SAND20XX-XXXXR
LDRD PROJECT NUMBER: 195307
LDRD PROJECT TITLE: Statistically Rigorous Uncertainty Quantification for Physical Parameter Model Calibration with Functional Output
PROJECT TEAM MEMBERS: Lauren Hund (PI), Justin Brown

Abstract

In experiments conducted on the Z-machine at Sandia National Laboratories, dynamic material properties cannot be analyzed using traditional analytic methods, necessitating solving an inverse problem. Bayesian model calibration is a statistical framework for solving an inverse problem to estimate parameters input into a computational model in the presence of multiple uncertainties. Disentangling input parameter uncertainty and model misspecification is often poorly identified problem. When using computational models for physical parameter estimation, the issue of parameter identifiability must be carefully considered to obtain accurate and precise estimates of physical parameters. Additionally, in dynamic material properties applications, the experimental output is a function, velocity over time. While we can sample an arbitrarily large number of points from the measured velocity, these curves only contain a finite amount of information about the calibration parameters. In this report, we propose modifications to the Bayesian model calibration framework to simplify and improve the estimation of physical parameters with functional outputs. Specifically, we propose scaling the likelihood function by an effective sample size rather than modeling the discrepancy function; and modularizing input nuisance parameters with weakly identified parameters. We evaluate the performance of these proposed methods using a statistical simulation study and then apply these methods to estimate parameters of the tantalum equation of state. We conclude that these proposed methods can provide simple, fast, and statistically valid alternatives to the full Bayesian model calibration procedure; and that these methods can be used to estimate parameters of the equation of state for tantalum.

1 Introduction

Computational simulation modeling is a commonly used tool in the physical sciences for characterizing the relationship between inputs and outputs. Using such models, outputs can be predicted as a function of inputs, or conversely, inputs can be inferred by pairing computational predictions with observed experimental data. The latter process, referred to as model calibration, is the process of inferring the distribution of computer model inputs based on how well the computational model output matches the observed experimental data.

In this report, we consider calibration of dynamic material properties using velocity profiles predicted from a computational simulation model coupled with experimental observations. Specifically, in dynamic experiments conducted on the Z-machine at Sandia

National Laboratories, materials of interest can be compressed to MBar pressures. These experiments are used for the validation and development of physics models relating to the equation of state, strength, and phase transition kinetics of materials in these extreme conditions. The increasing complexities of these experiments, however, are resulting in data which can longer be analyzed using traditional analytic techniques. As such, an inverse problem must be solved by coupling hydrocode computational simulations of velocity profiles with experimental measurements.

The objective of this analysis is estimation and uncertainty quantification for two parameters of equation of state for tantalum, the bulk modulus (B_0) and pressure derivative (BP_0). Velocity profiles were measured with error across 9 different experiments (Figure 1). Hydrocode simulations were run to calculate computational predictions of velocity as a function of time, considering as inputs material properties and other experimental uncertainties (e.g. sample thickness, sample density, and boundary condition). Uncertainties in the material property estimates are driven by unknown inputs to the computational simulation code, experimental measurement error, and potential physics model misspecification. The material properties, along with other uncertain parameters, are input into the computer model, with the objective of estimating and quantifying uncertainty associated with these material properties in the presence of other uncertainties. In this calibration problem (as is common in practice), choosing an optimal approach for uncertainty quantification on model inputs is not straightforward. The input parameter space is high dimensional, experimental data are measured with error, and the computational models have model form misspecification. Further, the output from the computer model is a function (velocity over time).

Bayesian model calibration is a popular framework for estimating the values of input parameters into computational simulation models in the presence of these multiple uncertainties (Kennedy and O'Hagan 2001; Bayarri et al. 2012). However, solving for computer model inputs is typically a poorly identified problem (Kennedy and O'Hagan 2001; Brun et al. 2001). That is, multiple values of the model parameters can produce the equally valid solutions. In the presence of model misspecification, calibration parameters will typically be biased. Such model misspecification could include systematic model discrepancy, e.g. monotone bias function over time (Brynjarsdóttir and O'Hagan 2014), as well as misspecification of the model likelihood. For instance, if the Gaussian process approximation for the residuals is incorrect or if the parameters of the Gaussian process are misspecified, estimates may be biased. The problem of identifiability is often bypassed when computer models are used for prediction, defined as predicting an output at areas of the design space different from where data have been collected. In the prediction setting, model calibration is typically employed to a set of 'best fitting' parameters that do not typically have a physical interpretation but improve the predictive capability of the model. When using computer models for physical parameter estimation, the issue of parameter identifiability must be carefully considered in order to obtain accurate and precise estimates of physical parameters (Arendt et al. 2012a; Brynjarsdóttir and O'Hagan 2014; Arendt et al. 2016).

Another challenge associated with calibration of dynamic material properties is calibration of physical parameters with functional outputs. While we can sample an arbitrarily large number of points from the measured function, these curves only contain a finite amount of information about the values of the calibration parameters; naturally, uncertainties on cal-

ibration parameter estimates are highly sensitive to the model parameters that dictate the amount of information provided from a single curve, though selection of these parameters is not straightforward. While calibration with functional outputs has been discussed in the literature (Williams et al. 2006; Bayarri et al. 2007; McFarland et al. 2008), previous work only addresses calibration of tuning parameters for prediction of model output and does not address calibration of physical parameters. Arendt et al. (2012b) discuss using multiple outputs to improve calibration of physical parameters but do not provide an explicit approach for calibration of physical parameters with functional output. Brynjarsdóttir and O'Hagan (2014) discusses the issue of artificially decreasing posterior variance by increased sampling from the design space. Using a toy example with a low dimensional input parameter space and a monotone increasing discrepancy function, they conclude that introducing a model discrepancy function can improve the accuracy of physical parameter estimates, but only in the presence of highly information priors about the shape of the discrepancy function. To our knowledge, there is a gap in the literature pertaining to best-practice recommendations for model calibration of physical parameters with functional outputs for practical calibration problems (multiple nonidentifiable inputs, measurement uncertainties, and model form misspecification).

The goal of this report is to describe our efforts to explore different approaches to physical parameter estimation using Bayesian model calibration with functional output, with applications to estimating dynamic material properties. The uncertainty in the physical parameter estimates is driven by how much information about the physical parameters is contained within the experiment-specific velocity profiles. In our example, this information is primarily driven by variability in the non-physical calibration parameters, which we deem 'nuisance parameters', model discrepancy, and measurement uncertainty. We describe different approaches for quantifying this information, and list the pros and cons for each approach. This report is structured as follows. First, we describe the methods we considered for robust calibration of the physical parameters with functional output and describe the simulations and data analysis we conducted to evaluate the performance of these methods (Section 2). We describe results from the simulation study as well as apply the methods to estimate parameters of the equation of state of tantalum (Section 3). Lastly, we discuss the results (Section 4), the anticipated impact of these results (Section 5), and summarize final conclusions (Section 6).

2 Methods

Our work was motivated by concerns about the performance of the standard-of-practice Bayesian model calibration procedure for physical parameter estimation with functional data. Because identification of physical parameter posteriors is not possible in the presence of 'nuisance' calibration parameters and model discrepancy, we propose an approach to improving the robustness of the physical parameter posterior estimates with functional outputs and nuisance calibration parameters.

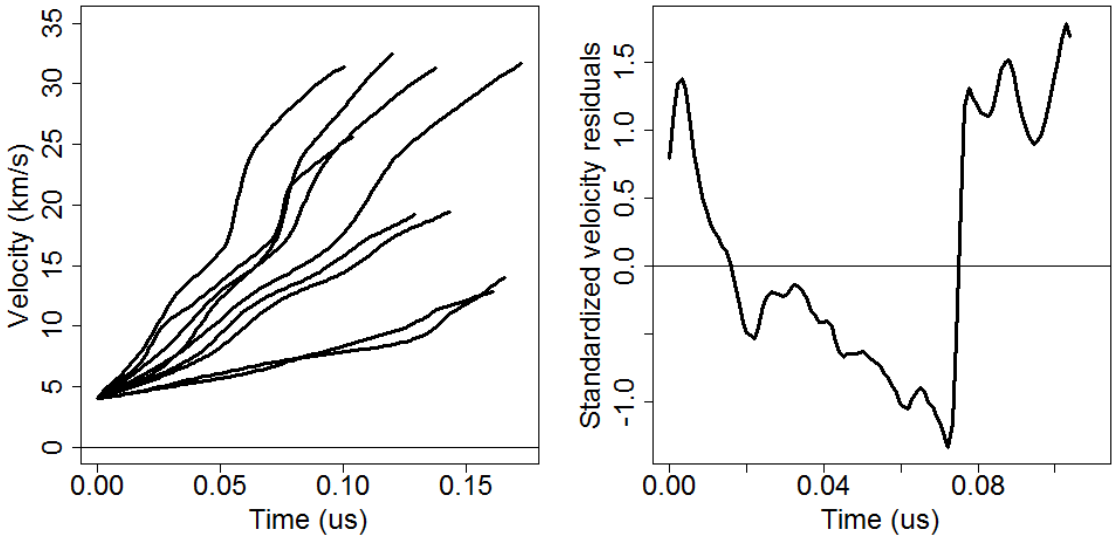

Figure 1: (Left) Experimental velocity curves for the 9 experiments. (Right) Experimental minus predicted velocity for a single experiment, standardized by the predicted variability in velocity attributable to the known experimental uncertainties.

2.1 Calibration model

First, we review the current 'best-practice' calibration model for functional outputs across multiple experiments, as described in Kennedy and O'Hagan (2001); Williams et al. (2006) and Brynjarsdóttir and O'Hagan (2014).

Notation: We use notation consistent with Kennedy and O'Hagan (2001) and Williams et al. (2006). We denote the output of the simulations as $\eta(x,t)$ given input vector (x,t), where the vector x contains observable inputs and the vector t contains unobservable calibration and tuning parameters to run the code. In our model, x includes pressure, a controllable experimental condition, and time. For each experiment, n equally spaced velocity measures are sampled from each velocity curve, such that $x_j = [P_j, x_{1j}, ..., x_{nj}]$ where P_j is pressure and $x_{1j}, ..., x_{nj}$ are times. The parameters t are the set of input parameters, made up of nuisance parameters (aluminum and tantalum thicknesses, sample density, and boundary condition scaling) and physical parameters (bulk modulus and pressure derivative). Note that the physical parameters and sample densities are constant between experiments (because all samples were cut from the same block of tantalum), whereas the thicknesses and boundard scaling changes between experiments. For notational convenience, we partition t into $t = (\gamma, \alpha)$, where $\gamma = [\gamma_j]$ is the set of nuisance parameters for each experiment and α is the set of physical parameters of interest.

Model: We model the i^{th} observation in the j^{th} experiment as:

$$y(x_{ij}) = \zeta(x_{ij}) + \epsilon(x_{ij}) \tag{1}$$

where $y(x_{ij})$ is the observed velocity at time x_{ij}, $\zeta(x_{ij})$ denotes the actual unobserved velocity at time x_{ij}, and $\epsilon(x_{ij})$ is measurement uncertainty at time x_{ij}. The measurement

uncertainty in velocity is known and specified through $\epsilon_j \sim N(0, \Sigma_j^\epsilon)$, where $diag(\Sigma_j^\epsilon) = \sigma_V^2 = .002\zeta(x_j)^2 + (0.03 * 280)^2$.

The output is modeled as a function of time and the calibration parameters using the simulator $\eta(\cdot)$.

$$\zeta(x_j) = \eta(x_j, \gamma_j^T, \theta) + \delta_j(x_j) \tag{2}$$

where θ is the true but unknown value of the physical parameters α, $\gamma^T = [\gamma_j^T]$ is the true but unknown value of the nuisance parameters, and $\delta(x_j)$ is a model discrepancy term.

Simulation runs and emulator. Because the simulation is computationally expensive, we use Gaussian process emulators to inexpensively generate new values of $\eta()$ as a function of x and t. The calibration parameters t are design values and we use Latin hypercube sampling to choose the $m = 500$ design points for each time point and experiment in x. Because uncertainty associated with the GP estimation is negligible relative to nuisance parameter and model form uncertainty, we do not account for GP uncertainty in the calibration procedure (sensitivity analyses suggest results do not change upon accounting for this source of uncertainty).

Discrepancy term. Model discrepancy can arise when the hydrocode simulations do not adequately capture the true physics, for instance due to numerical problems, grid size approximations, or order of the equation of state approximation. If present, model discrepancy can result in biased estimates of θ (Kennedy and O'Hagan 2001; Arendt et al. 2012a; Brynjarsdóttir and O'Hagan 2014). We assume the discrepancy function can be approximated using a Gaussian process, where $\delta_j \sim N(0, \Sigma_j^\delta)$ and

$$Cor[\delta(x_{ij}), \delta(x_{i'j})] \quad = \quad exp\left[-\rho_j(x_{i'j} - x_{ij})^2\right] \tag{3}$$

following Arendt et al. (2012b) and $Var(\delta_{ij}) = \phi_j\sigma_{ij}^2$, where $\{\sigma_{ij}^2\}$ are fixed variance weights (when relevant) and ϕ_j is an unknown variance scaling factor. Following Kennedy and O'Hagan (2001) and Arendt et al. (2012b), we estimate the autocorrelation parameter (here ρ_j) prior to calibration and therefore assume the autocorrelation parameter is fixed and known.

For functional data calibration, a discrepancy term must be included in the model to ensure that inferences are independent of n, the number of points sampled from the curve. Without including a discrepancy term, the standard errors of the calibration parameter estimates are $O(n^{-1/2})$. Hence, for functional data, the form of the discrepancy function has a large impact on the accuracy and precision of the physical parameter estimates.

Final model. Combining Equations 1 and 2, the observed velocity is linked to the simulation through:

$$y(x_{ij}) = \eta(x_j, \gamma_j^T, \theta) + \delta(x_{ij}) + \epsilon(x_{ij}) \tag{4}$$

We use Bayesian framework to estimate the model parameters, $[\alpha, \gamma, \phi]$ given the experimental data $y = [y_1, ..., y_J]$. Denote $\Sigma_j = \Sigma_j^\delta + \Sigma_j^\epsilon$. The model likelihood is:

$$l(y|\alpha, \gamma, \phi) \quad = \quad \prod_{j=1}^{J}(2\pi)^{-n/2}|\Sigma_j|^{-1/2}\text{exp}\left[-.5(y_j - \eta(\alpha, \gamma_j))^T(\Sigma_j)^{-1}(y_j - \eta(\alpha, \gamma_j))\right] \tag{5}$$

Denoting prior probability distributions by π, the posterior is:

$$\pi(\alpha, \gamma, \phi|y) \propto l(y|\alpha, \gamma, \Sigma^{\delta})\pi(\alpha)\pi(\gamma)\pi(\phi) \tag{6}$$

2.2 Scaling the likelihood

The standard of practice calibration model accommodates temporal autocorrelation in the velocity residuals using a Gaussian process discrepancy function. Rather than modeling the discrepancy, we instead assess the validity of scaling the likelihood function to account for the limited amount of information contained in the velocity profiles due to model discrepancy/residual autocorrelation. Specific issues we aim to address by scaling the likelihood include:

- Modeling the discrepancy function is numerically instable and computationally intensive, with the instability of the procedure increasing with the sample size (Gramacy and Lee 2012; MacDonald et al. 2015). Further, the discrepancy function is inherently non-identifiable (Brynjarsdóttir and O'Hagan 2014).

- Without including a discrepancy term in the model, the variances of the physical parameters are on the order of n^{-1}; that is, the physical parameter standard errors decrease with the number of points sampled from the velocity curve for each experiment. At some point, increasing sampling from the velocity curve does not increase the information about the physical parameters, and thus physical parameter standard errors should not scale with n.

Therefore, we aim to find an estimation procedure that constrains the standard errors of the physical parameters to be constant as a function of n but does not explicitly require modeling the discrepancy function. We could only find one other instance of scaling the likelihood in the literature (Mosbach et al. 2014). The authors chose an arbitrary scaling factor of 5 to avoid having their standard errors shrink toward 0 and did not link likelihood scaling to model discrepancy or provide any justification for selecting a scaling factor.

One reasonable scaling factor is the ratio of the effective sample size (ESS) of the velocity profile to the number of sampled points. This scaling factor results in inferences that are independent of the n, number of velocity residuals sampled from the curve. The ESS for an autocorrelated time series is:

$$\begin{aligned} n_{ej} &= n/\tau_j \\ \tau_j &= 1 + 2\sum_{k=1}^{\infty} \nu_j(k) \end{aligned} \tag{7}$$

where $\nu_j(k)$ is the autocorrelation at lag k, i.e. the correlation between $\epsilon(x_{ij})$ and $\epsilon(x_{i'j})$ where $|i - i'| = k$.

To estimate the ESS, we fit a model assuming velocities are independent, i.e.

$$\log l^*(y_j|\alpha, \gamma_j, \phi_j) = \left[-\frac{n}{2} log(2\pi) - \frac{n}{2} \sum_i log(\phi_j \sigma_{ij}^2) - \frac{1}{2} \sum_{i=1}^{n} \left(\frac{y(x_{ij}) - \eta(x_{ij}, \alpha, \gamma_j)}{\phi_j \sigma_{ij}} \right)^2 \right] \tag{8}$$

and calculate the autocorrelation $\nu_j(k)$ for the residuals from this model. Specifically, to estimate $\nu_j(k)$, we compare using the sample covariance to specifying a parametric form for the correlation structure.

The scaled likelihood is then defined as:

$$\log\, l(y_j|\alpha,\gamma,\phi) \;=\; \frac{n_{ej}}{n}\log\, l^*(y_j|\alpha,\gamma_j,\phi_j) \tag{9}$$

Estimation of the calibration parameters can proceed by using the scaled likelihood for estimation. Alternatively, if calibrating using maximum likelihood estimation, the variance of the maximum likelihood estimates (MLEs) from Equation 8 can simple be scaled by n_{ej}/n.

Scaling the likelihood by an ESS may be preferable to modeling autocorrelation in the residuals (discrepancy) for three reasons: simplicity, stability, and interpretability.

Regarding simplicity, with a sufficiently large n, it is rather straightforward to estimate τ_j. We no longer have to specify a functional form for the autocorrelation structure the residuals in order to obtain unbiased inferences. Using the standard approach of modeling the autocorrelation structure requires estimation of poorly identified parameters of the variance-covariance matrix as well as inverting an ill-conditioned matrix of size n, adding unnecessary computational complexity and instability.

Regarding stability, the estimates of the physical parameters do not change according to the form of the residual autocorrelation. Additionally, standard errors of these parameters are independent of n (insofar as a sufficiently large n is selected to capture the relevant information), providing inferences that are independent of the number of time points sampled from the velocity curve.

Regarding interpretability, n_{ej} is loosely interpretable as the number of independent pieces of information available for curve matching in experiment j after accounting for autocorrelation. Thus, one can compare n_{ej} to the number of parameters being estimated to evaluate potential identifiability issues.

2.3 Modularization

While calculating the ESS n_{ej} can help inform when parameter identifiability issues may arise, scaling the likelihood does not solve the issue of parameter identifiability. To address identifiability, we assume that the data cannot inform the true values of the nuisance parameters and subsequently 'modularize' nuisance parameter uncertainty (Liu et al. 2009; Plummer 2015). Specifically, we simply do not update the values of the nuisance parameter priors using the data and assume $\pi(\gamma|y) = \pi(\gamma)$. Modularization has been suggested as a tool to improve stability of estimates under model misspecification in computer modeling (Liu et al. 2009), but we could not find evidence of modularization with respect to sampling from prior distributions of nuisance calibration parameters.

Using this approach, the posterior is:

$$\begin{aligned}
\pi(\alpha,\gamma,\phi|y) &= P(\alpha,\phi|y,\gamma)\pi(\gamma|y) \\
&= P(\alpha,\phi|y,\gamma)\pi(\gamma) \\
P(\alpha,\phi|y) &= \int_\gamma P(\alpha,\Sigma|y,\gamma)\pi(\gamma)d\gamma
\end{aligned} \tag{10}$$

where P denotes a posterior and f denotes priors and

$$P(\alpha, \phi|y, \gamma) \propto l(y|\alpha, \phi, \gamma)\pi(\alpha)\pi(\phi) \tag{11}$$

We can modularize inference on all nuisance parameters or only a subset of the nuisance parameters. Modularizing parameters that are non-identifiable based on a sensitivity analysis will improve computational efficiency. In the presence of model discrepancy, parameters that are shared across multiple experiments may be more feasible to identify parameters than those that vary between experiments, if we are willing to assume that experiments are unbiased 'on average' (across all experiments); in this case, modularizing only experiment-specific parameters may be appropriate. In this report, we consider modularizing all nuisance parameters as well as only those that are non-identifiable based on results of a sensitivity analysis.

Modularization does not result in full Bayesian inference, and the estimated posterior distributions to not converge to a true posterior (Plummer 2015). However, using modularization, the estimated posteriors on the physical parameters should be more robust to model misspecification than full Bayesian inference. Updating nuisance parameters in the presence of model misspecification will artificially decrease uncertainty and introduce bias in physical parameter estimates (Brynjarsdóttir and O'Hagan 2014). However, modularization can also introduce bias into the physical parameter estimates (analogous to failing to include the outcome in a multiple imputation model). Therefore, neither full Bayesian inference or modularization of certain parameters is a perfect solution, and we recommend fitting both models in practice and comparing the results to understand variability in physical parameters under different modeling assumptions. Modularization has been applied in different Bayesian statistics applications, with the rationale that when models are mis-specified, updating parameters within an analysis may actually induce bias in other spaces of the model rather than improving estimates (Zigler et al. 2013; Plummer 2015).

2.4 Analysis of tantalum data

To test the performance of the proposed methods, we estimated parameters of the tantalum equation of state as a case example.

Data processing. We processed data for 9 different tantalum experiments. We thinned the datasets to have $n = 100$ time points per experiment for computational efficiency. We approximated the computational simulation model using a Gaussian process emulator, assuming a Gaussian correlation structure (Equation 3) across time points, with common hyperparameters set to the average over the time points. We estimated the emulator using the DiceKriging R package (Roustant et al. 2012). We also used a lower-fidelity but faster spline-based emulator. Results were not sensitive to the use of or choice of the emulator. We fit the emulators using 500 build points, selected using Latin hypercube sampling across the nuisance and physical parameter space.

Sensitivity analysis. We conducted a sensitivity analysis to understand how the sensitivity to the calibration parameters changes across the velocity curves. We calculated first-order and total Sobol indices (Saltelli et al. 2000) to estimate the amount of variance in velocity at a single time point explained by each input. We used the known input distribu-

tions for the input parameters and restricted the physical parameters to their approximate MLE sampling distributions: $B_0 \sim N(189, 2.9)$ and $BP_0 \sim N(3.92, .075)$.

Details of model calibration estimation procedures. We used both Bayesian and frequentist calibration procedures to estimate the physical parameters. Throughout, we use the likelihood function specified in Equation 9 to represent the likelihood of an observed velocity curve for a given experiment. We estimate the variance weights (σ_j) by calculating the variance in velocity at each time point attributable to the nuisance parameters and measurement uncertainty. Results were not sensitive to the choice of variance weights.

For the maximum likelihood estimation, we use brute force optimization of the likelihood, restricting the nuisance parameters to ± 3 standard deviations, with the standard deviations calculated from the known experimental uncertainties. We apply a REML correction $n_{ej}/(n_{ej}-p)$ to $\hat{\phi}$ to adjust the estimated variance for small sample bias in the maximum likelihood estimate.

For Bayesian model calibration (BMC), we used Monte Carlo Markov Chain (MCMC) sampling to estimate the posterior distribution of the model parameters. We specified the following prior distributions on the model parameters:

- Bulk modulus (GPa): $N(189, 17^2)$
- Pressure derivative: $N(3.90, .6^2)$
- Density (g/cm^3): $N(16.55, .066^2)$
- Thickness: $N(0, 1.5e^{-6})$
- Boundary scaling: $N(1, 4^{e-3})$
- $\phi_j \sim InvGamma(1.75, .35)$

We use a somewhat informative prior on ϕ_j to encourage the variance parameter to be near the expected variance in velocity, σ_j. We updated the variance parameter ϕ using the conjugate Inverse-gamma prior and updated the calibration parameters using a Metropolis step. For modularized parameters, we simply sampled from the priors rather than updating a posterior distribution. We computed 10,000 posterior samples with a burn in of 2,000 samples and assessed convergence using trace plots.

All analyses were conducted in the R statistical software package (R Core Team 2015).

2.5 Simulation study to evaluate proposed methods

We constructed a simple simulation study to evaluate the performance of the likelihood scaling approach. Specific questions addressed in the study include:

1. Can the effective sample size be accurately estimated using Equation 7?
2. Does the likelihood scaling estimation procedure result in valid inferences for a single physical parameter under a mean zero Gaussian process discrepancy function? for multiple physical parameters?

2.5.1 Estimation of ESS

To address question (1) above, we did not use the tantalum data, but simply generated realizations from a mean 0 Gaussian process (GP) with a Gaussian correlation structure (Equation 3). We selected autocorrelation parameters corresponding to effective sample sizes of $n_{ej} = 3, 5, 10$, and 20. For each generated GP, we sampled $n = 100, 200$, or 500 equally spaced points as our dataset and then estimated the effective sample size. We considered two methods for estimating the autocorrelation ν_j: (1) using the non-parametric sample correlations and (2) fitting a GP to the simulated data assuming the correct (Gaussian) form of the correlation structure. We compared the two procedures by generating 500 GP realizations and comparing the estimates of n_{ej} to the true value across the different scenarios.

2.5.2 Likelihood scaling for statistical inference on physical parameters

To address the second simulation question, we used a statistical simulation study based on the tantalum data to evaluate the statistical properties of scaling the likelihood when estimating calibration parameters with autocorrelated residuals. We first consider the simple case of estimating a *single physical parameter from a single experiment, ignoring nuisance parameter uncertainty*. We generate simulated datasets from the model in Equation 2, acting as though the emulator for the computational simulation model representing the true underlying physics. We assumed $\phi = 1$ and used the variance weights σ_j described in Section 2.4 to determine the variance of the discrepancy term. We used a single experiment for this simulation study, arbitrarily selecting the data for the chronologically first experiment in the dataset. We assume measurements are collected over $n = 100$ time points. We set the true values of the calibration parameters (physical and nuisance) to the prior mean values. We vary the effective sample size by generating mean zero GP discrepancies with ρ selected such that the effective sample size is approximately 4, 6, 11, and 20. We assume all physical and nuisance parameters are fixed and known, aside pressure derivative.

We fit the scaled likelihood model using maximum likelihood estimation (for computational efficiency) and calculated 95% confidence intervals for α using a t-distribution with $n_{ej} - 1$ degrees of freedom. We compare the average estimated standard error of the physical parameter $E[se(\hat{\alpha})]$ to the empirical standard error $se(\hat{\alpha})$ as well as calculate the 95% confidence interval coverage across ~ 1000 simulation runs.

We also considered the case where we jointly estimate 4 parameters - the 2 material properties and 2 sensitive nuisance parameters. We considered effective sample sizes of ~ 6 and 11 as case examples.

2.5.3 Tantalum analyses

Calibration of single experiments. We analyzed the individual experiments separately, to understand how estimates and uncertainties changed according to the estimation procedure when only one experiment is available (to inform future situations when fewer experiments are conducted). First, we inferred the physical parameters for each experiment separately using maximum likelihood estimation, ignoring the nuisance parameters by fixing all nuisance parameters to their nominal values. Next, we selected a single experiment

and compared three different estimation procedures: (1) the ML estimates ignoring nuisance parameter uncertainty; (2) BMC estimates with sensitive nuisance parameters (boundary scaling and density) estimated; and (3) BMC with all nuisance parameters modularized.

Calibration combining across experiments. To understand the impact of different modeling choices on uncertainty quantification for the tantalum material properties, we fit two models: BMC inferring physical parameters and sensitive nuisance parameters and BMC inferring only physical parameters. We compare the estimates and uncertainties across these two approaches and compared the results to the single experiment calibration.

3 Results

We have structured the results section to parallel the described methods section. First, we discuss the simulation study results and then we discuss the data application results.

3.1 Simulation study results

In this section, we describe results pertaining to the simulation study described in Section 2.5 above.

3.1.1 Estimation of the effective sample size

The ESS can be accurately estimated using Equation 7, but the correct form of the autocorrelation structure is needed for accurate inferences with smaller effective sample sizes (Figure 2). For low values of the ESS (< 10), the sample correlation is biased low, making the non-parametric ESS biased high. Upward bias in the ESS will result in underestimation of the calibration parameter standard errors (assumes more information than actually exists). For higher values of the ESS (≥ 10), the bias in the nonparametric estimator is negligible, and we suggest using the non-parametric estimator in this case.

3.1.2 Estimation of physical parameters

The likelihood scaling estimation procedure results in approximately valid inferences on a single physical parameter under a mean zero Gaussian process discrepancy function. Results of the simulation are summarized in Table 1. The estimates of the pressure derivative are unbiased. The estimated standard errors are slightly too low due to the upward bias in the non-parametric ESS described in Section 3.1.1, resulting in confidence interval coverage that is approximately too low for smaller ESS values (Figure 3). The standard errors of the physical parameters are now proportional to n_{ej} rather than n, such that the standard errors increase as the amount of autocorrelation in the residuals increases but are constant with respect to the number of points sampled n.

The simulation results suggest that obtaining valid inferences when inferring multiple physical parameters is more challenging, as expected. With 4 inferred parameters (2 material properties, density, and boundary scaling) and effective sample sizes of 6 and 11, asymptotic maximum likelihood procedures seem to begin to break down and identifiability

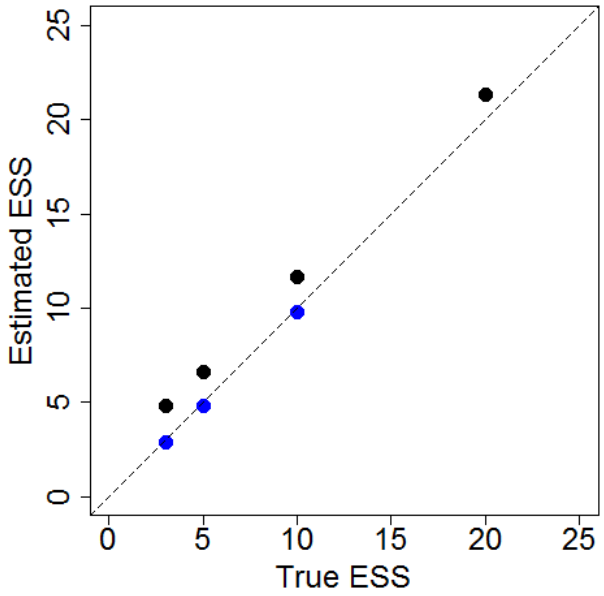

Figure 2: True versus median estimated ESS, comparing parametric (blue) to nonparametric (black) estimate of ESS. The parametric estimate for ESS = 20 failed to converge at least half of the time and is therefore not included in the plot.

issues arise. The point estimates of the material properties remain unbiased. Inferences on the bulk modulus are also valid, with confidence interval coverage 94% and 93% for ESS 6 and 11, respectively. These results suggest likelihood scaling is providing valid inferences for strongly identifiable parameters. However, confidence interval coverage for the pressure derivative is lower (90% and 85% for ESS 6 and 11), likely attributable to the fact that the pressure derivative is much more weakly identified than the bulk modulus (due to the strong influence of boundary scaling). Hence, the estimation procedure results in reasonable inferences when inferring strongly identifiable physical parameters under a mean zero Gaussian process discrepancy function. However, for weakly identifiable parameters (pressure derivative), performance seems to deteriorate due to identifiability issues. Incorporating prior information via Bayesian inference could improve the identifiability issues, but evaluating the performance of the Bayesian approach would require a large number of much more computationally expensive Bayesian model runs. This simulation study was simply designed to show proof of concept regarding scaling the likelihood as a viable alternative to modeling the discrepancy, and more extensive simulations are needed to evaluate performance across different ESS values and different numbers of inferred parameters.

3.2 Tantalum data analysis results

Sensitivity analysis. The material properties, density, and boundary scaling account for almost all of the variability in velocity, and the sensitivity of the model to these parameters changes over time (Figure 4). Material thicknesses explain a very small fraction of variability in velocity. The impact of the material properties and boundary scaling changes over time;

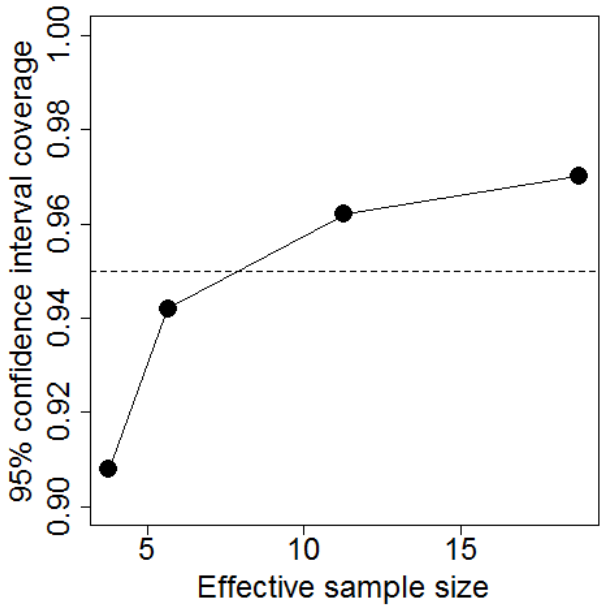

Figure 3: 95% confidence interval coverage for single physical parameter (pressure derivative) as a function of the true ESS, calculated using a statistical simulation study. With mean 0 discrepancy, scaling the likelihood results in approximately valid statistical inferences with functional outputs when inferring a single parameter, though coverage declines with the ESS due to small-sample bias.

True ESS	Estimate Bias	SE Bias	Coverage	ESS Bias
3.8	0.0	-12.6	0.908	60.1
5.6	-0.3	-11.7	0.942	37.9
11.3	0.2	-6.0	0.962	15.2
18.8	-0.1	3.5	0.970	8.8

Table 1: Simulation results for assessing the validity of scaling the likelihood by a non-parametrically estimated ESS when estimating a single physical parameter.

the bulk modulus is more important initially, and pressure derivative and boundary scaling are more important at later time points.

Calibration. Examining the results from the single-experiment analysis (Figure 5), the material property estimates are similar across the different calibration procedures (ignoring nuisance parameters, inferring sensitive nuisance parameters, and modularizing all nuisance parameters), while the variances change across the calibration procedures. Ignoring nuisance parameter uncertainty results in the smallest variance for the estimated material properties. Accounting for nuisance parameter uncertainty using BMC substantially inflates the variances. The variance of the bulk modulus is much smaller when inferring nuisance parameters versus modularizing the nuisance parameters, while the variance of the pressure derivative is similar across the settings. There is a strong negative correlation in the joint posterior for the bulk modulus and pressure derivative.

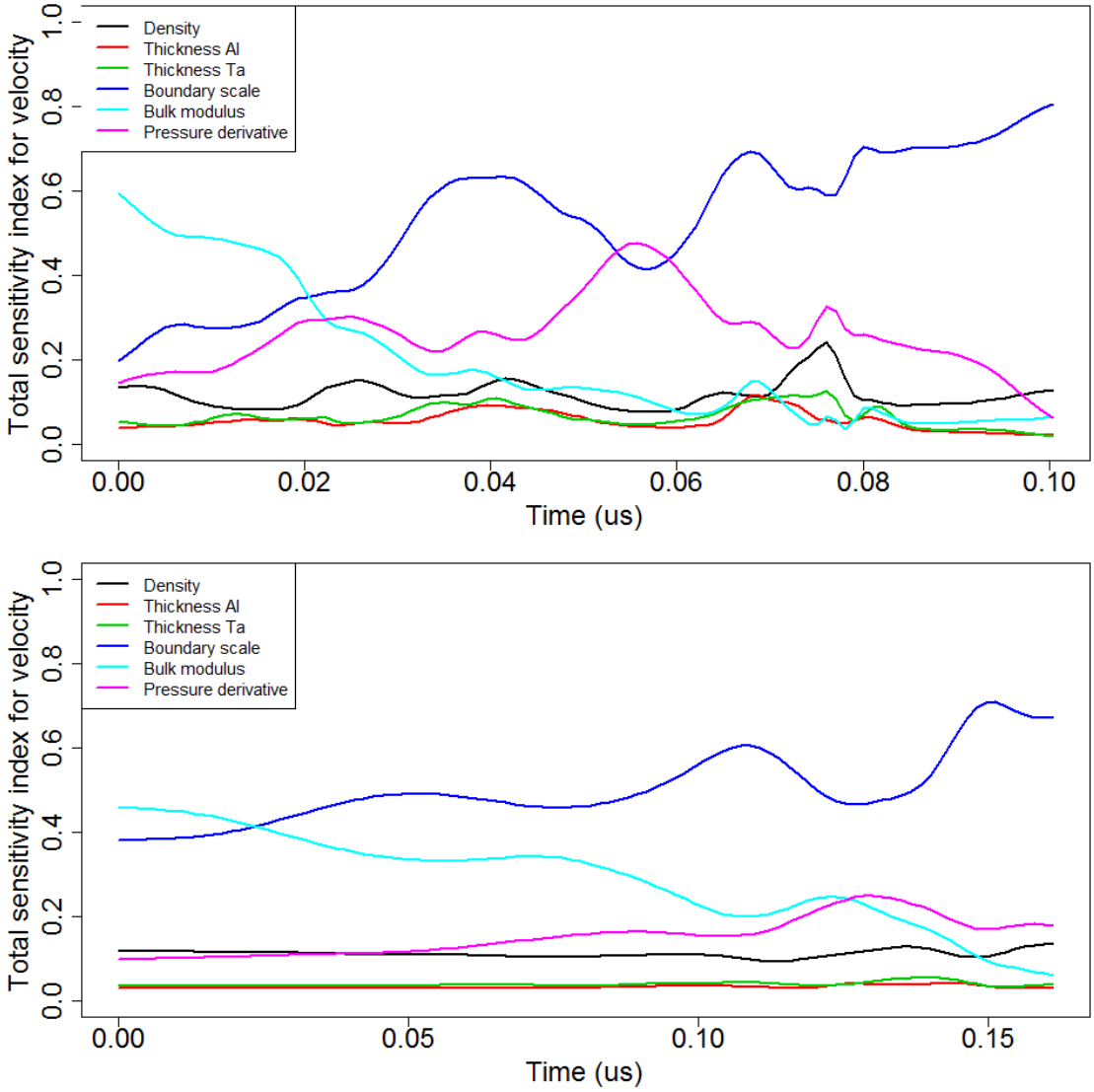

Figure 4: Total sensitivity index over time for input nuisance parameters and material properties in two experiments: high pressure (top) and lower pressure (bottom).

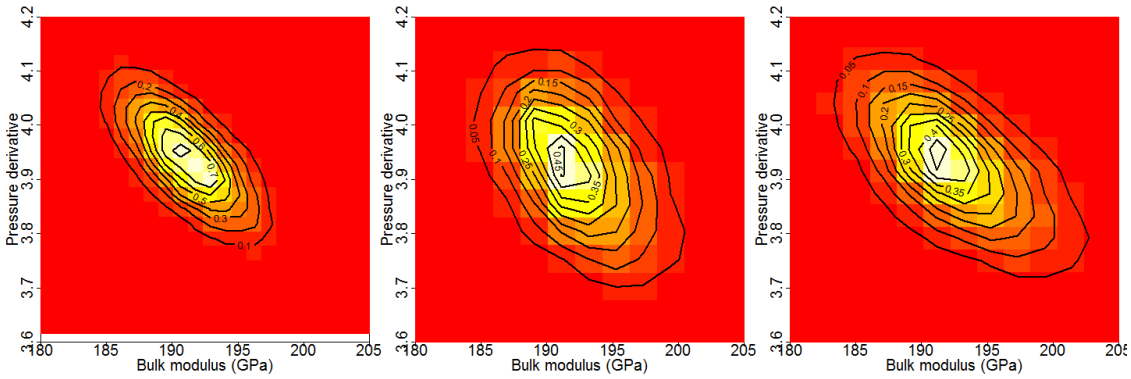

Figure 5: Posterior distribution of the bulk modulus and pressure derivative from a single experiment using different estimation procedures: (left) maximum likelihood approximation ignoring nuisance parameter uncertainty; (middle) Bayesian model calibration inferring only sensitive nuisance parameters (boundary scaling and density); and (3) modularizing all nuisance parameters.

Examining the maximum likelihood estimates for all 9 experiments (Figure 6), we observe substantial variability in both the point estimates and standard errors across the experiments. The average estimated bulk modulus across experiments is 191 GPa , with standard error 4.8 GPa; the average estimated pressure derivative is 3.8 with standard error .17. The estimated bulk modulus tends to decrease with input pressure while the pressure derivative increases. The results of the second experiment are somewhat different compared to the remaining 8 experiments; the second experiment has the highest bulk modulus and lowest pressure derivative estimates, with relatively tight 95% confidence intervals. The ESS for the individual experiments ranges from 5 to 12, with an average of 7, suggesting small sample bias in the ESS could result in some underestimation of the experiment-specific ESS values.

Using BMC, we combine across all 9 experiments to estimate the material properties. When we update the sensitive nuisance parameters (boundary scaling and density), the estimated bulk modulus is 189 GPa (SE 1.1) and the estimated pressure derivative is 3.8 (SE .05). Modularizing all nuisance parameters, the estimated bulk modulus is 190 (SE 3.0) and estimated pressure derivative is 3.8 (SE .11). While the point estimates are similar between the two estimation procedures, the standard errors are much higher when all nuisance parameters are modularized, as anticipated. Choice of which estimates are more credible depends on the assumptions about the underlying model. Specifically, if we believe the model in Equation 4 is correct, namely that the model is 'on average' correct across time (mean 0 discrepancy function), then the data should be able to inform the nuisance parameters and there is no need to modularize. On the other hand, if we question the accuracy of the model form in Equation 4, then there is a risk that we could update the nuisance parameters incorrectly, inducing more bias in the physical parameter estimates; in this case, modularization may be a more appropriate option.

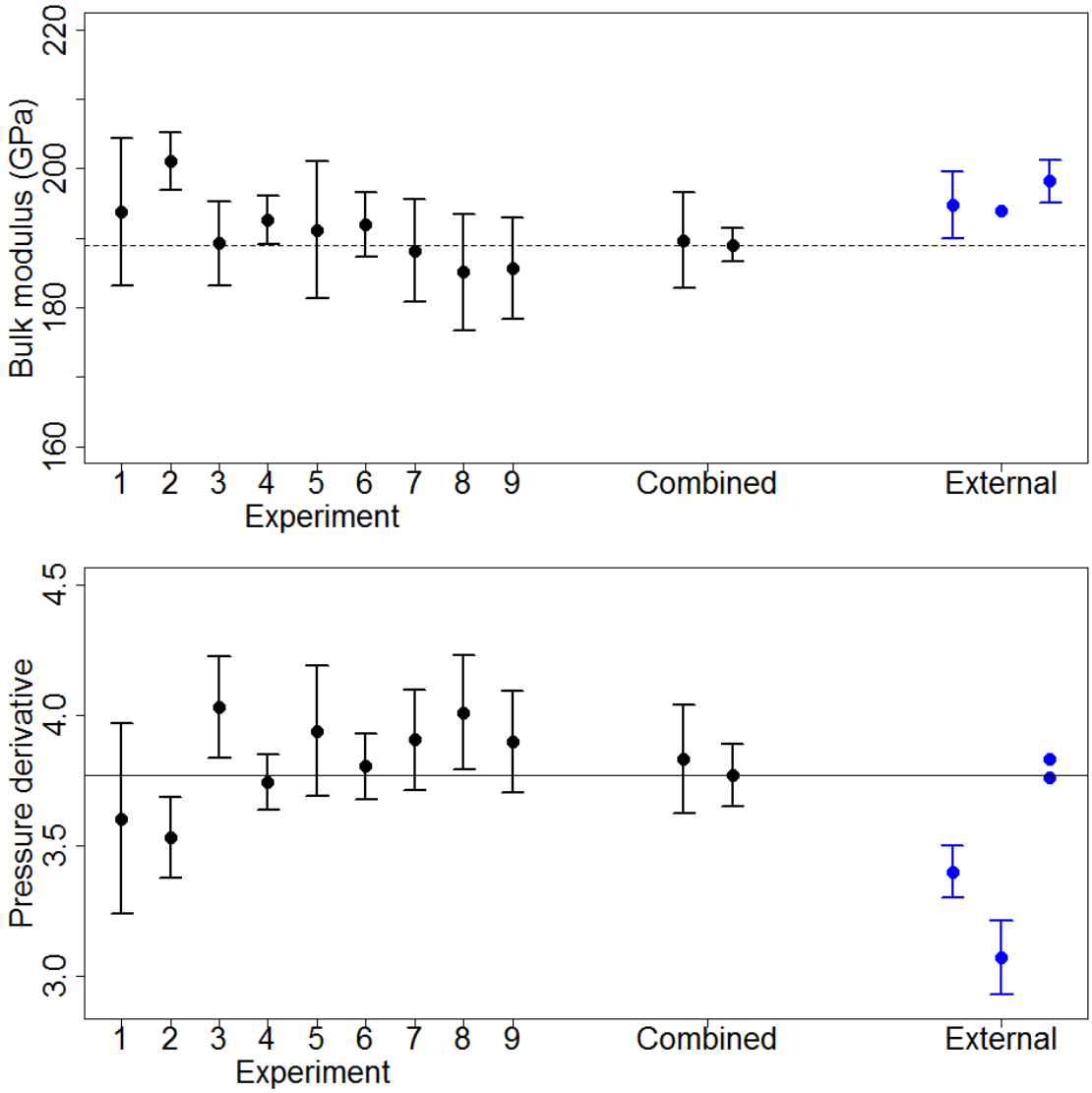

Figure 6: Estimates of bulk modulus (top) and pressure derivative (bottom). (Right) Maximum likelihood estimates with 95% CIs are provided for the 9 individual experiments, ordered by pressure ramping input P_j (experiment-specific). (Middle) BMC estimates pooling across all experiments (combined). The first combined estimate is based on modularizing all nuisance parameters; and the second is based on estimating sensitive nuisance parameters (boundary scaling and desnity). The external estimates come from published estimates of the material properties (Söderlind and Moriarty 1998; Cynn and Yoo 1999; Dewaele et al. 2004).

4 Discussion

In summary, we have proposed simple alternatives to the full Kennedy and O'Hagan (2001) calibration model and illustrated that these alternative can produce valid statistical inferences for estimating physical parameters using Bayesian model calibration. Specifically, we proposed scaling the likelihood function as an alternative to explicitly modeling the discrepancy function. Likelihood scaling is only valid assuming a mean zero discrepancy function. While the Kennedy and O'Hagan (2001) model can theoretically accommodate non-mean-zero discrepancies, estimates of physical parameters are biased without strong prior knowledge about the discrepancy in this setting (Brynjarsdóttir and O'Hagan 2014); assuming mean-zero discrepancy is thus necessary for inferring physical parameters without this prior knowledge. Future work could formally compare the full Kennedy and O'Hagan (2001) model to the scaled likelihood model. Further, we could explore different methods for estimating the ESS, such as pre-posterior analysis (Arendt et al. 2016) to find a value that produces valid inferences under discrepancy functions that are not mean zero but are constrained within some range of functional forms.

Estimating the 'effective sample size' of the functional output provides an interpretable measure of the number of independent pieces of information provided from the output, facilitating calculating the model degrees of freedom to assess whether the model is over-determined. Nonparametric estimates of the ESS are biased high, with the bias increasing as the ESS decreases. Parametric estimation of the ESS eliminates bias, though specification of the parametric form is somewhat arbitrary. Future work could explore bias corrections to the ESS as well as whether ESS estimates are approximately correct with mis-specified correlation structures.

In the presence of an over-determined system, modularization of the nuisance parameters may improve uncertainty estimation by avoiding identifiability problems. Specifically, if a model is mis-specified, full Bayesian inference will update the calibration parameters incorrectly, resulting in biased physical parameter estimates; modularization assumes that, because the model is potentially mis-specified, the experimental data cannot inform some (or all) of the nuisance parameters. Note that modularization does not guarantee conservatism in the presence of a mis-specified model. Modularization requires deviating from full Bayesian inference, and the parameter distributions lose technical interpretation as Bayesian posteriors. The experimental data may be able to inform some or all of the nuisance parameter values, in which case full Bayesian inference would be preferable. We considered modularizing all nuisance parameters, experiment-specific parameters, and parameters with negligible impact on the outcome. Future work should explore how to decide which parameters to modularize. Future work could also assess the theoretical validity of modularization as proposed as well as determine the optimal sampling algorithm for inference (sampling within or outside of the MCMC).

We used statistical simulation studies to validate the statistical properties of our proposed approaches. Statistical simulation can provide substantial added value when understanding the performance of calibration models under certain assumptions, but many proposed calibration approaches lack formal validation through statistical simulation. Because Bayesian estimation via MCMC is computationally expensive, we conducted the simulations using

maximum likelihood estimation. While maximum likelihood estimation should approximately agree with the Bayesian inference results, future work could explore replicating the simulation studies using Bayesian inference.

We applied the model calibration procedure to estimate the bulk modulus and pressure derivative of tantalum. The estimated material properties were consistent with hypothesized values in the literature, and the uncertainties in the properties combining across experiments were consistent with the variability in the experiment-specific estimates. Therefore, we conclude that Bayesian model calibration is a useful tool for estimating and characterizing uncertainty in the bulk modulus and pressure derivative of tantalum. The estimated uncertainties in the physical parameters varied according to the modeling decisions, such as when to modularize nuisance parameters and how to scale the likelihood. While future work could explore determining an optimal estimation procedure for the material properties, we recommend estimating the parameters across different settings to assess the sensitivity of the results to the subjective modeling decisions.

High correlation was observed between the bulk modulus and pressure derivative, as well as between the pressure derivative and boundary scaling, suggesting weak joint identifiability of these parameters. Reducing uncertainty in boundary scaling would improve precision in physical parameter estimates. We observed some trending of the estimates according to the pressure of the experiment as well as strong autocorrelation in the residuals of the fitted calibration model, suggesting presence of some model discrepancy that varies according to pressure. Future work should prioritize developing model diagnostics to identify the form of the model discrepancy. Further, future work can explore the validity of the proposed methods in the presence of more physical parameters to estimate, such as materials with phase transitions.

5 Anticipated Impact

This project provides valuable preliminary information needed to work toward an accepted inverse analysis technique within the field of dynamic material properties. We demonstrated the feasibility of estimating dynamic material properties by solving an inverse problem using Bayesian model calibration. In experiments conducted on the Z-machine at SNL, the collected data cannot be analyzed using traditional analytic methods, necessitating solving an inverse problem to estimate the material properties of interest. Credible uncertainty quantification in these estimates is needed to understand how much information each experiment provides about the properties of interest and to propagate these uncertainties forward. We applied the methods to tantalum as a proof of concept, but aim to extend the work to other materials without an analytic solution and whose properties are of direct pertinence to the Sandia nuclear weapons mission area. As next steps, we plan to continue expanding our work within dynamic material properties. Within this project, we worked to develop generalized software coded in Matlab that can be applied across different calibration problems. Subsequently, the methods can immediately be applied to other applications in dynamic material properties. Experiments have already been conducted, such that we can immediately begin analyzing additional datasets. However, some of the upcoming applications require inferring more physical parameters, so the validity of calibration framework will need to be

evaluated in these settings. We will continue partnering with Organization 1646 (Dynamic Material Properties) to finalize the calibration framework and assess the generalizability across different dynamic material properties applications.

Research into the calibration of physical parameters is limited but growing (Arendt et al. 2012b,a; Brynjarsdóttir and O'Hagan 2014; Arendt et al. 2016). Interest in Bayesian model calibration also seems to be increasing at Sandia National Laboratories, with a recent implementation of the methodology in Dakota and an increase in uncertainty quantification for computational modeling applications in general. Subsequently, there is a need for practical and implementable calibration solutions for analysts, as well as a need to distinguish between best-practice calibration for prediction of outputs versus learning about physical parameters. The methods proposed herein sacrifice some efficiency for robustness in physical parameter estimates, but we argue that this may be preferable when estimating physical parameters. Scaling the likelihood is simpler, more interpretable, faster, and more computationally stable than full BMC. Modularization may increase robustness of physical parameter estimates in the presence of nuisance parameters and model discrepancy.

Existing best-practice calibration methods are conceptually complex for non-statistical audiences and can also be computationally instable. The questions asked within this proposal are not unique to the calibration of dynamic material properties. There are many cross-cutting themes within calibration activities such as:

- How should the likelihood function be formulated when combining data across multiple experiments? How should individual experiments be weighted?
- Should all calibration parameters (physical and nuisance) be estimated when trying to infer physical parameters?
- How should model discrepancy be diagnosed and accounted for when inferring physical parameters with functional outputs?

The decision-making space for conducting model calibration in practice is exceptionally high dimensional and complex. Understanding the relative importance of the set of decisions is key to communicating credibility of results, but no generalized guidance has been provided for analysts.

As next steps, we aim to submit a proposal to provide practical guidance for implementing calibration, targeted toward analysts. We plan to target either ASC or LDRD funding mechanisms for pursuing this research. Additionally, we are part of an on-going project with Organization 1544 (V&V, UQ, Credibility Processes) to explore end-to-end ModSim uncertainty quantification recommendations for analysts. The lessons learned within this exploratory express project will inform the path forward for providing such end-to-end recommendations as well as produce additional preliminary data for building a calibration-themed proposal for LDRD submission or ASC funding. We are also presenting the results of our work at the University of New Mexico (UNM) Applied Math and Statistics Colloquium, raising the external visibility of the work as well as build partnerships with UNM faculty who are currently pursuing research in uncertainty quantification for ModSim. We have engaged a UNM statistics graduate student intern in the project, aiming to further build our relationship with the UNM statistics department and ultimately pursue joint research. We

will also submit an abstract to present the work at a statistics conference. We are currently preparing a publication for submission to a peer-reviewed journal during the first quarter of FY17.

6 Conclusion

In conclusion, Bayesian model calibration appears to be a viable methodology for estimating dynamic material properties from experimental data collected on the Z-machine. Existing approaches to Bayesian model calibration in these types of applications (Williams et al. 2006) are computationally expensive and may underestimate uncertainty in the presence of model form misspecification. Subsequently, we proposed simpler alternatives to the existing approaches: scaling the likelihood function as an alternative to explicitly modeling the discrepancy function and modularization of calibration nuisance parameters as an alternative to full Bayesian inference. Using statistical simulation coupled with analysis of experimental tantalum data, we illustrated that these alternatives can produce valid and potentially more robust statistical inferences for estimating physical parameters than the existing approach. Future work will explore assessing the validity of the proposed calibration approaches in the presence of a larger number of physical parameters as well as developing model diagnostics to identify model discrepancy. Further, this initial work can inform future work aiming to develop best-practice guidance for using Bayesian model calibration for the estimation of physical parameters.

References

Arendt, P. D., Apley, D. W., and Chen, W. "Quantification of model uncertainty: Calibration, model discrepancy, and identifiability." *Journal of Mechanical Design*, 134(10):100908 (2012a).

—. "A preposterior analysis to predict identifiability in the experimental calibration of computer models." *IIE Transactions*, 48(1):75–88 (2016).

Arendt, P. D., Apley, D. W., Chen, W., Lamb, D., and Gorsich, D. "Improving identifiability in model calibration using multiple responses." *Journal of Mechanical Design*, 134(10):100909 (2012b).

Bayarri, M., Berger, J., Cafeo, J., Garcia-Donato, G., Liu, F., Palomo, J., Parthasarathy, R., Paulo, R., Sacks, J., and Walsh, D. "Computer model validation with functional output." *The Annals of Statistics*, 1874–1906 (2007).

Bayarri, M. J., Berger, J. O., Paulo, R., Sacks, J., Cafeo, J. A., Cavendish, J., Lin, C.-H., and Tu, J. "A framework for validation of computer models." *Technometrics* (2012).

Brun, R., Reichert, P., and Künsch, H. R. "Practical identifiability analysis of large environmental simulation models." *Water Resources Research*, 37(4):1015–1030 (2001).

Brynjarsdóttir, J. and O'Hagan, A. "Learning about physical parameters: The importance of model discrepancy." *Inverse Problems*, 30(11):114007 (2014).

Cynn, H. and Yoo, C.-S. "Equation of state of tantalum to 174 GPa." *Physical Review B*, 59(13):8526 (1999).

Dewaele, A., Loubeyre, P., and Mezouar, M. "Equations of state of six metals above 94 GPa." *Physical Review B*, 70(9):094112 (2004).

Gramacy, R. B. and Lee, H. K. "Cases for the nugget in modeling computer experiments." *Statistics and Computing*, 22(3):713–722 (2012).

Kennedy, M. C. and O'Hagan, A. "Bayesian calibration of computer models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):425–464 (2001).

Liu, F., Bayarri, M., Berger, J., et al. "Modularization in Bayesian analysis, with emphasis on analysis of computer models." *Bayesian Analysis*, 4(1):119–150 (2009).

MacDonald, B., Ranjan, P., Chipman, H., et al. "GPfit: An R package for fitting a gaussian process model to deterministic simulator outputs." *Journal of Statistical Software*, 64(i12) (2015).

McFarland, J., Mahadevan, S., Romero, V., and Swiler, L. "Calibration and uncertainty analysis for computer simulations with multivariate output." *AIAA journal*, 46(5):1253–1265 (2008).

Mosbach, S., Hong, J. H., Brownbridge, G. P., Kraft, M., Gudiyella, S., and Brezinsky, K. "Bayesian Error Propagation for a Kinetic Model of n-Propylbenzene Oxidation in a Shock Tube." *International Journal of Chemical Kinetics*, 46(7):389–404 (2014).

Plummer, M. "Cuts in Bayesian graphical models." *Statistics and Computing*, 25(1):37–43 (2015).

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2015).
URL `https://www.R-project.org/`

Roustant, O., Ginsbourger, D., and Deville, Y. "DiceKriging, DiceOptim: Two R Packages for the Analysis of Computer Experiments by Kriging-Based Metamodeling and Optimization." *Journal of Statistical Software*, 51(1) (2012).

Saltelli, A., Chan, K., Scott, E. M., et al. *Sensitivity analysis*, volume 1. Wiley New York (2000).

Söderlind, P. and Moriarty, J. A. "First-principles theory of Ta up to 10 Mbar pressure: Structural and mechanical properties." *Physical Review B*, 57(17):10340 (1998).

Williams, B., Higdon, D., Gattiker, J., Moore, L., McKay, M., Keller-McNulty, S., et al. "Combining experimental data and computer simulations, with an application to flyer plate experiments." *Bayesian Analysis*, 1(4):765–792 (2006).

Zigler, C. M., Watts, K., Yeh, R. W., Wang, Y., Coull, B. A., and Dominici, F. "Model feedback in bayesian propensity score estimation." *Biometrics*, 69(1):263–273 (2013).