is paper describes objective technical results and analysis. Any subjective views or opinions that might be expres
the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Governm

National Laboratories

SAND2018-7580C

Overfitting in Bayesian Model Calibration of Functional Data Under Misspecified Models

PRESENTED BY

Lauren Hund[1]

[1] Department of Statistical Sciences
Sandia National Laboratories

ENERGY NNSA

ories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of San
well International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under cont

Nuclear Security Administration under contract DE-NA-0003525. SAND NO. 2017-00000

This presentation represents joint work with Justin Brown.

This work was supported by a Sandia National Laboratories Laboratory Directed Research and Development (LDRD) grant. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Outline

- The backstory
- Power likelihood methods for model calibration
- Different strategies for implementation
- Future directions

Backstory

Dynamic material properties experiments provide access to the most extreme temperatures and pressures attainable in a laboratory setting.

- Sandia National Laboratories's Z-machine is a pulsed power driver that can deliver massive electrical currents over very short timescales (of the order of 1 μs).

Goal: Improve our understanding of material models at extreme conditions by pairing computational simulation predictions with experimental data.

Backstory

Goal: *Generalized solution* for calibrating dynamic material models.

- First: Calibrate a well-understood model - two parameters of the equation of state of tantalum.

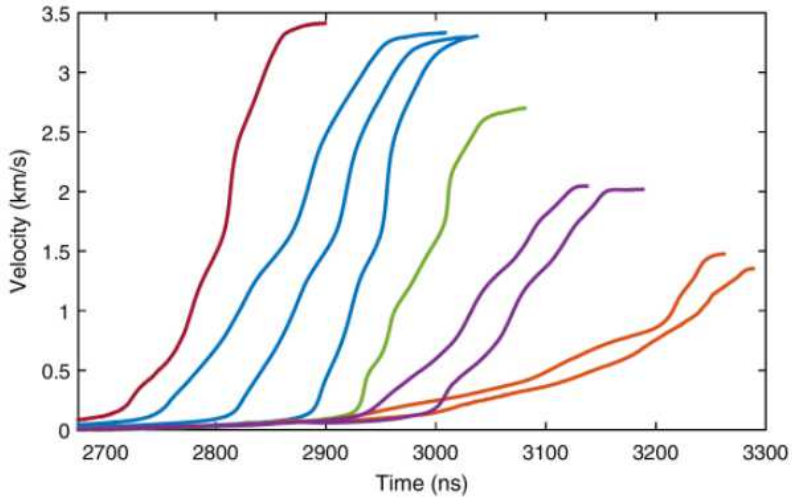


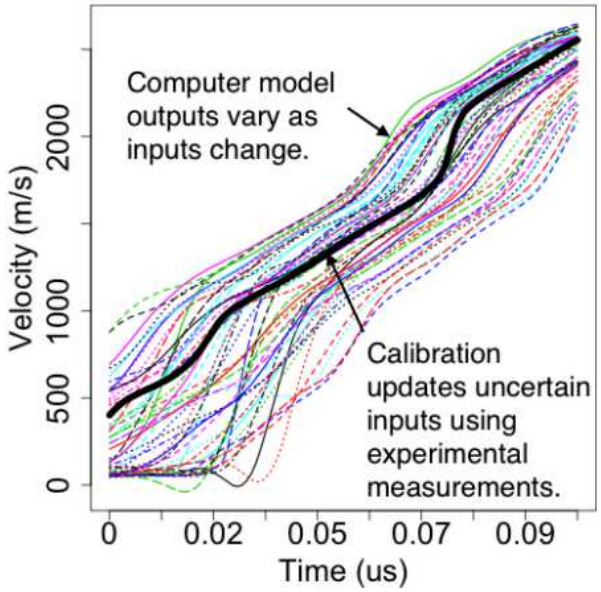
Fig. 2. Experimental measurements considered in this study and reported previously by Brown *et al.* (2014); each colour represents a different experiment

Calibration data are functional - velocity traces over time.

Backstory

Uncertain model inputs:

- Parameters of interest: bulk modulus and pressure derivative of tantalum
- Experimental uncertainties: boundary condition, material thicknesses, timing, velocity.



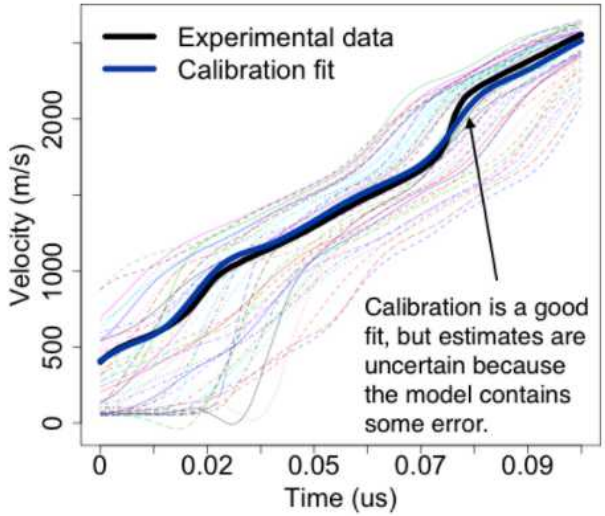
Backstory

Bayesian calibration is a natural solution...

- Can incorporate more experimental uncertainties (i.e., boundary condition) than previous approach.
- Strong prior information about experimental uncertainties constrains problem.

Model discrepancy

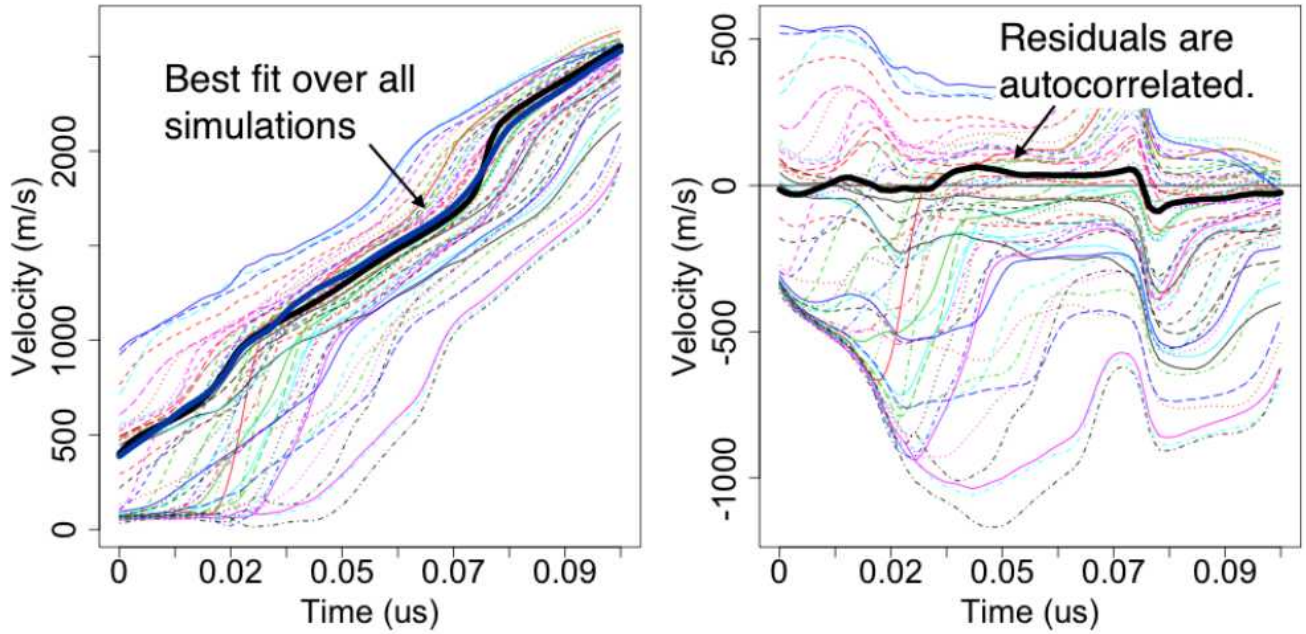
Model misspecification (discrepancy) must be accounted for to properly quantify input uncertainty.



Paraphrasing, we know that: ‘the models aren’t perfect, but they are pretty good.’

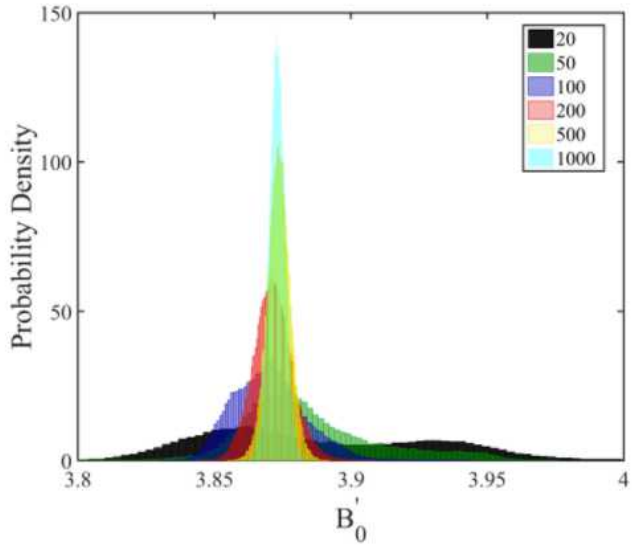
Model discrepancy

Discrepancy manifests as autocorrelation in residuals from calibration.



Model discrepancy

Accounting for discrepancy is not optional with functional output.



Variance of calibrated inputs scales with the number of points sampled from the curve.

KOH

The most commonly used model for discrepancy is Kennedy and O'Hagan (2001) (KOH):

$$\begin{aligned} y(x) &= \eta(x,\theta) + \delta(x) + \epsilon(x) \\ \delta &\sim N(\mu_\delta, \Sigma^\delta) \\ \epsilon &\sim N(0,\sigma) \end{aligned}$$

where

- x are known control inputs that are controlled by design (e.g. experimental test conditions, time)
- θ are calibration parameters
- η is the true value of the outcome as a function of x and θ
- ϵ is random measurement error, with σ known
- δ is a Gaussian process model discrepancy term

KOH

KOH is very popular for interpolative prediction.

- Predicts the outcome well within the data range.

KOH should not be expected to provide correct inferences for extrapolative prediction or physical parameter estimation, because the discrepancy function and calibration parameters are not jointly identifiable

- Loeppky et al. (2006); Arendt et al. (2012); Brynjarsdóttir and O'Hagan (2014); Tuo and Wu (2016).
- Our calibrated inputs have physical meaning.

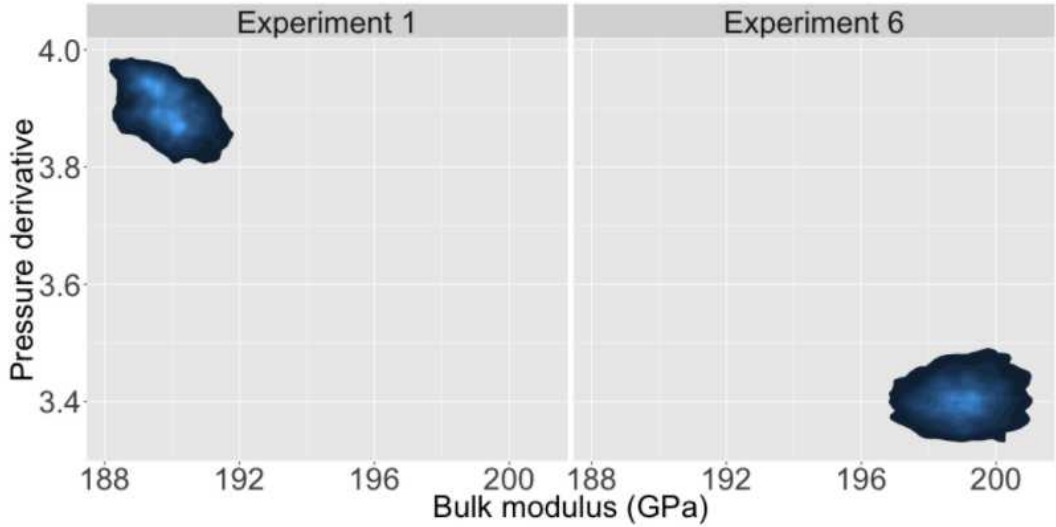
Model discrepancy

Important properties of the problem:

- No prior knowledge about where the model is wrong.
- Interested in physical interpretations of calibrated parameters - *not an interpolative prediction problem*.

KOH

KOH results for two different tantalum experiments.



KOH underestimates uncertainty for a single experiment.

- Some application may have only 1-2 experiments, others may have more (e.g. tantalum, with 9).

KOH

Recent solutions:

- Produce calibration results that are consistent under a desired loss function using a prior on the discrepancy that is orthogonal to the computer model gradient (Plumlee, 2017; Tuo, 2017) or two stage estimation procedure (Wong et al., 2017).

We still haven't solved the physical parameter estimation problem:

- Solution converges to minimizer of a loss function, not the best extrapolative prediction solution.
- Still estimating a discrepancy function.

KOH

Caveat: Physical parameter estimation with no assumptions about model discrepancy is an inherently intractable problem.

Approach: Try out alternatives to KOH.

- Solve the problem in different ways and compare results.

Power likelihoods

Rather than modeling the discrepancy, we can scale the log-likelihood to discount the statistical information for model misspecification (Bissiri et al., 2016).

In power-likelihood models, the posterior is a product of the prior and the likelihood raised to a power:

$$p(\theta|y) = exp(-wl(y|\theta))\pi(\theta)$$

Power likelihoods

With no discrepancy, the likelihood for a single experiment is simply an independent normal model:

$$y(x) \sim N(\eta(x, \theta), \sigma(x))$$

rather than multivariate normal:

$$y \sim MVN(\eta(\theta), \Sigma_\epsilon + \Sigma_\delta)$$

Misspecification in the model is handled (mostly) via w.

- How to choose the weight?

Power likelihoods - background

Relationship to prior work in calibration:

- Somewhat related to parameter uncertainty inflation methods used in model calibration (Sargsyan et al., 2015; Pernot and Cailliez, 2016; Pernot, 2017)
- Some precident for fractional likelihoods in calibration: Jackson et al. (2004); Mosbach et al. (2014), but no clear guidance on selection of w.

Power likelihoods - weight selection

How to choose the weights? Lots of recent discussion...

- Cross-validation - Gibbs posteriors (Jiang and Tanner, 2008)
- Condition on a relative entropy neighborhood (Miller and Dunson, 2015)
- **Change in information (Holmes and Walker, 2017; Brown and Hund, 2018)**
- **Calibrate a general credible region (Syring and Martin, 2017)**
- Prequentially minimize log-loss (Grünwald et al., 2017)

Power likelihoods - weight selection

Weight selection for functional calibration depends on the context...

- One experiment (function)
- Multiple experiments

Focus on experiment-specific weights.

- Experiment-specific inferences can be recombined assuming independence (Scott et al., 2016) or use weights in joint calibration.

Power likelihoods - weight selection

Case 1: Single experiment

- Change in information
- Calibrate a credible region

Change in information

Choose w by matching Fisher information under the true model to Fisher information under the power-likelihood misspecified model (Brown and Hund, 2018):

- Similar in spirit to Holmes and Walker (2017), who also use a change in information criteria.

$$
\begin{aligned}
\mathcal{M}_T : y &\sim MVN(\eta(\theta), \Sigma_\epsilon + \Sigma_\delta) \\
\mathcal{M}_F : y &\sim MVN(\eta(\theta), \phi\Sigma_\epsilon) \\
I_{\mathcal{M}_T} &= w I_{\mathcal{M}_F}
\end{aligned}
$$

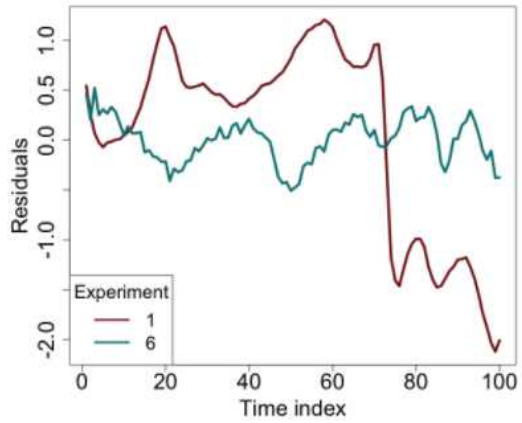
where $\mathcal{M}_T$ is the true model underlying the data and $\mathcal{M}_F$ is the fitted misspecified model.

Change in information

Under mean 0 stationary GP discrepancy, w asymptotically reduces to the inverse of the autocorrelation time from the model residuals.

$$\tau = 1 + 2\sum_{k=1}^{n-1}\left(1 - \frac{k}{n}\right)\nu(k)$$

where τ is the autocorrelation time, and $\nu(k)$ is the autocorrelation at lag k, i.e. the correlation between $\epsilon(x_i)$ and $\epsilon(x_{i'})$ where $|i - i'| = k$.



Change in information

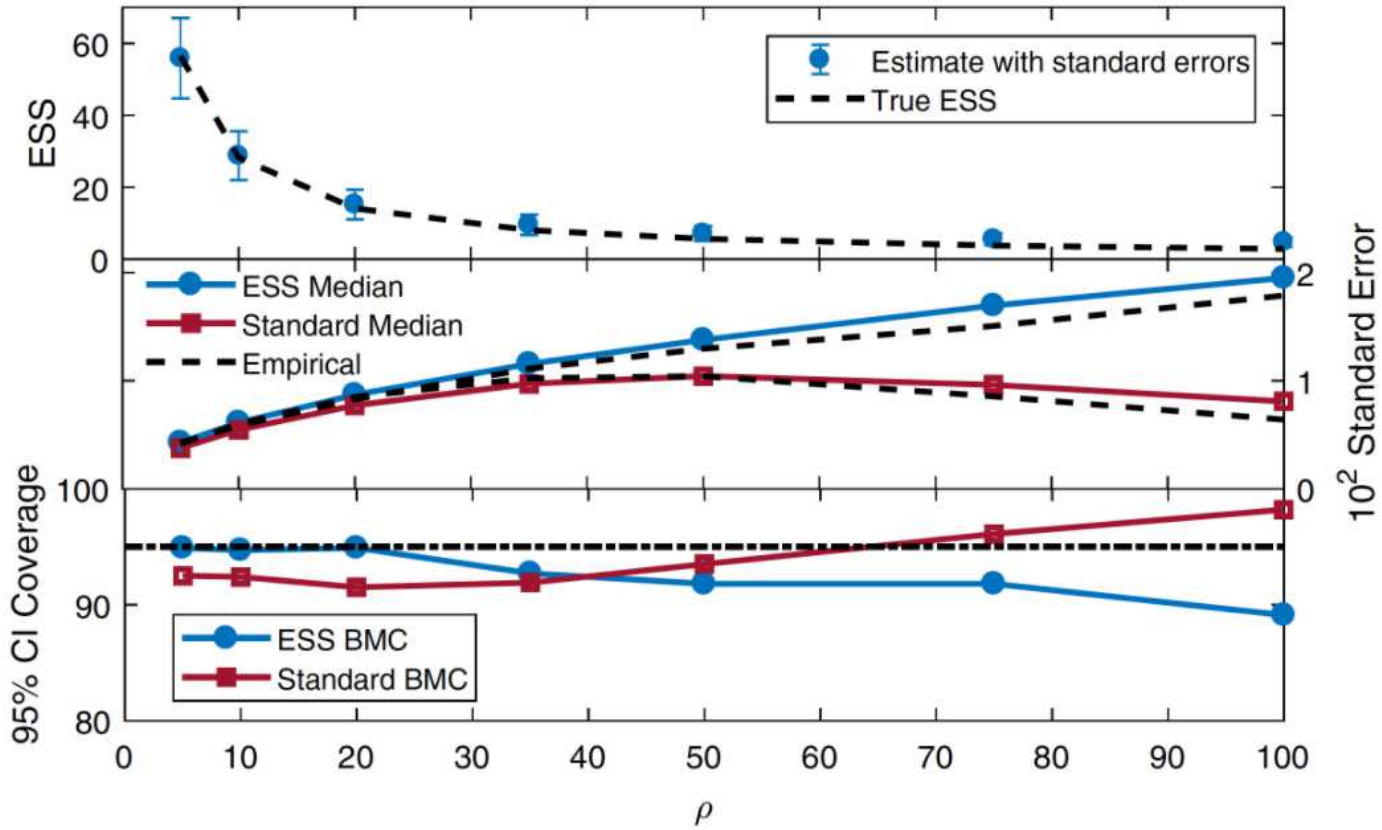
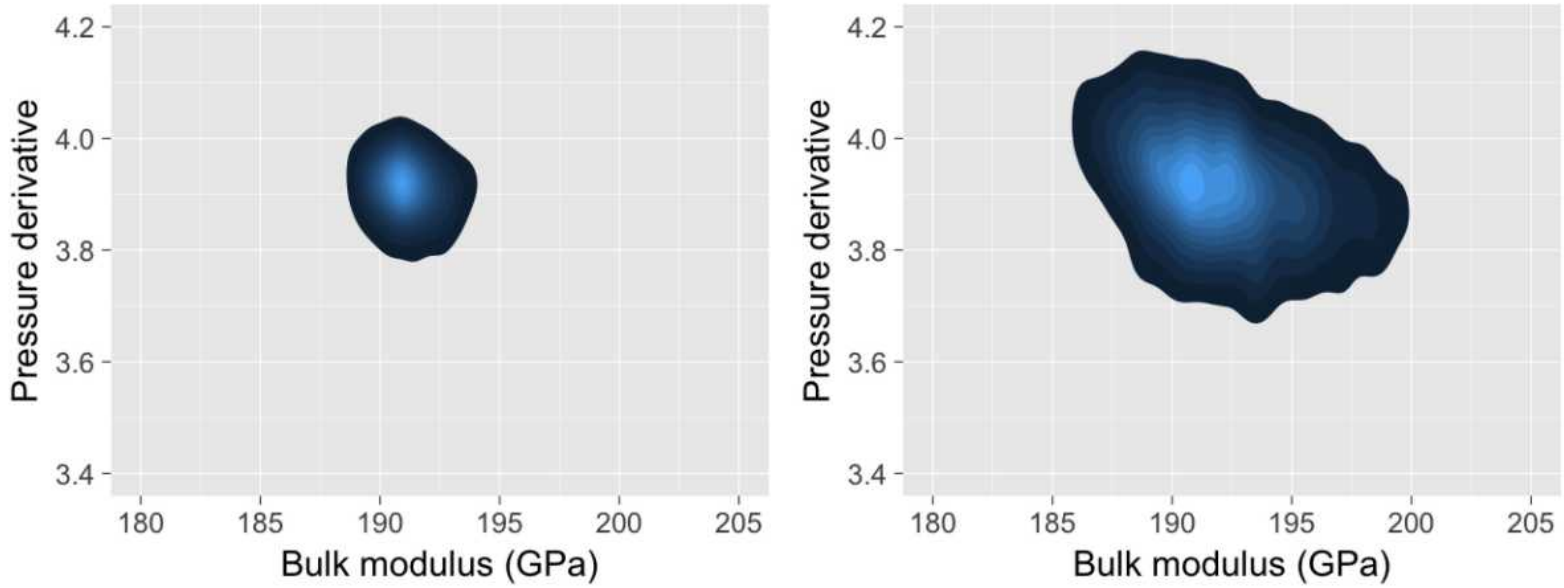
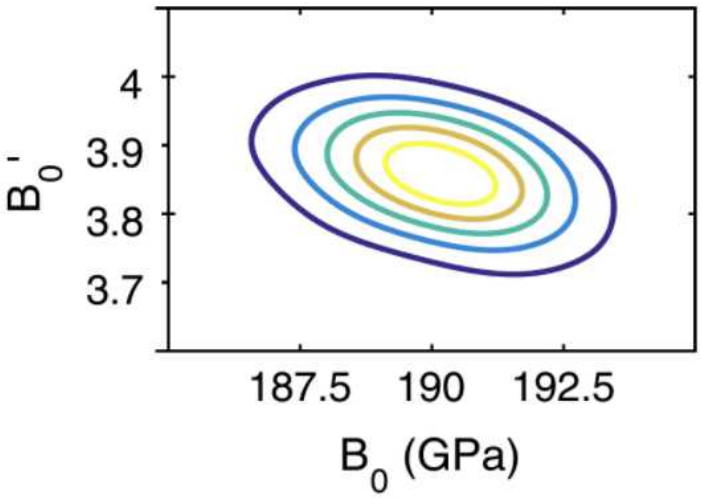


Fig. 3. Results of the simulation study: the likelihood scaling approach is comparable with the standard BMC approach and results in reasonable estimates for the physical parameter standard errors and coverage



Calibration for experiment 1 under (left) KOH model and (right) equivalent change in information model with $w = .05$ (effective sample size 5).

Using change in information model applied to all 9 experiments, results are consistent with previous analytic predictions for tantalum equation of state.



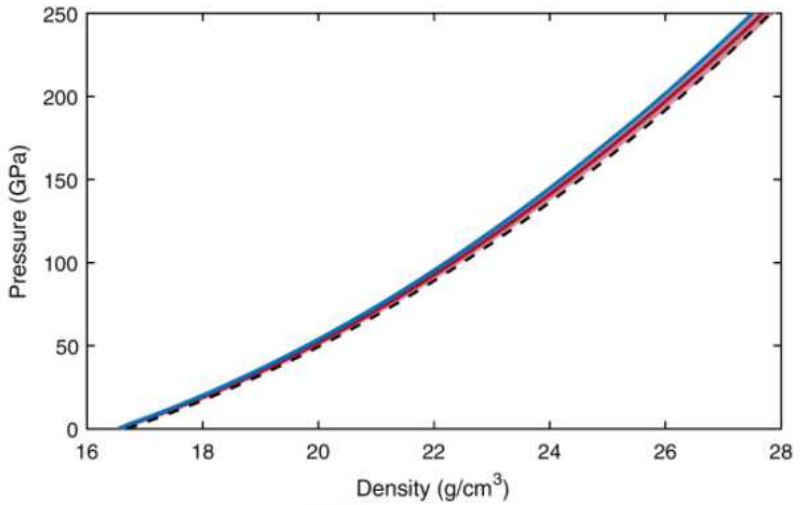


Fig. 9. Calibrated Ta material response (——) (equation (1)) compared with the analytic analysis in Brown *et al.* (2014) (——) and the theoretical calculations in Greeff *et al.* (2009) (– – –) (▬, calibration 95% credible interval)

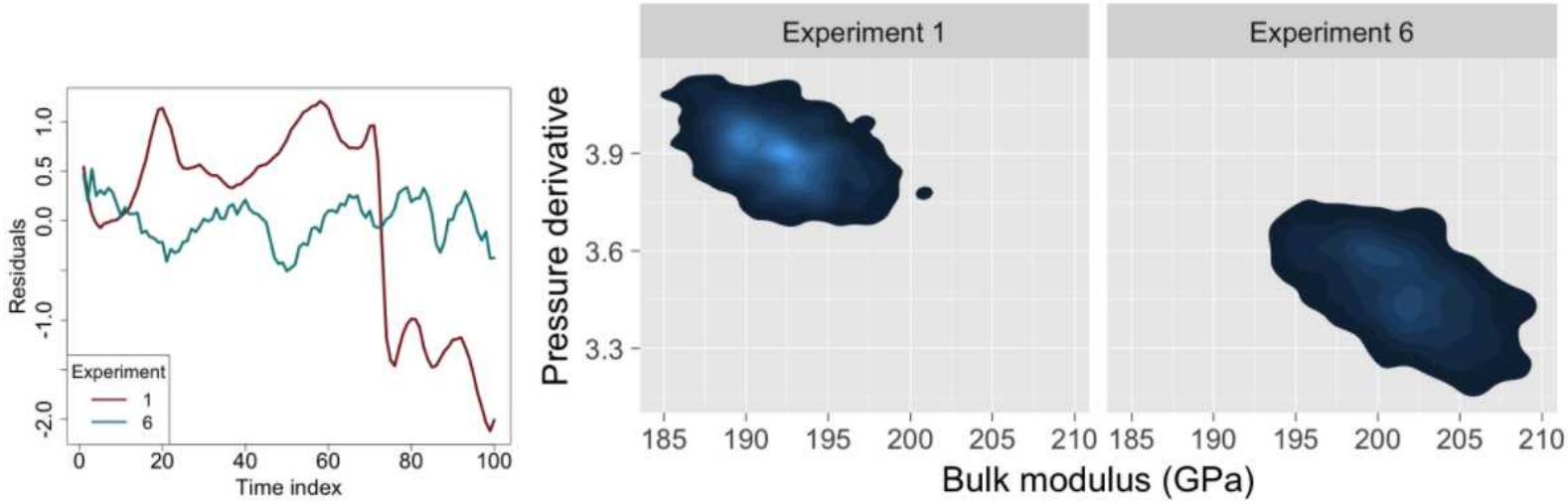
Change in information

Pros:

- Computationally nicer than KOH.
- More uncertainty than KOH (don't condition on discrepancy).

Limitations:

- Like GP hyperparameters, autocorrelation time is tough to estimate.
- The discrepancy is a mean 0 Gaussian process that is independent of the calibration parameters, a strong and typically unrealistic assumption.



(Left) Residuals from experiment 1 and 6 (standardized by measurement uncertainty; (middle) experiment 6 joint density estimate for $b0$ and $b0'$; (right) experiment 1 joint density estimate.

Power likelihoods - weight selection

Case 1: Single experiment

- Change in information
- **Calibrate a credible region**

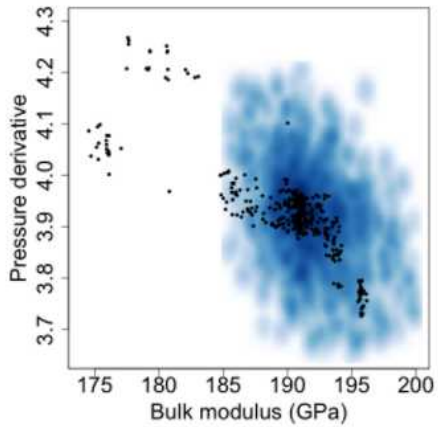
Calibrate credible region

Syring and Martin (2017) suggests choosing w to achieve $1 - \alpha$ coverage about θ.

$$\hat{w} \quad = \quad argmin_w |P(\theta(P) \in C_{w,\alpha}(Y)) - (1 - \alpha)|$$

where $C_{w,\alpha}(Y)$ is a credible region for $\theta(P)$ based on the data.

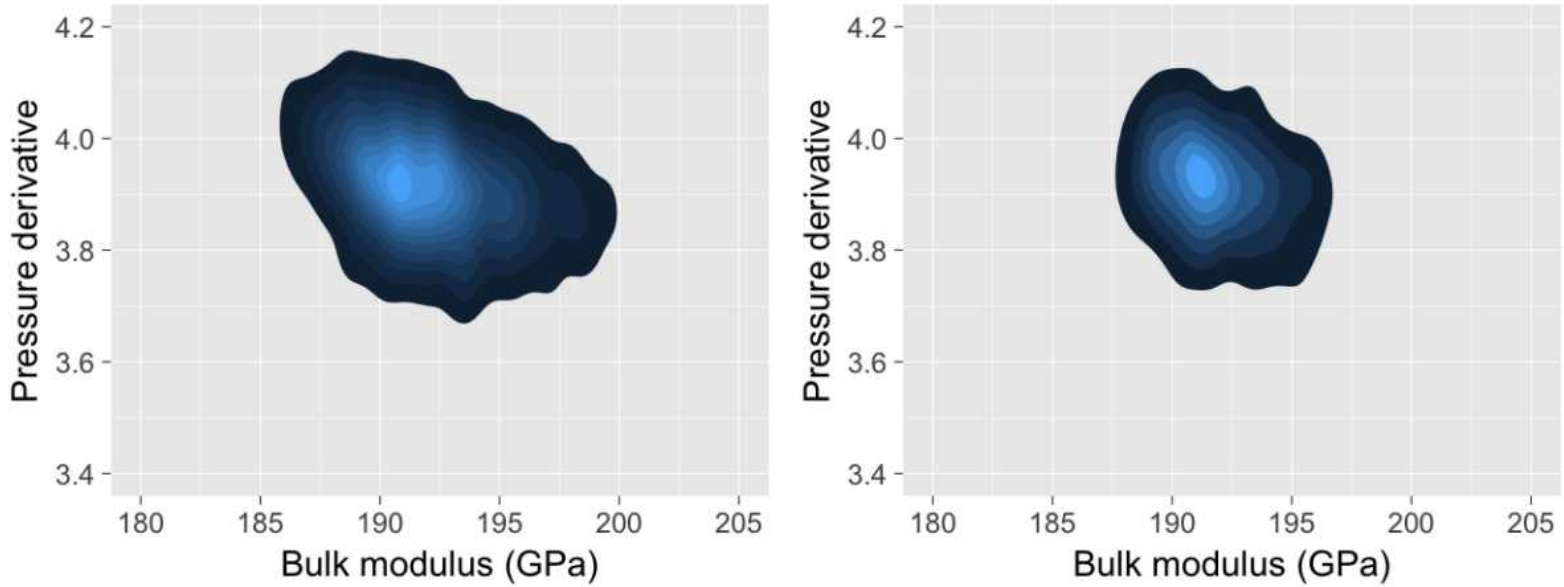
- $\theta(P)$ is approximated using bootstrap resampling.



Calibration data are not independent due to discrepancy, so use a block bootstrap.

- Currently choose blocks based on autocorrelation size.
- Method highly dependent on ability to partition into independent blocks - choosing blocks too small will underestimate uncertainty.
- Bootstrap could perform poorly in small samples.

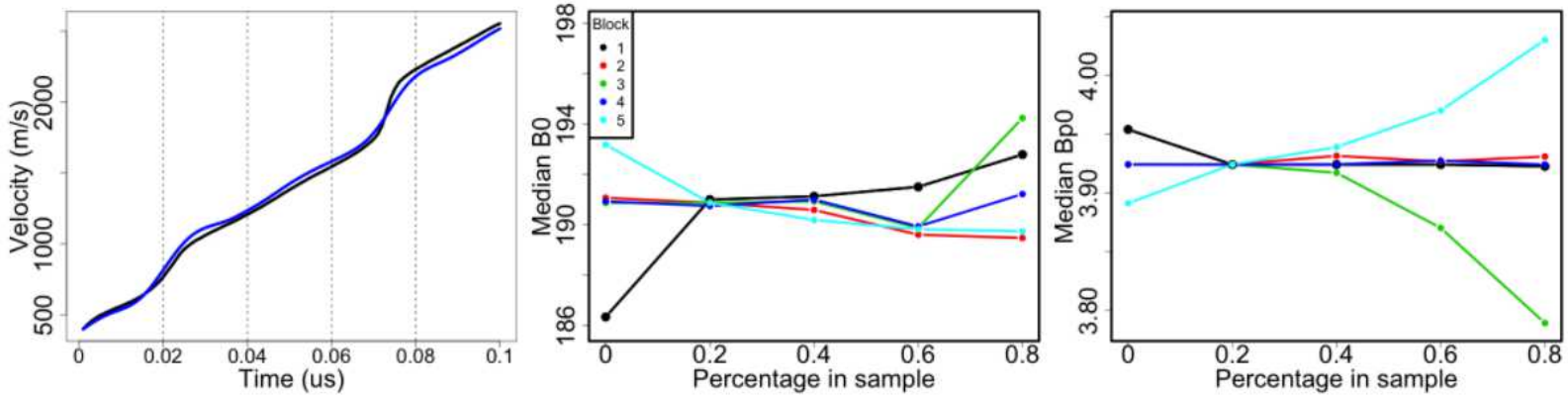
Calibrate credible region



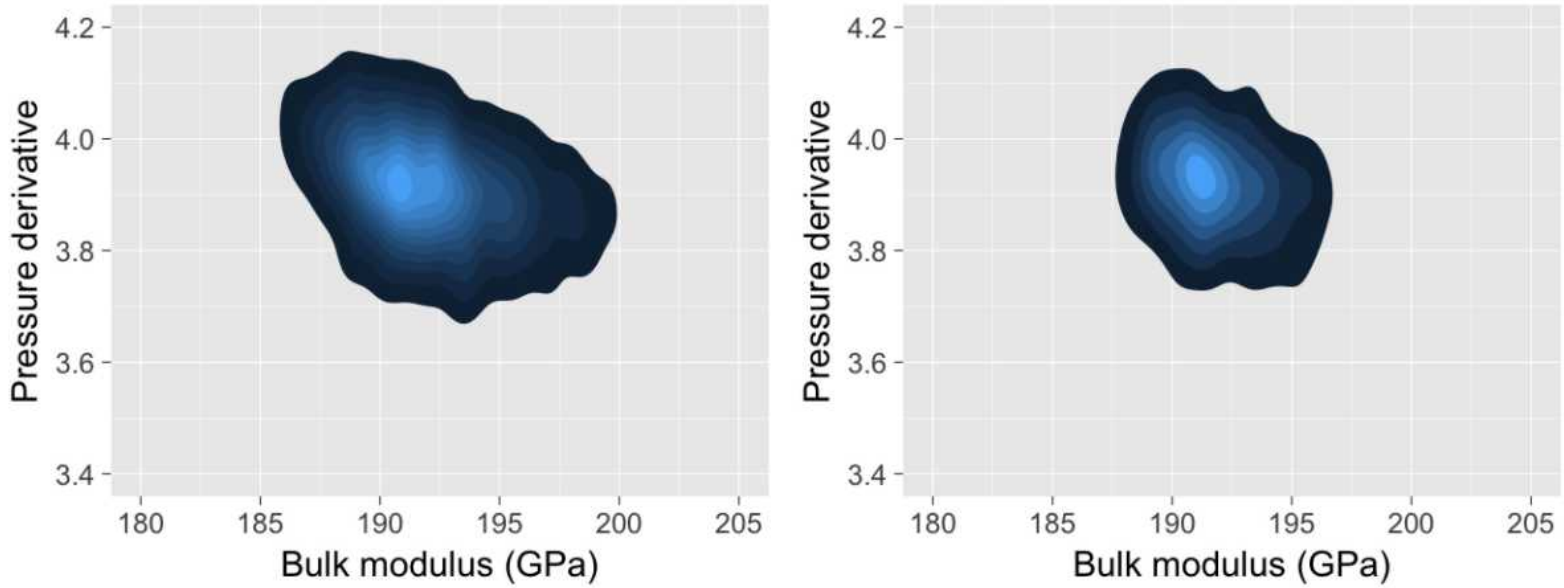
Calibration for experiment 1 under (left) change in information model ($w = .05$, ESS 5) and (right) equivalent block bootstrap model ($w = .075$, ESS 7.5).

Calibrate credible region

Block bootstrap provides nice model diagnostics regarding how blocks influence parameter values.



Calibrate credible region



Calibration for experiment 1 under (left) change in information model ($w = .05$) and (right) equivalent block bootstrap model ($w = .075$).

Alternative 'block' procedures could be used:

- Cross validation with metrics on posterior predictive distribution of the outcome (coverage or loss).
- Just use block bootstrap to directly estimate θ, rather than estimate w then θ.

Case 2: Multiple experiments

- Posterior prediction

Idea: Inferences from experiment k should be able to predict the outcomes of the remaining $k-1$ experiments.

Choose w for experiment k such that remaining $k-1$ experiments achieve $1-\alpha$ posterior predictive coverage.

$$\hat{w}_k \quad = \quad argmin_w |P(Y_{(-k)} \in C_{w,\alpha}(Y_k)) - (1-\alpha)|$$

Similar in spirit to Grünwald et al. (2017) prequential approach, but different metric and not actually prequential.

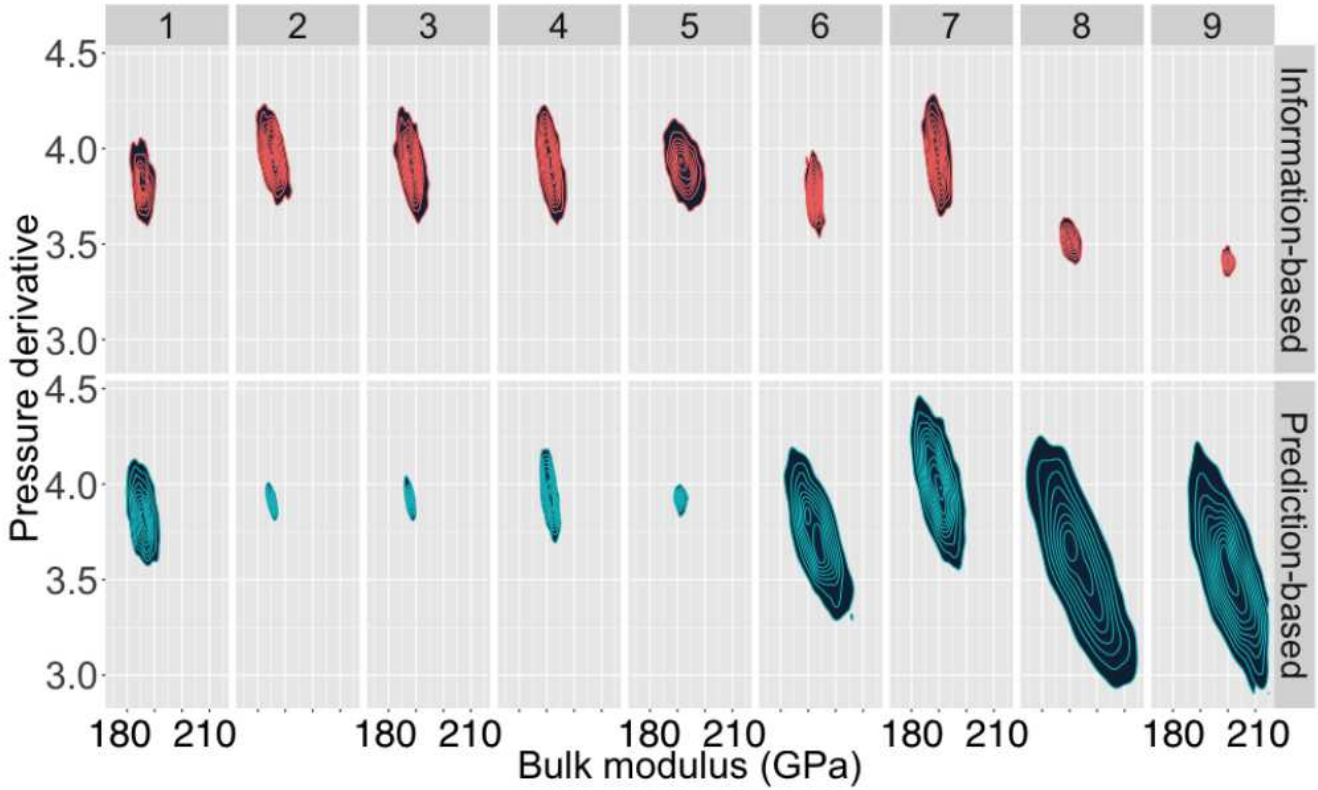
Challenges in implementation:

- How to incorporate experiment specific parameters into weight estimation?
- How to define σ in the likelihood?

$$
\begin{aligned}
y(x) &\sim N(\eta(x,\theta),\sigma(x)) \\
\sigma(x) &= \phi\sigma_{ME}(x) \text{ or} \\
\sigma(x) &= \sigma_{ME}(x)
\end{aligned}
$$

Measurement error is known and ignore experiment-specific parameters for now.

Higher pressure experiments, on average, result in more precise posteriors compared to information-based weights.



Conclusions

Limitations:

- Power likelihood methods still require some prior knowledge about the form of the discrepancy:
  - Change in information - mean 0 GP
  - Block bootstrap and predictive approach - independent partitions
- Computationally expensive
  - Find w using a grid search (aside from change in information weights).
  - Grünwald et al. (2017) prequential approach requires fitting n different models for each w.

Future work

- Consider different partitions
- Incorporate experiment-specific parameters in multiple experiment prediction

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 (2012).

Bissiri, P. G., Holmes, C., and Walker, S. G. "A general framework for updating belief distributions." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):1103–1130 (2016).

Brown, J. and Hund, L. "Model calibration via deformation." *Journal of the Royal Statistical Society, Series C* (2018).

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

Grünwald, P., van Ommen, T., et al. "Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it." *Bayesian Analysis*, 12(4):1069–1103 (2017).

Holmes, C. and Walker, S. "Assigning a value to a power likelihood in a general Bayesian model." *Biometrika*, 104(2):497–503 (2017).

Jackson, C., Sen, M. K., and Stoffa, P. L. "An efficient stochastic Bayesian approach to optimal parameter and uncertainty

estimation for climate model predictions." *Journal of Climate*, 17(14):2828–2841 (2004).

Jiang, W. and Tanner, M. A. "Gibbs posterior for variable selection in high-dimensional classification and data mining." *The Annals of Statistics*, 2207–2231 (2008).

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).

Loeppky, J., Bingham, D., and Welch, W. "Computer model calibration or tuning in practice." (2006).

Miller, J. W. and Dunson, D. B. "Robust Bayesian inference via coarsening." *arXiv preprint arXiv:1506.06101* (2015).

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).

Pernot, P. "The parameter uncertainty inflation fallacy." *The Journal of chemical physics*, 147(10):104102 (2017).

Pernot, P. and Cailliez, F. "A critical review of statistical calibration/prediction models handling data inconsistency and model inadequacy." *AIChE Journal* (2016).

Plumlee, M. "Bayesian calibration of inexact computer models." *Journal of the American Statistical Association*, 1–12 (2017).

Sargsyan, K., Najm, H., and Ghanem, R. "On the statistical calibration of physical models." *International Journal of Chemical Kinetics*, 47(4):246–276 (2015).

Scott, S. L., Blocker, A. W., Bonassi, F. V., Chipman, H. A., George, E. I., and McCulloch, R. E. "Bayes and big data: The consensus Monte Carlo algorithm." *International Journal of Management Science and Engineering Management*, 11(2):78–88 (2016).

Syring, N. and Martin, R. "Calibrating general posterior credible regions." *arXiv preprint arXiv:1509.00922* (2017).

Tuo, R. "Adjustments to Computer Models via Projected Kernel Calibration." *arXiv preprint arXiv:1705.03422* (2017).

Tuo, R. and Wu, J. "A theoretical framework for calibration in computer models: parametrization, estimation and

convergence properties." *SIAM/ASA Journal on Uncertainty Quantification*, 4(1):767–795 (2016).

Wong, R. K., Storlie, C. B., and Lee, T. "A frequentist approach to computer model calibration." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(2):635–648 (2017).