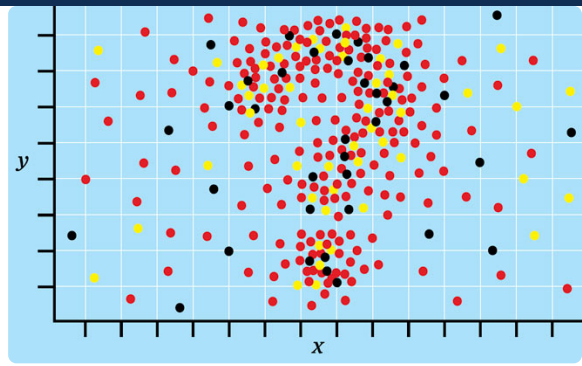
SAND2017-11066PE

Exceptional service in the national interest

Sandia National Laboratories

SANDIA STATISTICAL SCIENCES

Plan. Analyze. Inform.



Impacts of model uncertainty in data-driven decision-making at Sandia National Laboratories

Lauren Hund

October 13, 2017

U.S. DEPARTMENT OF ENERGY

NNSA National Nuclear Security Administration

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.

Acknowledgments

- This presentation represents joint work with Ben Schroeder and Justin Brown.

- The research was funded by the Laboratory Directed Research and Development program at Sandia National Laboratories.

- Sandia National Laboratories is a multi-mission laboratory managed and operated by National Technology & 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

- Introduction to Sandia and problem space
- Model uncertainty
- Examples
  - Statistical model uncertainty
  - Computational simulation model uncertainty

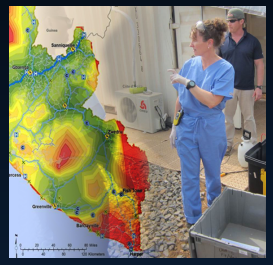
Sandia National Laboratories

SANDIA NATIONAL LABORATORIES AND THE MOTIVATING PROBLEM

Sandia National Laboratories

- "National security is our business" (SNL website).
  - Defense, Energy and Climate, Global Security
- Statistical Sciences group: 5 Masters-level, 6 PhD-level statisticians, 2 interns, and growing.
- UT-Austin is an Academic Alliance (AA) school.
  - "These are top-tier universities whose leadership has a commitment to collaborate on projects that have a mutual benifit and are willing to invest in growing the relationship with Sandia" (SNL AA website).
  - Sandia is incentivizing collaboration with AA schools, e.g. LDRD grants.
- My opinion: Decision sciences can help inform how modeling informs decisions in the SNL mission space.

Sandia's Impact

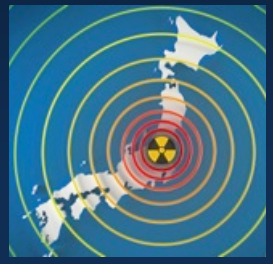


Ebola Outbreak

Sandia contributes to global response of Ebola outbreak by developing a sample delivery system cutting the wait time and potentially fatal exposure.

Cleanroom invented 1963

$50 billion worth of cleanrooms built worldwide. They're used in hospitals, laboratories and manufacturing plants today.

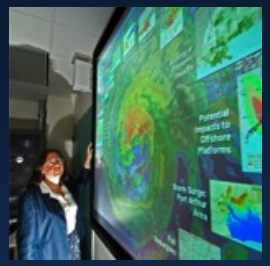


Fukushima Quake

Sandia helps clean up radioactive wastewater.

Detecting IEDs

Combat personnel now have a new tool for uncovering improvised explosive devices: Sandia's highly modified miniature synthetic aperture radar system, which is being transferred to the U.S. Army.



Hurricane Katrina

Sandia is called to assess flooding and infrastructure failures.

9/11

Sandia sets contingency plans for release of materials and aircraft attacks on critical facilities immediately after 9/11. Search dogs are equipped with cameras for search and rescue K-9 handlers. The capability allowed search efforts to be carried out in spaces inaccessible to humans.

Nuclear Deterrence

Sandia National Laboratories

Sandia assumes an increasingly pivotal role in sustaining the nation's nuclear deterrent.

QMU

- Quantification of Margins and Uncertainties (QMU) is a framework for **risk informed-decision making** about the nuclear weapon stockpile.
  - Assess the health of the stockpile in the absence of full system testing.

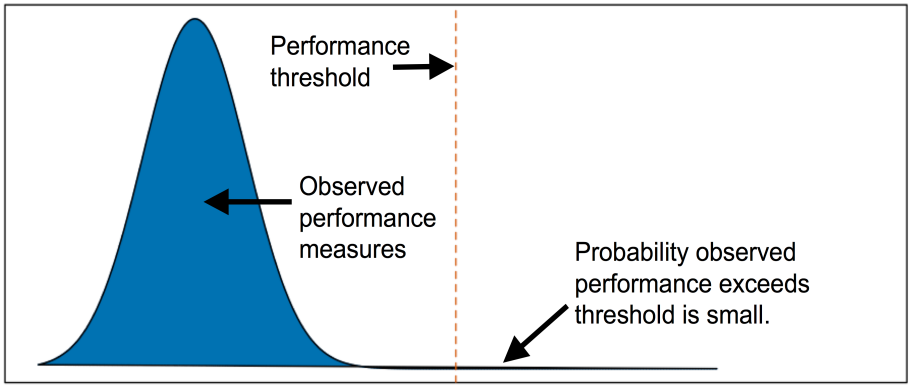
- Four components of information (Pilch, Trucano, & Helton, 2006):

1. What can go wrong?
2. How likely is it? (likelihood)
3. What are the consequences if it happens? (consequence)
4. How much confidence do we have in this risk assessment? (credibility)

QMU

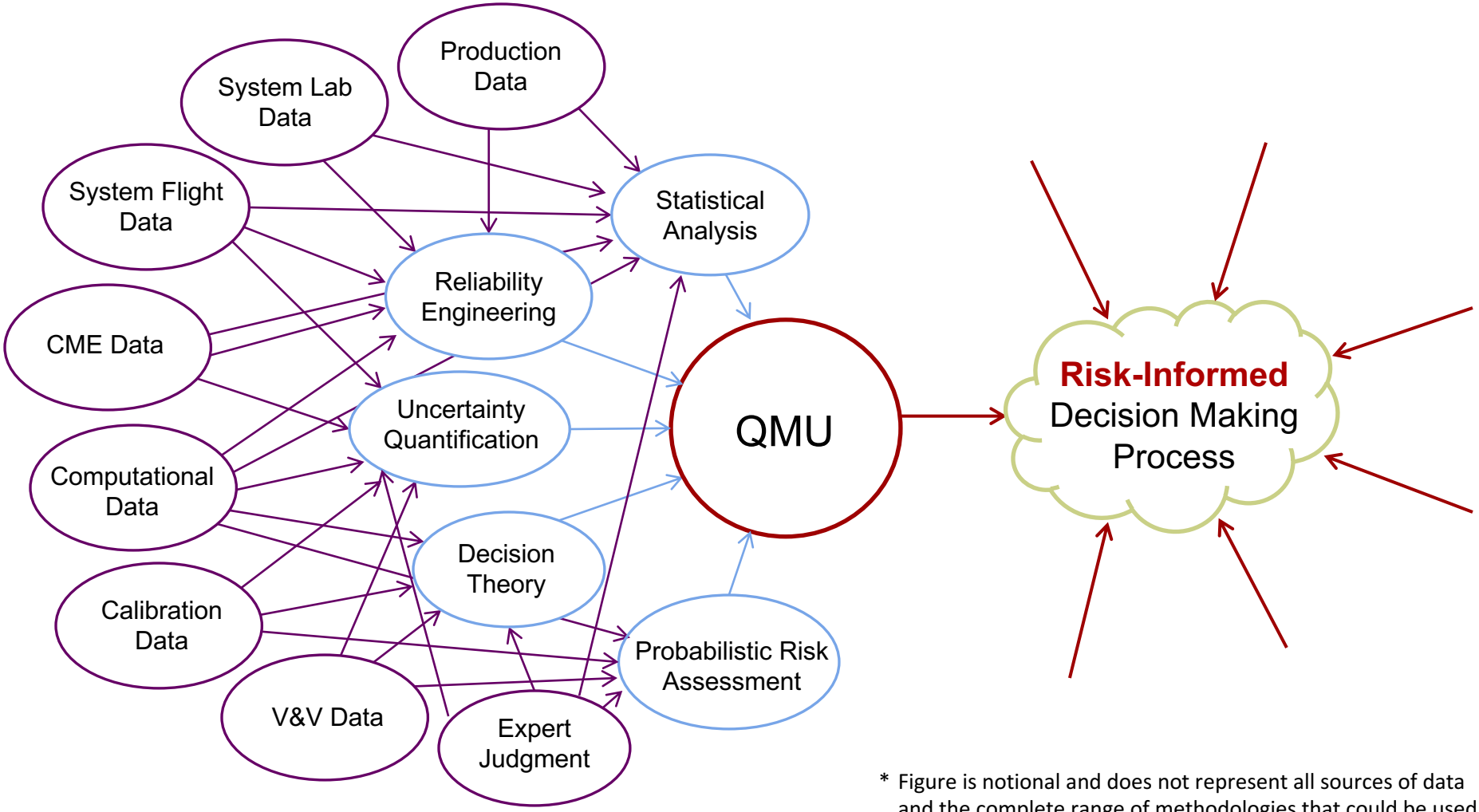
- **Goal:** Show that a performance measure will not exceed a performance threshold with high confidence.
  - A ***performance measure*** is a measure of performance related to system functionality and requirements.
  - A ***performance threshold*** is the required performance for the measure.

- **Complications:** Requirements must be met across a wide range of inputs (e.g. electrical), environments (e.g. mechanical, thermal), and over the specified life (e.g. 20 years).



The Complexity of QMU at Sandia

- QMU is a conceptual framework that ***is evolving*** over time that outlines a process for ***communicating*** the ***confidence*** in the stockpile.



* Figure is notional and does not represent all sources of data and the complete range of methodologies that could be used.

Complexity of QMU

Sandia National Laboratories

- QMU is a difficult prediction problem, with the goal of assembling an evidence package that:
  - Provides quantitative evidence of positive margin and
  - Includes a comprehensive treatment of uncertainty.

- Challenges are similar to nuclear power and climate.
  - Risk assessment, risk communication, decision-making under deep uncertainty.

"QMU is a collection of methods supporting NW decision making under uncertainty."

"QMU emphasizes and quantifies the presence of subjectivity."

"QMU is not a machine that produces numbers that highly influence a decision."

Pilch, Trucano, and Helton (2006)

Complexity of QMU

- **Ideal:** We are XX% confident that YY% of units will meet the requirement.
- **In practice:** We have high confidence that units will meet requirement. Here's why [insert evidence package].

- Barriers to "full quantification" in practice:

Test data have measurement uncertainty, are limited to a restrictive set of inputs and conditions, and are relatively few in number.

Models have model form uncertainty, numerical uncertainty, coding errors, and input distribution specification uncertainty.

Experts state of knowledge is imperfect.

Information comes from heterogeneous sources (test data, modeling, and expert judgment) that are difficult to combine.

Separation of aleatory and epistemic uncertainty is infeasible.

Not all uncertainties can be straightforwardly quantified.

Quantity of interest is poorly defined.

Sandia National Laboratories

MODEL UNCERTAINTY

Model uncertainty

- **My primary area of interest** is risk communication under model uncertainty (with a strong statistics slant).
  - How to choose a model/inferential procedure that is robust to model misspecification.
  - How to communicate the results from that model.

- **Earlier research**: Model selection in health effects estimation, survey design for resource allocation using lot quality assurance sampling.

- **Nowadays:** Communicating model uncertainty in QMU, developing a principled framework for QMU at SNL, model uncertainty in Bayesian model calibration of computer experiments.

Overall research question

- Does a modeling effort improve decision-making?
- How can we assess the added-value of increasing model complexity in QMU predictions?
  - Wrong models with little uncertainty
  - Better models with high uncertainty

> "Increasing a model's complexity… may actually increase the uncertainty."
>
> Pidgeon and Fischoff (2011)

Model uncertainty

- Is there a need for understanding the added value of statistical modeling more broadly?

> "Formal (statistical) models… serve a crucial but limited role in providing hypothetical scenarios that establish what would be the case if the assumptions made were true and the input data were both trustworthy and the only data available… Overconfident inferences follow when the hypothetical nature of these inputs is forgotten and the resulting outputs are touted as unconditionally sound scientific inferences instead of the tentative suggestions that they are (however well-informed)."
>
> Greenland (2017)

Sandia National Laboratories

EXAMPLES

Example: Percentile estimation

Sandia National Laboratories

- My interest in this research question stems from my experiences with one of the simplest examples: fitting a normal distribution to a set of data.*

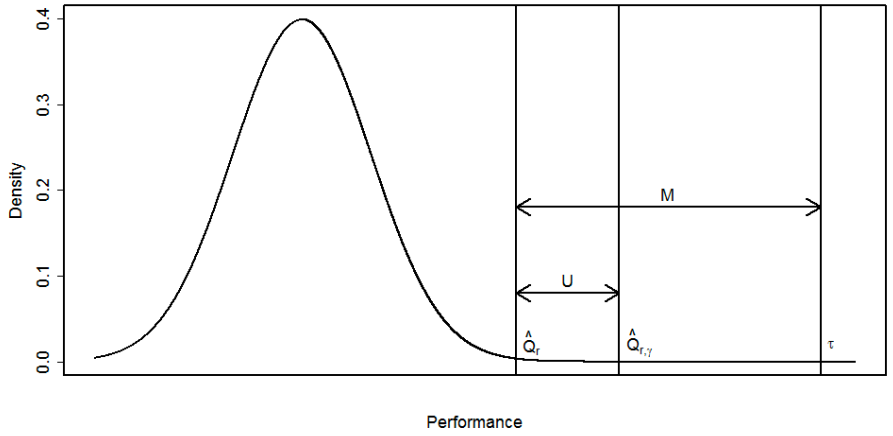
> "In nature, there never was a normal distribution" (Box 1976).

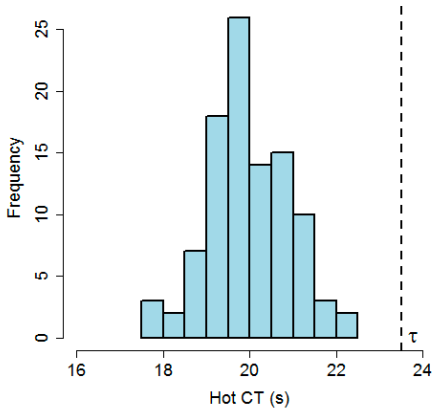
- "Simple" QMU applications fit parametric distributions to data and extrapolate extreme percentiles from this model.
  - This problem generated my hypothesis that education about and communication of model uncertainty is lacking.

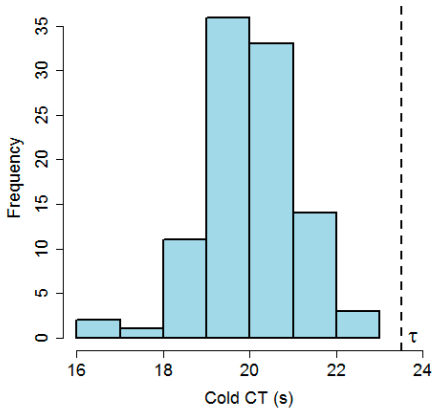
*We will talk about more complicated models later.

Model uncertainty in tolerance interval estimation

- Characterizing the "tails," i.e. extreme percentiles, of probability distributions plays a large role in QMU applications.
- Tolerance intervals are often used to "demonstrate margin" to a requirement when test data are available.
  - A one-sided tolerance bound is simply a confidence bound on a percentile.
- Example: a hypothetical launch safety device on a missile has a requirement to close within 23.5s of launch with 99.5% reliability.
  - Device is tested at hot and cold, n = 100 times per condition.
  - Are we 95% confident that 99.5% of units will pass the 23.5s requirement?
- When is there enough information or data such that this analysis is credible?







Current practice

- Use statistical tools to select a model for the data (often normal model).
  - QQ plots
  - Goodness of fit tests

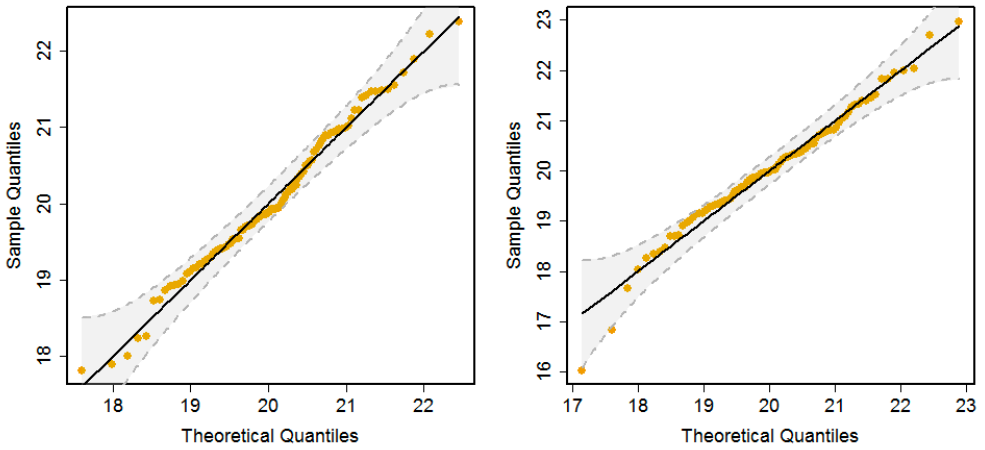


Figure: QQ plots for hot (left) and cold (right) closure time under a normal model.

- Extrapolate to the tails using the model.
  - Conclusion: we have 95% confidence that 99.5% of units will pass the requirement, *assuming the model is correct.*
  - Correct answer: Hot has margin, cold does not.

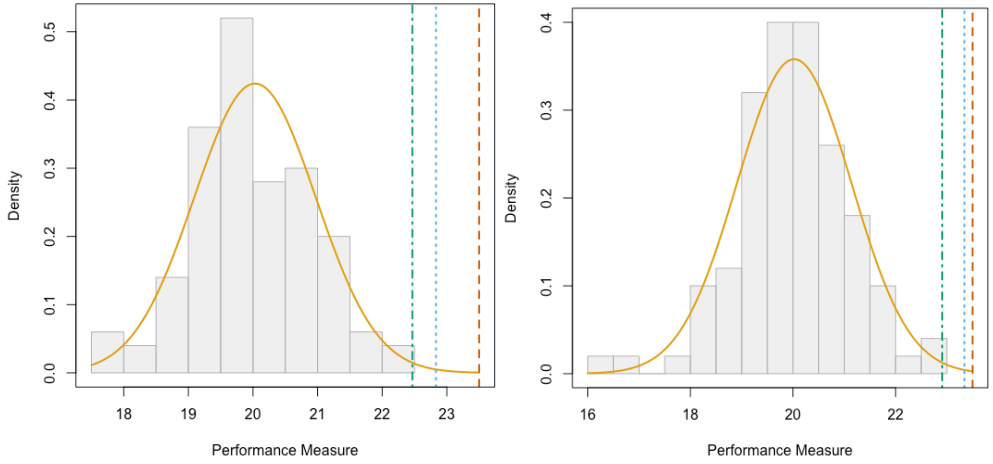


Figure: Percentile estimates (green) and 95% tolerance bounds (blue) relative to requirement (red) for hot (left) and cold (right) closure time under a normal model.

Model uncertainty in tolerance interval estimation

- **Problem:** modelers know model uncertainty decreases confidence in results, but lack a method to communicate this uncertainty.

> The [normal] tolerance interval... is not distributionally robust to even small deviations from normality" (Fernholz and Gillespie 2001).

> “Estimating tail parameters is analogous to estimating parameters `exterior to the data'... Many times estimates are made by assuming the data is sampled independently from a parametric family. This can lead to disastrous results” (Scholz 2005).

> “[Obtaining] a numerical estimate of reliability based on knowledge of full probability distributions in conjunction with QMU would place great demands on our ability to characterize uncertainties. In view of this, it is inevitable that there would be pressure to adopt ‘short cuts’ by simply assuming the forms of PDFs or using PDFs that are not based on some but inadequate supporting data. The response to such pressure would **make or break nuclear certification**. No analysis that is based on speculation or that neglects significant possibilities can lead to genuine confidence, but instead will frequently lead to over-confidence or under-confidence, both of which carry severe costs” (Sharp et. al 2003).

Model validation

- **QQ plots and goodness of fit tests** do not address model validity for extrapolative prediction, which is the objective.
  - **Goodness of fit test** can tell you there is evidence of lack of model fit, but cannot validate a model.
  - **QQ plots** tell you about the fit of your data over the entire range, rather than "for the intended uses of the model."

- **Solution:** relate statistical model selection to **model validation** in engineering applications.
  - **Validity:** Is the model an accurate representation of the real world for the intended uses of the model? (Oberkampf and Barone 2006).

Model validation

- **Metrics for model validation:**
  1. **Degree of extrapolation:** is extrapolation outside the range of the observed data occurring?
  2. **Model fit in the tails:** How consistent are the observed tails of the data with the fitted model?
  3. **Sensitivity to model choice:** How much do the tail estimates change when the modeling assumptions are relaxed?

- These metrics supplement/replace:
  - GoF test: Is there evidence that my model is not a good fit?
  - QQ plot: How well does the model fit all of the data?

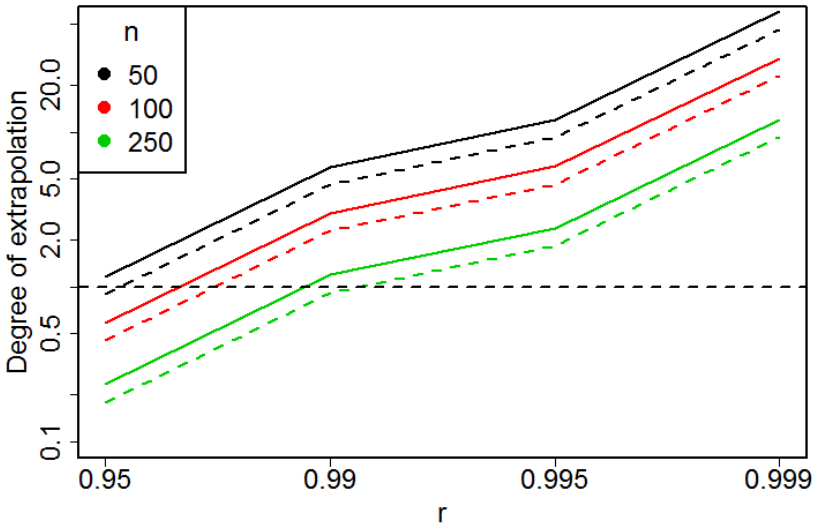
Degree of extrapolation

- How many samples would be needed for non-parametric estimation (i.e. no model required)?

$$\mathrm{n}^* = \log(1-\gamma)\,/\log(r)$$

where γ is the confidence level and r is the percentile (Wilks 1941).

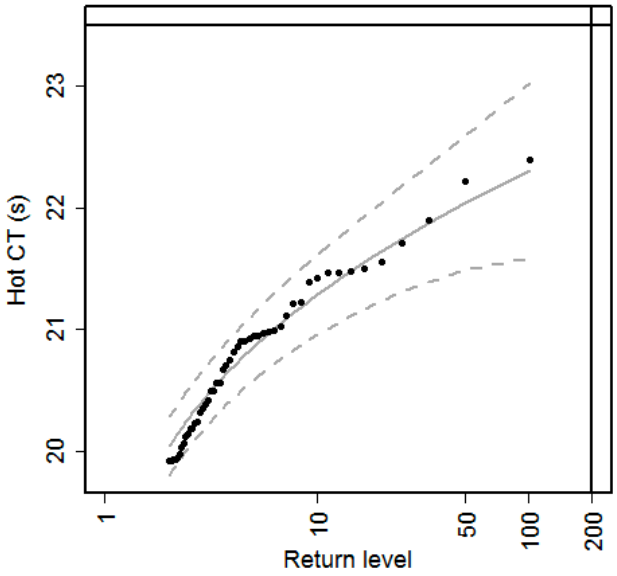
- Degree of extrapolation: n^*/n

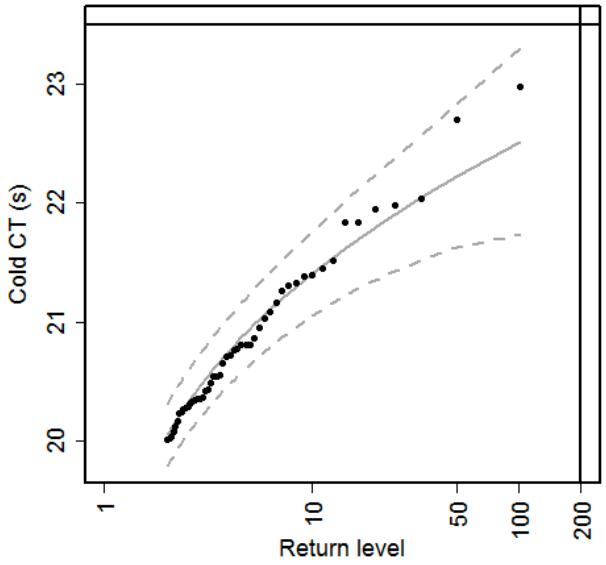


Degree of extrapolation as a function of r. Dashed line is 90% confidence, solid line is 95%.

Model fit in the tails

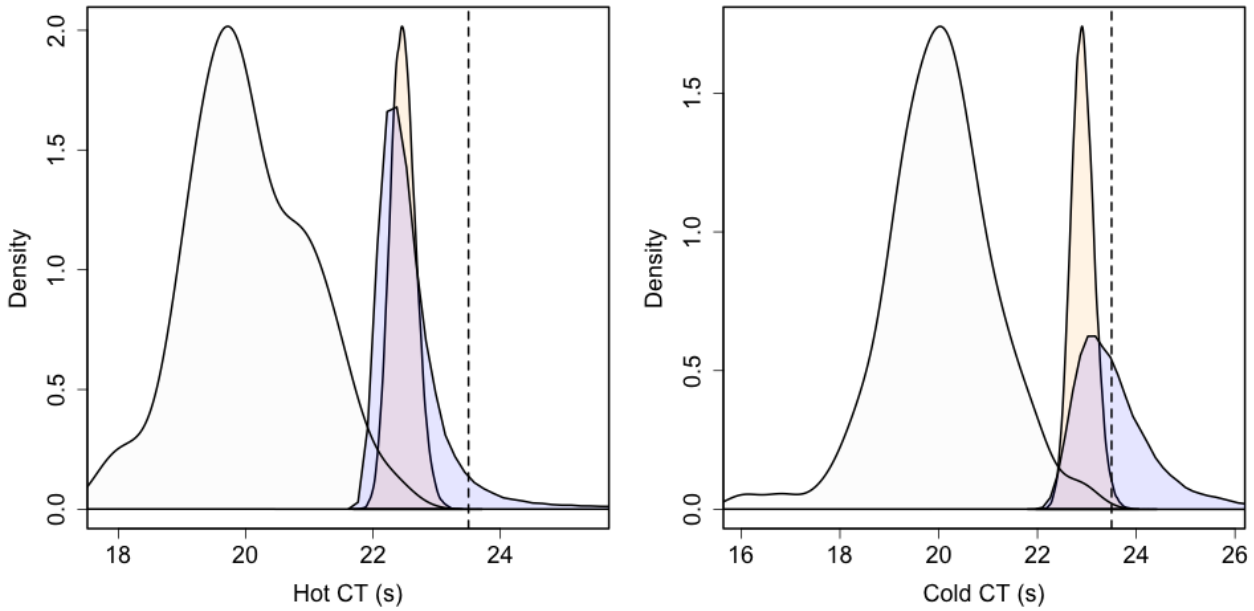
- Plot the outcome and model predictions (with CI) as a function of the return-level (Coles et. al 2001).
  - The return-level for percentile r is the number of units for which we expect one failure.
  - Mathematically, $n_R = \frac{1}{1-r}$.
  - Heuristically, a 99th percentile pertains to a 1 out of 100 failure rate.





Sensitivity to model choice

- How much do the tail estimates change when the modeling assumptions are relaxed?
- Developing a metric is a tougher question.
  - Need a relevant comparison model that relies only on the tail behavior; we used the generalized Pareto distribution from extreme value theory, fitted to the tails (e.g. upper 10% of data) only.
  - Important caveat: requires enough data to be able to realistically model the tails.



Validation metrics

- Need a metric for deciding whether the percentile estimates change substantively comparing parametric to extreme value model.
  - Area-based metric:

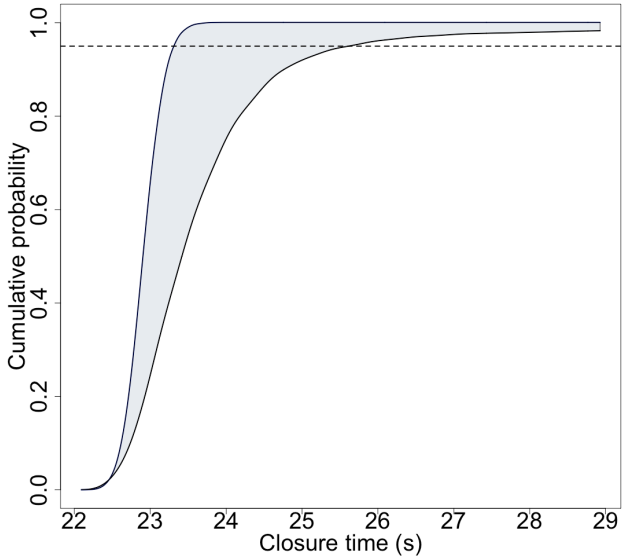
    $$\int |F^{Q_2}(x) - F^{Q_1}(x)| dx$$

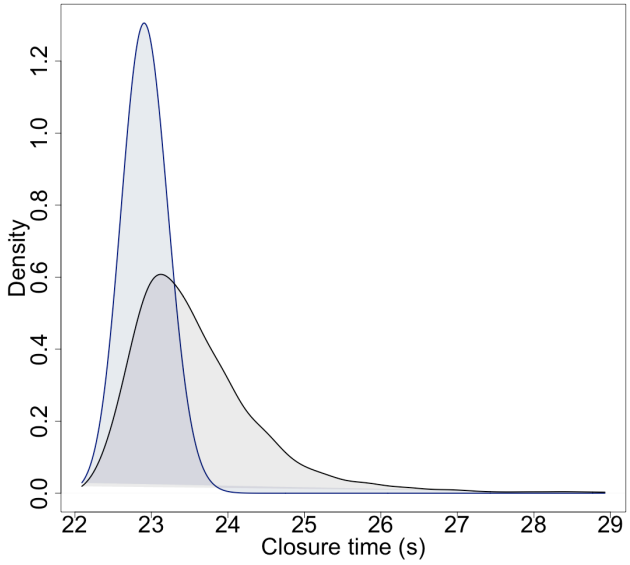
    - Pros: Equals 0 if distributions are identical.
    - Cons: Value difficult to interpret, must pick x range.
    - Alternative: Simple comparison of tolerance bounds
  - Reliability metric

    Two sided: $R = P(|Q_1 - Q_2| < \epsilon)$

    One-sided: $R = P(Q_1 - Q_2 < \epsilon)$

    - Pros: Interpretable as a probability.
    - Cons: Does not equal 0 when distributions are identical; must choose ϵ based on SME.





Metric properties

- Reliability metric has good properties:
  - Less data are required when we are willing to tolerate more error in the model predictions (ϵ).
  - More data are required as the target percentile moves further out in the tails of the distribution (p gets closer to 0 or 1).

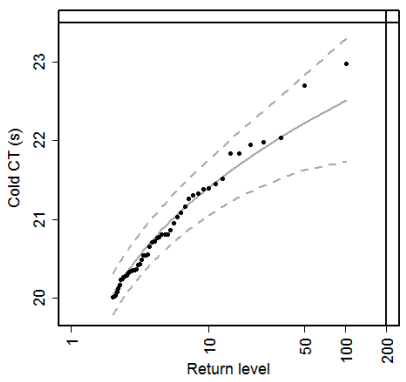
- Compare to goodness of fit tests:
  - Evidence of lack of model fit increases with sample size.
  - Does not consider how the model is being used.

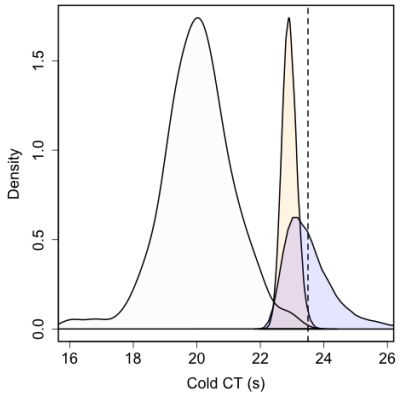
Example

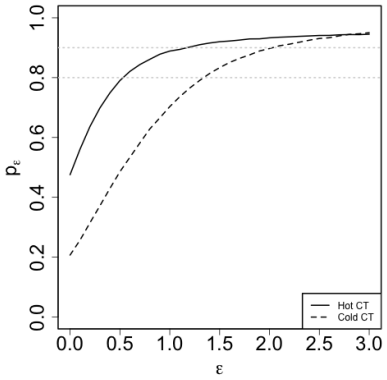
- **Closure time at cold temp:** estimate 99.5th percentile with 95% confidence.
  - **Degree of extrapolation**: When n = 100, extrapolation is occurring beyond the 97th percentile. We would a 6 times larger sample to avoid extrapolation.
  - **Validation metrics**: Median of the 99.5th percentile estimate is 23.4 s under tail-based model versus 22.9 s under normal.

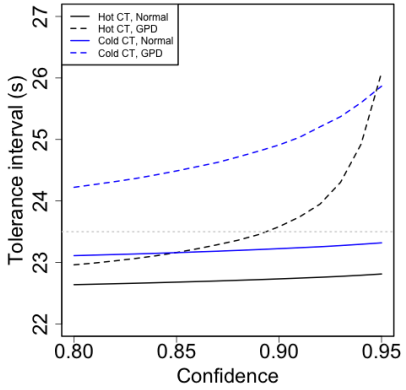
    If the decision-maker is willing to tolerate a 2-3s error in percentile estimation, then normal is likely a sufficient approximation.

    Difference between 95% tolerance interval estimates is 2.7s.









Conclusions

Sandia National Laboratories

- **Main message**: communication of statistical model uncertainty in engineering language may have added value in QMU applications.

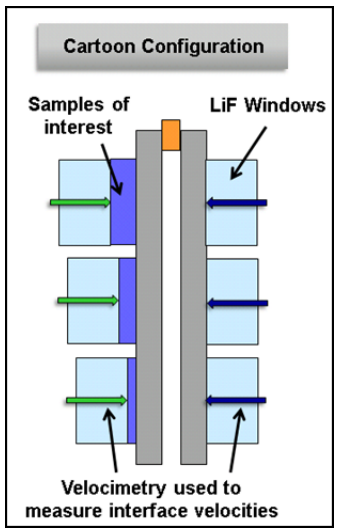
- **Future work:** We have only looked at the simplest case possible – the challenge of model uncertainty communication increases as models become more complex.
  - Does modeling have added value under model uncertainty?

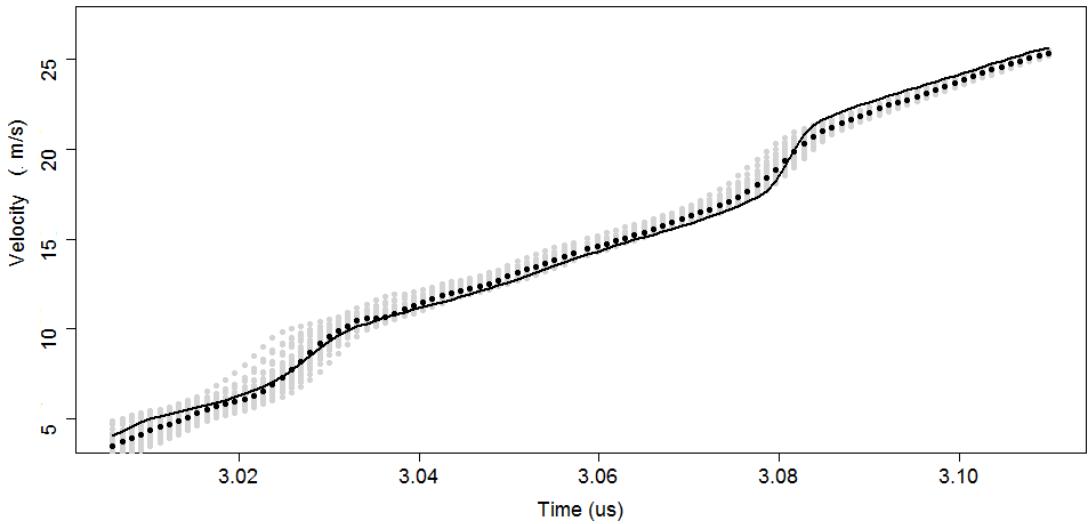
Sandia National Laboratories

A HARDER EXAMPLE

Background

- Sandia's Z machine 'provides the fastest, most accurate, and cheapest method to determine how materials will react under high pressures and temperatures' (SNL Z-machine website).
- The increasing complexities of these experiments are resulting in data which can longer be analyzed using traditional analytic techniques.
- An **inverse problem** must be solved by coupling **modeled** velocity profiles with experimental measurements. *Models are good, but not perfect.*





Background - model calibration

Bayesian model calibration is a popular framework for estimating the values of input parameters into computational simulation models in the presence of multiple uncertainties (Kennedy and O'Hagan, 2001; Bayarri et al., 2012).

- Kennedy and O'Hagan (2001) framework is widely used to estimate calibration parameters θ and model discrepancy δ given experimental output y, design points x, a simulation model η, and measurement error ϵ:

$$y(x) = \eta(x, \theta) + \delta(x) + \epsilon(x)$$

Background - Dynamic material properties

Goal: Estimate of 2 parameters of the equation of state for tantalum, the bulk modulus (B_0) and pressure derivative (BP_0), which inform how tantalum's density changes as a function of pressure and density.

- Velocity profiles were measured across 9 experiments at varying input pressure.
- Hydrocode simulations produce computational predictions of velocity as a function of time, considering as inputs:
  - Material properties: B_0 and BP_0 (shared), and
  - Nuisance parameters: Sample density, sample thicknesses, boundary condition/pressure.
- Uncertainties in the material property estimates are driven by:
  - Unknown model inputs,
  - Experimental measurement error, and
  - **Potential physics model misspecification.**

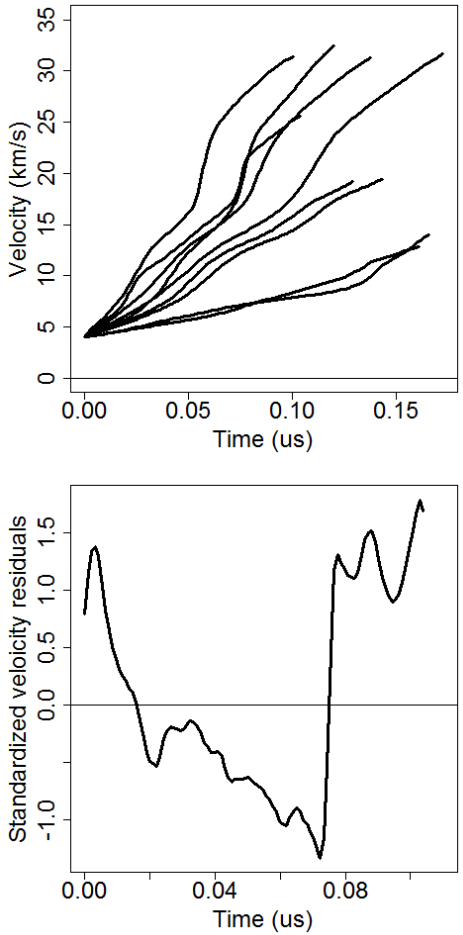


Figure: (Top) 9 experimental velocity curves. (Bottom) Standardized velocity residuals for a single experiment.

Data application

There are $9 \times 4 + 3$ total input parameters that could be calibrated.

Primary question: What is a reasonable calibration approach to estimate the material properties?

Challenges: How much statistical information about material properties is provided from experimental data under model uncertainty?

Model calibration - identifiability

Nonidentifiability: multiple values of the calibration parameters produce the equally valid solutions.

- Bayesian framework handles identifiability issues well when model is correctly specified.
- In the presence of model discrepancy, calibration parameters will typically be biased (Kennedy and O'Hagan, 2001; Brynjarsdóttir and O'Hagan, 2014).

In the presence of discrepancy, the problem of identifiability is often bypassed when models are used for **prediction**.

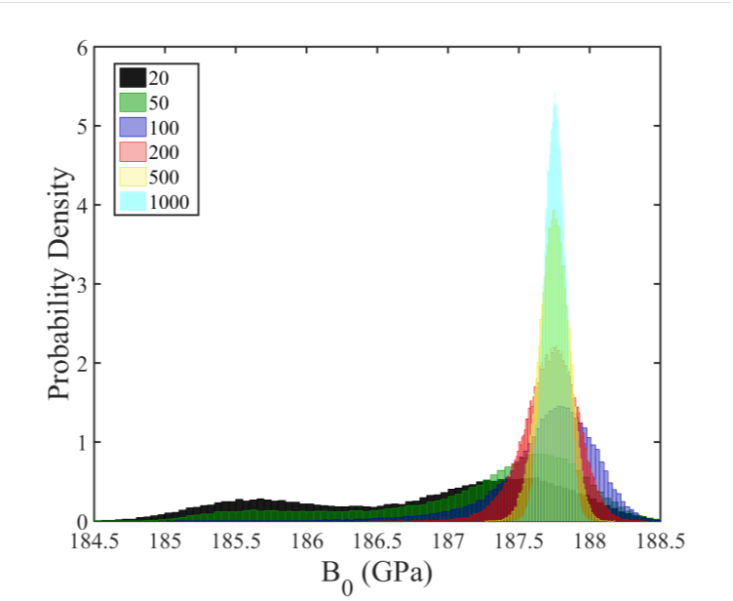
- Model is calibrated to a set of 'best fitting' (Kennedy and O'Hagan, 2001) parameters that do not typically have a physical interpretation but improve predictive capability.
- For **physical parameter estimation**, parameter identifiability must be carefully considered in order to obtain accurate and precise estimates (Arendt et al., 2012, 2016; Brynjarsdóttir and O'Hagan, 2014).

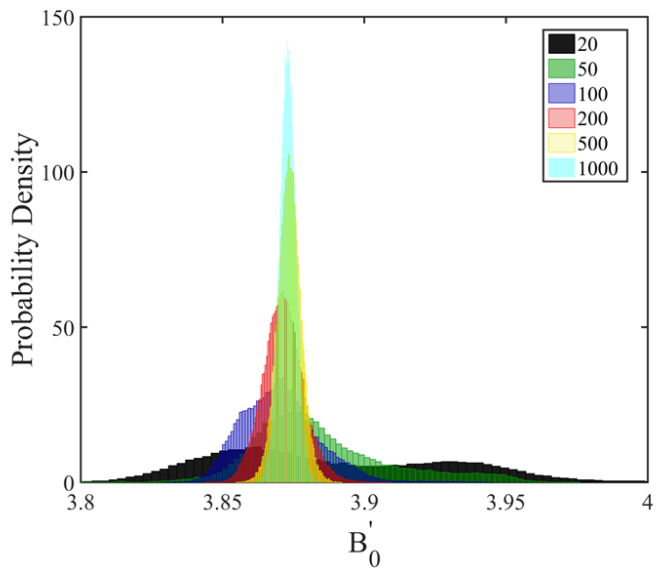
We are interested in the **true physical values** of the material properties.

Model calibration - functional output

Experimental outputs are functions - velocity profiles over time.

- While we can sample an arbitrarily large number of points from the measured velocity profiles, these curves only contain a finite amount of information about the values of the calibration parameters.
- Previous work on calibration with functional outputs focuses on prediction (McFarland et al., 2008; Bayarri et al., 2007; Williams et al., 2006).

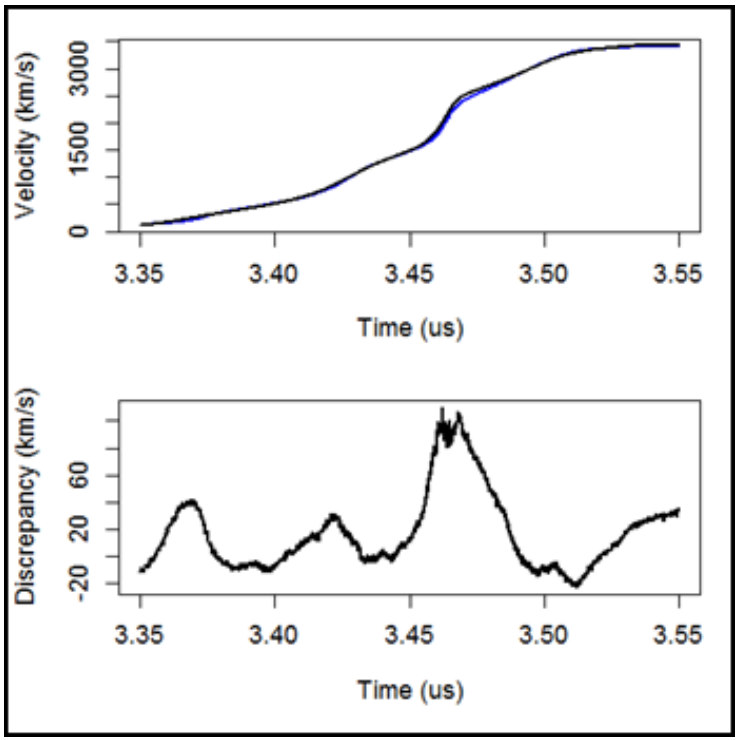




Discrepancy function

Discrepancy must be accounted for when quantifying uncertinaty on the physical parameters.

- Model discrepancy = temporal autocorrelation.
- Residuals are clearly autocorrelated over time.
- Pressure seems to impact parameter estimates.



Discrepancy function

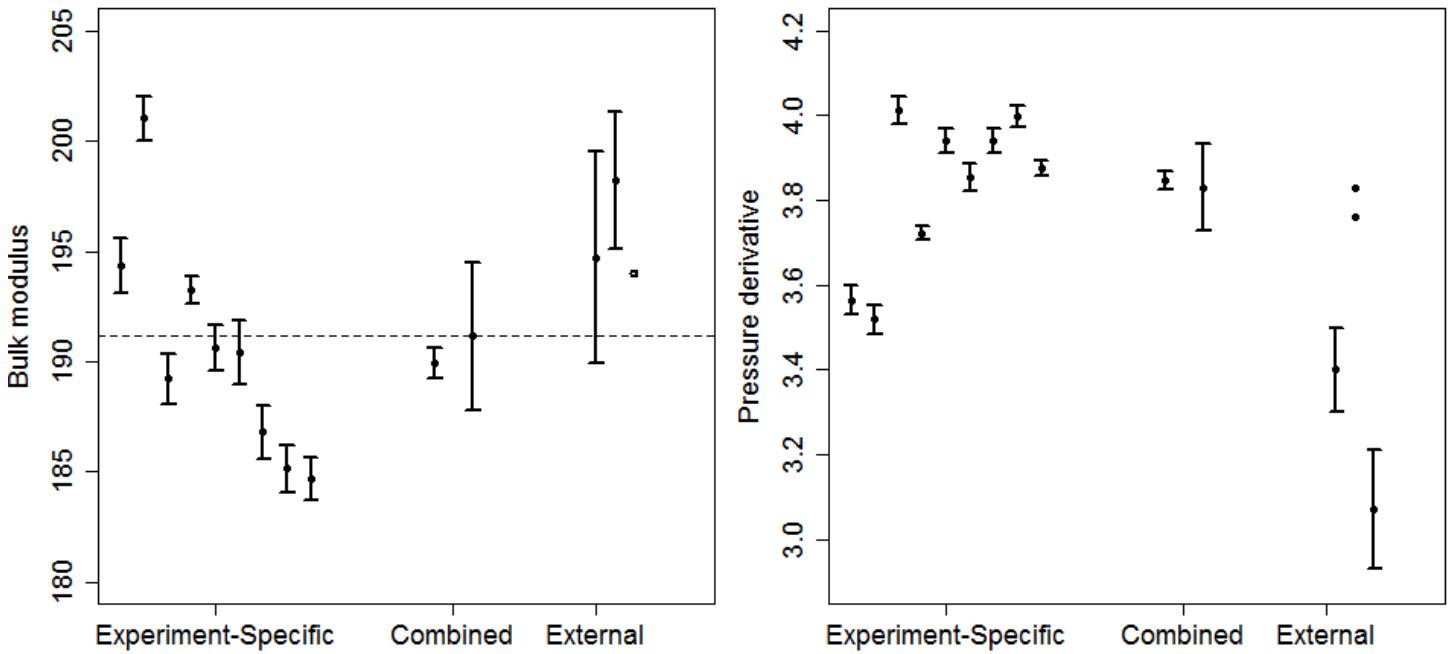
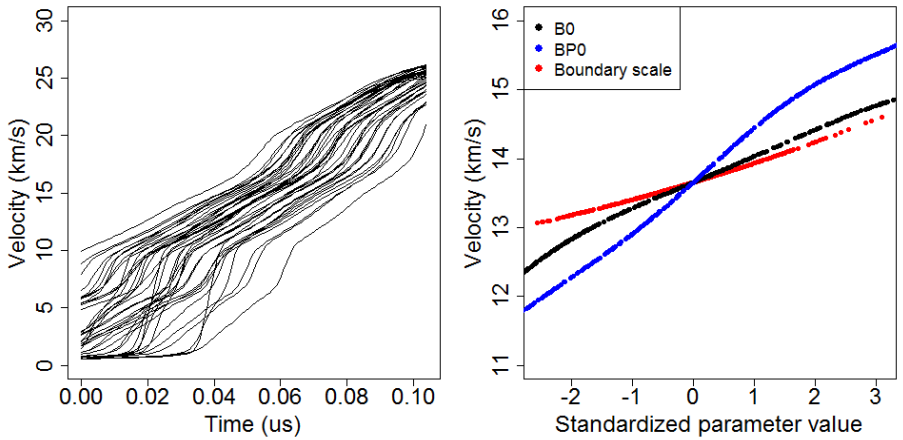


Figure: MLEs of bulk modulus (left) and pressure derivative (right) assuming no nuisance input parameter uncertainty and no model discrepancy. Estimates are provided for the 9 individual experiments, ordered by pressure ramping input P_j (experiment-specific), as well as pooled across all experiments (combined). For the combined estimates, the narrower estimate is based on the assumption that all experiments have the same material properties whereas the wider estimate allows experiment to experiment variability in the material property estimates. The external estimates come from published estimates.

Estimation approach

Steps for estimating the equation of state parameters are:

1. Generate multiple Monte Carlo realizations of the output given known uncertainties on the computer model inputs.
2. Build a Gaussian process emulator to obtain computationally cheap samples from the velocity output curve.
3. Determine how sensitive the velocity output is to each of the model inputs.
4. Calibrate the model to estimate the material properties.



(Left) Step 1: Simulate output realizations; (Right) Step 2: Build Gaussian process emulator for each experiment/time.

Calibration model

For physical parameter estimation, partition input paramters into: physical parameters and nuisance parameters.

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

$$y(x_{ij}) = \eta(x_j, \gamma_j^T, \theta) + \delta_j(x_j) + \epsilon(x_{ij}) \tag{1}$$

where

- y is experimental data, η is computer model simulator, ϵ is measurement error, x is time.
- θ is the true but unknown value of the **physical parameters**,
- $\gamma^T = [\gamma_j^T]$ is the true but unknown value of the **nuisance parameters**, and
- δ is a model discrepancy term.

Should we calibrate all nuisance parameters γ? How should we incorporate the model discrepancy term, δ?

Scale the likelihood

Rather than modeling the discrepancy function, we consider scaling the likelihood function to adjust for model discrepancy/temporal autocorrelation in the residuals when estimating physical parameters.

Rather than modeling the autocorrelation through Σ^{δ}, formulate the likelihood assuming velocity observations are IID normal:

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

Then, the likelihood is:

$$\log\, l(y|\alpha,\gamma,\phi) \quad \approx \quad \frac{n_e}{n}\log\, l^*(y_j|\alpha,\gamma_j,\phi) \quad (3)$$

Interpretation of n_e: the number of independent pieces of information provided from an experiment in the presence of model discrepancy.

Scale the likelihood

How to estimate the scaling factor n_e:

- Pick something that looks decent (Mosbach et al., 2014).
- Scale by the number of sampled points - assume all experiments are wrong but all are useful.
- Scale by n_e - assume each experiment has a mean 0 discrepancy function.

To estimate n_e, note that the effective sample size for an autocorrelated time series is:

$$\begin{aligned} n_e &= n/\tau \\ \tau &= 1 + 2\sum_{k=1}^{\infty} \nu(k) \end{aligned}$$

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

- $\nu(k)$ can be estimated using the sample correlation (non-parametric) or assuming a parametric form.

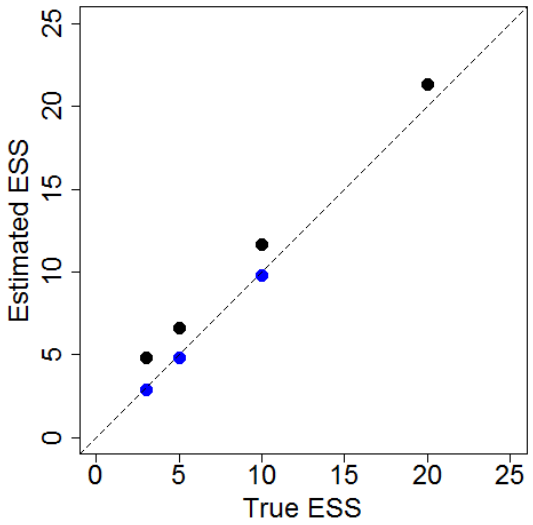


Figure: True versus median estimated ESS, comparing parametric (blue) to nonparametric (black) estimate of ESS. The parametric estimate for ESS $= 20$ failed to converge at least half of the time and is therefore not included in the plot.

Scale the likelihood

Advantages:

- Computationally cheap
- Computationally stable
- Estimates do not change with amount of discrepancy/temporal correlation.
- Variance of estimates increases with the amount of discrepancy/temporal correlation.
- GP discrepancy is non-identifiable anyway.

Disadvantages:

- Ad hoc?
- Discrepancy function is not directly estimated (but easy to obtain).

Application to tantalum EOS

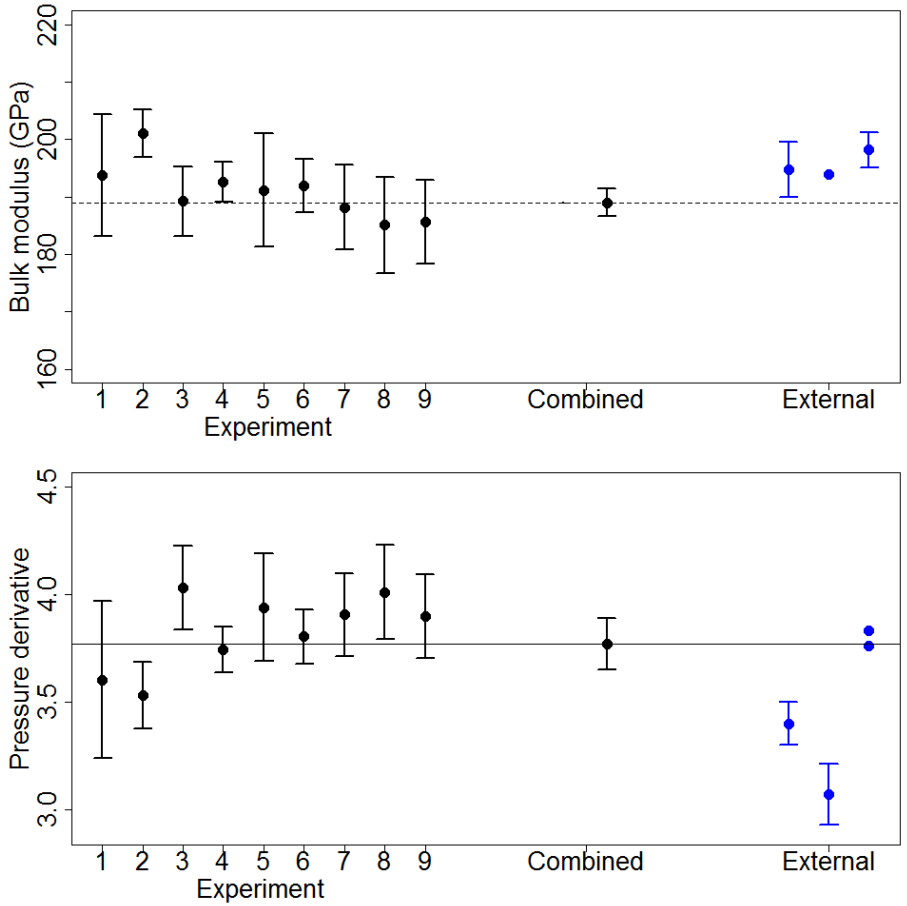


Figure: Estimates of bulk modulus (top) and pressure derivative (bottom). (Right) Maximum likelihood estimates with 95% CIs are provided for the 9 individual experiments, ordered by pressure ramping input P_j (experiment-specific). (Middle) BMC estimates pooling across all experiments (combined). (Left) The external estimates come from published estimates of the material properties.

Are we underestimating uncertainty using full Bayesian model calibration?

- Updating 37 nuisance calibration parameters using BMC will decrease variance in physical parameter estimates.
- In the presence of non-mean 0 model discrepancy, the calibration parameters will be biased.
- The model is definitely wrong (but we are unsure of how wrong).
- Should all nuisance parameters be updated in the MCMC?

Modularization

To 'modularize' the nuisance parameters from the experimental data, we simply do not update the values of the nuisance parameters using the data and assume $\pi(\gamma|y) = \pi(\gamma)$.

- The idea of modularization is used throughout the calibration process: building emulators; estimating parameters using MLEs (Liu et al., 2009), for the computational advantages.

Using this approach, the 'posterior' is:

$$\begin{aligned} \pi(\alpha, \gamma, \phi|y) &= P(\alpha, \phi|y, \gamma)\pi(\gamma|y) \\ &= P(\alpha, \phi|y, \gamma)\pi(\gamma) \\ P(\alpha, \phi|y) &= \int_{\gamma} P(\alpha, \Sigma|y, \gamma) f_{\gamma}(\gamma)\, d\gamma \end{aligned} \tag{4}$$

where P denotes a posterior and f denotes priors and

$$P(\alpha, \Sigma|y, \mathcal{D}^s, \gamma) \propto l(y|\alpha, \Sigma, \mathcal{D}^s, \gamma) f_{\alpha}(\alpha) f_{\Sigma}(\Sigma) \tag{5}$$

Modularization does not result in full Bayesian inference and the estimated posterior distributions to not converge to a true posterior (Plummer, 2015).

- Philosophically appealing to assume the collected data cannot inform the values of the nuisance parameters under discrepancy.
- Induces bias similar to leaving the outcome out of a multiple imputation model.
- Modularization has been applied in other areas to reduce model misspecification bias (Plummer, 2015; Zigler et al., 2013).

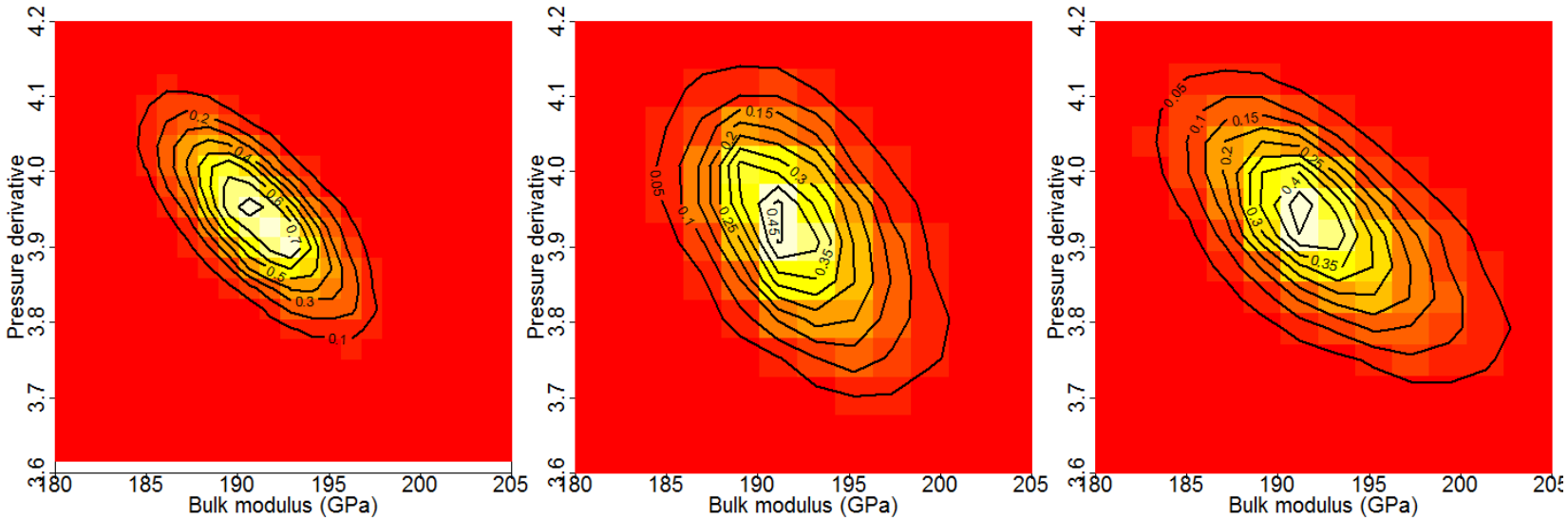


Figure: Posterior distribution of the bulk modulus and pressure derivative from a single experiment using different estimation procedures: (left) maximum likelihood approximation ignoring nuisance parameter uncertainty; (middle) Bayesian model calibration inferring only sensitive nuisance parameters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

Summary

In summary, we are working toward developing a general inverse analysis technique for dynamic material properties rooted in Bayesian model calibration.

- The estimates and uncertainties on the tantalum EOS parameters are reasonable and consistent with those from traditional analytic techniques.
- Future work
  - Improve likelihood scaling factor.
  - Determining if, when, and how to 'modularize' in general.
  - Make functional calibration more 'functional.'

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 (2012).

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

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).

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

Coles, S., Bawa, J., Trenner, L., and Dorazio, P. *An introduction to statistical modeling of extreme values*, volume 208. Springer (2001).

Greenland, S. "For and against methodologies: some perspectives on recent causal and statistical inference debates." *European journal of epidemiology*, 32(1):3–20 (2017).

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).

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).

Newcomer, J. T. "A new approach to quantification of margins and uncertainties for physical simulation data." *SAND2012-7912* (2012).

Oberkampf, W. L. and Barone, M. F. "Measures of agreement between computation and experiment: validation metrics." *Journal of Computational Physics*, 217(1):5–36 (2006).

Pilch, M., Trucano, T. G., and Helton, J. C. "Ideas underlying the quantification of margins and uncertainties." *Reliability Engineering & System Safety*, 96(9):965–975 (2011).

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

Sharp, D., Wallstrom, T., and Wood-Schulz, M. "Physics package confidence: "one" vs. "1.0"." *Proceedings of the NEDPC 2003* (2003).

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*

Questions?

Sandia National Laboratories

Thank you!