SAND2015-1854C

Exceptional service in the national interest

Sandia National Laboratories

Use of Parallel MCMC Methods with the Community Land Model

Jaideep Ray, Laura Swiler, Maoyi Huang, Jason Hou

SIAM Computational Science and Engineering Meeting
March 14-18, 2015. Salt Lake City.

U.S. DEPARTMENT OF ENERGY NNSA National Nuclear Security Administration

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000. SAND NO. 2011-XXXXP

Overview

- MCMC Methods, DRAM
- Community Land Model
- Results and Implementation
- Next Steps

Project Goal: Given observational data, and the CLM model, invert for parameters of CLM using a Bayesian formulation

Bayesian Formulation

- Generate posterior distributions on model parameters, given
  - Experimental data
  - A prior distribution on model parameters
  - A presumed probabilistic relationship between experimental data and model output that can be defined by a likelihood function

$$f(\theta \mid d) \propto \pi(\theta) L(d \mid \theta)$$

Model parameters

Observed Data

Likelihood function which Incorporates the model

Prior parameter distribution

Bayesian Formulation

- Experimental data = Model output + error

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

- 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 Markov Chain Monte Carlo (MCMC)
  - 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.

Markov Chain Monte Carlo

- Metropolis-Hastings is a commonly used algorithm
- It has the idea of a "proposal density" which is used for generating X_{i+1} in the sequence, conditional on $X_{i.}$

  Sample a candidate Y from the proposal density function $q_Y(Y|X_i)$

  Calculate the acceptance ratio $\alpha(X,Y) = \mathbf{min}\left[1, \frac{f_X(Y)q_Y(Y|X_i)}{f_X(X)q_X(X_i|Y)}\right]$

  If $\alpha(X_i,Y) \geq U$, set $X_{i+1}=Y$, else set $X_{i+1}=X_{i.}$

  Increment 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
  - How do you tune it for an optimal acceptance rate, etc.?

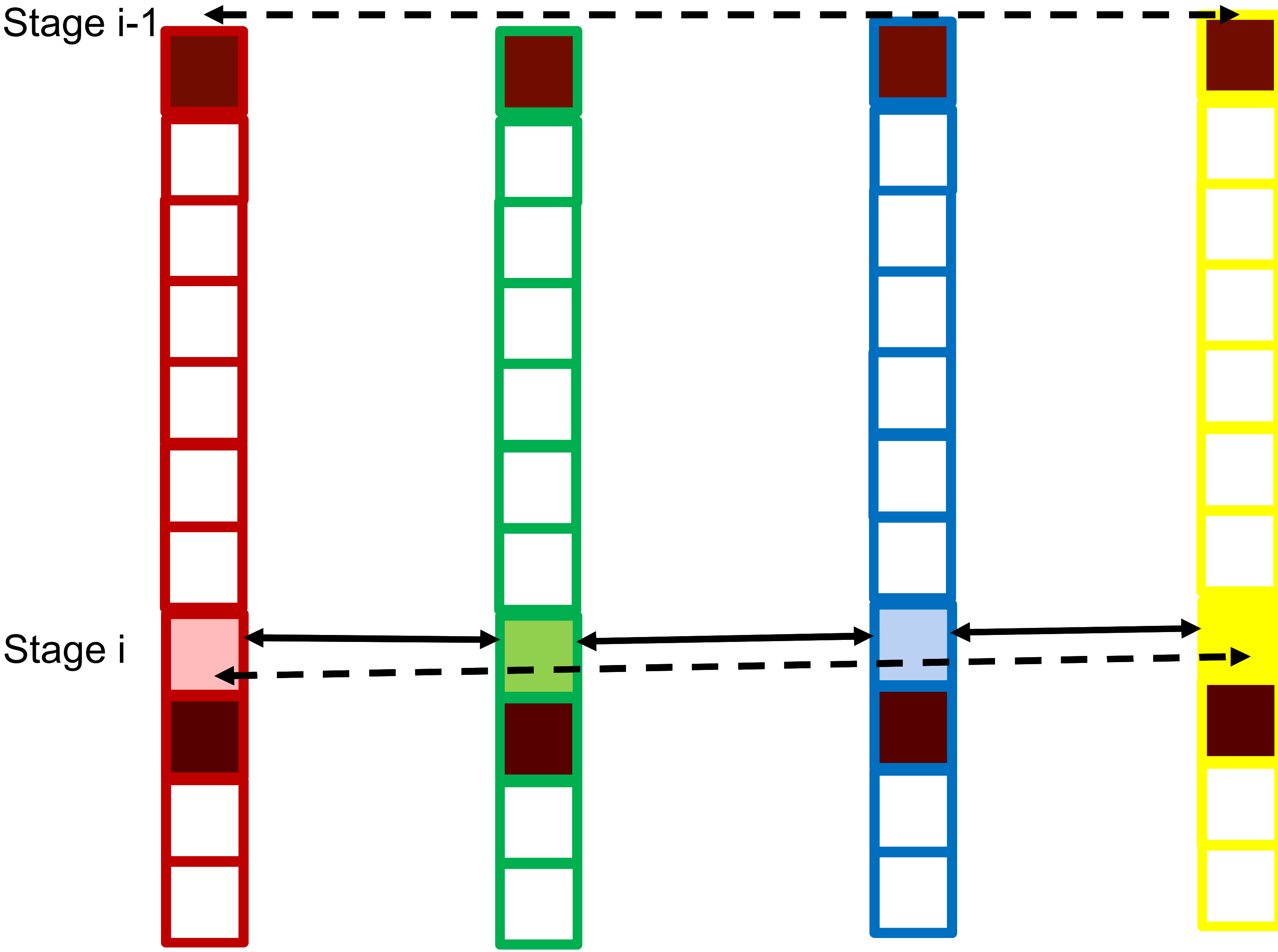
Markov Chain Monte Carlo

- MCMC depends on asymptotic behavior of the chain. Ideally, you want to run for 100,000+ samples. COMPUTATIONALLY VERY EXPENSIVE!
  - Typically, a limited number of model runs are used to generate a surrogate model and the MCMC sampling is performed on the surrogate
  - We want to avoid surrogates
- Limitation of MCMC: it is inherently sequential.
- We want to exploit some parallelism by using multiple chains

SOLUTION: PARALLEL DRAM on the actual CLM model

- DRAM: Delayed Rejection Adaptive Metropolis
- MCMC algorithm with two features:
  - Delayed Rejection: don't reject right away…another chance
  - Adaptive Metropolis: Update the proposal covariance periodically based on the accepted samples from the chain

Parallel DRAM



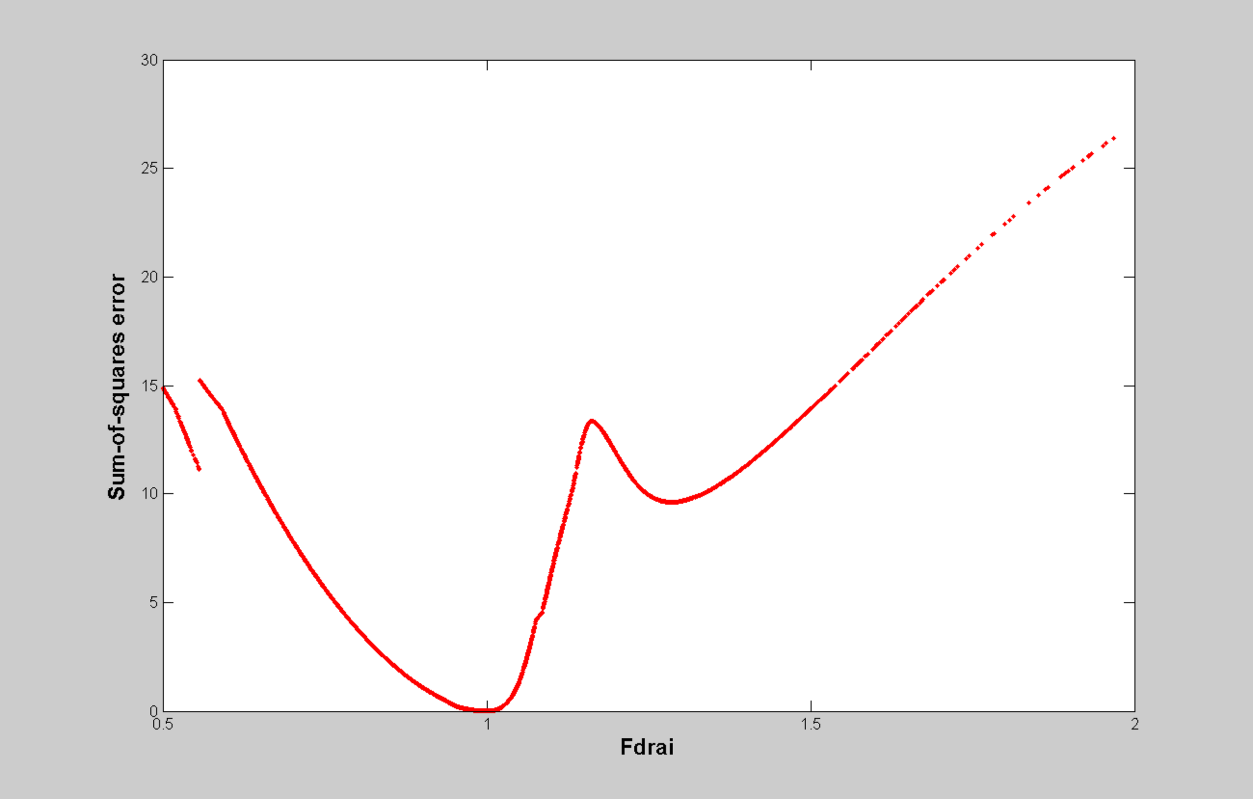
Full formula given in "Solonen et al. "Efficient MCMC for Climate Model Parameter Estimation: Parallel Adaptive Chains and Early Rejection. Bayesian Analysis (2012) 7(2), pp. 1-22.

$$\Sigma_i = \frac{(i-1)}{i}\Sigma_{i-1} + \frac{1}{n}[\theta_n - \bar{\theta}]$$

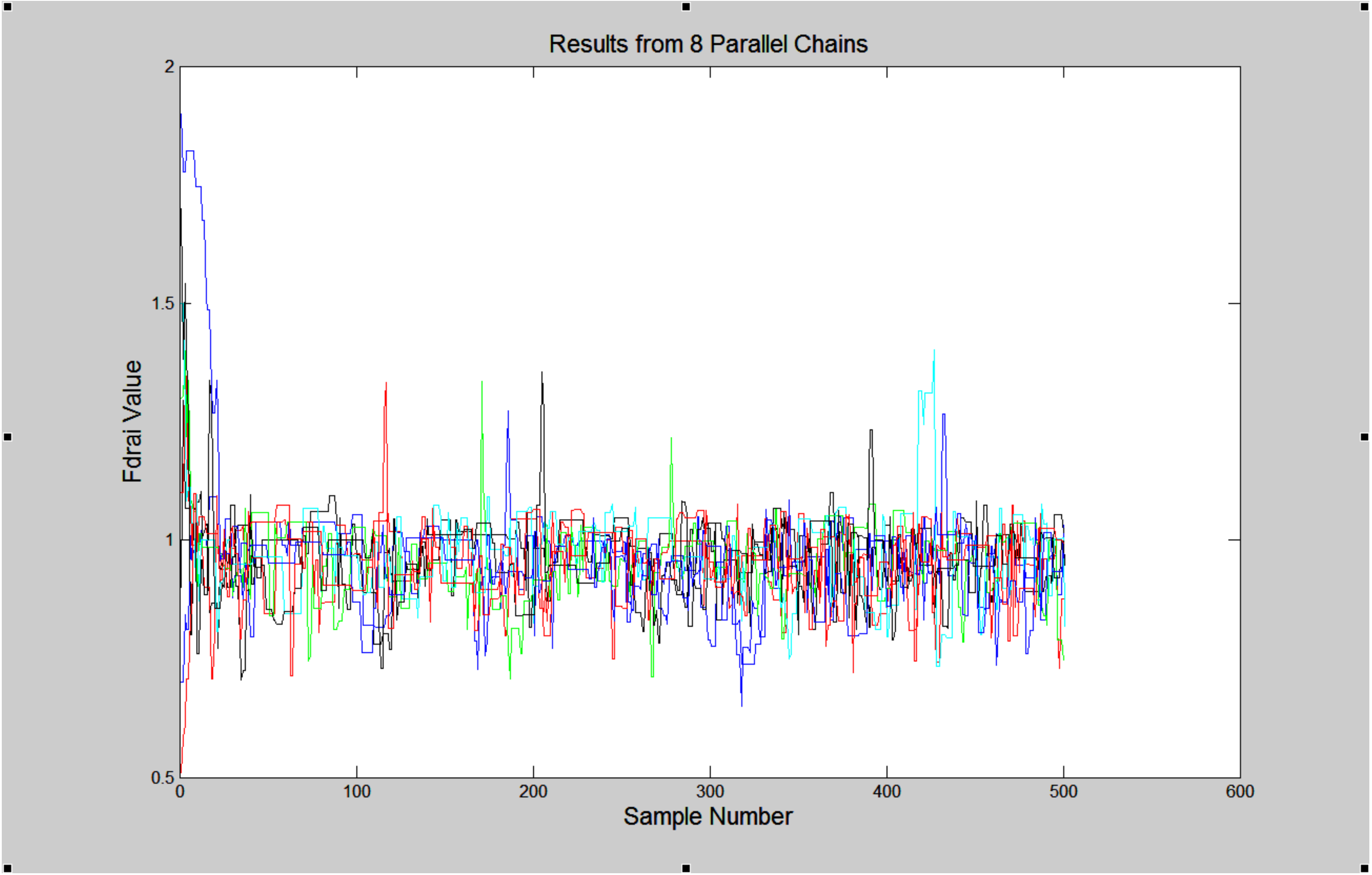
CLM Model with simulated observations

Sandia National Laboratories

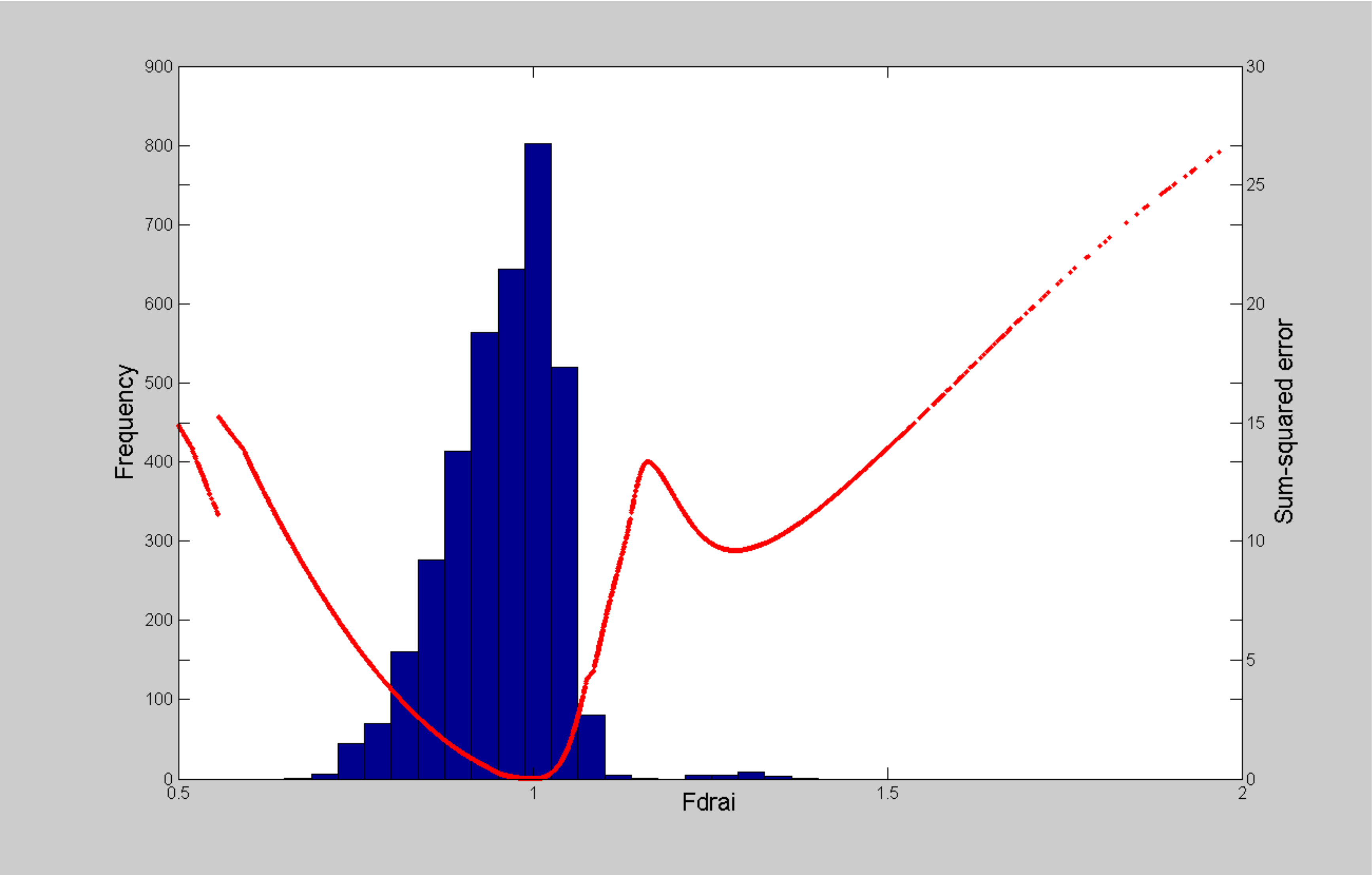
- Varying Fdrai from 0.5 to 2.0
- Simulated observations at Fdrai = 1.0
- Likelihood involves differences of Latent Heat over 12 months
- Double-humped and discontinuous likelihood function can be a challenge



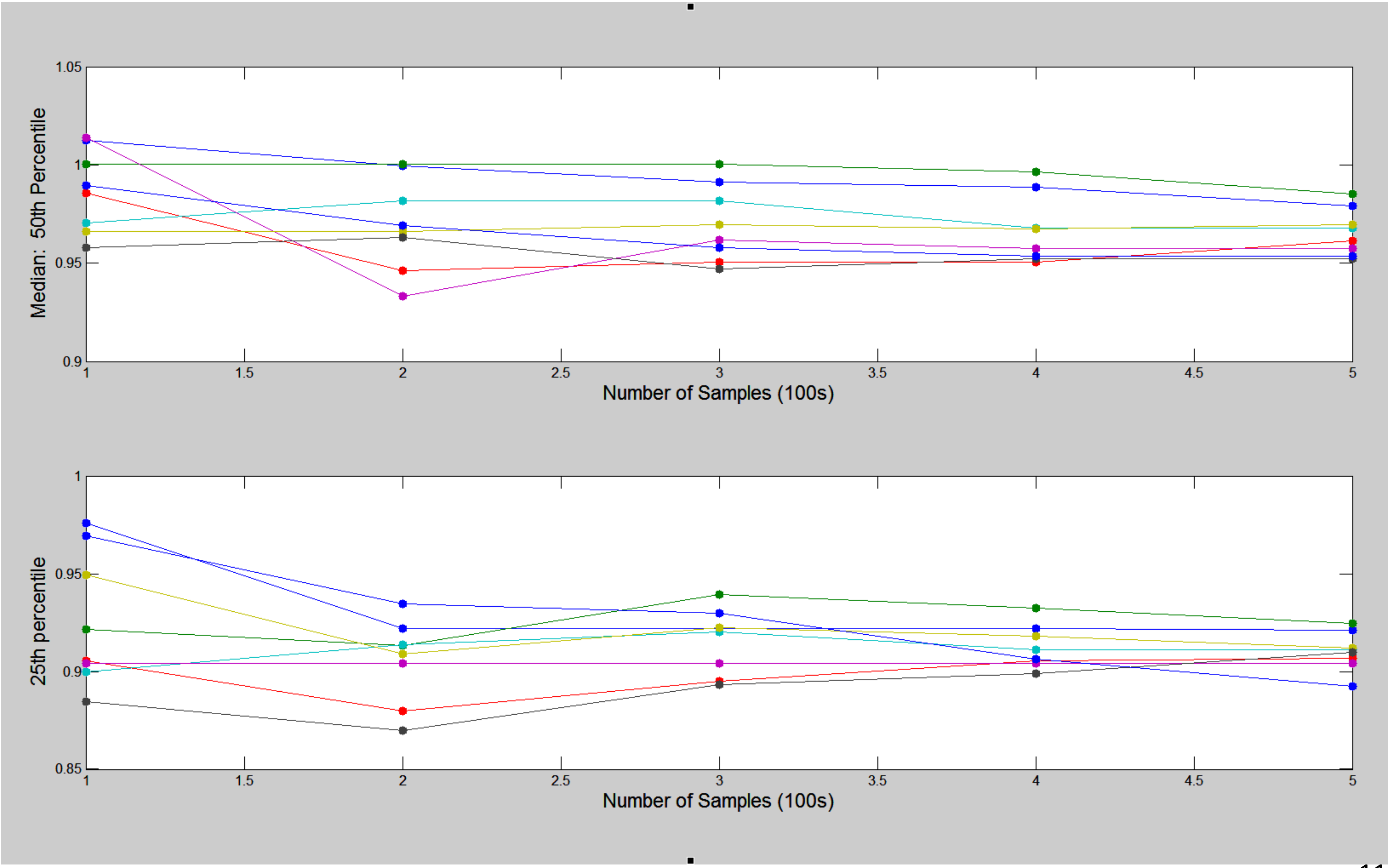
CLM Model: 8 chain MCMC



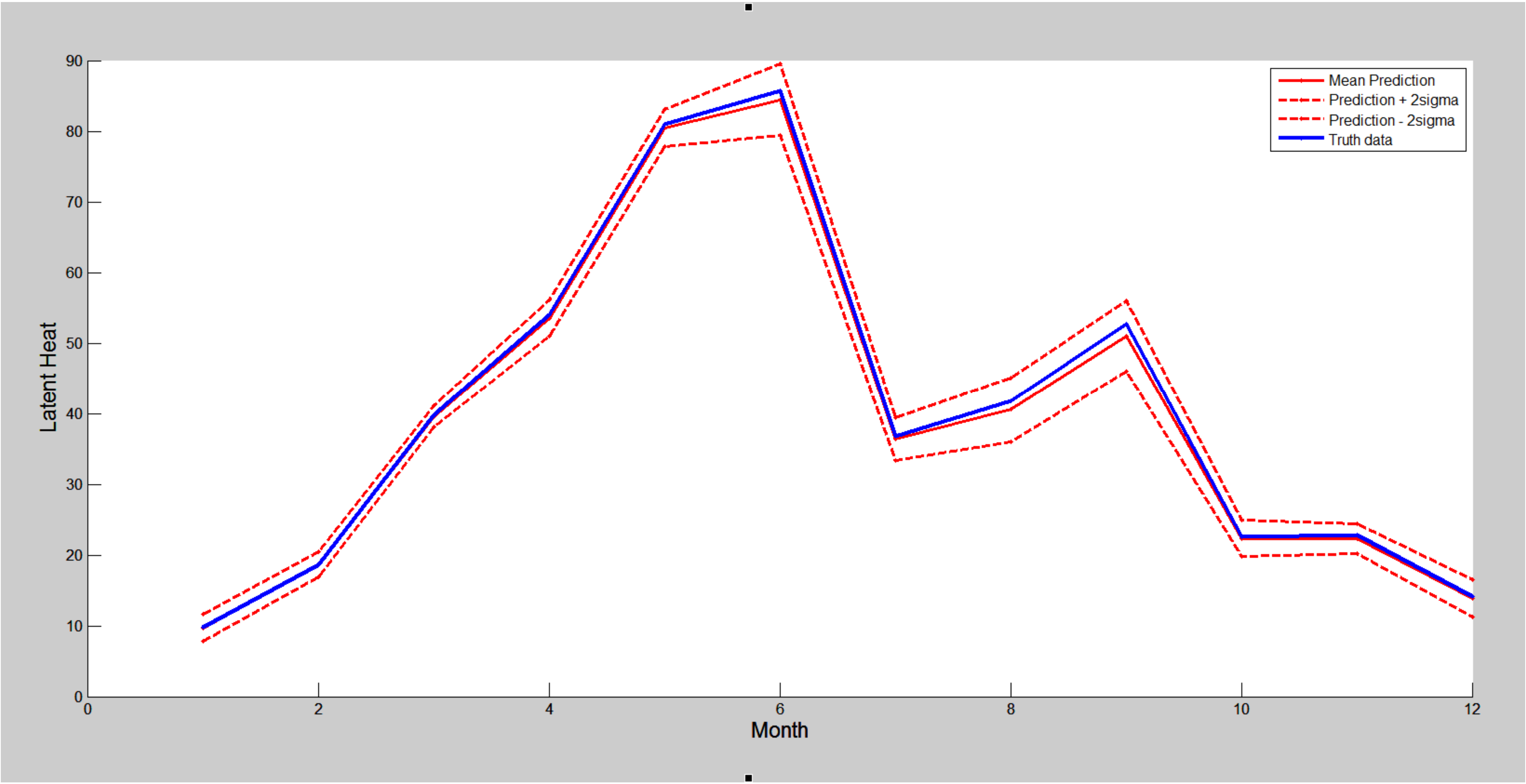
CLM Model: Posterior histogram



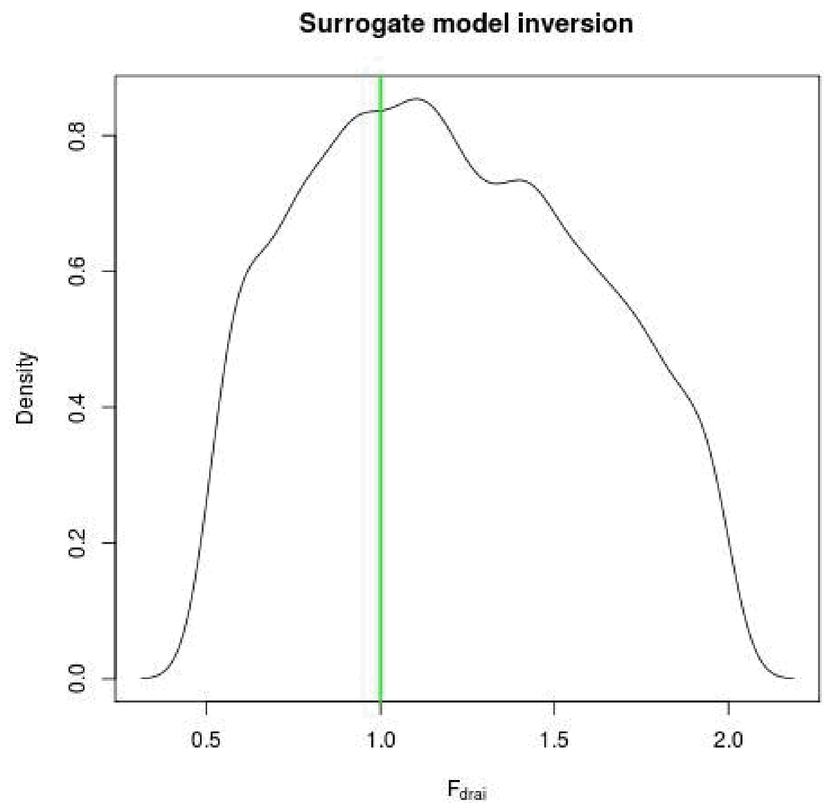
Convergence of percentiles from posterior



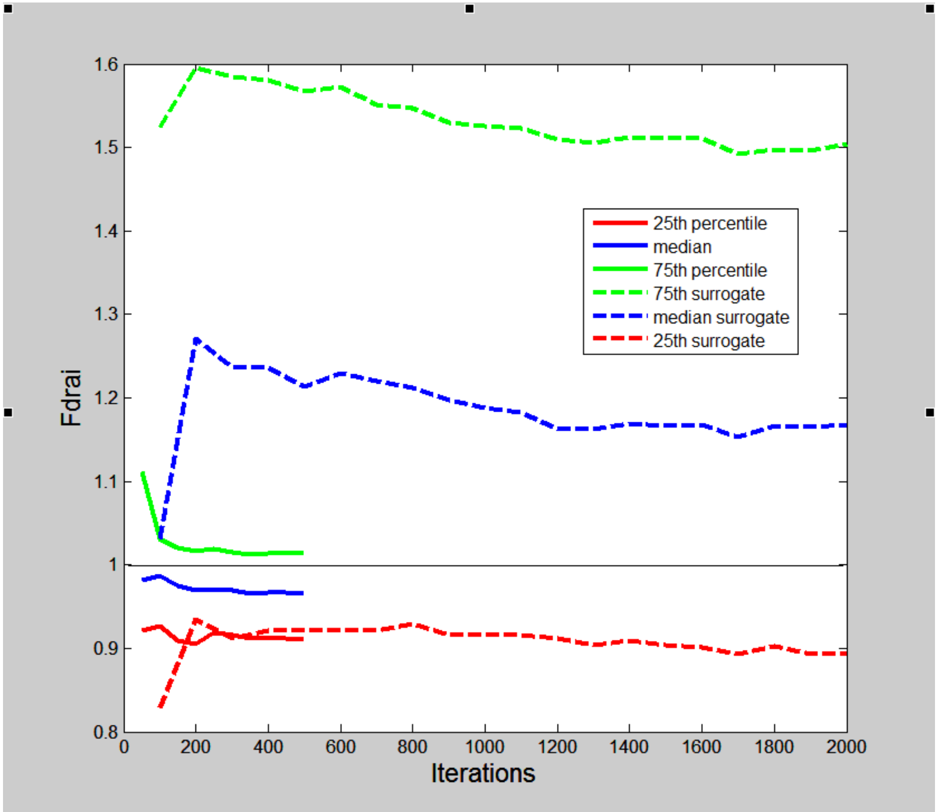
Pushed-forward Posterior



Results w/ surrogate – synthetic data



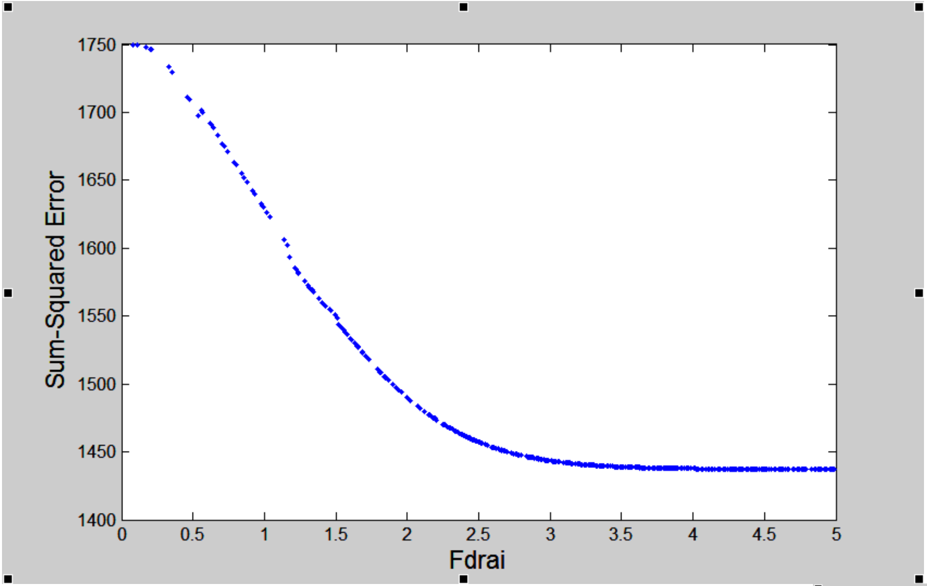
PDF of F_{drai} estimated with surrogates



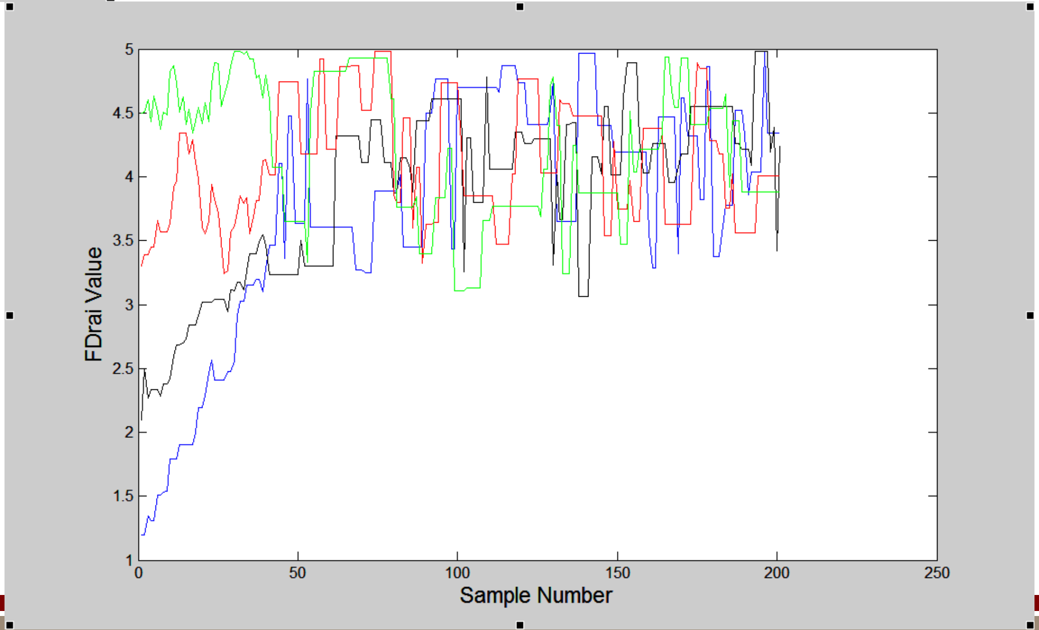
Convergence of quantiles of F_{drai}

- Surrogate infers the right F_{drai} value (see MAP values) but ..
- Convergence is slow – take 4x longer

CLM Model with Actual 2003 data



Difficulty with flat likelihood function and parameter insensitivity over a large region: convergence hard to assess.



Conclusions

- Bayesian calibration even with 1-parameter is non-trivial with a multi-modal likelihood function
- Differences between the actual and surrogate CLM are important: in many cases, surrogates will not be sufficient, could take longer to converge, and could converge to incorrect values
- Parallelism necessary for running MCMC on expensive simulations with no surrogate
- We need to run larger scaling studies
- Next steps: DREAM and DRAM integration. We will “preconditon” the proposal covariance by running DREAM for some number of samples, using the individual chains to generate a high-quality proposal covariance for DRAM.