LA-UR-14-25968
Approved for public release; distribution is unlimited.

Title: Data Assimilation - Advances and Applications

Author(s): Williams, Brian J.

Intended for: use as future presentation materials

Issued: 2014-07-30

MeV

Modeling • Experimentation • Validation

SUMMER SCHOOL

www.MeVSchool.org

Data Assimilation – Advances and Applications

Brian Williams

Statistical Sciences Group

Los Alamos National Laboratory

Session #7

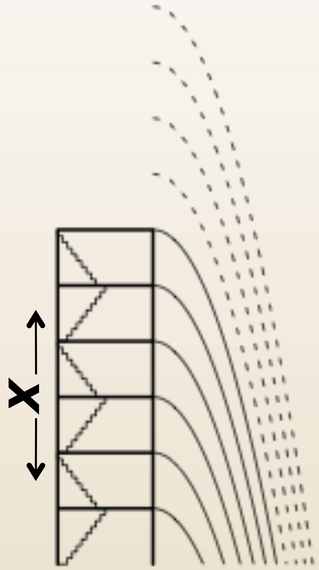
23 July 2014

10 – 11:30 AM

Abstract

This presentation provides an overview of data assimilation (model calibration) for complex computer experiments. Calibration refers to the process of probabilistically constraining uncertain physics/engineering model inputs to be consistent with observed experimental data. An initial probability distribution for these parameters is updated using the experimental information. Utilization of surrogate models and empirical adjustment for model form error in code calibration form the basis for the statistical methodology considered. The role of probabilistic code calibration in supporting code validation is discussed. Incorporation of model form uncertainty in rigorous uncertainty quantification (UQ) analyses is also addressed. Design criteria used within a batch sequential design algorithm are introduced for efficiently achieving predictive maturity and improved code calibration. Predictive maturity refers to obtaining stable predictive inference with calibrated computer codes. These approaches allow for augmentation of initial experiment designs for collecting new physical data. A standard framework for data assimilation is presented and techniques for updating the posterior distribution of the state variables based on particle filtering and the ensemble Kalman filter are introduced.

Calibration With Model Form Error



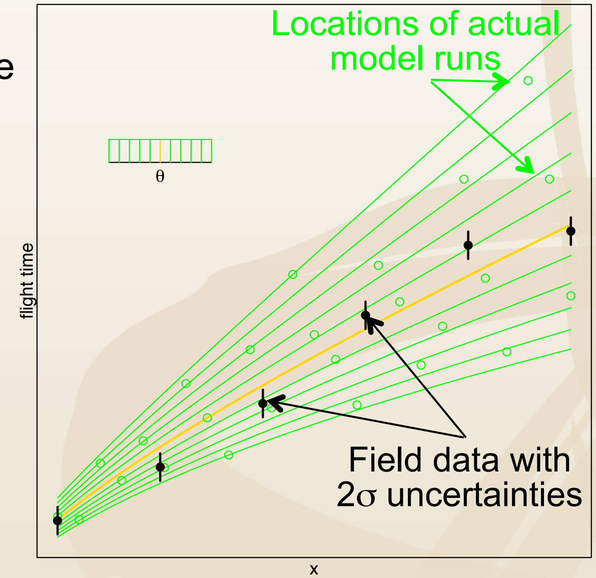
computer model

s = position; τ = time

$$\frac{d^2 s}{d\tau^2} = -1 - \theta \frac{ds}{d\tau}$$

initial conditions

$$s(0) = x, \quad \left.\frac{ds}{d\tau}\right|_{\tau=0} = 0$$

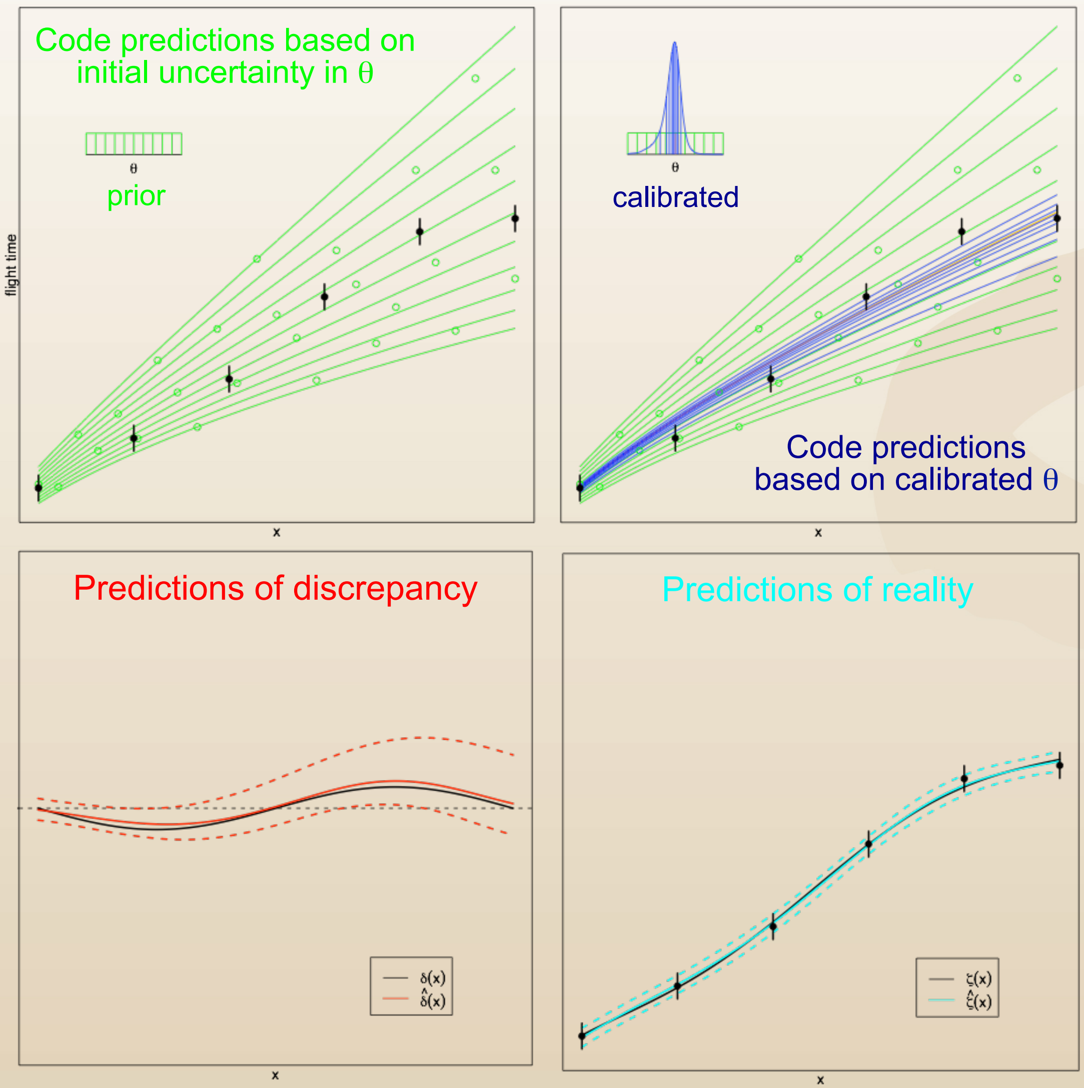


flight time

$\eta(x, \theta)$ is the root of the equation $s(\tau) = 0$.

- Experiment: Drop a solid ball from a specified height
  - Output: Measured flight time (y)
- Computer Model: Implements Newton's Law with drag coefficient
  - Two parameters: x = height (controlled)
    θ = drag coefficient (uncertain physics)
  - Output: Calculated flight time ($\eta(x, \theta)$)

Code Calibration: Statistical Model and Inference



<u>Inputs</u>
x controllable
t uncertain physics
θ best, unknown value of t

$$y(x) = \zeta(x) + \varepsilon(x)$$

field data: $y(x)$; reality: $\zeta(x)$; observation error: $\varepsilon(x)$

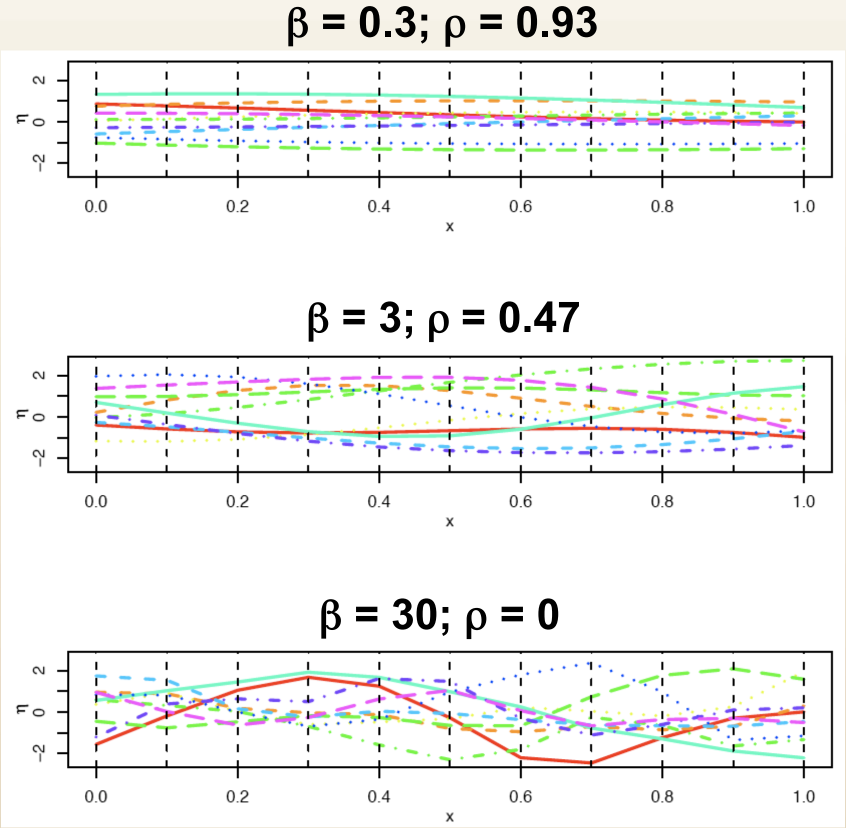
$$\zeta(x) = \eta(x, \theta) + \delta(x)$$

computer model: $\eta(x, \theta)$; discrepancy: $\delta(x)$

Basic steps in calibration analysis:

1. Assume prior probability distribution for physics uncertainties θ.
2. Calibrate parameters θ to field data and simultaneously infer model form error.

Gaussian Process Review and Notation



Semiparametric regression model for emulating code $\eta(x)$

- Joint distribution of surrogate outputs is multivariate Gaussian
- Mean zero, precision λ
- Correlation function:

$$R\left(\eta(\mathbf{x}_1),\eta(\mathbf{x}_2)\,|\,\boldsymbol{\beta}\right)=\exp\left(-\sum_{j=1}^{d}\beta_j\left(x_{1,j}-x_{2,j}\right)^2\right)$$

- Define correlation length:
- $\rho_j = \exp(-\beta_j / 4)$
- Notation: GP(0; λ, ρ)

Correlation lengths ρ_j determine *complexity* of process realizations

Calibration Framework: Scalar Output

- Experiments: $\mathbf{x}_1, \ldots, \mathbf{x}_n$
- Code Runs: $(\mathbf{x}_1^c, \mathbf{t}_1), \ldots, (\mathbf{x}_m^c, \mathbf{t}_m)$
- $\eta(\cdot) \sim GP\left(0; \lambda_\eta, \rho_\eta\right)$ independent of $\delta(\cdot) \sim GP\left(0; \lambda_\delta, \rho_\delta\right)$
- Correlation functions $R_\eta(\mathbf{x}_1 - \mathbf{x}_2, \mathbf{t}_1 - \mathbf{t}_2)$ and $R_\delta(\mathbf{x}_1 - \mathbf{x}_2)$
- $\epsilon \sim \mathcal{N}\left(0, \Sigma_y\right)$

Centered by Average Code Output
Scaled by SD of Code Output

- Output Vector: $\mathcal{D} = (y(\mathbf{x}_1), \ldots, y(\mathbf{x}_n), \eta(\mathbf{x}_1^c, \mathbf{t}_1), \ldots, \eta(\mathbf{x}_m^c, \mathbf{t}_m))$

Likelihood Function: Scalar Output

Likelihood Function

$$L\left(\theta, \lambda_\eta, \rho_\eta, \lambda_\delta, \rho_\delta, \Sigma_y | \mathcal{D}\right) \propto |\Sigma_{\mathcal{D}}|^{-1/2} \exp\left\{-\frac{1}{2}\mathcal{D}^T \Sigma_{\mathcal{D}}^{-1} \mathcal{D}\right\}$$

$$\Sigma_{\mathcal{D}} = \Sigma_\eta + \begin{pmatrix} \Sigma_y + \Sigma_\delta & 0 \\ 0 & 0 \end{pmatrix}$$

$\lambda_\eta \Sigma_\eta$: Correlation matrix between $(\mathbf{x}_1, \theta), \ldots, (\mathbf{x}_n, \theta), (\mathbf{x}_1^c, \mathbf{t}_1), \ldots, (\mathbf{x}_m^c, \mathbf{t}_m)$ — R_η

$\lambda_\delta \Sigma_\delta$: Correlation matrix between $\mathbf{x}_1, \ldots, \mathbf{x}_n$ — R_δ

$\lambda_y \Sigma_y = \mathbf{I}_n$ in many applications

λ_y fixed or random

Prior Distributions and Posterior Sampling

- Prior Distributions

→ Correlation parameters (η and δ)

$$\pi(\boldsymbol{\rho}) \propto \prod_{j=1}^{n_\rho} (1-\rho_j)^{(b_\rho - 1)}, 0 < \rho_j \le 1$$

– Control degree of prior smoothness (variable importance)

→ Precision parameters (η, δ, and ε)

$$\pi(\lambda) \propto \lambda^{(a_\lambda - 1)} \exp(-b_\lambda \lambda), \ \lambda > 0$$

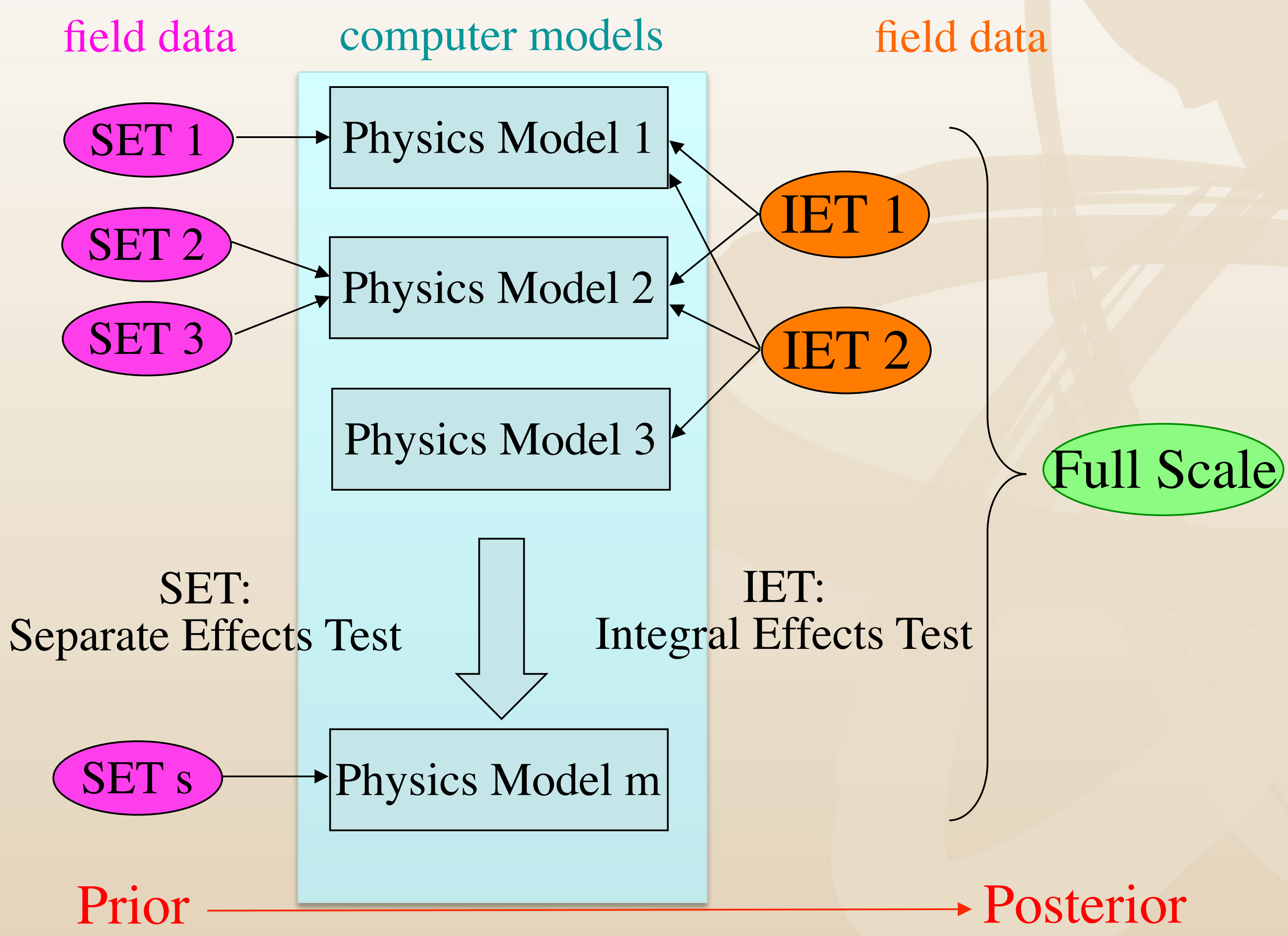
– Set $a_\eta = b_\eta$ (prior mean 1; larger b_η smaller prior variance)

– Set $b_\delta / a_\delta \approx 0$, i.e. noninformative with large prior mean

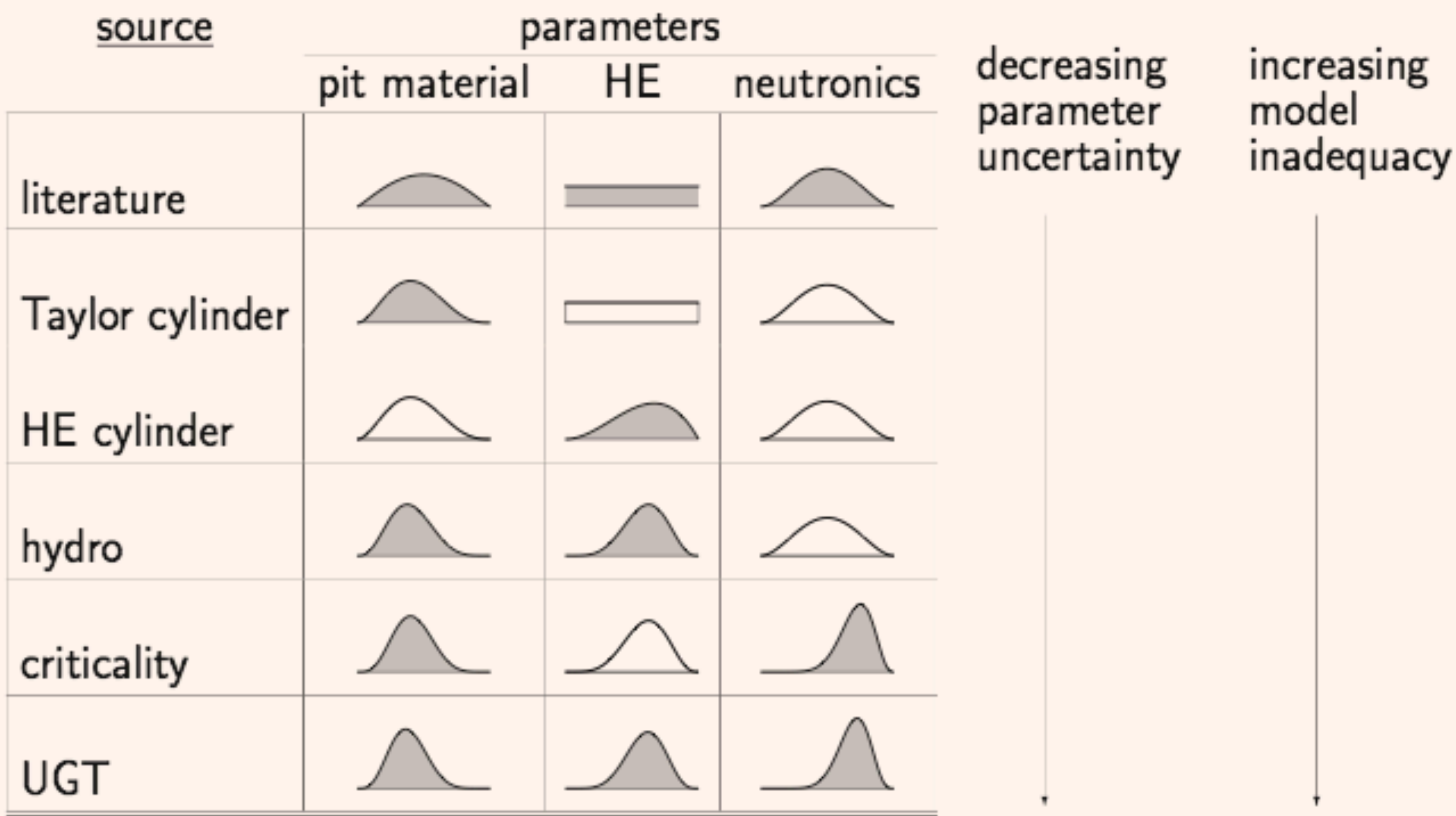
– Settings for a_ϵ and b_ϵ depend on assumptions for observation error

- Posterior Sampling
  - Metropolis within Gibbs MCMC
  - Burn-in + logistic regression to estimate step sizes

Simultaneous Code Calibration for Multi-Physics Applications

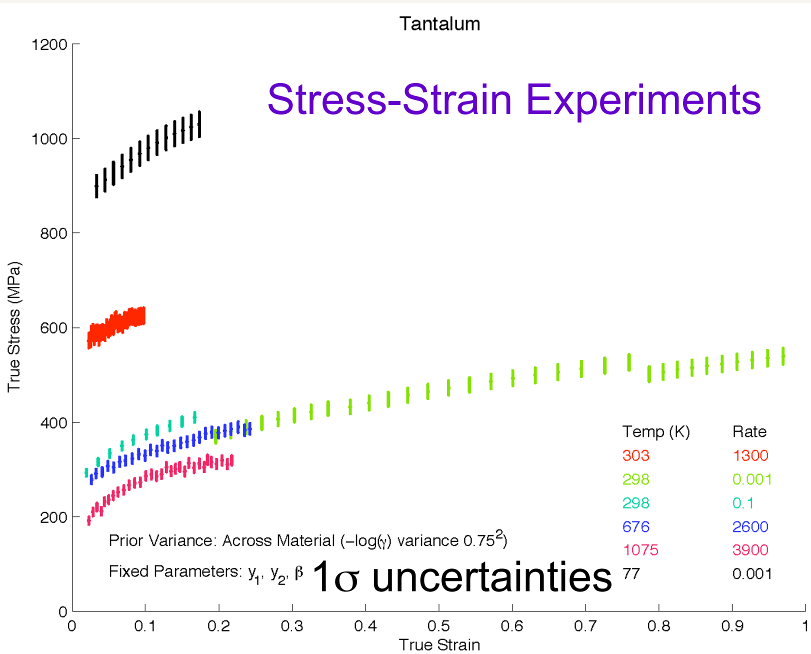


Simultaneous Code Calibration for Weapon Performance Applications

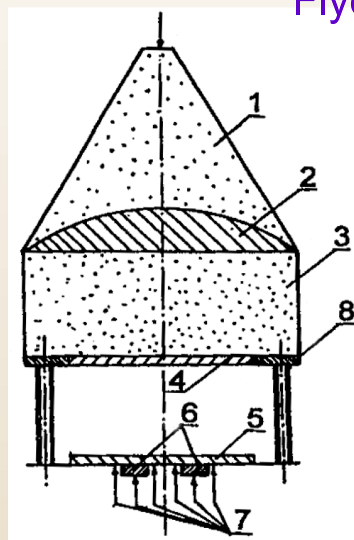


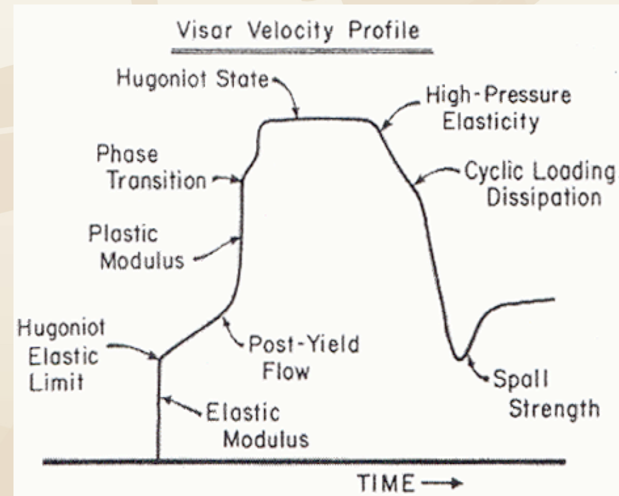
- conditioning on more experiments ⇒ less parametric uncertainty
- prediction uncertainty becomes more affected by model inadequacies

Code Calibration: Multiple Datasets and Functional Output



Flyer Plate Experiments





- SET: Stress-strain experiments are conducted to infer material strength
    - Physics model: PTW (Preston-Tonks-Wallace)
- IET: Flyer plate experiments are conducted to infer material equation of state (EOS), strength and damage simultaneously
    - Physics models: tabular EOS, PTW, tension limit

PTW Plastic Deformation Model

activation energy — strain rate

$$\hat{\tau}_y = y_0 - (y_0 - y_\infty)\,\mathrm{erf}\left[\kappa \hat{T} \ln\left(\gamma \dot{\xi} / \dot{\psi}\right)\right]$$

yield stress

T/ $T_m(\rho)$

$$\hat{\tau}_s = s_0 - (s_0 - s_\infty)\,\mathrm{erf}\left[\kappa \hat{T} \ln\left(\gamma \dot{\xi} / \dot{\psi}\right)\right]$$

saturation stress

T = temperature

$T_m(\rho)$ = melting temp.

atomic vibration time

strain

$$\hat{\tau} = \hat{\tau}_s + \frac{1}{p}\left(s_0 - \hat{\tau}_y\right)\ln\left[1 - \left[1 - \exp\left(-p\frac{\hat{\tau}_s - \hat{\tau}_y}{s_0 - \hat{\tau}_y}\right)\right]\exp\left\{-\frac{p\theta_0 \psi}{\left(s_0 - \hat{\tau}_y\right)\left[\exp\left(p\frac{\hat{\tau}_s - \hat{\tau}_y}{s_0 - \hat{\tau}_y}\right) - 1\right]}\right\}\right]$$

$$\theta = (\theta_0, p, \kappa, \gamma, y_0, y_\infty, s_0, s_\infty)$$

calibration parameters

Calibration of PTW Model

stress

strain

dataset-specific parameters (temp., strain rate)

dataset-specific estimated uncertainty

$$y_{ij} = ptw(s_{ij}, \mathbf{x}_i, \boldsymbol{\theta}) + \varepsilon_{ij}; \boldsymbol{\varepsilon}_i \sim N\left(\mathbf{0}, \frac{\sigma_i^2}{\lambda}\mathbf{I}\right);$$

uncertain model parameters

$$\boldsymbol{\theta} \sim N\left(\mathbf{b}_0, \mathbf{V}_0^{-1}\right); \lambda \sim Gamma\ (a, b);$$

adjustment to uncertainty

prior mean

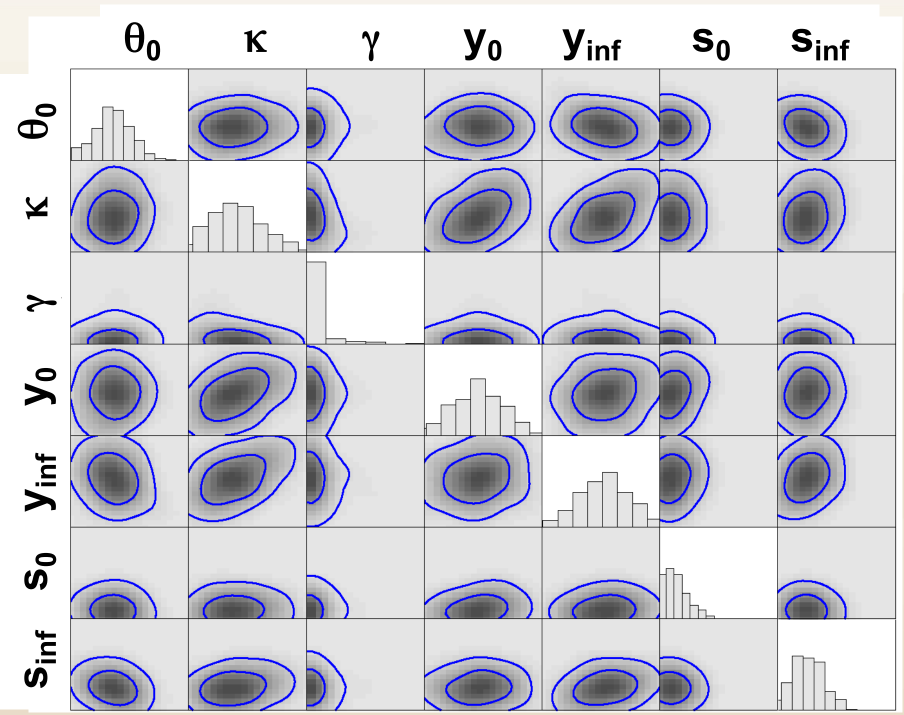
prior covariance (depends on analysis)

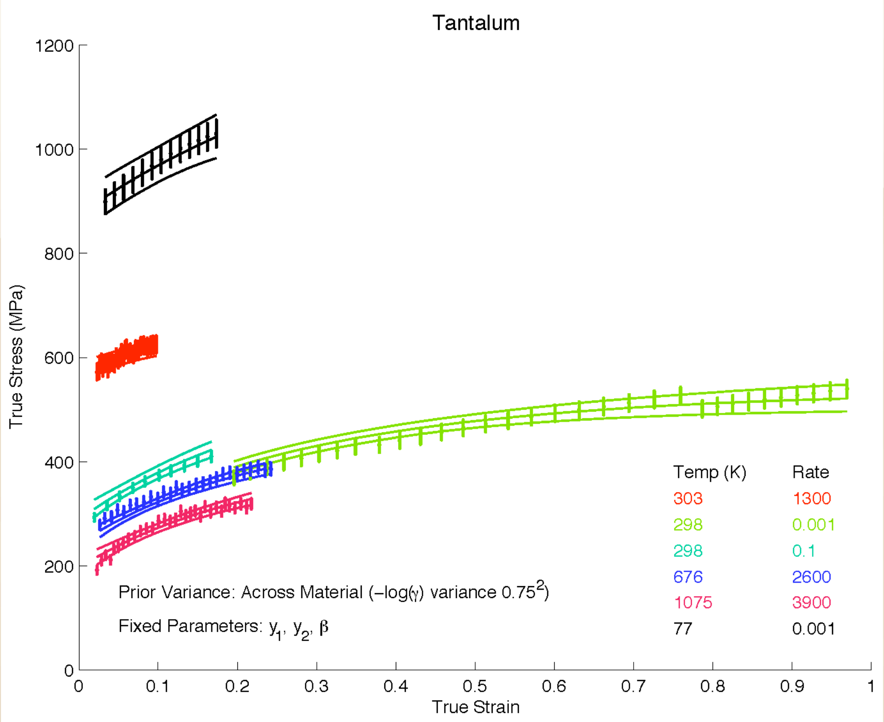
PTW

$\theta = (\theta_0, p, \kappa, -\ln(\gamma), y_0, y_\infty, s_0, s_\infty)$

Uncertain parameters θ “common” to all datasets

Phase I: Calibration to Small-Scale Data





Prior constraint on PTW parameters for calibration of all parameters to integral data

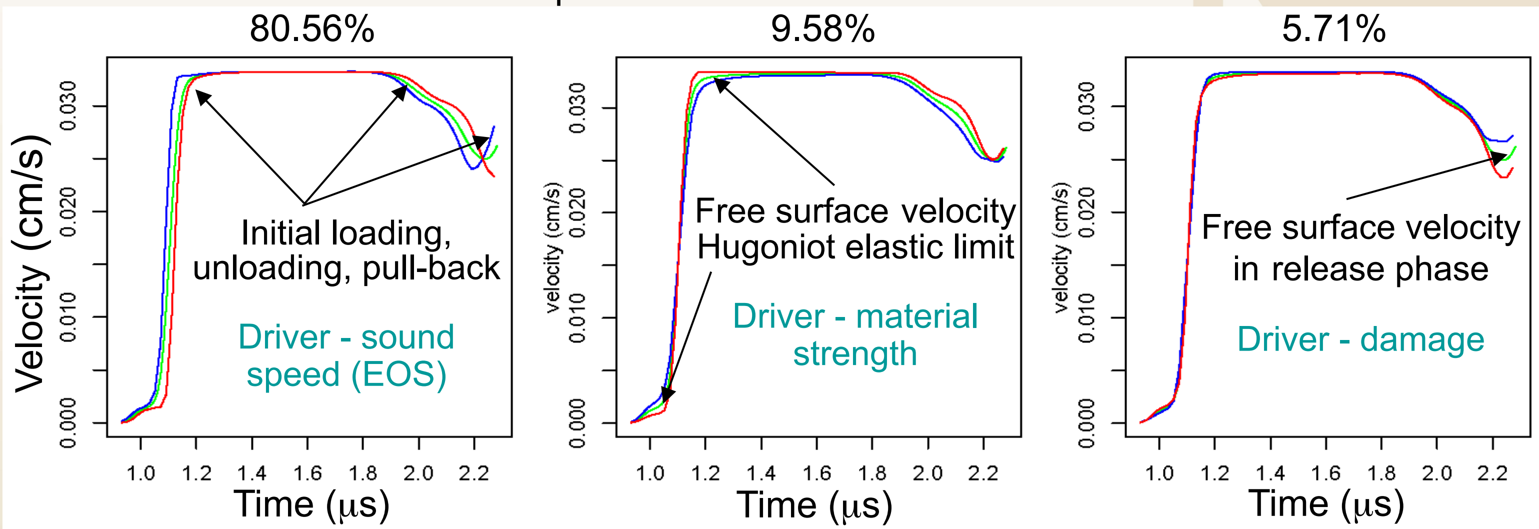
Physics Models and Parameters

Input	Description	Domain	
		Min	Max
ε	Perturbation of EOS table from nominal	-5%	5%
θ_0	Initial strain hardening rate	2.78×10^{-5}	0.0336
κ	Material constant in thermal activation energy term – relates to the temperature dependence	0.438	1.11
γ	Material constant in thermal activation energy term – relates to the strain rate dependence	6.96×10^{-8}	6.76×10^{-4}
y_0	Maximum yield stress (at 0 K)	0.00686	0.0126
y_∞	Minimum yield stress (~ melting)	7.17×10^{-4}	0.00192
s_0	Maximum saturation stress (at 0 K)	0.0126	0.0564
s_∞	Minimum saturation stress (~ melting)	0.00192	0.00616
P_{min}	Spall strength	-0.055	-0.045
v_s	Flyer plate impact velocity	329.5	338.5

Calibrate all parameters to integral (flyer plate) data
128 flyer plate runs defined by an OA-based LH design

Modeling Functional Computer Model Output

Total dispersion in the mean-centered simulations



- $n_\eta \times m$ matrix of simulator output ("time" by "space")

$\rightarrow$ each row mean centered; entire matrix scaled so output has variance 1

- Statistical model:

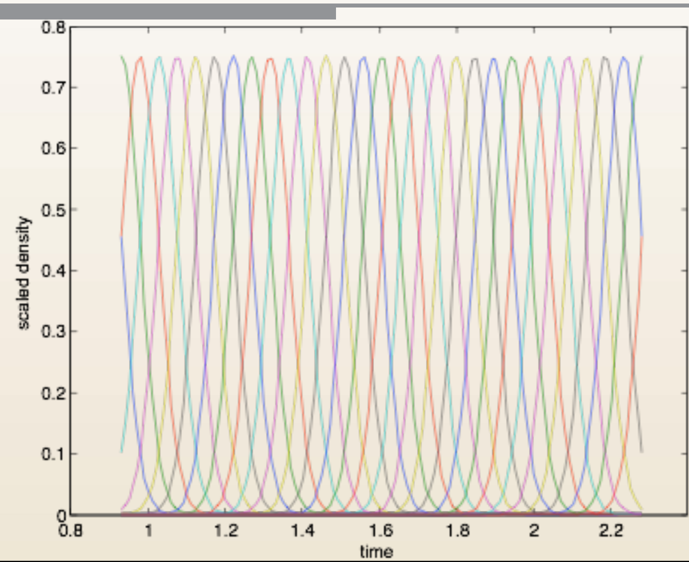
$$\boldsymbol{\eta}(\boldsymbol{x}, \boldsymbol{t}) = \sum_{i=1}^{p_\eta} \boldsymbol{k}_i w_i(\boldsymbol{x}, \boldsymbol{t}) + \boldsymbol{\epsilon}$$

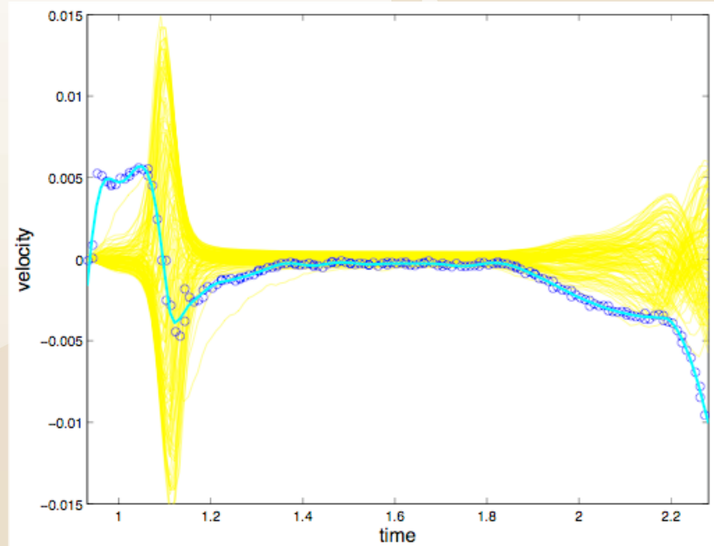
$\rightarrow$ $\boldsymbol{k}_1, \ldots, \boldsymbol{k}_{p_\eta}$ are $n_\eta \times 1$ orthogonal basis vectors (e.g. principal components)

$\rightarrow$ $w_i(\boldsymbol{x}, \boldsymbol{t})$: basis coefficients; modeled as GP($\boldsymbol{\rho}_{wi}$,λ_{wi}); independent

$\rightarrow$ $\boldsymbol{\epsilon}$: model error; modeled as GP($\boldsymbol{0}$,λ_η); independent of basis coefficients

Modeling Functional Experimental Data





- $\boldsymbol{y}(\boldsymbol{x}_i)$ is $n_{y_i} \times 1$ vector of centered/scaled experimental data, $i = 1, \ldots, n$
- Statistical model:

$$\boldsymbol{y}(\boldsymbol{x}_i) = \boldsymbol{K}_i \boldsymbol{w}(\boldsymbol{x}_i, \boldsymbol{\theta}) + \boldsymbol{D}_i \boldsymbol{v}(\boldsymbol{x}_i) + \boldsymbol{\epsilon}_i$$

$\rightarrow$ $\boldsymbol{K}_i$ is $n_{y_i} \times p_\eta$ matrix of simulator basis vectors interpolated onto data grid

$\rightarrow$ $\boldsymbol{w}(\boldsymbol{x}_i, \boldsymbol{\theta})$: simulator basis coefficients evaluated at best, unknown $\boldsymbol{\theta}$

$\rightarrow$ $\boldsymbol{D}_i$ is $n_{y_i} \times p_\delta$ matrix of discrepancy basis vectors

$\rightarrow$ $\boldsymbol{v}(\boldsymbol{x}_i)$: discrepancy basis coefficients; modeled as GP($\boldsymbol{\rho}_v$,λ_v); independent

$\rightarrow$ $\boldsymbol{\epsilon}_i$: model error; modeled as GP($\boldsymbol{0}$,λ_y); independent of basis coefficients

Joint Prior Distribution of Coefficients

$$\mathbf{v} = \mathrm{vec}\left([\mathbf{v}(\mathbf{x}_1); \cdots ; \mathbf{v}(\mathbf{x}_n)]^T\right)$$

Define $$\mathbf{u}(\theta) = \mathrm{vec}\left([\mathbf{w}(\mathbf{x}_1, \theta); \cdots ; \mathbf{w}(\mathbf{x}_n, \theta)]^T\right)$$

$$\mathbf{w} = \mathrm{vec}\left([\mathbf{w}(\mathbf{x}_1, \mathbf{t}_1); \cdots ; \mathbf{w}(\mathbf{x}_m, \mathbf{t}_m)]^T\right)$$

For $\mathbf{z} = \left(\mathbf{v}^T, \mathbf{u}^T(\theta), \mathbf{w}^T\right)^T$,

$$\mathbf{z} \sim \mathcal{N}\left(\mathbf{0}\,,\ \Sigma_z = \begin{pmatrix} \Sigma_v & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \Sigma_u & \Sigma_{u,w} \\ \mathbf{0} & \Sigma_{u,w}^T & \Sigma_w \end{pmatrix}\right)$$

Representation of Data and Error Model

Define
$$\mathbf{y} = \left(\mathbf{y}^T(\mathbf{x}_1), \ldots, \mathbf{y}^T(\mathbf{x}_n)\right)^T$$
$$\eta = \left(\eta^T(\mathbf{x}_1^c, \mathbf{t}_1), \ldots, \eta^T(\mathbf{x}_m^c, \mathbf{t}_m)\right)^T$$

Then
$$\begin{pmatrix} \mathbf{y} \\ \eta \end{pmatrix} = \begin{pmatrix} \mathbf{B} & \mathbf{0} \\ \mathbf{0} & \mathbf{K} \end{pmatrix} \mathbf{z} + \begin{pmatrix} \epsilon \\ \varepsilon \end{pmatrix}$$

where
$$\begin{pmatrix} \epsilon \\ \varepsilon \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} (\lambda_y \mathbf{W}_y)^{-1} & \mathbf{0} \\ \mathbf{0} & \lambda_\eta^{-1}\mathbf{I} \end{pmatrix}\right)$$

$$\mathbf{W}_y = \mathrm{diag}\left(\mathbf{W}_1, \ldots, \mathbf{W}_n\right)$$

Likelihood Function: Functional Output

Likelihood Function

$$L(\theta, \lambda_\eta, \lambda_w, \rho_w, \lambda_y, \lambda_v, \rho_v | \mathbf{y}, \eta) \propto$$

$$|\Sigma_{\widehat{\mathbf{z}}}|^{-1/2} \exp\left\{ -\frac{1}{2} \widehat{\mathbf{z}}^T \Sigma_{\widehat{\mathbf{z}}}^{-1} \widehat{\mathbf{z}} \right\}$$

$$\widehat{\mathbf{z}} = \text{vec}\left(\left[\left(\mathbf{B}^T \mathbf{W}_y \mathbf{B} \right)^{-1} \mathbf{B}^T \mathbf{W}_y \mathbf{y}; \left(\mathbf{K}^T \mathbf{K} \right)^{-1} \mathbf{K}^T \eta \right] \right)$$

$$\Sigma_{\widehat{\mathbf{z}}} = \Sigma_z + \begin{pmatrix} \left(\lambda_y \mathbf{B}^T \mathbf{W}_y \mathbf{B} \right)^{-1} & \begin{matrix} \mathbf{0} \\ \mathbf{0} \end{matrix} \\ \begin{matrix} \mathbf{0} & \mathbf{0} \end{matrix} & \left(\lambda_\eta \mathbf{K}^T \mathbf{K} \right)^{-1} \end{pmatrix}$$

Prior Distributions: Functional Output

Parameter Prior Distributions

Extension of scalar case, with modified Gamma parameters for λ_η and λ_y

$$a'_\eta = a_\eta + \frac{m(n_\eta - p_\eta)}{2}$$

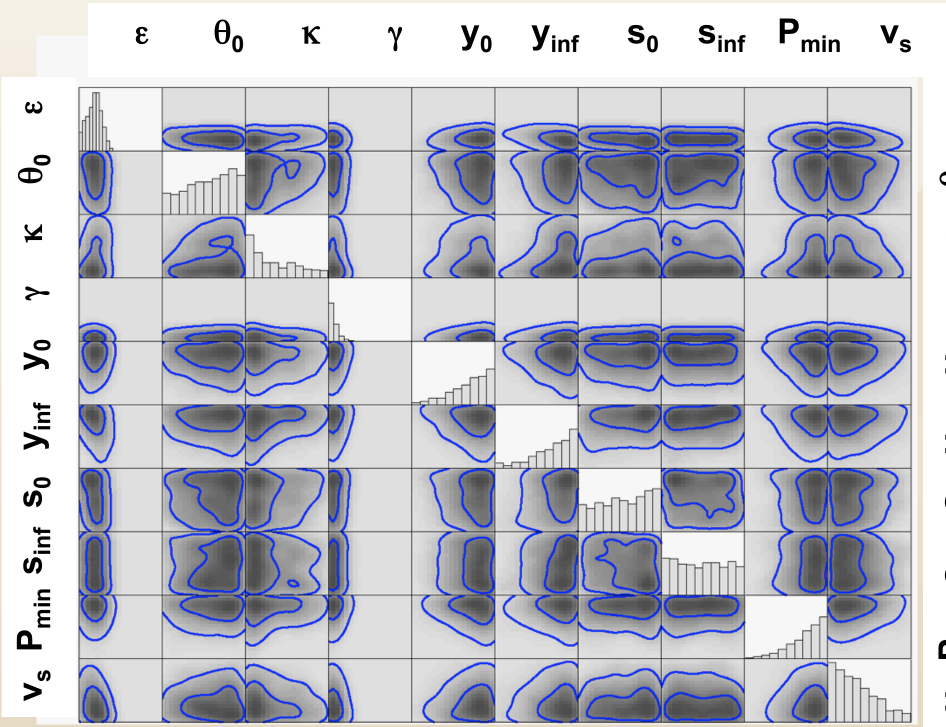
$$a'_y = a_y + \frac{n_y - \text{rank}(\mathbf{B})}{2}$$

$$b'_\eta = b_\eta + \frac{1}{2}\eta^T \left(\mathbf{I} - \mathbf{K}\left(\mathbf{K}^T\mathbf{K}\right)^{-1}\mathbf{K}^T\right)\eta$$

$$b'_y = b_y + \frac{1}{2}\mathbf{y}^T \left(\mathbf{W}_y - \mathbf{W}_y\mathbf{B}\left(\mathbf{B}^T\mathbf{W}_y\mathbf{B}\right)^{-1}\mathbf{B}^T\mathbf{W}_y\right)\mathbf{y}$$

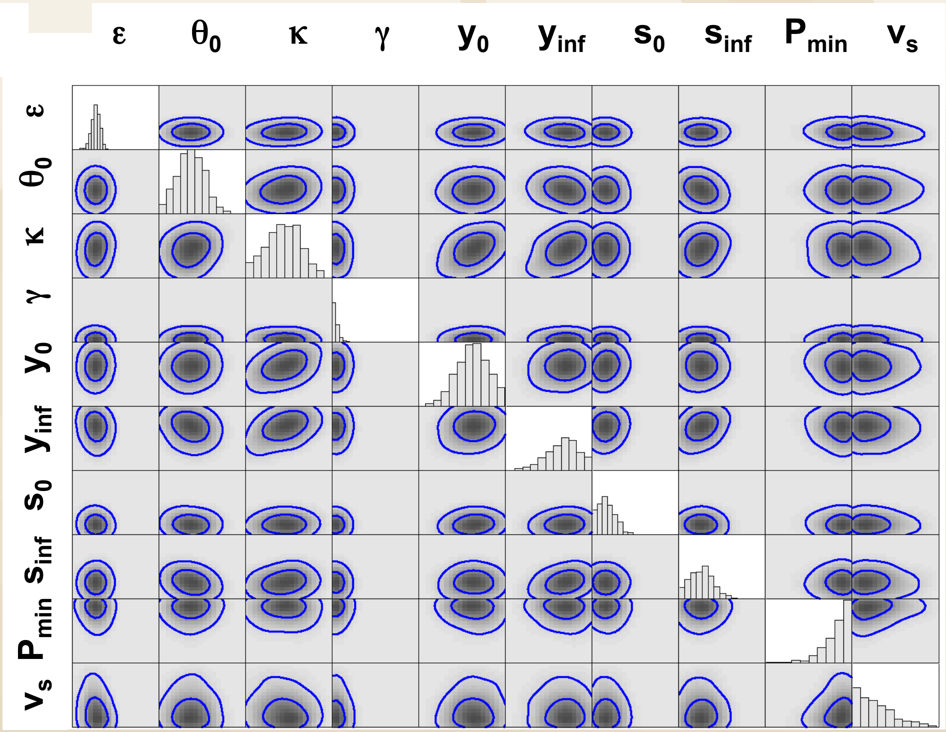
Phase II: Calibration to Integral Data

Uniform prior for all parameters



Small scale posterior is prior for material strength parameters

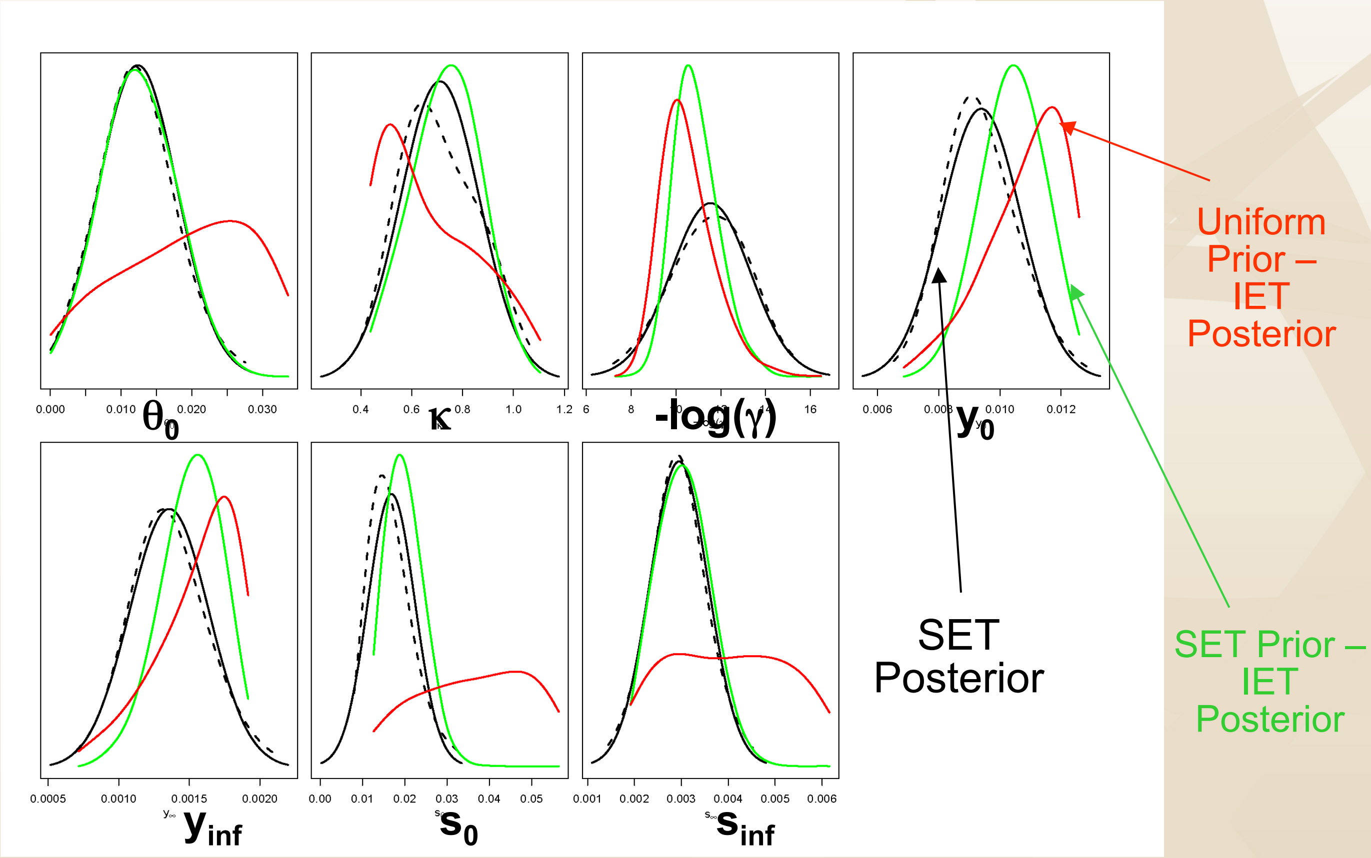
Uniform prior for other parameters



Small-scale data often helps reduce compensating errors

Marginal Effects of Parameter Prior Assumptions

- Sensitivity to priors
- Same prior and posterior indicates no value for IET

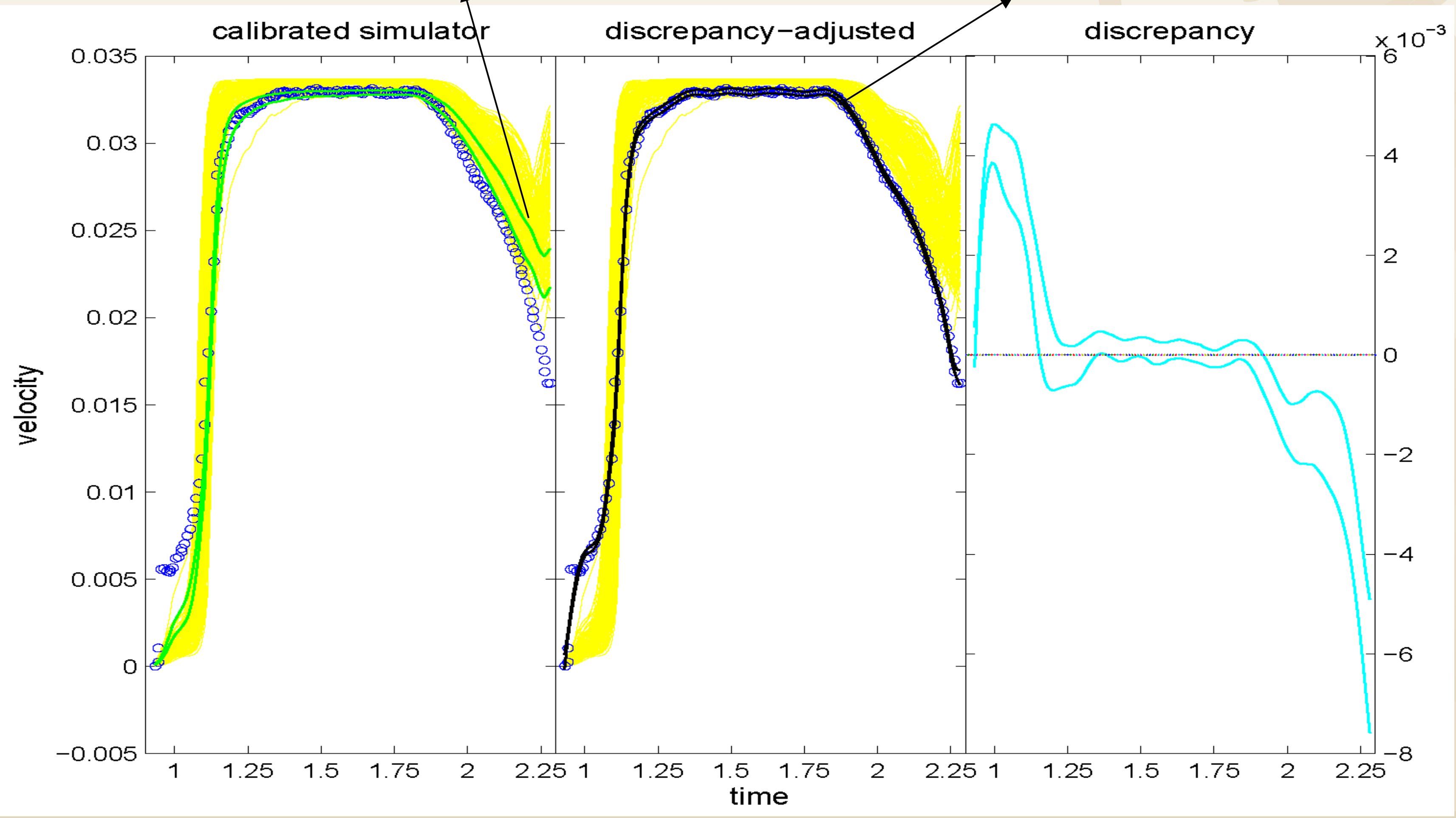


Flyer plate data refines knowledge about activation energy and yield stress parameters

Calibrated Prediction

5%-95% bounds on calibrated code predictions (no discrepancy)

5%-95% bounds on calibrated code predictions adjusted for discrepancy

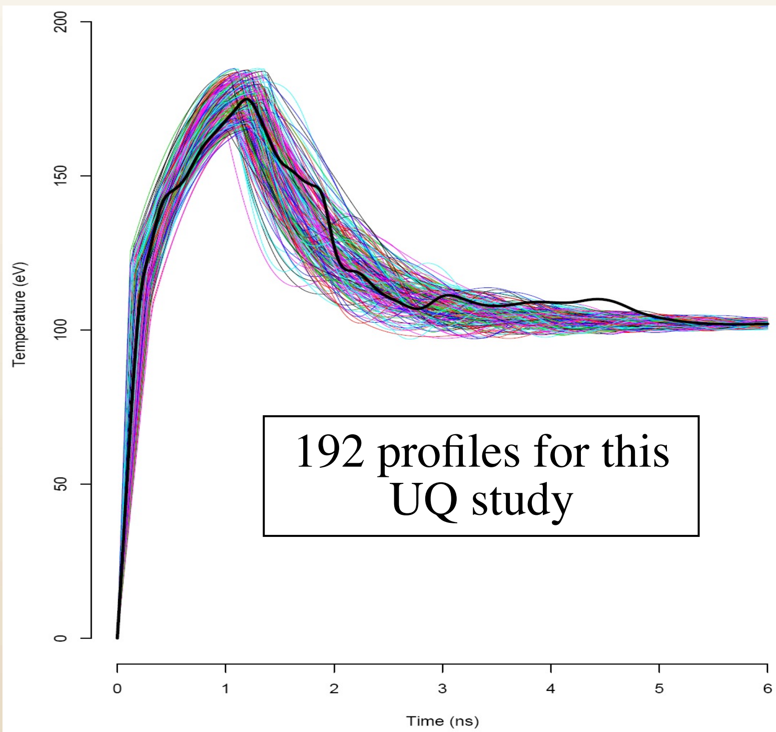


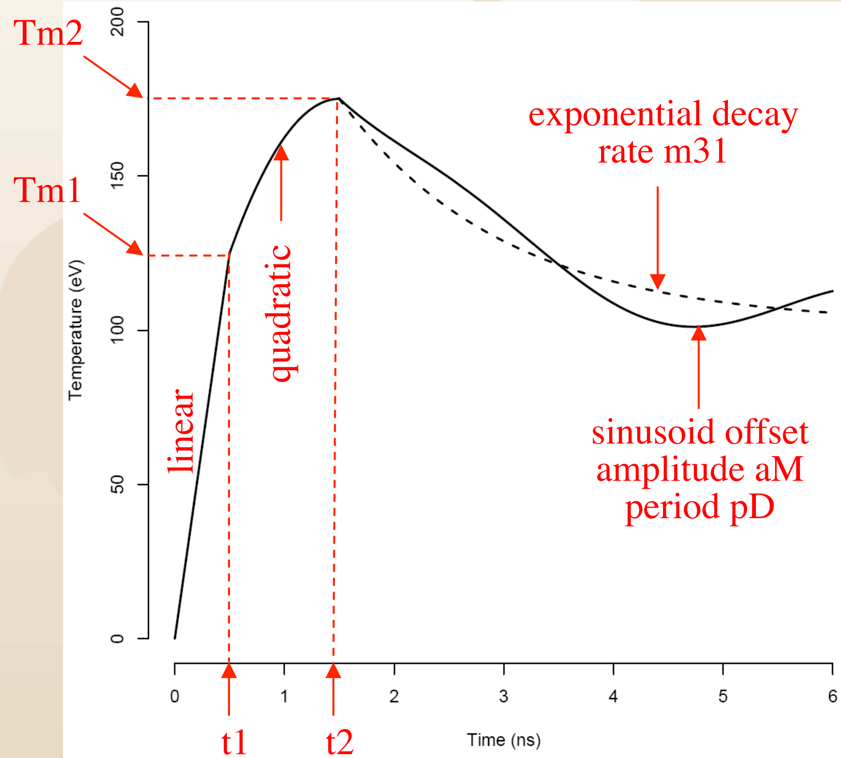
Implementation Considerations

- Check sensitivity analysis on prior ranges
  - Parameter screening may be important
- Observational error model
- Discrepancy model (if included)
  - Multiple scalars different than functional
- Prior distribution for calibration parameters and statistical model parameters
- Check emulator performance (if code surrogate is required)
  - Cross-validation, out-of-sample validation
- Check posteriors and predictions carefully

Calibration: A Cautionary Tale

Simulation Input

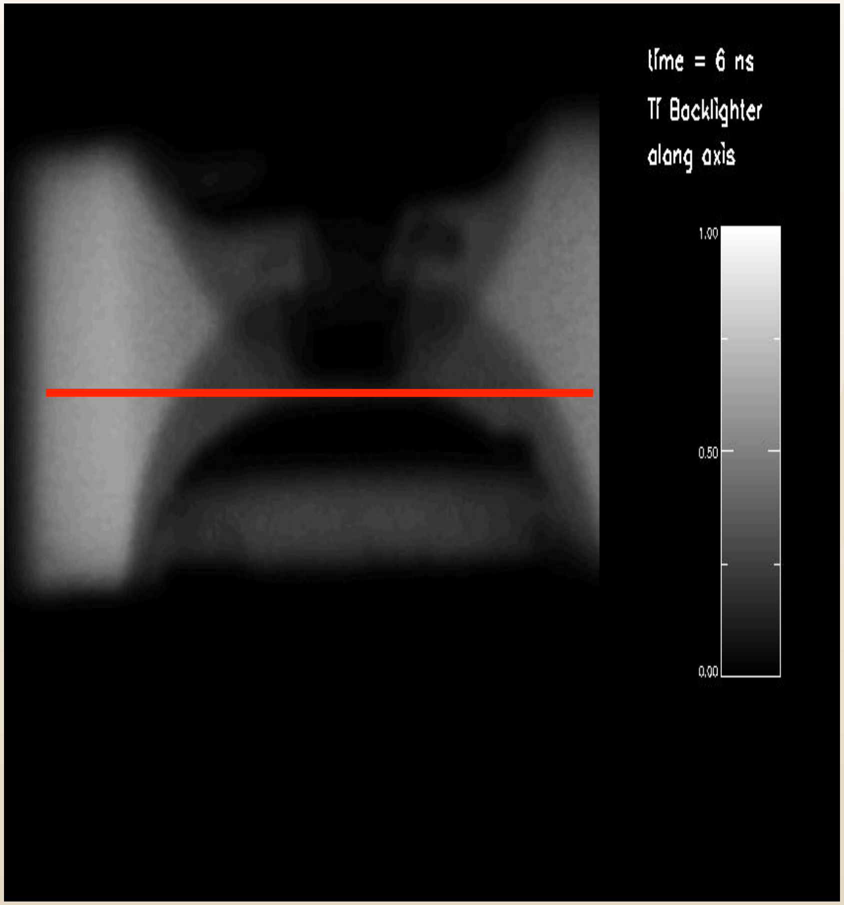


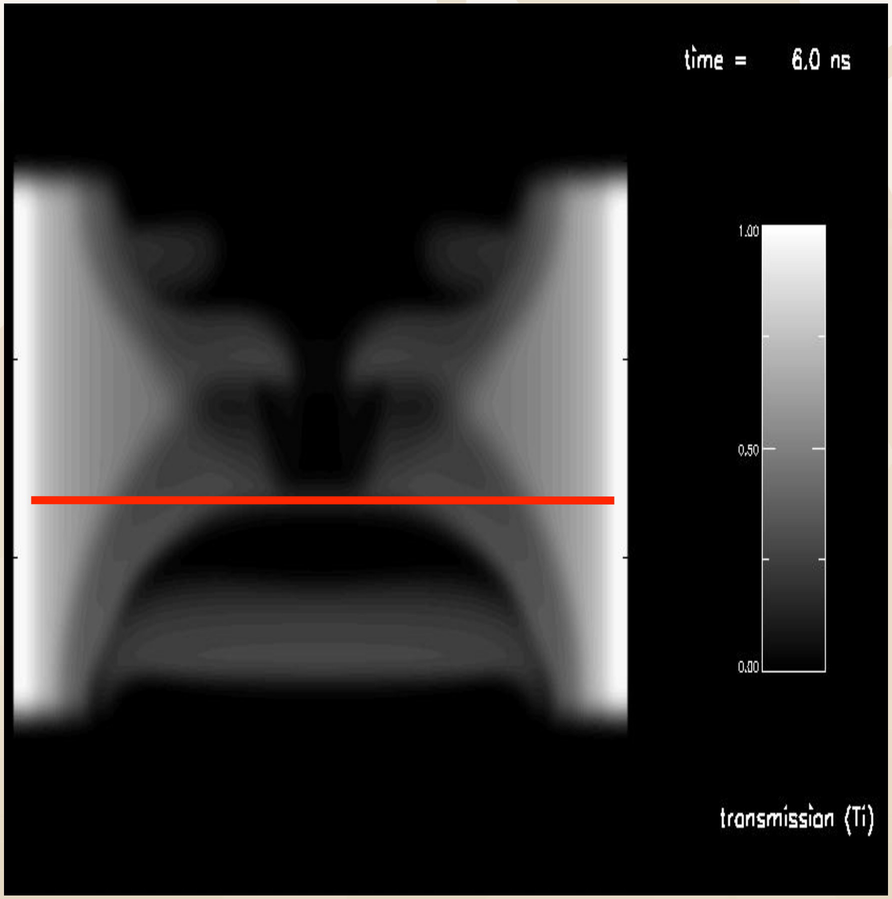


Drive conditions specified by temperature profiles

- Continuous family of functions indexed by 7 parameters

Data and Simulations are Radiographic Images

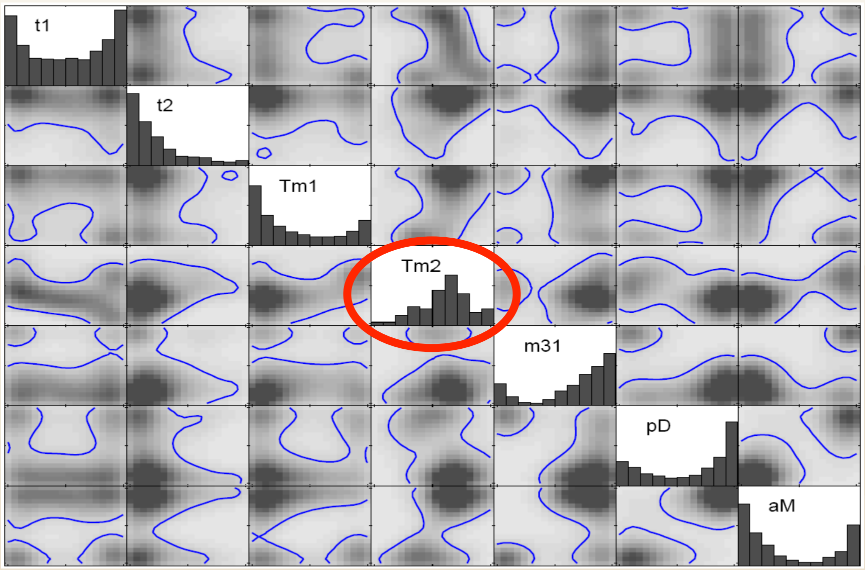




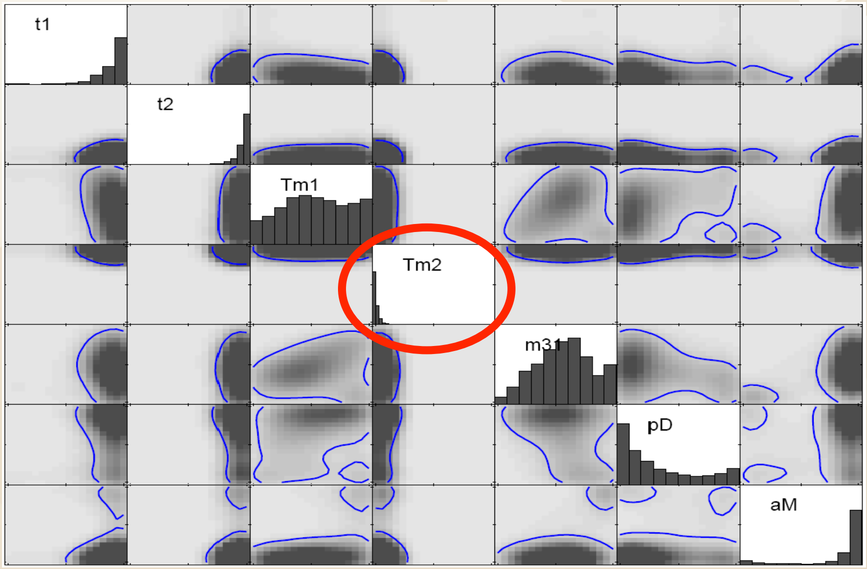
Quantity of Interest is transmission along a selected lineout as a function of distance from centerline

Statistical Model Permits Tradeoff Between Model Fit and Model Form Error

No Discrepancy



Discrepancy



Substantial difference in the calibrated Tm2 marginal distributions

Would like "data - best code", $\chi(\theta) = y - \eta(\theta)$, small (absolute sense)

To first order, accomplished for small values of quadratic error:

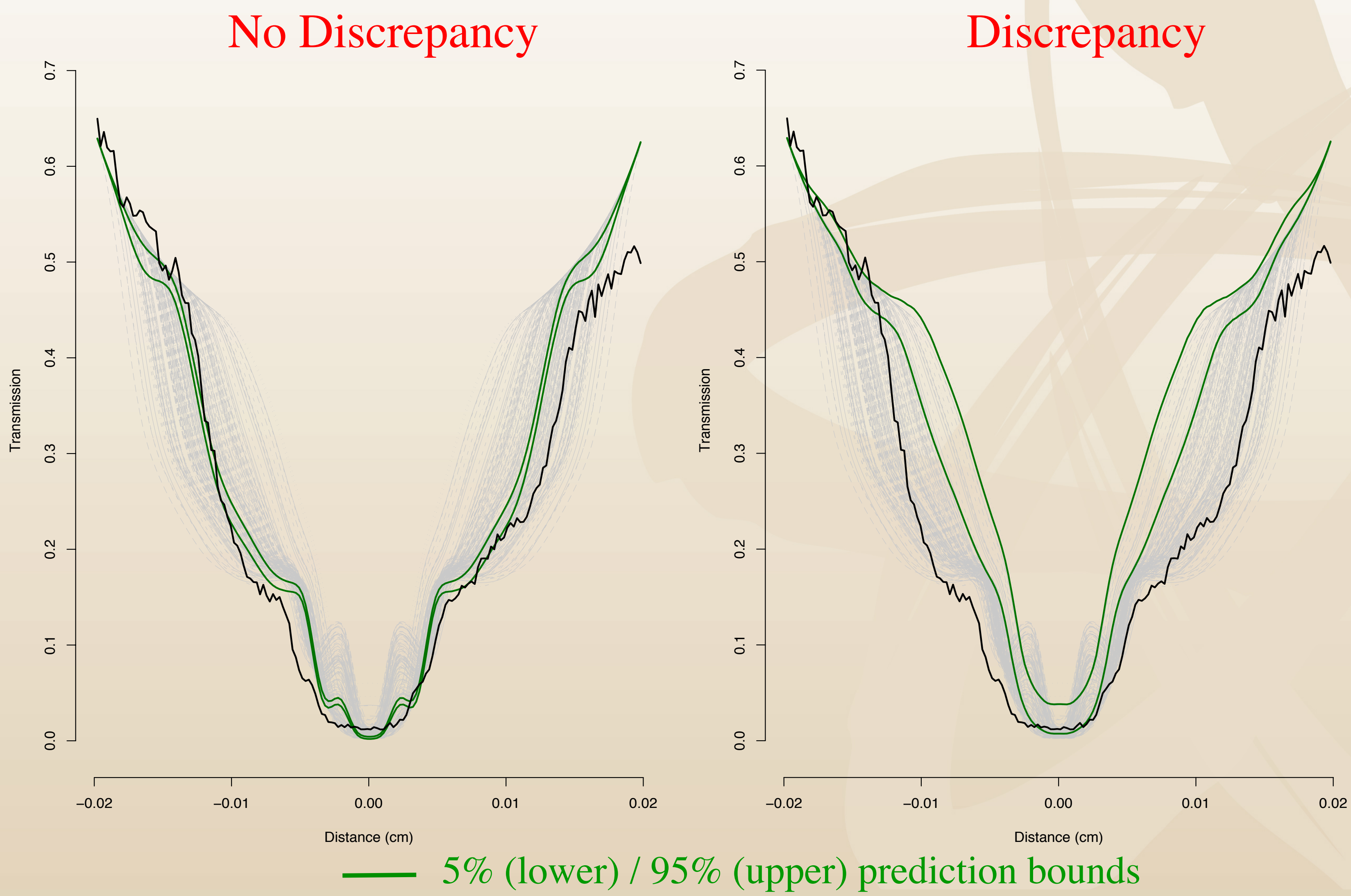
$$Q(\theta, \delta) = (\chi(\theta) - \delta)^T W_y (\chi(\theta) - \delta) + \delta^T W_\delta \delta$$

discrepancy

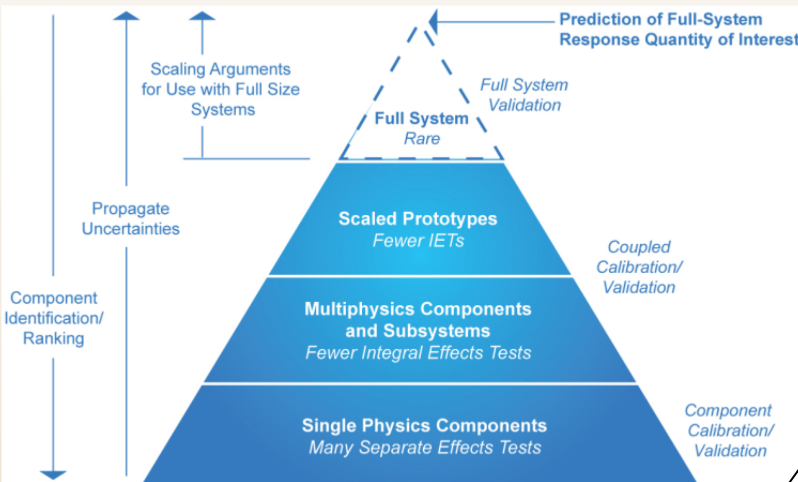
field data precision

discrepancy precision

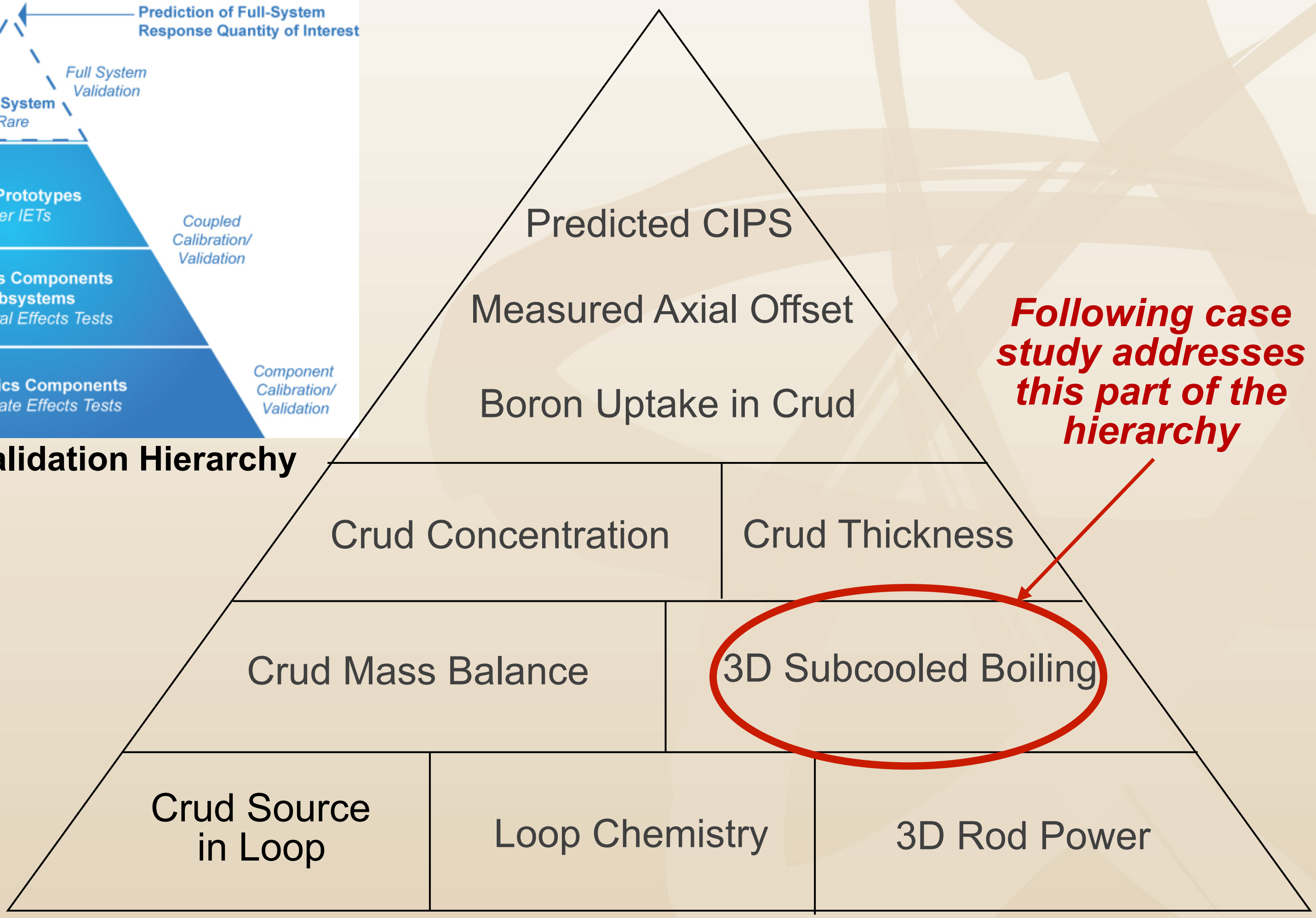
Undesirable Tradeoff Between Model Fit and Model Form Error



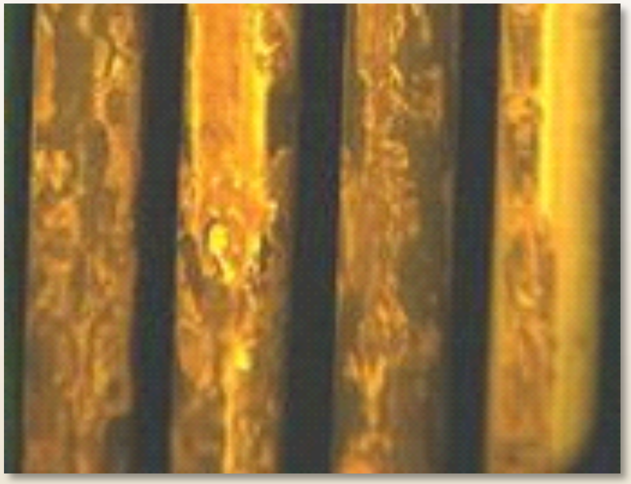
Calibration Supports Validation of Computational Models



The "Generic" Validation Hierarchy



Validation is a Key Component of Predictive Capability Assessment

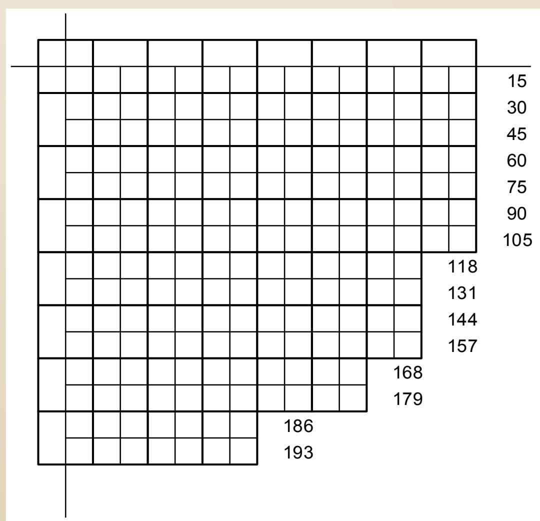


Crud deposits

Assess predicted mass evaporation (boiling) rate and compare to Plant B Crud index data

- **Calibration** to assemblies F71, F22, and F88
- **Validation** to assembly F09

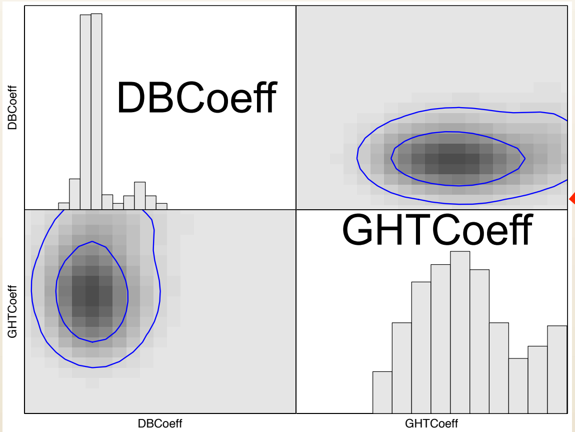
Westinghouse VIPRE-W Thermal-Hydraulics Simulator quarter-core geometry and axial channel layout



Parameter	Range
Lead coefficient of Dittus-Boelter Correlation (DBCoeff)	0.019 – 0.033
Lead Coefficient of Grid Heat Transfer Model (GHTCoeff)	2 - 6
Axial Friction Correlation Coefficient (AFCCoeff)	0.1 - 0.25
Lateral Resistance Correlation Coefficient (LRCCoeff)	1.5 - 4
Exponent of Partial Boiling Model (ExpPBM)	1 - 4

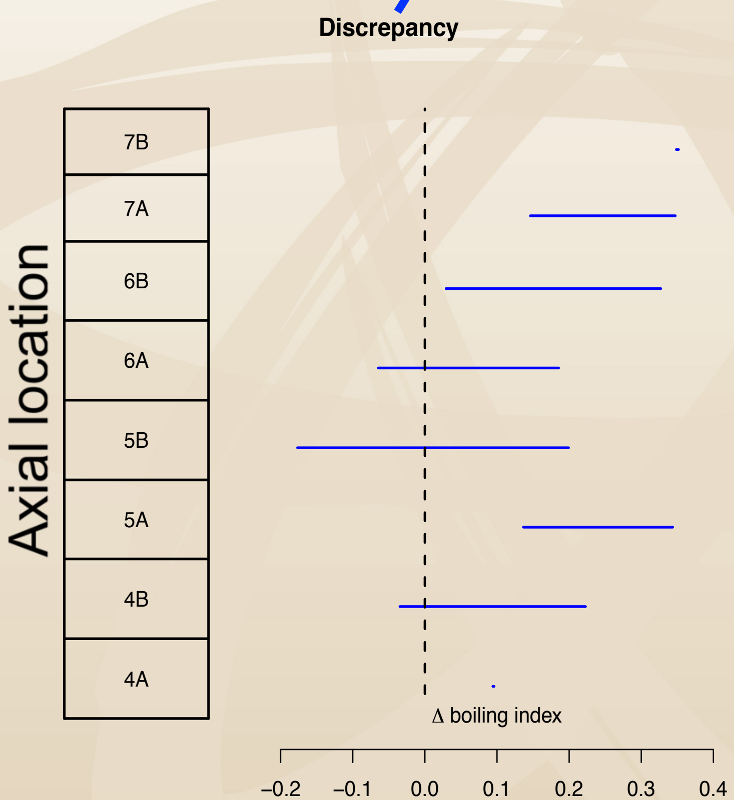
Calibration Methodology Implemented Treating Boiling/Crud Index as Functional

Crud Index = VIPRE-W Boiling Index (calibration parameters) + Discrepancy + Error

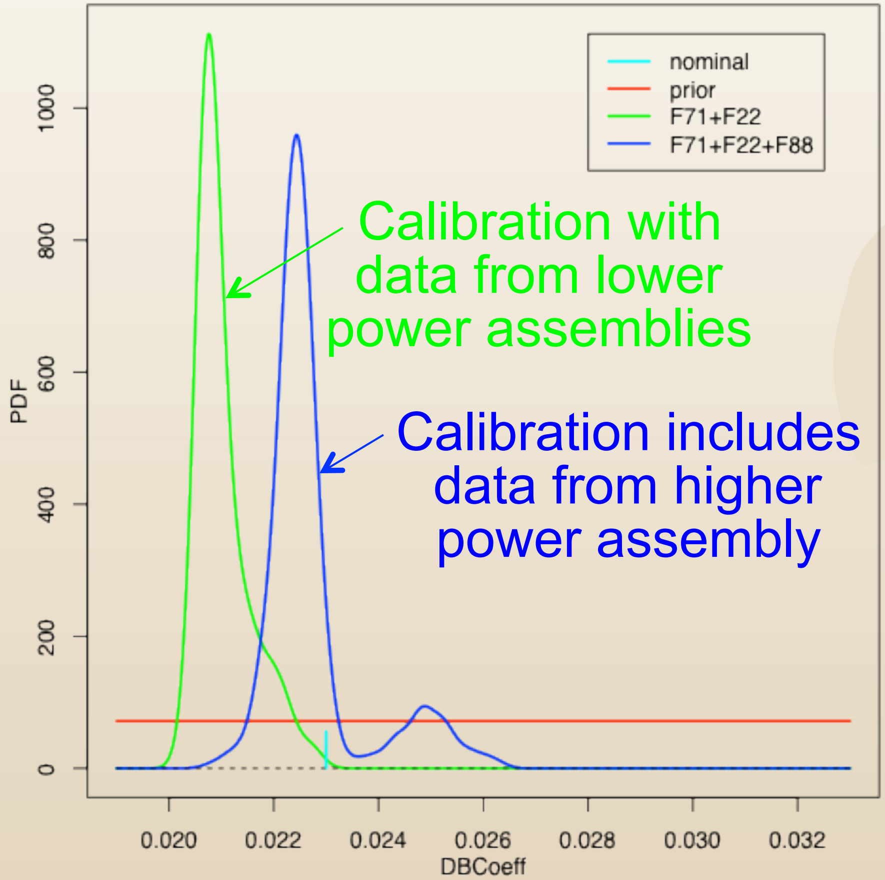


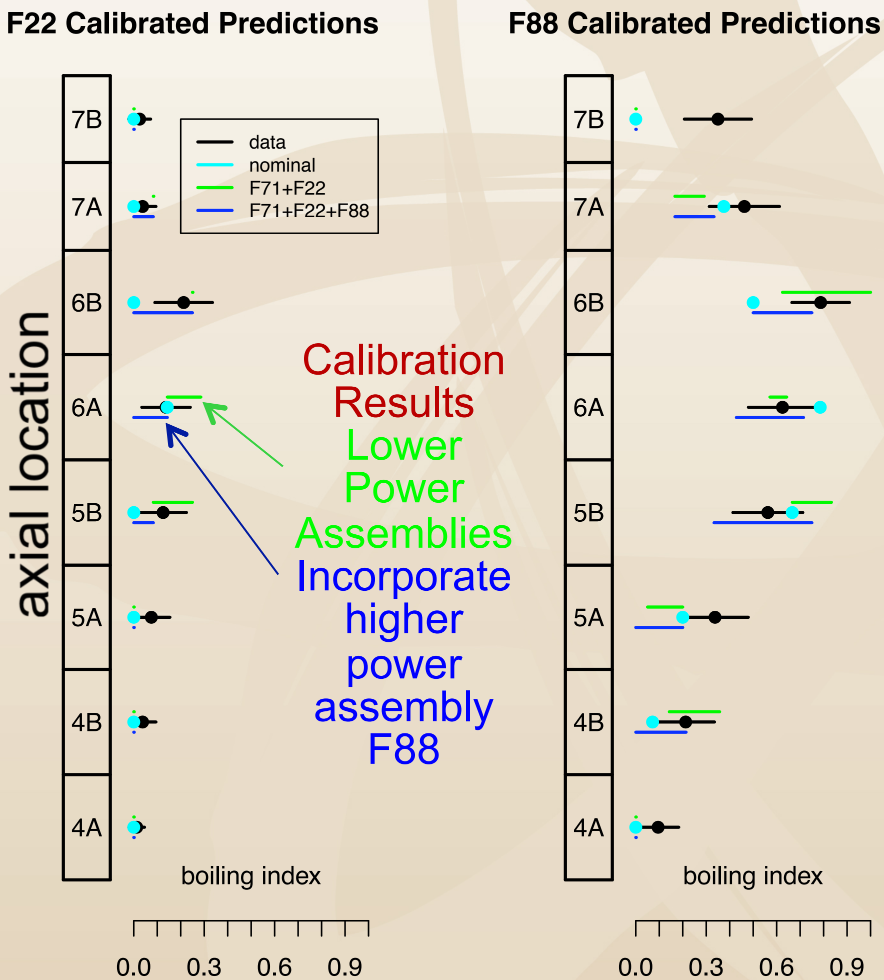
Sensitivity Analysis

	F71		F22		F88		F09	
Parameter	ME (%)	TE (%)	ME (%)	TE (%)	ME (%)	TE (%)	ME (%)	TE (%)
DBCoeff	94.8	98.1	93.6	98.1	93.1	96.8	96.6	98.4
GHTCoeff	1.7	4.8	0.4	3.3	2.5	5.6	1.2	2.8
AFCCoeff	0	0.4	0.2	3.2	0.4	1.6	0.3	0.7
LRCCoeff	0	0.4	0.1	2.8	0	0.6	0	0.2
ExpPBM	0	0.3	0.1	2.8	0	0.4	0	0.2

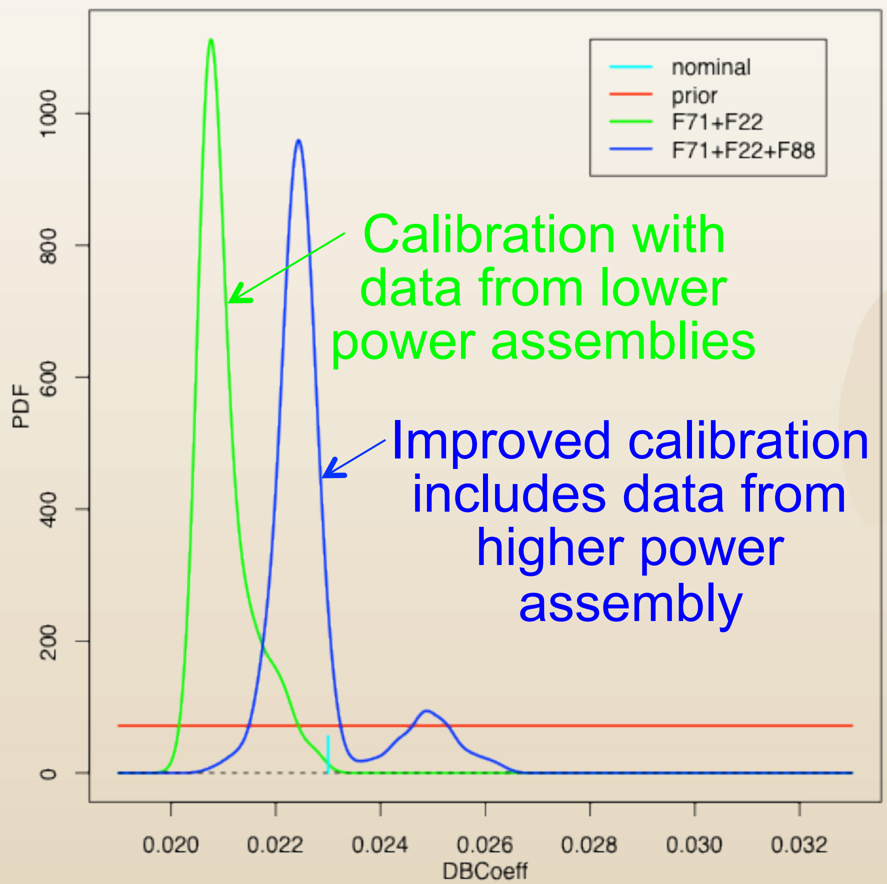


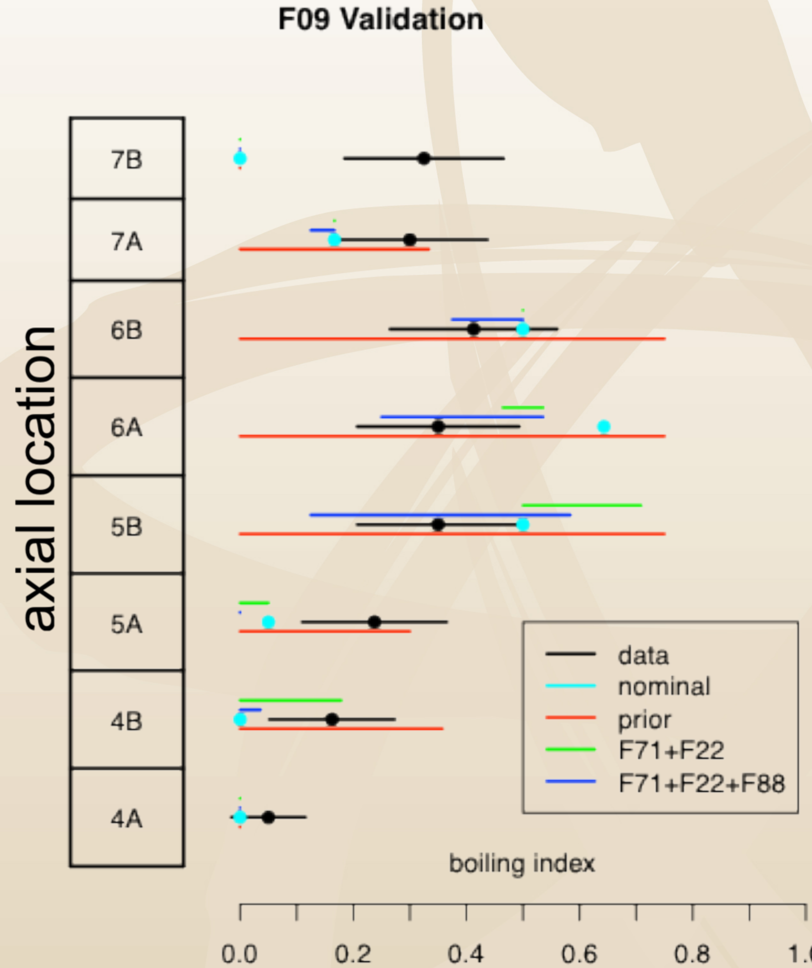
Calibration Results Strongly Dependent on Reference Experimental Data





Validation Establishes Domain of Applicability





Uncertainties have been reduced (blue vs. red), but predictability not uniformly attained at all locations

Quantitative Validation Metric

- Let $p(\,\mathbf{y}_V \mid \mathbf{y}_C\,)$ refer to the predictive distribution of validation data $\mathbf{y}_V$, given calibration data $\mathbf{y}_C$
- Let $q(\,\mathbf{y}\,)$ denote a specified reference distribution
- Let $\mathbf{Y}_V$ denote the observed validation data and define

$$S = \left\{ \mathbf{y} \,:\, \frac{p(\,\mathbf{y}\,|\,\mathbf{y}_C\,)}{q(\mathbf{y})} \geq \frac{p(\,\mathbf{Y}_V\,|\,\mathbf{y}_C\,)}{q(\mathbf{Y}_V)} \right\}$$

- Compute

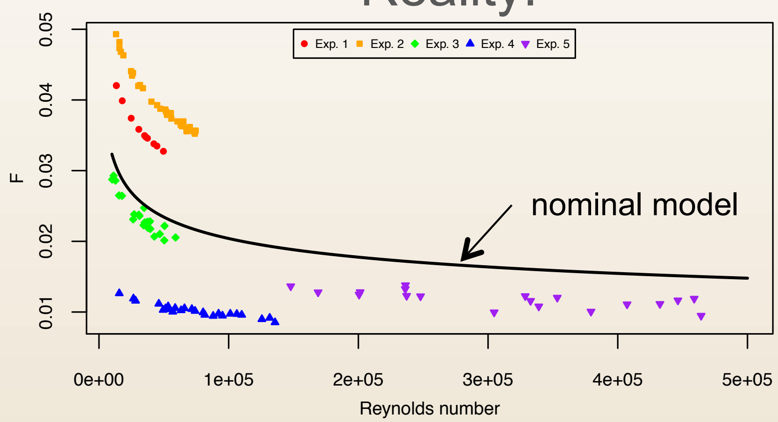
$$\gamma(\,\mathbf{Y}_V) = 1 - \int_S p(\,\mathbf{y}\,|\,\mathbf{y}_C\,)\, d\mathbf{y}$$

- If $\gamma(\,\mathbf{Y}_V\,) < T$, validation data are implausible
  - Threshold T set to, e.g., 0.05 or 0.01

Courtesy: Bob Moser

Code Calibration in Mixed Effect Settings

Reality:



- **Calibration process:**
- Prior distributions on θ and λ
- Prior distribution on group model parameter perturbations (d_i) and covariance matrix parameters (ϕ)

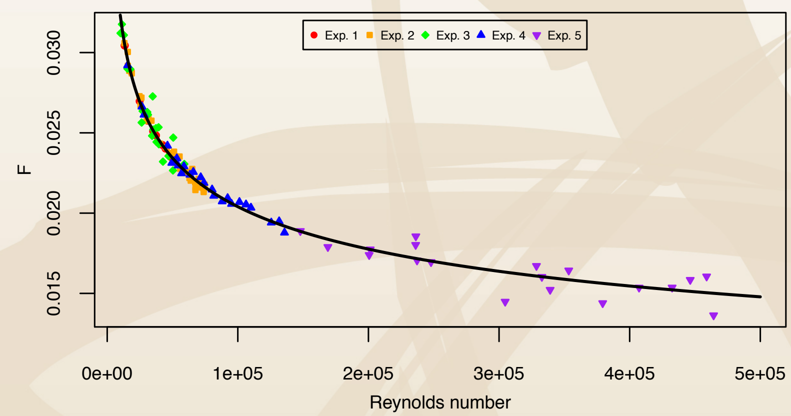
$$d_i|\phi \sim \mathcal{N}\left(0, \Phi(\phi)\right)\ ,\ \phi \sim \pi(\phi)$$

- Statistical model for experimental data

$$y_i(x_j) = \eta(x_j, \theta + d_i) + \epsilon_{ij}$$

- Posterior distributions on θ, λ, d_i, ϕ
- Statistical model has special case of common θ ($d_i = 0$)

Desirable situation:



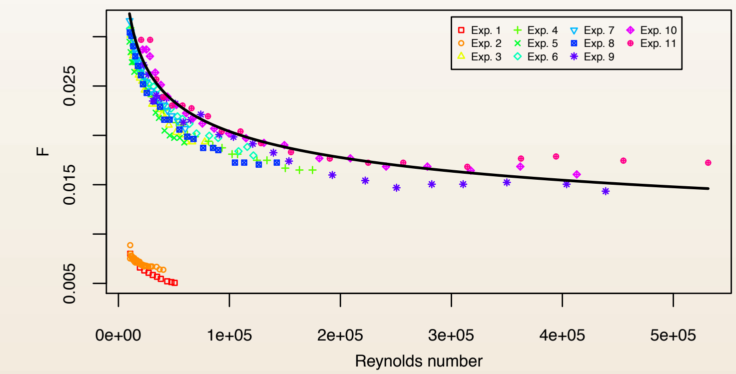
- **Calibration process:**
- Prior distributions on model parameters (θ) and error precisions (λ)
- Statistical model to explain variation in experimental data

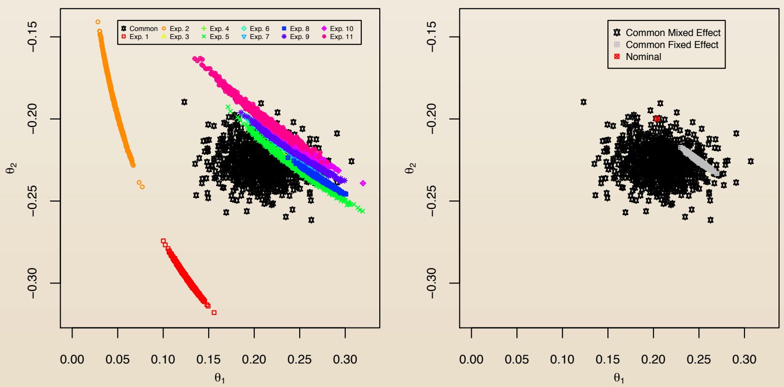
$$y_i(x_j) = \eta(x_j, \theta) + \epsilon_{ij}$$

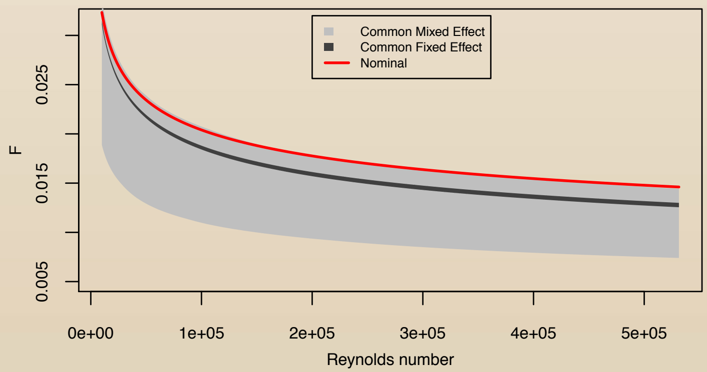
$$\epsilon_i \sim \mathcal{N}\left(0, \lambda_i^{-1} I_{n_i}\right)$$

- Posterior distributions on model parameters (θ) and error precisions (λ)

Calibration of the McAdams Correlation







- McAdams is an empirical model for friction due to the boundary layer in forced convection and turbulent flow
  - Friction factor (*f*): Proportionality constant in pressure loss correlation
  - Reynolds number (Re): Ratio of inertial to viscous forces

$$f = \theta_1 Re^{\theta_2}$$

- Prior for θ: Uniform on SME provided ranges
- In some cases, residual correlation persists after random effects adjustment
  - Modify error model:

$$\epsilon_{ij} = \delta_i(x_j) + \varepsilon_{ij}$$

- Code calibration accounts for differences among relevant experiments while also accounting for model form error

Software for Code Calibration

- Software for Code Calibration
  - Gaussian Process Models for Simulation Analysis
    - Gaussian process-based surrogate models

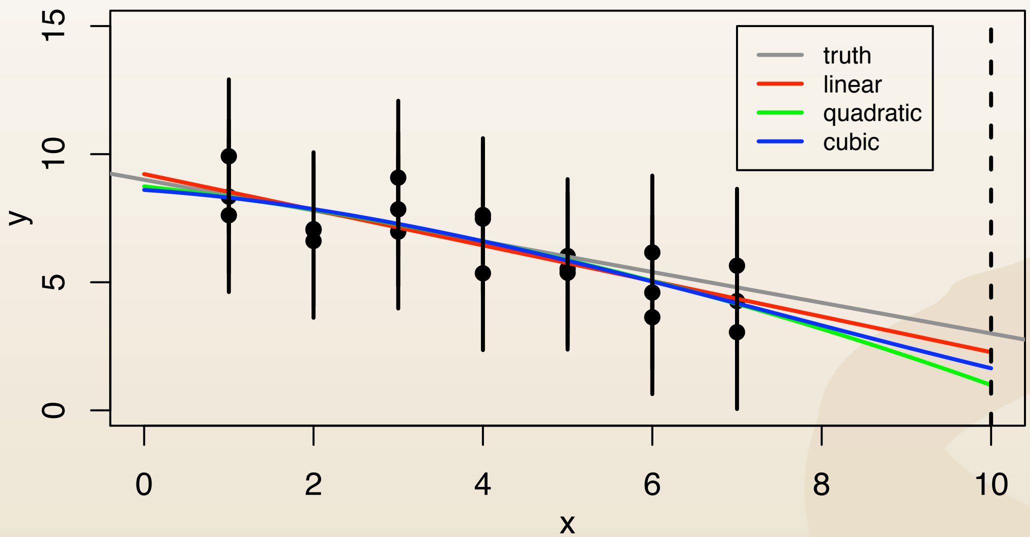
      http://www.stat.lanl.gov/source/orgs/ccs/ccs6/gpmsa/gpmsa.html
  - Bayesian Analysis of Computer Code Output (BACCO)
    - R package implementation of Kennedy-O'Hagan

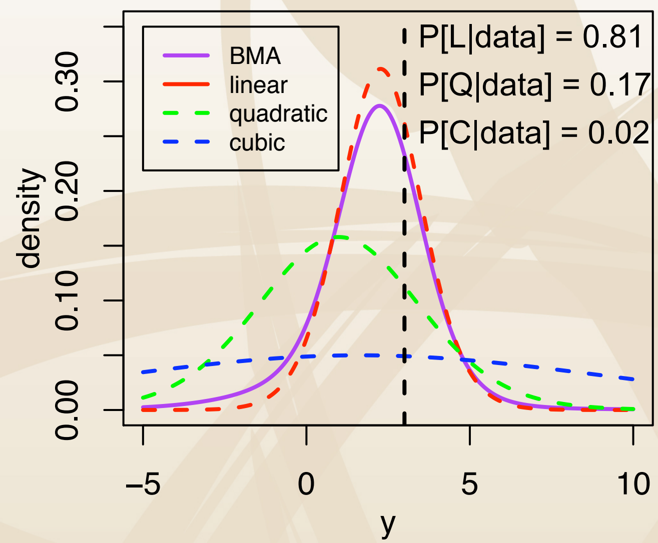
      http://cran.r-project.org/web/packages/BACCO
  - Dakota
    - Sandia National Laboratories optimization and UQ

      http://dakota.sandia.gov
  - QUESO
    - UT Austin calibration and UQ

      https://red.ices.utexas.edu/projects/software/wiki/QUESO

Role of Model Form Uncertainty in UQ



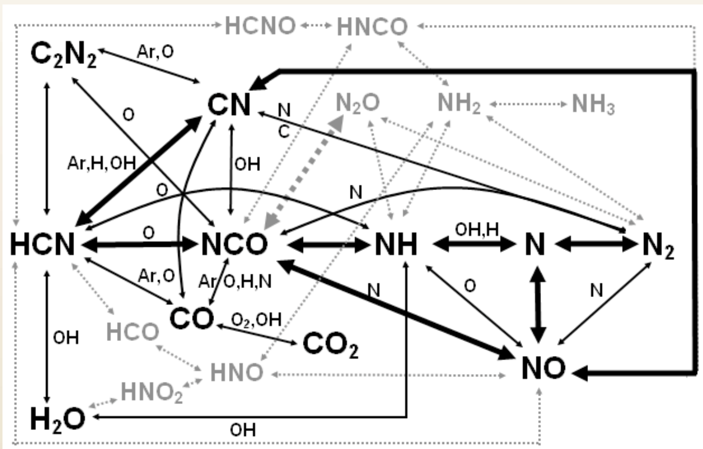


- Model form uncertainty should be incorporated in formal UQ
  - Model form *and* parametric uncertainties should be *simultaneously* calibrated to experimental data
  - Principled methodologies such as Bayesian model averaging

$$\underbrace{p(Y_P \mid \text{data})}_{\text{predictive performance distribution}} = \sum_{k=1}^{K} \underbrace{p(M_k \mid \text{data})}_{\text{posterior model probability}} \int \underbrace{p(Y_P \mid \boldsymbol{\theta}_k, M_k, \text{data})}_{\text{predictive performance distribution given model } M_k \text{ with parameters } \theta_k} p(\boldsymbol{\theta}_k \mid M_k, \text{data})\, d\boldsymbol{\theta}_k$$

Example: Calibration and Multi-Model Inference

- HCN/O_2/Ar kinetics



6 Reactions

R1: $HCN + Ar \rightarrow H + Cn + Ar$
R2: $O_2 + H \rightarrow OH + O$
R3: $O_2 + CN \rightarrow NCO + O$
R4: $HCN + O \rightarrow NCO + H$
R5: $NCO + Ar \rightarrow CO + N + Ar$
R6: $O_2 + N \rightarrow NO + O$

- Mass reaction rate of *m*-th species ($N_r = 6$, $N_s = 11$)

$$\frac{d[X_m]}{dt} = \sum_{r=1}^{N_r} \left\{ \left(\nu''_{m,r} - \nu'_{m,r}\right) k_{f,r} \prod_{m=1}^{N_s} [X_m]^{\nu'_{m,r}} + \left(\nu'_{m,r} - \nu''_{m,r}\right) k_{b,r} \prod_{m=1}^{N_s} [X_m]^{\nu''_{m,r}} \right\}$$

stoichiometric coefficient

- Reaction rate of *r*-th reactions

$$k_{f,r} = 10^{A_r} T^{m_r} \exp\left(-\frac{\Theta_r}{T}\right), k_{f,b} = \frac{k_{f,r}}{K_{C,r}} \qquad r = 1,\ldots,N_r$$

- State equation: $p = \rho R T$

Courtesy: Bob Moser

Experimental Data

T (K)	$[Ar] \times 10^6$ (mol cm^{-3})	$[HCN]/[O_2]$ (ppm/ppm)	$[N] \times 10^{12}$ mol cm^{-3} 100 µs	200 µs	300 µs	400 µs	600 µs	800 µs
2233	10.30	200/200	—	—	0.71	1.28	3.30	5.83
2382	9.50		0.97	2.30	4.33	6.50	12.50	21.60
2566	8.62		2.50	9.50	17.00	23.30	29.00	—
2447	9.13	100/100	—	0.82	1.73	2.83	6.50	11.70
2551	8.63		—	1.67	4.08	7.50	13.80	19.00
2690	8.17		1.78	7.00	13.30	20.00	31.60	—
2595	8.37	50/50	—	0.85	1.58	2.76	5.42	8.33
2600	8.36		—	0.70	1.50	2.63	5.67	8.50
2888	7.15		1.55	5.17	10.30	15.80	24.10	25.80
2998	6.80		2.75	9.67	15.80	23.30	—	—
2512	8.67	50/250	—	0.83	1.37	1.92	3.08	4.33
2594	8.28		0.67	1.58	2.83	3.83	5.58	7.75
2760	7.55		2.00	5.00	7.50	9.17	11.70	12.70
3028	6.65		8.50	16.30	20.80	20.00	14.50	9.17
3169	6.12		16.67	25.00	25.00	20.00	10.00	2.75
3391	5.42		23.00	33.30	25.00	14.20	2.42	—
2655	8.08	100/1000	2.50	4.50	6.50	7.83	9.50	8.33
2690	7.93		1.92	4.67	7.83	10.30	10.00	5.33
2718	7.97		3.16	6.17	9.17	11.50	11.30	5.83
2720	7.70		2.42	5.33	8.34	10.30	11.70	8.67
2894	7.15		8.34	13.30	15.80	14.17	5.50	1.00
2962	6.83		9.34	16.20	17.30	12.80	4.40	0.67
2989	6.85		12.00	16.70	17.50	11.70	3.00	1.17
3013	6.62		14.20	18.30	15.00	9.67	1.17	—
3110	6.23		19.00	19.20	13.30	5.83	—	—
2833	7.42	25/1000	—	1.41	1.58	1.67	1.70	1.41
3009	6.83		2.50	2.92	2.83	2.47	1.67	0.83
3241	5.97		5.67	4.83	3.33	1.67	—	—

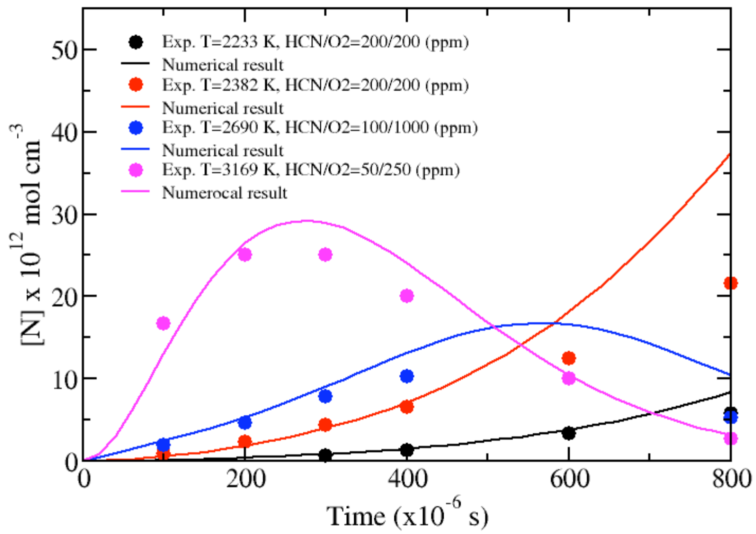


Figure: Comparison experimental data and model result

- Experimental data available ([O], [N], and [H]) for different initial conditions
  - N_{exp} = 79 data sets, each set has time history data at N_T = 6 time points

Courtesy: Bob Moser

Stochastic Models

- General form of stochastic models

$$\mathbf{x}(0) = \mathbf{x}(0; \mathbf{u}(0))$$

$$\mathbf{x}(t) = \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t), \gamma(t), \boldsymbol{\theta}) \in \Re^{N_s} \qquad \text{[stochastic dynamic model]}$$

$$\mathbf{y}(t) = \mathbf{h}(t, \mathbf{x}(t), \mathbf{u}(t), \varepsilon(t), \boldsymbol{\theta}) \in \Re^{N_o} \qquad \text{[observation equation]}$$

- Definitions
  - t = time
  - $\mathbf{x}(t)$ = model state vector at time t
  - $\mathbf{y}(t)$ = measured output vector at time t
  - $\mathbf{u}(t)$ = system input vector (eg. temperature, pressure, etc.) at time t
  - θ = uncertain model parameters (calibration)
  - $\gamma(t) \in \Re^{Ns}$ = model equation error function/noise at time t
  - $\varepsilon(t) \in \Re^{No}$ = output equation error function/noise at time t

Courtesy: Bob Moser

Model Class

- M_1-M_4
  - Multiplicative error function in output equations $y(t) = x(t)\,\gamma(t)\,\varepsilon(t)$
  - 11 uncertain parameters (including the physical parameters A_i, m_i, and Θ_i (i = 1, 3, 4) and the error function variances σ_s^2 and σ_o^2)
- M_5
  - Same as M_3 except different error structure $y(t) = x(t)\,(1+\gamma(t)) + \varepsilon(t)$
- M_6
  - Same as M_3 except obs. errors have a covariance structure in time

$$\operatorname{cov}(\varepsilon(t_m),\varepsilon(t_n)) = \sigma_o^2 \exp\left[-\left(\frac{|t_m - t_n|}{l_o}\right)^{r_0}\right]; \qquad \sigma_o^2, l_o, r_0 \text{ uncertain}$$

- M_7
  - Same as M_6 with modified state equation $x_k(t) = f_k(t, \mathbf{x}(t), \mathbf{u}(t), \boldsymbol{\theta}) \exp(\gamma_k(t))$
  - $\gamma_k(t)$ modeled as a Gaussian process
  - Cross-correlation structure

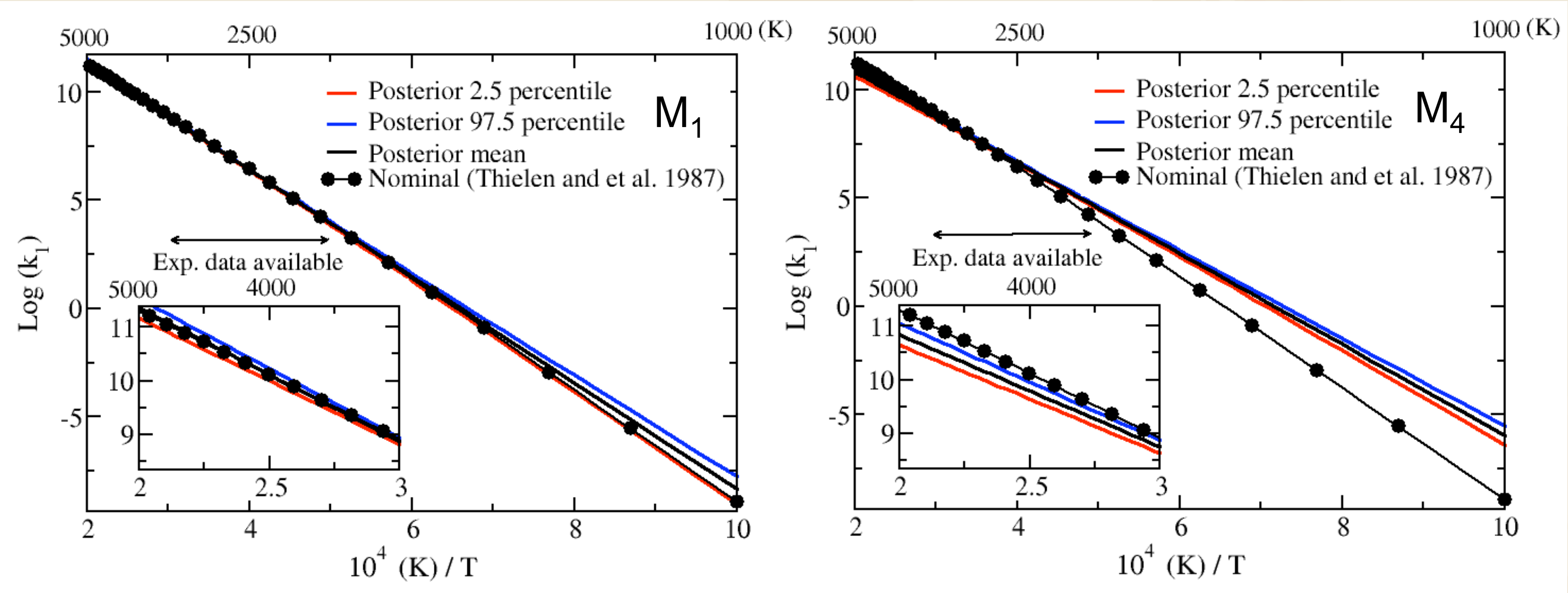
$$\operatorname{cov}(\gamma_k(t_m),\gamma_l(t_n)) = \sigma_s^2 \exp\left[-\left(\frac{|t_m - t_n|}{l_s}\right)^{r_s}\right]; \qquad \sigma_s^2, l_s, r_s \text{ uncertain}$$

Courtesy: Bob Moser

Model Plausibility

- Reaction rates are sensitive to choice of hypotheses
  - Uncertainties quantified: parametric, model and observational errors, and model form

	M_1	M_2	M_3	M_4	M_5	M_6	M_7
$\log(L(M_j \mid data)$	-264.0	-264.6	-256.9	-269.2	-313.5	-260.6	-249.0
$P(M_j \mid data)$	3.1e-7	1.7e-7	3.7e-4	1.7e-9	0	9e-6	0.9996



Courtesy: Bob Moser

Predictive Maturity

- Efficiently achieve accuracy in global predictions of discrepancy, calibrated code, or physical reality

$$y(x_i) = \eta(x_i;\theta) + \delta(x_i) + \varepsilon(x_i) \quad \text{for } i = 1 \ldots N_{Tests}$$

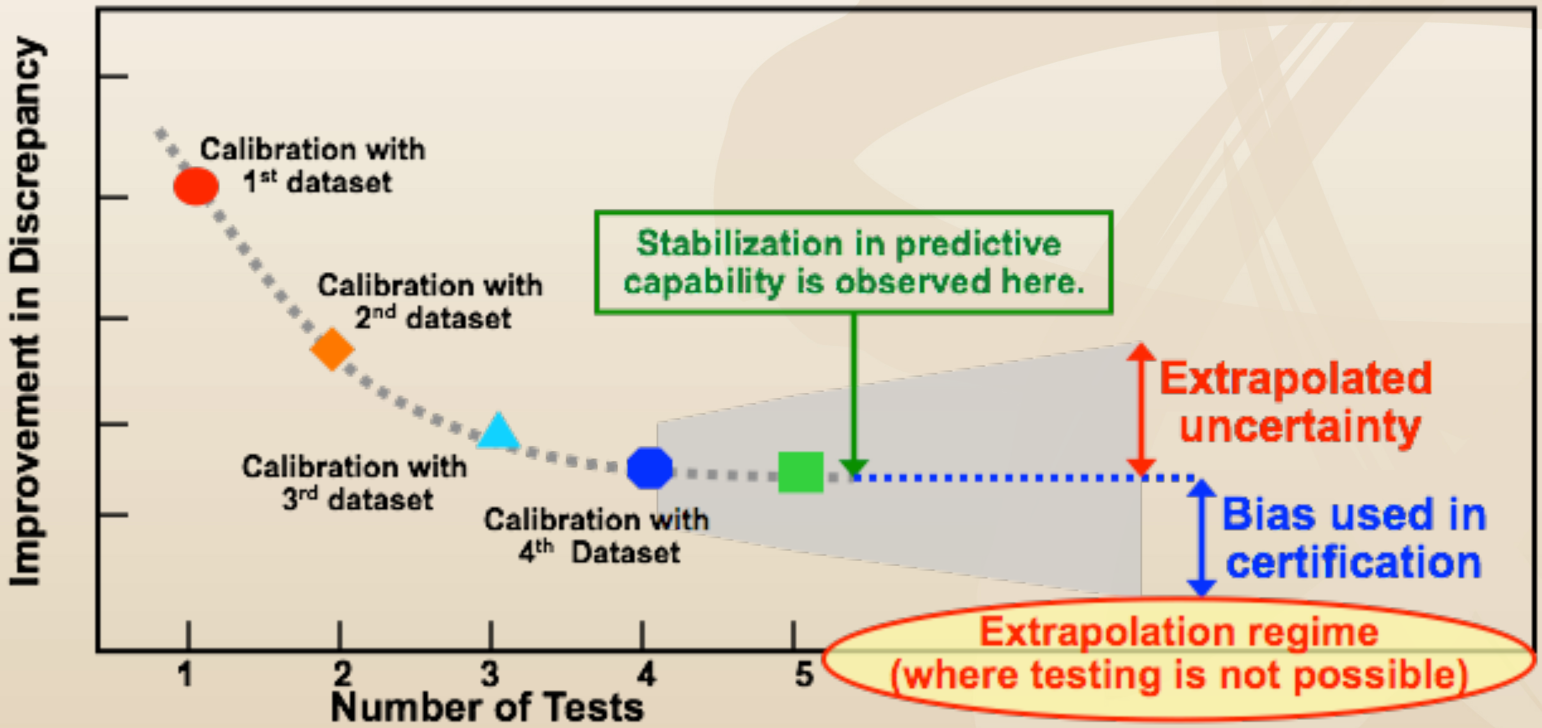


Figure: F. Hemez

Experiment Design Strategies

- Several options exist if given a fixed experimental budget
  - single-stage design
    - space-filling LHD, Sobol' sequence, scrambled Sobol' sequence
  - sequential design
    - Initial space-filling design followed by sequential augmentation
- Sequential design required for augmenting existing data
  - Design criteria used for augmentation
    - distance-based criteria
    - IMSE, MMSE
    - Minimum information gain, Maximum entropy
    - Lam and Notz (2008) EIGF criterion
  - Batches of runs desired

A Batch Sequential Algorithm

- Estimate model parameters using runs from initial design X_0
- Set $X_1 = (X_0^t, X_b^t)^t$ and obtain X_b by optimizing a design criterion with respect to the proposed *b* additional runs
- Collect runs from X_b and re-estimate model parameters using the entire set of runs from the augmented design X_1
- Set X_0 to the augmented design X_1 and repeat steps (2)-(3) until termination
  - stopping criteria: experiment budget expended, insignificant improvement in design criterion value

Optimization: Modified Federov Exchange

Initialize $\mathbf{x}_1^*, \ldots, \mathbf{x}_b^*$.

While (ΔCriterion > Stopping Value & Count < MaxCount)

For $i = 1, \ldots, b$

Optimize Criterion w.r.t. $\mathbf{x}_i^*$,
holding all other input vectors fixed.

End

Compute ΔCriterion.

End

Return optimized $\mathbf{x}_1^*, \ldots, \mathbf{x}_b^*$ and Criterion.

Batch Sequential Expected Improvement Criteria

- Expected improvement criteria are typically formulated as one-step iterations

$$\text{Choose next design site } \boldsymbol{x} \text{ to } \textit{maximize } E\left[I(\boldsymbol{x})|\boldsymbol{y}_2\right]$$

Improvement Function

Current Data

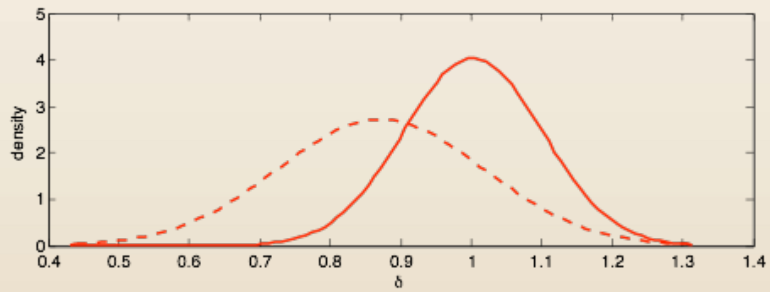
- Straightforward extension allows for batch updates

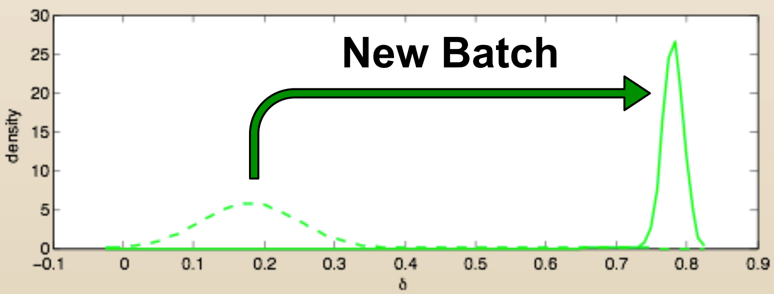
$$\text{Choose next design sites } \boldsymbol{x}_1^*, \ldots, \boldsymbol{x}_b^* \text{ to } \textit{minimize}$$
$$\text{the } \textit{maximum } E\left[I(\boldsymbol{x}|\boldsymbol{x}_1^*, \ldots, \boldsymbol{x}_b^*)|\boldsymbol{y}_2\right]$$

Bayesian Design

- Consider impact of a new batch of data on prediction at arbitrary input **x** in design region
- Improvement function:

$$I = -\log\left(\frac{\pi(\mathbf{y}_1|\mathbf{y}_2, \mathbf{y}_3)}{\pi(\mathbf{y}_1|\mathbf{y}_2)}\right)$$





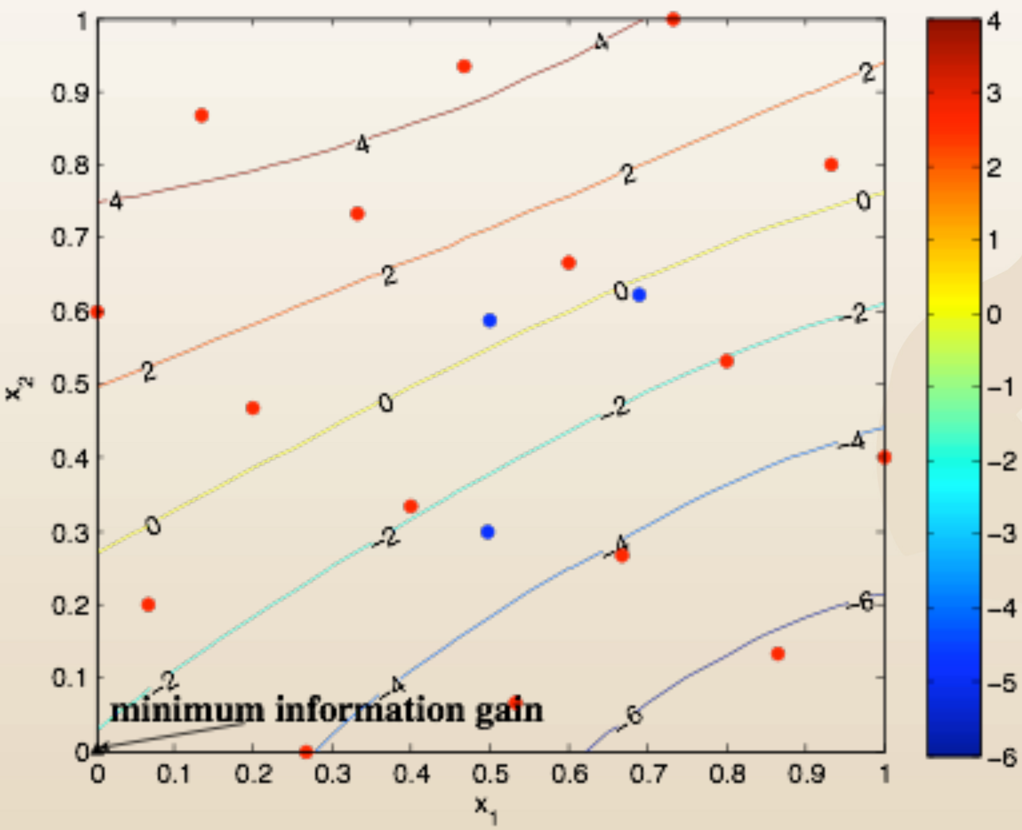
"Maximize the minimum information gain"

y_1: output predicted at input **x**
y_2: current data
y_3: hypothetical data from new batch

Choose batch $\boldsymbol{x}_1^*, \ldots, \boldsymbol{x}_b^*$ to

minimize $\max_{\boldsymbol{x}} E(I|\boldsymbol{y}_2)$

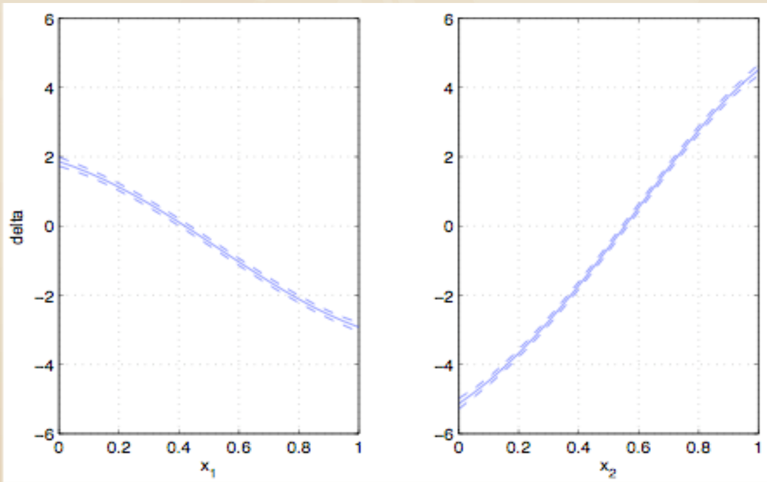
Example: Bayesian Design



Contours of predicted discrepancy function based on 16 initial experiments

- 2 x, 1 θ
- 16 initial experiments
- 3 new experiments
- Use IG criterion
- Minimum IG: 0.044

Main Effect Functions



Sequential Design for Optimal Calibration

- Utility function defined as entropy loss due to new data $\mathbf{y}_3$

$$U(\mathbf{y}_3) = \int_\Theta \pi(\theta \,|\, \mathbf{y}_2, \mathbf{y}_3) \log \pi(\theta \,|\, \mathbf{y}_2, \mathbf{y}_3)\, d\theta - \int_\Theta \pi(\theta \,|\, \mathbf{y}_2) \log \pi(\theta \,|\, \mathbf{y}_2)\, d\theta$$

- Compute expected utility with respect to unknown future observations

$$E\left[U(\mathbf{y}_3) \,|\, \mathbf{y}_2\right] = \int \int_\Theta \pi(\theta, \mathbf{y}_3 \,|\, \mathbf{y}_2) \log \frac{\pi(\theta, \mathbf{y}_3 \,|\, \mathbf{y}_2)}{\pi(\theta \,|\, \mathbf{y}_2)\, \pi(\mathbf{y}_3 \,|\, \mathbf{y}_2)}\, d\theta\, d\mathbf{y}_3$$

  - Mutual information between model parameters θ and new data $\mathbf{y}_3$ given available data $\mathbf{y}_2$
  - Smaller expected utility implies new data $\mathbf{y}_3$ does not inform as well about model parameters θ given available data $\mathbf{y}_2$

Choose batch $\mathbf{x}_1^*, \ldots, \mathbf{x}_b^*$ to maximize $E\left[U(\mathbf{y}_3) \,|\, \mathbf{y}_2\right]$

Data Assimilation Framework

- Time steps $t = 1,2, \ldots$
  - State vector $\{\theta_1, \theta_2, \ldots\}$ and observations $\{\mathbf{y}_1, \mathbf{y}_2, \ldots\}$
- Prior model for state process
  - $\pi(\theta_1)$ and $\pi(\theta_t \mid \theta_{t-1})$
- Observational data model
  - $\mathbf{y}_t = \eta(\theta_t) + \mathbf{e}_t$; $\mathbf{e}_t$ iid $N(\mathbf{0}, \Sigma_y)$
- **Goal**: At each time t, produce draws from $\pi(\theta_t \mid \mathbf{y}_{1:t})$
  - $\mathbf{y}_{1:t}$ denotes all data up to time t

Data Assimilation Algorithm

1. At time $t-1$, the ensemble of state vectors $\left\{\theta^{(1)}_{t-1,1},\ldots,\theta^{(1)}_{t-1,M}\right\}$ are treated as draws from $\pi\left(\theta_{t-1}\,|\,\mathbf{y}_{1:t-1}\right)$
2. Propagate each $\theta^{(1)}_{t-1,k}$ according to $\pi\left(\theta_t\,|\,\theta_{t-1}\right)$, producing an ensemble of draws $\{\theta^{\circ}_{t,1},\ldots,\theta^{\circ}_{t,M}\}$, from $\pi\left(\theta_t\,|\,\mathbf{y}_{1:t-1}\right)$
3. Given observations $\mathbf{y}_t$, update each $\theta^{\circ}_{t,k}$, producing an ensemble $\left\{\theta^{(1)}_{t,1},\ldots,\theta^{(1)}_{t,M}\right\}$ from $\pi\left(\theta_t\,|\,\mathbf{y}_{1:t}\right)$

Filtering Methods

- Particle Filter

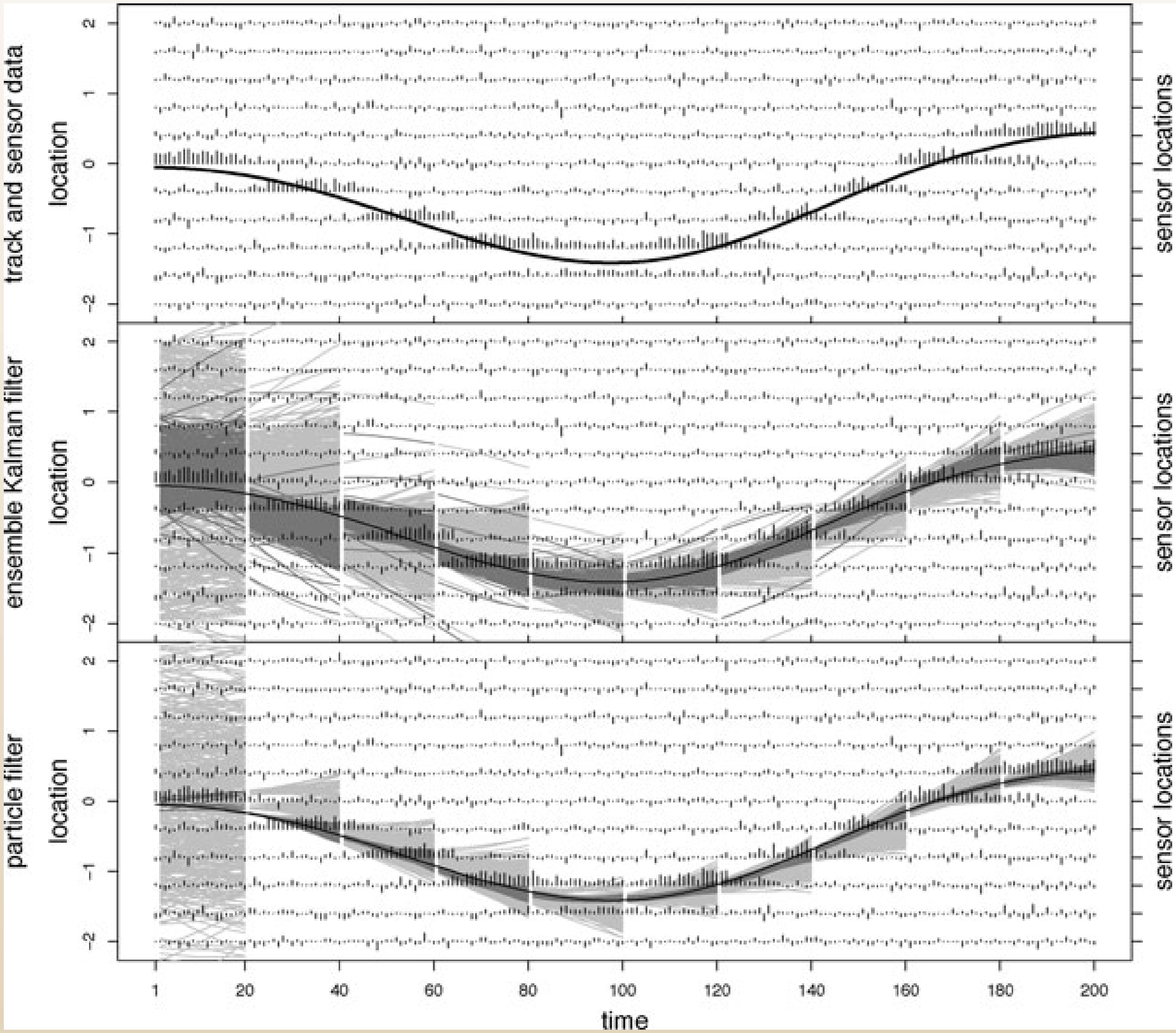
Sample from $\{\theta^\circ_{t,1}, \ldots, \theta^\circ_{t,M}\}$ according to importance weights $\{w_1, \ldots, w_M\}$ given by

$$w_k \propto \exp\left[-\frac{1}{2}\left(\mathbf{y}_t - \eta\left(\theta^\circ_{t,k}\right)\right)^T \boldsymbol{\Sigma}_y^{-1}\left(\mathbf{y}_t - \eta\left(\theta^\circ_{t,k}\right)\right)\right]$$

- Ensemble Kalman Filter (EnKF)

1. From $\left\{\begin{pmatrix}\theta^\circ_{t,1} \\ \eta\left(\theta^\circ_{t,1}\right)\end{pmatrix}, \ldots, \begin{pmatrix}\theta^\circ_{t,M} \\ \eta\left(\theta^\circ_{t,M}\right)\end{pmatrix}\right\}$ construct the sample covariance $\boldsymbol{\Sigma}_{\text{pr}} = \begin{pmatrix}\boldsymbol{\Sigma}_{\theta\theta} & \boldsymbol{\Sigma}_{\theta\eta} \\ \boldsymbol{\Sigma}_{\eta\theta} & \boldsymbol{\Sigma}_{\eta\eta}\end{pmatrix}$
2. Draw perturbed data vector $\mathbf{y}_k \sim \mathcal{N}\left(\mathbf{y}_t\,, \boldsymbol{\Sigma}_y\right)\,, k = 1, \ldots, M$
3. Set $\theta^{(1)}_{t,k} = \theta^\circ_{t,k} + \boldsymbol{\Sigma}_{\theta\eta}\left(\boldsymbol{\Sigma}_{\eta\eta} + \boldsymbol{\Sigma}_y\right)^{-1}\left(\mathbf{y}_k - \eta\left(\theta^\circ_{t,k}\right)\right)\,, k = 1, \ldots, M$ (*)

Data Assimilation Example



An object moves vertically over time in the presence of 11 sensors whose locations are shown in the right, vertical axis. Its path is given by the solid black line. The sensor signals are given by the 11 horizontal time series. As the object nears a sensor, the signal becomes elevated.

Every 20 seconds, the object's path is predicted 20 seconds into the future using the EnKF (middle) and the particle filter (bottom).

Prior paths θ_t (light lines) are extended from the previous time period's posterior paths θ_{t-1} (dark lines) according to a stochastic model for the object's path. Given the sensor reading for the current time period y_t, these prior paths are updated.

A model $\eta(\theta_t)$ produces an expected signal given a path θ_t that is comparable to the sensor signals y_t for the current time period.

The EnKF perturbs each prior path to produce a posterior path according to the formula in (*). The particle filter samples posterior paths using likelihood weighted draws over the prior paths.

References

Kennedy, M. and O'Hagan, A. (2001). "Bayesian calibration of computer models (with discussion)," *Journal of the Royal Statistical Society, Series B*, 63, 425-464.

Higdon, D., Kennedy, M., Cavendish, J. C., Cafeo, J. A., and Ryne, R. D. (2004). "Combining field data and computer simulations for calibration and prediction," *SIAM Journal on Scientific Computing*, 26, 448-466.

Higdon, D., Gattiker, J., Lawrence, E., Jackson, C., Tobis, M., Pratola, M., Habib, S., Heitmann, K., and Price, S. (2013). "Computer model calibration using the ensemble Kalman filter," *Technometrics*, 55, 488-500.

Higdon, D., Gattiker, J., Williams, B., and Rightley, M. (2008). "Computer model calibration using high-dimensional output," *Journal of the American Statistical Association*, 103, 570-583.

Terejanu, G., Upadhyay, R.R., and Miki, K. (2012). "Bayesian experimental design for the active nitridation of graphite by atomic nitrogen," *Experimental Thermal and Fluid Science*, 36, 178-193.

Wasserman, L. (2000). "Bayesian model selection and model averaging," *Journal of Mathematical Psychology*, 44, 92-107.

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

Williams, B.J., Loeppky, J.L., Moore, L.M., and Macklem, M.S. (2011). "Batch sequential design to achieve predictive maturity with calibrated computer models," *Reliability Engineering and System Safety*, 96, 1208-1219.