Supplementary materials for this article are available online. Please go to http://www.tandfonline.com/r/TECH

Prediction and Computer Model Calibration Using Outputs From Multifidelity Simulators

Joslin GOH and Derek BINGHAM

Department of Statistics and Actuarial Science
Simon Fraser University
Burnaby, BC V5A 1S6, Canada
(*joslin_goh@sfu.ca; bingham@stat.sfu.ca*)

James Paul HOLLOWAY, Michael J. GROSSKOPF, Carolyn C. KURANZ, and Erica RUTTER

Center for Radiative Shock Hydrodynamics
University of Michigan
Ann Arbor, MI 48109
(*hagar@umich.edu; mikegros@umich.edu; ckuranz@umich.edu; ruttere@umich.edu*)

Computer simulators are widely used to describe and explore physical processes. In some cases, several simulators are available, each with a different degree of fidelity, for this task. In this work, we combine field observations and model runs from deterministic multifidelity computer simulators to build a predictive model for the real process. The resulting model can be used to perform sensitivity analysis for the system, solve inverse problems, and make predictions. Our approach is Bayesian and is illustrated through a simple example, as well as a real application in predictive science at the Center for Radiative Shock Hydrodynamics at the University of Michigan. The Matlab code that is used for the analyses is available from the online supplementary materials.

KEY WORDS: Computer experiment; Gaussian process; Markov chain Monte Carlo.

1. INTRODUCTION

Deterministic computer models are used to simulate a wide variety of physical processes (Sacks et al. 1989; Santner, Williams, and Notz 2003; Welch et al. 1992). Oftentimes, a single run of the code requires considerable computational effort, making it infeasible to continually exercise the simulator. Instead, experimenters attempt to explore the computer model response (and to some extent the physical process) using a limited number of computer model runs.

In some applications, several simulators of the physical process are available, each with different levels of fidelity. The varying levels of fidelity can occur, for example, because of the presence of reduced order physics in lower fidelity models, different levels of accuracy specified for numerical solvers, or solutions obtained on finer grids. In these cases, a higher fidelity model is assumed to better represent the physical process than a lower fidelity model, but also takes more computer time to produce an output than a lower fidelity model. Combining relatively cheap lower fidelity model runs with more costly high-fidelity runs to emulate the high-fidelity model has, thus, been a significant problem of interest (Kennedy and O'Hagan 2000; Qian et al. 2006; Qian and Wu 2008; Cumming and Goldstein 2009).

The most common approach to combining the outputs of multifidelity simulators was proposed by Kennedy and O'Hagan (2000). Their work writes a high-fidelity model as a linear combination of the next lowest fidelity model and a discrepancy term. With an alteration of the linear combination used to model the high-fidelity simulator, Qian et al. (2006, Qian and Wu 2008) also used the Bayesian hierarchical Gaussian process to model the response surfaces. Cumming and Goldstein (2009) further generalized the model and used a Bayes linear approach.

Another important application of computer simulators is that of *calibration* (e.g., Kennedy and O'Hagan 2001; Higdon et al. 2004) where the aim is to combine simulator outputs with physical observations to build a predictive model and also estimate unknown parameters that govern the behavior of the mathematical model. The latter endeavor amounts to solving a sort of inverse problem, while the former activity is a type of regression problem. In this setting, it is common to write the physical observations as a sum of the simulator output, a systematic discrepancy, and observational error. This approach has been adapted to consider a variety of output data structures (e.g., Higdon et al. 2008; Paulo, Garca-Donato, and Palomo 2012). These approaches to model calibration have used only a single computer model and have not considered the use of multifidelity simulators.

Motivated by applications at the Center for Radiative Shock Hydrodynamics (CRASH) at the University of Michigan, the aim of this work is to develop new methodology to combine outputs from simulators with different levels of fidelity and field observations to make predictions of the physical system with associated measurements of uncertainty. The CRASH simulators also require estimation of optimal values for several input parameters (i.e., *calibration parameters*), and the simulators have different calibration parameters depending on the level of fidelity, thereby complicating the calibration problem. In the spirit similar to Kennedy and O'Hagan (2000, 2001) and Higdon et al. (2004), we propose a predictive model that incorporates computer model outputs and field data, while attempting to estimate the calibration parameters. The approach calibrates each computer model to the next highest level of fidelity model, and the simulator of the highest fidelity is then calibrated to

DOI: 10.1080/00401706.2013.838910

the field measurements. All response surfaces are modeled using Gaussian process (GP) models, and the various sources of information that inform predictions of the physical system are combined with a Bayesian hierarchical model.

The article is organized as follows: In Section 2, we introduce the proposed methodology and the GP models, along with the relevant prior distributions. The framework for prediction will be discussed at the end of the section. A simple example from the literature and an application from CRASH are used to demonstrate the proposed approach in Section 3. Further discussion follows in Section 4, with some concluding remarks in Section 5.

2. A HIERARCHICAL MODEL FOR MULTIFIDELITY MODEL CALIBRATION

In this section, a Bayesian hierarchical model that calibrates multifidelity computer simulators is proposed. Throughout, higher fidelity codes are assumed to better represent the real-world process but require more computing resources to simulate the system. For ease of exposition and notation, we present the case where there are only two computer simulators—a high-fidelity and a low-fidelity model. It is conceptually easy to extend the proposed methodology to cases with more than two simulators, and this setting is discussed in Section 4.

2.1 The Hierarchical Model

Throughout this work, the simulators are assumed to be deterministic mathematical functions that map inputs to outputs. The computer codes have two types of inputs: (i) *design variables*, $\mathbf{x}$, that are adjustable or measurable in the field experiments; and (ii) *calibration parameters*, $\mathbf{t}$, whose values are thought to impact the physical system, but are unknown a priori. The latter inputs can only be adjusted within the simulator, but are not measurable in the field. We use $\mathbf{t}$ to denote inputs for calibration parameters used to run the code and let $\boldsymbol{\theta}$ stand for the true, or optimal, values of the calibration parameters in the field. In model calibration problems, the issue is to build a predictive model for the field process and also to estimate the unknown calibration parameters.

An important feature of the application that motivated the current work is that the calibration parameters for the computer models are not all the same. Some of the calibration parameters, $\mathbf{t}_f$, are shared among the simulators, whereas others are required inputs only to individual computer models. The vectors of calibration inputs exclusive to the high- and low-fidelity models are denoted as $\mathbf{t}_h$ and $\mathbf{t}_l$, respectively.

First consider the low-fidelity computer model with inputs $(\mathbf{x}, \mathbf{t}_f, \mathbf{t}_l)$ (i.e., the design variables and calibration parameters that are shared and unshared with the high-fidelity simulator), where $\mathbf{x} = (x_1, \ldots, x_p)$, $\mathbf{t}_f = (t_{f,1}, \ldots, t_{f,m_f})$, and $\mathbf{t}_l = (t_{l,1}, \ldots, t_{l,m_l})$. An output $Y_l(\cdot)$ from the low-fidelity simulator, $\eta_l(\cdot)$, is univariate and written as

$$Y_l(\mathbf{x}, \mathbf{t}_f, \mathbf{t}_l) = \eta_l(\mathbf{x}, \mathbf{t}_f, \mathbf{t}_l). \tag{1}$$

Similarly, the high-fidelity simulator, $\eta_h(\cdot)$, has inputs $(\mathbf{x}, \mathbf{t}_f, \mathbf{t}_h)$, where $\mathbf{t}_h = (t_{h,1}, \ldots, t_{h,m_h})$, and univariate output $Y_h(\cdot)$:

$$Y_h(\mathbf{x}, \mathbf{t}_f, \mathbf{t}_h) = \eta_h(\mathbf{x}, \mathbf{t}_f, \mathbf{t}_h).$$

Both simulators are used to describe the same process, but will not always give the same response. There are a few obvious reasons why this is the case. The lower fidelity model is inferior to the high-fidelity simulator since it may, for example, fail to capture some processes that the high-fidelity code can more accurately model. Furthermore, the two codes do not share all of the same inputs. The input vector $\mathbf{t}_h$ only appears in the high-fidelity model and thus, any impact that these variables have on the output cannot be captured by the low-fidelity model. Similarly, the inputs $\mathbf{t}_l$ appear only in the low-fidelity model. To address these issues, we take the approach of writing the high-fidelity simulator as a discrepancy-adjusted version of the low-fidelity model (e.g., Kennedy and O'Hagan 2000; Qian et al. 2006)

$$Y_h(\mathbf{x}, \mathbf{t}_f, \mathbf{t}_h) = \eta_l(\mathbf{x}, \mathbf{t}_f, \boldsymbol{\theta}_l) + \delta_2(\mathbf{x}, \mathbf{t}_f, \mathbf{t}_h). \tag{2}$$

Specifying the first term in Equation (2) as $\eta_l(\mathbf{x}, \mathbf{t}_f, \theta_l)$ amounts to partially calibrating (partially in the sense that the other calibration parameters must still be estimated) the first simulator to the second simulator. Recall, we will generally use t's to denote inputs to the code and θ's to denote the optimal values for the calibration parameters, respectively. Furthermore, the discrepancy, $\delta_2(\cdot)$, represents the systematic differences between the partially calibrated low-fidelity model and the high-fidelity code. Finally, notice that $\delta_2(\cdot)$ is a function of not only the design variables—as in Kennedy and O'Hagan (2001)—but also $(\mathbf{t}_f, \mathbf{t}_h)$. The calibration parameters are included in this discrepancy term because they can be modified in the high-fidelity code. Therefore, this discrepancy term captures the systematic differences in the outputs from the two computer models over values of the design variables and the changes in the calibration inputs $\mathbf{t}_f$ and $\mathbf{t}_h$.

In addition to simulator output, there are also field observations that are used to inform predictions. Since the higher fidelity simulator is assumed to better represent the physical process than the low-fidelity simulator, it is natural to model the field observations with the simulator of highest fidelity. Similar to Kennedy and O'Hagan (2001), a discrepancy function, $\delta_f(\cdot)$, is used to capture the systematic inadequacy of the high-fidelity simulator. The field observations are noisy versions of the mean process, and thus independent and identically distributed (iid) observational errors are included in our specification. For design variable setting, $\mathbf{x}$, the univariate field process is written as

$$Y_f(\mathbf{x}) = \eta_h(\mathbf{x}, \boldsymbol{\theta}_f, \boldsymbol{\theta}_h) + \delta_f(\mathbf{x}) + \epsilon, \tag{3}$$

where $\epsilon \sim N(0, 1/\lambda_y)$. Substituting Equation (2) into Equation (3) allows the field observations to be written as

$$Y_f(\mathbf{x}) = \eta_l(\mathbf{x}, \boldsymbol{\theta}_f, \boldsymbol{\theta}_l) + \delta_2(\mathbf{x}, \boldsymbol{\theta}_f, \boldsymbol{\theta}_h) + \delta_f(\mathbf{x}) + \epsilon. \tag{4}$$

So, the response surface for the field data is written as the sum of the calibrated low-fidelity simulator, the calibrated discrepancy between the two different simulators, the discrepancy between the high-fidelity model and the data, as well as observational error. From here on out, we describe the response surfaces for the low- and high-fidelity simulators and the field data using the framework described in Equation (1), Equation (2), and Equation (4), respectively.

It is possible at this point to envision applications with more than two simulators, each ranked from lowest to highest levels

of fidelity. For example, consider the case where there are three simulators of different fidelity, $\eta_1(\cdot)$, $\eta_2(\cdot)$, and $\eta_3(\cdot)$, where $\eta_1(\cdot)$ is of the lowest fidelity and $\eta_3(\cdot)$ is best at describing the physical process. Outputs from $\eta_1(\cdot)$ and $\eta_2(\cdot)$ are modeled as described earlier. Next, the relationship between $\eta_2(\cdot)$ and $\eta_3(\cdot)$ is similarly described by Equations (1) and (2). Through substitution, $\eta_3(\cdot)$ can then be written as $\eta_1(\cdot)$ and two discrepancy functions. Finally, through the same sort of substitution, the field observations can be written as the sum of lowest fidelity simulator, a sequence of discrepancy terms, and observational error. More on this in Section 4.

2.2 Gaussian Process Models

To make predictions of the physical system, the response surfaces for the low-fidelity simulator and discrepancies need to be estimated. We follow the common practice of using independent GPs to model the response surfaces (e.g., see Sacks et al. 1989; Kennedy and O'Hagan 2001). The reason for using these models, in general, boils down to the success of the GP as a nonparametric regression estimator and also the ability of the GP model to provide a basis for statistical inference for the outputs of deterministic computer codes. From a Bayesian viewpoint in this context, one can think of the GP as a prior distribution over the class of functions produced by the low-fidelity simulator and the discrepancies, respectively.

We begin by first considering the specification for the low-fidelity simulator. The outputs are treated as a realization of a random function of the form

$$Y_l(\mathbf{x}, \mathbf{t}_f, \mathbf{t}_l) = \sum_{i=1}^{p} f_i(\mathbf{x}, \mathbf{t}_f, \mathbf{t}_l)\beta_i + Z(\mathbf{x}, \mathbf{t}_f, \mathbf{t}_l),$$

where $f_1, \ldots, f_p$ are regression functions, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is the vector of unknown regression coefficients, and Z is a mean zero GP. We follow the convention of most simulator applications by specifying the mean function as a constant, μ, and model the response surface through the covariance structure. The covariance between observations at inputs $(\mathbf{x}, \mathbf{t}_f, \mathbf{t}_l)$ and $(\mathbf{x}', \mathbf{t}'_f, \mathbf{t}'_l)$ is specified as

$$\begin{aligned}&\operatorname{cov}[Z(\mathbf{x}, \mathbf{t}_f, \mathbf{t}_l), Z(\mathbf{x}', \mathbf{t}'_f, \mathbf{t}'_l)] \\ &\quad = \frac{1}{\lambda_{\eta_l}} \prod_{s=1}^{p} \rho_{\eta_l,s}^{4(x_s - x'_s)^2} \prod_{s=1}^{m_f} \rho_{\eta_l,p+s}^{4(t_{f,s} - t'_{f,s})^2} \prod_{s=1}^{m_l} \rho_{\eta_l,p+m_f+s}^{4(t_{l,s} - t'_{l,s})^2}, \quad (5)\end{aligned}$$

where λ_{η_l} is the marginal precision of the GP for the low-fidelity simulator. The $(p + m_f + m_l)$-vector $\boldsymbol{\rho}_{\eta_l}$ is the vector of correlation parameters that govern the dependence in each of the component directions of $\mathbf{x}$, $\mathbf{t}_f$, and $\mathbf{t}_l$ (e.g., Linkletter et al. 2006; Higdon et al. 2008).

The discrepancy, $\delta_2(\cdot)$, captures the systematic differences between the high- and low-fidelity simulators as a function of the inputs, $(\mathbf{x}, \mathbf{t}_h, \mathbf{t}_f)$, that are adjustable in the high-fidelity model. Continuing as earlier, $\delta_2(\cdot)$ is modeled as mean zero GP with covariance

$$\begin{aligned}&\operatorname{cov}[Z(\mathbf{x}, \mathbf{t}_f, \mathbf{t}_h), Z(\mathbf{x}', \mathbf{t}'_f, \mathbf{t}'_h)] \\ &\quad = \frac{1}{\lambda_2} \prod_{s=1}^{p} \rho_{2,s}^{4(x_s - x'_s)^2} \prod_{s=1}^{m_f} \rho_{2,p+s}^{4(t_{f,s} - t'_{f,s})^2} \prod_{s=1}^{m_h} \rho_{2,p+m_f+s}^{4(t_{h,s} - t'_{h,s})^2}, \quad (6)\end{aligned}$$

where λ_2 is the marginal precision of the discrepancy function, and the vector of correlation parameters for this discrepancy function is $\boldsymbol{\rho}_2$.

A zero-mean GP is chosen for the discrepancy, $\delta_f(\cdot)$, between the response from high-fidelity simulator and the mean physical process. Let λ_f denote the marginal precision of the discrepancy function, $\delta_f(\cdot)$, and $\boldsymbol{\rho}_f$ be the vector of correlation parameters for the p design variables. The covariance function for $\delta_f(\cdot)$ has the form

$$\operatorname{cov}[Z(\mathbf{x}), Z(\mathbf{x}')] = \frac{1}{\lambda_f} \prod_{s=1}^{p} \rho_{f,s}^{4(x_s - x'_s)^2}. \quad (7)$$

Denote the number of field observations and high-fidelity and low-fidelity simulation trials by n_f, n_h, and n_l, respectively. Furthermore, define the vector of all observations and simulation outputs as $\mathbf{Y} = (\mathbf{Y}_f^T, \mathbf{Y}_h^T, \mathbf{Y}_l^T)^T$, where $\mathbf{Y}_f$ is the $n_f \times 1$ vector of field measurements, $\mathbf{Y}_h$ is the $n_h \times 1$ vector of high-fidelity simulator responses, and $\mathbf{Y}_l$ is the $n_l \times 1$ vector of low-fidelity simulator outcomes. The field measurements have associated calibration parameters $(\mathbf{t}_f, \mathbf{t}_h, \mathbf{t}_l) = (\boldsymbol{\theta}_f, \boldsymbol{\theta}_h, \boldsymbol{\theta}_l)$. Similarly, the high-fidelity simulator has associated calibration inputs $(\mathbf{t}_f, \mathbf{t}_h, \boldsymbol{\theta}_l)$, where $(\mathbf{t}_f, \mathbf{t}_h)$ can vary for each run of high-fidelity code specified by the experimental design.

To simplify notation, denote $\boldsymbol{\theta} = (\boldsymbol{\theta}_f, \boldsymbol{\theta}_h, \boldsymbol{\theta}_l)$, $\boldsymbol{\lambda} = (\lambda_{\eta_l}, \lambda_2, \lambda_f)$, and $\boldsymbol{\rho} = (\boldsymbol{\rho}_{\eta_l}, \boldsymbol{\rho}_2, \boldsymbol{\rho}_f)$. The likelihood for $\mathbf{Y}$ is

$$L(\mathbf{Y}|\boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{\rho}) \propto |\Sigma_{\mathbf{Y}}|^{-\frac{1}{2}} \exp\{(\mathbf{Y} - \boldsymbol{\mu})^T \Sigma_{\mathbf{Y}}^{-1} (\mathbf{Y} - \boldsymbol{\mu})\},$$

where $\boldsymbol{\mu}$ is the constant mean vector and

$$\begin{aligned}\Sigma_{\mathbf{Y}} = \Sigma_{\eta_l} &+ \begin{pmatrix} \Sigma_2 & 0_{(n_f+n_h)\times n_l} \\ 0_{n_l \times (n_f+n_h)} & 0_{n_l \times n_l} \end{pmatrix} \\ &+ \begin{pmatrix} \Sigma_f + \Sigma_y & 0_{n_l \times (n_f+n_h)} \\ 0_{(n_f+n_h)\times n_l} & 0_{(n_f+n_h)\times(n_f+n_h)} \end{pmatrix}, \quad (8)\end{aligned}$$

where $0_{a\times b}$ is the $a \times b$ matrix of zeroes. The covariance matrix Σ_{η_l} for the low-fidelity simulator GP is obtained by applying Equation (5) to each pair of the $(n_f + n_h + n_l)$ observations and simulation outputs in $\mathbf{Y}$. Similarly, the covariance matrix Σ_2 is obtained by applying Equation (6) to each pair of the $n_f + n_h$ observations and model high-fidelity simulator responses, $(\mathbf{Y}_f^T, \mathbf{Y}_h^T)^T$. Equation (7) is applied only to each pair of field observations to construct the covariance matrix Σ_f. Finally, the covariance matrix for the measurement error, ϵ, is given by the $n_f \times n_f$ diagonal matrix $\Sigma_y = (1/\lambda_y) I_{n_f}$.

2.2.1 Prior Distributions and MCMC. The posterior distribution of calibration and statistical model parameters, $(\boldsymbol{\theta}, \mu, \boldsymbol{\lambda}, \boldsymbol{\rho})$, takes the form

$$\begin{aligned}\pi(\boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{\rho}|\mathbf{Y}) \propto L(\mathbf{Y}|\boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{\rho}) \times \pi(\boldsymbol{\theta}) \times \pi(\mu) \\ \times \pi(\boldsymbol{\lambda}) \times \pi(\boldsymbol{\rho}), \quad (9)\end{aligned}$$

where we abuse notation and denote the prior distributions for $\boldsymbol{\theta}$, $\boldsymbol{\lambda}$, and $\boldsymbol{\rho}$ as

$$\begin{aligned}\pi(\boldsymbol{\theta}) &= \prod_{i=1}^{m_f} \pi(\theta_{f,i}) \times \prod_{i=1}^{m_h} \pi(\theta_{h,i}) \times \prod_{i=1}^{m_l} \pi(\theta_{l,i}), \\ \pi(\boldsymbol{\lambda}) &= \pi(\lambda_{\eta_l}) \times \pi(\lambda_2) \times \pi(\lambda_f) \times \pi(\lambda_y),\end{aligned}$$

and

$$\pi(\boldsymbol{\rho}) = \prod_{i=1}^{p+m_f+m_l} \pi(\rho_{\eta_l,i}) \times \prod_{i=1}^{p+m_f+m_h} \pi(\rho_{2,i}) \times \prod_{i=1}^{p} \pi(\rho_{f,i}),$$

respectively.

Since the calibration parameters are scaled uniformly to the unit interval [0, 1] (e.g., Linkletter et al. 2006), the prior for each of the component in $\boldsymbol{\theta}$ is chosen to be an independent Normal distribution centered at 0.5 (center of the unit interval) with standard deviation 10. The choice of large standard deviation results in a weak information prior, allowing the data to move the calibration parameters. The prior for the precision of the marginal variance, λ_{η_l}, is chosen to encourage its values to be close to 1—the idea being that the low-fidelity model should capture much of the signal in the observations. We use a gamma distribution (denoted generally as Gam(a, b)) for the prior for λ_{η_l}

$$\pi(\lambda_{\eta_l}) \propto \lambda_{\eta_l}^{a_{\eta_l}} \exp\{-b_{\eta_l}\lambda_{\eta_l}\}.$$

When expert knowledge is unavailable, we have found that $a_{\eta_l} = b_{\eta_l} = 5$ works reasonably well as the choice centers the prior distribution at 1 with a reasonably large variance, thereby allowing for a fairly broad exploration of the posterior. Similarly, the priors chosen for the remaining precision parameters are also gamma distributions. Similarly, we use the default prior distributions, Gam$(1, 0.001)$, suggested by Higdon et al. (2004) for the hyperparameters of priors for the remaining precision parameters. This specification implies a relatively uninformative prior for these precision parameters and encourages the data to choose a suitable value.

The components in $\boldsymbol{\rho}$ are bounded within the unit interval. Hence, a natural choice of prior for any $\rho \in \boldsymbol{\rho}$ is an independent Beta distribution (denoted Beta(c, d)) of the following form:

$$\pi(\rho) \propto (\rho)^{c-1}(1-\rho)^{d-1}.$$

Conventionally, the Beta priors are flat, with a mean near 1 and a small variance (e.g., Williams et al. 2006). This is based on the prior belief that all the inputs are equally uncorrelated to the simulator and allows the data to decide upon the dependence of the simulator on the different inputs by moving the ρ's away from 1 in the posterior. In our experience, the default choice of Beta$(1, 0.001)$, suggested by Higdon et al. (2004) and Williams et al. (2006), encourages strong enough dependence in each of the parameters and works well in general.

The posterior distribution for each parameter is explored using MCMC. Specifically, single-site Metropolis updates (Metropolis et al. 1953) are used for the components of $\boldsymbol{\rho}$ and $\boldsymbol{\theta}$. Proposals are made for each of these parameters from a uniform distribution centered at the parameter's current value. The widths of the uniform distributions (one for each component parameter) are precomputed by running short MCMC runs and choosing a width that gives an acceptance rate of about 0.44 (Gelman et al. 2004). Although this adjustment does not guarantee an acceptance probability of 0.44, we have found this procedure to be helpful at choosing widths resulting in acceptance ratios between 0.25 and 0.75 and, more importantly, encourages the MCMC to converge. Good default choices for the widths for the updates can also be found using the method proposed by Graves (2005). For each of the precision parameters, we used Hastings updates (Hastings 1970), where the proposed value is drawn from a uniform distribution centered at the current parameter value, with a width that is proportional to the parameter's current value. We have found that a width that is 0.3 times the current parameter value (originally proposed by Higdon et al. 2008) works fairly well in general. It is feasible to use Metropolis updates for the precision parameters as well, but in the problems we have encountered, we have found the Hastings updates result in faster convergence.

2.3 Prediction

The main goal of this endeavor is prediction. Using the MCMC draws for the parameters, we estimate the posterior predictive distribution for a new field measurement at new inputs, $\mathbf{x}^{\text{new}}$. Given the posterior realizations from Equation (9), predictions of the field measurement, $Y_f(\mathbf{x}^{\text{new}})$, can be made at a new input setting $\mathbf{x}^{\text{new}}$.

The joint distribution between $\mathbf{Y}$ and $Y_f(\mathbf{x}^{\text{new}})$, conditional on the parameters $\boldsymbol{\theta}$, $\boldsymbol{\lambda}$, and $\boldsymbol{\rho}$, is

$$\begin{pmatrix} \mathbf{Y} \\ Y_f(\mathbf{x}^{\text{new}}) \end{pmatrix} \Bigg| (\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\rho}) \sim \text{MVN}(\mathbf{0}, \Sigma^{\text{new}}),$$

where the covariance matrix, Σ^{new}, is analogous to the covariance in Equation (8)—there is an extra row and column in Σ^{new} as a result of appending $Y_f(\mathbf{x}^{\text{new}})$ to $\mathbf{Y}$.

Through the usual properties of the multivariate normal distribution, the predictive distribution of $Y_f(\mathbf{x}^{\text{new}})$, conditional on $\mathbf{Y}$ and the parameters, is

$$Y_f(\mathbf{x}^{\text{new}}) \mid (\mathbf{Y}, \boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\rho}) \sim \text{MVN}(\mu_{\text{pred}}, \Sigma_{\text{pred}}), \tag{10}$$

where $\mu_{\text{pred}} = \Sigma_{21}^{\text{new}}(\Sigma_{11}^{\text{new}})^{-1}\mathbf{Y}$ and $\Sigma_{\text{pred}} = \Sigma_{22}^{\text{new}} - \Sigma_{21}^{\text{new}}(\Sigma_{11}^{\text{new}})^{-1}\Sigma_{12}^{\text{new}}$. The matrices Σ_{ij}^{new} are submatrices of Σ^{new} where

$$\Sigma^{\text{new}} = \begin{pmatrix} \Sigma_{11}^{\text{new}} & \Sigma_{12}^{\text{new}} \\ \Sigma_{21}^{\text{new}} & \Sigma_{22}^{\text{new}} \end{pmatrix}.$$

The submatrix Σ_{11}^{new} is an $(n_f + n_h + n_l) \times (n_f + n_h + n_l)$ matrix, while Σ_{12}^{new} and Σ_{21}^{new} are of dimension $(n_f + n_h + n_l) \times 1$ and $1 \times (n_f + n_h + n_l)$, respectively. The remaining submatrix, Σ_{22}^{new}, is a scalar.

To make predictions, we first sample a vector of parameters from Equation (9). Next, conditional on the sampled parameters, a prediction is sampled from Equation (10). The sampling of parameters and predictions is repeated many times to provide estimated posterior quantities (e.g., posterior mean, variance, or prediction intervals).

3. EXAMPLES

In this section, two examples are presented. The first example is a simple simulator that is used to demonstrate the proposed approach. After illustrating our implementation and some diagnostics to assess the adequacy of the model fit, a small simulation study is carried out to investigate the predictive performance of the proposed methodology. The second example is the application that motivated this work, and involves a radiative shock experiment conducted at CRASH. The main goal is to predict

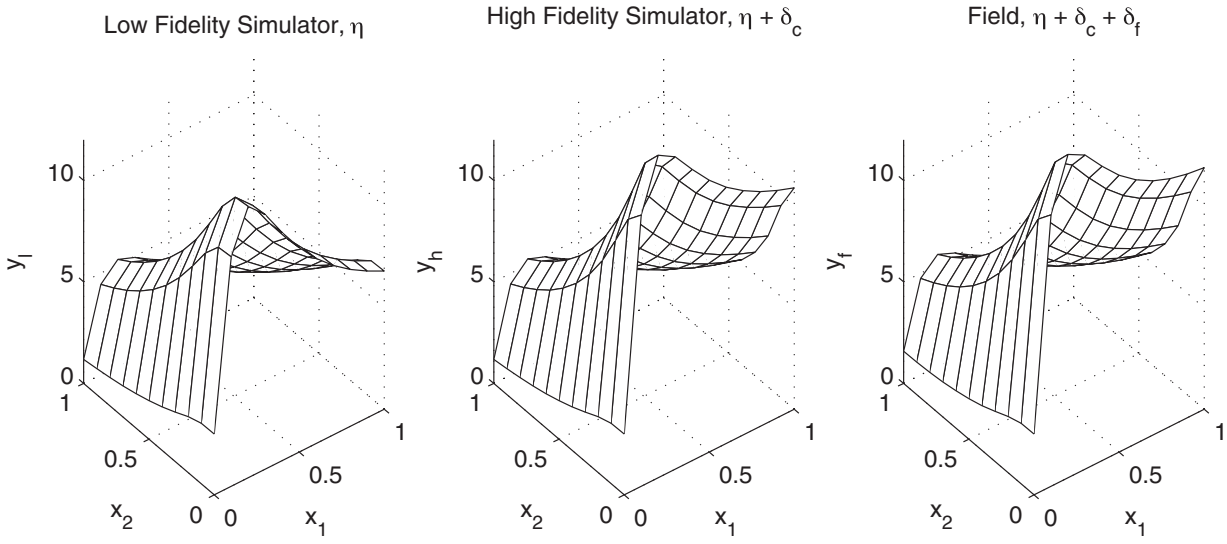


Figure 1. Response surface of the low-fidelity simulator, the high-fidelity simulator, and the mean of the physical process as outlined in Equations (11)–(13).

the observed field measurements given the outputs from two simulators and a set of field trials.

3.1 Toy Example

We begin with the "toy" example in Bastos and O'Hagan (2009), with some slight alterations. That is, the setting has been modified to accommodate two simulators and field experiments. In addition, we refashion the computer models to include two design variables, a common calibration parameter and calibration parameters that exist in each simulator, respectively. For simplicity, all the input settings and calibration parameters are chosen from the unit interval.

We specify the low-fidelity model as:

$$\begin{aligned} y_l(\mathbf{x}, t_f, t_l) &= \eta_l(\mathbf{x}, t_f, t_l) \\ &= \left(1 - \exp\left(-\frac{1}{2x_2}\right)\right) \\ &\quad \times \frac{1000 t_f x_1^3 + 1900 x_1^2 + 2092 x_1 + 60}{1000 t_l x_1^3 + 500 x_1^2 + 4 x_1 + 20}. \end{aligned} \quad (11)$$

The high-fidelity model is defined as the low-fidelity response model plus a discrepancy term:

$$\begin{aligned} y_h(\mathbf{x}, t_f, t_h) &= \eta_l(\mathbf{x}, t_f, \theta_l) + 5\exp(-t_f)\frac{x_1^{t_h}}{100\left(x_2^{2+t_h} + 1\right)} \\ &= \eta_l(\mathbf{x}, t_f, \theta_l) + \delta_c(\mathbf{x}, t_f, t_h). \end{aligned} \quad (12)$$

To illustrate the proposed approach, we simulate outputs from the respective models. Following Loeppky, Sacks, and Welch (2009), we used a 40-run random Latin hypercube design (Mackay, Beckman, and Conover 1979) for the low-fidelity simulator. Since, in practice, the high-fidelity model is likely to be more computationally expensive than the low-fidelity model, only 10 runs are generated—also chosen using a random Latin hypercube design.

In most simulator applications, there are relatively few field observations. Consequently, to mimic this setting, only three field observations were simulated from the mathematical model

$$\begin{aligned} y_f(\mathbf{x}) &= \eta_l(\mathbf{x}, \theta_f, \theta_l) + \delta_c(\mathbf{x}, \theta_f, \theta_h) + \frac{10x_1^2 + 4x_2^2}{50x_1x_2 + 10} + \epsilon \\ &= \eta_l(\mathbf{x}, \theta_f, \theta_l) + \delta_c(\mathbf{x}, \theta_f, \theta_h) + \delta_f(\mathbf{x}) + \epsilon, \end{aligned} \quad (13)$$

where $\epsilon \sim N(0, 0.5^2)$.

For this example, the true value of the common calibration parameter is chosen to be $\theta_f = 0.2$, while the calibration parameter appearing only in the high- and low-fidelity models is chosen to be $\theta_h = 0.3$ and $\theta_l = 0.1$, respectively.

Figure 1 displays the response surfaces for the two simulators and also the mean response surface for the field process. A quick glance at the figure reveals that the high-fidelity model appears closer to the mean process than the low-fidelity model. This represents the framework we are working within insofar as the high-fidelity model is assumed to be more like the true system than the low-fidelity model.

The posterior distribution of the model parameters was sampled using MCMC as outlined in Section 2.2.1. The MCMC chain is initialized with $\theta_f = \theta_h = \theta_l = 0.5$ (i.e., the center of the input space), $\lambda_{\eta_l} = 1$, $\lambda_2 = \lambda_f = \lambda_y = 20$ and all the correlation parameters, $\boldsymbol{\rho}$ are chosen to be 0.1 as we assume that the simulator and discrepancies are dependent on all the inputs. Through visual inspection of the traceplots (not shown), we found that, for the data encountered in this example, convergence is achieved in the first 1000 steps or so. The MCMC was run for 10,000 steps, where the first 2000 steps are treated as burn-in and discarded in further analysis. The code for this analysis can be found in the online supplementary materials.

In addition to the data simulated from Equations (11)–(13) used to fit the proposed model (i.e., the training set), a validation dataset was generated from Equation (13), so that the predictive performance can be evaluated. The validation set consisted of 25 field observations with input settings, $\mathbf{x}$, chosen using random Latin hypercube sampling. We use the posterior mean prediction at $\mathbf{x}$ to estimate $Y_f(\mathbf{x})$. Figure 2 shows the predicted versus actual values for each of the validation points. The figure shows that

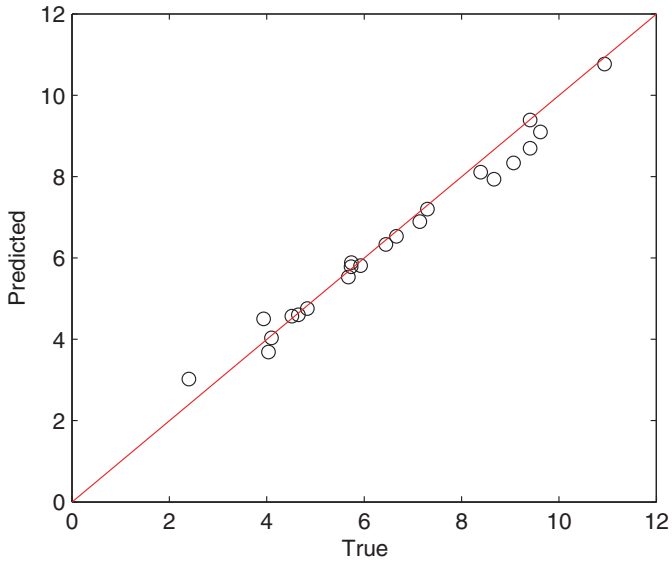


Figure 2. Predicted versus actual field measurements of the validation set (with the $y = x$ line). The online version of this figure is in color.

the predictive model performs reasonably well since the points center around the $y = x$ line.

Figure 3 displays the deviations of the predictions from the true values plotted against the predictions and also the input settings in each dimension. In each case, no obvious pattern is found in the plots, suggesting the outputs have similar degree of smoothness across the input space and that no obvious systematic behavior was unaccounted for.

While not the specific goal of the proposed methodology, we now consider the estimation of the calibration parameters. Figure 4 shows the estimated one-dimensional and two-dimensional marginal posterior distributions of the calibration parameters. Vertical dashed lines are plotted at the true values of the calibration parameters. In general, these posterior distributions can be interpreted as representing the uncertainty in the calibration parameters given the very limited number of observations and small numbers of simulations from imperfect simulators. A quick glance at the plots reveals that, except for θ_l, the calibration parameters are not being constrained by the data. It is not too surprising that we can constrain θ_l, but not the calibration of the other parameters, since there are more outputs (comparisons between the low- and high-fidelity models) to inform this parameter. The inability to constrain the other calibration parameters is likely due to the well-known issue of confounding between the calibration parameters and the discrepancy functions (e.g., Loeppky, Bingham, and Welch 2006) and the dearth of data. The confounding can be mitigated to some extent by the use of more informative prior distributions and also more observations.

The diagonals show the marginal posterior distributions of the calibration parameters, with the true values marked with vertical dashed lines. The off-diagonals subplots contain the two-dimensional marginal posterior distributions for the three calibration parameters. The solid lines represent the 95% high posterior density region.

Figures 4(b), 4(c), and 4(d) display the estimated posterior distributions of the calibration parameters for other sample sizes. The panels are the results of the simulations with (i) $n_l = n_h = 20$, $n_f = 3$, (ii) $n_l = n_h = n_f = 40$, and (iii) $n_l = n_h = n_f = 100$. The first case was chosen as a more simulation rich version of the earlier example. Comparing Figure 4(b) with the results in Figure 4(a), we see that the mode of the posterior distribution of θ_l is closer to the true value (solid line) and there is less variability in the posterior distribution when there are more simulations. However, very little is learned about the calibration parameters θ_h and θ_f. To gain more information on these parameters, there needs to be more field observations. Panels (c) and (d) consider cases where the number of simulations and field trials is larger than before. As the number of observations and simulations increases, the model is able to better estimate the calibration parameters. An interesting observation is that the shared calibration parameter θ_f is better constrained in panel (c) than θ_h. The reason for this, we surmise, is that given the same number of field trials both the low and high-fidelity models help inform θ_f, but only the high-fidelity model directly informs θ_h. When there are relatively many simulations and observations, all of the calibration parameters tend to be well constrained (panel (d)).

A subsequent simulation study is performed to compare predictions of the new model with approaches that only use some of the simulations. Models ML and MH are implementations of the Kennedy and O'Hagan (2001) approach using the data obtained from the low-fidelity simulator and experiments, and outputs obtained from the high-fidelity simulator and experiments,

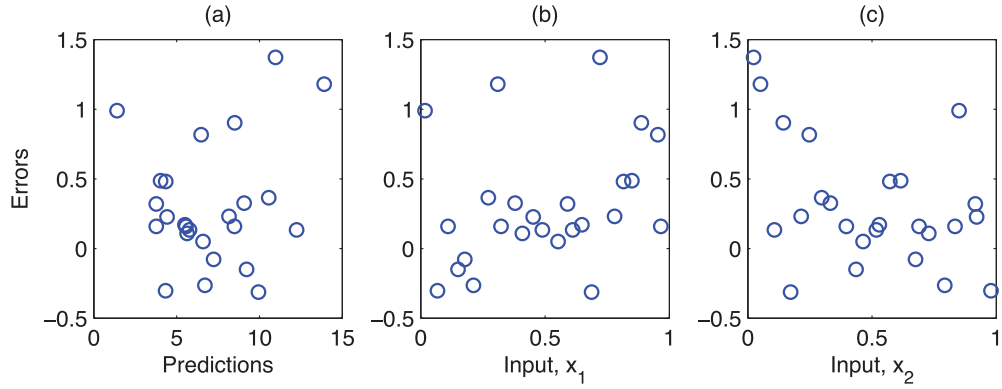
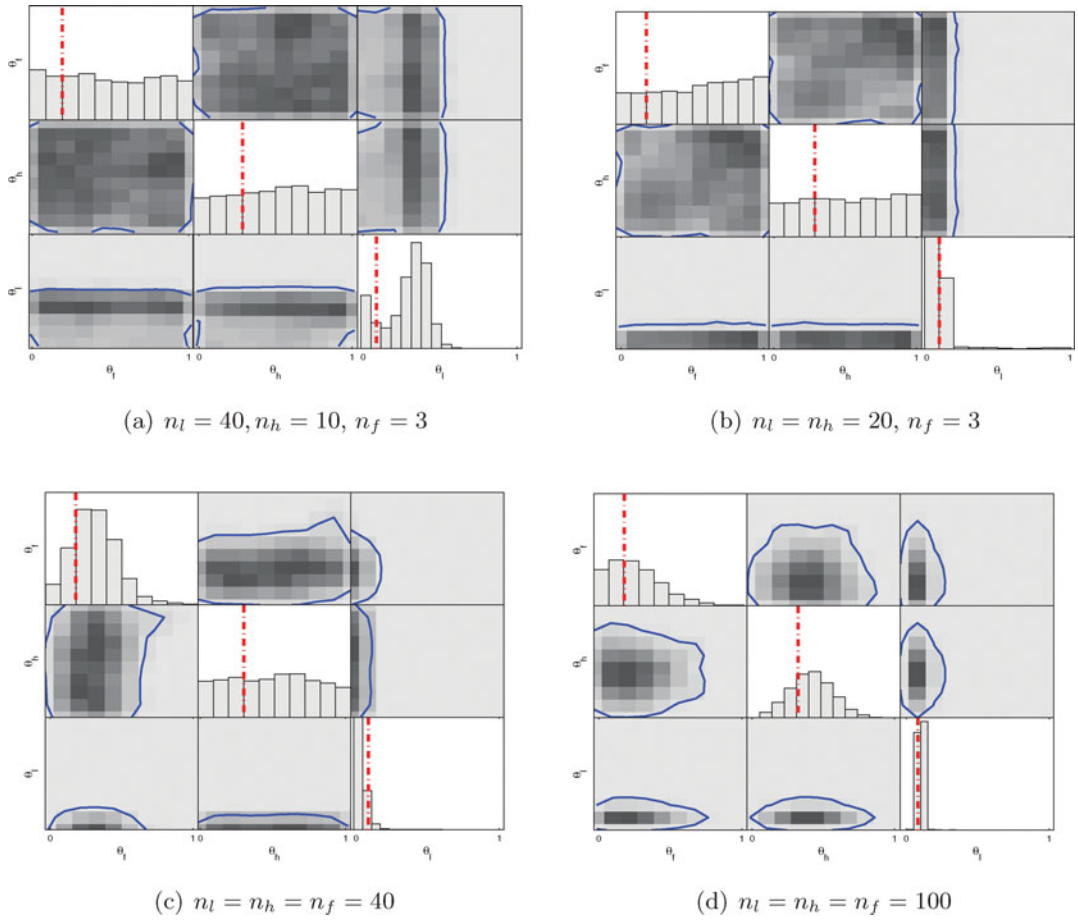


Figure 3. Diagnostics plots for the simple example: (a) Prediction error against predictions; (b) prediction error against x_1; (c) prediction error against x_2. The online version of this figure is in color.



(a) $n_l = 40, n_h = 10, n_f = 3$ (b) $n_l = n_h = 20, n_f = 3$

(c) $n_l = n_h = n_f = 40$ (d) $n_l = n_h = n_f = 100$

Figure 4. Plot of the two-dimensional marginals for the posterior distributions of the three parameters for different sample sizes. The online version of this figure is in color.

respectively. Predictions from these models are compared with those from the proposed approach—denote MLH. In other words, we are investigating whether the proposed methodology of combining all simulations and observations is better in some sense than the Kennedy and O'Hagan (2001) method using one of either the low-fidelity model or high-fidelity model outputs alone.

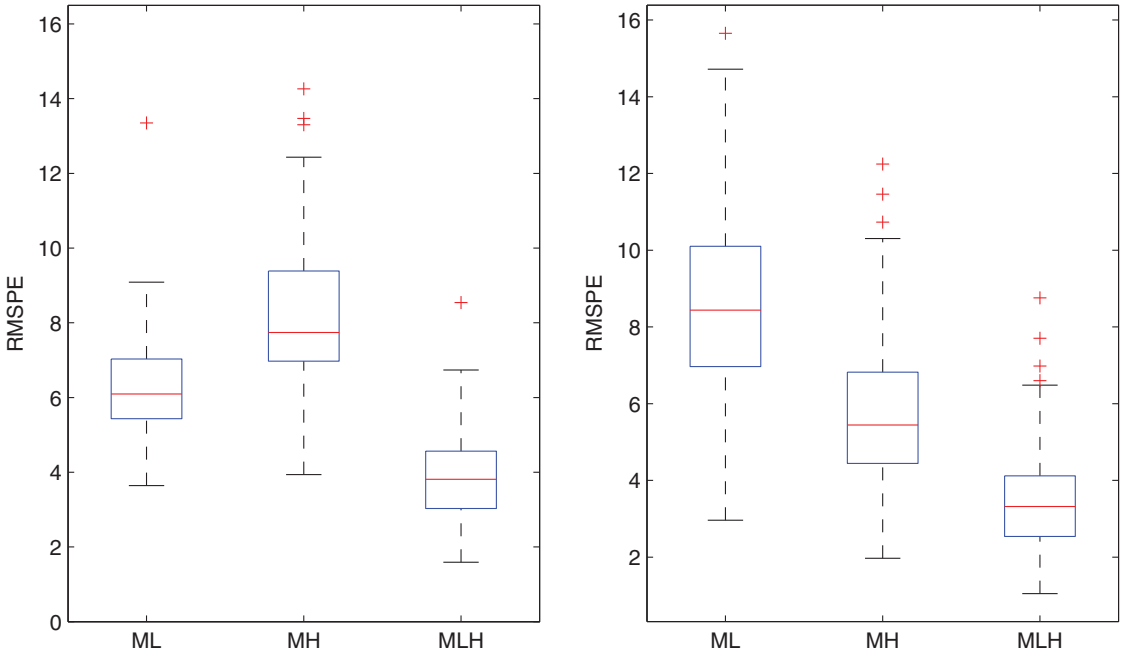
The simulation study is carried out as follows. Using random Latin hypercube sampling, 100 sets of training and validation data are first generated independently. Each training set contains the same number of outputs as the first illustration: 40 simulated values from the low-fidelity simulator, 10 computer runs from the high-fidelity simulator and 3 field observations. For each simulated training set, models ML, MH, and MLH are estimated, and predictions of the validation set are obtained from each model. The predictions are evaluated by computing the root mean squared prediction errors (RMSPE) for the validation data. This is done for each of the 100 simulated training and validation datasets. The simulation study results are summarized in Figure 5.

Figure 5 reveals that the RMSPE from the proposed model is consistently smaller than the RMSPE of the other two models. Interestingly, in panel (a), we notice that the RMSPE is larger for the high-fidelity model than the low-fidelity model. This is the result of having relatively few runs of the high-fidelity code. Looking at Figure 5(b), when $n_l = n_h = 20$ and $n_f = 3$, prediction using the higher fidelity outputs does better than prediction using only the low-fidelity outputs. In either case, the proposed approach that uses all sources of data tends to do better in terms of RMSPE.

In general, we found that the proposed model that makes use of all the simulations works well for making predictions for the physical system. The simulation demonstrates that more efficient estimation is gained through this approach. Although calibration is not the priority, we come across a similar issue encountered by Kennedy and O'Hagan (2001)—calibration is difficult with limited amounts of data. However, as the number of outputs and observations increases, more information is available to calibrate the parameters of interest. In the case of calibration in our setting, it is important to note what is being achieved. That is, the posterior distributions reflect the uncertainty in the calibration parameters given the observations and the imperfect simulators.

3.2 CRASH Application

The application that motivated the proposed methodology arises from radiative shock experiments at CRASH. Figure 6 gives a diagram of the system that we want to predict. In the physical experiments, a high-energy laser pulse irradiates a thin disk of beryllium at the front end of a xenon-filled tube. The energy deposited in the surface causes the beryllium to ablate. A shock wave is then driven by the ablation pressure through the beryllium disk. After the shock wave breaks out of the



(a) $n_l = 40, n_h = 10$ and $n_f = 3$ (b) $n_l = n_h = 20$ and $n_f = 3$

Figure 5. Boxplots of the RMSPE obtained from the 100 simulated datasets analyzed using models ML, MH, and MLH. The online version of this figure is in color.

beryllium disk, the disk acts as a piston, propagating the shock at a high speed into the xenon. When the xenon is shocked, it is heated to temperatures well over 100,000°K and emits thermal X-ray radiation. These shocks are considered radiative when the radiation energy flux from the shock is high enough to impact the structure of the shock wave. Details regarding the radiative shock physics can be found in Drake et al. (2011). The radiating shock experiments that we are concerned with can be viewed as small-scale experiments for understanding astrophysical shock waves and other high-temperature phenomena (Drake et al. 2011; McClarren et al. 2011).

Several measurements of interest are taken from each shock experiment and also simulations. We focus here on the time taken for the shock wave to exit the beryllium disk (breakout time). Our experiments were carried out at the OMEGA Laser Facility at the University of Rochester (Boehly et al. 1997). Two metrics were used to obtain measurements of the shock breakout. The streaked optical pyrometer (SOP) records a two-dimensional image with optical light emission in one spatial dimension and is streaked in time in the other direction (Miller et al. 2007). To measure shock breakout, SOP views beryllium disk on the side opposite the laser irradiation. The detector sees no signal from the hot, shocked material until the shock reaches the back surface of the disk, indicating shock breakout. The other metric is the active shock breakout (ASBO) diagnostic (Barker and Hollenback 1972). ASBO reflects a probe laser off the back surface of the beryllium disk and uses interferometry for measuring the relative distance from the target to the diagnostic over time. When the shock reaches the back surface of the disk, it accelerates the surface, changing the distance from the detector to the target. Both the SOP and ASBO metrics can be used to provide measurements of the shock breakout time. In practice, we have used both to compute the breakout time for a shock and taken the average of the two inferred computed values as the measurement. Of course, the shocks move very quickly (more than 100 km/s) and the breakout from the beryllium happens in a very short time period. For this setup, the time measurement system measures the breakout time in picoseconds (10^{-12} s).

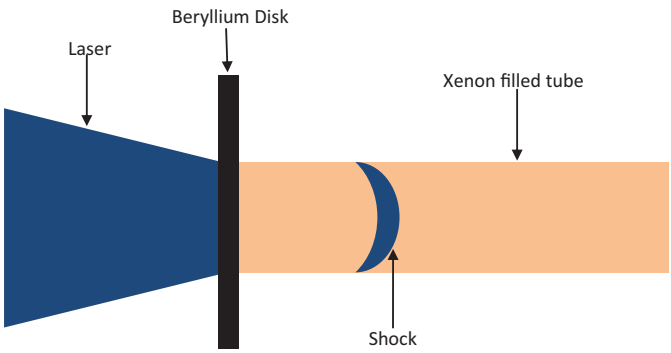


Figure 6. A pictorial version of the apparatus used in the radiative shock experiments. The vertical bar represents the beryllium disk where the laser deposits energy. The shock wave breaks through the beryllium disk and moves down the xenon-filled tube (horizontal bar). The online version of this figure is in color.

Using two different radiation-hydrodynamics codes (1D-CRASH and 2D-CRASH), we aim to predict the shock breakout time. The 2D-CRASH code includes two-dimensional processes and interactions that the one-dimensional code, or 1D-CRASH, does not. As a result, the 2D-CRASH model is not only assumed to be able to model the experiments better than the 1D-CRASH code, but it is also more computationally expensive.

The design variables for this experiment are the thickness of the beryllium disk (x_1) and laser energy (x_2). The electron flux limiter is calibration input to both simulators and is denoted

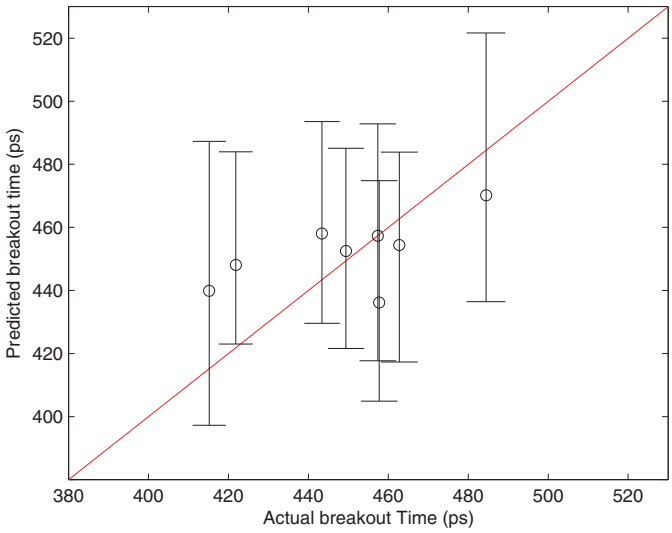


Figure 7. Predicted versus actual breakout times and 95% prediction intervals. The online version of this figure is in color.

as t_f. The laser energy scale factor is an additional calibration parameter, t_l, required by the 1D-CRASH code but not the 2D-CRASH simulator. The high-fidelity computer code has two calibration inputs—beryllium gamma ($t_{h,1}$) and wall opacity scale factor ($t_{h,2}$). All the inputs are scaled to the unit interval before fitting the data to the proposed model.

We have 365 simulations from 1D-CRASH and 104 2D-CRASH runs available. The designs for each computer experiment were Latin hypercube designs, optimized using a space-filling criterion (Johnson, Moore, and Ylvisaker 1990). There are also eight experiments that were conducted using the OMEGA laser where the breakout time was recorded.

The MCMC was set up as in the previous examples, with one exception. From previous usage of the laser, it was known that the observational standard error was about 50×10^{-12} s—or approximately 1 after standardizing. A gamma distribution with shape and scale parameter (10,000, 10,000) was chosen for the prior of λ_y. This is an informative prior that tightly centers the gamma distribution at 1. The widths for the Metropolis updates are chosen as outlined in Section 2.2.1. We found that convergence was achieved shortly after 1000 MCMC steps. So, the MCMC was run for 10,000 steps and the first 2000 were discarded as burn-in.

Like the previous example, the deviations of the predictions from the observed breakout times are plotted against the predictions and the two input settings (diagnostic plots not shown). No obvious pattern is found in any of the diagnostic plots, thereby suggesting that the model fit is adequate. The code for this analysis can be found in the online supplementary materials.

A leave-one-out study is conducted to evaluate the predictive ability of the new approach. That is, we delete an observation, fit the proposed model, and predict the deleted observation. This is done for each of the eight observations. Figure 7 is a plot of the resulting predictions against the observed breakout time. The 95% posterior prediction interval for each point is shown in the figure. The predictions are fairly close to the observed values and, thus most points are near to the $y = x$ line. However, the second observation from the left gives a prediction interval that almost fails to capture the observation.

Similar to the previous example, the proposed approach of combining all simulations and observations is compared to the Kennedy and O'Hagan (2001) method using one of either the 1D-CRASH or 2D-CRASH outputs alone. The results are shown in Figure 8. Looking at the Figure, the proposed methodology

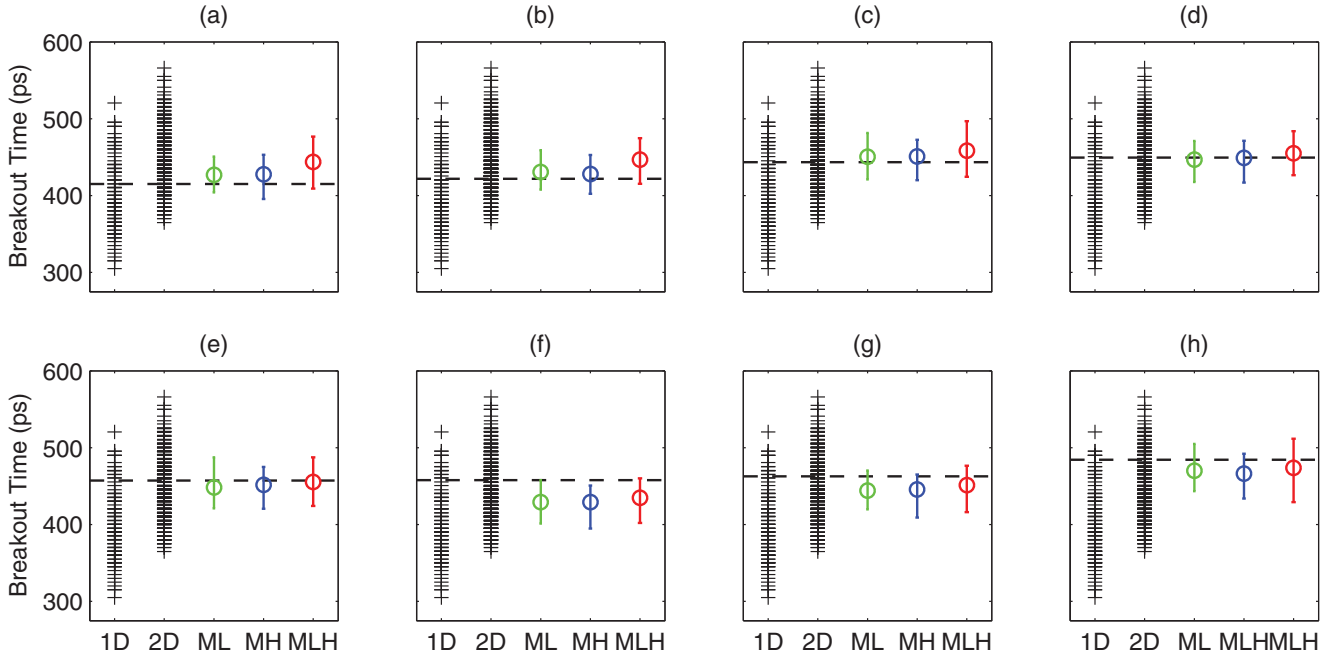


Figure 8. The horizontal dotted line in each subplot draws the actual observed breakout time. The first two bars of "+" on the far left of the plots are the simulated outputs from 1D and 2D-CRASH, respectively. The 95% prediction intervals, denoted as vertical intervals, are obtained from fitting models ML, MH, and MLH, respectively. The mean of each prediction interval is denoted with circle. The online version of this figure is in color.

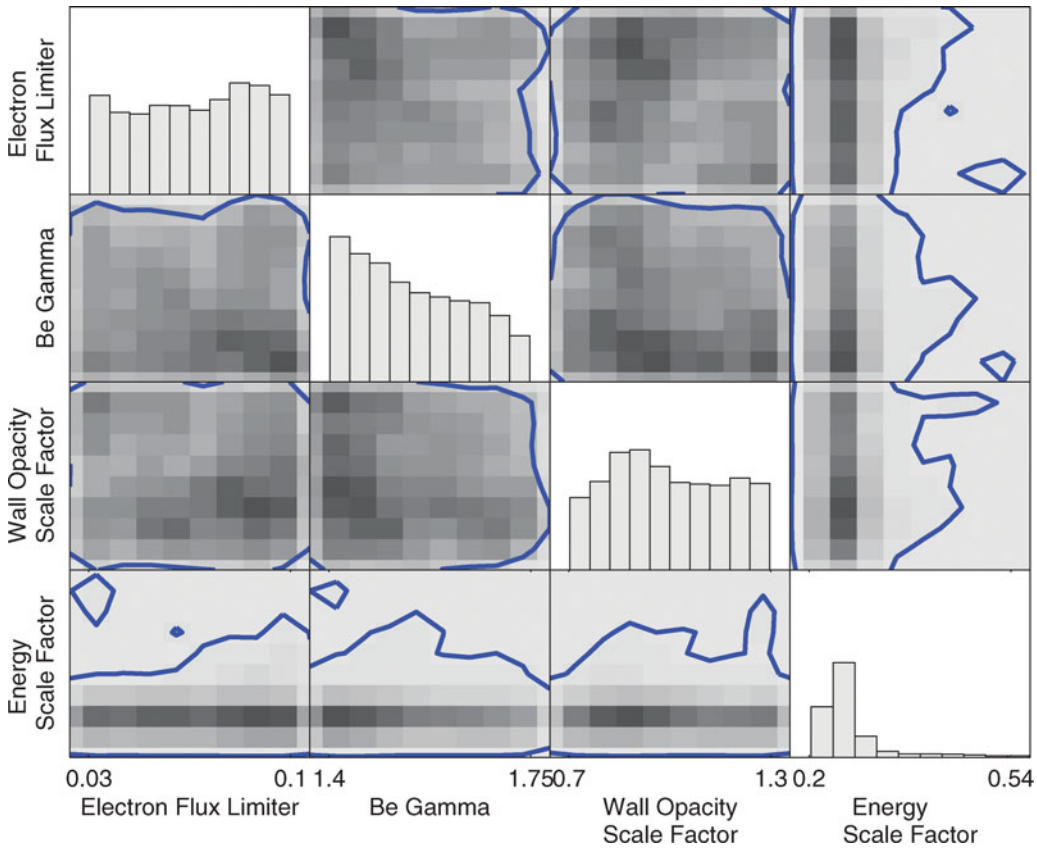


Figure 9. Plot of the two-dimensional marginals for the posterior distribution of the four parameters. The diagonals show the marginal posterior distributions of the calibration parameters. The off-diagonals subplots contain the two-dimensional marginal posterior distributions for the four calibration parameters. The solid lines represent the 95% high posterior density region. The online version of this figure is in color.

has prediction intervals that are usually smaller than prediction intervals obtained from model ML and MH—though not universally so. The prediction intervals from model MH are generally the widest. The prediction intervals from the new methodology contained the observations, but the prediction of breakout time for shot (f) was outside of both prediction intervals from ML and MH.

The results in Figure 8 point to the proposed approach being generally more successful at predicting the breakout time than using the observations with the low- or high-fidelity simulators alone. With that said, the results are not as striking as in the previous section. The benefits of the proposed methodology will depend on issues related to the specific application. In the end, the quality of the predictions will be based on features such as the ability of the simulators to mimic the real process, the form of the discrepancies, and the number of simulations and observations. In this example, we had quite a few runs from both simulators and emulated both simulators fairly effectively—we could not know we could do so beforehand—but did not have very many observations.

Plots of the marginal posterior distributions of the calibration parameters are shown in Figure 9. The posterior distributions for all the calibration parameters, except the energy scale factor, are not constrained in this application. This is expected because of the limited number of experiments ($n_f = 8$) that inform these parameters.

4. DISCUSSION

In this section, some extensions and limitations of the proposed approach are discussed. In addition, we identify some avenues for future work.

So far, the focus has been on the setting where there are only two simulators. The new methodology, however, can easily be extended to model applications that involve more than two simulators. Suppose that there are H simulators denoted as $\eta_k(\cdot)$ for $k = 1, \ldots, K$, where $\eta_k(\cdot)$ is the next highest level of fidelity model from $\eta_{k-1}(\cdot)$. The simulators share the same design variables, $\mathbf{x}$, and some common calibration parameters, $\mathbf{t}_f$. The remaining calibration parameters required by each of the respective simulators are denoted as $\mathbf{t}_k$, for $k = 1, \ldots, K$. The intersection between $\mathbf{t}_k$ and $\mathbf{t}_k'$ $(k \neq k')$ is empty and thus calibration parameters are included in either exactly one or all of the simulators.

The lowest fidelity simulator outputs are denoted as

$$Y_1(\mathbf{x}, \mathbf{t}_f, \mathbf{t}_1) = \eta_1(\mathbf{x}, \mathbf{t}_f, \mathbf{t}_1).$$

The outputs from the higher fidelity simulators can then be written as a combination of the lowest fidelity simulator and discrepancy functions that capture the systematic differences between pairs of simulators. For $k = 2, \ldots, K$, the simulated outputs are written as

$$Y_k(\mathbf{x}, \mathbf{t}_f, \mathbf{t}_h) = \eta_k(\mathbf{x}, \mathbf{t}_f, \mathbf{t}_k) = \eta_1(\mathbf{x}, \mathbf{t}_f, \boldsymbol{\theta}_1) + \sum_{j=2}^{k-1} \delta_j(\mathbf{x}, \mathbf{t}_f, \boldsymbol{\theta}_j) + \delta_k(\mathbf{x}, \mathbf{t}_f, \mathbf{t}_k).$$

The experimental observations are written as the sum of the low-fidelity simulator and discrepancy functions

$$Y_f(\mathbf{x}) = \eta_K(\mathbf{x}, \boldsymbol{\theta}_f, \boldsymbol{\theta}_K) + \delta_f(\mathbf{x}) + \epsilon = \eta_1(\mathbf{x}, \boldsymbol{\theta}_f, \boldsymbol{\theta}_1) + \sum_{j=2}^{K} \delta_j(\mathbf{x}, \boldsymbol{\theta}_f, \boldsymbol{\theta}_j) + \delta_f(\mathbf{x}) + \epsilon,$$

where $\delta_f(\mathbf{x})$ measures the discrepancy between the highest fidelity simulator and physical process. The response surfaces of the different sources of data are modeled with GPs with mean and covariance functions discussed in Section 2.2.

While it is conceptually simple to extend the setting to more than two simulators, note that as the number of models grows, likely so too will the number simulations. As a result, the covariance matrices (e.g., for the low-fidelity model) can become so large that matrix inversion poses an additional computation problem. This issue occurs for all applications of GPs where many outputs are available. One way to deal with large datasets is to change the GP specifications and use a compactly supported covariance that reduces the computational effort through sparse matrix techniques (e.g., Kaufman et al. 2011). Alternatively, a multistage approach can be considered. That is, one would first emulate the lowest fidelity simulator using only outputs from that computer model. The next highest level of fidelity model is calibrated using outputs from this simulator and the lowest level of fidelity emulator. One can continue building models in a hierarchical fashion. This has the impact on reducing the size of the covariance matrices at each stage and thus the computational burden. A similar approach is used for implementing the Kennedy and O'Hagan model in Bayarri et al. (2007).

Our choice of sampling method from the posterior distribution—univariate random walk MCMC—has worked fairly well in the examples encountered. However, as pointed out by one of the referees, we would expect that, as the parameter space grows, a relatively large number of MCMC steps are required to allow the random walks to converge. In these cases, more efficient MCMC procedures are needed.

Some care should be taken in the prior specification for the precision parameters for the GPs. We have found that the default choices of prior distributions outlined in Section 2.2 work fine in most cases (e.g., the simulations in Section 3.1). However, for some datasets, extremely large values of λ_y are observed. This amounts to essentially a model with no measurement error and discrepancies that are interpolating the noise. We noticed the phenomenon when the default priors are used for the CRASH example. This can also happen with the model proposed by Kennedy and O'Hagan (2001). In our case, we avoided this problem because we had a more informative prior distribution for λ_y. Alternatively, one can address this issue by rejecting small values of a precision parameter in the MCMC (this was done in Higdon et al. 2004), or at the design stage by taking replicate field observations.

A further note of caution with respect to the experimental design is that the design regions for the computer experiments should coincide to avoid uncertainty due to extrapolation in the discrepancies between models. Suppose for example, the design for $\mathbf{t}_f$ in the low-fidelity simulator explores a much larger region than the design for the high-fidelity model. When predictions are made, the proposed approach averages over the posterior distribution of the calibration parameters. For values of $\boldsymbol{\theta}_f$ from the posterior that are outside of the range explored by the design of the high-fidelity model, the proposed approach extrapolates $\delta_2(\cdot)$. This results in larger prediction intervals.

Finally, there are some avenues for future research that can be envisioned in other applications. For example, the proposed model could be adapted to consider different output data-structures such as functional or spatial responses. Furthermore, our approach is not obviously suitable to settings where the simulators are not ranked by fidelity (e.g., climate models arising from different research groups), in applications with more than two simulators that involve common calibration parameters that appear among a proper subset of the models or in settings with design variables that appear in only some, but not all, simulators.

5. CONCLUSION

A new methodology, which combines outputs from multifidelity simulators and field observations, is proposed. The approach successfully uses a Bayesian hierarchical model to make predictions of the physical system with associated measurements of uncertainty (e.g., posterior variance or prediction intervals). Different GPs are used to model the various response surfaces. The real example that motivated this work used two simulators of the process, but methodology can be easily extended to cases with more than two simulators.

SUPPLEMENTARY MATERIALS

Code and data: a zip folder containing Matlab code for the toy example and the CRASH application, as well as the data from the CRASH application.

ACKNOWLEDGMENTS

This work is funded by the Predictive Sciences Academic Alliances Program in NNSA-ASC via grant DEFC52-08NA28616 and the Natural Sciences and Engineering Research Council of Canada. The authors are grateful for the encouraging and helpful comments made by the referees, Associate Editor, and Editor. The implementation of the proposed methodology is built upon the Gaussian Process Modeling for Simulation Analysis software (Gattiker, Higdon, and Williams 2008) developed at Los Alamos National Laboratory. The authors would like to thank the Los Alamos National Laboratory Statistical Sciences Group for sharing their libraries.

[Received June 2012. Revised August 2013.]

REFERENCES

Barker, L., and Hollenback, R. (1972), "Laser Interferometer for Measuring High Velocities of Any Reflecting Surface," *Journal of Applied Physics*, 43, 4669–4675. [508]

Bastos, L. S., and O'Hagan, A. (2009), "Diagnostics for Gaussian Process Emulators," *Technometrics*, 51, 425–438. [505]

Bayarri, M. J., Berger, J. O., Paulo, R., Sacks, J., Cafeo, J., Cavendish, J., Lin, C., and Tu, J. (2007), "A Framework for Validation of Computer Models," *Technometrics*, 49, 138–154. [511]

Boehly, T. R., Brown, D. L., Craxton, R. S., Keck, R. L., Knauer, J. P., Kelly, J. H., Kessler, T. J., Kumpan, S. A., Loucks, S. J., Letzring, S. A., Marshall, F. J., McCrory, R. L., Morse, S. F. B., Seka, W., Soures, J. M., and Verdon, C. P. (1997), "Initial Performance Results of the OMEGA Laser System," *Optics Communications*, 133, 495–506. [508]

Cumming, J. A., and Goldstein, M. (2009), "Small Sample Bayesian Designs for Complex High-Dimensional Models Based on Information Gained Using Fast Approximations," *Technometrics*, 51, 377–388. [501]

Drake, R. P., Doss, F. W., McClarren, R. G., Adams, M. L., Amato, N., Bingham, D., Chou, C. C., DiStefano, C., Fidkowski, K., Fryxell, B., Gombosi, T. I., Grosskopf, M. J., Holloway, J. P., van der Holst, B., Huntington, C. M., Karni, S., Krauland, C. M., Kuranz, C. C., Larsen, E., van Leer, B., Mallick, B., Marion, D., Martin, W., Morel, J. E., Myra, E. S., Nair, V., Powell, K. G., Rauchwerger, L., Roe, P., Rutter, E., Sokolov, I. V., Stout, Q., Torralva, B. R., Toth, G., Thornton, K., and Visco, A. J. (2011), "Radiative Effects in Radiative Shocks in Shock Tubes," *High Energy Density Physics*, 7, 130–140. [508]

Gattiker, J., Higdon, D., and Williams, B. (2008), *GPM/SA (Gaussian Process Modeling for Simulation Analysis)*; software available from the Statistical Sciences Group at Los Alamos National Laboratory. Available at *http://www.stat.lanl.gov/source/orgs/ccs/ccs6/gpmsa/gpmsa.html* [511]

Gelman, A., Carlin, J. B., Stern, H. S., and Dublin, D. B. (2004), *Bayesian Data Analysis* (2nd ed.), Boca Raton, FL: Chapman & Hall. [504]

Graves, T. (2005), "Automatic Step Size Selection in Random Walk Metropolis Algorithm," Technical Report, Los Alamos National Laboratory. [504]

Hastings, W. K. (1970), "Monte Carlo Sampling Methods Using Markov Chains and Their Aplications," *Biometrika*, 57, 97–109. [504]

Higdon, D., Gattiker, J., Williams, B., and Rightley, M. (2008), "Combining Field Data and Computer Simulations for Calibration and Prediction," *Journal of the American Statistical Association*, 103, 570–583. [501,503,504]

Higdon, D., Kennedy, M., Cavendish, J., Cafeo, J., and Ryne, R. D. (2004), "Combining Field Data and Computer Simulations for Calibration and Prediction," *SIAM Journal of Scientific Computing*, 26, 448–466. [501,504,511]

Johnson, M. E., Moore, L. M., and Ylvisaker, D. (1990), "Minimax and Maximin Distance Designs," *Journal of Statistical Planning and Inference*, 26, 131–148. [509]

Kaufman, C. G., Bingham, D., Habib, S., Heitmann, K., and Frieman, J. A. (2011), "Efficient Emulators of Computer Experiments Using Compactly Supported Correlation Functions, With an Application to Cosmology," *Annals of Applied Statistics*, 5, 2470–2492. [511]

Kennedy, M., and O'Hagan, A. (2000), "Predicting the Output From a Complex Computer Code When Fast Approximations Are Available," *Biometrika*, 87, 1–13. [501,502]

Kennedy, M., and O'Hagan, A. (2001), "Bayesian Calibration of Computer Models" (with discussion), *Journal of the Royal Statistical Society*, Series B, 68, 425–464. [501,502,503,506,507,509,511]

Linkletter, C., Bingham, D., Hengartner, N., Higdon, D., and Ye, K. Q. (2006), "Variable Selection for Gaussian Process Models in Computer Experiments," *Technometrics*, 48, 478–490. [503,504]

Loeppky, J. L., Bingham, D., and Welch, W. L. (2006), "Computer Model Calibration or Tuning in Practice," Technical Report, University of British Columbia. [506]

Loeppky, J. L., Sacks, J., and Welch, W. L. (2009), "Choosing the Sample Size of a Computer Experiment: A Practical Guide," *Technometrics*, 51, 366–376. [505]

Mackay, M. D., Beckman, R. J., and Conover, W. J. (1979), "A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output From a Computer Code," *Technometrics*, 21, 239–245. [505]

McClarren, R. G., Ryub, D., Drake, P., Grosskopf, M., Bingham, D., Chou, C.-C., Fryxell, B., van der Holst, B., Holloway, J. P., Kuranz, C. C., Mallick, B., Rutter, E., and Torralva, B. (2011), "A Physics Informed Emulator for Laser-Driven Radiating Shock Simulations," *Reliability Engineering and System Safety*, 96, 1194–1207. [508]

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. (1953), "Equations of State Calculations by Fast Computing Machines," *Journal of Chemical Physics*, 21, 1087–1091. [504]

Miller, J. E., Boehly, T. R., Melchior, A., Meyerhofer, D. D., Celliers, P. M., Eggert, J. H., Hicks, D. G., Sorce, C. M., Oertel, J. A., and Emmel, P. M. (2007), "Streaked Optical Pyrometer System for Laser-Driven Shock-Wave Experiments on OMEGA," *Review of Scientific Instruments*, 78, 034901–034901-6. [508]

Paulo, R., Garca-Donato, G., and Palomo, J. (2012), "Calibration of Computer Models With Multivariate Output," *Computational Statistics and Data Analysis*, 56, 3959–3974. [501]

Qian, Z. G., Seepersad, C. C., Joseph, V. R., Allen, J. K., and Wu, J. C. F. (2006), "Building Surrogate Models Based on Detailed and Approximate Simulations," *ASME Journal of Mechanical Design*, 128, 668–677. [501,502]

Qian, P. Z. G., and Wu, J. C. F. (2008), "Bayesian Hierachical Modeling for Integrating Low-Accuracy and High-Accuracy Experiments," *Technometrics*, 50, 192–204. [501]

Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P. (1989), "Design and Analysis of Computer Experiments," *Statistical Science*, 4, 409–423. [501,503]

Santner, T. J., Williams, B. J., and Notz, W. I. (2003), *The Design and Analysis of Computer Experiments*, New York: Springer-Verlag. [501]

Welch, W. J., Buck, R. J., Sacks, J., Wynn, H. P., Mitchell, T. J., and Morris, M. D. (1992), "Screening, Predicting, and Computer Experiments," *Technometrics*, 34, 15–25. [501]

Williams, B., Higdon, D., Gattiker, J., Moore, L., McKay, M., and Keller-McNulty, S. (2006), "Combining Experimental Data and Computer Simulations, With an Application to Flyer Plate Experiments," *Bayesian Analysis*, 1, 765–792. [504]