

Exceptional service in the national interest



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.

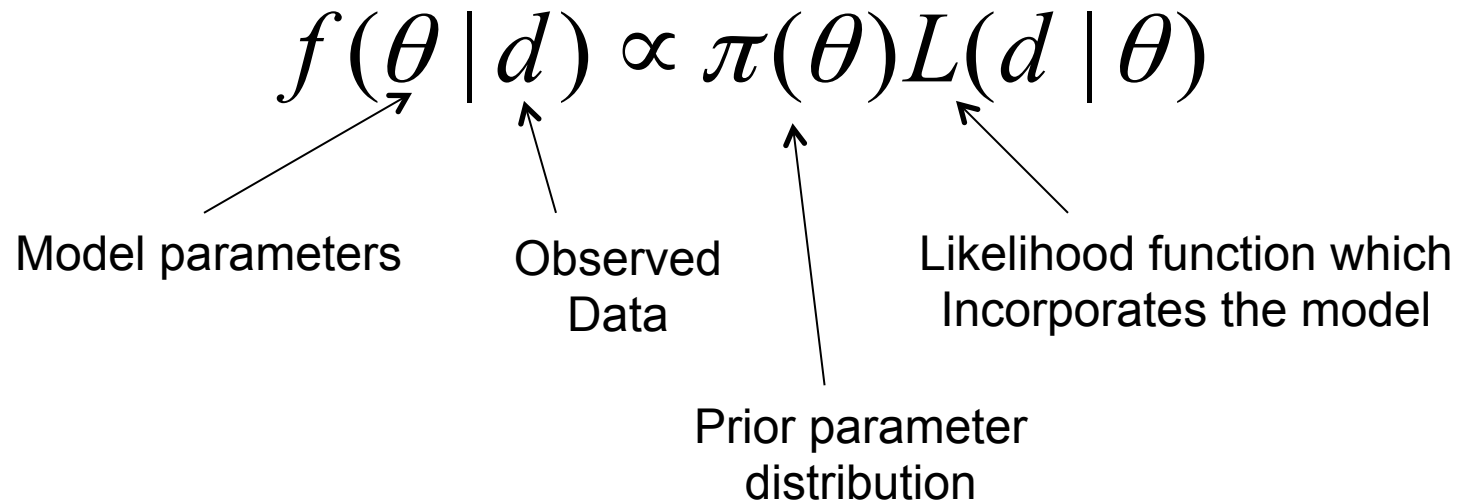
Overview

- 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

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


Model parameters

Observed Data

Prior parameter distribution

Likelihood function which Incorporates the model

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

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

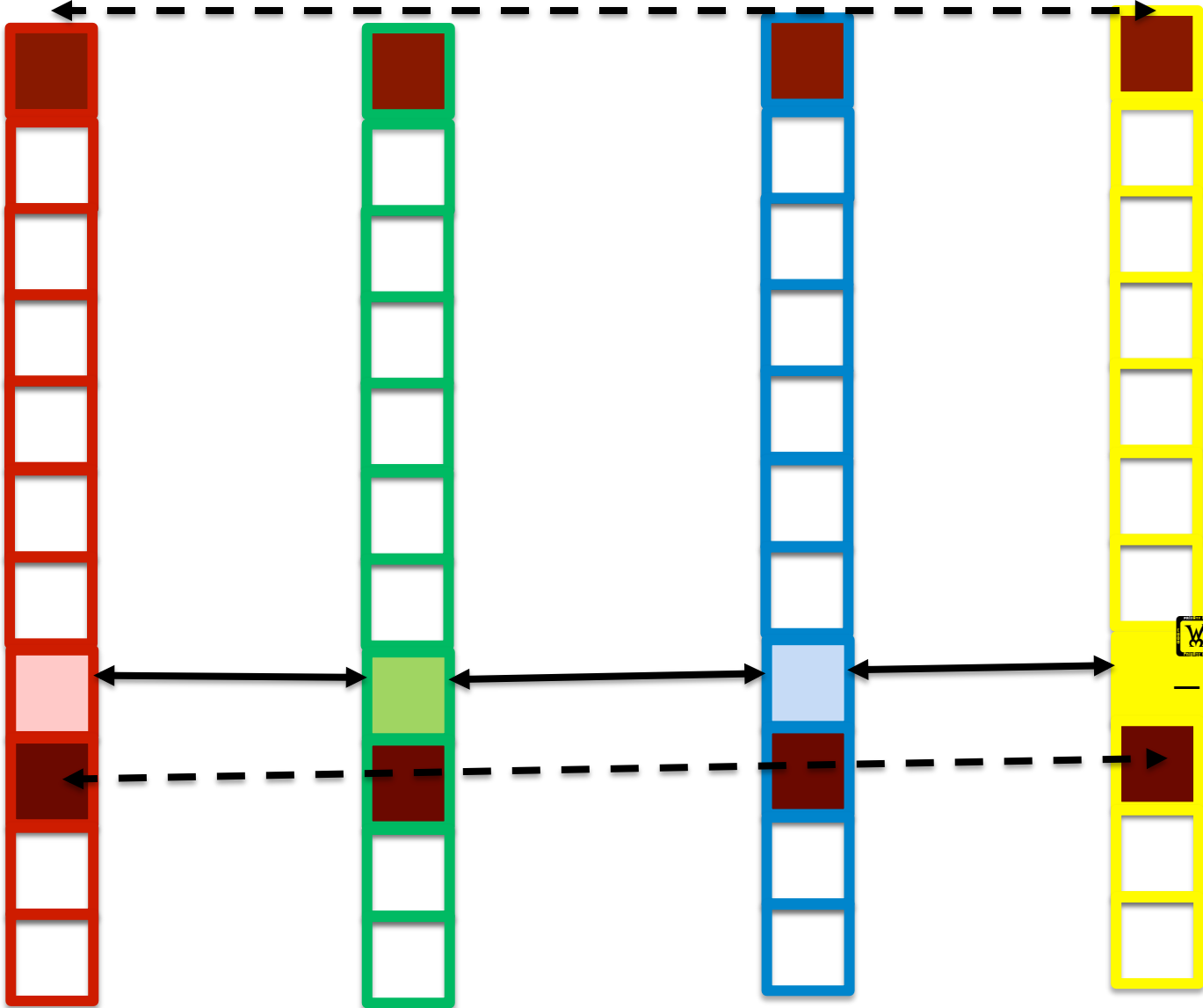
- MCMC depends on asymptotic behavior of the chain. Ideally, you want to run for 100,000+ samples. **COMPUTATIONALLY VERY EXPENSIVE!**
 - Typically, a limited number of model runs are used to generate a surrogate model and the MCMC sampling is performed on the surrogate
 - We want to avoid surrogates
- 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

Stage i-1



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.

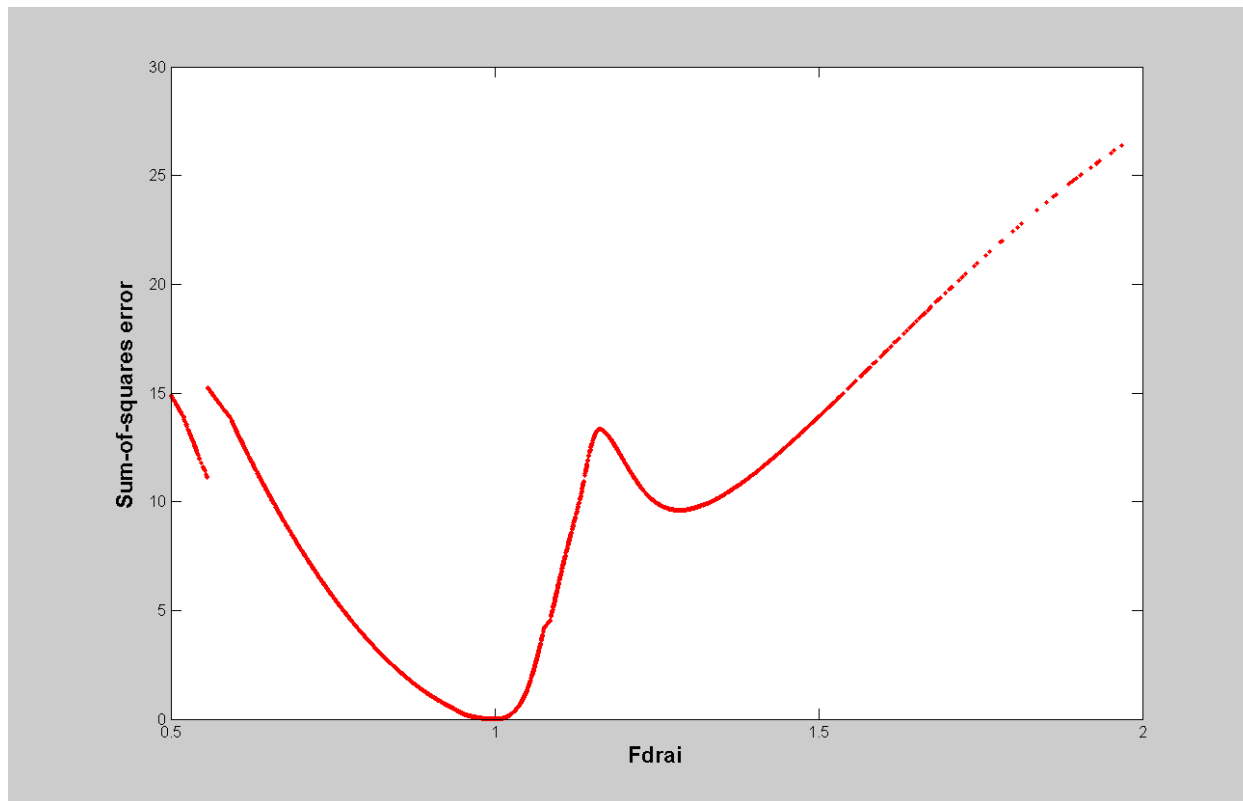
$$\left[\mathbb{W} \downarrow i = (i-1)/i \mathbb{W} \right]$$

$$-1 + 1/n [\theta \downarrow n - \theta]$$

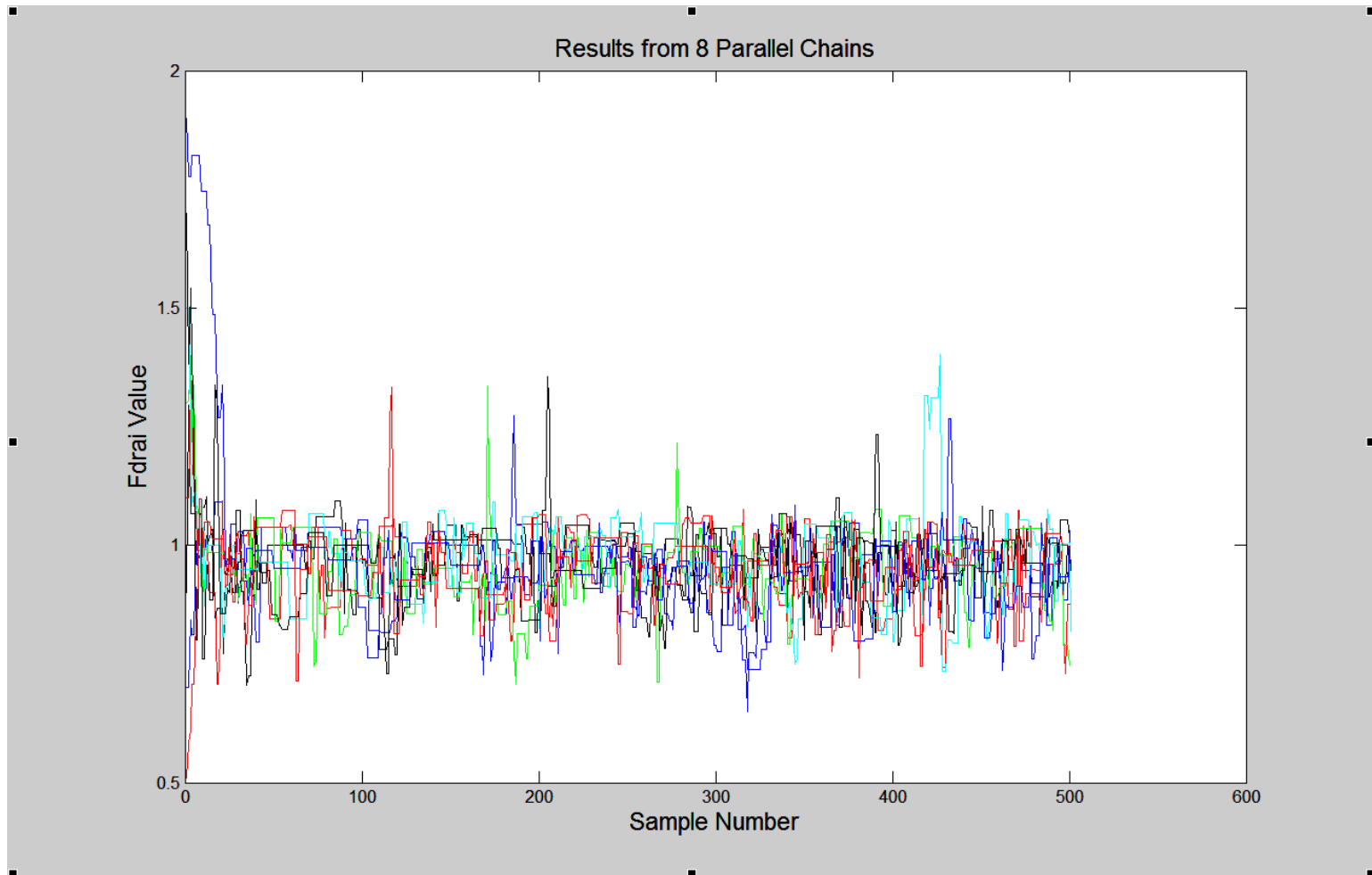
Stage i

CLM Model with simulated observations Sandia National Laboratories

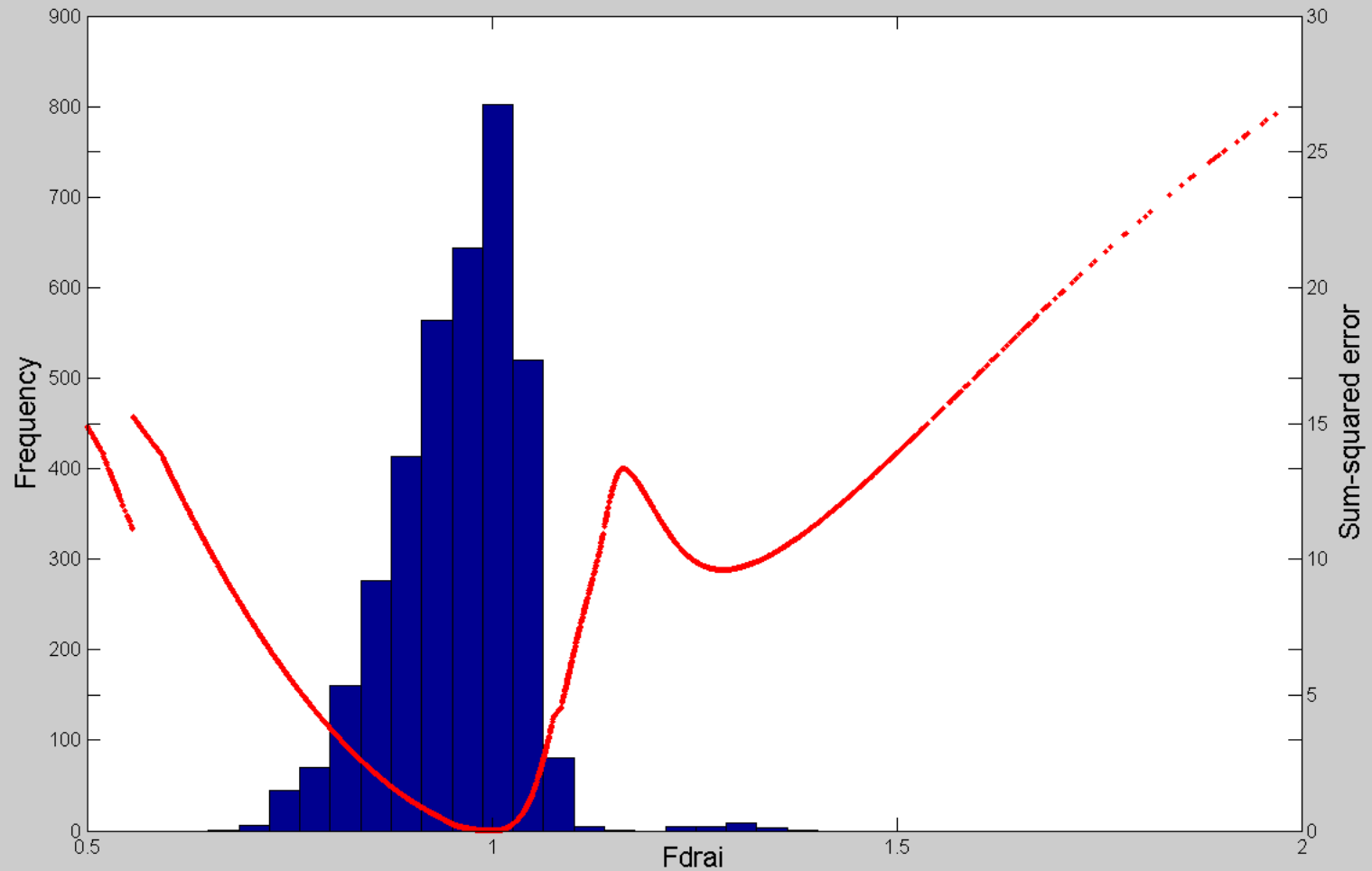
- Varying F_{drai} from 0.5 to 2.0
- Simulated observations at $F_{\text{drai}} = 1.0$
- Likelihood involves differences of Latent Heat over 12 months
- Double-humped and discontinuous likelihood function can be a challenge



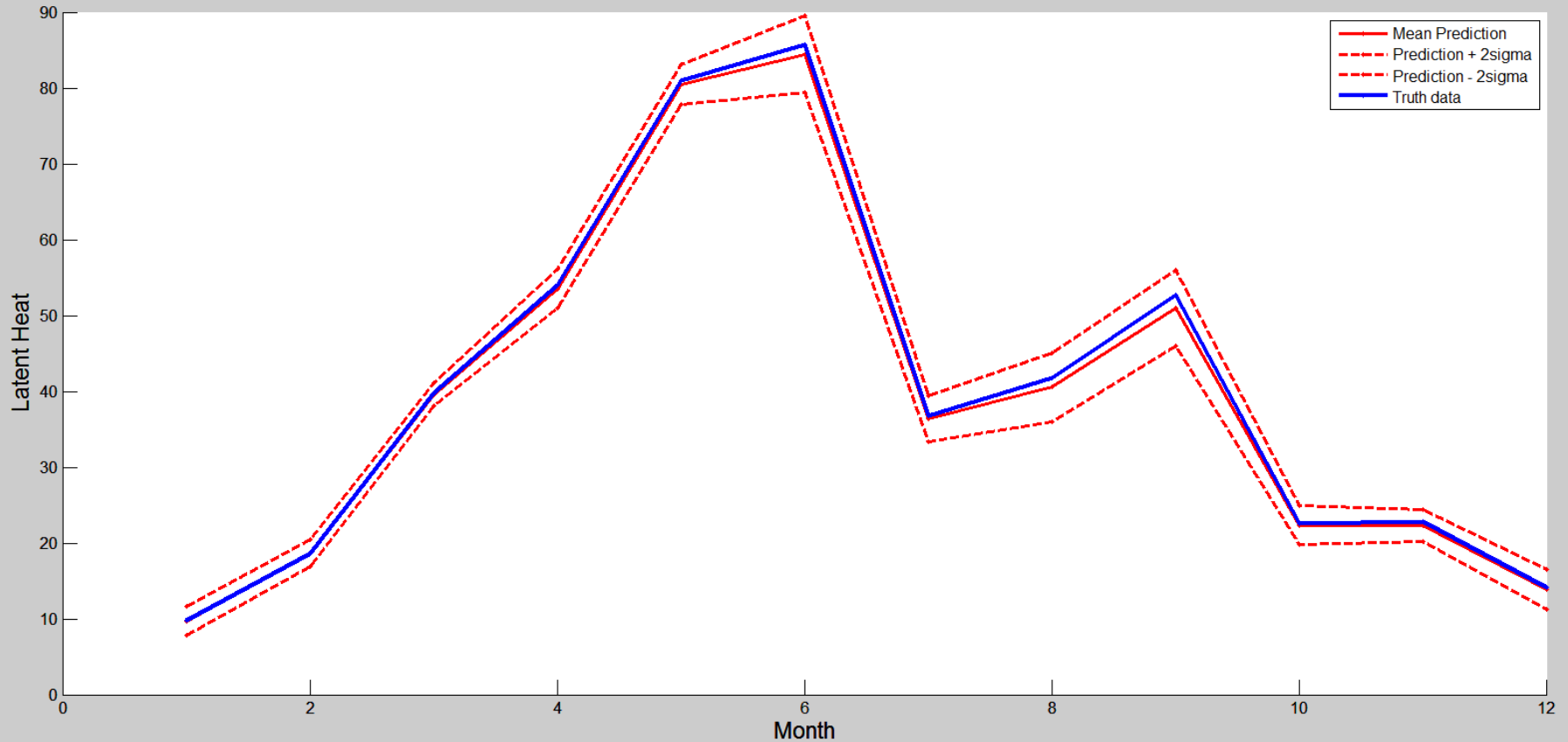
CLM Model: 8 chain MCMC



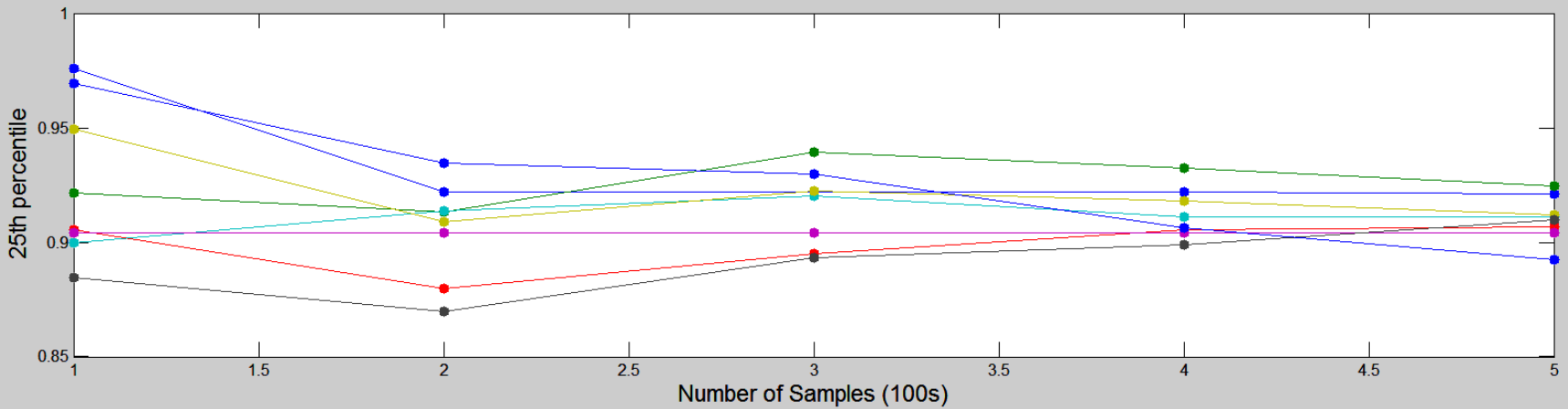
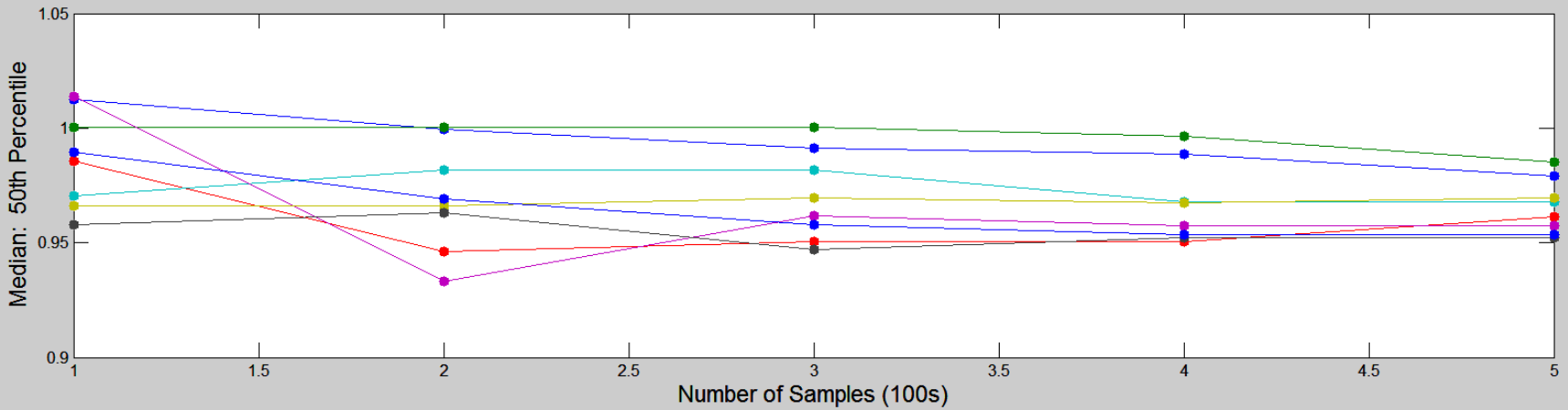
CLM Model: Posterior histogram



Pushed-forward Posterior

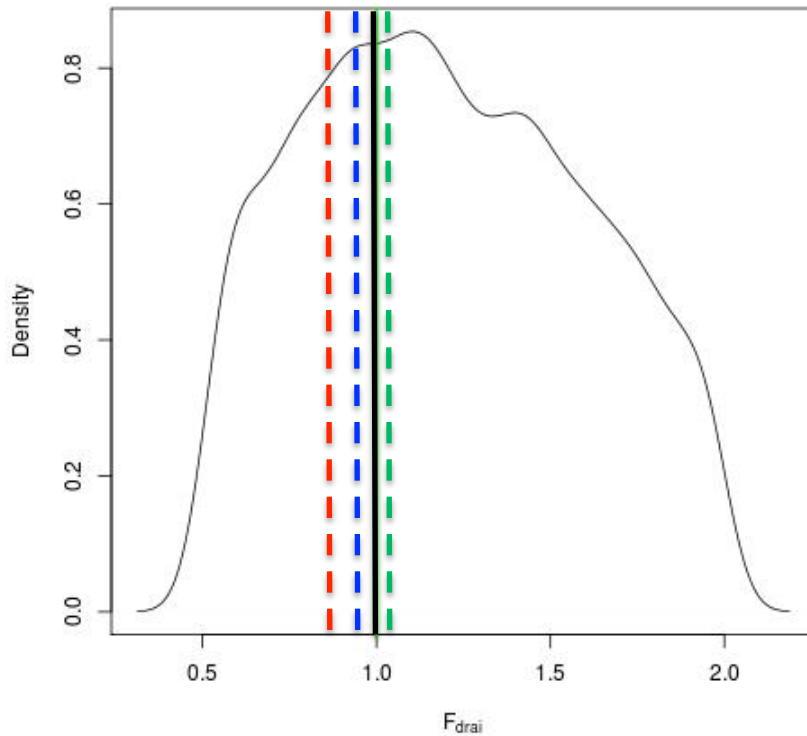


Convergence of percentiles from posterior

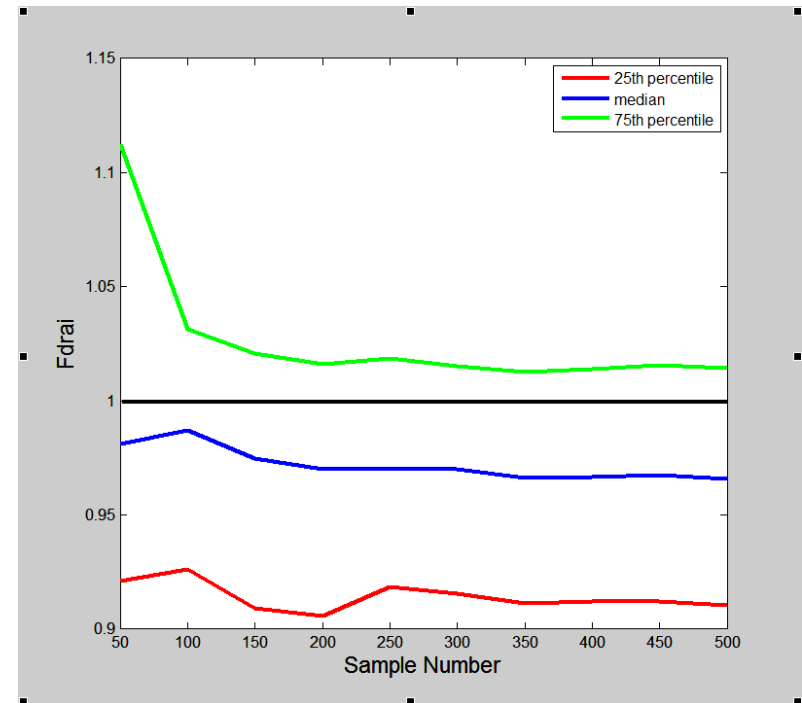


Results with a surrogate

Surrogate model inversion



PDF of F_{drai} estimated with surrogates

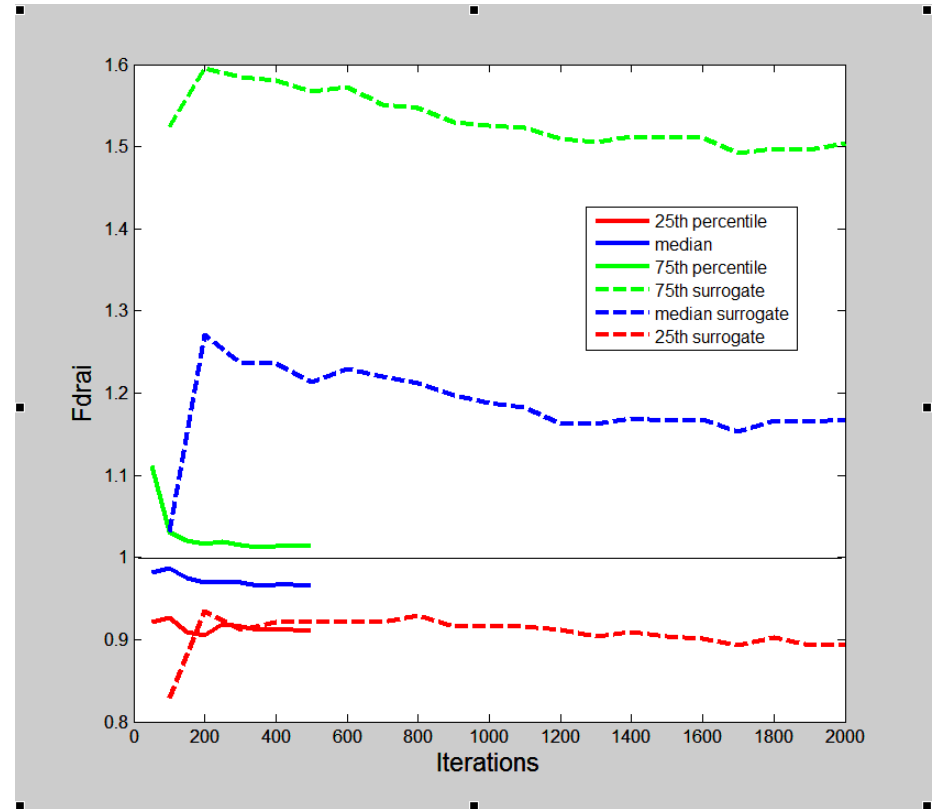


Convergence of quantiles of F_{drai} for CLM

- Gelman-Rubin statistic on chains from CLM: 1.08, indicates convergence
- While quartiles from CLM look converged, this does not imply the entire PDF has converged
- The surrogate-generated posterior PDF differs significantly from the CLM one

Results with a surrogate

- Both CLM4 and surrogate's estimates of the quartiles have converged
- The behavior of the surrogates are quite different from CLM4 in the region $0.75 < F_{\text{drai}} < 1.25$, which leads to a broader PDF
- Since this is a synthetic data inversion with identical σ , there is no reason that the surrogate should not give the same tight quartiles that CLM4 does
- In this case, the LH surrogate is different (and wrong) from CLM → gives a well-converged but incorrect posterior PDF.

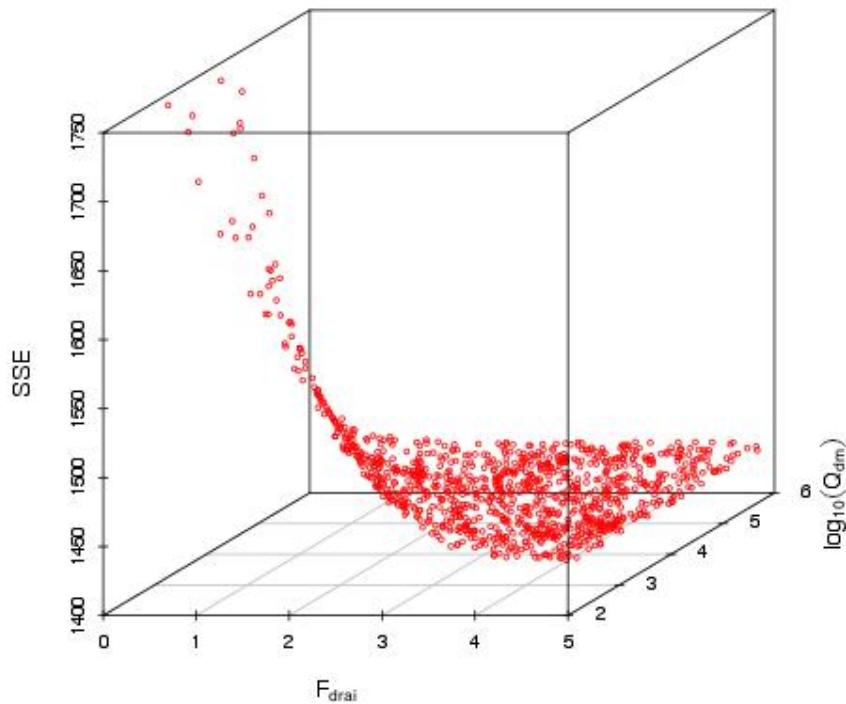


Convergence of quantiles of F_{drai} for CLM and surrogate

**SURROGATE PDF CAN APPEAR CONVERGED
BUT IS ONLY APPROXIMATE!**

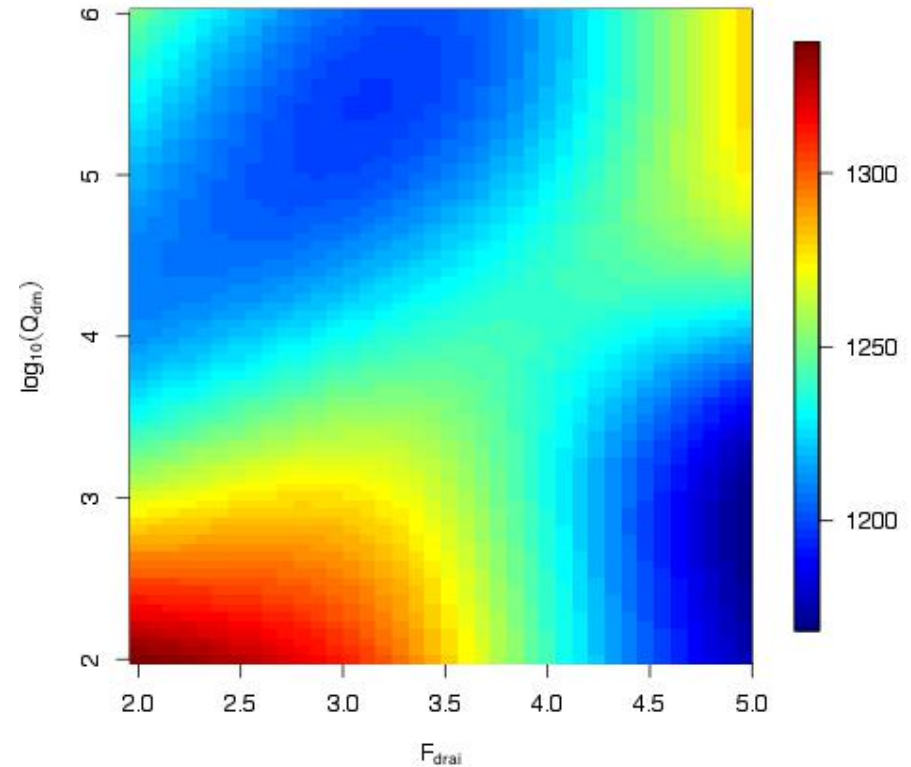
CLM Model with actual 2003 data

SSE for US-ARM, using real data from 2003



Scatter plot

SSE for US-ARM

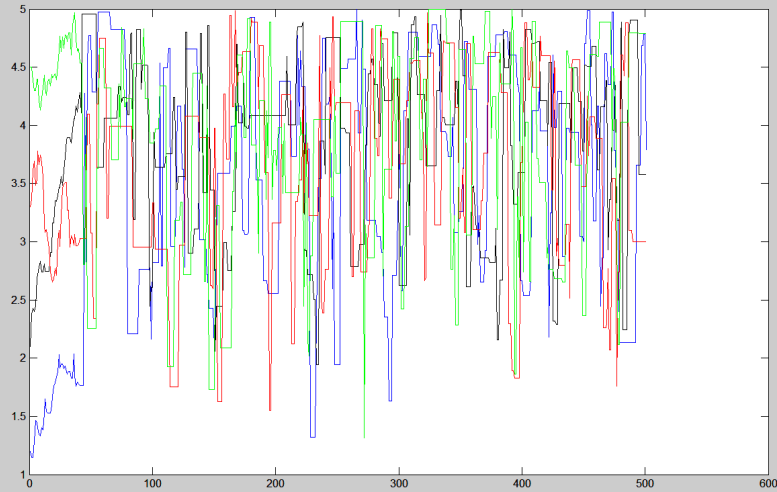


Flood plot

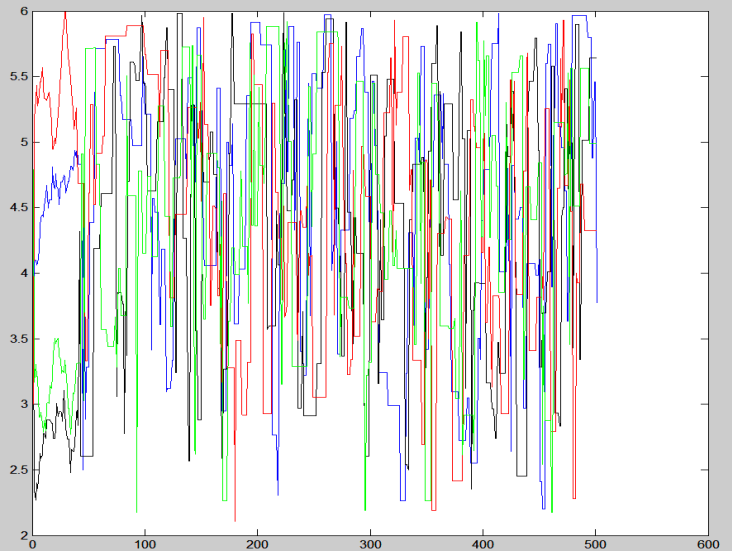
- Difficulty with flat likelihood function for F_{drai} above 3, and parameter insensitivity over a large region for Q_{dm} . ($\log_{10}(Q_{\text{dm}}) > 3$)
- Convergence is difficult to assess.

CLM Model with actual 2003 data

Posterior of F_{drai}



Posterior of $\log_{10}(Q_{\text{dm}})$

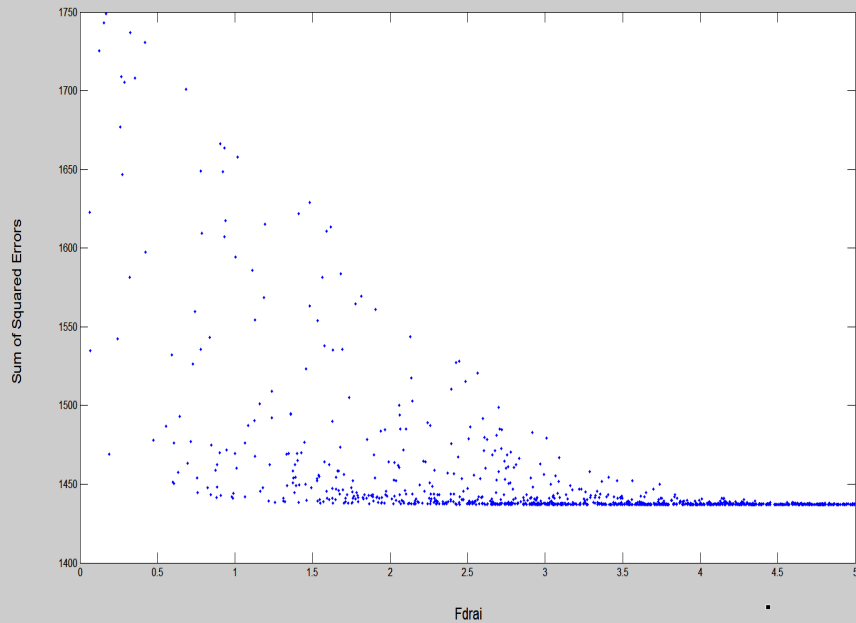


Conclusions

- 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 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 “precondition” the proposal covariance by running DREAM for some number of samples, using the individual chains to generate a high-quality proposal covariance for DRAM.

BACKGROUND

CLM Model with Actual 2003 data



Difficulty with flat likelihood function for F_{drai} above a value of 2.5, and parameter insensitivity over a large region for Q_{dm} .

Convergence is difficult to assess.

