

# Bayesian Calibration of the QASPR Simulation

John McFarland\* and Sankaran Mahadevan†

*Vanderbilt University, Nashville, TN 37235*

Laura Swiler‡ and Anthony Giunta§

*Sandia National Laboratories,¶ Albuquerque, NM 87185*

This report discusses the application of a Bayesian calibration analysis to data from the QASPR project at Sandia. The goal is to use experimental measurements of a response value to obtain information about the “best” values for some of the inputs which go into the corresponding simulator. The simulator is treated as an expensive black-box model, so that only a finite number of runs are available. Towards this end, a fast Gaussian process response surface approximation is used as an emulator for the simulator. It is shown how the Bayesian framework will allow us to explicitly account for uncertainty present in the experiments, the response surface approximation, and the results. In addition to simple point estimates, we can obtain information on marginal and joint confidence intervals/regions as well as correlations among the various parameters. A method for handling multiple, correlated, response measures (such as occur over time) is also developed.

## Nomenclature

$Y_{obs}$	Experimental measurement of response
$M(\cdot)$	Simulator output
$\hat{M}(\cdot)$	Response approximation to simulator output
$\mathbf{x}$	Scenario descriptor inputs to simulator
$\boldsymbol{\theta}$	Calibration inputs to simulator
$N(\mu, \sigma)$	Normal probability distribution with mean $\mu$ and standard deviation $\sigma$
$N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	Multivariate normal distribution of dimension $p$ , with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$
$\pi(\cdot)$	Prior probability distribution
$L(Y   \boldsymbol{\theta})$	Likelihood function for $\boldsymbol{\theta}$ , based on observed data $Y$

### *Gaussian process modeling*

$n$	Number of training points
$d$	Number of input variables
$q$	Number of trend basis functions
$Y$	Response value
$\mathbf{x}$	Vector of input variables
$\mathbf{f}(\mathbf{x})$	Vector of $q$ trend basis functions
$\boldsymbol{\beta}$	Coefficients of the trend function
$\sigma^2$	Process variance of the GP
$\boldsymbol{\xi}$	Parameters of the correlation function

---

\*Graduate Student, Department of Mechanical Engineering, Student Member AIAA.

†Professor, Department of Civil and Environmental Engineering, Member AIAA.

‡Principal Member of Technical Staff, Optimization and Uncertainty Estimation Dept., Member AIAA.

§Principal Member of Technical Staff, Validation and Uncertainty Quantification Dept., Senior Member AIAA.

¶Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed-Martin Company, for the United States Department of Energy under Contract DE-AC04-94AL85000.

## I. Introduction

This report discusses the results of a “calibration” analysis of data from the QASPR (Qualification Alternatives to the Sandia Pulsed Reactor) project at Sandia National Laboratories, which models the effects of radiation on electronics. Although the term calibration has had various interpretations, a specific meaning will be used here. By calibration, we mean using experimental observations of a response value to inform upon, or learn about, uncertain input parameters which go into a computational simulation.

The science of model calibration, where the calibration analysis accounts for uncertainty and/or variability, is relatively immature. See Refs. 1,2 for overviews, and Refs. 3–5 for examples. Whereas most of the simpler calibration methods are based on an optimization search for “best-fitting” values of the inputs, there is much extra insight that can be had by approaching the problem from a probabilistic standpoint. By doing so, we will be able to account for uncertainties which inevitably exist in the experimental observations, the response surface approximation (if one is used), and also the results of the calibration. Thus, we can obtain such information as confidence intervals for the calibrated parameters, as well as interaction information, such as correlations.

It turns out the Bayesian framework provides a natural method for performing a calibration analysis in a probabilistic setting. Not only can we account for the various uncertainties mentioned above, but we will also be able to incorporate any prior information we may have about the parameters being calibrated. The landmark paper dealing with Bayesian calibration was published by Kennedy and O’Hagan,<sup>3</sup> and their particular Bayesian model for calibration is often referred to as the “Kennedy and O’Hagan” framework.

Section II will introduce the Bayesian formulation for calibration, and Section III describes the implementation of the Gaussian process response surface approximation model. Section IV discusses the results of the application to data obtained from the QASPR project at Sandia. This will include several different analyses, including a study of the effect of the various boundary conditions and response measures. Finally, Section IV.C illustrates a method for calibrating based on multiple (possibly correlated) response values simultaneously. Additional analyses based on the results of Section IV are discussed in Section V.

## II. The Bayesian model for calibration

Kennedy and O’Hagan<sup>3</sup> begin by defining a very useful classification of the parameters which are inputs to the simulation. They divide these inputs into two categories: “variable inputs” (what we will call “scenario descriptors”), denoted by  $\mathbf{x}$ , and “calibration inputs”, denoted by  $\boldsymbol{\theta}$ . The scenario descriptor inputs are things like boundary or initial conditions, which are assumed to have known, and possibly even controllable, values for each of the experiments. The calibration inputs are those that we wish to learn about, and we want to think of them as having fixed but unknown values. Under this framework, we think of the calibration inputs as being constant across the various scenarios, and not observable in the experiments.

The probabilistic model used by Kennedy and O’Hagan is

$$Y_{obs}^{(i)} = \rho M(\mathbf{x}_i, \boldsymbol{\theta}) + \delta(\mathbf{x}_i) + \epsilon_i, \quad (1)$$

where  $Y_{obs}^{(i)}$  is the  $i$ th observation of the response value,  $\rho$  is an unknown scale parameter,  $M(\mathbf{x}_i, \boldsymbol{\theta})$  denotes the response of the simulator (model) for inputs  $\mathbf{x}_i$  and  $\boldsymbol{\theta}$ ,  $\delta(\cdot)$  is the “model inadequacy function”, and  $\epsilon_i$  is a random variable representing measurement error. Kennedy and O’Hagan assume the  $\epsilon_i$  to be independently normally distributed as  $N(0, \sigma_{exp}^2)$ .

The probabilistic model implemented in the following analyses is a simplification of the above model, so that we have

$$Y_{obs} = M(\boldsymbol{\theta}) + \epsilon. \quad (2)$$

(For now, we disregard the scenario variables,  $\mathbf{x}$ , and consider that we are only calibrating based on observations for one scenario.) We have dropped the multiplicative and additive “inadequacy” terms  $\rho$  and  $\delta(\cdot)$ . We are thus assuming that for some “good” value of  $\boldsymbol{\theta}$ , our model is an accurate representation of reality. We will again assume that the measurement error  $\epsilon$  is normally distributed with a known variance,  $\sigma_{exp}^2$ .

The model given by Eq. (2) implies what is known as a likelihood function for the unknown parameters  $\boldsymbol{\theta}$ , based on the observed data  $Y_{obs}$ . Since the observational data are fixed, the likelihood function depends on  $\boldsymbol{\theta}$  only and is commonly written  $L(\boldsymbol{\theta})$ ; for clarity, however, we express it here as  $L(Y_{obs} | \boldsymbol{\theta})$ . This notation emphasizes that the likelihood function is based on the distribution of the observed data. For the model given by Eq. (2), the likelihood function comes from the following simple result:

$$Y_{obs} | \boldsymbol{\theta} \sim N(M(\boldsymbol{\theta}), \sigma_{exp}^2). \quad (3)$$

By applying Bayes' theorem, we obtain the following posterior distribution for the calibration inputs:

$$f(\boldsymbol{\theta} | Y_{obs}) \propto \pi(\boldsymbol{\theta})L(Y_{obs} | \boldsymbol{\theta}), \quad (4)$$

where  $\pi(\boldsymbol{\theta})$  is the prior distribution for  $\boldsymbol{\theta}$ .

Although it is necessary to specify some joint prior distribution for the calibration inputs, this does not mean that it is necessary to incorporate prior knowledge. It is possible to use vague, non-informative prior distributions to represent a complete lack of knowledge with respect to the inputs, before observing the outputs. For inputs which have support over the entire real line, this is most commonly done using independently uniform prior distributions, so that the joint prior distribution for all calibration inputs is

$$\pi(\boldsymbol{\theta}) = \text{constant}. \quad (5)$$

This prior distribution will result in a posterior which is exactly proportional to the likelihood function, so that all of our knowledge comes from the data.

We may also incorporate bounds for any of the inputs into this prior distribution, so that we have

$$\pi(\boldsymbol{\theta}) = \begin{cases} \text{constant}, & \boldsymbol{\theta} \in \Omega \\ 0, & \boldsymbol{\theta} \notin \Omega \end{cases} \quad (6)$$

where the region  $\Omega$  defines the bounds for the calibration inputs  $\boldsymbol{\theta}$ . This formulation will still yield a posterior that is proportional to the likelihood, but it will also require that the posterior distribution lies inside  $\Omega$ .

The process of obtaining the posterior distributions of the input variables is undertaken using Markov Chain Monte Carlo (MCMC) sampling (specifically, the Metropolis algorithm is used; refer to Refs. 6 and 7, and also to the Appendix for computational notes). This requires thousands of evaluations of the model. Since the model is typically expensive to evaluate (as with the QASPR model), a surrogate model, or response surface approximation, may be used. For this analysis, a Gaussian process (Kriging) response surface is implemented, and its parameters are estimated using the method of maximum likelihood (see Section III).

One advantage of the Gaussian process response surface approximation model is that it gives information regarding the usefulness of each of the inputs in predicting the output. This information is obtained through the maximum likelihood estimates of the correlation parameters. Standardizing each input variable to have the same variance will allow for a meaningful comparison of these sensitivities with each other. However, since the likelihood function tends to be multimodal, the optimal parameters are sensitive to the starting point used in the optimization algorithm. In addition, each optimal set of parameters will indicate somewhat different "sensitivities". However, the most important variables will usually be evident from one run of the MLE code. This report will briefly explore the input/output sensitivity for the QASPR model, as well as its dependence on the starting point of the MLE optimization algorithm.

Also, the uncertainty due to the predictions given by the Gaussian process model can be explicitly accounted for and incorporated into the Bayesian analysis. Thus, we replace the true model  $M(\boldsymbol{\theta})$  by the Gaussian process model surrogate,  $\hat{M}(\boldsymbol{\theta})$ , which is not deterministic, but is itself a random variable. The distribution of  $\hat{M}(\boldsymbol{\theta})$  is defined by the parameters of the Gaussian process, and is written as

$$\hat{M}(\boldsymbol{\theta}) \sim N(\mu_{GP}(\boldsymbol{\theta}), \sigma_{GP}^2(\boldsymbol{\theta})). \quad (7)$$

( $\mu_{GP}(\cdot)$  and  $\sigma_{GP}^2(\cdot)$  are defined in Section III by Eqs. (13) and (14).) In many applications,  $\mu_{GP}(\boldsymbol{\theta})$  alone is used as the emulator for the true model, but the Bayesian analysis allows us to incorporate the uncertainty associated with each prediction,  $\sigma_{GP}^2(\boldsymbol{\theta})$ . Thus, by replacing the slow simulator by a Gaussian process surrogate model, we have the new probabilistic model

$$Y_{obs} = \hat{M}(\boldsymbol{\theta}) + \epsilon. \quad (8)$$

Since  $\hat{M}(\cdot)$  and  $\epsilon$  have independent normal distributions, it is easy to show that the likelihood function is now based on the following distribution:

$$Y_{obs} | \boldsymbol{\theta} \sim N(\mu_{GP}(\boldsymbol{\theta}), \sigma_{GP}^2(\boldsymbol{\theta}) + \sigma_{exp}^2). \quad (9)$$

### III. Implementation of the Gaussian process response surface approximation

Gaussian process models have several features which make them an attractive choice for a response surface approximation. The primary feature of interest is the ability of the model to “account for its own uncertainty”. That is, each prediction obtained from a Gaussian process model also has an associated variance, or uncertainty. This prediction variance primarily depends on the closeness of the prediction location to the training data, but it is also related to the functional form of the response.

#### III.A. Prediction with a Gaussian process model

The basic idea of the GP model is that the response values,  $Y$ , are modeled as a group of multivariate normal random variables. A parametric covariance function is then constructed as a function of the inputs,  $\mathbf{x}$ . The covariance function is based on the idea that when the inputs are close together, the correlation between the outputs will be high. As a result, the uncertainty associated with the model’s predictions is small for input values which are close to the training points, and large for input values which are not close to the training points. In addition, the GP model may incorporate a systematic trend function, such as a linear or quadratic regression of the inputs. The effect of the mean function on predictions is typically only important when the model is used for extrapolation.

We will denote by  $Y$  a Gaussian process with mean and covariance given by

$$E[Y(\mathbf{x})] = \mathbf{f}^T(\mathbf{x})\boldsymbol{\beta} \quad (10)$$

and

$$\text{Cov}[Y(\mathbf{x}), Y(\mathbf{x}^*)] = \sigma^2 c(\mathbf{x}, \mathbf{x}^* | \boldsymbol{\xi}), \quad (11)$$

where  $\mathbf{f}^T(\mathbf{x})$  defines  $q$  basis functions for the trend, and is given by 1 for a constant trend and  $[1 \ \mathbf{x}^T]^T$  for a linear trend,  $\boldsymbol{\beta}$  gives the coefficients of the regression trend,  $c(\mathbf{x}, \mathbf{x}^* | \boldsymbol{\xi})$  is the correlation between  $\mathbf{x}$  and  $\mathbf{x}^*$ , and  $\boldsymbol{\xi}$  is the vector of parameters governing the correlation function.

Consider that we have observed the process at  $n$  locations (the training or design points)  $\mathbf{x}_1, \dots, \mathbf{x}_n$  of a  $d$ -dimensional input variable, so that we have the resulting observed random vector  $\mathbf{Y} = (Y(\mathbf{x}_1), \dots, Y(\mathbf{x}_n))^T$ . By definition, the joint distribution of  $\mathbf{Y}$  satisfies

$$\mathbf{Y} \sim \mathbf{N}_d(\mathbf{f}^T(\mathbf{x})\boldsymbol{\beta}, \sigma^2 \mathbf{R}), \quad (12)$$

where  $\mathbf{R}$  is the  $n \times n$  matrix of correlations between the training points. Under the assumption that the parameters governing both the trend function and the covariance function are known, the expected value and variance (uncertainty) at an untested location  $\mathbf{x}^*$  are calculated as

$$E[Y(\mathbf{x}^*)] = \mathbf{f}^T(\mathbf{x}^*)\boldsymbol{\beta} + \mathbf{r}^T(\mathbf{x}^*)\mathbf{R}^{-1}(\mathbf{Y} - \mathbf{F}\boldsymbol{\beta}) \quad (13)$$

and

$$\text{Var}[Y(\mathbf{x}^*)] = \sigma^2 (1 - \mathbf{r}^T \mathbf{R}^{-1} \mathbf{r}), \quad (14)$$

where  $\mathbf{F}$  is an  $n \times q$  matrix with rows  $\mathbf{f}^T(\mathbf{x}_i)$  (the trend basis functions at each of the training points), and  $\mathbf{r}$  is the vector of correlations between  $\mathbf{x}^*$  and each of the training points. When the coefficients of the trend function are not known, but are estimated using a generalized least squares procedure, or equivalently, maximum likelihood, the variance becomes:<sup>8</sup>

$$\text{Var}[Y(\mathbf{x}^*)] = \sigma^2 \left\{ 1 - \mathbf{r}^T \mathbf{R}^{-1} \mathbf{r} + \left[ \mathbf{f}(\mathbf{x}^*) - \mathbf{F}^T \mathbf{R}^{-1} \mathbf{r} \right]^T \left[ \mathbf{F}^T \mathbf{R}^{-1} \mathbf{F} \right]^{-1} \left[ \mathbf{f}(\mathbf{x}^*) - \mathbf{F}^T \mathbf{R}^{-1} \mathbf{r} \right] \right\}, \quad (15)$$

which can also be written as

$$\text{Var}[Y(\mathbf{x}^*)] = \sigma^2 - \left[ \mathbf{f}^T(\mathbf{x}^*), \mathbf{r}^T \right] \begin{bmatrix} \mathbf{0} & \mathbf{F}^T \\ \mathbf{F} & \mathbf{R} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{f}(\mathbf{x}^*) \\ \mathbf{r} \end{bmatrix}. \quad (16)$$

There are several different methods of parameterizing the correlation function. The form implemented by this author is the squared exponential form, given by

$$c(\mathbf{x}, \mathbf{x}^*) = \exp \left[ - \sum_{i=1}^d \xi_i (x_i - x_i^*)^2 \right], \quad (17)$$

where  $d$  is the dimension of  $\mathbf{x}$ , and the  $d$  parameters  $\xi_i$  must be non-negative. The exponent must lie in the range  $[0, 2]$  in order for the covariance matrix to be positive definite, but the value 2 is usually chosen because it produces a function that is infinitely differentiable. This form of the correlation function dictates that the degree of correlation of the outputs depends on the closeness of the inputs.

Also, the relative magnitudes of the parameters  $\xi$  specify the amount of importance each dimension of  $\mathbf{x}$  has in predicting the output  $Y$ : a large value for  $\xi_i$  (which is akin to a small correlation length) indicates a high amount of “activity” (and likewise a low amount of correlation) in that direction. For example, if the response is independent of one of the inputs, then that input will have an infinite correlation length (because the response does not change in that direction) and a  $\xi$  of 0. Thus, the relative magnitudes of the  $\xi_i$  connote the global sensitivity of the response to the inputs.

Finally, a constant trend function is implemented for all GP models reported here. This is because no systematic trends are evident from simple input/output plots, and the constant trend gave the best performance over linear and quadratic trends when compared using a simple cross validation exercise.

### III.B. Parameter Estimation

Before applying the Gaussian process model for prediction, values for the parameters  $\sigma^2$ ,  $\xi$ , and  $\beta$  must be set. There are several methods used in practice to estimate good values of the parameters governing the GP.<sup>9</sup> This paper will employ the method of maximum likelihood estimation (MLE).

Maximum likelihood estimation is based on the assumption that the response values follow a multivariate normal distribution. We parametrize the mean and covariance of this distribution, and then estimate the “best fitting values” of these parameters giving the training data. Thus, the likelihood function is simply given by the joint PDF of the observed responses, as in Eq. (12).

For computational reasons, it is easier to work with the log of the likelihood when performing maximum likelihood estimation. We express our objective function using the negative log of the likelihood (in consideration of minimization), and dropping constants, as

$$NL = n \log \sigma^2 + \log |\mathbf{R}| + \frac{1}{\sigma^2} (\mathbf{Y} - \mathbf{F}\beta)^T \mathbf{R}^{-1} (\mathbf{Y} - \mathbf{F}\beta). \quad (18)$$

The gradients of  $NL$  with respect to  $\sigma^2$ ,  $\mathbf{x}_i$ , and  $\beta$  can all be derived analytically.<sup>9,10</sup> Further, we can solve for the zero gradients with respect to  $\sigma^2$ , and  $\beta$ , yielding their conditional optima as:

$$\hat{\sigma}^2 = \frac{1}{n} (\mathbf{Y} - \mathbf{F}\beta)^T \mathbf{R}^{-1} (\mathbf{Y} - \mathbf{F}\beta), \quad (19)$$

and

$$\hat{\beta} = \left( \mathbf{F}^T \mathbf{R}^{-1} \mathbf{F} \right)^{-1} \mathbf{F}^T \mathbf{R}^{-1} \mathbf{Y}. \quad (20)$$

We thus minimize the negative of the “profile log likelihood”, which can be defined as

$$NL^* = n \log \hat{\sigma}^2 + \log |\mathbf{R}|, \quad (21)$$

with  $\beta$  replaced by  $\hat{\beta}$ .

The numerical minimization of Eq. (21) is carried out with respect to the parameters  $\xi$ . Since analytical gradients are also available with respect to  $\xi$ , we can make use of gradient based optimization routines. For this work, a form of the BFGS algorithm is used to find the optimum values for  $\xi$ . Recall that the values of  $\xi_i$  have an interpretation regarding the global sensitivity of the response to each input. Thus, we take the MLE estimates of each  $\xi_i$  as a “relative importance” measure of the  $i$ th input in predicting the response.

## IV. Application to QASPR data

We now discuss the application of the above calibration methodology to data from the QASPR simulation. The QASPR (Qualification Alternatives to the Sandia Pulsed Reactor) project consists of several levels of code from the atomic scale to the circuit level which model radiation effects on electronics. The data we work with comes from the “1D-code”, which deals with effects at the device level. The response, or output, for this code is gain (a current ratio) versus time. For both the experiments and the simulations, we have the response value at four different time instances (these times correspond between the experiments and simulations).

First, consider the classification of the inputs, as defined by Kennedy and O’Hagan.<sup>3</sup> For the scenario descriptors, we have one “variable”, which takes three discrete scenarios, which we will call “Q1, Q2, and Q3”. These conditions represent three different bias voltages applied to the transistor at the time of the radiation pulse. Note that some authors might classify time as a scenario descriptor, since the response is a function of time. However, this can be dangerous because each individual experiment will yield measurements of the response at various discrete time intervals. If the experiment is repeated, the responses at these time intervals will have some correlation structure. This is not the case for the usual scenario variables, because each scenario will correspond to a completely different experiment. Thus, we do not want to assume that experimental measurements of the response at different times are independent.

As for the calibration inputs, the 1D-code has 12 variables (reduced from an originally much larger number, using a preliminary sensitivity analysis). Our prior information on these parameters consists only of the reasonable bounds for each. For each of the Q1, Q2, and Q3 scenarios, 300 runs of the simulator are available, corresponding to randomly chosen values of  $\theta$ . Thus, we choose the prior distribution for  $\theta$  to be independently uniform over the bounds, so that we have the prior given by Eq. (6). The use of a bounded prior distribution is helpful when using a response surface approximation because the predictions given by the response surface can only be expected to be valid within the ranges of the training data (which will usually correspond to the prior bounds).

For each analysis, the value of  $\sigma_{exp}$  is set to ten percent of the corresponding experimental observation. This is admittedly arbitrary, and likely underestimates the uncertainty associated with each experimental observation. However, in many cases, the experimenter will be able to characterize, at least roughly, the uncertainty, repeatability, or error associated with his or her measurements. Such estimates could then be used to determine a value for  $\sigma_{exp}$ . Alternatively, if data are available from repeated experiments,  $\sigma_{exp}$  could be estimated directly from the data or treated as an uncertain variable inside the Bayesian analysis.

The following section will report the results for a “nominal” case. This will be for the Q1 data set, considering response 1 only. In later sections, various other scenarios are considered so that the results can be compared to the nominal case. In the ideal world, if the simulation model captures the physics correctly, the results of the calibration exercise would be similar for each data set, since in theory the “true” value of  $\theta$  does not depend on the scenario  $x$ .

### IV.A. Nominal case

For the “nominal case” we consider the Q1 data, and update the inputs based only on the observation of response 1. Since the optimal MLE parameters of the GP response surface approximation depend somewhat on the starting point given to the optimizer, some method is needed to decide which parameters will be used for the nominal GP model. Here, the decision is based both on comparing the objective function values at different starting points and cross-validation type exercises in which the model is used to predict points which are left out of its design.

The apparent sensitivities of the response to the inputs obtained from the nominal Gaussian process model correlation parameters are illustrated graphically in Figure 1. It appears that for this data set, inputs number 6 and 11 are the most important in predicting the value of response 1. Also, inputs 3, 7, 10, and 12 seem to have very little effect on the output, and could probably be removed from the analysis.

To illustrate the form of the response, several input/output plots based on the Gaussian process model are constructed. We will look at response 1 versus inputs 6 and 11 (the 2 most important inputs) and response 1 versus inputs 2 and 8 (the next 2 most important inputs). In each case, the values of the 9 other inputs are held constant<sup>a</sup>. Figure 2 plots response 1 as a function of inputs 6 and 11 as both a mesh plot and

---

<sup>a</sup>The particular values of the non-varying inputs are chosen based on the results of the following calibration analysis. The values used are the estimated joint mode of the inputs, based on the observation of response 1. This means that the response

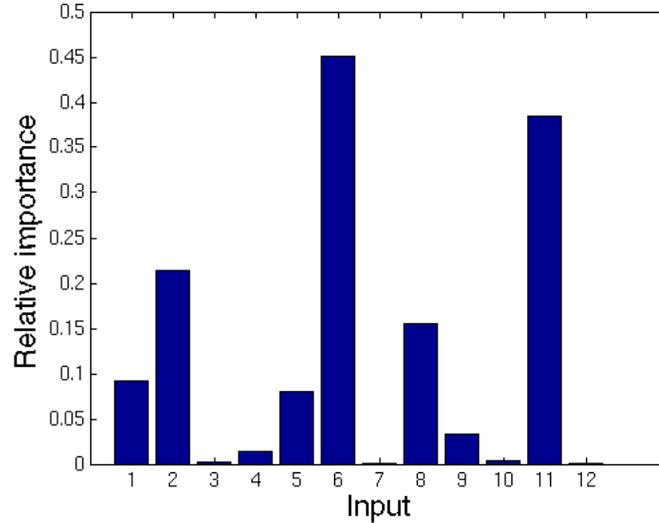


Figure 1. Sensitivity analysis based on nominal GP correlation parameters

a contour plot. The contour plot helps to illustrate the particular region of these inputs which matches well with the experimental observation (which is 0.41 for this case). A mesh/contour plot of  $\sigma_{GP}$  is also given to illustrate how the response surface approximation uncertainty varies in this domain. The corresponding 3 plots are also given for inputs 2 and 8 in Figure 3.

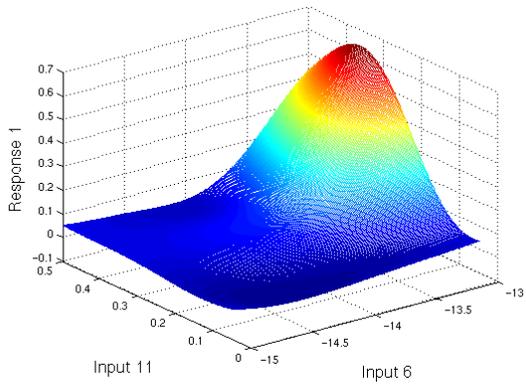
After specifying the GP response surface model, MCMC simulation is used to estimate the updated distributions of the inputs. The resulting estimated posterior probability distributions for the 12 inputs are plotted below in Figure 4. The range of each plot is the same as the prior bounds used for that input variable. The dotted black line represents the prior probability distribution, for comparison. For each input, the posterior is plotted by fitting a beta distribution to 25,000 random MCMC samples. The beta distribution is chosen because of its flexibility in fitting a variety of shapes, and also because the distribution for each variable has an upper and lower bound.

Notice that for inputs 3, 4, 7, 10, and 12, the the updated distributions have not deviated significantly from their uniform priors. This means that marginally, all values within the respective ranges are equally effective at yielding a response consistent with the observation. Also, the most profound changes are for variables 2, 5, 6, 8, and 11, indicating that these are the most important variables to consider when calibrating the model (this is as expected, given the sensitivity analysis of Figure 1). We can see that for effectively all of the variables, “better” values of the inputs correspond to larger values.

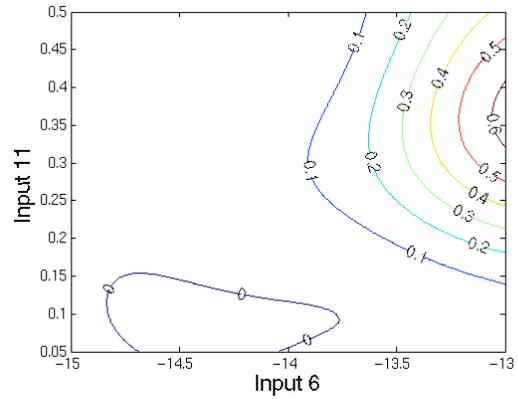
In addition, the updating of the response value is illustrated in Figure 5. This figure shows the approximate distribution of the response for the 300 true model evaluations, the assumed likelihood function for the single observation (the experimental uncertainty/variability), and the posterior distribution of the response (the posterior distribution of the response is an output of the MCMC simulation: it is the value of  $\mu_{GP}(\boldsymbol{\theta})$  for each accepted sample of  $\boldsymbol{\theta}$ ). Since we have used independent uniform priors for the inputs, we expect the posterior to be proportional to the likelihood function of the experiment. This would be the case exactly, except that we are using a response surface approximation, which has its own uncertainty (recall that this uncertainty is accounted for by the full likelihood function of Eq. (9)). The additional uncertainty added by the response surface approximation causes the variance of the posterior to be greater than the variance/uncertainty of the experiments. In addition, the posterior is “pulled” very slightly away from the data towards the area where there is less response surface approximation uncertainty. Since most of the original model runs correspond to values of the response which are less than the observation, the posterior shifts slightly away from the data in this direction.

---

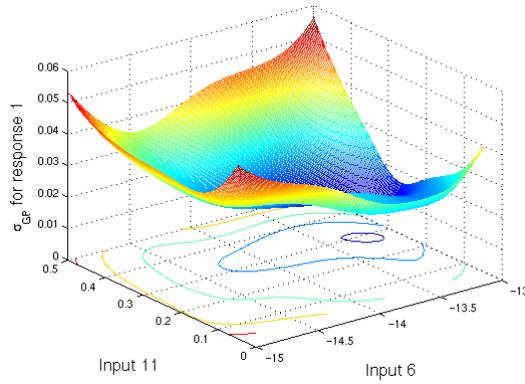
plots will be (at least somewhat) relevant to the posterior distributions of the inputs



(a) Response based on Gaussian process model ( $\mu_{GP}$ )

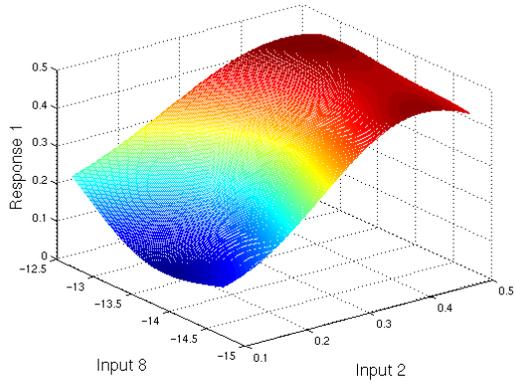


(b) Contour plot of response based on Gaussian process model ( $\mu_{GP}$ )

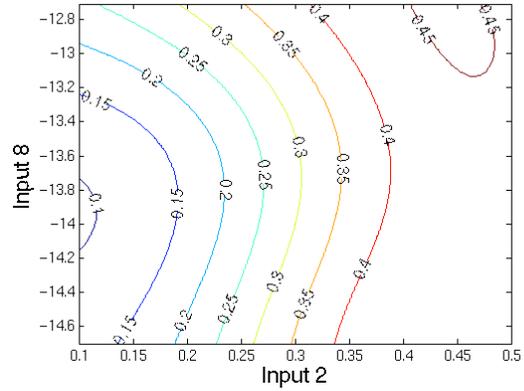


(c) Uncertainty ( $\sigma_{GP}$ ) associated with Gaussian process model

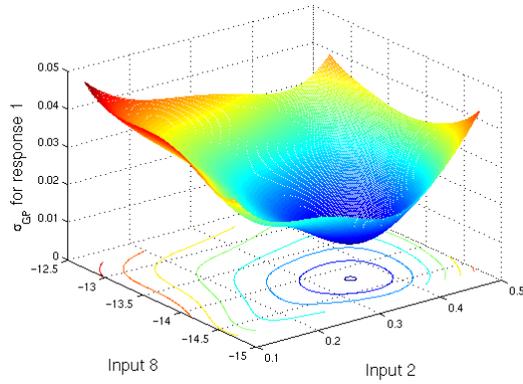
**Figure 2. Gaussian process approximation to response 1 based on inputs 6 and 11**



(a) Response based on Gaussian process model ( $\mu_{GP}$ )



(b) Contour plot of response based on Gaussian process model ( $\mu_{GP}$ )



(c) Uncertainty ( $\sigma_{GP}$ ) associated with Gaussian process model

**Figure 3. Gaussian process approximation to response 1 based on inputs 2 and 8**

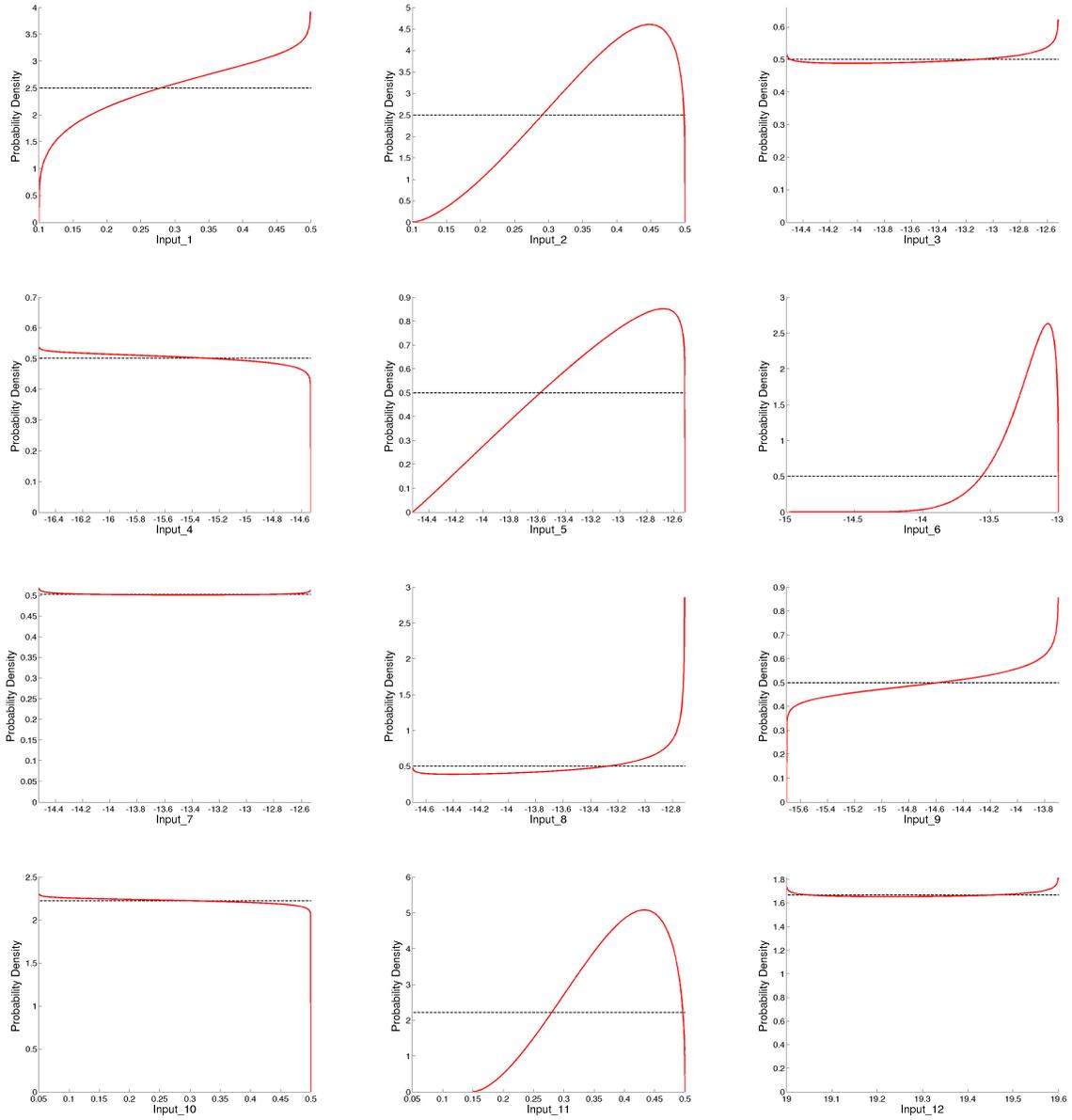


Figure 4. Posterior probability distributions for Q1 data based on response 1, one observation: “nominal case”. (Prior distributions given by dotted lines.)

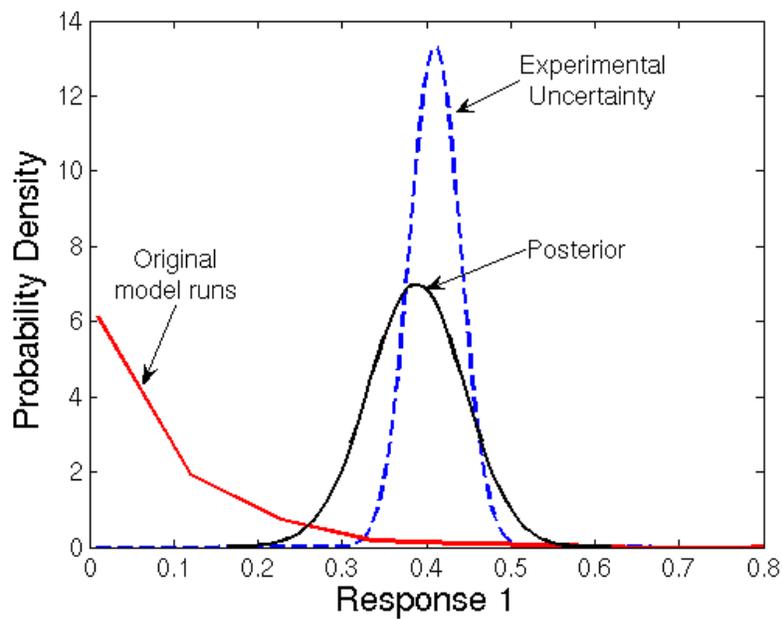


Figure 5. Distribution of original model runs, experimental uncertainty, and posterior for response 1

## IV.B. Comparative studies

In this section we consider how the results compare when we consider different parameters for the GP model, using the data for other scenarios (Q2 and Q3), and updating the inputs based on a different response. It is important to consider the effects of using different parameters for the GP model because the calibration results will depend completely on the response surface approximation, and the parameters governing the GP model found by MLE can vary depending on the starting point given to the optimizer. Further, we would like to see that the calibration results are the same regardless of what scenarios or response functions are used.

### IV.B.1. Effect of Gaussian process model parameters

Here we will re-run the “nominal” calibration process using different values for the parameters governing the GP model. The optimal parameters for the GP model in the “nominal” case were found by giving the optimizer a starting value of 0.4 for all correlation parameters. However, if this initial value is changed slightly to 0.2 or 0.6, a different local maximum of the likelihood function will be found. Thus, we want to study how much the calibration results will change if different GP parameters are used.

We will consider giving the correlation parameters,  $\xi$ , starting values of 0.2 and 0.6 in the MLE optimization routine, whereas 0.4 was used for the nominal case. The resulting MLE estimates found by the optimizer for these starting values are summarized in Figure 6. The y-axis shows the value of each  $\xi_i$  found by the optimizer, which is an indication of that inputs relative importance in predicting the output. The value of “Lhood” indicates the relative magnitude of the likelihood function, so that we can compare the quality of the various local optima.

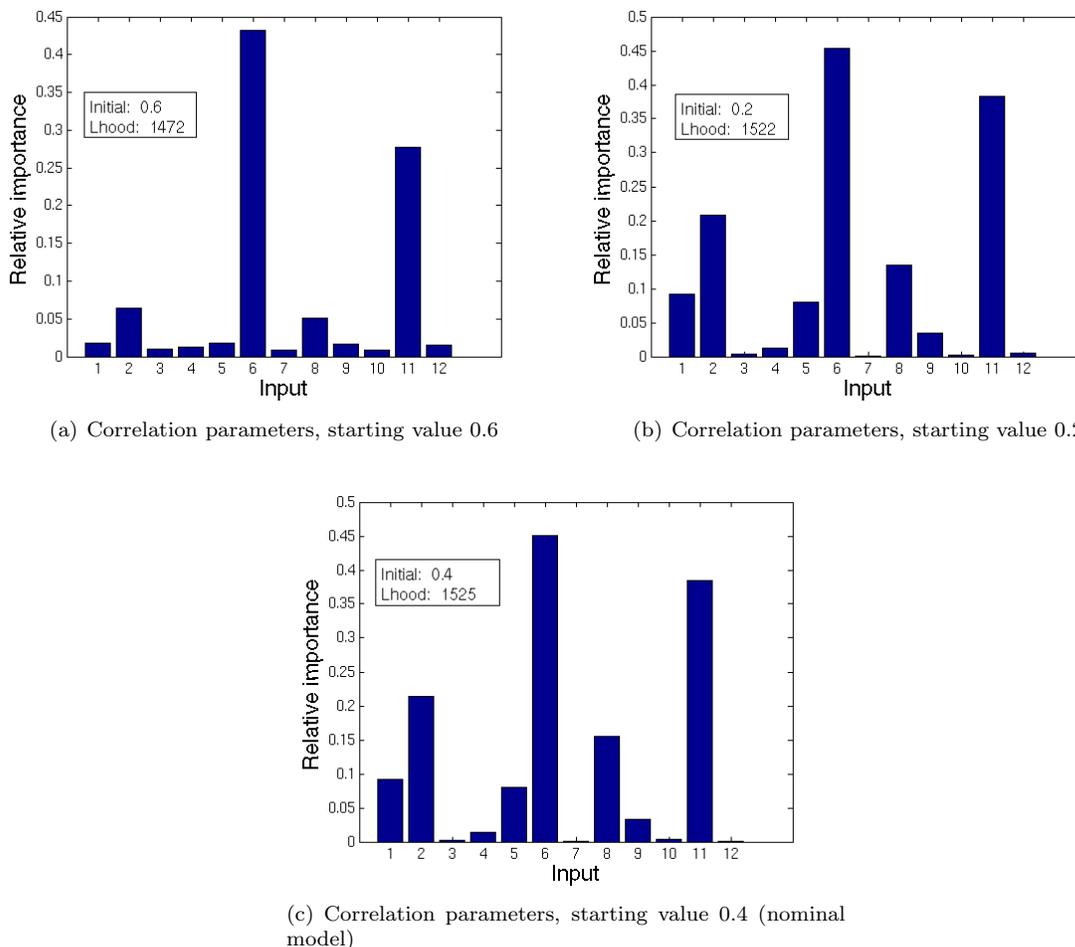


Figure 6. Comparison of optimal correlation parameters found by MLE, given different starting values

We can see that the optimal parameters are almost identical for the starting values of 0.2 and 0.4. However, a slightly different optimum is found for the starting value of 0.6. The optimum found for 0.6 has a lower likelihood, and more of the correlation parameters are given very small values, indicating they are not important. However, both this case and the nominal case still indicate that inputs 6 and 11 are the most important in predicting the output.

Since the GP parameters for the starting value of 0.2 differ significantly from the nominal case, we will perform the same calibration analysis as was done in Section IV.A and compare the results. Based on the marginal posteriors, the difference between the results based on the two different GP models is not discernible, giving an indication that the results may not be too sensitive to the formulation of the Gaussian process model.

#### IV.B.2. Effect of scenarios

Here we compare the results of calibrations based on different scenarios. The nominal case corresponds to the scenario Q1, and we will compare these results to calibrations based on data for scenarios Q2 and Q3. For all cases we use 300 runs of the simulation model, however, they do not correspond to the same values of  $\theta$ .

The results of the GP model parameter estimation are shown below in Figure 7. Here we see that the correlation parameters change somewhat when the scenario is varied. Overall, inputs 6, 11, 2, and 8 seem to be very important to the response function described by the simulation.

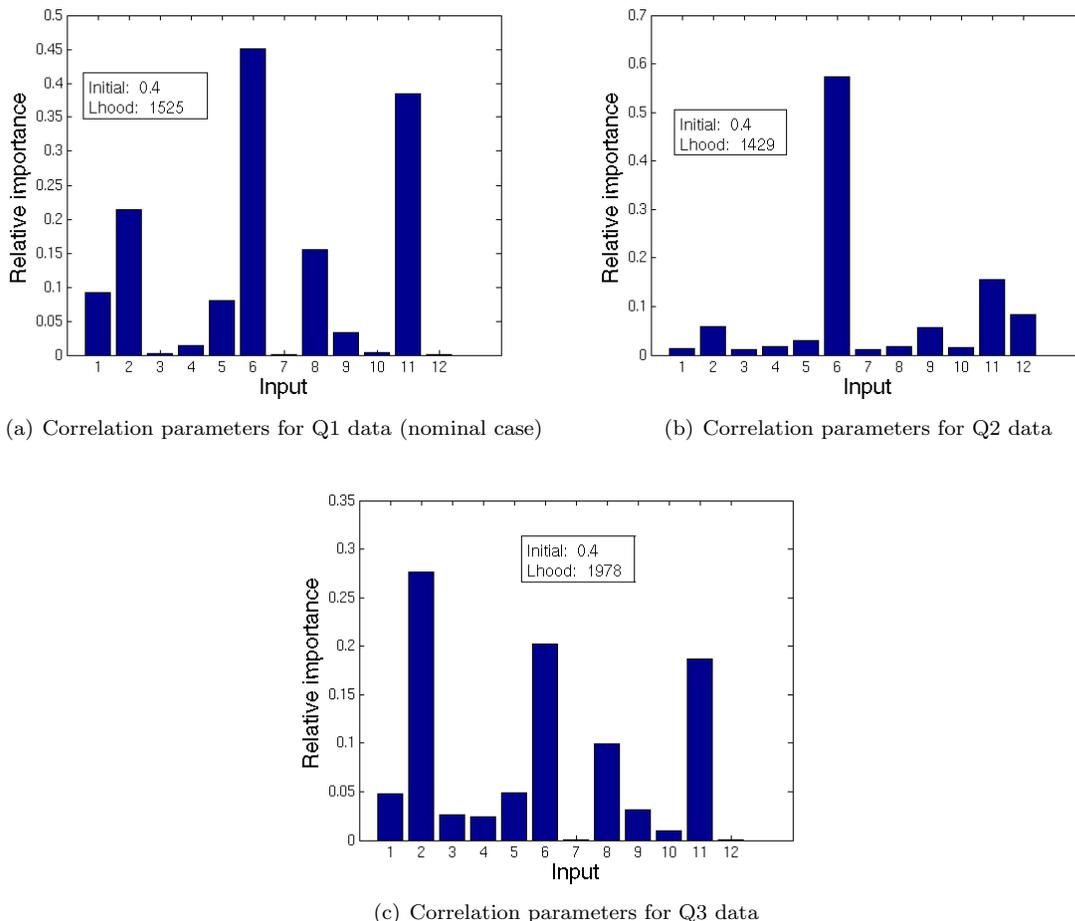


Figure 7. Comparison of correlation parameters for different scenarios

Now consider a plot of the predicted and measured responses, as a function of the scenario, shown in Figure 8. We quickly see that, on average, the simulator is under-predicting for Q1, shows little bias for Q2, and over-predicting for Q3. This indicates a strong systematic dependence of the bias on the particular

scenario. It is easy to see that, without the full Kennedy and O’Hagan model of Eq. (1), this type of bias variation will be very difficult for the Bayesian model to handle, since “good” values corresponding to one scenario may very well be “bad” for another scenario.

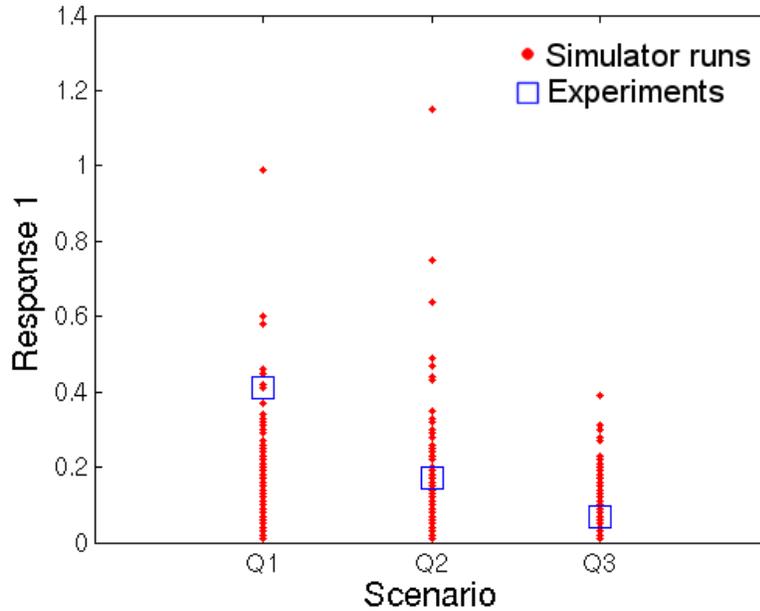


Figure 8. Predicted and measured values of response 1, as a function of the scenario

A comparison of the Bayesian calibration for each scenario is shown below in Figure 9. The results indicate that there is more disagreement here than when changing only the Gaussian process model. We see that in particular, inputs 1, 2, 5, 6, and 12 do not calibrate to the same region based on the 3 different data sets. There appear to be possible discrepancies for inputs 1 and 6. These inputs calibrate towards medium values when using the Q3 data, but they calibrate towards high values for the Q1 and Q2 data. This is not surprising, given the systematic dependence of the bias on the scenario. A single calibration based on multiple scenarios will be considered in Section IV.C.2.

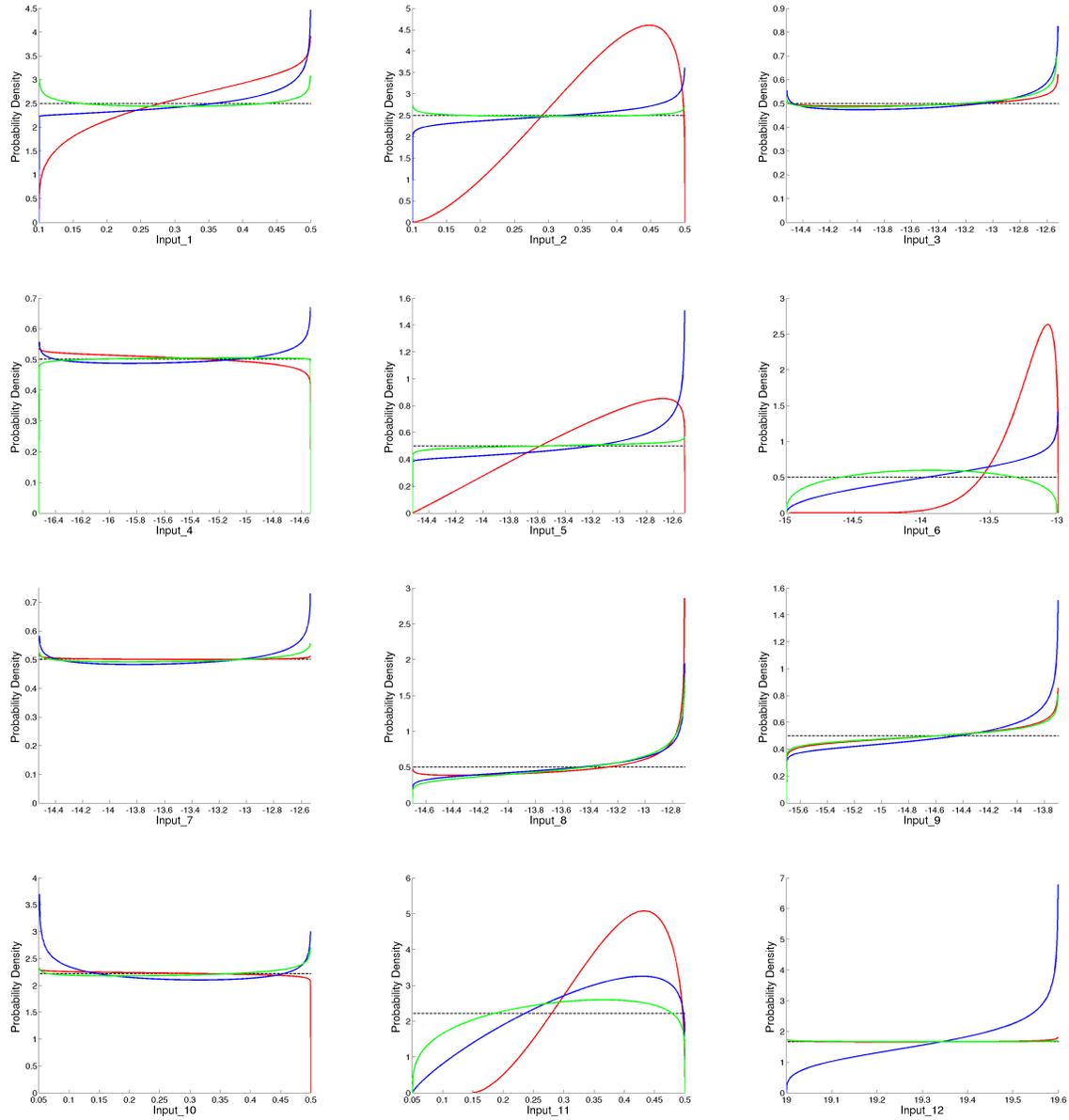


Figure 9. Comparison of calibration results for boundary conditions Q1, Q2, and Q3. The red curves represent the Q1 (nominal) condition, the blue represent Q2, and the green Q3. All results based on response 1 only.

### IV.B.3. Effect of time

In this section we consider, all else equal, the effect of updating the distributions based on response 4 as opposed to response 1. Figure 10 plots the response as a function of time for the Q1 scenario. As with the various scenarios, we see that the discrepancy between the simulator and the experiments changes over time. For “response 1” ( $10^{-4}$  seconds), the simulator tends to under-predict the experiment, but on average this bias goes towards zero as time increases. The change of bias is not as pronounced as with the various scenarios.

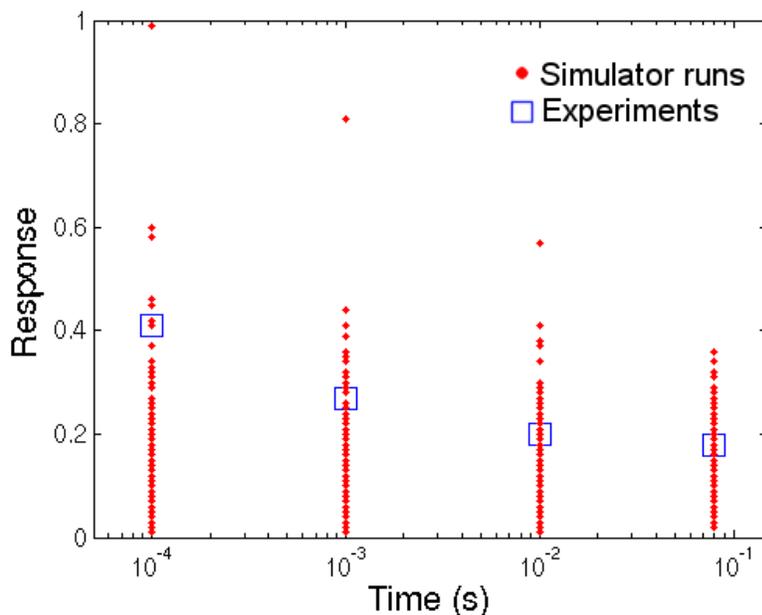


Figure 10. Predicted and measured response for Q1, as a function of time

However, we must keep in mind that these responses are measurements of the same quantity at different time instances, so we would expect the experimental measurements to have some correlation. Since there are no repeated experiments available, we estimate the correlation based on the simulator runs. Based on the 300 simulation runs, the linear correlation coefficient between responses 1 and 4 is 0.56, indicating moderate correlation. However, the two responses are related to different physics phenomena, so it is possible that the calibrated parameters based on each could be different.

The results of the GP maximum likelihood estimation are shown in Figure 11. We see that more of the inputs appear to be relevant to response 4 than response 1. In particular, input 9 appears to play a large role in predicting response 4.

The results of the Bayesian calibration, updating based on response 4 (using the Q1 data) are shown in Figure 12. The results here are similar to those of the scenario comparison, in that different inputs have become important, and these new inputs become central to the calibration updating. The main differences here are for inputs 1, 5, 6, 7, 9, and 11. When calibrating based on response 1, inputs 1, 5, 6, and 11 want to increase, whereas when using response 4, inputs 7 and 9 want to increase. But this does not necessarily mean that the two calibrations are inconsistent with each other, since flat posterior distributions indicate an input which takes on all values in the range. A joint analysis will be required to better explore how the two conditions interact with each other (see Section IV.C.1).

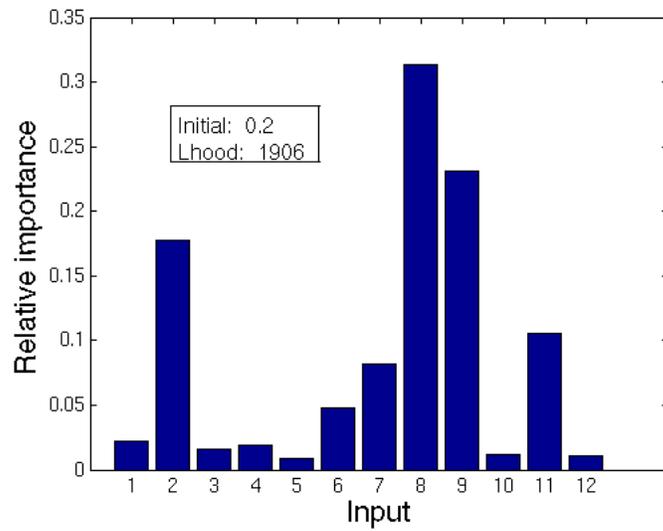
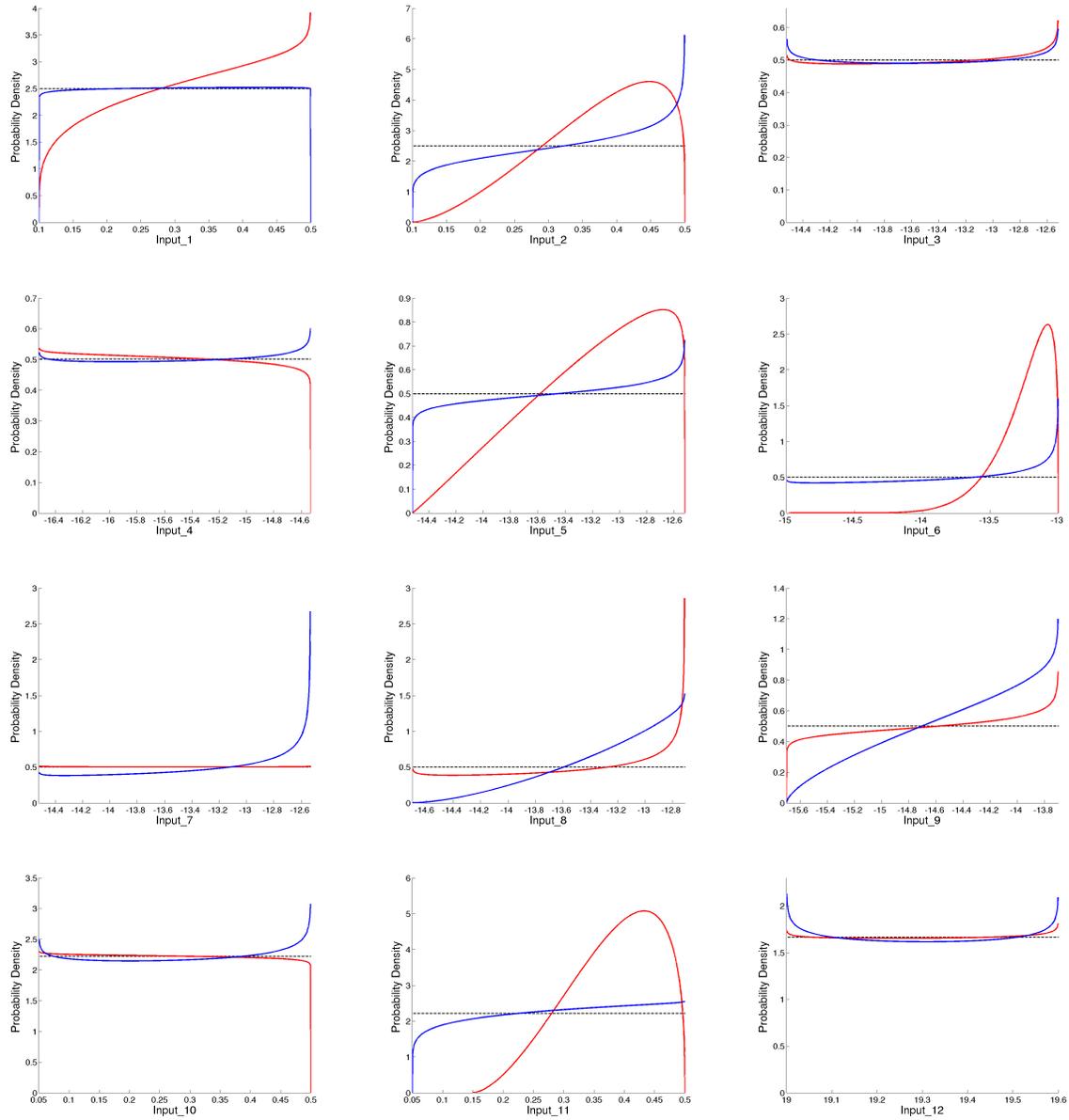


Figure 11. Correlation parameters (sensitivity) for predicting response 4



**Figure 12.** Comparison of calibration results based on response 4 versus response 1, for Q1 data. The red curves represent response 1 (nominal), while the blue represent response 4.

## IV.C. Calibration based on multiple response values

### IV.C.1. Multiple time responses

Note that all of the preceding analyses were based on one response value only, whereas both the simulation model and the experimental observations consist of measurements of the response at 4 distinct time instances. Ideally, we would like for our calibration to be based on all response outputs, instead of just 1. This section will discuss a method for updating the inputs based on all 4 response measures, simultaneously.

First, we must consider the statistical properties of the 4 responses. If they are correlated (as we would expect, since they are measurements of the same quantity at different times), then we can not treat them as if each response is independent of the others. For example, if the 4 responses are perfectly correlated, then three out of four of the responses do not contain any additional information, and it is incorrect to treat them as if they do. The information about the response correlations exists in the Bayesian calibration process through the joint likelihood function. For example, when updating based on multiple correlated responses, the likelihood function is based on the following distribution (whereas it was previously based on the distribution of Eq. (3):

$$\mathbf{Y}_{obs} | \boldsymbol{\theta} \sim N_p(\mathbf{M}(\boldsymbol{\theta}), \boldsymbol{\Sigma}), \quad (22)$$

where the responses,  $\mathbf{Y}$  and  $\mathbf{M}(\boldsymbol{\theta})$ , are now vectors of dimension  $p$ , and  $\boldsymbol{\Sigma}$  is the covariance matrix of the observations. Thus, the information about the correlations is captured by  $\boldsymbol{\Sigma}$ . Just like we assumed that the observations followed a normal distribution in the original model, Eq. (22) states that we are now assuming that the responses together follow a multivariate normal distribution.

In many cases we will not have enough experimental observations to directly estimate  $\boldsymbol{\Sigma}$ . In such situations, one possibility is to estimate the correlation structure based on the simulator runs, and use the assumed values for the individual variances based on experimental error assumptions. Thus, we might construct  $\boldsymbol{\Sigma}$  as

$$\Sigma_{i,j} = \rho_{i,j} \sigma_{exp}^{(i)} \sigma_{exp}^{(j)}, \quad (23)$$

where  $\rho_{i,j}$  is the correlation between responses  $i$  and  $j$ , as estimated from the simulator data, and  $\sigma_{exp}^{(i)}$  is the (assumed) standard deviation corresponding to the experimental observation of response  $i$ . The reason for not estimating  $\boldsymbol{\Sigma}$  exclusively from the simulator data is that we do not necessarily expect the variances of the simulator output to correspond to those of the experiments.

With the addition of the response surface approximation, the likelihood function for  $\boldsymbol{\theta}$  is now based on the distribution

$$\mathbf{Y}_{obs} | \boldsymbol{\theta} \sim N_p(\hat{\mathbf{M}}(\boldsymbol{\theta}), \boldsymbol{\Sigma} + \boldsymbol{\Sigma}_{GP}), \quad (24)$$

where  $\boldsymbol{\Sigma}_{GP}$  is a diagonal matrix with elements  $\sigma_{GP}^2(\boldsymbol{\theta})$  corresponding to the Gaussian process approximation for each response. Using Eq. (24), we could proceed with the Bayesian analysis. For the QASPR data, this would not be excessively expensive, since there are only 4 responses. However, we will present a method here which is applicable even when the time response is highly multivariate. Since the development of the response surface can be difficult when there are a large number of time intervals indexing the response<sup>b</sup>, we will illustrate how principal components analysis (PCA) can be used to reduce the dimensionality of the response.

To illustrate the method, consider the QASPR data, for the Q1 scenario. PCA effects a transformation into an uncorrelated space based on the eigenvectors of the covariance or correlation matrix (the correlation matrix is used to reduce the effect of variance on the transformation). The correlation matrix of the 4 time responses based on the 300 true model evaluations is estimated as

$$\mathbf{R} = \begin{bmatrix} 1.00 & 0.97 & 0.86 & 0.56 \\ 0.97 & 1.00 & 0.95 & 0.69 \\ 0.86 & 0.95 & 1.00 & 0.86 \\ 0.56 & 0.69 & 0.86 & 1.00 \end{bmatrix} \quad (25)$$

<sup>b</sup>In such a case there are two options, both of which are non-trivial: either develop a separate response surface for each time response, or include time as an extra input variable.

Using PCA, the transformation is given by the eigenvectors,  $\mathbf{A}$ ; the corresponding eigenvalues,  $\boldsymbol{\lambda}$ , of  $\mathbf{R}$  represent the amount of variance explained by each principal component:

$$\mathbf{A} = \begin{bmatrix} 0.49 & 0.54 & -0.55 & 0.40 \\ 0.53 & 0.29 & 0.18 & -0.78 \\ 0.53 & -0.13 & 0.69 & 0.47 \\ 0.45 & -0.78 & -0.43 & -0.09 \end{bmatrix} \quad (26)$$

$$\boldsymbol{\lambda} = \begin{bmatrix} 3.458 & 0.504 & 0.0336 & 0.0042 \end{bmatrix}$$

First, the eigenvalues tell us that using the first two principal components only, we can explain 99.1% of the variance of the original variables. This allows us to reduce the number of variables from 4 to 2. Also, the columns of  $\mathbf{A}$  represent the transformations corresponding to each component. We see that the first component is effectively an average of the 4 original variables. This is typical when the variables are highly correlated. The second component is made up mostly of the 1st and 4th original variables, which makes sense because they each contain slightly different information.

Thus, using only the first two principal components, our transformation matrix is given by

$$\mathbf{A}_{(2)} = \begin{bmatrix} 0.49 & 0.54 \\ 0.53 & 0.29 \\ 0.53 & -0.13 \\ 0.45 & -0.78 \end{bmatrix} \quad (27)$$

Since our analysis is based on the correlation matrix as opposed to the covariance matrix, we must first standardize the original variables by dividing by their standard deviations. Secondly, we note that less information is lost in the reverse transformation when the variables are close to the origin. Because of this, we want to shift the original variables to have a mean of  $\mathbf{0}$ . For the calibration exercise, we will actually shift by the observed response values because this is what we are calibrating to. Thus, we can represent the forward and reverse transformations as

$$\mathbf{z} = \mathbf{A}_{(2)}^T \mathbf{y}' \quad (28)$$

and

$$\mathbf{y}' = \mathbf{A}_{(2)} \mathbf{z}, \quad (29)$$

where  $\mathbf{y}' = \mathbf{D}_s^{-1}(\mathbf{y} - \mathbf{y}_{obs})$ , (lower case  $\mathbf{y}_{obs}$  is used here to represent the fixed realization of the experiment), and  $\mathbf{D}_s$  is the diagonal matrix of standard deviations (corresponding to  $\boldsymbol{\Sigma}$  above, so that  $\mathbf{D}_s$  contains the standard deviations of the experimental observations).

Based on Eq. (22), it can be shown that we have the following distributions for the transformed experimental observations:

$$\mathbf{Y}'_{obs} | \boldsymbol{\theta} \sim N_4 \left( \mathbf{D}_s^{-1} \left( \hat{\mathbf{M}}(\boldsymbol{\theta}) - \mathbf{y}_{obs} \right), \mathbf{R} \right) \quad (30)$$

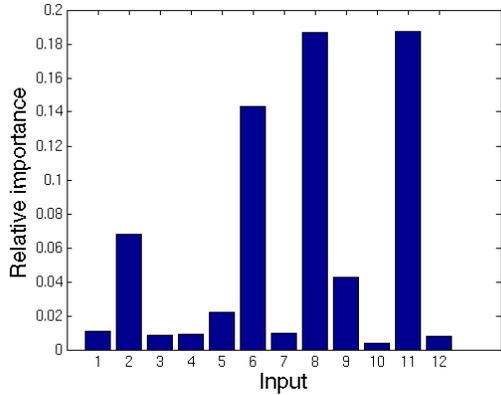
and

$$\mathbf{Z}_{obs} | \boldsymbol{\theta} \sim N_2 \left( \mathbf{A}_{(2)}^T \mathbf{D}_s^{-1} \left( \hat{\mathbf{M}}(\boldsymbol{\theta}) - \mathbf{y}_{obs} \right), \mathbf{A}_{(2)}^T \mathbf{R} \mathbf{A}_{(2)} \right). \quad (31)$$

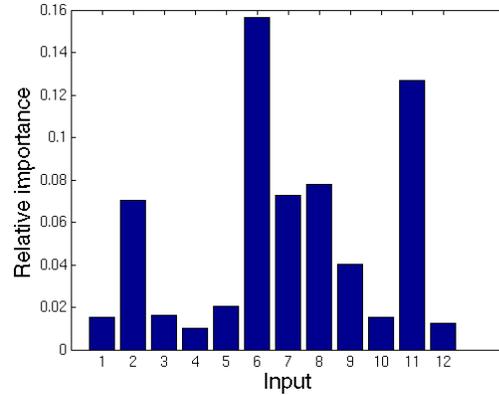
In addition to reducing the response dimensionality from 4 to 2, the principal components analysis also yields variables ( $\mathbf{Z}_{obs}$ ) which are uncorrelated. Since we are assuming joint normality, as per Eq. (22), this also implies that the components are independent, which simplifies the likelihood calculations somewhat.

Now that we have established our variable transformation to 2 principal components,  $\mathbf{z}$ , we build two response surface approximations based on the transformed simulator data,  $z_1$  and  $z_2$ , given by Eq. (28). In addition, we also transform our experimental observations, but in light of the fact that we are shifting by the value of the experimental observations, we conveniently have  $\mathbf{z}_{obs} = \mathbf{0}$ .

The resulting maximum likelihood estimates for the correlation parameters,  $\boldsymbol{\xi}$ , for the 2 Gaussian process models are illustrated below in Figure 13. We notice that the parameters based on the 1st component are very similar to the nominal model, with input 8 having more importance. However, the correlation parameters based on the 2nd component indicate that almost all of the inputs are relevant. Thus, the second component may be capturing a somewhat more complicated behavior than the first. Also notice that the parameters for the 2nd principal component are similar to those based on response 4 (see Figure 11).



(a) Correlation parameters for GP model of  $z_1$



(b) Correlation parameters for GP model of  $z_2$

**Figure 13. Correlation parameters for Gaussian process models based on 1st 2 principal components of all 4 responses**

The results of the calibration process are given in Figure 14. Compared to the calibration results based only on 1 response, it is clear that it takes a much more precise combination of inputs to get good agreement for all 4 response values simultaneously. Unlike before, almost all of the inputs now have specific posterior ranges, as opposed to having support along their entire bounds.

Figure 15 illustrates the agreement between the model predictions and the experimental observations, after updating the input distributions. The posterior distributions of three of the response values are plotted (response 3 is omitted for clarity), along with the likelihood function corresponding to each response. We can see that the updated predictions match very well with the observations for all 4 responses. Recall that MCMC simulation was conducted on the transformed variables (principal components). Thus, to plot the posterior distributions of the original variables, the reverse transformation given by Eq. (29) is employed.

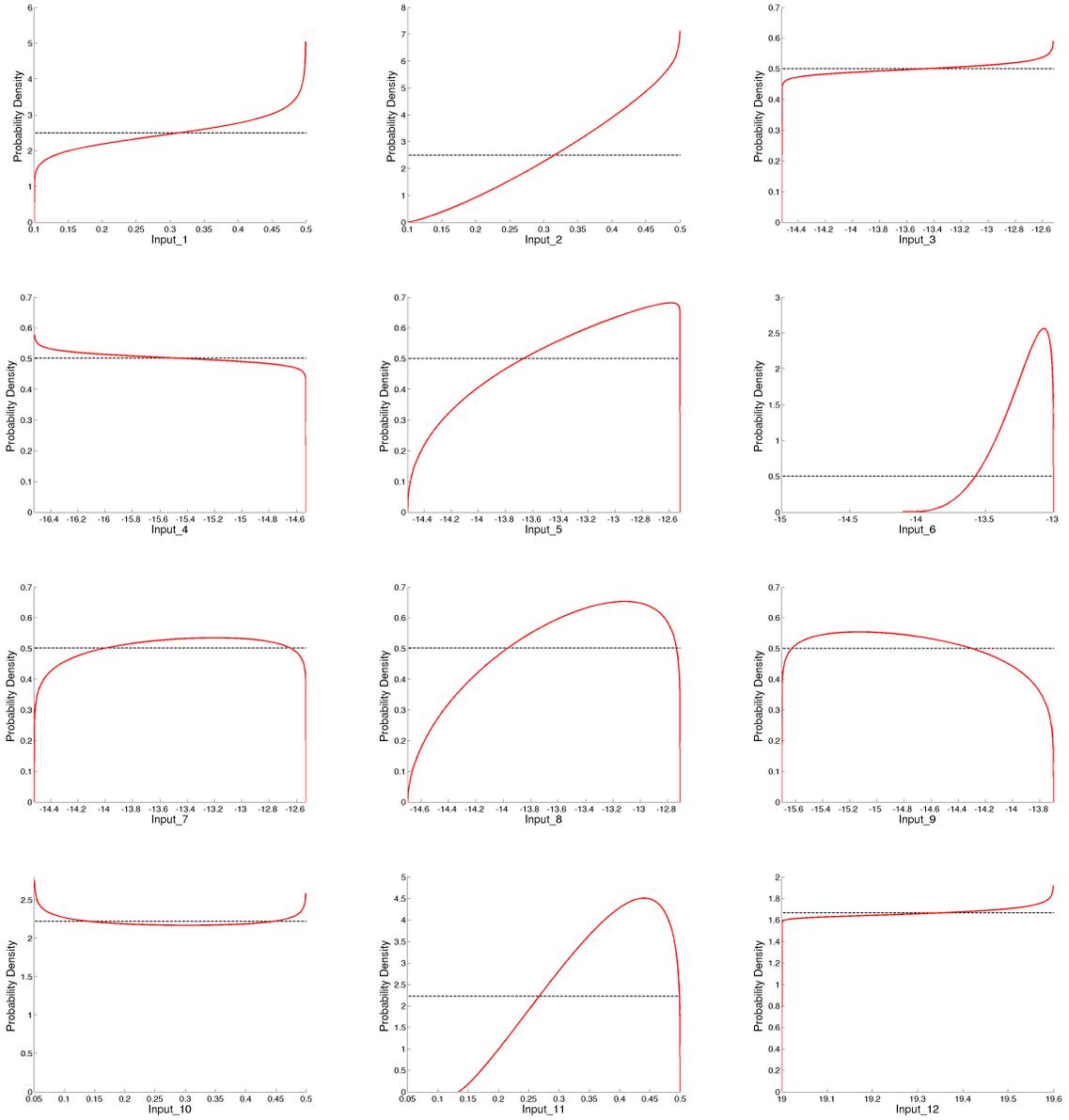


Figure 14. Calibrated input distributions based on all 4 responses, using the first 2 principal components

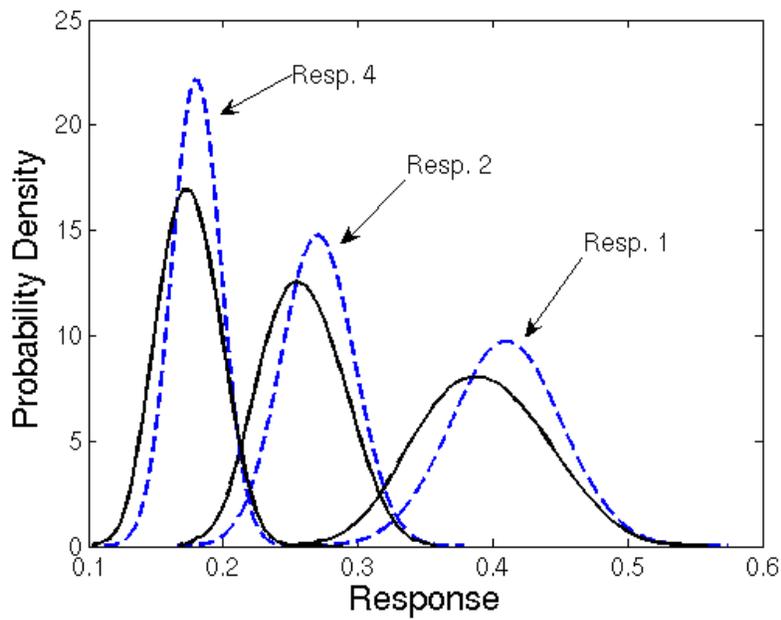


Figure 15. Updated output distributions for responses 1, 2, and 4 (response 3 omitted for clarity) resulting from the calibration based on 2 principal components of all 4 response measures. Updated model outputs given by solid black lines; experimental observations given by dashed blue lines.

#### IV.C.2. Data from multiple scenarios

We may also want to calibrate the model based on data from 2 or more scenarios simultaneously. The analysis will be similar to that for multiple response values. However, there is a difference in that we do not need to worry about correlations between the experiments for different scenarios. Since the experiments corresponding to each scenario are completely separate from each other, we can safely assume that the experimental measurements for Q1, Q2, and Q3 are independent of each other.

Here we perform the calibration based on response 1 only, using the Q1 and Q2 data. As mentioned before, this requires separate Gaussian process surrogate models for each the simulator response of each scenario. The results are shown below in Figure 16.

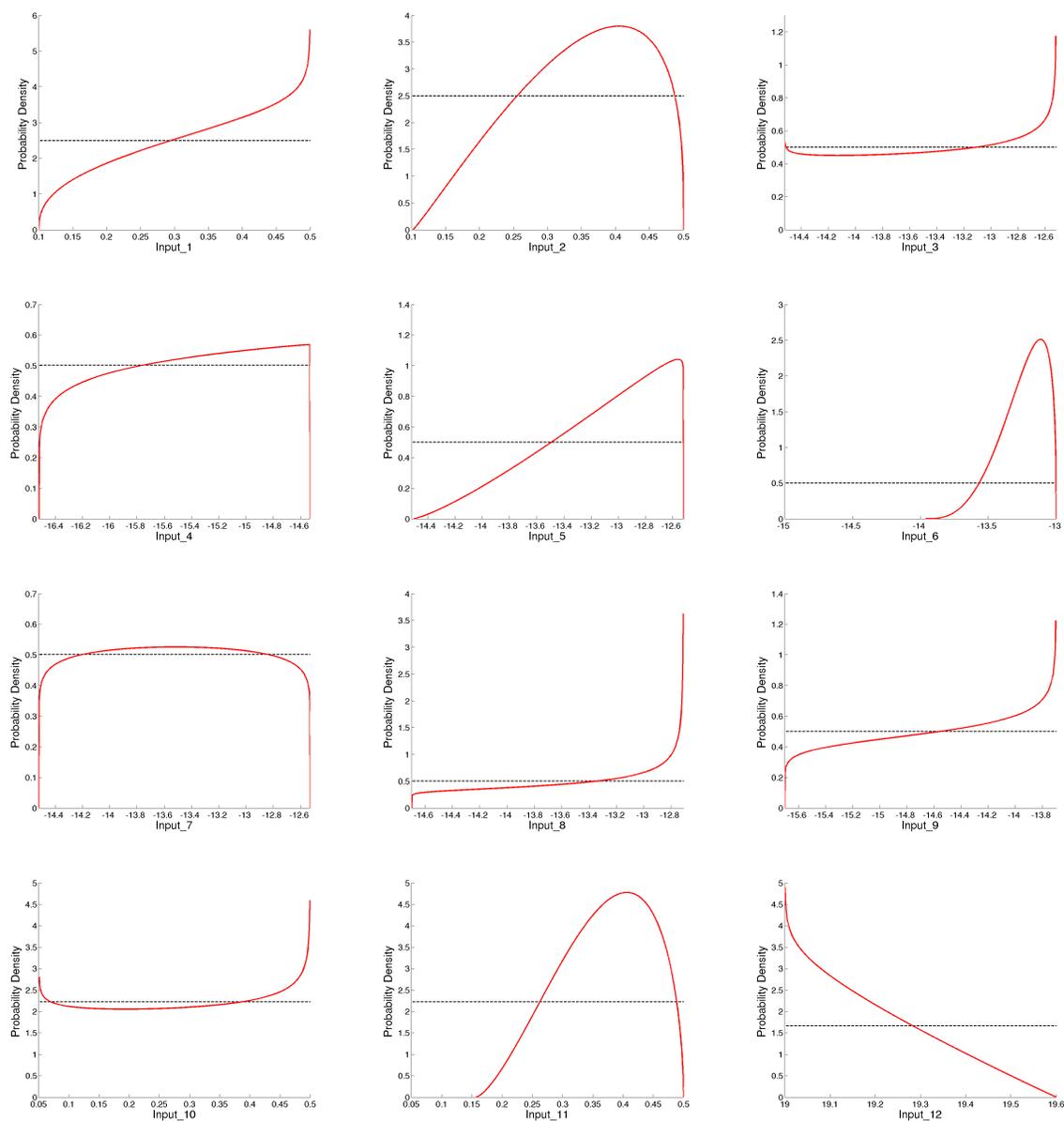


Figure 16. Calibrated input distributions based on both Q1 and Q2 data, response 1 only.

In addition, Figure 17 shows the updated response distributions for both the Q1 and Q2 scenarios. It appears that these calibration results can give predictions which are consistent with the data for both the Q1 and Q2 scenarios, simultaneously.

Recall from Section IV.B.2 that the calibration results based on Q1 and Q2 individually were not incon-

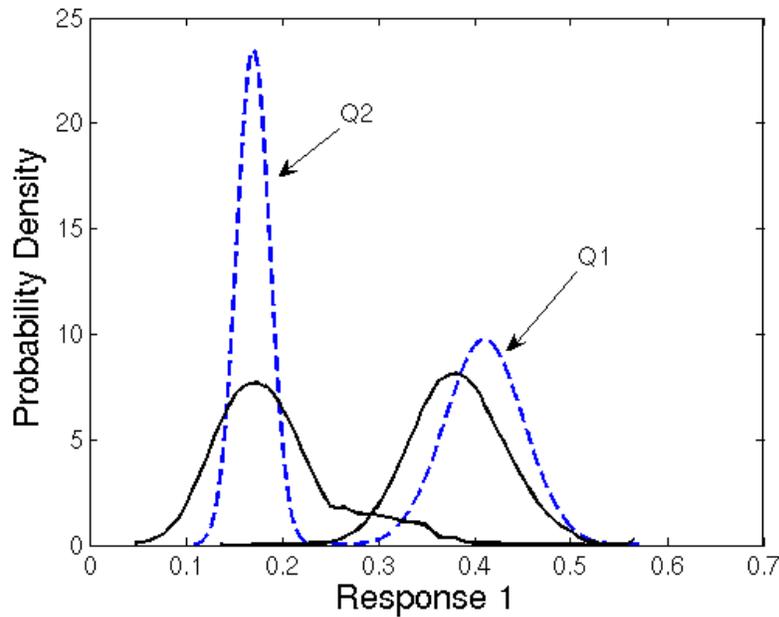


Figure 17. Updated output distributions for Q1 and Q2 scenarios, based on one calibration using both data sets. Updated model outputs given by solid black lines; experimental observations given by dashed blue lines.

sistent. However, from Figure 9, there appears to be a possible inconsistency between Q1 and Q3, for some of the inputs. This inconsistency is expected, given how different the discrepancy between the simulator and experiments is for Q1 and Q3 (see Figure 8). The result is that the calibration based on both the Q1 and Q3 data does not give satisfactory results (i.e., the posterior distributions of the responses are not consistent with the data). This is a strong indication that more attention should be given to the physics simulation to try and understand why the bias is so different for the various scenarios, because if the simulation is to be used for extrapolation to an untested scenario, such changes in the bias could render the predictions unusable.

## V. Further analysis of results

This section will briefly consider some cross-validation analyses of the calibration results. The purpose is to attempt to develop confidence in our interpretations and usages of the resulting input distributions.

### V.A. Interpretation of posterior distributions

First, we note that the 1-dimensional posterior distributions for the inputs shown in Section IV do not capture all of the information about the joint distribution of the inputs. This is because correlations can exist among the inputs, and this problem is further complicated by the fact that the joint posterior is not multivariate normal. Since the marginal posteriors are non-normal, it is even insufficient to specify the full joint distribution using the marginal distributions and correlation coefficients. Thus, the only reliable samples from the joint posterior are those from the MCMC chain generated by the calibration process.

This effect is illustrated in Figure 18, by fitting independent beta distributions to each input variable's posterior (beta distributions are used because they are extremely flexible at fitting a variety of distributional forms, and because they have support over a finite range, like the posteriors). 25,000 random values were then generated from these independent distributions and propagated through the GP model. Although the location of the response does not change too much, the uncertainty has increased significantly. This is because the use of independent beta distributions has not captured correctly the joint posterior distribution of the inputs and its associated correlation structure.

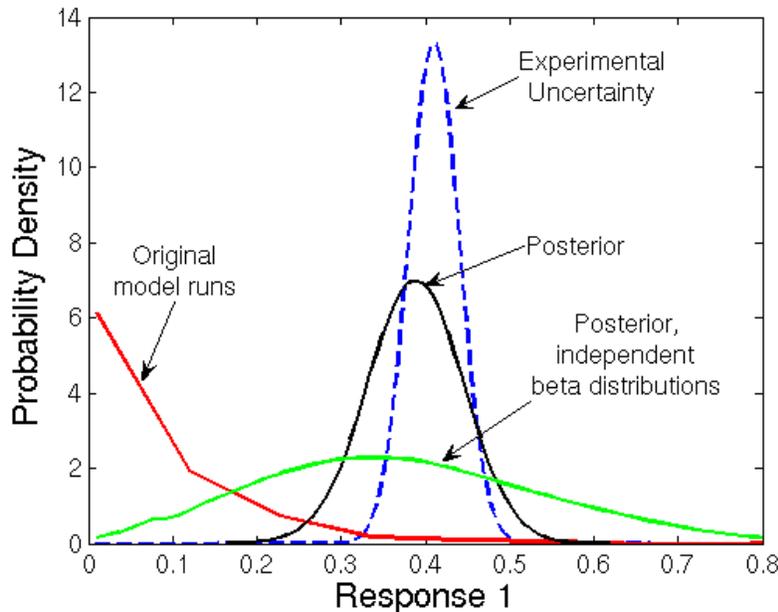


Figure 18. Updated response for nominal case, compared with re-propagation of inputs through GP model assuming independent beta distributions

Two of the largest correlations between the updated inputs are between inputs 2 and 6, and between 5 and 11. Contour plots are shown for the two pairwise joint densities using two-dimensional histograms in Figures 19 and 20. It is clear from the two-dimensional plots that the variables can not be treated as independent. Using Spearman's  $\rho$ , a non-parametric correlation measure, the correlations are  $-0.26$  and  $-0.29$ , which seem fairly mild. However, as more variables are considered together, the correlation structure can only become more complicated, further reinforcing the fact that it is dangerous to assume the updated probability distributions to be independent of each other.

Note that we can use the contour plots to express pairwise joint confidence regions for two parameters at a time. Each contour line represents a confidence region at a particular significance level. Although there are several different ways of expressing confidence regions, the use of contour lines from a joint density estimate corresponds to what is known in the Bayesian literature as “Highest Density Regions” (HDR's).<sup>11</sup> A HDR

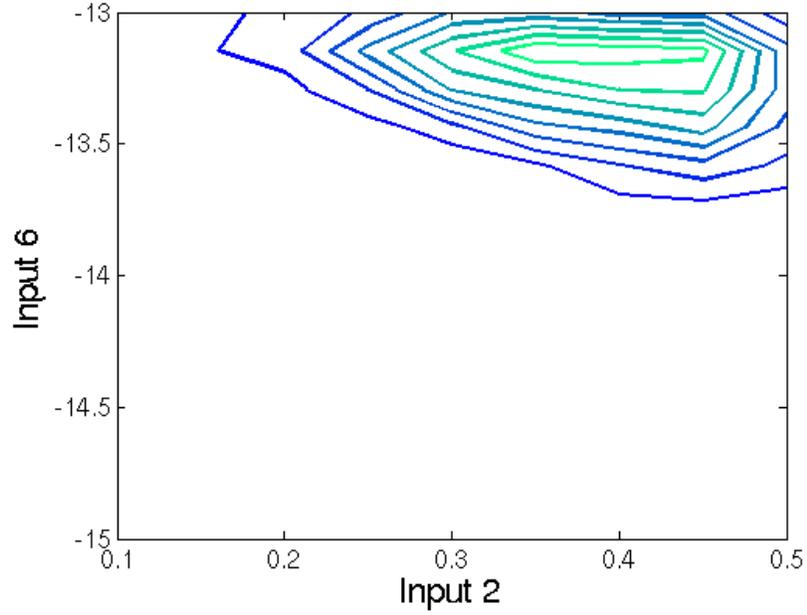


Figure 19. Estimated joint density of inputs 2 and 6

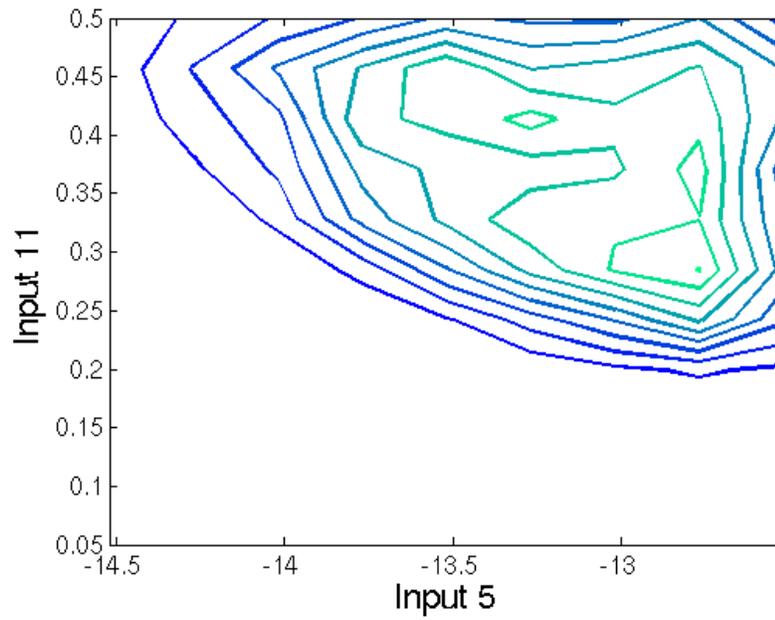


Figure 20. Estimated joint density of inputs 5 and 11

is a region such that the probability density at all points inside the region is greater than that for all points outside the region.

### V.B. Prediction across models

Here we will see how well we can use the calibration results from one model (i.e., a particular scenario or time response) to make predictions using a different model. (However, given Figures 8 and 10, we do not expect the results to predict well across time and scenario.) We will continue to use the “nominal” (Q1, response 1) model as our base case, and we will test whether the calibrations resulting from the other models can predict the experimental response for the nominal model. This is a test of both the calibration process and physical model itself, since these cross-predictions will only work if the physical model is in some sense correct.

First, we consider what happens if we take the updated input distributions obtained using one Gaussian process model, and propagate them through a different Gaussian process model of the same simulator data. We will do this by revisiting the analysis of Section IV.B.1. We will take the results obtained using the alternate GP model and plug them into the nominal GP model. Ideally, the resulting distributions should be the same, since the two response surfaces are modeling the same data. The results are shown in Figure 21. We see that the resulting distributions are only slightly different. This is an indication that the calibration process may be only slightly sensitive to the particular formulation of the response surface approximation model.

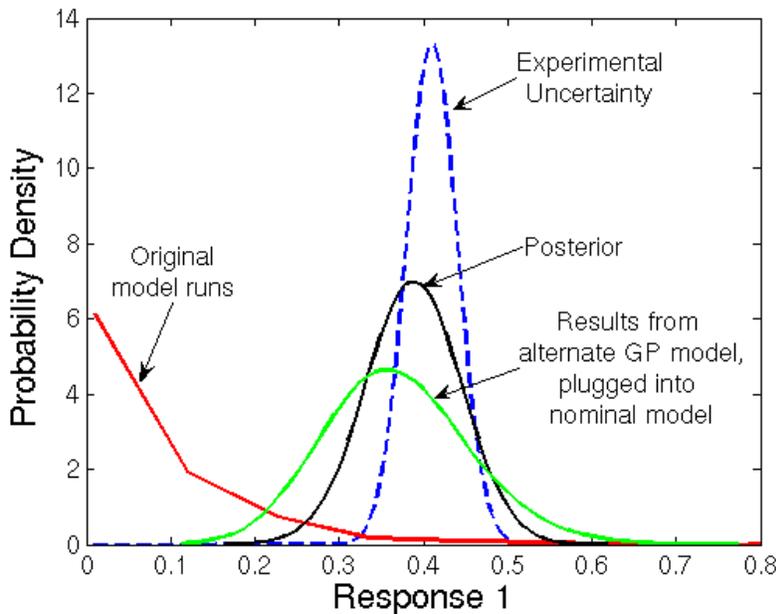
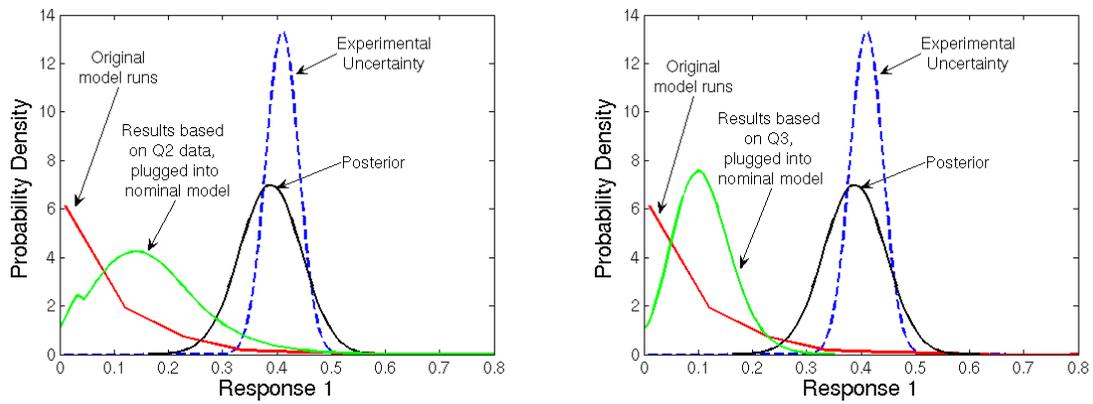


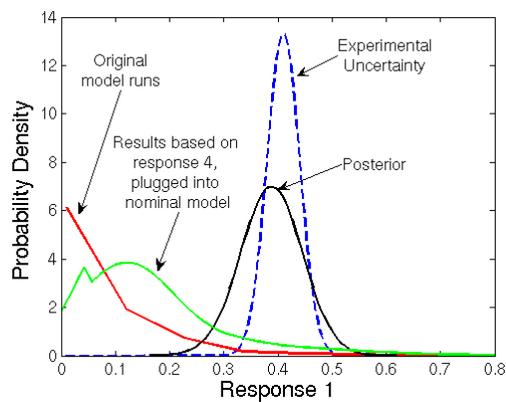
Figure 21. Resulting response distribution obtained by plugging the calibration results obtained from the alternate GP model into the nominal GP model

Next, we consider the performance when the results based on one particular scenario or response measure are used to predict at another. We illustrate this effect by using the results of the Q2 and Q3 scenarios (response 1), as well as those of Q1, response 4, to predict the output for the “nominal” case (Q1, response 1). This is simply done by plugging the MCMC output samples for the alternative cases into the *nominal* GP model. The results are shown below in Figure 22. Unfortunately, the calibrated input distributions do not give accurate performance for predicting scenarios or response measures other than those with which they were calibrated. This could be an indication that the physics simulation is not modeling the experimental data correctly.



(a) Comparison with results from Q2 condition

(b) Comparison with results from Q3 condition



(c) Comparison of results for response 4, Q1

**Figure 22.** Performance of updated input distributions for predicting the response at different scenarios and response measures.

## VI. Conclusions

A Bayesian calibration methodology is presented and applied to data from the Sandia QASPR project. The methodology is used to update belief about the “best” values of the uncertain input parameters to a simulation, using experimental measurements of its response value. Since only a finite number of runs of the simulator are available, Gaussian process response surface approximations are constructed as fast emulators for the true simulation. It is shown how the Bayesian framework allows us to account for uncertainty in the experimental measurements, the response surface approximation, and the results of the calibration.

One interesting conclusion is that the interpretation/presentation of the results of the calibration analysis is not trivial. As discussed in Section V.A, ignoring the full correlation structure of the updated distributions will result in a large overestimate of the uncertainty. Given the importance of the joint structure, marginal posterior distributions and confidence intervals should be used with caution. Additionally, it is difficult to express the resulting distribution information when there are multiple inputs being updated. Joint distributions can only be visualized in two dimensions, and the expression of confidence regions in more than two dimensions becomes very difficult. Thus, there exists the potential for future work to address the task of constructing “summary statistics” based on random samples of a high-dimensional random variable.

Also, only mild success is achieved in attempting to make use of multiple scenarios and/or responses simultaneously for calibration. However, this is an important step, since real world applications will want to make use of all available data when calibrating a simulation. The difficulty seen here is most likely due to the way the discrepancy between the simulation output and the experiments appears to depend on the particular scenario or response function (refer to Figures 8 and 10). One possible way to deal with this problem is to make use of the full probabilistic model of Eq. (1). This model, although more complicated, allows the analyst to model the discrepancy between the predictions and observations as a random process,  $\delta(\mathbf{x})$ . The implementation of this full model is certainly an area for future study.

## Acknowledgments

This work was supported by the Department of Energy’s Advanced Simulation and Computing (ASC) Verification and Validation Program and by the Nanoscience, Engineering, and Computation Institute (NECIS) at Sandia National Laboratories.

The authors would also like to acknowledge the helpful insight gained from valuable discussions with Youssef Marzouk on Bayesian calibration and Ramesh Rebba on the maximum likelihood estimation of Gaussian process models.

## Appendix

### MCMC implementation

For all calculations reported here, the posterior distributions were estimated using Markov Chain Monte Carlo (MCMC) sampling. The particular algorithm implemented here is known as the Metropolis algorithm.<sup>6</sup> The Metropolis algorithm can be used to generate random samples from any probability density which is known up to a proportionality constant (it turns out that the intractable part of expressing the Bayesian posterior analytically is the evaluation of a complicated integral, which is just a constant).

A sequential form of the algorithm is implemented here, so that candidate moves on each component (variable) are made one at a time. These candidate moves are generated from what is known as a proposal density, and random walk proposals are used here. For each component, the variance of the random walk is adjusted so that the observed acceptance ratio is approximately 0.3 for each variable.

Convergence of the sampling chain to the target distribution is often an issue when using MCMC methods. Convergence is usually assessed by looking at trace plots of the samples and checking for stationarity. It is often the case that some number of “burn-in” samples must be discarded, so that the chain is allowed time to reach its stationary distribution. However, convergence is not found to be an issue for the calculations done here, given that an appropriate starting value is used (a good choice is a value close to the “best” of the true simulator runs). For each case, 25,000 random samples were generated for the calibration inputs, and no burn-in samples were needed.

It is of note that when using a response surface approximation, bounded prior distributions help to keep the calibration inputs from straying into regions of large response surface uncertainty. This is especially

important if the response surface uncertainty is not being accounted for, because the MCMC chain may attempt to wander outside the range of the training points, in which case the predictions given by the response surface may not be trustworthy.

## References

- <sup>1</sup>Campbell, K., “A brief survey of statistical model calibration ideas,” *International Conference on Sensitivity Analysis of Model Output*, Santa Fe, NM, 2004.
- <sup>2</sup>Trucano, T., Swiler, L., Igusa, T., Oberkampf, W., and Pilch, M., “Calibration, validation, and sensitivity analysis: what’s what,” *Reliability Engineering and System Safety*, Vol. 91, 2006.
- <sup>3</sup>Kennedy, M. and O’Hagan, A., “Bayesian calibration of computer models,” *J. R. Statist. Soc. B*, Vol. 63, No. 3, 2001, pp. 425–464.
- <sup>4</sup>Beven, K. and Binley, A., “The future of distributed models: model calibration and uncertainty prediction,” *Hydrological Processes*, Vol. 6, 1992, pp. 279–298.
- <sup>5</sup>Vecchia, A. and Cooley, R., “Simultaneous confidence and prediction intervals for nonlinear regression models, with application to a groundwater flow model,” *Water Resources Research*, Vol. 23, No. 7, 1987, pp. 1237–1250.
- <sup>6</sup>Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E., “Equation of state calculations by fast computing machines,” *J. Chem. Phys.*, Vol. 21, 1953, pp. 1087–1092.
- <sup>7</sup>Chib, S. and Greenberg, E., “Understanding the Metropolis-Hastings algorithm,” *American Statistician*, Vol. 49, 1995, pp. 327–335.
- <sup>8</sup>Ripley, B., *Spatial Statistics*, John Wiley, New York, 1981.
- <sup>9</sup>Martin, J. and Simpson, T., “Use of kriging models to approximate deterministic computer models,” *AIAA Journal*, Vol. 43, No. 4, 2005, pp. 853–863.
- <sup>10</sup>Rasmussen, C., *Evaluation of Gaussian processes and other methods for non-linear regression*, Ph.D. thesis, University of Toronto, 1996.
- <sup>11</sup>Lee, P., *Bayesian Statistics, an Introduction*, Oxford University Press Inc., New York, 2004.