parameter θ). For Step 0a, the simulations were observed on a 4×4 evenly spaced grid ($N_m = 16$) over the input space ($1300 \le x \le 2300$ N and $150 \le \theta \le 300$ GPa). The simulation response observations $\mathbf{y}^m$ and the posterior mean of $y^m(\mathbf{x},\boldsymbol{\theta})$ from Step 0a are shown in Figure 2. Because of the smooth nature of $y^m(\mathbf{x},\boldsymbol{\theta})$, there was very small posterior uncertainty in $y^m(\mathbf{x},\boldsymbol{\theta})$ over the entire input space.

Regarding Step 0b, we will conduct different analyses for different values of N_e. For each N_e, $\mathbf{X}^e$ is evenly spaced over the input range $1300 \le x \le 2300$. For purposes of comparing the preposterior covariance to the posterior covariance, we generated actual "physical" experimental data via the same

FEA model but using a more elaborate material law that results in a discrepancy between the computer model and the experimental data. Observation error with variance λ = 6.63e-12 m^2 was added to the experimental data, and the true value of the Young's modulus calibration parameter was taken to be θ^* = 206.8 GPa. These physical experimental data are not used in the preposterior algorithm of Figure 1 and are used only to verify the results of the preposterior algorithm.

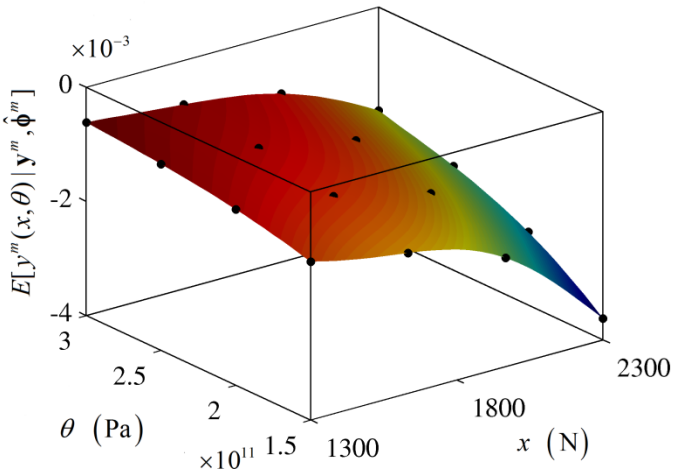

Figure 2. Posterior mean of the computer model GP for the beam example. The bullets indicate $\mathbf{y}^m$.

For Step 0c, we assigned to $\boldsymbol{\phi}^\delta$ a point mass prior with mass at $\{\boldsymbol{\beta}^\delta, \sigma_\delta^2, \boldsymbol{\omega}^\delta\}$ = {0.00, 1.11×10^{-7}, 0.002}. The value 1.11×10^{-7} m^2 for σ_δ^2 was chosen so that the bounds of the 99.7% prediction interval of the discrepancy function were approximately [-1,1]×10^{-3} m. The correlation parameters $\boldsymbol{\omega}^\delta$ were chosen to represent a relatively smooth GP model for the discrepancy function. We also assigned to λ a point mass prior with mass at 6.63e-12 m^2. Recall that the prior distributions for $\boldsymbol{\phi}^\delta$ and λ are used only in Steps 1—5 to generate random realizations of $\mathbf{y}^{e(i)}$. In general, the MLEs of $\boldsymbol{\phi}^\delta$ and λ are calculated in Step 6. For the examples in this paper, however, we treated λ as a known value and did not estimate it in Step 6. This is reasonable in systems for which the random error variance can be estimated externally, simply by taking replicate experimental measurements at the same input settings. Lastly, we chose $p(\theta)$ to be a noninformative uniform distribution over the full range of θ.

5.1.2 Comparing the Preposterior and Posterior Covariances

We begin by comparing the fixed-**θ** preposterior covariance $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta},\mathrm{pp}}(\theta^t)$ for various θ^t to the actual posterior covariance based on the experimental data from the more elaborate FEA model. As described in Section 4, for each θ^t, $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta},\mathrm{pp}}(\theta^t)$ was calculated using the algorithm of Figure 1 with $\theta^{(i)}$ from Step 1 replaced by the same value θ^t for all N_{mc} replicates. Figure 3 plots the fixed-**θ** preposterior standard deviation (STD) [i.e., the square root of $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta},\mathrm{pp}}(\theta^t)$] versus θ^t for two different experimental designs with N_e = 3 and N_e = 8. The individual points plotted at θ = 206.8 GPa and represented by the shaded circle (for the design with N_e = 3) and shaded square (for the design with N_e = 8) are the actuals posterior STDs of θ that result by applying Modules 1—3 of the standard modular Bayesian approach to a set of experimental data generated using the true value θ^* = 206.8 GPa. These individual points are included in Figure 3 for purposes of comparing to the preposterior STDs that our algorithm calculates, which are represented by the series of connected points. In Figure 3 both the fixed-**θ** preposterior STD and the posterior STD have been normalized (divided) by the STD of the prior $p(\theta)$.

Although the fixed-**θ** preposterior STDs differ from the posterior STD, their relative change as N_e increases is comparable. In particular, the relative difference in the fixed-**θ** preposterior covariance for the design with N_e = 3 and the design with N_e = 8 is reasonably independent of the fixed θ^t, and this relative difference is quite consistent with the relative difference in the actual posterior covariance for the true θ^* = 206.8 GPa. Consequently, the relative difference in the preposterior STD [which we calculated as the integral of the fixed-**θ** preposterior STD in Figure 3, with respect to $p(\theta)$] accurately reflects the relative difference in the actual posterior STD in this example. Specifically, when N_e is increased from 3 to 8 in Figure 3, the posterior STD decreases by 10.3% (from 0.8231 to 0.7383), and the preposterior STD decreases by 12.5% (from 0.6676 to 0.5840). This correctly indicates that the experimental design with N_e = 8 will result in somewhat better identifiability than the experimental design with N_e = 3, as measured by the posterior STD that would result after the experiment is conducted.

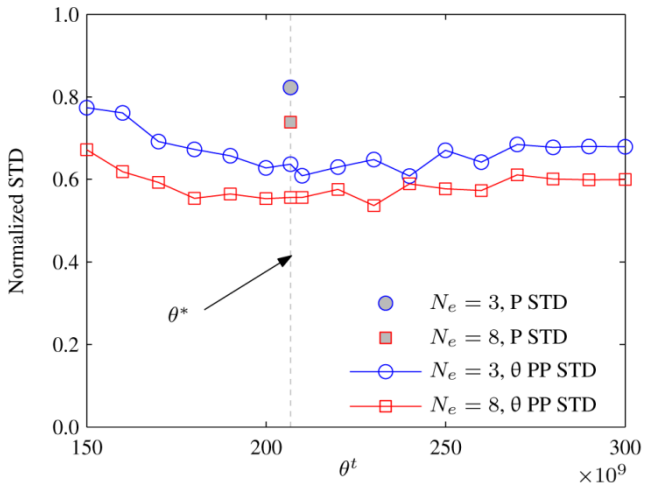

Figure 3. Plot of the fixed-θ preposterior STD and the posterior STD (both normalized by the prior STD of θ) verses θ^t for the beam example. "θ PP STD" is the fixed-θ preposterior STD and "P STD" is the posterior STD.

We then calculated the preposterior STD from the fixed-θ preposterior STDs in Figure 3, as described in Section 4. We repeated the analyses for various N_e, and Figure 4 shows a scatter plot of the preposterior STD versus the posterior STD [each normalized by prior STD θ)] for N_e = 3, 4, …, 15. The scatter plot conveys a number of interesting characteristics of the beam example. Although the preposterior STD slightly underestimates the posterior STD for the different designs, the trend is the same: When N_e increases from 3 to 4, the preposterior STD correctly predicts that the posterior STD will decrease. Then, as N_e increases from 4 through 15, the preposterior STD again correctly predicts that there will be negligible further decrease in the posterior STD. The overall conclusion that a user might draw from this preposterior analysis is that the system inherently suffers from a lack of identifiability, because an increase in N_e beyond 4 results in almost no further improvement in the preposterior STD. Arendt et al. (2012a, 2012b) discussed the reasons behind the lack of identifiability in detail and also demonstrated that identifiability is greatly improved when multiple responses are measured experimentally. In the next section we will explore a second example in which the identifiability of the system continues to improve as experimental data are added.

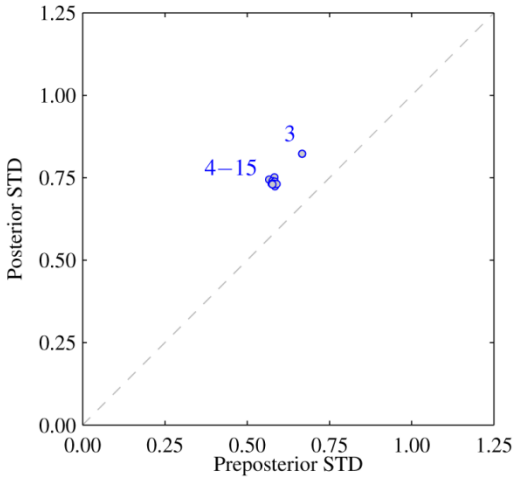

Figure 4. Plot of the preposterior STD versus the posterior STD (both normalized by the prior STD of θ) for the beam example. The numbers indicate N_e for each case.

5.2 Sinusoidal Example

In this section, we explore an example involving a mathematical test function to investigate various aspects of the preposterior analysis. We compare the fixed-**θ** preposterior covariance to the posterior covariance for two different experimental data input settings and show results analogous to those shown in Figure 3 and Figure 4. Further analyses are presented in the online supplement.

5.2.1 Computer Model, Physical Experiments, and Preliminaries

Suppose the computer model is the function

$$y^m(x,\theta) = \sin(\theta x)\,. \tag{4}$$

The simulation was run on an 8×8 evenly spaced grid (N_m = 64) over the input space ($1 \le x \le \pi$ and $1 \le \theta \le 4$). The simulation response observations $\mathbf{y}^m$ and the posterior mean of $y^m(\mathbf{x},\boldsymbol{\theta})$ from Step 0a are shown in Figure 5. Although the model shown in Eq. (4) was used to generate the computer model data used in Step 0a, no knowledge of this parametric model was used anywhere in the algorithm to calculate the preposterior standard deviations. In the Gaussian process modeling in Step 0a, these computer model data were treated as having come from a black-box, and in Step 2 the computer model responses were

generated from the posterior distribution of $y^m(\mathbf{x},\boldsymbol{\theta})$ obtained in Step 0a without using knowledge of the underlying sinusoidal relationship.

For the experimental data, we consider different values of N_e, and for each, $\mathbf{X}^e$ is evenly spaced over the input range $1 \le x \le \pi$. Within Steps 1—5 of the algorithm of Figure 1, the experimental data are generated according to the GP models. However, to generate the actual physical experimental data (used only for the purpose of calculating actual posterior STDs to compare to the preposterior STDs calculated by our algorithm), we use the model

$$y^e(x) = y^m(x,\theta^*) + \delta^k(x) + \varepsilon, \tag{5}$$

where we have added a superscript k on $\delta^k(x)$ because we consider a number of different discrepancy functions in the online supplement. Here we consider the discrepancy function

$$\delta^1(x) = 0.25(0.1\exp\{x\} - 0.05x^2), \tag{6}$$

which we use only to generate the actual experimental data (the discrepancy function is still generated from a GP model within the algorithm for calculating the preposterior STD). Observation error with variance $\lambda = 1\times10^{-14}$ was added to the experimental data.

Regarding Step 0c, we assigned to $\boldsymbol{\phi}^\delta$ a point mass prior with mass at $\{\boldsymbol{\beta}^\delta, \sigma_\delta^2, \boldsymbol{\omega}^\delta\} = \{0.00, 0.028, 3.09\}$. The value 0.028 for σ_δ^2 was chosen so that the bounds of the 99.7% interval for the discrepancy function were approximately [-0.5,0.5] (other values for σ_δ^2 are considered in the online supplement), and $\boldsymbol{\omega}^\delta$ was chosen to represent a relatively smooth GP model. We also assigned to λ a point mass prior with mass at 1×10^{-14}. As in the beam example, we treat λ as known, but estimate $\boldsymbol{\phi}^\delta$ in Step 6. We chose $p(\theta)$ to be a noninformative uniform distribution over the full range of θ.

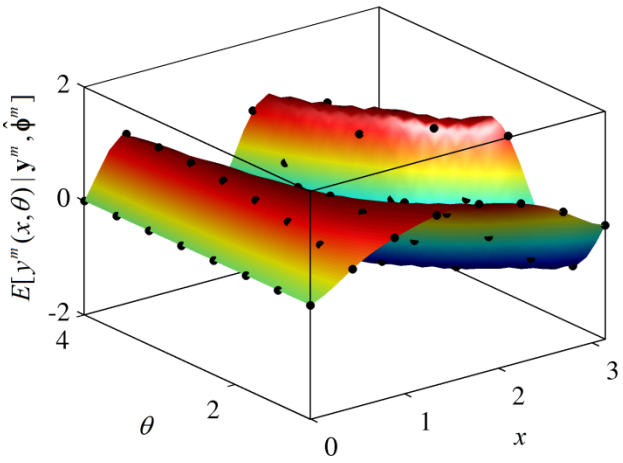

Figure 5. Posterior mean of the computer model GP for the sinusoidal example. The bullets indicate $\mathbf{y}^m$.

5.2.2 Comparing the Preposterior and Posterior Covariances

We start by comparing the fixed-**θ** preposterior STD for various θ^t to the actual posterior STD based on the experimental data generated from Eq. (6) for two different experimental designs with $N_e = 3$ and $N_e = 8$. When generating the experimental data, we considered a range of values for θ^*. Figure 6 plots the fixed-**θ** preposterior STD versus θ^t and the actual posterior STD versus θ^*. The posterior and preposterior STDs are again normalized by the prior STD of $p(\theta)$. In Figure 6, the fixed-**θ** preposterior STD and the posterior STD are similar when $\theta^t = \theta^*$, and the preposterior STD reflects the relative differences (for different N_e) in the actual posterior STD reasonably well for most θ^*. For example, if $\theta^* = 2.25$ the posterior STD decreases from 0.6302 to 0.2861 when N_e increases from 3 to 8, and the preposterior STD [the integral of the fixed-**θ** preposterior STD in Figure 6 with respect to $p(\theta)$] decreases from 0.8304 to 0.2650. Although these are different, the preposterior STD correctly indicates that the experimental design with $N_e = 8$ will result in substantially better identifiability compared to the identifiability achieved using the experimental design with $N_e = 3$.

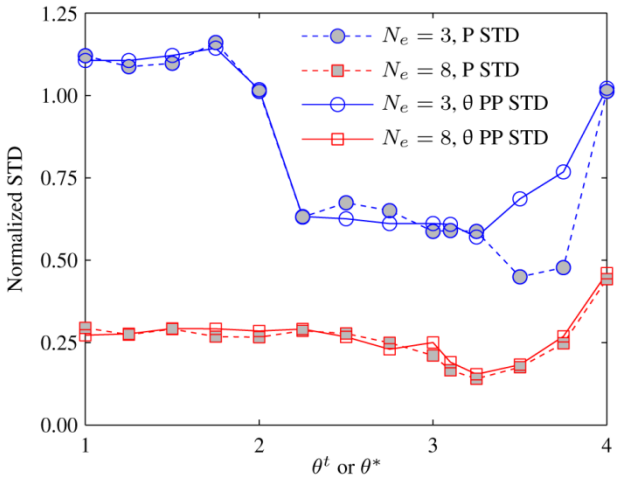

Figure 6. Plot of the fixed-θ preposterior STD versus θ^t and the posterior STD versus θ^* (the vertical axis is normalized by the prior STD of θ) for the sinusoidal example. "θ PP STD" is the fixed-θ preposterior STD and "P STD" is the posterior STD.

We next compare the preposterior and posterior STDs for various N_e. Instead of specifying a single value for θ^* as in Figure 4, we consider the mean posterior STD [the mean of the posterior STD with respect to a uniform distribution for θ^* over the range $1 \le \theta^* \le 4$]. And as a measure of the extent to which the posterior STD depends on θ^*, we also calculate the standard deviation of the posterior STD [also with respect to a uniform distribution for θ^* over the range $1 \le \theta^* \le 4$]. Figure 7 shows a scatter plot of the preposterior STD versus the mean posterior STD (the error bands are ± 1.0 standard deviations of the posterior STD) for N_e = 3, 4, …, 10. There is a very tight relationship between the preposterior STD and the mean posterior STD. Notice that the wide error bands for N_e = 3 and the much narrower error bands for N_e = 8 are consistent with Figure 6 in the following sense. For N_e = 8 the posterior STD in Figure 6 is much less dependent on θ^* than for N_e = 3. There is actually a stronger trend in Figure 7 than the wide error bands would seem to imply. If one fixed θ^* and plotted the posterior STD for the fixed θ^* versus the preposterior STD (as was done in Figure 4), there would be a very tight linear trend in the graph for each θ^*. Consequently, for this example we conclude that the preposterior STD provides an accurate indication of the relative improvements in identifiability that would be achieved by increasing the size of the physical experiment. See the online supplement for further analyses.

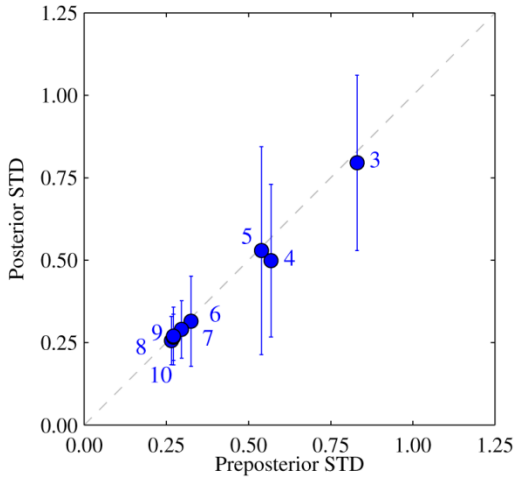

Figure 7. Plot of the posterior STD versus the preposterior STD (both normalized by the prior STD) for the sinusoidal example. The numbers indicate N_e for each case. In the vertical axis, the circles and error bars represent the mean posterior STD and ± 1 standard deviations of the posterior STD, with respect to a uniform distribution for θ^* over the range $1 \le \theta^* \le 4$.

5.3 Using the Preposterior Analysis to Select N_e

One way in which the preposterior analysis can be used is to help the user choose the number N_e of experimental runs. Figure 8, which plots the normalized preposterior STD versus N_e for the beam and sinusoidal examples, illustrates how this might be accomplished. For the beam example, one might conclude that increasing N_e beyond 4 will provide little further knowledge of θ^* and that the beam system is inherently difficult to identify with only the single experimentally measured response (which was strain at the beam midpoint). See Arendt et al. (2012a) for discussions on how experimentally measuring multiple responses (e.g., angle of beam deflection, internal energy, etc.) that share a mutual dependence on the calibration parameter may substantially improve identifiability. For the sinusoidal example (for which results with two different choices of σ_δ^2 are shown in Figure 8), one might conclude that the system is more identifiable than the beam system, but that there is little further improvement to be gained by using $N_e > 8$. See the online supplement for an additional example with two inputs.

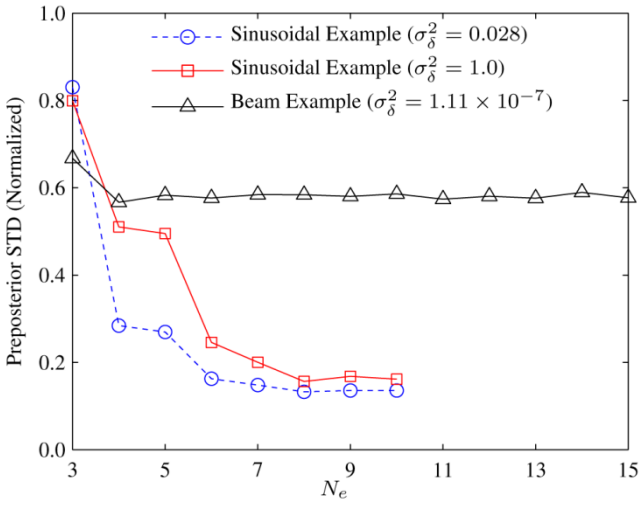

Figure 8. Preposterior STD (normalized) verses N^e for the beam and sinusoidal examples.

5.4 Is Calibration Identifiability Possible with the Kennedy and O'Hagan Model?

The focus of this work is on assessing the level of identifiability in the Kennedy and O'Hagan calibration model, prior to conducting the physical experiment. This is premised on the assumption that reasonable identifiability is sometimes possible for some applications. In the calibration literature, the nearly universal consensus is that identifiability is extremely difficult for many systems, because of the difficulty in distinguishing between the discrepancy function and the effects of the calibration parameters. However, there appears to be some debate regarding whether one can ever claim good identifiability with any level of confidence. We believe good identifiability is indeed possible in some cases, and in this section we provide an example to support this and demonstrate how it depends strongly on the experimental design. We also argue that the Kennedy and O'Hagan approach does a good job of letting the user know whether or not reasonable identifiability was achieved for a particular situation, via the posterior standard deviations for $\boldsymbol{\theta}$. We also discuss implications of the form of identifiablity studied in Loeppky, et al. (2006) and Tuo and Wu (2013), which is asymptotic in nature

To illustrate situations in which good identifiability is possible, depending on the experimental design, suppose there are two inputs, two calibration parameters, and the computer model is $y^m(\mathbf{x}, \boldsymbol{\theta}) = x_2 sin(\theta_1 x_1 + \theta_2)$ with $x_1 \in [-1,1]$, $x_2 \in [-1,1]$, $\theta_1 \in [10,20]$, and $\theta_2 \in [0,2\pi]$. Further suppose the

computer simulation is cheap, so that infinitely many observations of y^m are available, and the physical experimental data are generated from (1) with true values $\{\theta_1^* = 15, \theta_2^* = \pi/2\}$ (in radians, with $\mathbf{x}$ unitless), $\varepsilon \sim N(0, 0.1^2)$, and discrepancy function $\delta(\mathbf{x}) = 0.5 + x_1^2$. As always, we treat the parametric computer model and discrepancy function as a black box and use no knowledge of their parametric forms in the analyses, other than as a mechanism to generate the data. We used point mass priors for $\boldsymbol{\phi}^\delta$ and λ with masses at $\{\boldsymbol{\beta}^\delta, \sigma_\delta^2, \boldsymbol{\omega}^\delta\}$ = {0.00, 1.0, 1.0} and $\lambda = 0.1^2$ and a noninformative prior for $\boldsymbol{\theta}$ over its full range.

Each panel in the top row of Figure 9 shows $y^e(\mathbf{x})$ and $y^m(\mathbf{x}, \boldsymbol{\theta})$ (the latter for three different values of $\boldsymbol{\theta}$) as a function of x_1 for a particular physical experimental design. Each panel in the bottom row shows the three different $\hat{\delta}(\mathbf{x})$ corresponding to the three different potential values of $\boldsymbol{\theta}$, for the data in the top row panel immediately above. In panels (a) and (b), the physical experimental design is N_e = 60 points evenly spaced over the full range $x_1 \in [-1,1]$ with $x_2 = 1$ fixed. For the design in panel (a), one can observe a very close match (aside from a vertical offset due to the quadratic discrepancy) between the observed experimental data $y^e(\mathbf{x})$ and the computer model response $y^m(\mathbf{x}, \boldsymbol{\theta})$ only at the true values $\boldsymbol{\theta} = \boldsymbol{\theta}^* = \{15, \pi/2\}$. For the other two values of $\boldsymbol{\theta}$ ($\{10, \pi/2\}$ and {15,0}), the resulting $y^m(\mathbf{x}, \boldsymbol{\theta})$ is a far poorer match with $y^e(\mathbf{x})$.

Would the data in panel (a) constitute reasonably compelling evidence that the true $\boldsymbol{\theta}^*$ is close to the correct value $\{15, \pi/2\}$? This is debatable, but our opinion is that it would. The response $y^m(\mathbf{x}, \boldsymbol{\theta})$ in panel (a) exhibits very distinct behavior as a function of x_1, and this behavior is highly dependent on $\boldsymbol{\theta}$. Although one cannot rule it out with absolute certainty, it seems highly unlikely that the close match in this distinct behavior of $y^m(\mathbf{x}, \boldsymbol{\theta})$ and $y^e(\mathbf{x})$ only at $\boldsymbol{\theta} \approx \boldsymbol{\theta}^*$ is just coincidence and the true $\boldsymbol{\theta}$ is really some other value for which the match between $y^m(\mathbf{x}, \boldsymbol{\theta})$ and $y^e(\mathbf{x})$ is far inferior.

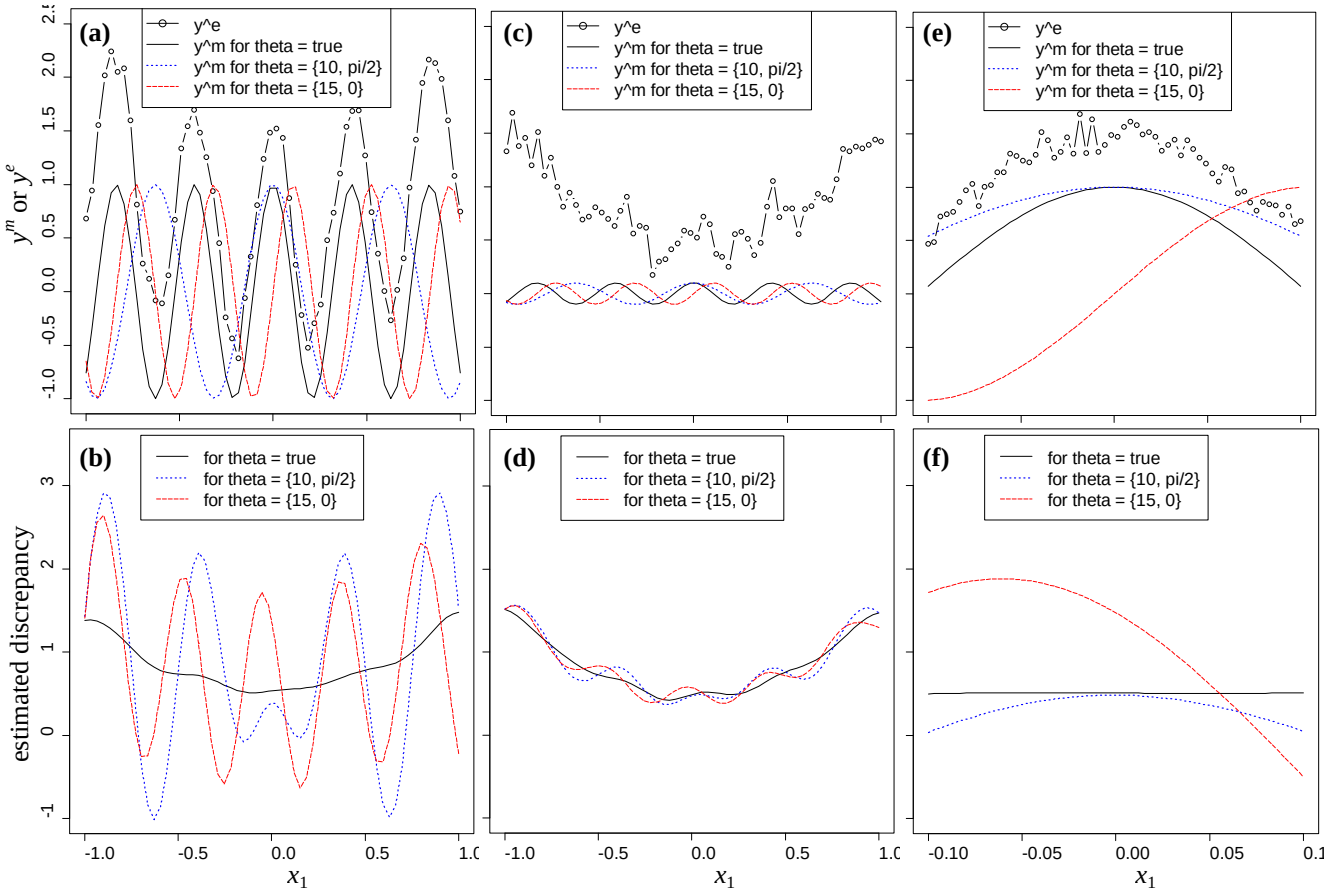

Figure 9. Top row: y^e and y^m for three different values of $\boldsymbol{\theta}$, as a function of x_1 for (a) $x_1 \in [-1,1]$ and $x_2 = 1$, (c) $x_1 \in [-1,1]$ and $x_2 = 0.1$, and (e) $x_1 \in [-0.1,0.11]$ and $x_2 = 1$. Bottom row: Estimated discrepancy functions for the three values of $\boldsymbol{\theta}$, for the data shown in the top row panel immediately above the corresponding bottom row panel.

As an analogy, much of the inferential statistical objective when fitting linear regression models to limited-size samples of data is distinguishing between what is likely a meaningful relationship versus what is just coincidental random chance. There are established methods of quantifying the statistical significance and uncertainty (P-values and confidence intervals), and even though one may not be able to conclude with absolute certainty that an observed relationship in the sample data is more than just coincidence, one may be able to conclude this with reasonable certainty.

Fortunately, the Kennedy and O'Hagan approach has an inherent mechanism for quantifying the statistical uncertainty and assessing (what it thinks is) the level of identifiability; namely, the posterior standard deviations of $\boldsymbol{\theta}$. For the data in Figure 9(a,b), the posterior standard deviations of θ_1 and θ_2 were

0.081 and 0.050, respectively, which are quite small and indicate that there is reasonably good identifiability for this example and this design (point estimates were $\hat{\theta}_1 = 14.978$ and $\hat{\theta}_2 = 1.610$, which are close to the true values). We discuss the preposterior standard deviation shortly.

Contrast this with the data for the design shown in Figures 9 (c,d), for which there were still 60 points evenly spaced over the full range $x_1 \in [-1,1]$, but now with $x_2 = 0.1$ fixed at a smaller value than for Figures 9(a,b). With smaller x_2, the amplitude of oscillation in the response is much smaller. The match between $y^e(\mathbf{x})$ and $y^m(\mathbf{x}, \boldsymbol{\theta})$ is slightly better at the true values $\boldsymbol{\theta} = \boldsymbol{\theta}^* = \{15, \pi/2\}$ than at the two other values of $\boldsymbol{\theta}$, but the difference is not nearly as pronounced as in Figures 9(a,b). Visually, the data for the Figures 9 (c,d) design present far less compelling evidence that $\boldsymbol{\theta}^* \approx \{15, \pi/2\}$ than the corresponding data for the Figures 9(a,b) design. Quantitatively, the Kennedy and O'Hagan approach correctly assessed the poorer level of identifiability, producing posterior standard deviations of 0.834 and 0.521 for θ_1 and θ_2, respectively, which are a full order of magnitude larger than for Figures 9(a,b) (point estimates were $\hat{\theta}_1 = 14.88$ and $\hat{\theta}_2 = 1.232$).

The design shown in Figures 9(e,f), for which the 60 design points were evenly spaced over a much smaller range $x_1 \in [-0.1,0.1]$ with $x_2 = 1.0$, resulted in even poorer identifiability for θ_1. Intuitively, the reason for the poorer θ_1 identifiability is that the distinct oscillatory behavior of $y^e(\mathbf{x})$ and $y^m(\mathbf{x}, \boldsymbol{\theta})$ is not present over the smaller range $x_1 \in [-0.1,0.1]$. The Kennedy and O'Hagan approach was again able to assess the poorer level of identifiability, producing posterior standard deviations of 1.273 and 0.214 for θ_1 and θ_2, respectively (point estimates were $\hat{\theta}_1 = 14.77$ and $\hat{\theta}_2 = 1.501$).

Figure 10 is a scatter plot of the posterior standard deviation versus preposterior standard deviation for θ_1 and θ_2 for the three different experimental designs corresponding to Figures 9 (a), (c), and (e). There is quite strong correlation between the posterior and preposterior standard deviations, which indicates that the preposterior analysis has done a good job of predicting the level of identifiability that will result after conducting the physical experiment. Figure 10 also highlights the importance of the experimental design and the extent to which it can effect identifiability. In this case, the preposterior

analysis could have been used to choose the design in Figure 9(a), on the basis of providing far better identifiability than the designs in Figures 9(c) and 14(e).

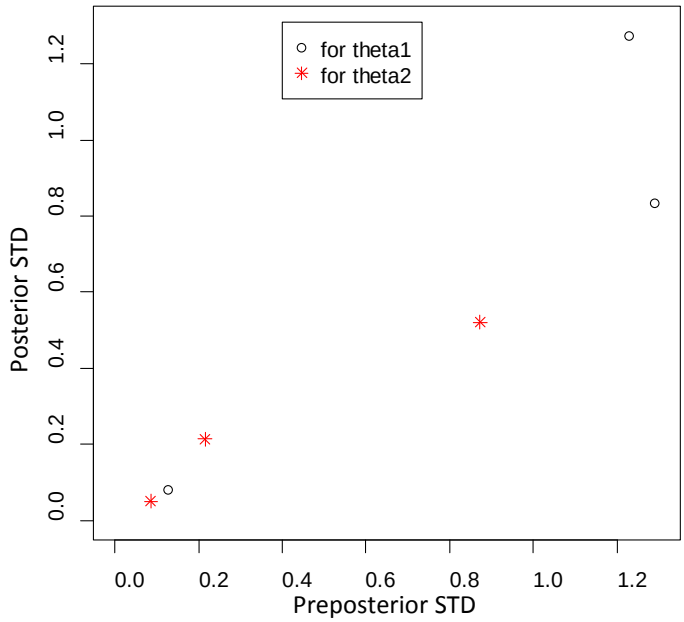

Figure 10. Scatter plot of posterior standard deviations versus preposterior standard deviations for θ_1 (open circles) and θ_2 (asterisks) for the three different designs corresponding to panels (a), (c), and (e) of Figure 9.

It is worth noting that Tuo and Wu (2013) investigated the identifiability issue via an asymptotic analysis. They ignored the random error ε in (1) and considered the likelihood-based estimator of $\boldsymbol{\theta}$, which (for cheap simulation code) minimizes

$$[\mathbf{y}^e - \mathbf{y}^m(\boldsymbol{\theta})]^T \mathbf{R}_\delta^{-1} [\mathbf{y}^e - \mathbf{y}^m(\boldsymbol{\theta})], \tag{7}$$

where $\mathbf{y}^e$ and $\mathbf{y}^m(\boldsymbol{\theta})$ are the vectors of experimental and simulation response observations at the same input sites $\mathbf{X}^e$, and $\mathbf{R}_\delta$ denotes the covariance matrix of the discrepancy function at $\mathbf{X}^e$. They showed that as $N_e \to \infty$ and the observations more and more densely cover the input domain (denoted by Ω) for $\mathbf{x}$, the minimizer of the discrete-sample likelihood criterion (7) converges to the minimizer $\boldsymbol{\theta}' \equiv argmin_{\boldsymbol{\theta}} \int_\Omega \int_\Omega [y^e(\mathbf{x}) - y^m(\mathbf{x}, \boldsymbol{\theta})] R_\delta^{-1}(\mathbf{x}, \mathbf{z}) [y^e(\mathbf{z}) - y^m(\mathbf{z}, \boldsymbol{\theta})] d\mathbf{x} d\mathbf{z}$ of its continuous-domain counterpart. The integral here is the reproducing kernel Hilbert space (RKHS) norm of the difference $y^e(\mathbf{x}) - y^m(\mathbf{x}, \boldsymbol{\theta})$ between the experimental and simulation response functions, where the kernel

$R_\delta(\mathbf{x},\mathbf{z})$ is the discrepancy covariance function, and its inverse $R_\delta^{-1}(\mathbf{x},\mathbf{z})$ can be informally viewed as the Karhunen-Loève expansion of $R_\delta(\mathbf{x},\mathbf{z})$ with the eigenvalues inverted [$y^e(\mathbf{x}) - y^m(\mathbf{x},\boldsymbol{\theta})$ must be an element of the same RKHS, for which its RKHS basis representation coefficients converge at a faster rate than the eigenvalues of $R_\delta(\mathbf{x},\mathbf{z})$, which was assumed in Tuo and Wu (2013)]. They also defined $\boldsymbol{\theta}_0 \equiv argmin_{\boldsymbol{\theta}} \int_\Omega [y^e(\mathbf{x}) - y^m(\mathbf{x},\boldsymbol{\theta})]^2 \, d\mathbf{x}$ [i.e., the minimizer of the L_2-norm of $y^e(\mathbf{x}) - y^m(\mathbf{x},\boldsymbol{\theta})$, which is the RKHS norm with a Dirac delta kernel] and termed the likelihood-based estimator as inconsistent because it approaches $\boldsymbol{\theta}'$, as opposed to $\boldsymbol{\theta}_0$, as the observations cover Ω more and more densely. Loeppky, et al. (2006) derived the similar, but less general, result that the likelihood-based estimator approaches $\boldsymbol{\theta}_0$ in the special case that $y^m(\mathbf{x},\boldsymbol{\theta}_0) = y^e(\mathbf{x})$ for all $\mathbf{x} \in \Omega$ (i.e., when there exists a value of $\boldsymbol{\theta}$ that makes the difference $y^e(\mathbf{x}) - y^m(\mathbf{x},\boldsymbol{\theta})$ identically zero). In light of this, in the context of tuning (for which there is no true value $\boldsymbol{\theta}^*$ of the parameters), Tuo and Wu (2013) recommended a modified estimator of $\boldsymbol{\theta}$ that essentially minimizes $[\mathbf{y}^e - \mathbf{y}^m(\boldsymbol{\theta})]^T[\mathbf{y}^e - \mathbf{y}^m(\boldsymbol{\theta})]$ instead of Eq. (7). Their modified estimator converges to $\boldsymbol{\theta}_0$ as the observations cover Ω more and more densely.

In the context of calibration (for which one would like to estimate the true value $\boldsymbol{\theta}^*$, as opposed to $\boldsymbol{\theta}_0$), one may be tempted to view the inconsistency result of Tuo and Wu (2013) as casting the prospect of calibration in a dismal light: To quote one of the anonymous referees for this paper, the (likelihood version of the) estimator in the Kennedy and O'Hagan model will "converge to the wrong value". However, we believe that this does not preclude reasonable calibration identifiability for a number of reasons. Although the Tuo and Wu (2013) result is asymptotic in sample size, the input domain Ω remains fixed. And for most, if not all, real simulation and experimental systems, the size of Ω is limited. Because of the highly spatially correlated nature of response surface data, when sample size increases beyond some value, very little new information is gained by more and more densely covering a limited-size input domain with additional observations. In this sense, one can view the Tuo and Wu (2013) result as a finite-sample result; and for finite samples, virtually all estimators will give the wrong value with probability

one. If the size of Ω were allowed to grow to infinity in a truly asymptotic scenario, it is likely that one could devise conditions on $y^m(\mathbf{x}, \boldsymbol{\theta})$ and $y^e(\mathbf{x})$ under which the estimator of $\boldsymbol{\theta}$ is guaranteed to converge to the true $\boldsymbol{\theta}^*$. However, considering the inherently limited size of Ω in real applications, it is doubtful that any such asymptotic analysis would be meaningful.

For the calibration problem, we believe the more relevant question is not whether the estimator is wrong for a given set of infinitely dense experimental and simulation data over a finite, limited input domain; but rather, whether reasonably good identifiability can be achieved for the finite sample of data that will be used. We also believe that the examples in this paper and elsewhere (e.g., Higdon, et al. (2004), Loeppky, et al. (2006), Han, et al. (2009), Arendt, et al. (2012a), and Arendt, et al. (2012b)) demonstrate that reasonably good identifiability is achievable in some situations and, in order to assess the level of identifiability, the posterior standard deviation (after conducting the experiment) and the preposterior standard deviation (prior to conducting the experiment) can be used.

6 Conclusions

Because of identifiability issues, it is often difficult to distinguish parameter uncertainty from uncertainty in the discrepancy function when using the GP-based method of Kennedy and O'Hagan (2001) for calibrating computer models. In this paper, we have shown that the preposterior covariance of $\boldsymbol{\theta}$ calculated via the algorithm of Figure 1 constitutes a reasonable prediction of the posterior covariance that will result when the physical experimental data are collected, at least for the examples considered. As such, the preposterior covariance can provide insight into the identifiability of a system and serve as a criterion for helping to design an effective physical experiment, based on the results of a previously conducted set of computer simulations.

Regarding the conditions on $\delta(\mathbf{x})$ required for reasonable identifiability, clearly there must be some, and $\delta(\mathbf{x})$ is not allowed to be any function. Our assumption that $\delta(\mathbf{x})$ can be represented by a Gaussian process model is not a particularly strong one, because for various choices of correlation function a Gaussian process prior can produce an extremely wide variety of stochastic realizations of $\delta(\mathbf{x})$.

This flexibility of the Gaussian process model is one of the reasons for its popularity in the calibration literature. As discussed in Arendt, et al. (2012b), good identifiability is more difficult to achieve if $\delta(\mathbf{x})$ is not relatively smooth, which is reflected by the correlation parameters $\boldsymbol{\omega}^{\delta}$ (smaller $\boldsymbol{\omega}^{\delta}$ corresponds to a smoother discrepancy function). *After* collecting the physical experimental data, one can always estimate $\boldsymbol{\omega}^{\delta}$ to gauge the smoothness of $\delta(\mathbf{x})$. But perhaps the best and most direct measure of whether $\delta(\mathbf{x})$ is sufficiently smooth to allow identifiability is to simply look at the posterior standard deviation of $\boldsymbol{\theta}$, which directly reflects the impact of the smoothness of $\delta(\mathbf{x})$ on identifiability. However, *prior* to collecting the physical experimental data, it may be difficult to accurately gauge the smoothness of $\delta(\mathbf{x})$. In this case, if one specifies a specific $\boldsymbol{\omega}^{\delta}$ (or a range of $\boldsymbol{\omega}^{\delta}$ with some prior distribution assigned to it) to use in the preposterior algorithm, then the estimated preposterior standard deviation of $\boldsymbol{\theta}$ should be interpreted as what would result if the discrepancy function truly had the assumed degree of smoothness. One should keep in mind that this complication (not knowing the actual smoothness of $\delta(\mathbf{x})$ until after the experimental data are collected, and hence not knowing how good the design will be until after the experiment is conducted) is certainly not unique to the calibration problem. In any experimental design problem, one never knows with certainty whether the design is good until after the experiment is conducted. For example, an experiment designed for estimating a linear model will end up being a very poorly designed experiment if, after conducting the experiment, it is discovered that the true response surface is really quadratic. The prevailing viewpoint in experimental design is to accept the fact that one may need to conduct a follow-up experiment based on what was observed in the initial experiment, and this same viewpoint applies to the calibration problem: If one designs a calibration experiment based on an assumed $\boldsymbol{\omega}^{\delta}$, and after conducting the experiment it is discovered that the actual $\boldsymbol{\omega}^{\delta}$ is much different, then one could design and conduct a better follow-up experiment based on the improved knowledge of $\delta(\mathbf{x})$ learned from the initial experiment.

As noted by a referee, for the examples summarized in Figure 8 there was little further improvement in the preposterior standard deviation when N_e increased beyond the number of settings for

x used in the computer experiment; and for the grid designs in those examples the x settings for the computer and physical experiments coincided. This suggests that some form of nested designs such as those proposed in Qian (2009) and Qian, et al. (2009) may be useful designs for calibration purposes, a point that was also noted by Han, et al. (2009). There are certainly other issues that would influence which design would provide the lowest posterior standard deviation for a particular calibration problem. For example, Arendt et al. (2012a, 2012b) have shown that the posterior standard deviation depends strongly on which response variables are measured experimentally when there are multiple responses, and they have also shown that regions of the $\mathbf{x}$ space over which the simulation response $y^m(\mathbf{x},\boldsymbol{\theta})$ depends more strongly on $\boldsymbol{\theta}$ are prime candidates for input settings at which to run physical experiments. A proper investigation of the relative merits of nested designs versus other designs in general should use existing methods (e.g., the Bayesian analyses of Kennedy and O'Hagan (2001), Williams, et al. (2006), Higdon, et al. (2004), Loeppky, et al. (2006), Bayarri, et al. (2007), Han, et al. (2009); etc.) to evaluate the actual posterior standard deviation and/or Bayesian point estimates of $\boldsymbol{\theta}$ that result *after* conducting experiments with various designs. The preposterior analysis of this paper is not intended to serve as a measure of performance after the design is conducted or as a method of comparing the general performance of different classes of experimental designs over a broad range of examples. It is only intended to serve as a rough proxy for performance that can be calculated prior to conducting the physical experiment, to aid in selecting a reasonable design for a particular problem. The examples of Section 5 illustrate how the preposterior standard deviation can be used for these purposes. For example, Figure 8 indicates that the designs with N_e = 4 and 8 are reasonable designs for the beam and sinusoidal examples, respectively, in the sense that the preposterior standard deviation does not further decrease if N_e is further increased. The actual posterior standard deviations shown in Figures 4, 7, 9, and 10 confirm that the relative performance differences between the different designs were accurately predicted by the preposterior standard deviations.

In the examples, we have used the preposterior standard deviation to compare and choose from among a small set of experiments that were designed using other methods. The examples considered evenly spaced grid designs and Latin hypercube designs of various sizes and focused on selecting the appropriate size. More generally, the algorithm in Figure 1 could be used directly to compare any set of given designs (one would specify the input settings for each design in Step 0b). The approach could in principle also be used as a formal design optimization criterion to customize the exact values for the experimental **x** settings, choose which of the simulation response variables to measure experimentally, etc. However, in its current form the computational expense of the preposterior algorithm is prohibitive for these purposes. For the beam (Section 3.1) and sinusoidal (Section 5.2) examples, each MC replicate (i.e., each iteration of Steps 1—7 in Figure 1) took roughly 0.018 minutes on average on a single-processor, 3 GHz, i7 machine. The computational expense is largely independent of the size N_e of the physical experiment if N_e is less than N_m, as will often be the case. Because Steps 1—7 account for most of the computational expense of the entire algorithm, the overall expense is roughly proportional to N_{mc}. For example, using 1,700 MC replicates for the beam example, the entire algorithm took roughly 31 minutes to calculate the preposterior standard deviation for each experimental design. Using 1,700 MC replicates was more than sufficient for this example, and we found no appreciable difference in the results when we increased to 8,500 MC replicates. The computational expense will also increase with the number of variables. For the example depicted in Figures 12 and 13, each MC replicate took roughly 0.028 minutes on average, and this was also largely independent of N_e. In order to use the approach in a formal design optimization algorithm, one would have to calculate the preposterior standard deviation for a great many different candidate designs as the exact locations of the **x** settings are varied. Hence, further research on reducing computational expense is needed to adapt the algorithm for these purposes. We are currently investigating methods of improving the computational expense, for example taking into account the fact that the computer simulation design is the same for each physical experimental design being evaluated.

Acknowledgements

Support for this work from the National Science Foundation (CMMI-1233403, CMMI-0928320 and CMMI-0758557) and the U.S. Army Tank-Automotive Research Development & Engineering Center (TARDEC) (contract number W911NF11D0001-0037) is gratefully acknowledged.

References

Akkaram, S., Agarwal, H., Kale, A., and Wang, L. (2010), "Meta Modeling Techniques and Optimal Design of Experiments for Transient Inverse Modeling Applications," *ASME International Design Engineering Technical Conference*.

Arendt, P., Apley, D., and Chen, W. (2012a), "Improving Identifiability in Model Calibration Using Multiple Responses," *Journal of Mechanical Design*, 134, 100909.

Arendt, P., Apley, D., and Chen, W. (2012b), "Quantification of Model Uncertainty: Calibration, Model Discrepancy, and Identifiability," *Journal of Mechanical Design*, 134, 100908.

Kumar, A. (2008), "*Sequential Calibration of Computer Models*," The Ohio State University, Statistics.

Bayarri, M. J., Berger, J. O., Paulo, R., Sacks, J., Cafeo, J. A., Cavendish, J., Lin, C. H., and Tu, J. (2007), "A Framework for Validation of Computer Models," *Technometrics*, 49, 138-154.

Berger, J. O. (1985), *Statistical Decision Theory and Bayesian Analysis*, New York, NY: Springer-Verlag.

Carlin, B. P., and Louis, T. A. (2000), "Empirical Bayes: Past, Present and Future," *Journal of the American Statistical Association*, 95, 1286-1289.

Cressie, N. (1993), *Statistics for Spatial Data*, New York: Wiley.

Drignei, D. (2009), "A Kriging Approach to the Analysis of Climate Model Experiments," *Journal of Agricultural, Biological, and Environmental Statistics*, 14, 99-112.

Han, G., Santner, T. J., and Rawlinson, J. J. (2009), "Simultaneous Determination of Tuning and Calibration Parameters for Computer Experiments," *Technometrics*, 51, 464-474.

Handcock, M., and Stein, M. (1993), "A Bayesian Analysis of Kriging," *Technometrics*, 35, 403-410.

Higdon, D., Gattiker, J., Williams, B., and Rightley, M. (2008), "Computer Model Calibration Using High-Dimensional Output," *Journal of the American Statistical Association*, 103, 570-583.

Higdon, D., Kennedy, M. C., Cavendish, J., Cafeo, J., and Ryne, R. (2004), "Combining Field Data and Computer Simulations for Calibration and Prediction," *SIAM Journal on Scientific Computing*, 26, 448-466.

Hoffman, R., Sudjianto, A., Du, X., and Stout, J. (2003), "Robust Piston Design and Optimization Using Piston Secondary Motion Analysis," *SAE Paper (Paper No. 2003-01-0148)*.

Huan, X., and Marzouk, Y. (2011), "Simulation-Based Optimal Bayesian Experimental Design for Nonlinear Systems," *arXiv:1108.4146v1*.

Kennedy, M. C., and O'Hagan, A. (2001), "Bayesian Calibration of Computer Models," *Journal of the Royal Statistical Society: Series B*, 63, 325-364.

Loeppky, J., Bingham, D., and Welch, W. (2006), "Computer Model Calibration or Tuning in Practice," Technical Report, University of British Columbia.

Maheshwari, A. K., Pathak, K. K., Ramakrishnan, N., and Narayan, S. P. (2010), "Modified Johnson-Cook Material Flow Model for Hot Deformation Processing," *Journal of Material Science*, 45, 859 - 864.

Malhotra, R., Zue, L., Belytschko, T., and Cao, J. (2012), "Mechanics of Fracture in Single Point Incremental Forming," *Journal of Materials Processing Technology*, 212, 1573-1590.

Montgomery, D. (2005), *Design and Analysis of Experiments*, Hoboken, NJ: John Wiley & Sons, Inc.

O'Hagan, A. (1978), "Curve Fitting and Optimal Design for Prediction," *Journal of the Royal Statistical Society: Series B*, 40, 1-41.

Qian, P. Z. G. (2009), "Nested Latin Hypercube Designs," *Biometrika*, 96, 957 - 970.

Qian, P. Z. G., Tang, B., and Wu, C. F. J. (2009), "Nested Space-Filling Designs for Computer Experiments with Two Levels of Accuracy," *Statistica Sinica*, 19, 287 - 300.

Ranjan, P., Lu, W., Bingham, D., Reese, S., Williams, B., Chou, C., Doss, F., Grosskopf, M., and Holloway, J. (2011), "Follow-up Experimental Designs for Computer Models and Physical Processes," *Journal of Statistical Theory and Practice*, 5, 119-136.

Rasmussen, C. E. (1996), "*Evaluation of Gaussian Processes and Other Methods for Non-Linear Regression*," University of Toronto, Computer Science.

Rasmussen, C. E., and Williams, C. K. I. (2006), *Gaussian Processes for Machine Learning*, ed. T. Dietterich, Cambridge, Massachusetts: The MIT Press.

Reese, C. S., Wilson, A. G., Hamada, M., and Martz, H. F. (2004), "Integrated Analysis of Computer and Physical Experiments," *Technometrics*, 46, 153-164.

Robert, C., and Casella, G. (1999), *Monte Carlo Statistical Methods*, New York, NY: Springer.

Song, E. and Nelson, B. L. (2015), "Quickly Assessing Contributions to Input Uncertainty," *IE Transactions*, 47, 893–909.

Schabenberger, O., and Gotway, C. (2005), *Statistical Methods for Spatial Data Analysis* eds. B. Carlin, C. Catfield, M. Tanner and J. Zidek, Boca Raton, Florida: Chapman & Hall/CRC.

Tuo, R., and Wu, C. F. J. (2013), "A Theoretical Framework for Calibration in Computer Models Parameterization, Estimation, and Convergence Properties," *Submitted for Publication*.

Williams, B., Higdon, D., Gattiker, J., Moore, L. M., McKay, M. D., and Keller-McNulty, S. (2006), "Combining Experimental Data and Computer Simulations, with an Application to Flyer Plate Experiments," *Bayesian Analysis*, 1, 765-791.

Williams, B., Loeppky, J., Moore, L., and Macklem, M. (2011), "Batch Sequential Design to Achieve Predictive Maturity with Calibrated Computer Models," *Reliability Engineering & System Safety*, 96, 1208-1219.

Wu, J., and Hamada, M. (2000), *Experiments: Planning, Analysis, and Optimization*, Hoboken, NJ: John Wiley & Sons, Inc.