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

Spline-Based Emulators for Radiative Shock Experiments With Measurement Error

Avishek CHAKRABORTY, Bani K. MALLICK, Ryan G. MCCLARREN, Carolyn C. KURANZ, Derek BINGHAM, Michael J. GROSSKOPF, Erica M. RUTTER, Hayes F. STRIPLING, and R. Paul DRAKE

Radiation hydrodynamics and radiative shocks are of fundamental interest in the high-energy-density physics research due to their importance in understanding astrophysical phenomena such as supernovae. In the laboratory, experiments can produce shocks with fundamentally similar physics on reduced scales. However, the cost and time constraints of the experiment necessitate use of a computer algorithm to generate a reasonable number of outputs for making valid inference. We focus on modeling emulators that can efficiently assimilate these two sources of information accounting for their intrinsic differences. The goal is to learn how to predict the breakout time of the shock given the information on associated parameters such as pressure and energy. Under the framework of the Kennedy–O'Hagan model, we introduce an emulator based on adaptive splines. Depending on the preference of having an interpolator for the computer code output or a computationally fast model, a couple of different variants are proposed. Those choices are shown to perform better than the conventional Gaussian-process-based emulator and a few other choices of nonstationary models. For the shock experiment dataset, a number of features related to computer model validation such as using interpolator, necessity of discrepancy function, or accounting for experimental heterogeneity are discussed, implemented, and validated for the current dataset. In addition to the typical Gaussian measurement error for real data, we consider alternative specifications suitable to incorporate noninformativeness in error distributions, more in agreement with the current experiment. Comparative diagnostics, to highlight the effect of measurement error model on predictive uncertainty, are also presented. Supplementary materials for this article are available online.

KEY WORDS: Adaptive spline; Computer model validation; Emulator; Measurement error model; Non-Gaussian error; Reversible jump Markov chain Monte Carlo.

1. INTRODUCTION

High-energy density physics (HEDP) studies the behavior of systems with a pressure at or above 1 million times atmospheric pressure (Drake 2006). Such high-energy-density systems occur in astrophysical phenomena [e.g., supernovae explosions (Chevalier 1997)]. Given advances in laser technology, the high-energy-density regime is now routinely accessible in laboratory experiments where a laser focused on a target accelerates material to create shock waves of about 10 kilometers per second. Shock waves traveling at these extreme speeds radiate light in the X-ray spectrum (as a result of black-body radiation emission) that fundamentally alters the propagation of the shock when compared with traditional shocks such as the shock wave created by supersonic aircraft. These shocks are said to be radiative shocks and described by the radiation-hydrodynamics physics model comprised of traditional hydrodynamics augmented with equations that govern the transport of radiation.

Using the Omega Laser Facility at Rochester University (Boehly et al. 1995), several experimental campaigns have been conducting HEDP experiments concerned with understanding radiative shocks. In these experiments, a disk of beryllium (atomic symbol Be) is placed at the end of a plastic tube of xenon (atomic symbol Xe). Then the laser is discharged onto the Be disk; the energy deposition of the laser causes a layer of Be to ablate thereby accelerating a shock into the Be. This shock travels through the Be and "breaks out" of the disk into the Xe gas where radiography is able to capture images of the shock structure. These experiments require the dedication of large amounts of resources in terms of laser time, fabrication costs, and experimentalist time and, as a result, only tens of such experiments are performed per year. Hence, to deal with the scarcity of experimental data, it is natural to turn to computer simulation to understand the behavior of the radiating shocks in the experiment and to predict the results of new experiments. The fidelity of the simulation must, nevertheless, be validated with experimental data, and therefore, our interest lies in assimilating the information obtained from these simulator runs and field experiments. In the present literature, we often have similar situations where computer algorithms or numerical procedures are designed to describe or closely approximate a real-life experiment or physical process. Examples arise in diverse areas of science including cosmic mass distribution (Habib et al. 2007), heat transfer (Higdon et al. 2008b), volcano analysis (Bayarri et al. 2009), atmospheric sciences (Kennedy et al. 2008), and hydrodynamics (Williams et al. 2006). Accuracy of the experimental data can be quantified in some of these cases (as ours, see Section 2), but all inputs to the experimental system

Avishek Chakraborty is Postdoctoral Associate (E-mail: *avishekc@stat.tamu.edu*) and Bani Mallick is Distinguished Professor (E-mail: *bmallick@stat.tamu.edu*) in the Department of Statistics, Texas A&M University, College Station, TX 77843-3143. Ryan G. McClarren is Assistant Professor (E-mail: *rgm@tamu.edu*) and Hayes F. Stripling is Ph.D. candidate (E-mail: *h.stripling@tamu.edu*) in the Department of Nuclear Engineering, Texas A&M University, College Station, TX 77843-3133. Carolyn C. Kuranz is Assistant Research Scientist (E-mail: *ckuranz@umich.edu*), Michael J. Grosskopf is Research Engineer (E-mail: *mikegros@umich.edu*), Erica M. Rutter is a Research Technician (E-mail: *ruttere@umich.edu*), and R. Paul Drake is Henry Smith Carhart Professor of Space Science (E-mail: *rpdrake@umich.edu*) in the Department of Atmospheric, Oceanic and Space Sciences, University of Michigan, Ann Arbor, MI 48109-2143. Derek Bingham is Associate Professor in the Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, BC V5A 1S6, Canada (E-mail: *dbingham@stat.sfu.ca*). This work was funded by the Predictive Sciences Academic Alliances Program in DOE/NNSA-ASC via grant DEFC52- 08NA28616. The authors thank the editors and referees for their insightful feedback and for helping to improve the content and presentation of this article.

Journal of the American Statistical Association
June 2013, Vol. 108, No. 502, Applications and Case Studies
DOI: 10.1080/01621459.2013.770688

cannot be controlled or even measured. Thus, the main challenge in this regime is to account for the uncertainty due to the unknown state of the natural parameters that influence outcome of the experiment. However, for the simulator, all the factors that are believed to be influential in determining the response are provided as controlled input to the code. Also the mathematical model used for the simulation may only partially represent the true physical process, leading to discrepancy between experimental and simulator outputs. These simulators themselves often represent complex mathematical models and can be expensive to run in terms of cost and time. An *emulator* is a stochastic surrogate for the simulator; it is fast to implement and enables the user to predict a large number of code outputs as desired input configurations. The main task of the emulator is to learn the relationship between inputs and response from the code results and then to use it for calibration of the unknown parameters of the real-world process by matching it with corresponding responses. This enables both (i) running the simulator with input configuration that resembles the real-world parameters better and (ii) predicting the outcome of a real-life event with knowledge of only a subset of the inputs.

The basic statistical framework for validating computer models was developed in the recent literature. Sacks, Schiller, and Welch (1989) used the Gaussian process (GP) as the model deterministic computer codes to obtain uncertainty estimates at untried input configurations. Later, Kennedy and O'Hagan (2001) built a joint model (referred to henceforth as the Kennedy–O'Hagan model) for the experimental data and the code output. Their model uses a GP prior for the emulator and, in a hierarchical structure, efficiently assimilates the two sources of information accounting for their intrinsic differences. In a similar setup, Higdon et al. (2004) discussed uncertainty quantification and potential discrepancy between the physical model and the real-world system. Of late, extension of the Kennedy–O'Hagan model to multivariate (and possibly high dimensional) outputs has been developed in Higdon et al. (2008a). For other works in this field, covering a wide range of applications and extensions, see, for example, Bayarri et al. (2007), Liu and West (2009), Fricker, Oakley, and Urban (2010), Bastos and OHagan (2009), and references therein. While majority of the literature uses GP emulators, here we propose two alternative specifications based on splines. Splines are common in regression problems (Smith and Kohn 1998; Denison et al. 2002). In Section 3.2, we argue about their advantages over GP, and in Section 5, we provide diagnostic results from the data analysis in favor of them.

Another area of focus in this article is to construct an appropriate model for measurement error associated with the experimental output. The measurement error is universal to almost every data collection procedure. Early work with the measurement error in a generalized linear model can be found in Stefanski and Carroll (1987). Mallick and Gelfand (1995) developed a Bayesian approach for this problem relating the data, parameters, and unmeasured variables through a hierarchical structure. A Gaussian distribution is the usual choice for noise modeling. However, it may not be robust in every situation, specifically, in the examples where the error is known to have a skewed or flat-shaped structure. Non-Gaussian or skewed error specifications were used by Chen, Gott, and Ratra (2003), Arellano-Valle et al. (2005), and Rodrigues and Bolfarine (2007). In Section 4, we discuss two alternative ways to combine multiple sources of error, accounting for potential noninformativeness. Comparative performance analysis against the Gaussian error specification, based on experimental data, is provided in Section 5.

The article is organized as follows. Section 2 describes the experiment and the associated simulator in detail. In Section 3, within the generic formulation of the full hierarchical model, multiple alternative specifications are suggested, discussed and compared against each other. Section 4 extends this model with non-Gaussian error proposals for field experiments. In Section 5, we present an in-depth analysis of the shock experiment dataset. Finally, in Section 6, along with a summary of our work, additional possibilities for improvement, in the context of the current problem as well as the general modeling in this area, are outlined as topics of future research.

2. DETAILS OF SHOCK EXPERIMENT

Let us start with the motivation for studying shock features in physics. Subsequently we discuss the details of the laboratory experiment as well as the associated simulator, developed by the Center for Radiating Shock Hydrodynamics (CRASH; *http://aoss-research.engin.umich.edu/crash/*), funded by the Department of Energy Predictive Science Academic Alliance Program (PSAAP).

2.1 Scientific Relevance

Radiative shocks are of significant interest to researchers due to their connection to astrophysical shocks. In any shock wave, heating at the shock transition leads to heating and compression of the matter entering the shock. If the shock wave is strong enough, the heating becomes so large that radiation from the heated matter carries a significant amount of energy upstream (because light travels faster than the shock wave) and, as a result, alters the unshocked material and, therefore, the structure of the shock. Such shocks are called radiative shocks, and in these shocks, the transport of radiative energy plays an essential role in the dynamics. In astrophysics, systems with large velocities or very hot interiors are common, so radiative shocks are ubiquitous. Moreover, the shock waves emerging from supernovae and the shock waves that can be produced in laboratory experiments have specific similarities such as the optical depth (the ratio of the average distance the radiation travels before being absorbed to the length of the system) and the shock strength, which relates the shock speed to the thermodynamic properties of the material.

2.2 Experimental Details

In the laboratory experiment, first a laser pulse irradiates a thin disk of Be metal at the end of a tube of Xe, as in Figure 1. The energy of the laser causes the surface of the Be to ablate. To balance the momentum of the ablating material, a shock wave is driven into the Be at about 50 km/s. When this shock wave breaks out of the Be into the Xe, it is propagating at a speed of roughly 200 km/s and heats the Xe to temperatures above 5×10^5 K. At these high temperatures, the Xe becomes a plasma and emits a great deal of energy in the form of soft X-ray

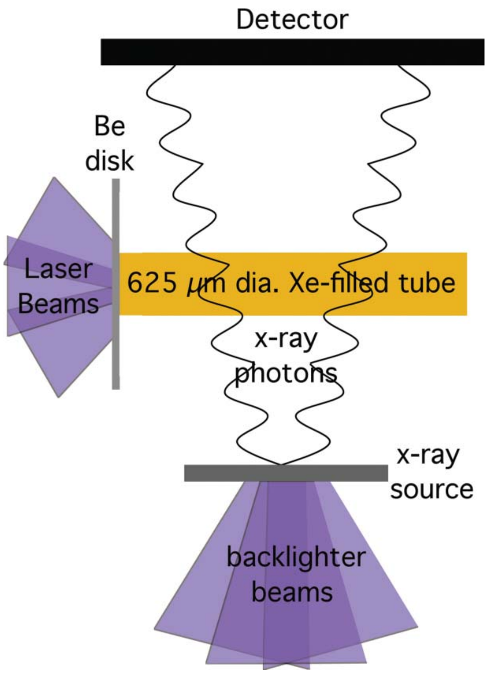


Figure 1. The Be disk is irradiated with several laser beams to drive a radiative shock in the tube of Xe gas. The main diagnostic for the radiative shock experiment is X-ray radiography. This technique is executed by additional laser beams irradiating a metal foil to create X-ray photons that pass through the target to an X-ray detector. The online version of this figure is in color.

radiation. This radiation travels ahead of the shock, heating the unshocked Xe and providing feedback to the shock dynamics.

In October 2008, a set of radiative shock experiments were conducted at the Omega laser. There were a number of controlling factors such as Be drive disk thickness, laser energy, Xe pressure, and observation time, which were varied from one experiment to another. Outcomes of interest were the shock features such as its position down the tube, time of breakout from the Be disk, and its speed. Measurement of these outcomes involved multiple sources of error. Specifically the shock breakout time, the feature of interest in this article, was measured using three different instruments of varying accuracy ranging from 10 to 30 picoseconds (ps). Two of them were velocity interferometer system for any reflector (VISAR; Barker and Hollenbach 1972) instruments that use a laser of 532 nanometer wavelength to probe a surface (see Figure 2) and detect the rate of change in the derivative of the optical path to a surface. These instruments are often used to measure the amount of time it takes for a shock to pass or break out of a material and are referred to as active shock breakout (ASBO) systems. They vary in sweep speed (3 and 5 nanoseconds, referred to as ASBO1 and ASBO2 respectively), which results in a different resolution on the images.

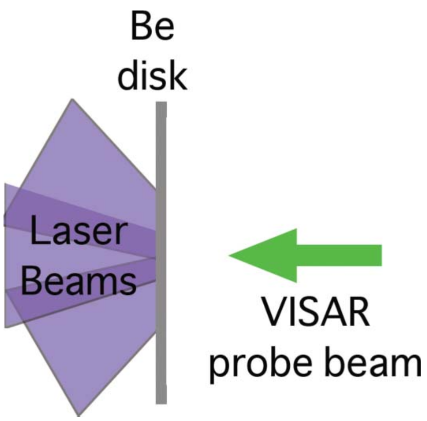


Figure 2. Shock breakout was measured using a VISAR to probe the rear surface of the Be disk. An SOP was also used. The online version of this figure is in color.

The third diagnostic used to measure the shock breakout time was a streaked optical pyrometer (SOP; Miller et al. 2007). An SOP is a passive detector that records thermal emission on a streak camera that results in a two-dimensional image showing the emission in space and time. Overall, they provide three independent measurements of the same experimental output. Though the VISAR and SOP diagnostics have their own uncertainties, the dominant uncertainty of ±50 ps was due to the systematic timing error for the laser firing mechanism, which was common to all the measurements. This uncertainty is derived from the time interval between the main laser and a diagnostic fiducial laser, which is used to calibrate the VISAR and SOP instruments.

These experiments are costly, and therefore, this project will only perform 1 or 2 experimental campaigns, each producing about a dozen experiments per year. Planning for these experiments begins about 6 months prior to the experimental day and involves multiple scientists, engineers, and technicians. Therefore, it is cost effective to perform a large number of simulations to complement the experimental dataset. One of the missions of CRASH is to design efficient computer algorithm for that purpose, as discussed in the following.

2.3 Simulation Algorithm and Design

Computer simulations are used to predict the several features of the shock, such as its position down the tube as a function of time and its speed. The main simulator uses the CRASH code, an adaptive mesh refinement Eulerian radiation hydrodynamics code that computes the shock behavior in the Xe. Because the CRASH code lacks a package to model laser deposition within the Be, it is not able to model the first nanosecond of the CRASH experiments, that is, the portion of the experiment where the laser is irradiating the target. This modeling is provided by a preprocessor, a Lagrangian radiation hydrodynamics code (HYADES; *http://www.casinc.com/hyades.html*), that computes the laser energy deposition and system evolution for the first 1.3 ns (the laser pulse width is 1 nanosecond at full-width half-maximum). Initial calculations from HYADES code provides matter and radiation fields to initialize the CRASH radiation hydrodynamics code.

In this article, we focus on building emulators for the validation of HYADES code. The original one-dimensional algorithm was extended to a two-dimensional version, known as H2D. The physics assumptions used in H2D are similar to those of one-dimensional HYADES, the major difference being the numerical implementation of these physics models. The fundamental quantity of interest is the time required for the ablation-induced shock to travel through a disk of Be (shock breakout time). We consider one set of H2D output for applying our model. In a previous analysis with one-dimensional output, McClarren et al.

(2011) found only five of the inputs to have significant effect on the response: Be disk thickness, laser energy, electron flux limiter (related to heat transport by the electrons), Be gamma constant (related to the compressibility of Be), and the computational mesh resolution in the Be. Later, the mesh resolution was fixed keeping in mind code stability and convergence and was used for all subsequent runs of HYADES. Along with the previous four parameters, an additional source of uncertainty specifically for the two-dimensional solution was the wall opacity of the plastic tube, which controls how strongly radiation is absorbed.

To construct a fast running preprocessor, a set of 104 runs was conducted to cover this five-dimensional input space. The choice of these input settings (i.e., the experimental design) was a combination of a smaller space filling Latin hypercube design, with additional points added according to a space-filling criterion. More specifically, the first 64 design points, or runs, were chosen using an orthogonal array-based Latin hypercube design (LHD) Tang (1993). The base orthogonal array was a replicated 2^5 factorial design, and these points converted to a Latin hypercube. The orthogonal array-based design guarantees that the cells (or strata) defined by the factorial design contain the same number of points and the space filling criterion aims to fill the input space as uniformly as possible. Since the construction of such designs involves random permutations, there are many possible designs that can be constructed in this manner. So, to identify a very good one, many designs were generated and the one that optimized the so-called maximin space filling criterion [maximizing the minimum distance between design points, see Johnson, Moore, and Ylvisaker (1990)] was chosen. Next, a batch of 10 new design points was added to the existing 64 run design. The points were allocated so that the maximin criterion was again optimized, conditional on the first set of the points. The procedure of adding 10 points at a time was repeated until a total of 104 design points were identified. The design was constructed in this way since it was not clear from the start whether it would be possible to complete 104 runs with the available financial resources. Therefore, if this limit got exceeded in any intermediate step between 64 and 104 trials, the resulting design would still have very good space filling properties. The analysis of this dataset is presented in Section 5.

3. HIERARCHICAL MODEL

We start with a description of the basic framework for a computer code validation problem, based on the specification in Kennedy and O'Hagan (2001). Under this setup, we build a multistage model for joint analysis of the experimental and simulation outputs. A variety of competing model choices are proposed for the emulator. We discuss their properties, limitations, and mutual differences. Finally, we provide details of estimation procedure.

3.1 Generic Hierarchical Specification

Suppose we have obtained data from n runs of the simulator and m trials of the experiment. Denote by $y_i^{(r)}, y_i^{(c)}$ the ith response from the experiment and the simulator, respectively; $x_i^{(r)}, x_i^{(c)}$ denote corresponding p-dimensional inputs known under both scenarios; $\theta^{(r)}$ denotes the q-dimensional vector of unobserved natural parameters that is believed to have an influence on the response. The simulator is run at n different known combinations $\theta_1^{(c)}, \theta_2^{(c)}, \ldots, \theta_n^{(c)}$ of these parameters. Usually, the combinations are chosen using a prior knowledge about range and likely values of the features.

We introduce the multistage model in an incremental manner. Whenever possible, we omit the subscripts indexing the data points. In the first stage, we specify a stochastic model $\mathcal{M}_C$ for the simulator as

$$\mathcal{M}_C: \ \ y^{(c)} = f(x^{(c)}, \theta^{(c)}) + \eta_c(x^{(c)}, \theta^{(c)}). \tag{1}$$

Next, $\mathcal{M}_C$ is hierarchically related to the model for the experiments $\mathcal{M}_R$ as

$$\mathcal{M}_R: \ \ y^{(r)} = f(x^{(r)}, \theta^{(r)}) + \delta_r(x^{(r)}). \tag{2}$$

In the above equation, f is the *link* function between the code and the experiment, which captures the major patterns common to both of them; η_c and δ_r represent their individual residual patterns. In practice, the experiment and the simulation have their own biases and uncertainties. Physical models often involve complex optimizations or set of equations that do not admit closed-form solutions, but require use of numerical algorithms. Irrespective of how the observations are simulated, any such code output is likely to have an associated uncertainty from multiple perspectives such as sensitivity to the choice of initial value, simplifying assumptions, prespecified criterion for convergence, etc. Similarly, the input–output relationship in an actual event can deviate from the theoretical prediction due to many reasons such as input uncertainty, lack of knowledge about possible factors (other than θ) influencing the experimental outcome, and, more importantly, partial accuracy of the mathematical model. Hence, we account for them with the inclusion of η_c and δ_r in $\mathcal{M}_C$ and $\mathcal{M}_R$, respectively. The focus of this article is to propose and compare different model specifications for f. We defer that entirely to Section 3.2.

The above interpretation relies on the specification of f and δ_r. From the modeling perspective, (1) and (2) represent two regression problems. A priori, we believe that the solution of the mathematical model is a good approximation of the experimental outcome, that is, the two response functions are expected to match. Accordingly, f can be seen as the shared *mean* surface and δ_r and η_c as zero-mean residual processes for the experiment and the simulator, respectively. As usual for any regression problem, the properties of the residuals depend on the specification of mean, for example, using a linear mean for a quadratic response function can produce heteroscedastic residuals, whereas choosing a second-order mean may lead to homoscedastic pattern. Hence, if f is constrained to have a fixed functional form or is chosen from a fixed family of functions, interpretation of the estimates of the residuals (η_c and δ_r) is subject to that selection. However, for the purpose of this problem, learning of the individual functions is not the primary goal of inference; the objective is to improve predictions from these models. The functions f, η_c, and δ_r serve as tools for that. Using the stochastic model, we want to predict the outcome, with reasonable accuracy, at an untried input configuration without the need to perform complex simulations (and even more expensive experiments) in the long run. Hence, we want f and δ_r to be flexible enough so that, through $\mathcal{M}_C$ and $\mathcal{M}_R$, we can efficiently approximate the

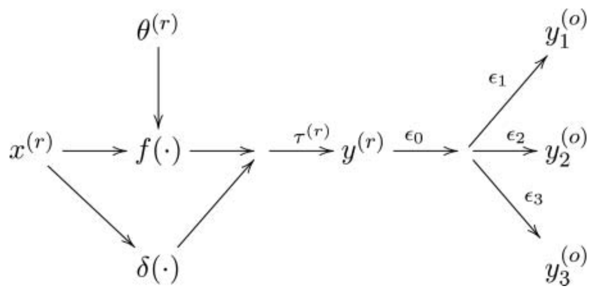


Figure 3. Graphical model for the experimental system: the controlled inputs $x^{(r)}$ and natural parameters $\theta^{(r)}$ influence the emulator f. Generation of the actual outcome $y^{(r)}$ also involves additional bias through input-dependent bias $\delta(\cdot)$ and pure error $\tau^{(r)}$. Three devices, all sharing a common bias ϵ_0 but different device-specific biases ϵ_j, are used to produce measurements $y_j^{(o)}$, $j = 1, 2, 3$, separately, of the same experimental outcome.

response functions of the simulator and the experimental system, respectively. Reasonable and well-defined specifications for f are subject to user judgment and practical considerations. As a simple example, we may decide to include polynomials only upto a certain order in the emulator specification. When $\eta_c \equiv 0$, we get back the usual Kennedy–O'Hagan specification with f as an interpolator for the computer output and δ_r as discrepancy capturing the potential inadequacy of the mathematical model to explain the dynamics of an actual experiment.

We return to the model. For modeling the residual process δ_r in $\mathcal{M}_R$, we decompose it into a zero-mean covariate-dependent random effect plus a nugget, similar to a spatial regression model (Banerjee, Carlin, and Gelfand 2004):

$$\delta_r\left(x^{(r)}\right) = \delta\left(x^{(r)}\right) + \tau^{(r)}.$$

In the literature, $\delta(\cdot)$ is modeled using a GP on $x^{(r)}$ (Higdon et al. 2004); $\tau^{(r)}$ accounts for variation in repeated experiments with identical $x^{(r)}$ due to the factors that were unobserved and unaccounted for in the model.

Although $y^{(r)}$ is the actual outcome of the real-world experiment, usually it is measured with some error. Specifically, for the current dataset, three different procedures were used to measure the output of each experiment with different degrees of accuracy (known a priori). There was also a common measurement error ϵ_0 of known scale, see Section 2.2. Hence, we augment a third stage to the hierarchy: the measurement error model $\mathcal{M}_E$ for the observed output $y^{(o)}$ conditional on $y^{(r)}$ as follows:

$$\mathcal{M}_E : y_j^{(o)} = y^{(r)} + \epsilon_j + \epsilon_0\,; \qquad j = 1, 2, 3, \tag{3}$$

where ϵ_j is the error specific to measurement procedure j. $\mathcal{M}_R$ and $\mathcal{M}_E$ can be simultaneously represented through Figure 3.

3.2 Choice of Functions With Interpretation

In scientific experiments, it is often desirable to use a specification for $\mathcal{M}_C$, that is, an interpolator. For code outputs of deterministic nature (i.e., different simulator runs with same input combination produce the exact same output), one expects the stochastic model to exactly fit the observed responses at the corresponding input levels. GP is an interpolator, but using any other residual distribution inside $\mathcal{M}_C$ (such as white noise (WN) or heteroscedastic but independent residuals) violates that condition.

By definition, an emulator is a stochastic substitute of the simulator and should be easy to run than the latter. Thus, it should be computationally robust with respect to the size of simulation dataset. For the shock experiment dataset, the H2D simulator is less expensive to run than the corresponding field experiment, hence over time it is possible to conduct more and more trials. In general, regression with independent errors are computationally convenient than GP regression where the inversion of covariance matrix gets increasingly complicated as n increases.

Next, we mention some possible choices for f and compare them on the basis of above criteria:

- GP emulator: The original Kennedy–O'Hagan specification Kennedy and O'Hagan (2001) used a GP prior for the emulator f. GP is by now common in nonparametric Bayes literature as prior for regression functions (Neal 1999; Shi and Choi 2011) and in geostatistics to model the response correlated over space (Banerjee, Carlin, and Gelfand 2004). The model is specified as

$$f(\cdot) \sim \mathrm{GP}(\mu(\cdot), C(\cdot,\cdot)), \quad \eta_c(\cdot) = 0. \tag{4}$$

The mean function $\mu(\cdot)$ can be chosen as a polynomial in (x, θ) and the covariance function $C(\cdot,\cdot)$ is often specified to have a stationary correlation structure along with a constant scale, that is, if $a = (x_1, \theta_1)$, $b = (x_2, \theta_2)$ are two inputs to the simulator, then

$$C_{a,b}(\sigma^2, \nu) = \sigma^2 \prod_{s=1}^{p+q} \kappa(|a_s - b_s|, \nu_s),$$

where κ is a valid correlation function on $\mathbb{R}$ and ν_s is the parameter specific to sth input. Together, forms of κ and $\{\nu_s\}$ control the correlation (hence smoothness) in the emulator surface over the input space. Choice for κ is welldiscussed in literature, the most popular being the Matérn family of functions Matérn (1960), for example, exponential and Gaussian correlation functions.

The specification in (4) makes f as well as $\mathcal{M}_C$ interpolators for the simulator outputs. The computation for the parameters of f are fairly standard, *but* with a large number of sample points, inversion of the sample covariance matrix gets difficult. Several approximation methods are available in that case including process convolution (Higdon 2002), approximate likelihood (Stein, Chi, and Welty 2004), fixed rank kriging (Cressie and Johannesson 2008), covariance tapering (Kaufman, Schervish, and Nychka 2008), predictive process (Banerjee et al. 2008), and compactly supported correlation functions (Kaufman et al. 2011).
- A spline-based alternative: As discussed earlier, we want f to be robust to the possibility of overfitting the simulation dataset. The emulator f should have a strong predictive property for experiments with new sets of inputs. In that regard, using a model that fits exactly the code samples may not be desirable. The complexity of the physics as well as numerical approximation to solve the computer model makes it unlikely to develop codes that can fully simulate the experiments. Also computational efficiency

in estimating parameters of f is desirable. Specifying f as sum of local interactions of varying order using a complete class of basis functions serves these needs and provides greater flexibility to model relationships among variables. A number of options are available, including multivariate adaptive regression splines (MARS; Friedman 1991; Denison, Mallick, and Smith 1998) as

$$f(x) = \sum_{h=1}^{k} \beta_h \phi_h(x), \quad \eta_c(x) \text{ ind noise},$$
$$\phi_1(x) = 1;\ \phi_h(x) = \prod_{l=1}^{n_h} \left[u_{hl}\left(x_{v_{hl}} - t_{hl}\right)\right]_+;\ h > 1, \quad (5)$$

where $(\cdot)_+ = \max(\cdot, 0)$ and n_h is the degree of the interaction of basis function ϕ_h. The sign indicators $\{u_{hl}\}$ are ± 1, v_{hl} gives the index of the predictor variable that is being split at the *knot* point t_{hl} within its range. The set $\{\phi_h(\cdot)\}_h$ defines an adaptive partitioning of the multidimensional predictor space; $\{\beta_h\}$ represents the vector of weights associated with the functions. The response is modeled to have a mean f and a WN term η_c, which averages out any additional pattern present in the data that is not accounted for by the components of f. For the rest of this article, we refer to this model as MARS+WN.

It is evident from the above specification that how well f can capture the local patterns present in the response depends on two factors: (i) the number and positions of the knots and (ii) class of polynomials associated with each knot. The larger number of knots implies a finer partition of the predictor space so that we have greater flexibility in f at the expense of reduced smoothness. Also, if we allow polynomials and interactions of all possible order, that constitutes a complete class of functions. Consequently, the likelihood of observed data increases as more and more of this type of local polynomials enter the model. That reduces the estimation bias but increases predictive uncertainty. Hence, to preserve the flexibility of the model allowing for adaptive selection of its parameters, we need restrictions that discourage such overfitting.

For choosing the number of knots in a penalized spline regression, a generalized cross-validation technique is presented in Ruppert (2002). The recent work of Kauermann and Opsomer (2011) uses a maximum-likelihood-based algorithm for this. In both of them, penalty is introduced in the form of a smoothing parameter attached to the weights. In our hierarchical approach, this is equivalent to using a precision parameter in the prior for $\{\beta_h\}$. Additionally, instead of determining the number of knots from an optimization algorithm, we adopt a fully model-based approach. Total number of knots in f is given by $\sum_{h=2}^{k} n_h$.

First, we restrict the class of polynomials that can appear in ϕ_h, that is, we may set $n_h = 1$ to include only piecewise linear terms and *no* interaction among variables at all; $n_h \leq 2$ ensures that interactions (of only first order in each participating variable) are allowed. Higher the allowed threshold for n_h is, smoother are the basis functions $\{\phi_h\}$. One may like to use a prior on the class of functions, so that functions of higher order (large n_h) are less likely to be chosen. Also important is to penalize large values of the number of basis functions k, which, in combination with $\{n_h\}$, discourages presence of *too many* knots inside f. We use a Poisson prior distribution for k to assign low probabilities to its high values. A more stringent form of penalty is enforced by letting the prior to be truncated at right at some prespecified value k_0, which essentially serves as the maximum number of basis functions that can be present inside f at a time. Friedman (1991) described an alternative approach for limiting the number of knots by constraining them to be apart by at least a certain distance (in terms of number of data points between them). For computer model-based validation problems, sometimes we can only have a limited number of simulator runs available, that is, n is small. There, it may be appropriate to opt for a sparse structure in f—using a simpler class of polynomials as well as a smaller k_0 to truncate the number of parameters entering the model. Treating the knots as parameters allows the model to locally adjust the smoothness of the fitted function as required, further encouraging sparsity without compromising the fit.

The above specification has an advantage over the traditional GP emulator in (4) in several aspects. First, the emulator is modeled as a sum of localized interactions allowing for nonstationarity. Kennedy and O'Hagan (2001) also mentioned the need to consider such local structures in the specification of emulator, which cannot be achieved with a stationary GP. One can attempt to use kernel-weighted mixture of stationary GPs, but computational efficiency of those type of models is still arguable. The flexibility of MARS lies in the fact that the interaction functions can be constructed adaptively, that is, the order of interaction, knot locations, signs, and even the number of such terms is decided by the pattern of the data during model-fitting, eliminating the need for any prior ad hoc or empirical judgment. In spite of having a flexible mean structure, MARS is easy to fit. For GP, even when stationarity holds, the covariance specification makes it computationally inconvenient for handling large number of simulator runs. The computation gets more complicated if one tries to introduce nonstationarity within covariance structure. However, unlike GP, MARS with independent noise does not have the interpolating property. If that is a constraint, we suggest modifying (5) as follows.

- A combined approach: If we want to retain the flexibility of f from (5) while enforcing the interpolating property on $\mathcal{M}_C$, we can combine the earlier specifications as

$$f(\cdot) = \sum_h \beta_h \phi_h(\cdot) + \mathrm{GP}(\mathbf{0}, C(\cdot,\cdot)), \quad \eta_c(\cdot) = 0, \quad (6)$$

where $\phi_h(\cdot)$ is from (5). In the subsequent discussion and data analysis, we refer to the above model as MARS+GP. Essentially, we are modeling the simulator output as a sum of two parts—a combination of local patterns of up to a certain order plus a residual pattern with a global stationary correlation structure. This makes $\mathcal{M}_C$ an interpolator. Choosing $\mu(\cdot) = \sum_i \beta_i \phi_i(\cdot)$ makes (4) and (6) identical. Hence, with an increase in simulator runs, (6) gets computationally as difficult to fit as (4) due to the problem with associated GP, discussed earlier. In Section 5.1, we outline a possible modification in the form of correlation function C to handle those situations more efficiently.

3.3 Estimation and Inference From the Joint Model

Now, we discuss how to make inference from the hierarchical model specified in Section 3.1. For the most exhaustive sampling scheme, below we choose MARS+GP as the specification for $\mathcal{M}_C$. With either of GP or MARS+WN as the choice, the corresponding steps can be reworked following this.

The model-fitting procedure used here is a Markov chain Monte Carlo (MCMC) scheme. The available data are $(y_l^{(c)}, x_l^{(c)}, \theta_l^{(c)})$, $l = 1, 2, \ldots, n$ and $(y_{ij}^{(r)}, x_{ij}^{(r)})$, $j = 1, 2, 3$, $i = 1, 2, \ldots, m$. The set of parameters consists of the ones appearing in the distributions for f, η_c, δ, and $\tau^{(r)}$ as well as $\theta^{(r)}$. We provide the full model specification below.

$$\begin{aligned}
y_{ij}^{(o)} &= y_i^{(r)} + \epsilon_{i0} + \epsilon_{ij}; \qquad\qquad j = 1, 2, 3,\\
y_i^{(r)} &= f\big(x_i^{(r)}, \theta^{(r)}\big) + \delta\big(x_i^{(r)}\big) + \tau_i^{(r)}; \;\; i = 1, 2, \ldots, m,\\
y_l^{(c)} &= f\big(x_l^{(c)}, \theta_l^{(c)}\big) + \eta_c\big(x_l^{(c)}, \theta_l^{(c)}\big), \;\; l = 1, 2, \ldots, n,\\
f(x, \theta) &= \sum_{h=1}^{k} \beta_h \phi_h(x, \theta) + \mathrm{GP}(\mathbf{0}, C(\sigma^2, \nu))\\
\delta\left(x_{1:m}^{(r)}\right) &\sim \mathrm{MVN}_m\big(\mathbf{0}_m, C\big(\sigma_\delta^2, \nu_\delta\big)\big), \;\; \tau_i^{(r)} \overset{\text{ind}}{\sim} \mathrm{N}(0, \tau^2),\\
\epsilon_{i0} &\overset{\text{ind}}{\sim} \mathrm{N}\big(0, \sigma_0^2\big); \quad \epsilon_{ij} \overset{\text{ind}}{\sim} \mathrm{N}\big(0, \sigma_j^2\big),\\
\pi(\beta, k) &\propto \mathrm{MVN}_k\big(\mathbf{0}, \sigma_\beta^2 \mathcal{I}_k\big) \; \times \; \mathrm{Pois}(k - 1|\lambda) I(k \le k_0). \quad (7)
\end{aligned}$$

In the above equation, we use the customary additive Gaussian structure to combine both sources of measurement error. Alternative choices of noise distributions are discussed in Section 4. We note that, unless informative prior distributions are chosen for the scales of ϵ_j, ϵ_0, scale of $\tau^{(r)}$ is *not* identifiable (evident if we marginalize out $y^{(r)}$). Since, for the shock experiment data, the scale of accuracy of all the measuring devices are known beforehand, $\{\sigma_j^2 : j = 0, \ldots, 3\}$ are fixed a priori in (7). In more general applications, one *must* use informative priors for each of them. Such priors can be constructed from the knowledge of the procedure as well as other studies where they have been used. As far as choice of priors is concerned, for spatial correlation in η_c and δ, we used exponential correlation functions. Corresponding spatial decay parameters (ν, ν_δ) are chosen uniformly from the interval that corresponds to a reasonable spatial range in the input space. Following the discussion on controlling the possibility of overfitting in MARS, $(k - 1)$ is given a Poisson (λ) prior truncated below k_0, depending on the maximum number of nonconstant basis functions we want to allow in the emulator specification. Finally, for the calibration parameter $\theta^{(r)}$, scientists often have a priori knowledge about the range of values it might have. Such knowledge may be derived from physical understanding of the parameter or any previous study involving $\theta^{(r)}$ or any small-scale survey conducted with the purpose of eliciting any prior information on its likely values. In fact, while running the simulator, the input configurations $\theta_l^{(c)}$, $l = 1, 2, \ldots, n$ are determined so as to imitate that learning as close as possible. For the shock experiment data, scientists used informative guess only about the range of each of the parameter and within that range equi-spaced values were used as input for the H2D simulator. We find it sensible to quantify that information with a uniform prior distribution for each component of $\theta^{(r)}$, restricted to the range of attempted configurations.

The model in (7) consists of latent vectors $y^{(r)}$, δ, and η_c. Marginalizing over any one or more of them produces different sets of conditional posterior distributions. In the following, we present a specific sampling scheme from model (7). We define $\bar{y}_i^{(o)} = \frac{1}{3}\sum_{j=1}^{3} y_{ij}^{(o)}$ and $\bar{\epsilon}_i = \epsilon_{i0} + \frac{1}{3}\sum_{j=1}^{3} \epsilon_{ij}$. Also reparametrize $\sigma_\beta^2 = \sigma^2\sigma_\beta^2$ and $\tau^2 = \sigma^2\tau^2$ for computational convenience to be utilized in the sampling. We use independent inverse gamma priors for σ^2, τ^2, σ_β^2, and σ_δ^2. Now, let

$$\begin{aligned}
\mathbf{y}_f &= \begin{bmatrix} \bar{y}_{1:m}^{(o)} \\ y_{1:n}^{(c)} \end{bmatrix}, \qquad \mathbf{x}_f = \begin{bmatrix} x_{1:m}^{(r)} & \theta^{(r)T} \otimes \mathbf{1}_m \\ x_{1:n}^{(c)} & \theta_{1:n}^{(c)} \end{bmatrix},\\
D &= C_{m+n}(1, \nu) + \begin{bmatrix} \tau^2 \mathbf{I}_m & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix},\\
D_\delta &= \left[\sigma_0^2 + \frac{1}{9}\sum_{j=1}^{3} \sigma_j^2\right]\mathbf{I}_m + C_m\big(\sigma_\delta^2, \nu_\delta\big),
\end{aligned}$$

and

$$P = [\phi_1[\mathbf{x}_f],\ \phi_2[\mathbf{x}_f],\ \ldots\ ,\ \phi_k[\mathbf{x}_f]],$$

where, for a matrix A, $\phi_h[A]$ denotes the vector obtained by applying ϕ_h on each row of A. $C_d(a, b)$ stands for a d-dimensional stationary GP covariance matrix with variance a and correlation parameter(s) b. Using these, the model from (7) is rewritten as

$$\begin{aligned}
\mathbf{y}_f &= P\beta + \begin{bmatrix} z_{1:m} \\ \mathbf{0}_n \end{bmatrix} + \mathrm{MVN}_{m+n}(\mathbf{0}, \sigma^2 D),\\
\mathbf{z} &\sim \mathrm{MVN}_m(\mathbf{0}, D_\delta). \qquad\qquad (8)
\end{aligned}$$

During the MCMC, the vector of parameters to be updated can be classified into broad categories as (i) MARS weights $\{\beta_h\}$ and parameters of the spline: k, $\{n_h, \mathbf{u}_h, \mathbf{v}_h, \mathbf{t}_h\}$ as in (5); (ii) calibration parameters $\theta^{(r)}$; (iii) the m-dimensional latent vector $\mathbf{z}$; (iv) other parameters σ^2, ν, σ_δ^2, ν_δ, τ^2, σ_β^2, and λ. We outline the sampling steps below.

At any particular iteration of MCMC, the updating distributions are as follows:

(a) MARS parameters: With k basis functions, let $\alpha_k = \{(n_h, \mathbf{u}_h, \mathbf{v}_h, \mathbf{t}_h) : h = 1, 2, \ldots k\}$ be the corresponding set of spline parameters; n_h and $\mathbf{v}_h$ control the type of basis function, whereas $\mathbf{u}_h$ and $\mathbf{t}_h$ determines the signs and knot points, respectively. We update (k, α_k) jointly using a reversible jump MCMC (RJMCMC; Richardson and Green 1997) scheme. First, we marginalize out β and σ^2 from the distribution of $\mathbf{y}_f$ as in Appendix A. Now, using a suitable proposal distribution q, propose a dimension changing move $(k, \alpha_k) \to (k', \alpha_{k'})$. We consider three types of possible moves (i) *birth*: addition of a basis function, (ii) *death*: deletion of an existing basis function, and (iii) *change*: modification of an existing basis function. Thus $k' \in \{k - 1, k, k + 1\}$. The acceptance ratio for such a move is given by

$$p_{k\to k'} = \min\left\{1,\ \frac{p(\mathbf{y}_f|k', \alpha_{k'}, \ldots)}{p(\mathbf{y}_f|k, \alpha_k, \ldots)}\, \frac{p(\alpha_{k'}|k')p(k')}{p(\alpha_k|k)p(k)} \times \frac{q((k', \alpha_{k'}) \to (k, \alpha_k))}{q((k, \alpha_k) \to (k', \alpha_{k'}))}\right\}.$$

The details of the priors and proposal distributions for the three different types of move are described in Appendix B. Set $k = k', \alpha_k = \alpha_{k'}$ if the move is accepted, leave unchanged otherwise. Subsequently, β can be updated using the k-variate t distribution with degrees of freedom $d = n + m + 2a_\sigma$, mean μ_k, and dispersion $\frac{c_{0k}\Sigma_k}{d}$. (These quantities are defined in Appendix A.)

(b) $\mathbf{z}$: Marginalize β in the first line of (8) and subsequently condition the experimental responses on the code output to get

$$\bar{y}^{(o)}_{1:m}|y^{(c)}_{1:n} = z_{1:m} + \tilde{D}_{m,n}\tilde{D}^{-1}_{n,n}y^{(c)}_{1:n} + \mathrm{MVN}_m(\mathbf{0}, \sigma^2\tilde{D}_{m|n}),$$

where $\tilde{D} = D + \sigma^2_\beta PP^T$ is partitioned into blocks as

$$\tilde{D} = \begin{bmatrix} \tilde{D}_{m,m} & \tilde{D}_{m,n} \\ \tilde{D}_{n,m} & \tilde{D}_{n,n} \end{bmatrix}$$

and $\tilde{D}_{m|n} = \tilde{D}_{m,m} - \tilde{D}_{m,n}\tilde{D}^{-1}_{n,n}\tilde{D}_{n,m}$. It follows that the posterior of $\mathbf{z}$ is $\mathrm{MVN}_m(\mu_z, \Sigma_z)$, where $\Sigma_z^{-1} = \frac{1}{\sigma^2}\tilde{D}^{-1}_{m|n} + D_\delta^{-1}$ and $\Sigma_z^{-1}\mu_z = \frac{1}{\sigma^2}\tilde{D}^{-1}_{m|n}\{\bar{y}^{(o)}_{1:m} - \tilde{D}_{m,n}\tilde{D}^{-1}_{n,n}y^{(c)}_{1:n}\}$. When the measurement error distributions are non-Gaussian, $\mathbf{z}$ do not have a standard posterior any more and one needs to use Metropolis–Hastings (MH) methods. In Section 4, we discuss a modified sampling scheme for $\mathbf{z}$ specific to the non-Gaussian measurement error distributions proposed there.

(c) $\theta^{(r)}$: Construct a prior $\tilde{\pi}$ for the components of $\theta^{(r)}$ based on their ranges, likely values, and mutual dependence. Knowledge of these quantities can be obtained from physical reasoning as well as prior studies, if available. The posterior distribution of $\theta^{(r)}$ is q-dimensional and nonstandard, necessitating an MH step. We use conditional distributions one at a time, that is, update $\theta^{(r)}_i$ given current states of $\theta^{(r)}_{-i}$. If $\tilde{\pi}_i$ denotes the ith full conditional corresponding to $\tilde{\pi}$, then the corresponding posterior density at $\theta^{(r)}_i = \theta_{0,i}$ is given by

$$\theta_{0,i}\left|\theta^{(r)}_{-i} \sim \tilde{\pi}_i\left(\theta_{0,i}\left|\theta^{(r)}_{-i}\right.\right) c_{0k}^{-\frac{d}{2}}\ |\Sigma_k|^{1/2},\right.$$

where we replace $\theta^{(r)}_i$ with $\theta_{0,i}$ inside $\mathbf{x}_f$. Subsequently, $\theta^{(r)}_i$ can be sampled using a random walk MH step. As a special case, when $\tilde{\pi}_i$ is discrete, the above conditional posterior distribution essentially becomes multinomial and is easy to sample from.

(d) Other parameters: With $\nu \sim \pi_\nu$ and $\sigma^2 \sim$ inverse gamma(a_σ, b_σ) a priori, it follows that $\pi(\nu|\ldots) \propto \pi_\nu(\nu)|D|^{-1/2}\, c_{0k}^{-\frac{d}{2}}\ |\Sigma_k|^{1/2}$ and

$$\pi(\sigma^2|\nu, \beta, \ldots) = \text{inverse gamma} \left(\frac{n+m+k}{2} + a_\sigma,\ \frac{S^TD^{-1}S + \beta^T\beta/\sigma^2_\beta}{2} + b_\sigma\right),$$

where

$$S = \mathbf{y}_f - \sum_h^k \beta_h\phi_h[\mathbf{x}_f] - \begin{bmatrix} z_{1:m} \\ \mathbf{0}_n \end{bmatrix}.$$

However, if the number of code output is much more than the number of experimental observations (i.e., $n \gg m$), we recommend expediting the MCMC computation by using an estimate of ν based on $\mathcal{M}_C$ alone. Kennedy and O'Hagan (2001, sec. 4.5) argues that this approximation is reasonable, since it only neglects the "second-order" uncertainty about the hyper parameters. The posterior distributions of τ^2 and σ^2_β are also inverse gamma. The Poisson parameter λ is provided with a gamma prior. An MH sampler is used to draw from its posterior that has a gamma form multiplied with the normalizing constant of the truncated Poisson distribution of $(k-1)$. Parameters present in the prior distribution of z are σ^2_δ, ν_δ, and, if the measurement uncertainties are not exactly known, σ^2_0 and $\{\sigma^2_j : j = 1, 2, 3\}$. They can be updated in an MH step using the multivariate normal density of $\mathbf{z}$ although, to ensure identifiability as discussed before, informative priors need to be used.

4. NON-GAUSSIAN MEASUREMENT ERROR MODELS

The measurement error model $\mathcal{M}_E$ in (3) combines multiple sources of error with an additive Gaussian structure. Although this is computationally convenient, there may be instances where one or more of the error patterns lacks Gaussianity (or, more generally, exponential decay pattern). As described in Section 3.1, the particular experiment analyzed in this article has two types of measurement error, the common first stage error ϵ_0 and then the group-specific error ϵ_j, $1 \le j \le 3$. Practical knowledge of the measurement procedure suggests that the group-specific errors indeed have exponentially decaying patterns *but* ϵ_0 has a noninformative structure. More explicitly, one can only detect the response up to $\pm\alpha$ accuracy and any further assumption on where exactly the true value should be within that 2α range lacks justification. The overall measurement error is a combination of these two types. Here we suggest two different approaches of constructing the model under this scenario.

First, we maintain the additive structure of (3) and introduce non-Gaussian distributions. From the discussion above, it only seems reasonable to suggest $\epsilon_0 \sim \mathrm{U}[-\alpha, \alpha]$, where U is the uniform distribution. For ϵ_j, we prefer replacing the Gaussian distribution with a Laplace (double-exponential) density of rate ρ_j. Although both have exponentially decaying tail behavior, (i) the later produces simpler closed-form analytic expression for the distribution of the total error $\epsilon_j + \epsilon_0$ than the former and, more importantly, (ii) this choice enables us to relate the current approach to the next one, described below. Thus for the jth type of measurement procedure, we have

$$f\left(y_j^{(o)} \middle| y^{(r)}; \rho_j, \alpha\right) \propto \begin{cases} 2 - e^{-\rho_j\left(y_j^{(o)} - y^{(r)} + \alpha\right)} - e^{-\rho_j\left(y^{(r)} - y_j^{(o)} + \alpha\right)} & \left|y_j^{(o)} - y^{(r)}\right| < \alpha, \\ e^{-\rho_j\left|y_j^{(o)} - y^{(r)}\right|} & \left|y_j^{(o)} - y^{(r)}\right| > \alpha. \end{cases} \tag{9}$$

From a different perspective, we want to build the measurement error model starting with a loss function that is intuitively appropriate for the above situation. The usual Gaussian error follows from the squared error loss. Here, we try to motivate the following approach. It is evident from above that when the true and the observed responses are within a distance α of each other, there is no further information on the error, thus we should have uniform loss everywhere on $[-\alpha, \alpha]$. Once they differ by more than α, we use the absolute error loss accounting for the information from the second type of measurement error, which has a decaying pattern. The resulting loss function, introduced by Vapnik (1995) has the form

$$L_\alpha\left(y_j^{(o)}, y^{(r)}\right) = \begin{cases} c & \left|y_j^{(o)} - y^{(r)}\right| < \alpha, \\ c + \left(\left|y_j^{(o)} - y^{(r)}\right| - \alpha\right) & \left|y_j^{(o)} - y^{(r)}\right| > \alpha \end{cases} \tag{10}$$

for some nonnegative c. We model $[y_j^{(o)} | y^{(r)}]$ on the basis of this loss function so that the likelihood of observing $y_j^{(o)}$ increases if $L_\alpha(y_j^{(o)}, y^{(r)})$ decreases. One way of preserving this duality between likelihood and loss is to view the loss as the negative of the log-likelihood as follows:

$$f\left(y_j^{(o)} \middle| y^{(r)}\right) \propto \exp\left(-\rho_j L_\alpha\left(y_j^{(o)}, y^{(r)}\right)\right). \tag{11}$$

This transformation of loss into likelihood is referred to in the Bayesian literature as the "logarithmic scoring rule" Bernardo (1979). Clearly, f is independent of choice for c, and one can assume $c = 0$ in the sense that any value of the observed response within α of the truth is equally "good." This error distribution can be rewritten as

$$f\left(y_j^{(o)} \middle| y^{(r)}; \rho_j, \alpha\right) = p_j \operatorname{Laplace}\left(\rho_j, y^{(r)}, \alpha\right) + (1 - p_j)\operatorname{Unif}\left(y^{(r)} - \alpha, y^{(r)} + \alpha\right), \tag{12}$$

where $\operatorname{Laplace}(\rho_j, y^{(r)}, \alpha)$ represents the Laplace distribution with decay rate ρ_j and location parameter $y^{(r)}$ truncated within $[-\alpha, \alpha]^C$. For derivation of this form as well as expression for p_j, see Appendix C. Use of this loss function effectively diffuses the information from errors of small magnitude. Equations (9) and (12) show both methods combine the same pair of distributions in two different ways. Depending on the knowledge of a particular problem, one may be more appropriate to use than the other.

Choice of either of these error distributions leads to small change in the sampling scheme described in Section 3.3, for $z_{1:m}$ only. To describe the simulation of $\mathbf{z}$ under a general error distribution $f_e(\cdot; \rho, \alpha)$ as above, let us start with a model with no discrepancy term $\delta(x^{(r)})$. Observe that, when we derived (8) starting from (7), only considering the mean of procedure-specific measurements $\bar{y}_{1:m}^{(o)}$ for each experiment was sufficient to write the likelihood. Instead, if we retain the procedure-specific measurements $y_{ij}^{(o)}$, then we can similarly derive

$$y_{ij}^{(o)} = l_i + \tilde{z}_{ij}, \quad \begin{bmatrix} l_{1:m} \\ y_{1:n}^{(c)} \end{bmatrix} = P\beta + \mathrm{MVN}_{m+n}(\mathbf{0}, \sigma^2 D), \quad \tilde{z}_{ij} \sim f_e(\tilde{z}_{ij}; \rho_j, \alpha). \tag{13}$$

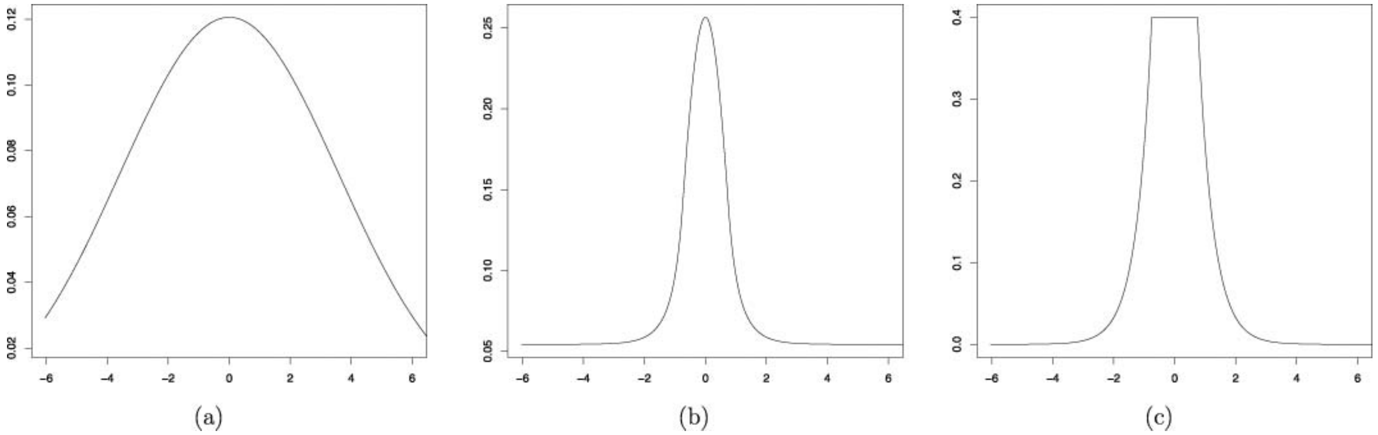
It follows that the z_i introduced in (8) is actually $\sum_{j=1}^3 \tilde{z}_{ij}/3$. If f_e is Gaussian as usual, then the prior for z_i is also Gaussian so it is sufficient to work with the average measurements $\bar{y}_{1:m}^{(o)}$ and $z_{1:m}$. But, with general choices of f_e, one needs to obtain samples of z_i only through samples of $\tilde{z}_{ij}$.

To draw samples of $\tilde{z}_{ij}$, use the fact that although f_e, the prior for $\tilde{z}_{ij}$, is non-Gaussian, the data-dependent part in the posterior is still Gaussian. When f_e is chosen as in (9) or (12), it has closed-form analytic expression and is easy to evaluate at a point. So one can perform an independent Metropolis sampler: draw $l^0 \sim l_{1:m} | y_{1:n}^{(c)}$ and set a proposal for $\tilde{z}_{ij}$ as $z_{ij}^{(p)} = y_{ij}^{(o)} - l_i^0$, simultaneously for all j and i. Then the accept–reject step can be carried out comparing the ratio of prior probabilities under f_e. When m is not small, a single Metropolis step for the entire family of $\tilde{z}_{ij}$ may lead to a poor acceptance rate. Instead, one can update only the set of $\tilde{z}_{ij}$ for a fixed i conditional on $\{z_{i'} : i' \neq i\}$ (similar to the posterior sampling of $\theta^{(r)}$) at a time, recalculate z_i and repeat the step for all $i = 1, 2, \ldots, m$. The rest of the sampling steps can be carried out exactly as described in Section 3.3. When, a covariate-specific discrepancy function $\delta(x_i^{(r)})$ is added, we need to modify only the last distribution in (13) as $\tilde{z}_{ij} = \delta(x_i^{(r)}) + w_{ij}$ with $w_{ij} \sim f_e(w_{ij}; \rho_j, \alpha)$ and a similar Metropolis scheme can be used for w_{ij}.

We conclude this section with an illustrative diagram for different choices of error distribution in $\mathcal{M}_E$. As above, let there be two types of noises, one with a support $[-\alpha, \alpha]$ and another with a decay rate ρ. Assume $\rho = 2, \alpha = 0.75$. Apart from the distributions in (9) and (12), we also considered the usual additive Gaussian framework as in Section 3. For that, we first replaced each of the above noise distributions with an *equivalent* Gaussian distribution. This can be done by first considering an interval (symmetric around zero) with probability 0.99 under the original distribution and then setting the variance of the zero mean Gaussian noise so that the probability of that interval remains unchanged under Gaussianity. Figure 4 shows the decaying pattern of error density around zero (in the same interval [−6,6]) for each of those models.

From the diagram, it is evident that the Gaussian error specification allows for larger values than the remaining two. The sum of Laplace and uniform shrinks the distribution toward errors of smaller magnitude, whereas the Laplace–uniform mixture provides a flat-top structure within $[-\alpha, \alpha]$ and decays beyond that.

In summary, the choice of the error distribution and its parameters depends solely on the prior knowledge of measurement procedure. Any noise distribution, known to have a decaying pattern (i.e., small values are more likely than large values), should be assigned a Gaussian or Laplace prior. On the other hand, if no such information is available, a uniform prior is the suitable choice for representing the flat-shaped noise. If there are multiple sources of uncertainty in the system, it is also up to the user to decide how to combine them. Additive models are the most common choice although, as Figures 4(a) and 4(b) show,



(a) (b) (c)

Figure 4. Probability density function of error for (a) sum of Gaussians, (b) sum of Laplace and uniform, and (c) mixture of Laplace and uniform.

eventually it leads to a distribution for $[y^{(o)}|y^{(r)}]$ that decays around zero. Hence, if one wants to retain the *noninformativeness* in the immediate vicinity of the true value, Figure 4(c) shows a nice formulation to achieve that using uniform and Laplace priors. In Section 5, we compare the sensitivity of the inference to these three choices of measurement error distributions.

5. DATA ANALYSIS

We proceed to the application of the model, developed in the preceding two sections, in the analysis of the shock experiment dataset. In Section 5.1, we start with the outputs from the H2D simulator. We explore different modeling options for this dataset. In Section 5.2, we present a combined analysis of the data pooled from the simulator and the actual shock experiments.

5.1 Analysis of H2D Dataset

In the H2D simulation dataset, the response (y) is the shock breakout time. The fixed inputs (x) are the Be disk thickness and laser energy. Three other inputs are included as the uncertain parameters (θ): Be gas gamma constant, a flux limiter, and tube-wall opacity. We have a total of 104 simulator runs generated from an LHD sampling of these five inputs.

We first want to evaluate different choices for the emulator $\mathcal{M}_C$. As we discussed in Section 3.2, one of the main advantages of using a MARS emulator is to allow for local patterns and nonstationarity in the response surface. Hence, it is of interest to know how this method performs relative to other possible choices of nonstationary models. We specifically chose two of the alternatives: Bayesian additive regression tree (BART; Chipman, George, and McCulloch 2010) and Bayesian treed GP (BTGP; Gramacy and Lee 2008). These two methods have been implemented using “BayesTree” and “tGP” packages inside R (*http://cran.r-project.org*). For software on MCMC-based methods for MARS, we refer to *http://www.stats.ox.ac.uk/∼cholmes/BookCode/*. As far as our models are concerned, for (5) and (6), we only allow interactions of up to the second order. For the GP emulator from (4), we use $\mu(\cdot)$ to be a quadratic trend surface with global parameters. The covariance function is chosen to be stationary but anisotropic with a separable Matérn form across each input dimension. We fix the smoothness parameter of the Matérn function at 2.5 to ensure mean-square differentiability of the output process.

The inference proceeds as follows: we leave out about 10% of the data, fit the model with the remaining points as training data, and then construct point estimate and 90% predictive interval for each point in the test sample. This completes one *trial*. Randomizing selection of the test samples, we performed 50 such trials, sufficient to cover all the simulator runs. Under each model, we provide boxplots of average absolute bias and average predictive uncertainty (estimated as width of predictive interval) over all trials in Figure 5.

Although the models performed comparably for prediction bias, the average predictive uncertainty varied to a larger extent from one model to another. BART generated very large predictive intervals for the test samples. BTGP performed relatively better with respect to predictive uncertainty. Although the traditional GP emulator of Kennedy and O’Hagan (2001) with global quadratic trend produced comparable bias estimates, using a MARS mean contributed to significantly smaller uncertainty estimates for both (5) and (6). We specifically note that (6) fits the training dataset exactly as (4) but was still able to produce much tighter credible intervals for the test samples. In Table 1, we summarize the bias and uncertainty results along with the empirical coverage rate of 90% posterior credible set constructed for each point in the test dataset.

All the models have provided satisfactory coverage rates. It shows that the tighter credible sets generated by MARS-based emulators are not subject to overconfidence inaccuracy. BART produced the highest coverage rate but this is expected given the increased uncertainty associated with its predictions. Based on this statistics, we continue to work with MARS-based emulators in subsequent analysis.

Next, we want to analyze the sensitivity of predictive performance to the choice of k_0, the maximum allowed number of basis functions in MARS. We have mentioned this as a method of strictly penalizing overfitting. For the H2D output, we carry out the estimation with six different values of k_0 ranging from 11 to 81 (including the constant function). The inverse gamma prior for σ_β^2 is chosen to be proper yet diffused. The histogram

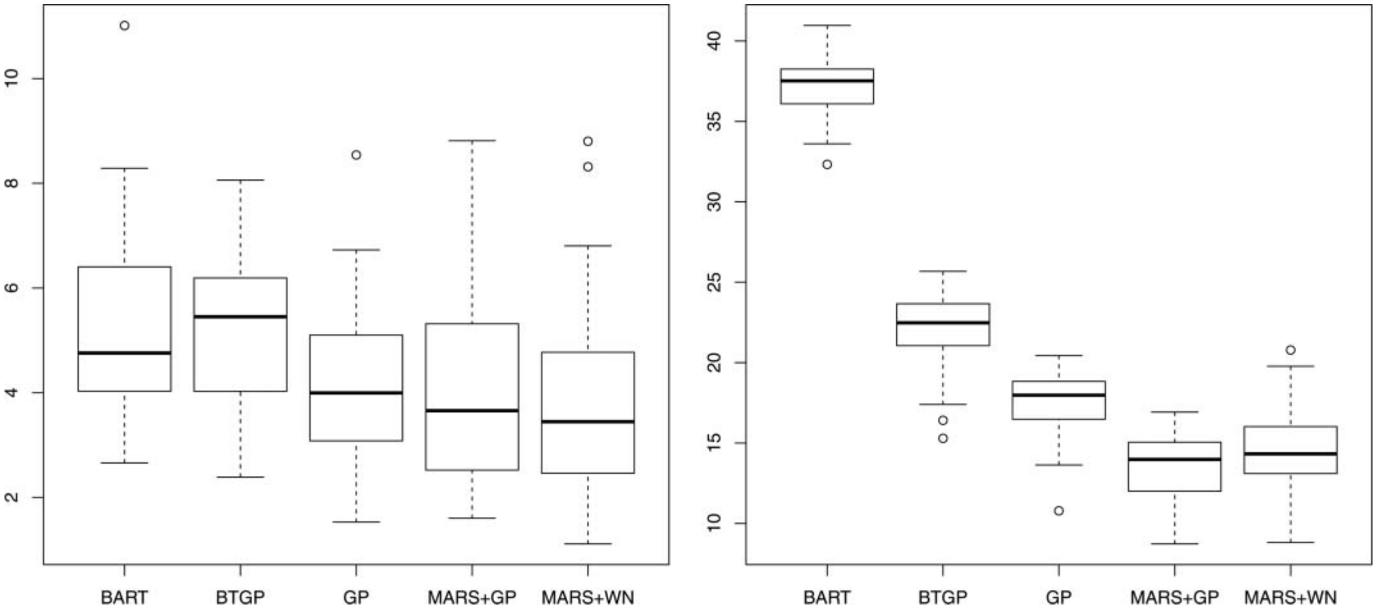


Figure 5. Model performance on (left) absolute error and (right) width of 95% interval width for the prediction of the test samples over 50 trials with the H2D output.

of the posterior samples of k in each of the trials are presented in Figure 6. We observe that, with $k_0 = 11$ (the constant term + up to 10 nonconstant polynomials), the model puts maximum of its weight at k_0, indicating that even larger values are also likely. With $k_0 = 21$, the decaying right tail-behavior of the posterior distribution towards k_0 indicates that this range is potentially sufficient to capture most of the variability. We decided to increase k_0 further to see if it goes on adding more and more functions. However, even with k_0 as large as 81, its behavior is essentially unchanged from $k_0 = 21$, that is, it continues to stay mostly within the region [8, 18]. This shows that the model can, on its own, control overfitting based on the prior for regression coefficients $\{\beta_h\}$ and number of basis functions k (as we discussed in Section 3.2) and we do not need to emphasize too much about choosing a strict bound like k_0.

Looking at the results of Table 1, it is relevant to question the usefulness of a MARS+GP model given the fact that a MARS+WN model, much simpler and faster to run, can show comparable prediction performance. We like to mention a couple of arguments in favor of that. First, as we have noted in Section 3.2, in scientific experiments, it is often preferred to have an emulator that can replicate the code output at previously tried input configurations. For that, we need an interpolator in $\mathcal{M}_C$ and only MARS+GP satisfies that requirement. The other reason comes from the perspective of sparsity in model fitting. In Section 3.2, we have mentioned that we can control the overfitting in MARS by restricting the total number of functions as well as order of each function. Now, if the true function is complex enough and we do not allow sufficient number of functions or restrict our choice only to the lower-order functions, the MARS component may not alone be sufficient in modeling the response. Having an additional GP component can be useful in this situation. To illustrate this argument, we take the H2D code output and carry out a validation study for different choices of k_0 with the first-order interactions *only*. Table 2 presents the comparison of the two procedures under different choices of k_0.

Table 2 highlights the potential usefulness of a MARS+GP specification. In the analysis presented in Table 1, we have used interactions of up to the second order. Figure 6 shows that the reasonable choice for k_0 should be between 15 and 20 for a second-order model. Hence, when we consider a MARS specification with $k_0 = 5$ or 10 and restrict ourselves to the first-order interactions only, that is likely too sparse to capture the important patterns of the response surface. This explains why in Table 2, the bias and uncertainty of prediction has gone up significantly for the MARS+WN type emulator in comparison to Table 1. However, MARS+GP has performed much better, and if we again compare to Table 1, the decrease in the prediction accuracy is much smaller even with k_0 as small as 5. This shows that the GP component can successfully compensate for the insufficient number of local functions. With very complex computer models, this advantage can be significant.

We also like to mention one computational aspect of the problem. As we have indicated in the end of Section 3.2, the MARS+GP model for emulator involves a GP covariance structure that gets computationally extensive with large number of

Table 1. Predictive performance for different emulator choices

Criterion	BART	BTGP	GP w/quadratic trend	MARS+GP	MARS+WN
Absolute bias	4.7577	5.4499	3.9960	3.6562	3.4462
Predictive uncertainty	37.5173	22.4752	17.9758	13.9765	14.3267
Empirical coverage rate of 90% credible set	0.9517	0.8883	0.8867	0.9017	0.9100

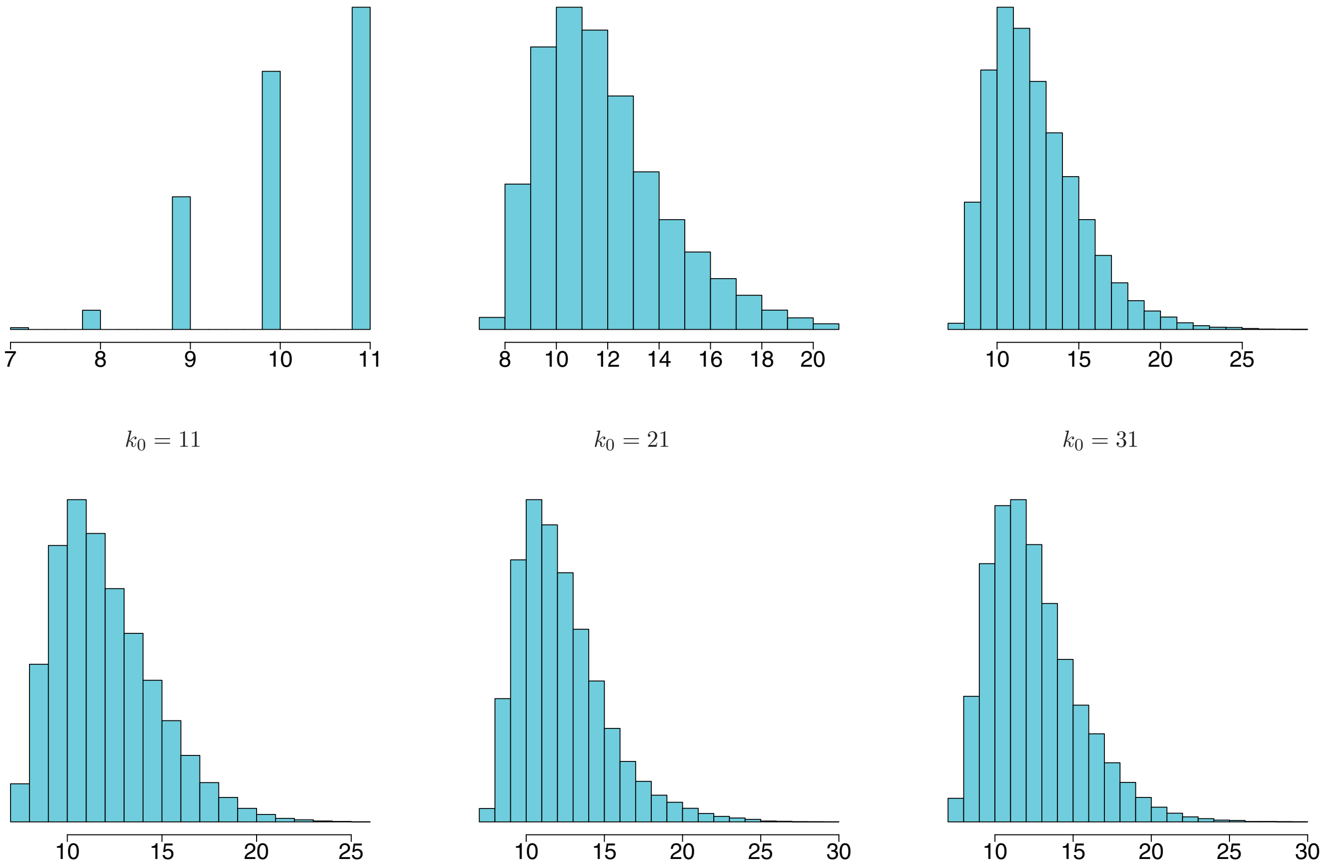
Table 2. Comparison of MARS+WN and MARS+GP methods with linear interactions

Method	MARS+WN			MARS+GP		
k_0	5	10	15	5	10	15
Absolute bias	3.6501	3.6098	3.4775	2.8795	2.7536	2.7512
Predictive uncertainty	21.9697	21.1067	21.2541	16.4132	16.9865	17.0799
Empirical coverage rate of 90% credible set	0.9250	0.9233	0.9283	0.9017	0.9033	0.8950

observations. Currently, we have only 104 outputs from the H2D code, but we expect to collect more and more of them over time. Thus it is relevant to discuss possible modifications we may have to implement in that scenario. Approximate computation methods, as indicated in Section 3.2, can be used there. However, unlike usual GP regression for two- or three-dimensional spatial datasets, this problem involves relatively higher-dimensional input space (five-dimensional input for the current experiment). We like to illustrate one specific choice that may be more convenient to use in this type of situations. The approach, presented in Kaufman et al. (2011) in the context of a simulator related to cosmological applications, is based on using compactly supported correlation functions to ensure sparsity of GP covariance matrix. The range of the sparsity is controlled hierarchically and can vary across different inputs. Since the size of the current dataset is not large enough to efficiently represent the computational gains from this method, we decide to perform a simulation study. A brief description of this idea along with an evaluation of its performance with respect to predictive performance as well as time efficiency is included in the online supplementary materials.

5.2 Validation With Laboratory Experiments

Now that we have evaluated different options for modeling the code output, the next step is to validate it using outcomes from actual laboratory experiments. In each of the eight experiments we have results from, the shock breakout time was measured by three different diagnostics (ASBO1, ASBO2, and SOP). For one of those experiments, SOP measurement was not available. The ranges of measurement inaccuracy for ASBO1 (±10 ps), ASBO2 (±20 ps), and SOP (±30 ps) are converted to standard



$k_0 = 11$ $k_0 = 21$ $k_0 = 31$

$k_0 = 41$ $k_0 = 61$ $k_0 = 81$

Figure 6. Variation in the posterior distribution of number of local functions k under different choices of the threshold k_0. The online version of this figure is in color.

Table 3. Mean absolute error and predictive uncertainty for different emulator choices

	Prediction criteria			
	Absolue bias		Uncertainty	
Device type	MARS+GP	MARS+WN	MARS+GP	MARS+WN
ASBO1	17.6760	17.3552	72.2343	71.7667
ASBO2	19.3281	18.3467	75.3237	73.4993
SOP	16.2600	16.3400	81.7914	80.2587

deviations for Gaussian error distribution with the 99% probability criterion used for Figure 4. Finally, in all measurements, there is a systematic timing error of ±50 ps for the laser firing mechanism. This is represented by ϵ_0 in the model (3).

We employ the full hierarchical model in (7) using the sampling scheme described in Section 3.3. Here, we follow the leave-one-out validation procedure, that is, at one run, we remove all measurements that belong to a particular experiment. Models are fitted to remaining data (simulator and experimental outcomes together), and similar to the above, point and interval estimates for left-out data points are obtained from posterior draws, which are converted to estimates of bias and uncertainty as before. Table 3 provides the posterior mean absolute bias and predictive uncertainty for the three models, for each of three different measurement procedures.

As expected, the uncertainty estimates for ASBO1, ASBO2, and SOP measurements follow the same (increasing) order as their individual precisions. Both of the emulators have produced comparable bias estimates. However, using MARS+WN results in a slight reduction in the uncertainty estimates across all types of measurements. We can attribute this to the additional uncertainty due to estimation of GP covariance parameters.

We also want to analyze, whether using an input-dependent bias function $\delta(x^{(r)})$ improves the prediction in MARS-based models. In Table 4, we compare the average absolute error and prediction uncertainty obtained by carrying out estimation with and without the $\delta(\cdot)$ function.

It can be seen that for neither of the emulators, adding a bias term does not significantly alter the magnitude of prediction error. On the other hand, the predictive uncertainty has slightly gone up in most of the cases due to the additional uncertainty in the parameters of $\delta(x^{(r)})$. So, we can conclude that the H2D algorithm provides a satisfactory approximation of the dynamics of the actual shock experiments. In the following, all subsequent analysis are carried out without the discrepancy term $\delta(x^{(r)})$ in $\mathcal{M}_R$.

Table 4. Impact of bias function on prediction error and uncertainty (in parentheses)

	MARS+GP		MARS+WN	
Device type	Without $\delta(\cdot)$	With $\delta(\cdot)$	Without $\delta(\cdot)$	With $\delta(\cdot)$
ASBO1	17.6760 (72.2343)	17.2680 (72.9123)	17.3552 (71.7667)	18.2971 (71.9894)
ASBO2	19.3281 (75.3237)	19.1712 (77.6631)	18.3467 (73.4993)	19.4010 (75.0506)
SOP	16.2600 (81.7914)	16.6120 (82.1937)	16.3400 (80.2587)	16.0868 (79.3420)

Table 5. Posterior summary for calibration parameters under MARS+WN model

Calibration parameter	Be gamma	Wall opacity	Flux limiter
Median	1.435	1.005	0.0595
90% credible interval	(1.402,1.483)	(0.719, 1.275)	(0.0504, 0.0722)

Another quantity of inferential interest is the vector of calibration parameters $\theta^{(r)}$. While designing the input configurations for the code, the scientists chose regularly spaced values for each component of $\theta^{(c)}$ within a prespecified range of likely values. For each component of $\theta^{(r)}$, we use independent uniform priors over the set of those input configurations. Post model-fitting, we found the corresponding posterior distributions not much sensitive to the choice of emulator model. Hence, it suffices to present the summary statistics only for the MARS+WN model in Table 5. Visual representation through posterior histograms and pairwise bivariate kernel density estimates are shown in Figures 7 and 8, respectively.

The extent of nonuniformity (and peakedness) in the posterior of each calibration parameter reflects whether the data have strong enough evidence to identify its "correct" value in the experiments. The posterior distribution of Be gamma, which is seen (in Figures 7 and 8) to be concentrated only within a subregion of its prior support, looks to be the most informative of all three quantities. However, the flux limiter has shown slightly more nonuniformity than the wall opacity. With respect to prediction, this implies that the model for shock breakout time is more sensitive to the uncertainty in learning Be gamma than flux limiter and wall opacity.

Finally, we move to the diagnostics related to the specification of measurement error. Each of the three measurement procedures (ASBO1, ASBO2, and SOP) is known to have a decaying error pattern (with different rates of decay), but the timing error for the laser firing mechanism is more noninformative. The laboratory does not have any further information on the actual shock breakout time within ±50 ps of the reported value. We take $\alpha = 50$. For $j = 1, 2, 3$, ρ_j was determined (as above) so that the Laplace distribution with rate ρ_j has 99% of its mass inside the desired range of accuracy. Now we fit (5) and (6) with each of (9) and (12) as the choice for the measurement error model. All the error distributions in Figure 4 have zero mean, but quite different patterns. The range of uniformity ($2\alpha = 100$ ps) is significantly large compared with the magnitude of the response (410–504 ps). Diagnostic outputs such as mean absolute predictive bias and uncertainty are provided in Table 6 across measurement types as well as choice of error distributions.

Notably, with both the emulators, the mean predictive absolute bias for all types of measurements does not vary significantly with the choice of the measurement error model. But the predictive uncertainty is significantly affected due to the noninformativeness of the latter two specifications. The use of the Laplace–uniform mixture has resulted in the largest uncertainty estimates, which is expected due to the flat-top nature of the

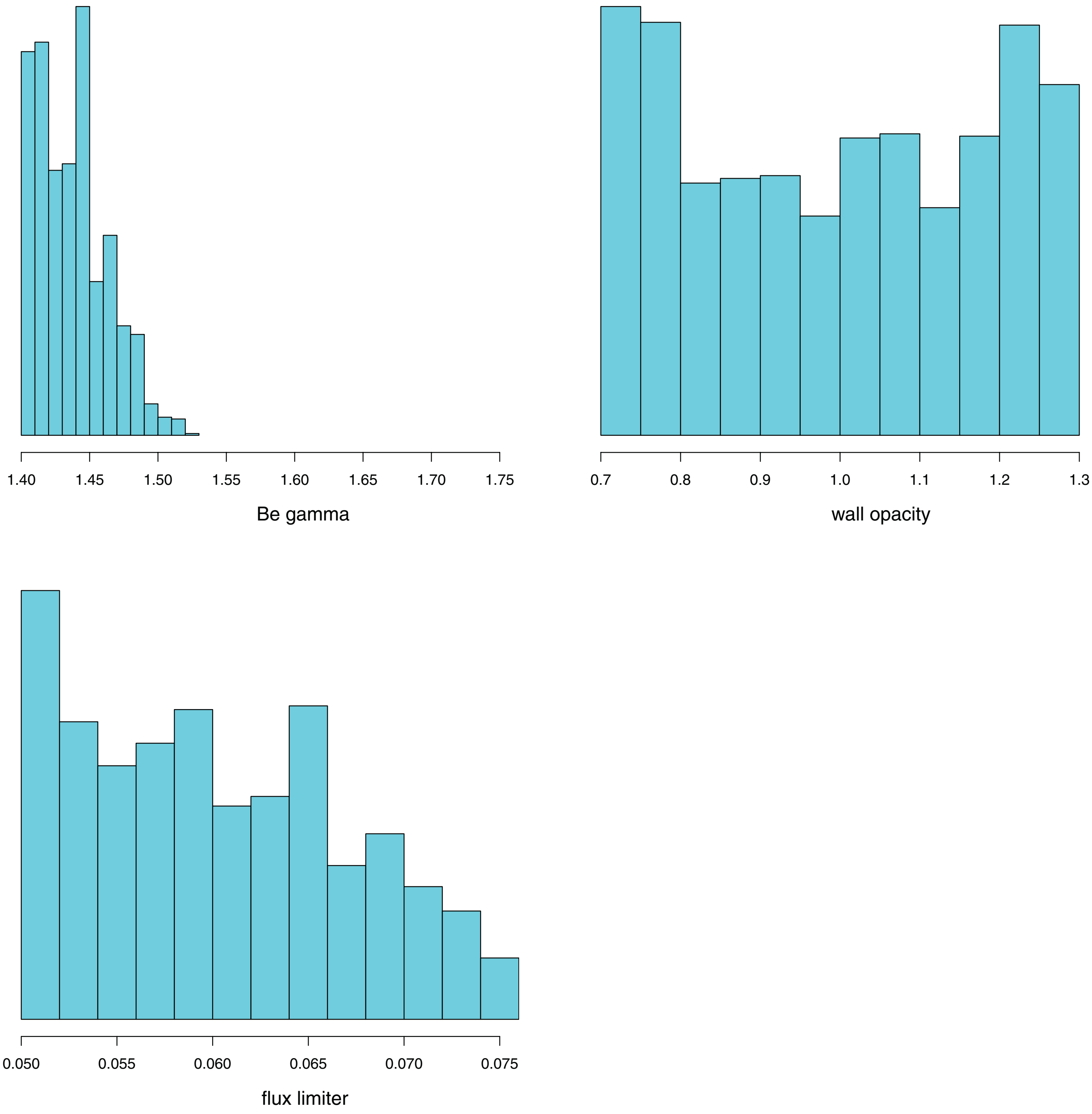


Figure 7. Posterior density estimates for the calibration parameters in the H2D dataset. The plots correspond to the model with MARS+WN specification. The online version of this figure is in color.

error within a ±50 ps interval around the true outcome. Both of the emulators have shown comparable performance with respect to prediction bias. However, it is important to observe that, unlike the Gaussian-error-based model, MARS+GP has produced slightly lower uncertainty estimates than MARS+WN with non-Gaussian measurement error distributions across all types of measuring devices. Since these error distributions have more uncertainty than an equivalent Gaussian model (as seen in Figure 4) and Table 4 shows that the simulator is a good approximate of the actual experiment, we can conclude that borrowing more information from the set of simulation outputs has actually helped in marginally reducing the uncertainty without compromising on the bias.

Based on the analysis in this section, MARS-based emulator turns out to be a flexible and efficient option for reducing the bias and uncertainty of prediction of the shock breakout time. It is also satisfactory to learn that the dynamics of the actual experimental system can be efficiently approximated by the numerical algorithm without the need for any discrepancy function. All of the performance diagnostics presented above indicate that MARS+WN and MARS+GP are comparable choices as the emulator for the H2D code. Although the former is

Table 6. Comparison of different choices of the measurement error models

		Measurement error model					
		Gaussian		Laplace+uniform		Laplace–uniform mixture	
Criterion	Device type	MARS+WN	MARS+GP	MARS+WN	MARS+GP	MARS+WN	MARS+GP
	ASBO1	17.3552	17.6760	16.5824	15.6187	17.0972	15.9965
Bias	ASBO2	18.3467	19.3281	18.7290	18.5009	18.3944	14.6391
	SOP	16.3400	16.2600	15.8779	17.8895	15.4721	19.1734
	ASBO1	71.7667	72.2343	96.8129	95.6592	104.3146	103.1104
Uncertainty	ASBO2	73.4993	75.3237	98.9259	97.3865	108.0810	107.1081
	SOP	80.2587	81.7914	102.7600	100.3264	112.2484	111.4389

computationally much more efficient, the latter is particularly useful with more complex systems that would otherwise require a large number of basis functions (as seen in Table 2). If the objective of modeling is to have an emulator which is a statistical interpolator, that is, (i) it matches the code output at previously used input combinations and (ii) provides an uncertainty estimate at a new point, MARS+GP is the only choice. It uses the flexibility of MARS to provide uncertainty estimates for prediction at new points and, at the same time, satisfies the criterion of interpolation due to the GP component. Consequently,

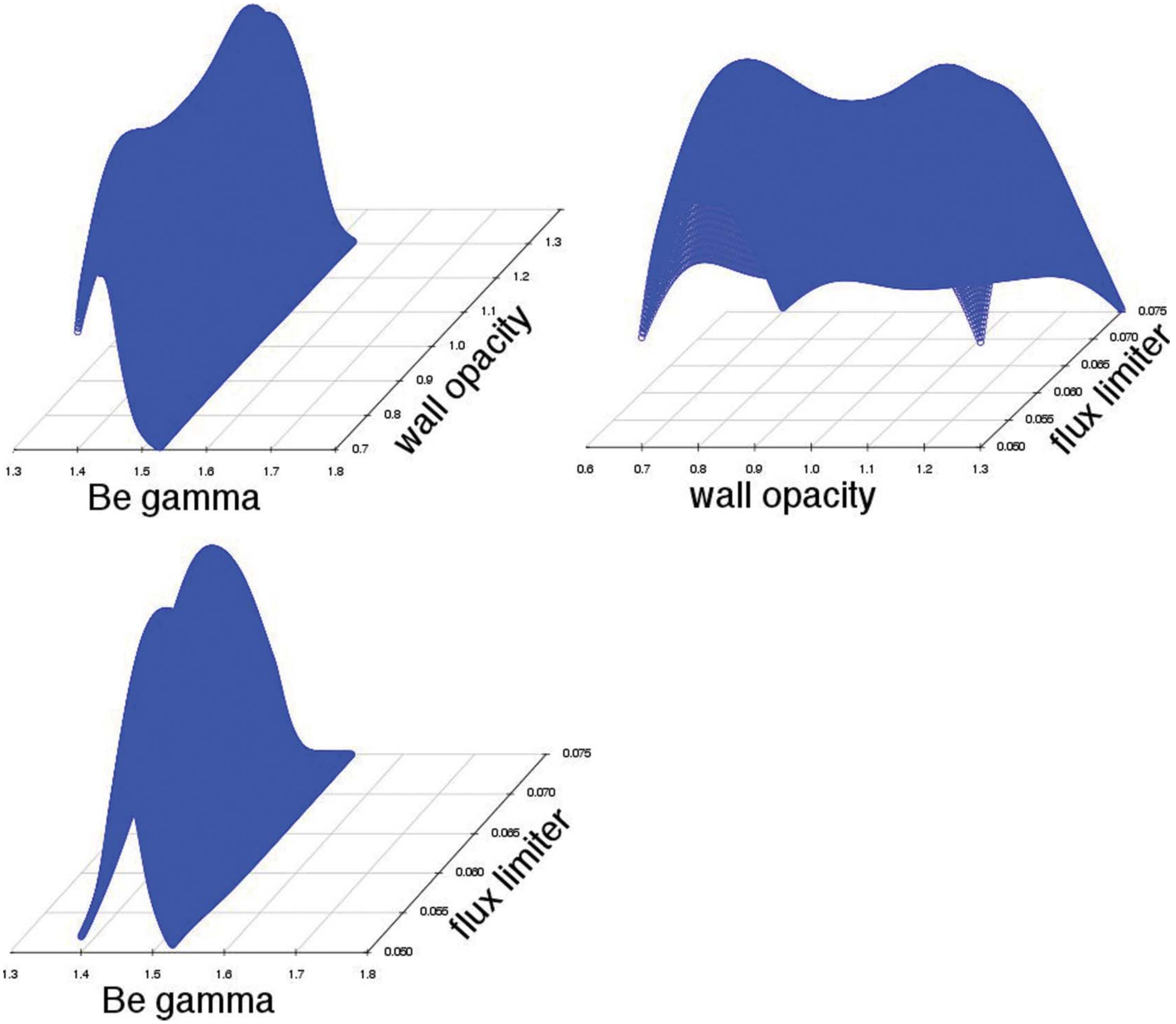


Figure 8. Pairwise bivariate kernel density estimates for the calibration parameters in the H2D dataset. The plots correspond to the model with MARS+WN specification. The online version of this figure is in color.

this leads to a more flexible approach to modeling nonstationarity within a full Kennedy–O'Hagan framework relative to the current literature.

6. DISCUSSION

In this article, we focused on validating the HYADES code for the radiative shock experiments. First, we presented two different types of emulators based on adaptive splines for predicting the shock breakout time. Their properties and relative advantage over the conventional GP emulator were discussed in Section 3, and a thorough performance analysis with the H2D output was done in Section 5. The data analysis shows the predictive power of the MARS regression compared with the GP alternatives in different circumstances. While (5) provides a computationally fast algorithm, the combined approach in (6) simultaneously uses this advantage of spline-based emulator and maintains the interpolating behavior of $\mathcal{M}_C$. Another concern for using the GP emulators rises from model-fitting perspective. For the experimental data, $\theta^{(r)}$ is unknown and needs to be estimated from the MCMC. Although for spline-based emulators, it is involved in the mean only, for the GP, it appears both in mean and covariance matrix. Learning $\theta^{(r)}$ from the not-so-well-behaved posterior distribution can be problematic. However, for specific applications, different emulators can be preferred depending on the practical considerations as well as the desired features of the emulator.

Another approach, similar to the MARS, is to specify f using radial basis functions (Holmes and Mallick 1998). There, instead of using the tensor product of univariate local basis functions, weighted multivariate kernels are used around the knots. It should be remembered that these basis function approaches are closely related to GP (Gibbs and Mackay 1997). Any combination of such functions corresponds to the eigen decomposition of some (nonstationary) covariance kernel. Introducing a WN in the model amounts to adding a nugget with the equivalent GP prior that destroys its interpolating property.

The work in this article fits into the larger goal of the CRASH center to develop the predictive capability for radiating shock experiments. In particular, the center seeks to build a framework that incorporates data from shocks traveling down cylindrical tubes and then use the simulation tools to predict experiments using tubes of elliptical cross-sections. HYADES is a preprocessor in the entire CRASH code in the sense that the output of the former is the input to the latter. Thus, a hierarchical extension of the existing model, which emulates the entire CRASH code, enables us to faithfully learn about particular aspects of the full experiment. Specifically, we must justify that

- our physics models are adequate,
- our codes are behaving as expected, and
- we understand the impact of uncertainties on code output and prediction.

In this case, we have looked at the shock breakout time and determined that our physics models are capable of reproducing experimental data after calibration; other experiment/simulation campaigns have considered different features such as the shock position and the total amount of shocked Xe and studied discrepancy between the physical model and the measurement (Holloway et al. 2011).

In the general context of the computer model validation, further improvement is possible. Any model, which combines a computer code with a real-life experiment, consists of three major components: an emulator function for the code output, a measurement error process for the experimental noise, and, if necessary, a residual function that accounts for the inadequacy of the physical model in predicting the input–output relationship in the real world. In this article, we have focussed on the first two components, suggesting specifications to enhance their applicability in different examples. However, for the residual function, we preferred to retain the zero-mean GP prior of Kennedy–O'Hagan's (2001) setup. Small number of experimental outputs as well as lack of substantial prior knowledge do not make the current dataset ideal for exploring a richer class of models for $\delta(\cdot)$ function. However, there are examples (e.g., experiments related to climate models) where a large number of real-world observations are available in addition to a complex simulator. There, it may be of significant interest to use models beyond GP, whcih allows for nonstationary patterns in the function surface and incorporates any prior knowledge on the form and contributing factors of such discrepancy. Another extension of the current model is to incorporate multivariate response (e.g., different features of the shock such as location, speed, etc.) Simultaneous or conditional modeling of the response can reveal the extent of association between those factors and possibly can improve the uncertainty of the prediction by borrowing of information through the prior as well as the calibration parameters. Finally, in the goal of incorporating all possible sources of bias inside the hierarchy, a direction of future work is to consider uncertainty in measurements of x; x is treated as the fixed input throughout this article and thus any inference of the distribution of y is conditional on x. However, even the known inputs can have measurement error in them. For example, laser energy is known to exhibit a quasi-Gaussian pattern, whereas a quasi-uniform distribution is likely for Be thickness. If a stochastic model for x is assumed, marginal predictive uncertainty for y can be subsequently obtained.

APPENDIX A. MARGINALIZING β AND σ^2

Denote by ... all parameters except β, σ^2. Let

$$S_y = \mathbf{y}_f - \begin{bmatrix} z_{1:m} \\ \mathbf{0}_n \end{bmatrix}, \quad S = S_y - P\beta.$$

We have,

$$\begin{aligned} p(y_f|\ldots) &= \int_\beta \int_{\sigma^2} p(y_f|\beta, \sigma^2, \ldots) p(\beta|\sigma^2) p(\sigma^2)\, d\sigma^2 d\beta, \\ &\propto \left(2\pi\sigma_\beta^2\right)^{-k/2} \int_\beta \int_{\sigma^2} (\sigma^2)^{-\frac{n+m+k}{2} - a_\sigma - 1} \\ &\quad \times \exp\left[-\frac{1}{2\sigma^2} \left(S^T D^{-1} S + \beta^T \beta/\sigma_\beta^2 + 2b_\sigma\right) \right] d\sigma^2 d\beta, \\ &\propto \left(2\pi\sigma_\beta^2\right)^{-k/2} \int_\beta \left(\frac{S^T D^{-1} S + \beta^T \beta \big/ \sigma_\beta^2}{2} + b_\sigma \right)^{-\frac{n+m+k}{2} - a_\sigma} d\beta. \end{aligned}$$

Now write $S^T D^{-1} S + \beta^T \beta/\sigma_\beta^2 = \beta^T A \beta - 2\beta^T B + C$, where $A = P^T D^{-1} P + \frac{\mathbf{I}_k}{\sigma_\beta^2}$, $B = P^T D^{-1} S_y$, and $C = S_y^T D^{-1} S_y$. Then we have, $S^T D^{-1} S + \beta^T \beta/\sigma_\beta^2 + 2b_\sigma = (\beta - \mu_k)^T \Sigma_k^{-1} (\beta - \mu_k) + c_{0k}$, where $\mu_k = A^{-1} B$, $\Sigma_k = A^{-1}$, and $c_{0k} = C - B^T A^{-1} B + 2b_\sigma$.

Denote $d = n + m + 2a_\sigma$. Hence,

$$p(y_f \mid \ldots) \propto$$

$$\left(\pi\sigma_\beta^2\right)^{-k/2} c_{0k}^{-\frac{d+k}{2}} \int_\beta \left[\frac{1}{d}(\beta-\mu_k)^T\left(\frac{c_{0k}\Sigma_k}{d}\right)^{-1}(\beta-\mu_k)+1\right]^{-\frac{d+k}{2}} d\beta.$$

The integrand is the pdf (upto constant) for the k-variate t distribution with mean μ_k, dispersion $\frac{c_{0k}\Sigma_k}{d}$, and degrees of freedom d. Hence, we obtain the closed-form expression:

$$p(y_f \mid \ldots) \propto \left(\sigma_\beta^2\right)^{-k/2} c_{0k}^{-\frac{d}{2}} |\Sigma_k|^{1/2}.$$

APPENDIX B. RJMCMC MOVES

We first mention the prior for (k, α_k) in the form of $p(\alpha_k|k)p(k)$. As specified in (7), $(k-1)$ has a Poisson(λ) prior truncated at some upper bound k_0. As specified, for fixed k, $\alpha_k = \{(n_h, \mathbf{u}_h, \mathbf{v}_h, \mathbf{t}_h) : h = 1, 2, \ldots, k\}$. The first term being the intercept, $2 \le h \le k$ correspond to the $(k-1)$ nonconstant functions present in the model. We require that these members are distinct, that is, if $h \ne h'$, then the quadruples $(n_h, \mathbf{u}_h, \mathbf{v}_h, \mathbf{t}_h)$ and $(n_{h'}, \mathbf{u}_{h'}, \mathbf{v}_{h'}, \mathbf{t}_{h'})$ are different. If p is the total number of covariates and we allow interactions up to the second order (n_h = 1 or 2), then number of possible choices for a nonconstant basis function is $N = p + p + \binom{p}{2} = \frac{p^2+3p}{2}$. Allowing a particular function to be chosen repeatedly, the number of *ordered* choices of $(k-1)$ members is N^{k-1}. For each of those functions, the required knots (n_h many of them for hth basis function) can be chosen uniformly from the available data points (since a change in pattern can only be detected at data points) and its sign can be selected in two ways—either positive or negative with equal probability. Thus, in all, there are $N^{k-1}(2n)^{\sum_{h=2}^k n_h}$ ordered ways of constructing α_k. Now, α_k is a set, so that order of the quadruples does not matter, and we are looking at only those choices with distinct members, so the same α_k can be generated by selecting those $(k-1)$ terms in any of the $(k-1)!$ possible permutations. Thus, prior probability of any candidate α_k can be written as

$$p(\alpha_k|k) \propto \frac{(k-1)!}{N^{k-1}}(1/2n)^{\sum_{h=2}^k n_h}.$$

The expression for $p(\alpha_k|k)$ does *not* put zero probability on any configuration whose members are not all distinct (as defined above). However, with moderately large values of N and n, the probability of any such duplication becomes very small. We also assume that all covariates have n distinct values to locate a knot at; modification can be made easily when this is not the case.

Next we specify the proposal distribution $q(\cdot,\cdot)$ for each of the three moves as follows:

(i) First decide on the type of move to be proposed with probabilities b_k (birth), d_k (death), and c_k (change), $b_k + d_k + c_k = 1$. We put $d_k = 0$, $c_k = 0$ if $k = 1$, $b_k = 0$ if $k = k_0$.

(ii) For a *birth* move, choose a new basis function randomly from the N-set. Calculate its order n_h. If this selected function does not agree with the existing $(k-1)$ nonconstant basis functions, then choose its knots and signs as before with probability $(\frac{1}{2n})^{n_h}$. Otherwise, to avoid repetition of the same polynomial form, choose the knots and signs in $(2n)^{n_h} - 1$ possible ways.

(iii) The *death* move is performed by randomly removing one of the $(k-1)$ existing basis functions (excluding the constant basis function).

(iv) A *change* move consists of altering the sign and knot variables of a randomly chosen nonconstant basis function.

From above, we have

$$q((k,\alpha_k)\to(k',\alpha_{k'})) = \begin{cases} b_k \dfrac{1}{N}\dfrac{1}{(2n)^{n_{k+1}} - r_k} & k' = k+1, \\ d_k \dfrac{1}{k-1} & k' = k-1 \\ c_k \dfrac{1}{k-1}\dfrac{1}{(2n)^{n_h} - 1} & k' = k. \end{cases}$$

In above, $r_k = 1$ if there exists an integer $h \in [2, k]$ such that $n_h = n_{k+1}$ and $\mathbf{v}_h = \mathbf{v}_{k+1}$, 0 otherwise. For the "change" step, h denotes the index of basis function that has been randomly chosen for change. The acceptance ratios for different types of move can be derived from this.

APPENDIX C. DERIVATION OF THE MIXTURE FORM FOR VAPNIK'S LOSS FUNCTION

We have the error density

$$f\left(y_j^{(o)}\right) \propto \begin{cases} 1 & \left|y_j^{(o)} - y^{(r)}\right| < \alpha, \\ \exp\left(-\rho_j\left(\left|y_j^{(o)} - y^{(r)}\right| - \alpha\right)\right) & \left|y_j^{(o)} - y^{(r)}\right| > \alpha. \end{cases}$$

For the first case above, the integral is $c_u = 2\alpha$. For the latter case, the integral

$$\begin{aligned} c_l &= \int_{|y-y^{(r)}|>\alpha} \exp\left(-\rho_j|y - y^{(r)}| + \rho_j\alpha\right)\, dy \\ &= \int_{y>y^{(r)}+\alpha} \exp\left(-\rho_j(y - y^{(r)}) + \rho_j\alpha\right)\, dy \\ &\quad + \int_{y<y^{(r)}-\alpha} \exp\left(\rho_j(y - y^{(r)}) + \rho_j\alpha\right)\, dy \\ &= \frac{2}{\rho_j} \end{aligned}$$

Thus we have $f(y_j^{(o)}) = \frac{1}{c_l+c_u}[g_1(y_j^{(o)}) + g_2(y_j^{(o)})]$, where g_1, g_2 are unnormalized Laplace$(\rho_j, y^{(r)}, \alpha)$ and Unif$(y^{(r)} - \alpha, y^{(r)} + \alpha)$ density, respectively. After appropriate rescaling by c_l and c_u, we have $f(y_j^{(o)}) = \frac{c_l}{c_l+c_u} f_1(y_j^{(o)}) + \frac{c_u}{c_l+c_u} f_2(y_j^{(o)})$, where f_i is the normalized version of g_i, $i = 1, 2$. Hence, the mixture weight of first component $p_j = \frac{c_l}{c_l+c_u} = \frac{1}{1+\rho_j\alpha}$.

SUPPLEMENTARY MATERIALS

The supplementary materials discuss the use of compactly supported covariance function approach of Kaufman et al. (2011) inside the MARS+GP framework; it can be useful for working with very large number of computer simulations.

[Received July 2011. Revised December 2012.]

REFERENCES

Arellano-Valle, R. B., Ozan, S., Bolfarine, H., and Lachos, V. H. (2005), "Skew Normal Measurement Error Models," *Journal of Multivariate Analysis*, 96, 265–281. [412]

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2004), *Hierarchical Modeling and Analysis for Spatial Data*, Boca Raton, FL: Chapman & Hall/ CRC. [415]

Banerjee, S., Gelfand, A., Finley, A., and Sang, H. (2008), "Gaussian Predictive Process Models for Large Spatial Data Sets," *Journal of the Royal Statistical Society*, Series B, 70, 825–848. [415]

Barker, L. M., and Hollenbach, R. E. (1972), "Laser Interferometer for Measuring High Velocities of Any Reflecting Surface," *Journal of Applied Physics*, 43, 4669–4675. [413]

Bastos, L., and O'Hagan, A. (2009), "Diagnostics for Gaussian Process Emulators," *Technometrics*, 51, 425–438. [412]

Bayarri, M., Berger, J., Cafeo, J., Garcia-Donato, G., Liu, F., Palomo, J., Parthasarathy, R., Paulo, R., Sacks, J., and Walsh, D. (2007), "Computer Model Validation With Functional Output," *The Annals of Statistics*, 35, 1874–1906. [412]

Bayarri, M. J., Berger, J. O., Calder, E. S., Dalbey, K., Lunagomez, S., Patra, A. K., Pitman, E. B., Spiller, E. T., and Wolpert, R. L. (2009), "Using Statistical and Computer Models to Quantify Volcanic Hazards," *Technometrics*, 51, 402–413. [411]

Bernardo, J. M. (1979), "Expected Information as Expected Utility," *The Annals of Statistics*, 7, 686–690. [419]

Boehly, T. R., Craxton, R. S., Hinterman, T. H., Kelly, J. H., Kessler, T. J., Kumpan, S. A., Letzring, S. A., McCrory, R. L., Morse, S. F. B., Seka, W., Skupsky, S., Soures, J. M., Verdon, C. P. (1995), "The Upgrade to the OMEGA Laser System," *Review of Scientific Instruments*, 66, 508–510. [411]

Chen, G., Gott III, J., and Ratra, B. (2003), "Non-Gaussian Error Distribution of Hubble Constant Measurements," *Publications of the Astronomical Society of the Pacific*, 115, 1269–1279. [412]

Chevalier, R. A. (1997), "Type II Supernovae SN 1987A and SN 1993J," *Science*, 276, 1374–1378. [411]

Chipman, H. A., George, E. I., and McCulloch, R. E. (2010), "BART: Bayesian Additive Regression Trees," *The Annals of Applied Statistics*, 4, 266–298. [420]

Cressie, N., and Johannesson, G. (2008), "Fixed Rank Kriging for Very Large Spatial Data Sets," *Journal of the Royal Statistical Society*, Series B, 70, 209–226. [415]

Denison, D. G. T., Holmes, C. C., Mallick, B. K., and Smith, A. F. M. (2002), *Bayesian Methods for Nonlinear Classification and Regression*, Chichester, UK: Wiley. [412]

Denison, D. G. T., Mallick, B. K., and Smith, A. F. M. (1998), "Bayesian Mars," *Statistics and Computing*, 8, 337–346. [416]

Drake, R. P. (2006), *High-Energy-Density Physics: Fundamentals, Inertial Fusion, and Experimental Astrophysics*, Berlin/Heidelberg, Germany: Springer-Verlag. [411]

Fricker, T., Oakley, J., and Urban, N. M. (2010), "Multivariate Emulators With Nonseparable Covariance Structures," Technical Report, MUCM Technical Report, University of Sheffield. [412]

Friedman, J. (1991), "Multivariate Adaptive Regression Splines," *The Annals of Statistics*, 19, 1–67. [416]

Gibbs, M., and Mackay, D. J. C. (1997), "Efficient Implementation of Gaussian Processes," Technical Report, Cavendish Laboratory, Cambridge, UK. [426]

Gramacy, R. B., and Lee, H. K. H. (2008), "Bayesian Treed Gaussian Process Models With an Application to Computer Modeling," *Journal of the American Statistical Association*, 103, 1119–1130. [420]

Habib, S., Heitmann, K., Higdon, D., Nakhleh, C., and Williams, B. (2007), "Cosmic Calibration: Constraints From the Matter Power Spectrum and the Cosmic Microwave Background," *Physical Review Letters D*, 76, 083503. [411]

Higdon, D. (2002), "Space and Space-Time Modeling Using Process Convolutions," in *Quantitative Methods for Current Environmental Issues*, eds. C. Anderson, V. Barnett, P. C. Chatwin, and A. H. El-Shaarawi, London: Springer, pp. 37–56. [415]

Higdon, D., Gattiker, J., Williams, B., and Rightley, M. (2008a), "Computer Model Calibration Using High-Dimensional Output," *Journal of the American Statistical Association*, 103, 570–583. [412]

Higdon, D., Kennedy, M., Cavendish, J., Cafeo, J., and Ryne, R. D. (2004), "Combining Field Observations and Simulations for Calibration and Prediction," *SIAM Journal of Scientific Computing*, 26, 448–466. [412,415]

Higdon, D., Nakhleh, C., Gattiker, J., and Williams, B. (2008b), "A Bayesian Calibration Approach to the Thermal Problem," *Computer Methods in Applied Mechanics and Engineering*, 197, 2431–2441. [411]

Holloway, J. P., Bingham, D., Chou, C., Doss, F., Drake, R. P., Fryxell, B., Grosskopf, M., van der Holst, B., Mallick, B. K., McClarren, R., Mukherjee, A., Nair, V., Powell, K. G., Ryu, D., Sokolov, I., Toth, G., and Zhang, Z. (2011), "Predictive Modeling of a Radiative Shock System," *Reliability Engineering and System Safety*, 96, 1184–1193. [426]

Holmes, C. C., and Mallick, B. K. (1998), "Bayesian Radial Basis Functions of Variable Dimension," *Neural Computation*, 10, 1217–1233. [426]

Johnson, M. E., Moore, L. M., and Ylvisaker, D. (1990), "Minimax and Maximin Distance Designs," *Journal of Statistical Planning and Inference*, 26, 131–148. [414]

Kauermann, G., and Opsomer, J. D. (2011), "Data-Driven Selection of the Spline Dimension in Penalized Spline Regression," *Biometrika*, 98, 225–230. [416]

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," *The Annals of Applied Statistics*, 5, 2470–2492. [415,422]

Kaufman, C. G., Schervish, M. J., and Nychka, D. W. (2008), "Covariance Tapering for Likelihood-Based Estimation in Large Spatial Data Sets," *Journal of the American Statistical Association*, 103, 1545–1555. [415]

Kennedy, M., Anderson, C., O'Hagan, A., Lomas, M., Woodward, I., Gosling, J. P., and Heinemeyer, A. (2008), "Quantifying Uncertainty in the Biospheric Carbon Flux for England and Wales," *Journal of the Royal Statistical Society*, Series A, 171, 109–135. [411]

Kennedy, M. C., and O'Hagan, A. (2001), "Bayesian Calibration of Computer Models," *Journal of the Royal Statistical Society*, Series B, 63, 425–464. [412,414,415,416,418,420]

Liu, F., and West, M. (2009), "A Dynamic Modelling Strategy for Bayesian Computer Model Emulation," *Bayesian Analysis*, 4, 393–412. [412]

Mallick, B., and Gelfand, A. E. (1995), "Bayesian Analysis of Semiparametric Proportional Hazards Models," *Biometrics*, 51, 843–852. [412]

Matérn, B. (1960), *Spatial Variation (Lecture Notes in Statistics)*, Berlin, Germany: Springer-Verlag. [415]

McClarren, R. G., Ryu, D., Drake, P., Grosskopf, M., Bingham, D., Chou, C., Fryxell, B., Van der Holst, B., Holloway, J. P., Kuranz, C. C., Mallick, B., Rutter, E., and Torralva, B. R. (2011), "A Physics Informed Emulator for Laser-Driven Radiating Shock Simulations," *Reliability Engineering and System Safety*, 96, 1194–1207. [414]

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, 034903. [413]

Neal, R. M. (1999), "Regression and Classification Using Gaussian Process Priors" (with discussion), in *Bayesian Statistics* (Vol. 6), eds. A. P. Dawid, J. M. Bernardo, J. O. Berger, and A. F. M. Smith, New York: Oxford University Press, pp. 475–501. [415]

Richardson, S., and Green, P. J. (1997), "On Bayesian Analysis of Mixtures With an Unknown Number of Components," *Journal of the Royal Statistical Society*, Series B, 59, 731–792. [417]

Rodrigues, J., and Bolfarine, H. (2007), "Bayesian Inference for an Extended Simple Regression Measurement Error Model Using Skewed Priors," *Bayesian Analysis*, 2, 349–364. [412]

Ruppert, D. (2002), "Selecting the Number of Knots for Penalized Splines," *Journal of Computational and Graphical Statistics*, 11, 735–757. [416]

Sacks, J., Schiller, S., and Welch, W. (1989), "Designs for Computer Experiments," *Technometrics*, 31, 41–47. [412]

Shi, J., and Choi, T. (2011), *Gaussian Process Regression Analysis for Functional Data*, Boca Raton, FL: Chapman and Hall. [415]

Smith, M., and Kohn, R. (1998), "Nonparametric Estimation of Irregular Functions With Independent or Autocorrelated Errors," in *Practical Nonparametric and Semiparametric Bayesian Statistics*, eds. D. Dey, Múller, and D. Sinha, New York: Springer-Verlag. [412]

Stefanski, L. A., and Carroll, R. J. (1987), "Conditional Scores and Optimal Scores for Generalized Linear Measurement-Error Models," *Biometrika*, 74, 703–716. [412]

Stein, M., Chi, Z., and Welty, L. (2004), "Approximating Likelihoods for Large Spatial Data Sets," *Journal of the Royal Statistical Society*, Series B, 66, 275–296. [415]

Tang, B. (1993), "Orthogonal Array-Based Latin Hypercubes," *Journal of the American Statistical Association*, 88, 1392–1397. [414]

Vapnik, V. N. (1995), *The Nature of Statistical Learning Theory*, New York: Springer-Verlag. [419]

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. [411]