Annals of Nuclear Energy 52 (2013) 103–112

Contents lists available at SciVerse ScienceDirect

Annals of Nuclear Energy

journal homepage: www.elsevier.com/locate/anucene

A calibration and data assimilation method using the Bayesian MARS emulator

H.F. Stripling [a,*], R.G. McClarren [a], C.C. Kuranz [b], M.J. Grosskopf [b], E. Rutter [b], B.R. Torralva [b]

[a] Nuclear Engineering Department, Texas A&M University, 3133 TAMU, College Station, TX 77843-3133, USA
[b] Atmospheric, Oceanic, and Space Science Department, University of Michigan, 2455 Hayward, Ann Arbor, MI 48109-2104, USA

ARTICLE INFO

Article history:
Available online 20 November 2012

Keywords:
Calibration
Uncertainty quantification
Predictive science
Bayesian MARS

ABSTRACT

We present a method for calibrating the uncertain inputs to a computer model using available experimental data. The goal of the procedure is to estimate the posterior distribution of the uncertain inputs such that when samples from the posterior are used as inputs to future model runs, the model is more likely to replicate (or predict) the experimental response. The calibration is performed by sampling the space of the uncertain inputs, using the computer model (or, more likely, an emulator for the computer model) to assign weights to the samples, and applying the weights to produce the posterior distributions and generate predictions of new experiments with confidence bounds. The method is similar to Metropolis–Hastings calibration methods with independently sampled updates, except that we generate samples beforehand and replace the candidate acceptance routine with a weighting scheme.

We apply our method to the calibration of a Hyades 2D model of laser energy deposition in beryllium. We employ a Bayesian Multivariate Adaptive Regression Splines (BMARS) emulator as a surrogate for Hyades 2D. We treat a range of uncertainties in our application, including uncertainties in the experimental inputs, experimental measurement error, and systematic experimental timing errors. The resulting posterior distributions agree with our existing intuition, and we validate the results by performing a series of leave-one-out predictions. We find that the calibrated predictions are considerably more accurate and less uncertain than blind sampling of the forward model alone.

1. Introduction and Motivation

1.1. Calibration of Uncertain Model Inputs

Calibration is the task of using field data to improve the predictive capability of a simulation model. Field data typically takes the form of experimental measurements and its associated uncertainty and/or variability. The simulation model, which may be used to replicate experimental results or predict the outcome of untried experiments, may be a function of a large vector of inputs. These inputs can generally be classified into two categories: those that correspond to measurable experimental inputs; and those that are empirical, numerical, or otherwise unmeasurable parameters required to close the mathematical model. We denote the former category with $\vec{x}$ and the latter with $\vec{\theta}$; we also make a distinction between $\vec{x}^{(\mathrm{exp})}$ and $\vec{x}^{(\mathrm{sim})}$, the measurable experimental variables and corresponding simulation inputs, respectively.

The correct or most appropriate values for model inputs $\vec{x}^{(\mathrm{sim})}$ and $\vec{\theta}$ are rarely known with full certainty. As an example, consider the task of using the simulation model to replicate field data resulting from experimental inputs $\vec{x}_0^{(\mathrm{exp})}$. The choice of $\vec{x}^{(\mathrm{sim})}$ may be hindered by imperfect or uncertain measurements of $\vec{x}_0^{(\mathrm{exp})}$; more importantly, once a choice for $\vec{x}_0^{(\mathrm{sim})}$ is made, the simulation output may be highly sensitive to the choice of $\vec{\theta}$. The contribution of this work is a method for "tuning" the $\vec{\theta}$ inputs with full regard for experimental, measurement, simulation, and regression uncertainty. The goal is to identify the components of $\vec{\theta}$ to which the simulation output is sensitive and to infer the values of these components that maximize the predictive capability of the model while reducing predictive uncertainty.

Our calibration method is based on the following model equation, which relates a measured quantity of interest (QOI) Y to a simulation $f(\cdot)$(Higdon et al., 2004)

$$Y_{\mathrm{meas}}\left(\vec{x}_i^{(\mathrm{exp})}\right) = F\left(\vec{x}_i^{(\mathrm{exp})}\right) + \epsilon \approx f(\vec{x}_i^{(\mathrm{sim})}, \vec{\theta}). \tag{1}$$

In Eq. (1), $\vec{x}_i \in \mathbb{R}^d, \vec{\theta} \in \mathbb{R}^p, F(\cdot) = Y_{\mathrm{true}}(\vec{x}_i)$, the true experimental response for input settings i, and ϵ is an error term accounting for stochastic behavior of nature, unknown unknowns, and/or experimental measurement error.

Our model equation does not include a model discrepancy term, a statistical function which compensates for systematic model error in the simulation (Kennedy and O'Hagan, 2001). Calibration in the presence of a non-zero model discrepancy is possible. In this

* Corresponding author. Tel.: +1 979 845 4161; fax: +1 979 845 6443.
E-mail addresses: h.stripling@tamu.edu (H.F. Stripling), rgm@tamu.edu (R.G. McClarren), ckuranz@umich.edu (C.C. Kuranz), mikegros@umich.edu (M.J. Grosskopf), ruttere@umich.edu (E. Rutter), bentorra@umich.edu (B.R. Torralva).

0306-4549/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.anucene.2012.08.025

case, however, $\vec{\theta}$ has only been "tuned" (in tandem with the model discrepancy) to improve the predictive accuracy of $f(\cdot)$ *plus* the discrepancy, not calibrated to improve the accuracy of $f(\cdot)$ alone. Our application calls for the use of the full forward model to predict experimental QOIs; therefore we do not allow for a model discrepancy term and seek calibrated $\vec{\theta}$s which improve our predictions and provide some insight into our model.

In order to generate reliable statistics, the calibration routine will require a large number of samples from the $\mathbb{R}^{d+p}$ input space and propagation of those samples through $f(\cdot)$. The cost and high-dimensionality of today's advanced simulations often limits the number of full model runs, leaving much of the input space unexplored and preventing the use of exhaustive sampling for sensitivity analysis. To generate estimates of the simulator response at untried inputs, we employ an emulator, or response surface, $\eta(\vec{x},\vec{\theta})$ to interpolate between these available results. We write this relationship as (dropping the superscript (sim)/(exp) notation)

$$f(\vec{x},\vec{\theta}) = \eta(\vec{x},\vec{\theta}) + \zeta,$$

where ζ is an estimated regression error. Inserting this result into Eq. (1) and rearranging terms, our model equation may also be written as

$$F(\vec{x}_i) \approx \eta(\vec{x}_i,\vec{\theta}_i) + \zeta - \epsilon. \tag{2}$$

The error terms ζ and ϵ are typically well characterized. For our emulator, the regression error ζ is estimated during the construction of the emulator, and ϵ can be calculated or estimated heuristically based on knowledge of the physical system and measurement equipment or instrumentation.

The application of Eqs. (1) and/or (2) becomes more interesting if we treat each term as a *distribution* instead of a single realization. The emulator term becomes a distribution if it is evaluated at a distribution of $\vec{\theta}$ values, and the result is a distribution of predictions for the true experimental response $F(\vec{x})$. Calibration algorithms which search for the most likely distribution of $\vec{\theta}$ have been the subject of extensive research for some time (see a thorough review in Chib and Greenberg, 1995). For example, the so-called random walk calibration algorithm searches the $\vec{\theta}$ space one candidate point at a time, accepting or rejecting each candidate based on a likelihood calculation, and "jumping" between candidates in a prescribed manner. In the end, the distribution of accepted candidates approximates the posterior distribution. Numerous software packages and applications are available for review Higdon et al. (2004).

The work presented here is similar to existing search algorithms with the exception that each sampled candidate in $\vec{\theta}$ space is assigned a weight which reflects its likelihood. This weight is a function of available field data, forward model results, and the uncertainty information that characterizes ζ and ϵ. The goal is to produce a predictive distribution for Y_{true} that is more accurate and less uncertain than blind sampling of the forward model alone.

1.2. Motivating Application: Initialization of Radiative-Shock Simulations

Our motivating application is related to the mission of the Center for Radiative Shock Hydrodynamics (CRASH) at the University of Michigan, one of five centers funded by the DOE Predictive Science Academic Alliance Program (PSAAP). The challenge for each PSAAP center is to use results and data from 4 years of increasingly complex simulations and experiments to predict the results of a "significantly different" or extrapolated 5th year experiment. Most importantly, each center must also develop methods to assess its own predictive capability and report defendable confidence bounds to support its year-5 prediction.

At CRASH, our goal is to simulate the development of a laser-driven shockwave traveling at high Mach number in a gas-filled plastic tube. The shock is formed after a laser strikes and ablates a beryllium disk at one end of the shock tube. The radiation energy traveling down the tube and ahead of the shock affects the shock evolution, resulting in highly non-linear physics. The CRASH code, a massively parallel Eulerian radiation hydrodynamics code with mesh adaptivity, is used to predict experimentally measurable quantities, such as the shock shape and location at specific times after the laser fires.

The CRASH experiments progress as follows: the years 1, 2, and 3 experiments involve only cylindrical tube geometries. The year 4 experiment complicates the tube geometry by "necking" the tube from a cylinder with a larger radius to one with a smaller radius. The year 5 experiment extrapolates the year 4 experiment by necking from a larger cylinder to a smaller but elliptical tube (requiring a fully-3D physics model). Crucial to our ability to extrapolate predictive capability to the year 5 experiment is a thorough understanding of the propagation of input uncertainties through the full physics to the output QOIs. The development of this understanding requires that we calibrate our year 1–4 predictive models and justify the use of that calibration to predict the year 5 experiment.

An interesting calibration question arises in the initialization of the CRASH code. For some time during the CRASH campaign, the CRASH code did not handle the laser energy deposition in the beryllium disk. Instead, we used a 1D (and later a 2D) Lagrangian radiation hydrodynamics code Hyades (Larsen and Lane, 1994) to simulate the energy deposition and initialize the CRASH code. Hyades takes a large number of parameter inputs and computes an even larger number of responses which are used to form the initial conditions for the CRASH code.

A previous study by McClarren et al. (2011) used physics-based arguments and a sensitivity study to filter the large number of Hyades 1D (H1D) inputs down to a manageable dimension. Further, this study found that the predictions made by the CRASH code were highly sensitive to its initialization and therefore highly sensitive to the H1D initialization. The center later moved from H1D to Hyades 2D (H2D) for CRASH initialization, and the need to reduce uncertainty arising from variability in H2D inputs became an imperative for improving the center's predictive capability.

The center designed an experimental campaign to generate field data for calibrating the inputs to H2D. Specifically, the experimental QOI was the shock breakout time (BOT), or the time required for the laser energy to propagate through the beryllium disk. We also generated a run-set of H2D predictions of the shock BOT by sampling from a five-dimensional H2D input space. Two of these inputs directly correspond to experimental variables:

$$\vec{x} = [\text{laser energy}, \text{disk thickness}].$$

Based on the previous sensitivity study by McClarren et al. (2011), heuristics, and computational limitations, three additional uncertain inputs were allowed to vary to generate the run-set:

$$\vec{\theta} = [\text{beryllium EOS gamma}, \text{plastic wall opacity}, \text{electron flux limiter}].$$

The task was to use the shock BOT data and predictions to infer posterior distributions for $\vec{\theta}$. The calibration is successful if H2D predictions of shock BOT become more accurate and less uncertain when the uncertain inputs are sampled from their posterior distribution. The task is complicated, however, by relatively large field data uncertainty and the interpretation thereof. This paper describes the calibration method, our approach for handling these challenges, and gives results which agree with physical intuition and have led to stronger predictive capability.

The outline of this paper is as follows. In Section 2, we describe the BMARS emulator, define and outline our calibration method, and specify the experimental data and simulation results available for our H2D calibration. In Section 3, we exercise our calibration method on first a simplified problem and then with the full uncertainty treatment. We conclude with a discussion of the method's advantages and potential challenges in future work.

2. Description of BMARS, the calibration method, and the application of interest

2.1. The Emulator: Bayesian Multivariate Adaptive Regression Splines (BMARS)

As mentioned in the introduction, an emulator is necessary to interpolate between available simulation samples when the forward model is difficult or expensive to evaluate at untried inputs. For this paper, we choose the BMARS response surface, described as follows.

The original multivariate adaptive regression splines (MARS) algorithm proposed by Friedman (1991) is a partition-based curve-fitting technique which attempts to emulate the mapping between a function's inputs and outputs as a summation of so-called "spline" functions. Multivariate spline functions are simply products of one-dimensional spline functions; these 1D spline functions are continuous and defined to be zero on part of the domain and a polynomial of some order on the remainder of the domain. The knot of the spline is the coordinate at which this definition changes, and the direction of the spline describes whether the non-zero portion of the spline is in the positive or negative direction from the knot.

Given a set of input or training data, the original formulation uses a semi-stochastic method to generate a basis function of the form

$$B(x) = \beta_0 + \sum_{k=1}^{\mathbf{K}} \beta_k \prod_{l=0}^{\mathbf{I}} (x_l - t_{k,l})_+^{o_k}, \tag{3}$$

where $\vec{x}$ is a vector of inputs (in our case, $x \in \mathbb{R}^{p+d}$), $t_{k,l}$ is the knot point in the l^{th} dimension of the k^{th} component, the function $(y)_+$ evaluates to y if $y > 0$, else it is 0, o is the polynomial degree of the kth component, β_k is the coefficient of the kth component, $\mathbf{K}$ is the number of components of the basis function, and $\mathbf{I}$ is the maximum allowed number of interactions between the $p + d$ dimensions of the input space. Note that the formulation does not require that each of the k components have a term in each dimension of $\vec{x}$ and that the optimization does penalize as the size of model, $\mathbf{K}$ becomes larger.

Denison et al. (1998) introduced a Bayesian extension of MARS (hereby named BMARS) which attempts to converge on a posterior distribution of predictive MARS functions. Then, each sample of the emulator results in a predictive distribution of the response instead of a single prediction. Step zero of the algorithm generates a classical basis function of the form (3). A Markovian process proposes a change to the current model in the form of an addition, deletion, or modification of a spline. When a new spline is created, the algorithm randomly chooses its order, knot point, direction, and level of interaction. The algorithm iterates this random selection process and accepts/rejects proposal basis functions based on a likelihood calculation. This likelihood is a function of the candidate's fit to the training data and the number of splines in its basis function. The coefficients, β, are found using a Bayesian least-squares inversion. Similarly, the error in the regression (ζ in Eq. (2)) is estimated at each iteration based on the current model's approximation of the training data.

In many respects, BMARS is different from other popular emulation techniques. First, the function is not an interpolator, meaning that it does not reproduce the training data exactly, and it does not have a closed form for the variance in its predictions; Guassian processes (Rasmussen and Williams, 2006) and polynomial chaos (Ghanem and Spanos, 1991) each have both of these features. We argue, however, that interpolation in an environment of large uncertainties is not essential so long as the regression error can be shown to be small or can be accounted for correctly. In this work, we show that the regression error is small compared to experimental uncertainties and that the Bayesian variance approximation is accurate.

BMARS has a number of advantageous features. The computational work to form the emulator (that is, solve for the β coefficients) at each iteration grows as $\mathbf{K}^3$, the cube of the number of splines in the model. This is generally much smaller than the computational work required to form interpolating emulators, which grows as n^3, the cube of the size of the training set. We generally find that $\mathbf{K}$ is an increasingly smaller fraction of n as n grows large. Because of the discontinuous nature of the spline functions, BMARS does not require an assumption of smoothness of the response in parameter space; regression over sharp discontinuities is still difficult, but BMARS is more apt to fit such features than globally continuous functions. In terms of accuracy, Denison et al. (1998) report that BMARS is competitive across a standard suite of both low and high dimensional regression tests.

We illustrate the utility of this emulator in Fig. 1, where we show a one-dimensional BMARS fit to piecewise-linear training data perturbed by a $N(0, 1.5^2)$ error term. The classical MARS algorithm would find just one of these predictive functions, while the Bayesian extension produces a number of candidate predictive functions. The distribution (mean and variance) of these predictive functions provides the modeler with an estimate for the regression uncertainty and unexplained variance in the forward model.

2.2. The calibration method: A sampling and filtering algorithm

The goal of calibration is to find values or distributions of an approximate model's inputs such that the approximate model becomes a more accurate predictor of reality. As we outlined in the introduction, an experiment and its approximate model typically share a set of inputs, $\vec{x}$. The simulation will also have a number of other inputs, $\vec{\theta}$, which can take the form of non-physical tuning parameters or other physical inputs which are not experimentally

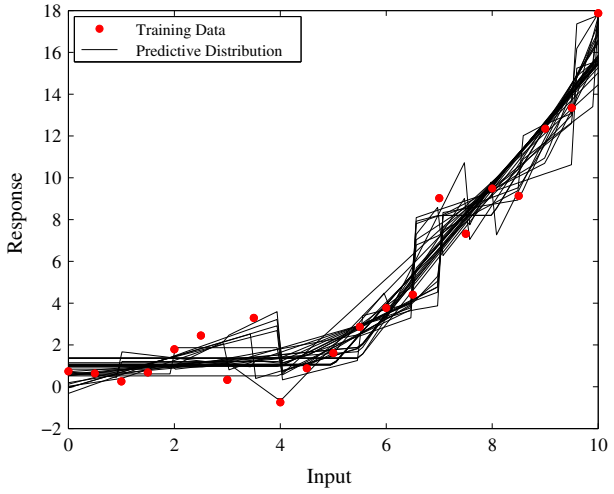


Fig. 1. An Example 1D BMARS fit to piecewise-linear data perturbed by a noise term.

varied. We seek distributions of these uncertain parameters such that the simulation produces a more accurate prediction when $\vec{x}^{(\mathrm{sim})} = \vec{x}^{(\mathrm{exp})}$.

We propose a two-step calibration procedure. Step one is to generate a BMARS model for the simulator, $f(\cdot) = B(\cdot) + \zeta$. The model will be a function of $d + p$ variables; that is, we make no distinction between independent and uncertain inputs in step one. Determining the adequacy of the emulator is an independent subject (see studies by Currin et al. (1991) and Sacks et al. (1989)); we note that the following analysis operates under the assumption that the emulator is an adequate representation of the simulator, or at least that it appropriately estimates the regression error ζ.

Step two relates the BMARS emulator to the available experimental measurements resulting from inputs $\{x_i\}_{i=1}^{I}$. We outline step 2 here and discuss the procedure in more detail in the next section of this paper.

1. Generate N samples of the uncertain input space. Samples should be contained within the convex hull of the available simulation runs to avoid extrapolating with the emulator and should be as dense as possible.
2. For each available experimental data point and each sample of the uncertain input space (*i.e.* for $i = 1 \cdots I$ and $n = 1 \cdots N$):
   (a) Randomly choose M indices from the posterior distribution of the BMARS model.
   (b) Generate and normalize a discrete probability distribution function, $g(\eta(\vec{x}_i, \vec{\theta}_n))$ by evaluating $\eta_m(\vec{x}_i, \vec{\theta}_n) = B_m(\vec{x}_i, \vec{\theta}_n) + \zeta_m, m = 1 \ldots M$.
   (c) Compute a local weight, $\omega_{i,n} = L(Y_{\mathrm{meas}}(x_i)|g,\epsilon)$. $L(\cdot)$ is a likelihood function which should return a measure of the accuracy of the simulator in predicting experiment i when using uncertain input $\vec{\theta}_n$ given the uncertainty in the regression model and field data.
3. Compute the global weight, $w_n = \prod_{i=1}^{I} \omega_{i,n}$.

The result of the algorithm is a global weight assigned to each $\vec{\theta}_n \in \mathcal{R}^p$, $n = 1 \cdots N$ sample of the uncertain input space. The weight w_n should be proportional to the likelihood that the simulator will replicate experimental results if it is run using $\vec{\theta}_n$ as its inputs. The global weight is a product over the likelihoods computed using each of the I experimental data points. Therefore, those $\vec{\theta}$'s with the largest global weight will be those that most consistently approximate the experimental data when pushed through the simulator.

The weights may also be used to generate posterior distributions in each dimension of $\vec{\theta}$. We choose to discretize the range of each uncertain input into "bins" containing an equal number of samples. Then, for each bin, we simply sum the global weights assigned to the samples in that bin. This distribution can be normalized to represent the posterior distribution. An uncertain input is said to be strongly calibrated if this posterior distribution is markedly different from the assumed prior distribution.

Finally, the modeler may apply the weights to generate predictions (with uncertainty) of new experiments using either the full simulation model or its emulator. If a few runs of the simulator are possible, the modeler will sample appropriately from the calibrated distributions to initialize the code. The type and quality of QOI statistics will be limited by the number of runs that can be afforded.

Alternatively, the modeler may choose to use the response surface to generate a large number of predictions, but will pay the price of regression uncertainty resulting from the use of the emulator. In this case, the expected value of an experiment is a weighted average:

$$\overline{Y}_{\mathrm{pred}}(x_i) = \frac{1}{M}\sum_{m=1}^{M}\sum_{n=1}^{N}\eta_m(\vec{x}_i, \vec{\theta}_n)w_n.$$

Each posterior BMARS realization includes a Gaussian estimate for its regression error, $\zeta_m \sim N(0, \chi_m^2)$. Therefore, the variance of the uncertain distribution about the globally weighted prediction for $\overline{Y}_{\mathrm{pred}}(x_i)$ can be estimated as a weighted average of the χ_m^2s.

Step 2(c) of the algorithm above will require interpretation of the specific application; that is, the modeler must choose which uncertainties to include in the analysis and how to properly account for them in the likelihood distribution $L(\cdot)$. At a minimum the system will have uncertainty from emulator regression and experimental measurement. As mentioned above, the BMARS regression error is estimated as a normal distribution, $\zeta \sim N(0, \chi^2)$. If the measurement error is also estimated as normal ($\epsilon \sim N(0, \tau^2)$), then the function L is equivalent to evaluating the normal pdf:

$$L(\vec{\theta}_{i,n}) \propto N\left(Y_{\mathrm{meas}}(x_i)\middle|\mu = \frac{1}{M}\sum_{m=1}^{M}B_m(\vec{x}_i, \vec{\theta}_n) + N(0, \chi_m^2), \sigma^2 = \chi^2 + \tau^2\right) \tag{4}$$

Other sources of uncertainty or distributions of those uncertainties will require the modeler to generate a tailored likelihood function. For example, the field data in our application is subject to a systematic, uniform timing error that is both large in magnitude and difficult to interpret. We outline our approach to handling this uncertainty in the following sections. In general, the form of $L(\cdot)$ will require careful consideration of the specific uncertainties and/or error present in the application.

2.3. The Application: Calibrating Hyades 2D to Experimentally Measured Shock Breakout Times

As described in Section 1.2 and summarized in Fig. 2, we seek to calibrate three uncertain inputs to the Hyades 2D laser deposition model. The results of this model are used to initialize the CRASH code, which is then used to generate predictions of late-time shock location and structure. Because these predictions are sensitive to the early-time energy deposition, an accurate and consistent initialization from H2D is required for accurate predictions.

A series of eight experiments were carried out on the OMEGA laser in the Laboratory for Laser Energetics at the University of Rochester. For each experiment, we took three measurements (using three different diagnostics: asbo1, asbo2, and sop) of the time required for the laser energy to propagate through a beryllium disk (this is the shock breakout time). Only the laser energy and disk thickness were experimentally varied. Table 1 summarizes the results of the eight shots.

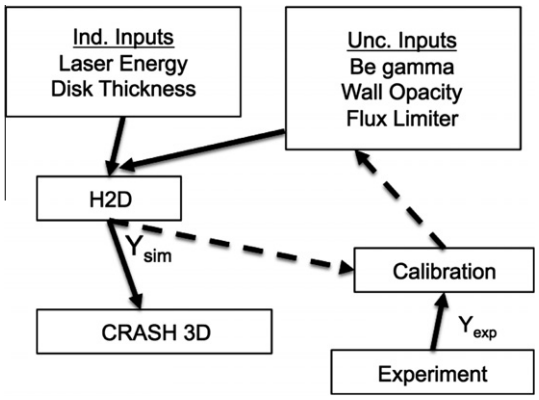


Fig. 2. Schematic of H2D/CRASH interaction.

Table 1
Experimental inputs and resulting shock breakout time measurements.

Experiment	Be disk thickness (μm)	Laser energy (J)	asbo1 (ps)	asbo2 (ps)	sop (ps)
1	21	3837.6	504	486	462
2	20	3925.2	467	475	430
3	20	3937.6	437	450	None[a]
4	19	3887.8	419	410	436
5	20	3914.6	425	467	456
6	20	3912.8	442	476	470
7	19	3923.3	447	456	470
8	19	3945.8	410	417	418

[a] No sop data recorded for this shot.

Table 2
Hyades 2D simulation input ranges and distributions.

Input	Distribution
Laser energy (J)	$U[3610, 3990]$
Disk thickness (μm)	$U[18, 22]$
Be gamma	$U[1.4, 1.75]$
Wall opacity	$U[0.7, 1.3]$
Flux limiter	$U[0.05, 0.075]$

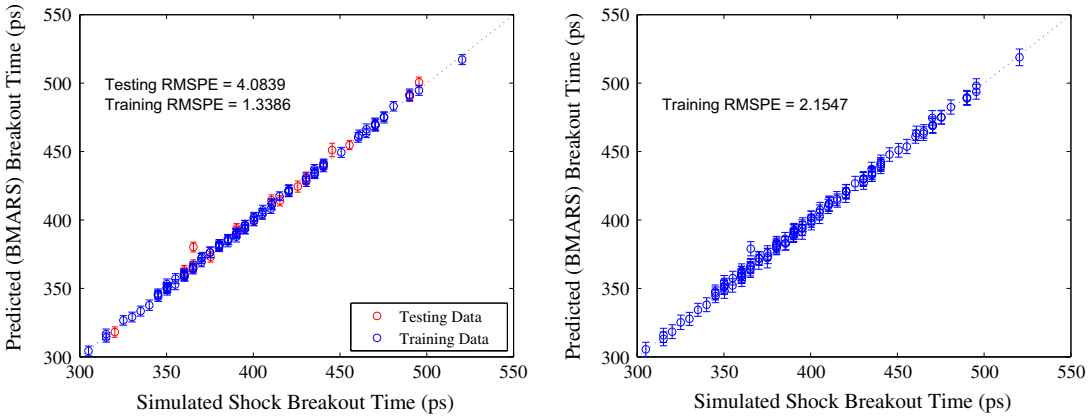
Each column of Table 1 has an associated uncertainty, which we describe as follows:

1. *Disk thickness*: the micrometer that measured the disk thickness reports only to the nearest micron; therefore we assume a $U[\pm 0.5\ \mu m]$ distribution about the reported value.
2. *Laser energy*: The total laser energy delivered to the beryllium disk is known very well *after* the shot; before the shot, however, the energy can only be estimated to within some range of the energy requested by the experimenter. Therefore, when predicting a response to a future experiment, this uncertainty must be included. During calibration, however, we can take the laser energy as a known parameter, which should help us reduce uncertainties on other parameters.
3. *Diagnostic uncertainties*: the asbo1, asbo2, and sop diagnostics have an intrinsic precision uncertainty reported as 10 ps, 20 ps, and 30 ps respectively. The best documentation indicates that these should be interpreted as standard deviations of a normal distribution. Also, since each is nominally a measurement of the same quantity, there is certainly some correlation between the errors in each diagnostics' reading of a given experiment.
4. *Systematic timing error*: we are also told that the facility has a systematic firing error of 50 ps which is to be interpreted as a top-hat distribution about any given reported value. In other words, the true breakout time is equally likely to be any value within 50 ps of the reported value of any diagnostic. This is a large uncertainty: on the order of 10% of the QOI!

For simulation data, we have 104 successful runs of H2D resulting from 104 samples from the 5 dimensional hypercube in $[\vec{x}, \vec{\theta}]$ space. Table 2 gives the input variables and the ranges in which they were varied to generate this run set. The independent variables, laser energy and disk thickness, again correspond to the experimentally varied variables to generate the field data. The uncertain variables are the beryllium equation-of-state parameter, opacity of the plastic in the shock-tube wall, and electron flux limiter regulating electron diffusion within the H2D code. These are the variables for which we seek a calibrated distribution.

Step one of the calibration procedure is to generate a response surface which can be cheaply sampled to provide estimates for the simulator response at untried inputs. One method for determining the adequacy of a chosen response surface is cross-validation, or training on a subset of the available data and predicting the rest. For example, we performed a series of emulator constructions wherein we trained the BMARS model on a random subset of 84 of the 104 available data points. We then used the resulting response surface to predict the "left out" 20 data points. After some tuning, we consistently found that the root-mean-square error in the 20 predictions was consistently on the order of 1%, which is *much* less than the experimental uncertainty in this particular problem. A final step is to formulate (and self-validate) a final BMARS model using all 104 data points. Figs. 3a and b illustrate an example cross-validation fit and the final fit, respectively.

In the following section, we give results of two calibration exercises: one with only measurement uncertainty and one with the full uncertainty treatment. We first show the posterior distributions resulting from the calibration using all eight experiments. We then give the results of 8 "leave-one-out" (L1O) tests/predictions. For a given L1O prediction, we will use the field data from 7 of the 8 experiments and all 104 H2D simulation runs to perform the calibration method. We then use the resulting calibrated $\vec{\theta}$s to predict the result of the 8th experiment. We will perform this 8



(a) Validation of the BMARS fit by leaving out then predicting 20 of the 104 H2D results.

(b) BMARS fit to all 104 H2D runs. This is the model used for calibration.

Fig. 3. We evaluated the accuracy of the BMARS emulator by performing a series of cross-validation tests. Note that in each plot, a perfect prediction would fall on the 45° line.

times, each repetition leaving out a different experiment's field data. In the end, we will have generated predictions for each of the 8 experiments. If successful, the L1O tests will provide some sense of validation of the method and confidence in subsequent predictions derived from its results.

3. Results of the Calibration of Hyades 2D

3.1. Calibration of a Simplified Problem

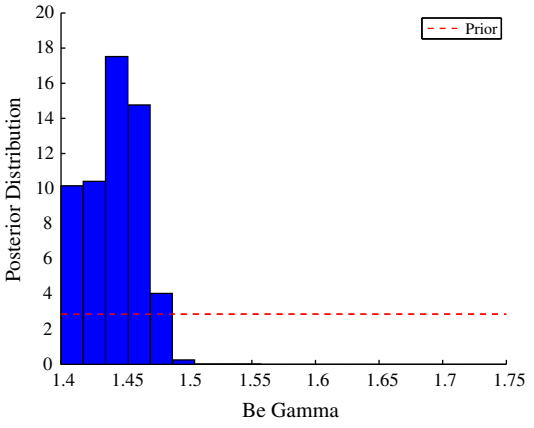
As an initial proof-of-concept of the calibration method, we first consider the case of zero regression error and a normally distributed measurement uncertainty model. We will only consider the asbo1 measurements, which have an associated measurement error term $\epsilon \sim N(0, 10\ \text{ps}^2)$.

Following the calibration steps, we generated a latin hypercube sample of 50,000 points in the three-dimensional $\vec{\theta}$ space. As we are only considering one diagnostic and its normally-distributed measurement error, our local likelihood distribution is as written in Eq. (4) with $\chi^2 = 0$. Fig. 4 shows the posterior distributions of the three components of $\vec{\theta}$ when they are calibrated to *all* of the asbo1 measurements.

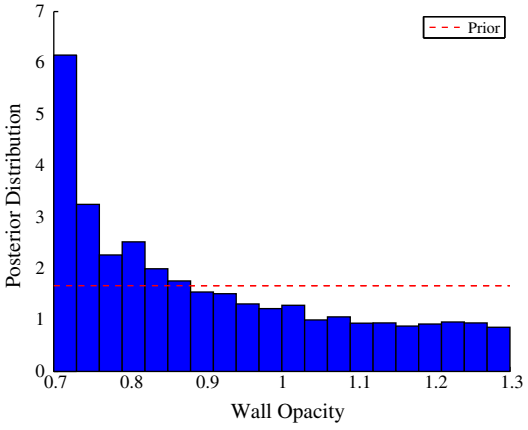
The immediate observation is that all three components calibrated to the lower end of their prior distributions and that the beryllium gamma input did so quite strongly. The interpretation of these plots is that the H2D simulations are more likely to replicate the experimental measurements when the uncertain inputs are sampled from the lower end of their priors. In other words, the calibration procedure is indicating that the true distribution of these parameters is not uniform (as were our prior guesses), but instead have more weight near the lower values of the inputs.

We may also seek to interpret the results and develop some intuition about our physical model. For example, the posterior distribution of the electron flux limiter may be indicating that H2D is over-driving the heat flux at the shock front. Similarly, the strong calibration of the Be-gamma input indicates a preferred bias of the equation-of-state by H2D. In some cases, however, physical intuition may disagree with the calibration results. For example, it is not clear that the wall-opacity should be a driving input for the shock BOT because the physics resulting from wall interactions are mostly downstream from the laser ablation.

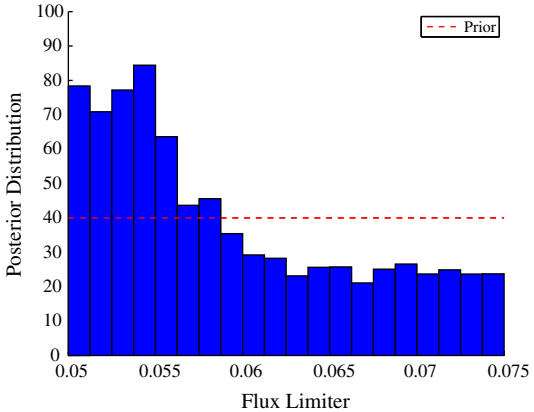
As mentioned, we provide some validation of our calibration method via leave-one-out experiments: we calibrate on seven of the experiments and attempt to use the resulting posteriors to predict the "left out" experiment. One manner by which to measure success is to compare these predictions to predictions that result from blind sampling of the $\vec{\theta}$ space. If the method is successful, we will improve the prediction of each experiment and reduce the confidence interval about that prediction. Fig. 5 compares the calibrated and uncalibrated predictions.



(a) Posterior distribution of the beryllium gamma EOS parameter



(b) Posterior distribution of the wall opacity



(c) Posterior distribution of the electron flux limiter

Fig. 4. The calibrated or posterior distributions of $\vec{\theta}$ in the simplified case.

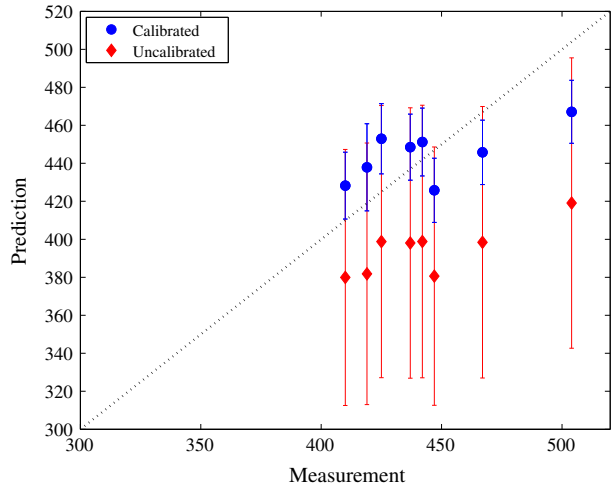


Fig. 5. The calibration routine improved the accuracy and reduced the uncertainty in experimental predictions. Note that a perfect prediction would fall on the 45° line.

First we note that the predictive accuracy was greatly improved when we use calibrated $\vec{\theta}$s to generate predictions. The figure also indicates that the *a priori* input ranges result in a general under-prediction of the experimental response by H2D. Recall that our model assumes that a discrepancy function (to account for systematic differences between the simulation and reality) is not necessary – in this case, we are very close to needing a discrepancy function to predict the asbo1 measurements. Indeed, one asbo1 measurements (~500 ps) is not encapsulated by the simulation runs. In the following section, we'll show that the added information from the inclusion of all three diagnostics aides in the predictive accuracy and moves us away from nearly needing a discrepancy function.

3.2. Leave-one-out Experiments with Full Treatment of Uncertainty

We now perform the same procedure of leave-one-out experiments but include the regression error ζ and the full treatment of the uncertainties described in Section 2.3. These uncertainties combine to form the likelihood function $L(\cdot)$ for each combination of experiment number and $\vec{\theta}$ sample. The manner in which they combine depends on the modelers understanding and interpretation of the uncertainties.

Previous and related studies from the CRASH center (Holloway et al., 2011; McClarren et al., 2011) have characterized the experimental uncertainties and their relative significance for predictive capability. These authors report that the largest source of error is the 50 ps systematic timing bias of the entire experimental apparatus, followed by the specific biases of the three individual diagnostics. We reflect these findings by taking conservative estimates (that is, erring on the side of more uncertainty) in constructing the likelihood function, $L(\cdot)$. Our final interpretations of the uncertainties are as follows:

1. *Regression error, ζ*: This is estimated to be normally distributed, $\zeta \sim N(0, \chi^2)$, where χ^2 is the average of the χ^2_m estimates from the posterior BMARS realizations. For our application, we find these estimates to be accurate using cross-validation studies as described in Section 2.3. The regression variance acts to widen the likelihood distribution.
2. *Disk thickness error*: The true disk thickness corresponding to any experiment is estimated to be within ±0.5 μm of the value reported in Table 1. We include this uncertainty by randomly sampling a value (uniformly) about the nominal value at each evaluation of the BMARS emulator. This results in a widening of the emulator's prediction of the simulation response (a net addition to ζ).
3. *Laser energy*: When calibrating, we take the laser energy to be known exactly (which is the case *post*-shock). In a later section, when we use the calibration to make predictions, the laser energy uncertainty must be included.
4. *Diagnostic uncertainties*: Each measurement diagnostic has an associated uncertainty, which we interpret as standard deviations of a normal distribution about the reported measurement. There are a number of ways by which we could combine the information provided by the three measurements of each experiment. Due to the overwhelmingly large systematic timing error (described next), we decided that the most conservative approach was to take the true experimental response as the mean of the diagnostics and to assume the worst-case standard deviation of 30*ps* for the nominal measurement error. This 30*ps* is reflected in the τ^2 term in $L(\cdot)$.
5. *Systematic timing error*: The entire experimental apparatus is estimated to have a ± 50 ps systematic timing error. We interpret this as follows: if the mean prediction of $Y_{sim}(x_i, \vec{\theta}_n)$ falls outside a 50 ps bound about any credible estimate of Y_{true} (accounting for uncertainties), then the weight $w_n = 0$. For experiment i, the bound of credible estimates of $Y_{true,i}$ is

$$\max_k\{Y_{meas,i,k} + 3\sigma_k\} - 50ps \leqslant Y_{true,i} \leqslant \min_k\{Y_{meas,i,k} - 3\sigma_k\} + 50ps$$

where k indexes the diagnostics, {k = asbo1, asbo2, sop}. To implement this constraint, we multiply the existing likelihood distribution by a top-hat distribution valued at 1 in the acceptable range and 0 elsewhere.

The result of this interpretation is a truncated normal likelihood distribution for each combination of $Y_{\text{meas}}(x_i)$ and $\vec{\theta}_n$ candidate. The mode of this distribution is the mean BMARS prediction of the simulator response, and the variance is a function of the regression error, disk thickness uncertainty, and diagnostic measurement error. An example likelihood distribution is given in Fig. 6. The weight assigned to $\vec{\theta}_n$ will be the value of the solid line at $Y_{\text{meas}}(x_i)$. This solid line follows a normal distribution but is truncated to value zero outside the acceptable predictive range.

We again use a 50,000 sample latin hypercube design to generate the $\vec{\theta}$ candidates. The posterior distribution estimates resulting from a calibration on all 8 experimental measurements are given in a bi-variate distribution plot in Fig. 7. Plots on the diagonal give the

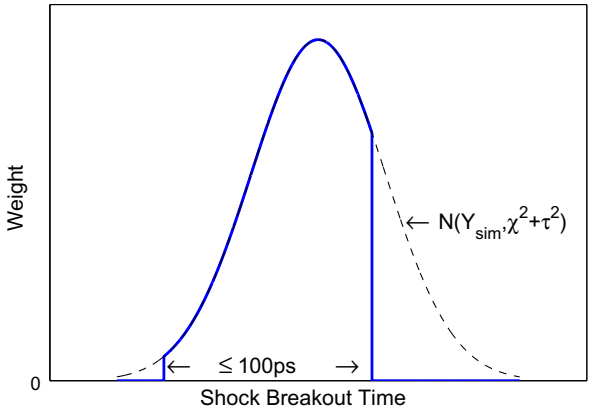


Fig. 6. An example likelihood representing all known uncertainties for our CRASH calibration. Note that the range of the non-zero valued likelihoods varies for each experiment but is always less than 100 ps.

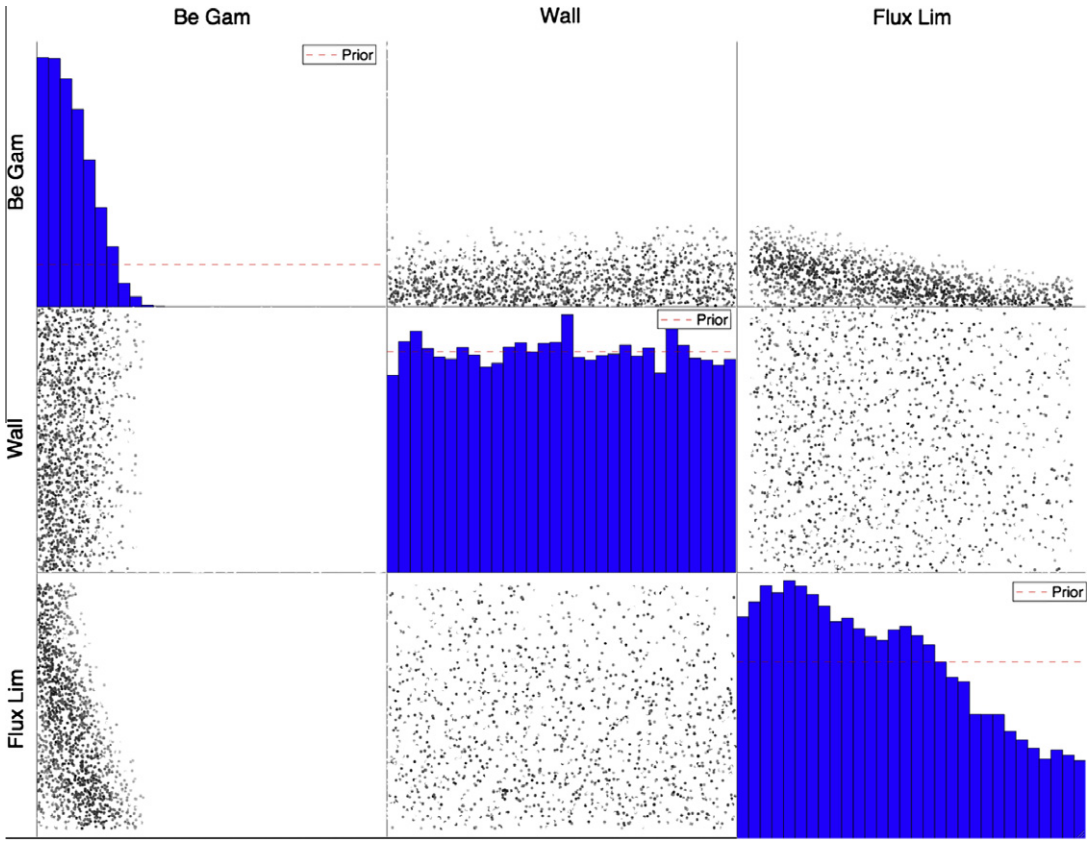


Fig. 7. Posterior distribution estimates for a calibration on all 8 experiments.

univariate posterior histograms and the value of the uniform prior distributions. The off-diagonal plots give the bivariate distributions, and the darkness of the data points is proportional to the magnitude of that point's global weight. We plot only a subset of the data for clarity in the figure.

The full uncertainty treatment resulted in weaker calibration for both the wall opacity and the electron flux limiter. As previously mentioned, we do not expect the wall opacity to be a driving input for BOT calculations because the wall is "downstream" from the laser energy deposition. The wall opacity posterior given in Fig. 7 is not statistically different than the prior distribution. The posterior distribution for the electron flux limiter indicates that smaller values of this parameter result in more accurate experimental predictions. This also agreed with our existing intuition, as we believe that Hyades is over-driving heat transfer at the shock front.

It is interesting that moving to the full treatment of uncertainty did not have a large effect on the estimated posterior of the beryllium gamma constant. We certainly expected this to be a driving input because the laser energy is ablating the beryllium, and we generate material properties using a γ-law equation of state. Thus, it seems intuitive that the value of the gamma constant would play a strong role in the calculation of shock breakout time. The fact that it remained strongly calibrated to the lower part of the prior indicates that the shock breakout time is highly sensitive to its value. This result is inconclusive about the true posterior distribution of this input; instead we only learned that our prior distribution was mostly too high and that future samples of this parameter should extend below the lower bound of 1.4 considered in this study.

The results of the leave-one-out predictions under the full uncertainty treatment are illustrated in Fig. 8. The plot shows 8 pairs of horizontal and vertical error bars. The intersection of each pair of error bars is at {x = nominal experimental measurement, y = mean experimental prediction}. The vertical error bars represent the 90% predictive confidence interval, and the horizontal error bars represent the bounds of possible values of Y_{true}, determined by the ±50 ps systematic timing uncertainty. To give an idea for the spread in the diagnostic measurements, each set of horizontal error bars contains the three (in one case two) diagnostic results.

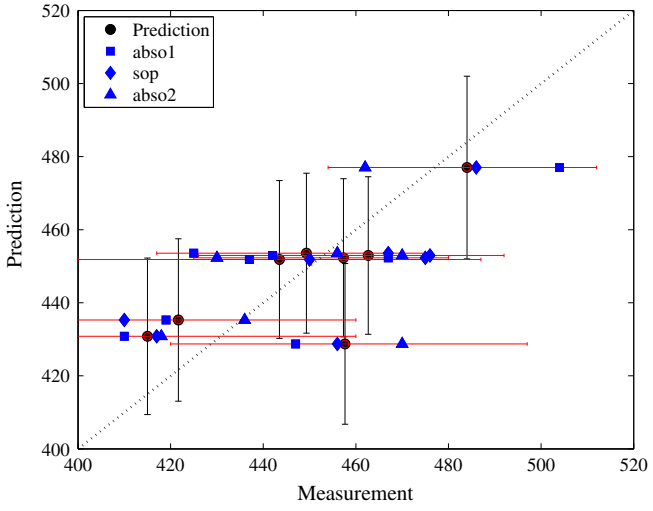


Fig. 8. Leave-one-out predictions under the full treatment of uncertainty.

Several differences between these predictions and those in Fig. 5 are noteworthy. First, as expected, the 90% predictive confidence intervals are larger, as expected, because the predictions are less certain under the full uncertainty treatment. Second, despite the inclusion of multiple sources of uncertainty, the root-mean-square predictive accuracy improved from 22.3 ps to 16.2 ps. As we alluded to above, we believe this is a result of using the mean of all three diagnostics as the target calibration value. Because the systematic timing error of ±50 ps is dominating this study, the measurements of an individual diagnostic are highly variable. In using an average of all three diagnostics, we were better able to characterize the true experimental response, and the Hyades emulator was able to more consistently predict those values. Finally, we assert that the predictions are fairly accurate and a large improvement over blind sampling of $\vec{\theta}$ to predict Y_{meas}. The "box" of uncertainty about each prediction contains the true experimental measurement.

3.3. Predictions of New Shock Breakout Time Experiments

The accurate prediction of the response of new experiments is the goal of any calibration method, and it is natural to extend the methods used to generate our leave-one-out predictions to predictions of the shock breakout time at new laser energies and/or disk thicknesses. In input space, the domain of valid predictions is at most the $\mathbb{R}^5$ space contained by the original 104 H2D input samples. Prediction outside this range would be an extrapolation of the emulator. We also must note that extrapolation outside the range of the experimental data requires an assumption that $\vec{\theta}$ would calibrate similarly at such an $\vec{x}$ *and* that H2D remains a valid model outside the range of the experimental data. The degree to which such assumptions are valid are usually a matter of expert judgement.

We generate predictions using the BMARS emulator, the 50,000 samples in $\vec{\theta}$ space, and the global weights computed by the calibration routine. As we are now predicting, we must include the laser energy uncertainty. Before a shot, the facility estimates $E_{actual} \sim N(E_{requested}, 19.4J^2)$. Therefore, as with the disk thicknesses, each time we evaluate a BMARS model, we randomly sample a number from this distribution as the true experimental input. This should add considerable width to our predictive confidence bounds. Figs. 9 and 10 show predictions of shock breakout time as a function of disk thickness and laser energy, respectively.

The figures indicate that in the range of inputs tested here, breakout time is a stronger function of disk thickness than of laser energy. The predictions are consistent with the experimental data (and intuition) in that breakout time increases with increasing disk thickness and decreasing laser energy. Our predictions mostly contain the experimental data points (recall that we calibrated on the mean of three measurements per experiment, so our predictions really trace the mean of each experiment). At new experiments, we see that our predictions are fairly smooth and extrapolate the general behavior of the experimental data points.

It is interesting to consider the use of the calibration results to inform both computer and physical experiment design. For example, in Fig. 9 we see evidence (increasing confidence bounds) of loss of predictive reliability at the larger disk thicknesses. This is an indication that either (a) the emulator is not achieving a good fit at these values and additional simulator runs may be required, or (b) the calibration results are not informed by the experimental data (indeed, these predictions are an extrapolation of the experimental data) and additional experiments should be planned at these inputs. The literature contains extensive studies of the design and analysis of computer experiments (e.g., Higdon et al., 2004; McKay et al., 1979; Santer et al., 2003). Literature regarding the impact of calibration on physical experiment design is less complete.

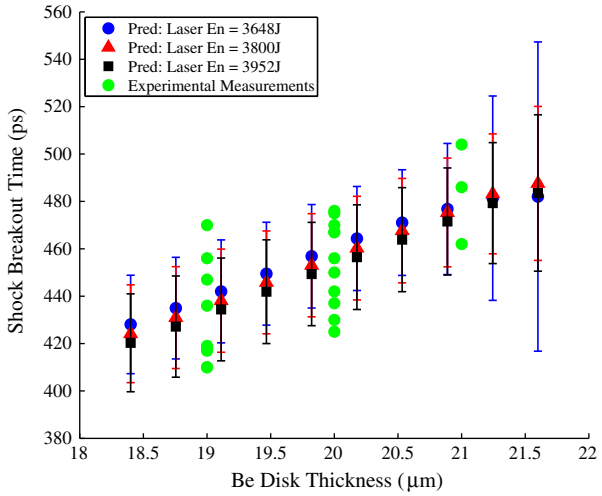


Fig. 9. Predictions of breakout time at new disk thicknesses for three different laser energies.

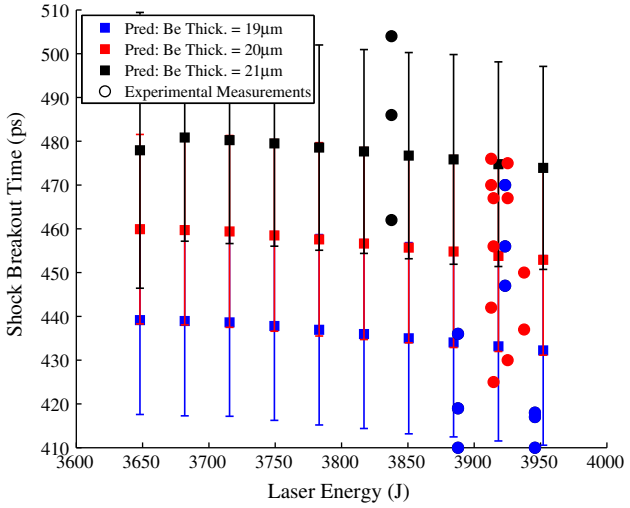


Fig. 10. Predictions of breakout time at new laser energies for three different disk thicknesses. In this figure, the color of each experimental data point corresponds to the disk thickness for that experiment, as indicated by the legend. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Authors generally agree that more experimental data is better and that calibration results are only valid in the vicinity of the design space near the existing experimental results. They often further assume that experimental data is a limited resource, due to historical, cost, or availability issues. As we have defined $\vec{\theta}$ as a set of uncertain inputs that cannot be experimentally varied, we do not believe that this method provides additional insight into the design of physical experiments.

4. Conclusions

We have outlined a method for calibrating the uncertain inputs to a computer model using experimentally measured data. The method requires sampling of the uncertain parameter space, and each of these samples is evaluated using a likelihood function which may be tailored to represent measurement and/or model error, physical input uncertainties, and other sources of uncertainty in the particular application. Most problems will require an emulator to generate estimates of the simulator response at untried

inputs and will also need to account for regression error. The weights that are generated by the algorithm are used to generate posterior distributions and weighted predictions (with confidence intervals) of new experiments.

We applied the method to a calibration of the Hyades 2D laser deposition model using experimentally measured shock breakout times. We employed the use of the BMARS emulator and tailored our weighting computation to include the realistic uncertainties involved with our experimental data. The results of leave-one-out type predictions indicated that the method greatly improved predictive accuracy and tightened the confidence intervals about those predictions. We used the successful leave-one-out predictions as justification to extend the method to produce predictions (and confidence bounds) of new experiments.

This method has a number of benefits. First, it is relatively straightforward to implement and could be useful for preliminary input-sensitivity and dimension reduction studies. Second, the method is analogous to existing calibration routines, but includes a measurement of the importance of the sampled uncertain inputs. This importance has a physical meaning which can be interpreted in the context of the application. The method is also flexible in the choice of emulator, the construction of the likelihood function, and to some extent the number of input variables which are being calibrated.

The method also has a number of areas for exploration and extension. For example, we only considered uniform priors in $\vec{\theta}$ space under the assumption that we had no prior information about our uncertain inputs. If some prior information can be estimated, then the modeler may consider alternate sampling strategies or a modification of the computed weights to reflect the likelihood of the sample given the prior distributions. We also hope to explore the calibration framework using other emulators, such as Gaussian processes. It is likely true that the nature of the physics model (in our case, laser energy deposition) will determine the most effective choice of emulator and ultimately the accuracy of the calibration.

This method may suffer heavily from the curse of dimensionality: the exponential decrease in sample density as the dimension of the uncertain inputs increases. For this application, we used a previous study to reduce the uncertain dimension to a manageable number ($p = 3$), which allowed for extremely dense sampling of the $\vec{\theta}$ space and a very large number of BMARS evaluations at each sample. The computational cost would obviously increase with the dimensionality of the $\vec{\theta}$ space; the evaluation of each $\vec{\theta}$ sample, however, is independent of all others, providing avenues for the use of parallel computing. The modeler may also use physics-based arguments to sample less frequently in non-driving input dimensions.

Acknowledgments

The first author is supported by the Department of Energy Computational Science Graduate Fellowship program under Grant No. DE-FD02-97ER25308. The first author also wishes to acknowledge a number of members of the research staff at Lawrence Livermore National Laboratory for helpful conversations regarding model calibration: P. Dykema, S. Brandon, G. Johannesson, and B. Johnson. The work of the other authors was supported by the DOE NNSA/ASC under the Predictive Science Academic Alliance Program by Grant No. DEFC52-08NA28616.

References

Chib, S., Greenberg, E., 1995. Understanding the Metropolis–Hastings algorithm. The American Statistician 49, 327–335.

Currin, C., Mitchell, T., Morris, M., Ylvisaker, D., 1991. Bayesian prediction of deterministic functions, with applications to the design and analysis of computer experiments. Journal of the American Statistical Association 86, 953–963.

Denison, D., Mallick, B., Smith, A., 1998. Bayesian MARS. Statistics and Computing 8, 337–346.

Friedman, J., 1991. Multivariate adaptive regression splines. The Annals of Statistics 19, 1–67.

Ghanem, R., Spanos, P., 1991. Stochastic Finite Elements: A Spectral Approach. Springer-Verlag, New York.

Higdon, D., Kennedy, M., Cavendish, J., Cafeo, J., Ryne, R., 2004. Combining field observations and simulations for calibration and prediction. SIAM Journal of Scientific Computing 26, 448–466.

Holloway, J., Bingham, D., Chou, C., Doss, F., Drake, R., Fryxell, B., Grosskopf, M., Holst, B., Mallick, B., McClarren, R., Mukherjee, A., Nair, V., Powell, K., Ryu, D., Sokolov, I., Toth, G., Zhang, Z., 2011. Predictive modeling of a radiative shock system. Reliability Engineering and System Safety, 1184–1193.

Kennedy, M., O'Hagan, A., 2001. Bayesian calibration of computer models. Journal of the Royal Statistical Society (Series B) 63, 425–450.

Larsen, J.T., Lane, S.M., 1994. HYADES: a plasma hydrodynamics code for dense plasma studies. Journal of Quantitative Spectroscopy and Radiative Transfer 51, 179–186 (Special Issue Radiative Properties of Hot Dense Matter).

McClarren, R.G., Ryu, D., Drake, R.P., Grosskopf, M., Bingham, D., Chou, C.C., Fryxell, B., van der Holst, B., Holloway, J.P., Kuranz, C.C., Mallick, B., Rutter, E., Torralva, B.R., 2011. A physics informed emulator for laser-driven radiating shock simulations. Reliability Engineering and System Safety 96, 1194–1207 (Quantification of Margins and Uncertainties).

McKay, M., Beckman, R., Conover, W., 1979. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics 21, 230–245.

Rasmussen, C., Williams, C., 2006. Gaussian Processes for Machine Learning. The MIT Press, Cambridge.

Sacks, J., Welch, W., Mitchell, T., Wynn, H., 1989. Design and analysis of computer experiments. Statistical Science 4, 409–423.

Santer, T., Williams, B., Notz, W., 2003. The Design and Analysis of Computer Experiments. Springer-Verlag, New York.