SAND2007-5696C

Calibration under Uncertainty: Comparison of Least-Squares and Bayesian methods

Laura P. Swiler, John McFarland, Kay Vugrin
Sandia National Laboratories
Optimization and Uncertainty Estimation
Albuquerque, NM
87185-1318
lpswile@sandia.gov
505-844-8093

Presentation at CSRI Workshop:
Large Scale Inverse Problems and Quantification of Uncertainty
Santa Fe, NM
September 10-12, 2007

*Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Goals of Presentation

- Present two approaches to parameter estimation
  - Classical Statistics
    - Methods for generating joint confidence regions in nonlinear regression problems
  - Bayesian Calibration Methods
    - Parameters are random variables, given a prior distribution, want to estimate a posterior distribution that incorporates the data
- Compare and contrast the approaches, identify assumptions/difficulties with each
- Apply the approaches to a real application
  - Calore thermal simulation of a component encapsulated by foam

Why is parameter estimation difficult?

- Parameter estimation / model calibration:
  - Use observations of response to make inference about model inputs
- Observations contain noise
- Model is imperfect
- Many combinations of parameter values yield comparable fits
- Model is expensive

Nonlinear Regression Methods

- **Extension of linear regression:** $d = f(\boldsymbol{\theta}, \mathbf{x}) + \varepsilon$
- **To determine the optimal parameters, one minimizes the error sum of squares S:**

$$S(\boldsymbol{\theta}) = \sum_{i=1}^{n} [f(\boldsymbol{\theta}, \mathbf{x}_i) - d_i]^2 = \sum_{i=1}^{n} r(\boldsymbol{\theta})^2$$

- **Specialized techniques have been developed to find the least squares estimator $\hat{\boldsymbol{\theta}}$ of the true minimum $\boldsymbol{\theta}^*$**

- **If one assumes that the residuals are close to zero near the solution, the Hessian matrix of *S* can be approximated using only first derivatives of the residuals *r*.**
- **Gauss-Newton methods are particularly effective on this type of problem**
- **We have three such methods in DAKOTA: Gauss-Newton based in OPT++, NLSSOL (SQP), NL2SOL (trust-region method)**

Joint Confidence Regions

- **Linear Approximation Method**
  - **In linear regression, SSE is quadratic, and contours of constant SSE are ellipsoids.**
  - **Can approximate the nonlinear function with Taylor series expansion about the parameter estimate** $\hat{\boldsymbol{\theta}}$

$$\{\boldsymbol{\theta} : (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) H(\hat{\boldsymbol{\theta}}) (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \leq S(\hat{\boldsymbol{\theta}}) (\frac{p}{n-p}) F^{\alpha}_{p,n-p}\}$$

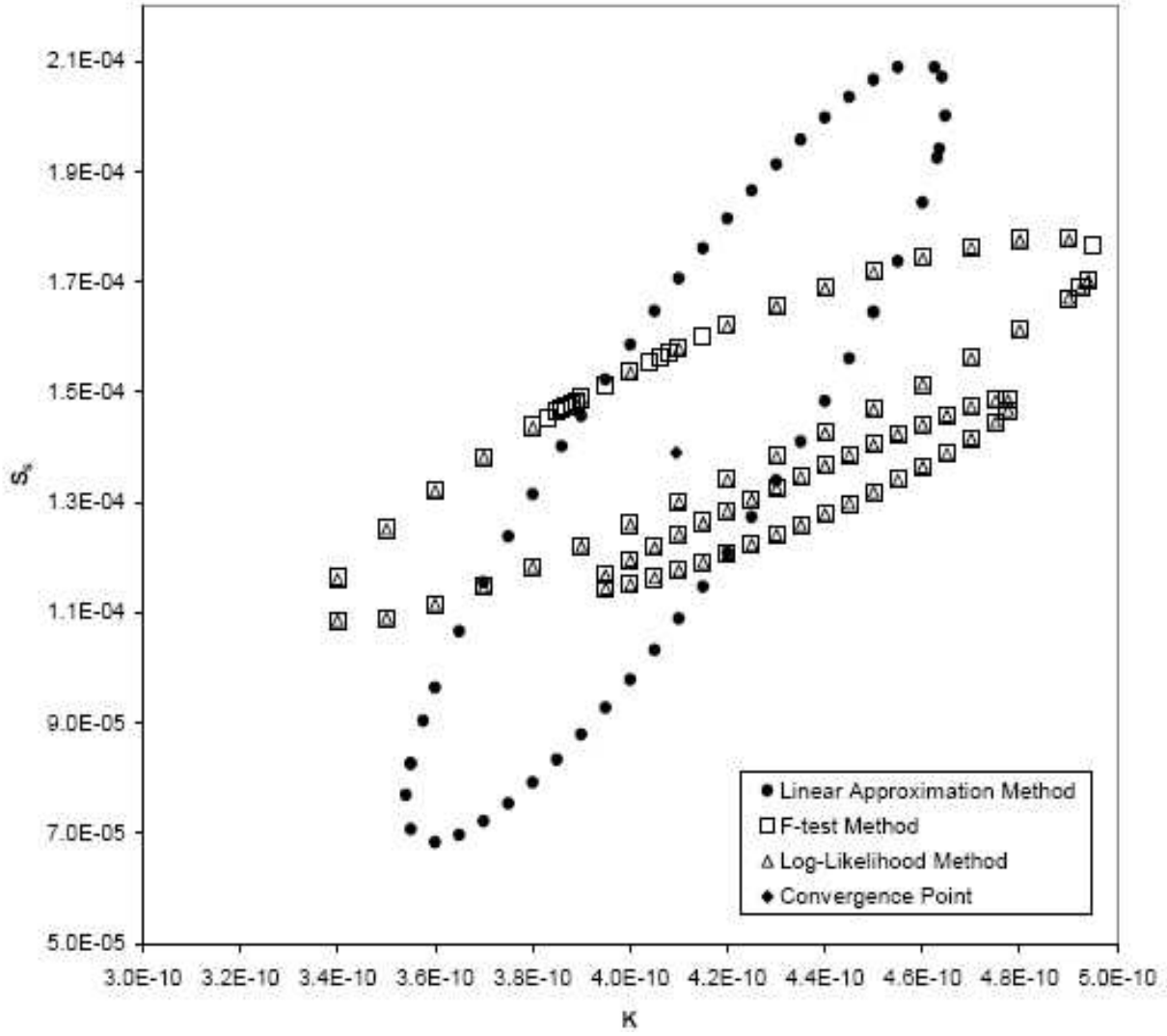
  - **For very nonlinear models or in situations where the Hessian approximation is poor, this can be inaccurate**
- **F-test Method**
  - **Based on the assumption that the error terms are jointly normally distributed**

$$\{\boldsymbol{\theta} : \frac{S(\boldsymbol{\theta}) - S(\hat{\boldsymbol{\theta}})}{S(\hat{\boldsymbol{\theta}})} \leq (\frac{p}{n-p}) F^{\alpha}_{p,n-p}\}$$

- **Log-likelihood method**
  - **Based on assumptions about the likelihood estimator of the parameters**

$$\{\boldsymbol{\theta} : n[\log S(\boldsymbol{\theta}) - \log S(\hat{\boldsymbol{\theta}})] \leq \chi^2_p(\alpha)\}$$

Quick example



Sandia National Laboratories

Bayesian Calibration

- **Uncertain input parameters are given prior distribution functions**
- **The priors are refined based on the data resulting in posterior distributions which represent the new state of knowledge**
- **Bayes theorem:** $$f(\theta \mid D) = \frac{\pi(\theta) f(D \mid \theta)}{\int_{\theta} \pi(\theta) f(D \mid \theta) \partial(\theta)}$$
  - **Posterior proportional to prior * likelihood**
    $$f(\theta \mid D) \propto \pi(\theta) L(\theta)$$
  - **Assumption that there is a probabilistic relationship between experimental data and model output that can be defined by a likelihood function**

Bayesian Calibration

- **Experimental data = Model output + error**

$$d_i = G(\boldsymbol{\theta}, \mathbf{x}_i) + \varepsilon_i$$

- **Error term incorporates measurement errors and modeling errors (can get more complex with a bias term)**
- **If we assume error terms are independent, zero mean Gaussian random variables with variance σ^2, the likelihood is:**

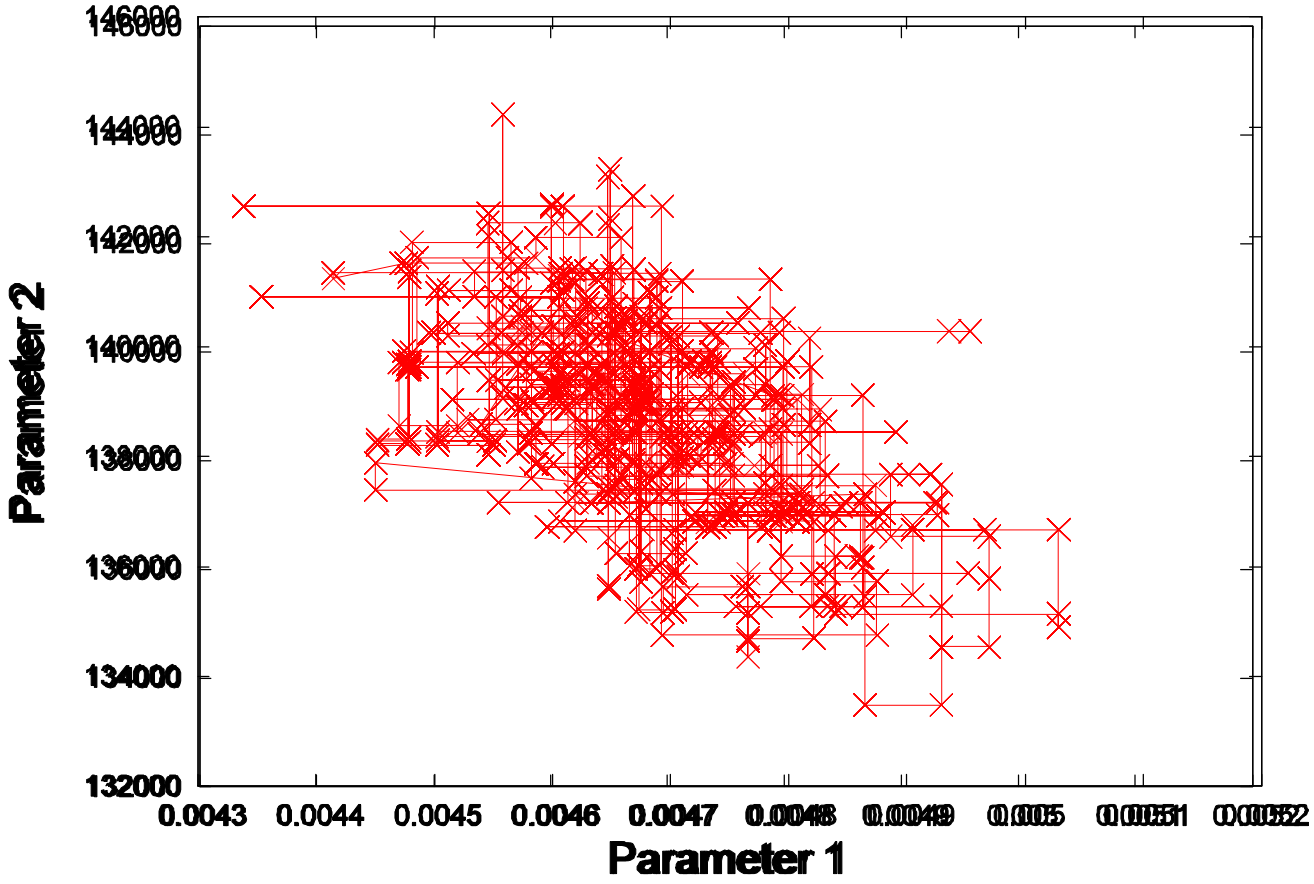
$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(d_i - G(\boldsymbol{\theta}, \mathbf{x}_i))^2}{2\sigma^2} \right]$$

- **How do we obtain the posterior?**
  - **It is usually too difficult to calculate analytically**
  - **We use a technique called Monte Carlo Markov Chain (MCMC)**

Markov Chain Monte Carlo

- In MCMC, the idea is to generate a sampling density that is approximately equal to the posterior. We want the sampling density to be the stationary distribution of a Markov chain.
- Metropolis-Hastings and Gibbs sampling are the most commonly used algorithms
- Both have the idea of a "proposal density" which is used for generating X_{i+1} in the sequence, conditional on X_i.
- Implementation issues: How long do you run the chain, how do you know when it is converged, how long is the burn-in period, etc.?
- ACCEPTANCE RATE is CRITICAL. Need to tune the proposal density to get an "optimal" acceptance rate, 45-50% for 1-D problems, 23-26% for high dimensional problems
- COMPUTATIONALLY VERY EXPENSIVE

Markov Chain Monte Carlo



Sandia National Laboratories

Gaussian Processes

- Since MCMC requires tens of thousands of function evaluations, it is necessary to have a fast-running surrogate model of the simulation
- Gaussian process surrogates are often used
- GPs are based on spatial statistics for interpolating data
- In GPs, the response values $Y(x_1)$, ..$Y(x_k)$ are modeled as a group of multivariate normal random variables
- A GP is fully specified by its mean function $\mu(x) = E[Y(x)]$ and its covariance function $C(x, x')$
- Often a constant mean is used with covariance:

$$C(\mathbf{x}, \mathbf{x}') = v_o \exp\{-\sum_{u=1}^{d} \rho_u^2 (\mathbf{x}_u - \mathbf{x}'_u)^2\}$$

- This covariance function involves the product of d squared-exponential covariance functions with different lengthscales on each dimension. The form of this covariance function captures the idea that nearby inputs have highly correlated outputs.

Gaussian Processes

- **Pros:**
  - Non-parametric
  - Handle multiple dimensions
  - Potentially accurate (potentially)
  - Can represent the uncertainty in prediction at new points

$$E[y_{n+1} \mid y_1, y_2, .. y_n] = k^T C^{-1} y$$
$$Var[y_{n+1} \mid y_1, y_2, .. y_n] = C(x_{n+1}, X_{n+1}) - k^T C^{-1} k$$

- **Cons:**
  - Estimation of the parameters defining the GP (hyperparameters) such as the lengthscales, the process variance, and mean parameters, is difficult
  - Can use maximum likelihood estimation or Bayesian approach
  - We use MLE
  - The MLE function may be multi-modal or have a flat landscape
  - GP Methods work well when the covariance matrix is not ill-conditioned but often it is
  - Jitter term is sometimes added to the diagonal terms of the covariance matrix to make it better conditioned
  - John McFarland developed a point selection algorithm which selects a subset of "optimal" points which minimize a cross-validation prediction error

Bayesian Calibration: Overall approach

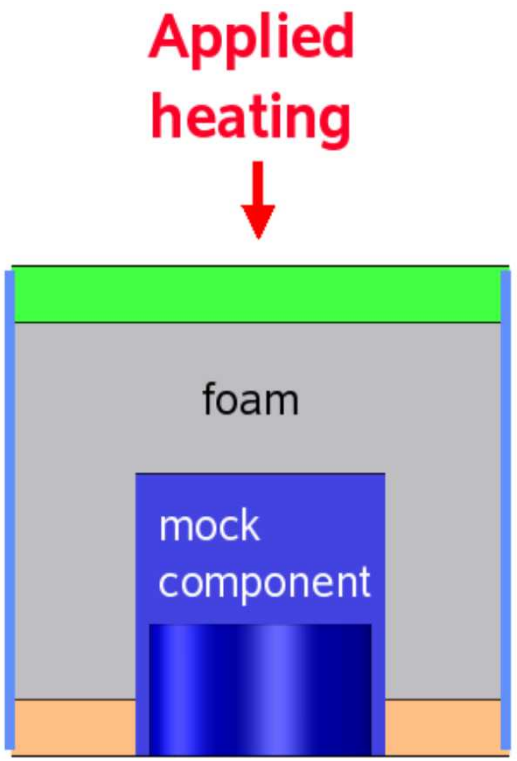
- Take initial set of samples from simulation
  - Use LHS or space-filling design
- Develop Gaussian process approximation of the simulation
- Put priors on the input parameters
- Perform Bayesian analysis using MCMC
- Generate and analyze posterior distributions
- NOTE: GP surrogate adds a layer of uncertainty. However, this is explicitly modeled in the revised likelihood:

$$L(\boldsymbol{\theta}) = 2\pi^{-n/2}\left|\Sigma\right|^{-1/2}\exp\left[-\frac{1}{2}(d_i-\mu_{GP})^T\Sigma^{-1}(d_i-\mu_{GP})\right]$$

$$\Sigma = \sigma^2 I + \Sigma_{GP}$$

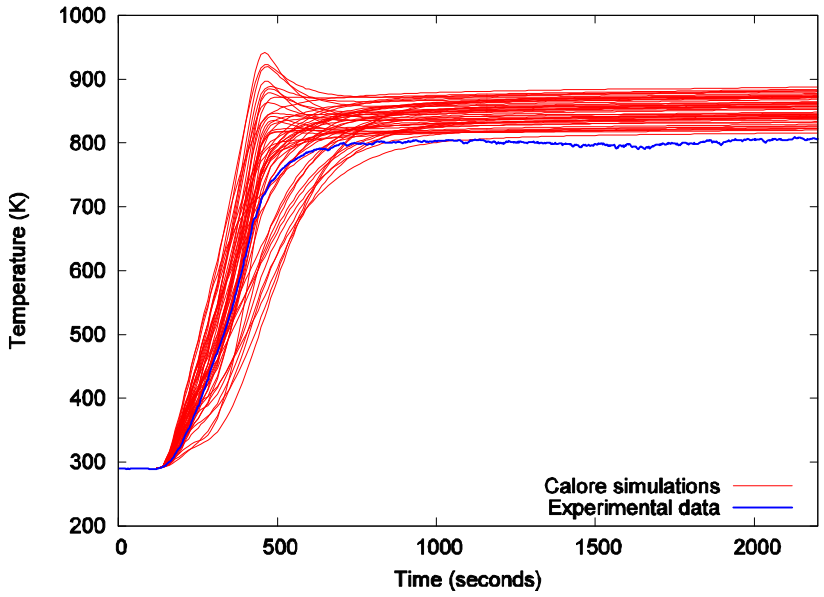
- Total uncertainty = (observation + model uncertainty) + code uncertainty

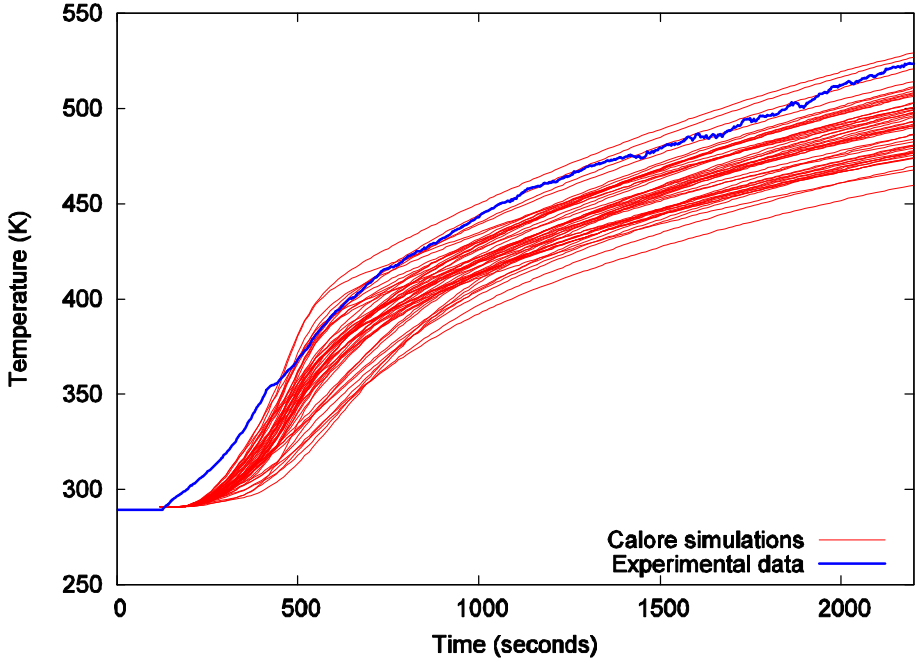
Calore Application: "Foam in a Can"



- Five calibration parameters:
  - Applied heat flux (*q*2, *q*3, *q*4, *q*5)
  - Foam Final Pore Diameter FPD
- Time-dependent temperature response
  - 20 thermocouple locations
  - Reduced to 9 "locations"
  - Data taken approximately every 1.8 seconds
  - Even with data reduction, still have over 10,000 data points

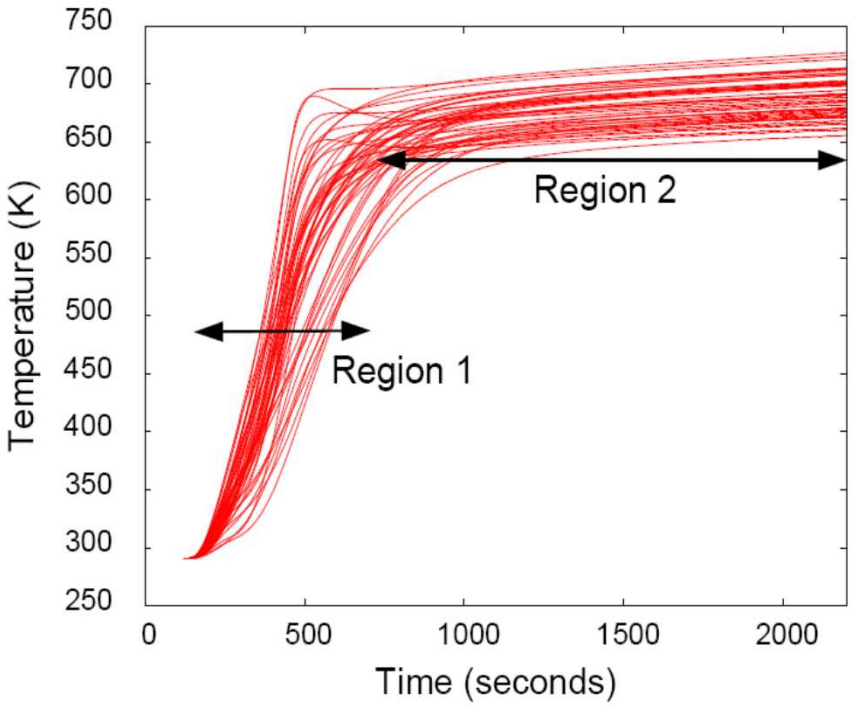
Response Data





Sandia National Laboratories

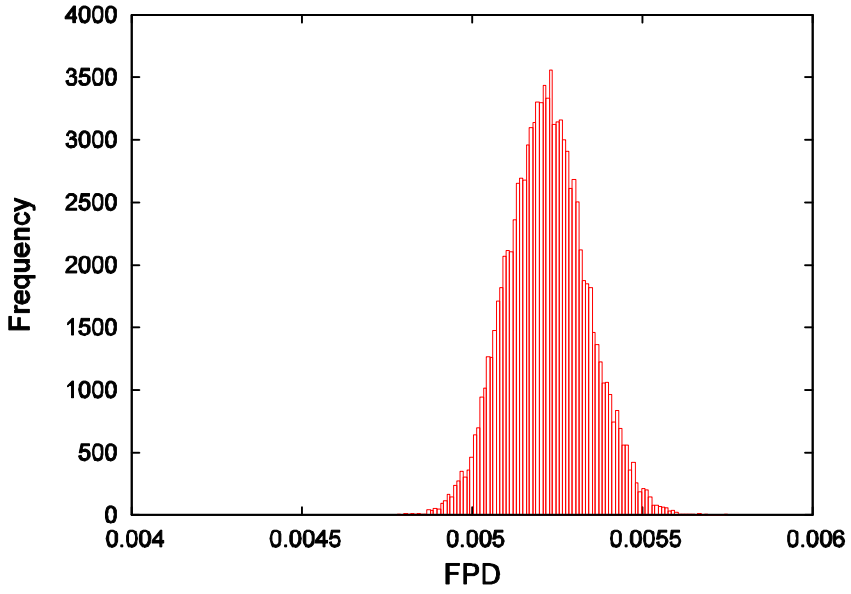
Response Data



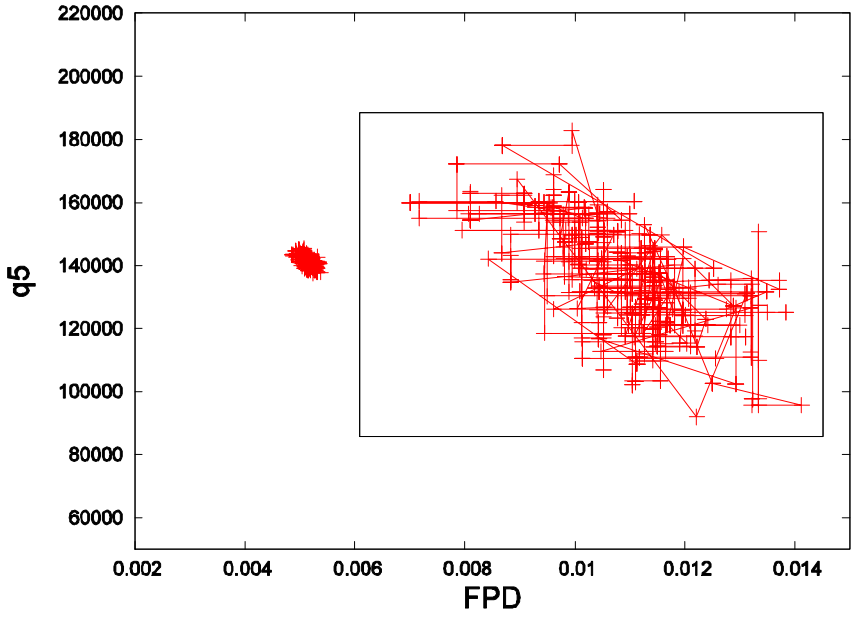
- 50 LHS samples of CALORE code
- Each sample predicts at each of 9 locations
- Response at each location is fit with 2 GPs representing different behavior after 500 seconds
- Reduced the code to 12.5 sec increments
- 18 GPs
- Used MLE and point selection to estimate GP parameters for each
- Then, use the 18 GPs in the likelihood function → MCMC to construct posterior

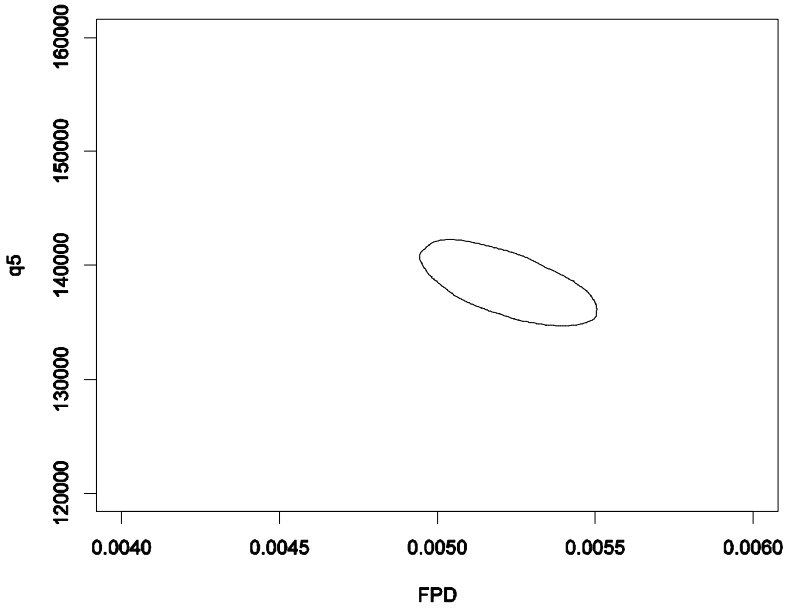
Bayesian calibration results

- Samples from $f(\theta | \boldsymbol{d})$
- Posterior statistics (best estimate=mean)
- Marginal histograms
- Joint confidence regions (kernel density estimates)



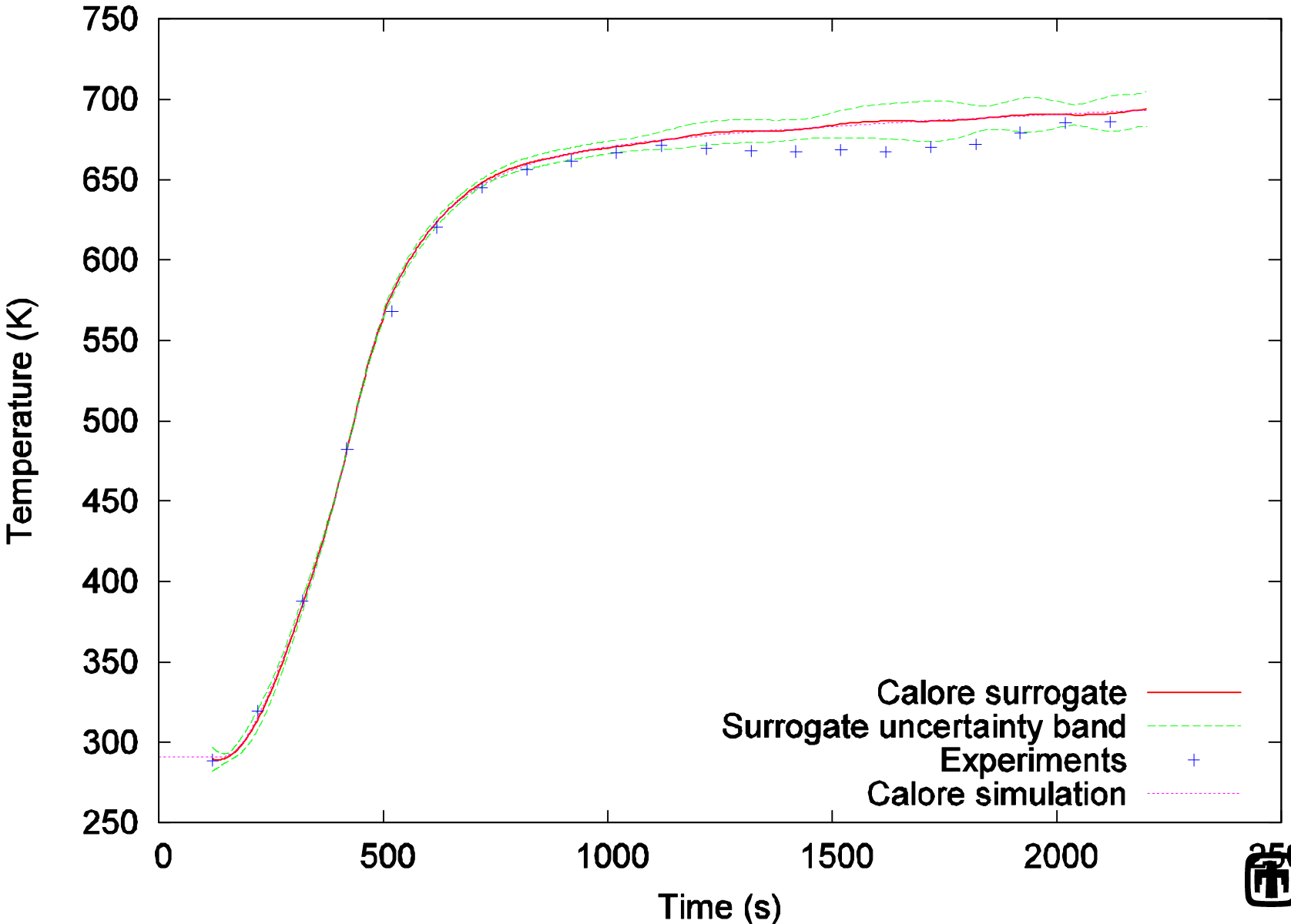
Bayesian calibration results





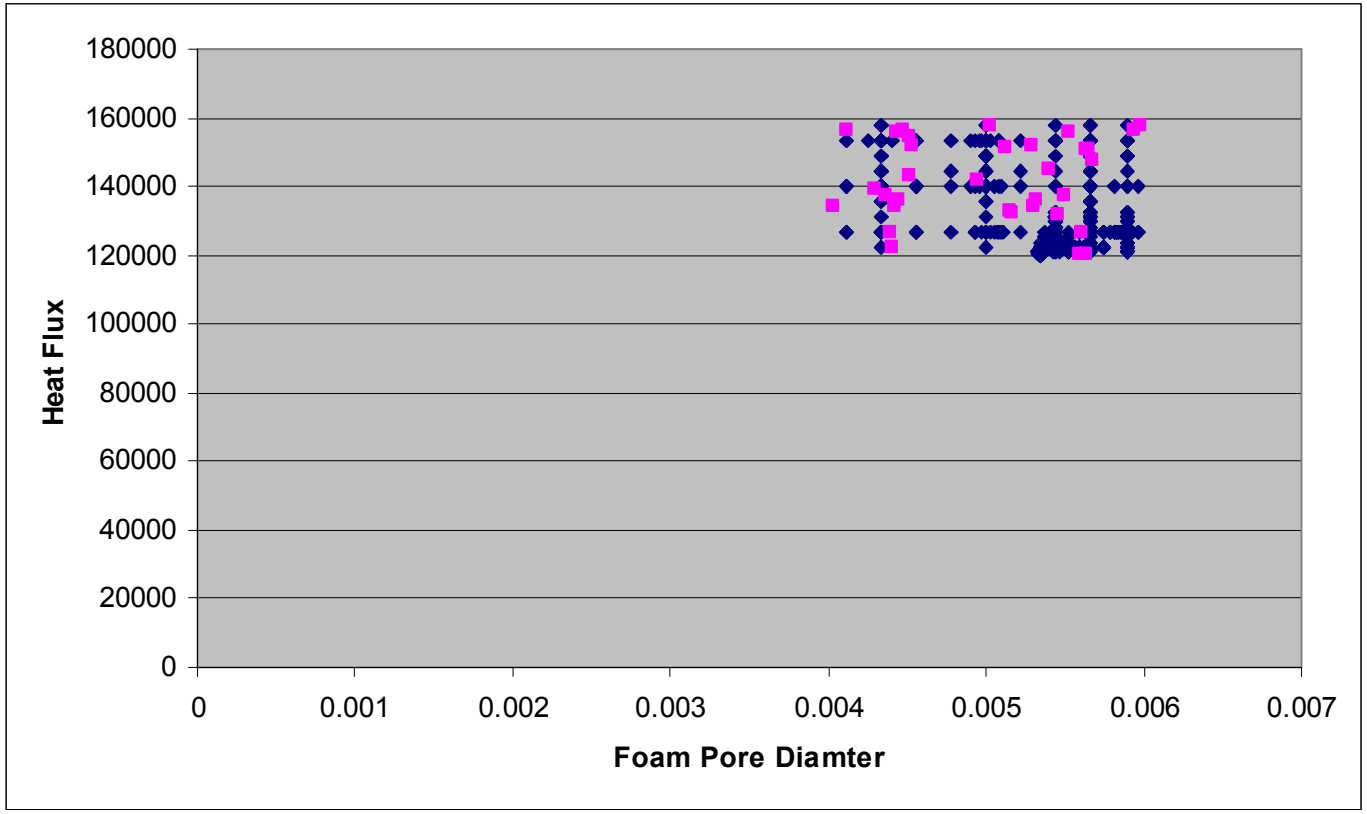
Method	RMS	Function Evaluations
Bayesian mean	19.4	50
DIRECT	32.3	65
Nominal	43.4	

Calibrated Predictions vs. Observations



Joint Confidence Region Estimation

- **Show results (NOTE: These are very preliminary, will be changed. The nature of the picture will be the same, but the points may be located in different places**



Summary

- Computational issues with each
  - Bayesian calibration
    - GP problems: MLE estimation of the parameters, ill-conditioning of the covariance matrix, stationary variance assumption
    - MCMC: testing for convergence done in an ad-hoc fashion
  - Nonlinear regression methods
    - Model assumptions
    - Linear approximations often require Jacobian or Hessian estimates, not always accurate
    - SSE contour estimation expensive
  - Both methods require surrogates, use optimization within the overall strategy (e.g., MLE estimation)
  - Both methods account for uncertainty in the parameters based on some idea of likelihood of the parameters in certain regions of the space given the data