Selecting a Representative Sample

Collaborators: Matt Taddy, Genetha Gray

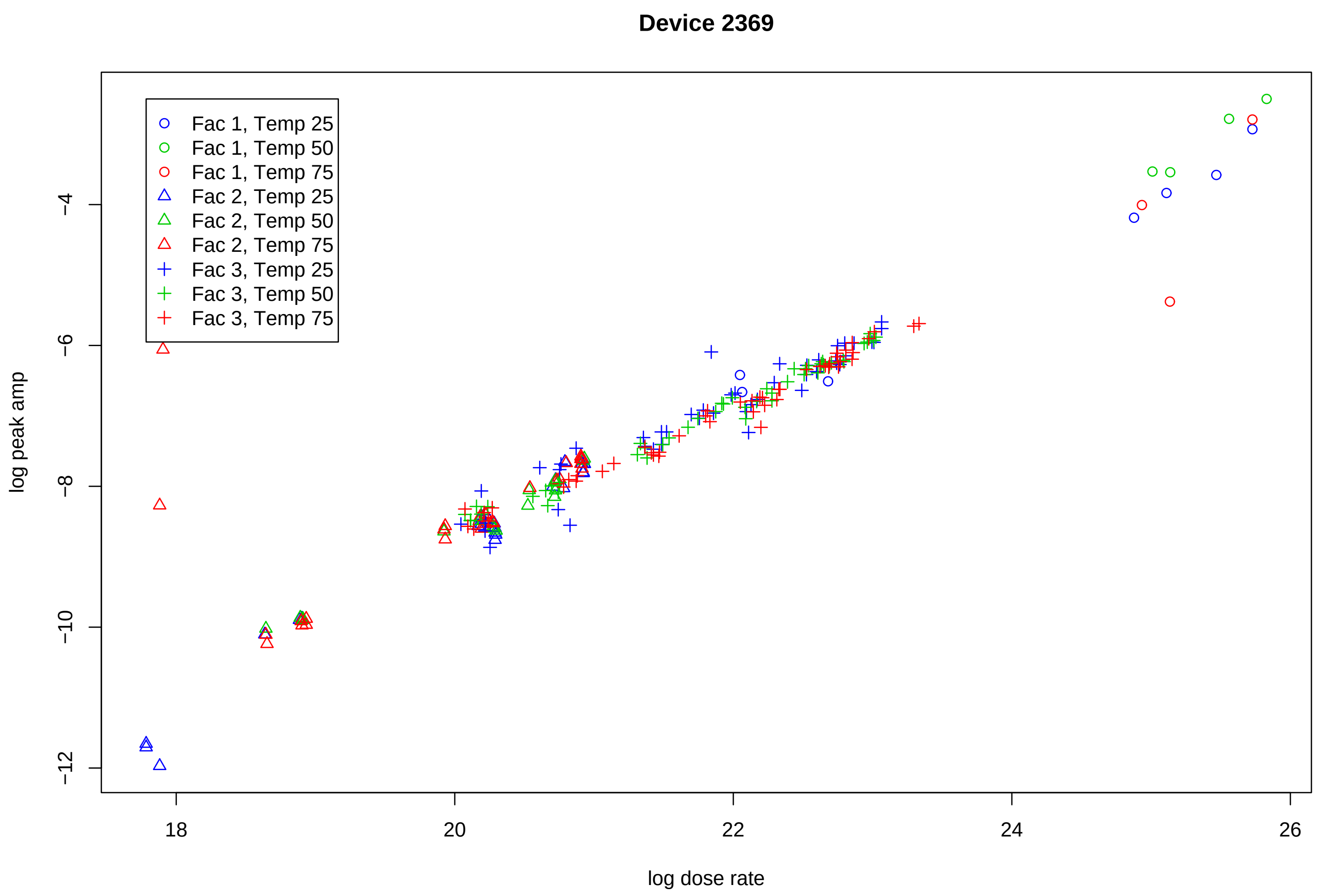
Circuit Design

- Overall goal: optimize circuit design — use data from a physical experiment to calibrate a computer simulation of the response of circuits to an electromagnetic pulse, optimize via simulation
- Experimental design in collecting the physical data
- Statistical choice of training samples for calibration
- Statistical emulation for guiding optimization

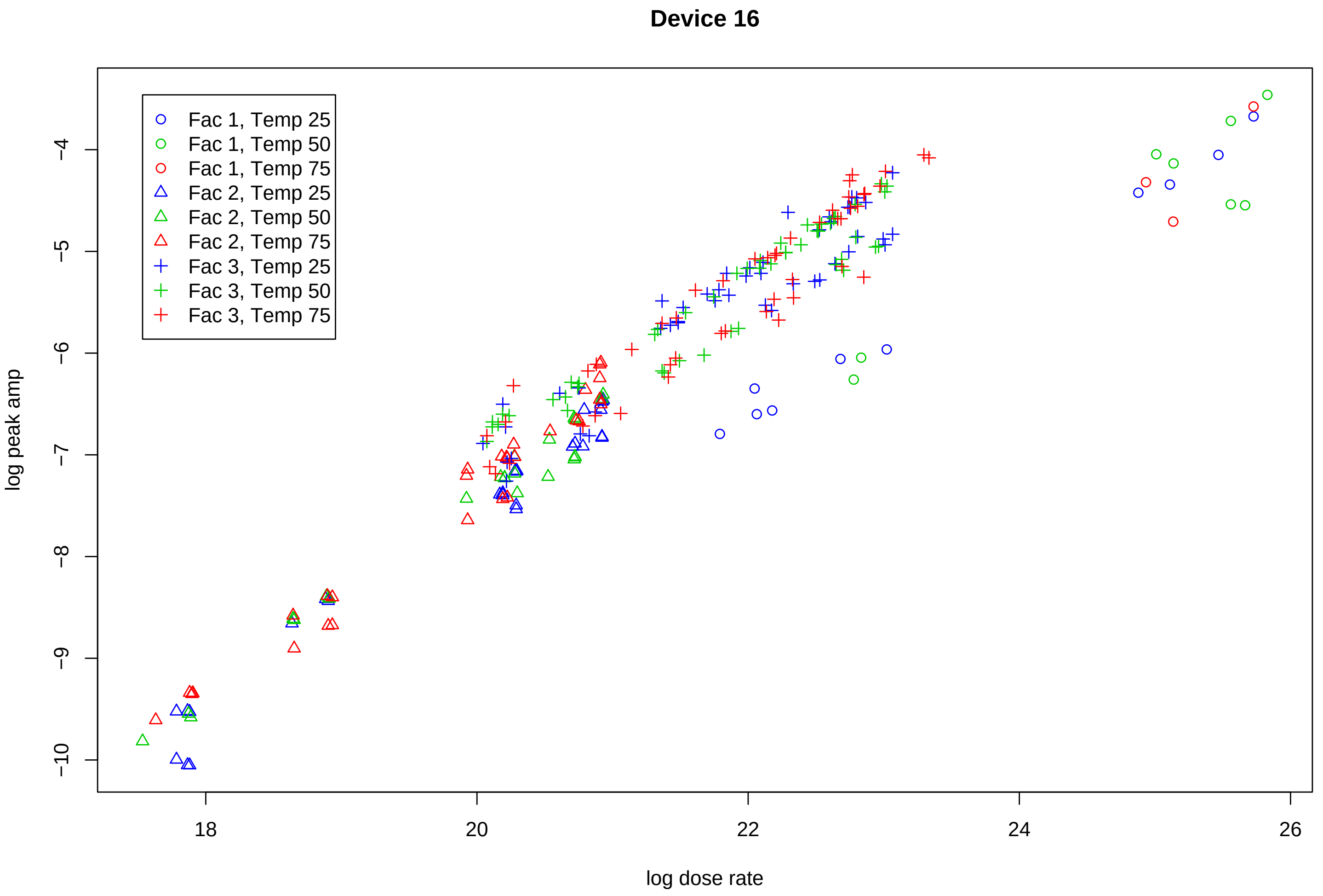
Selecting a Representative Sample

- After the physical data have been collected and cleaned, our collaborators want to choose six points from each experimental setting to use for calibrating their computer simulator
- Standard decision-theoretic approach is not possible because the computer simulator is unavailable to us, so the true loss function is unknown to us
- Data show either linear and/or non-functional forms

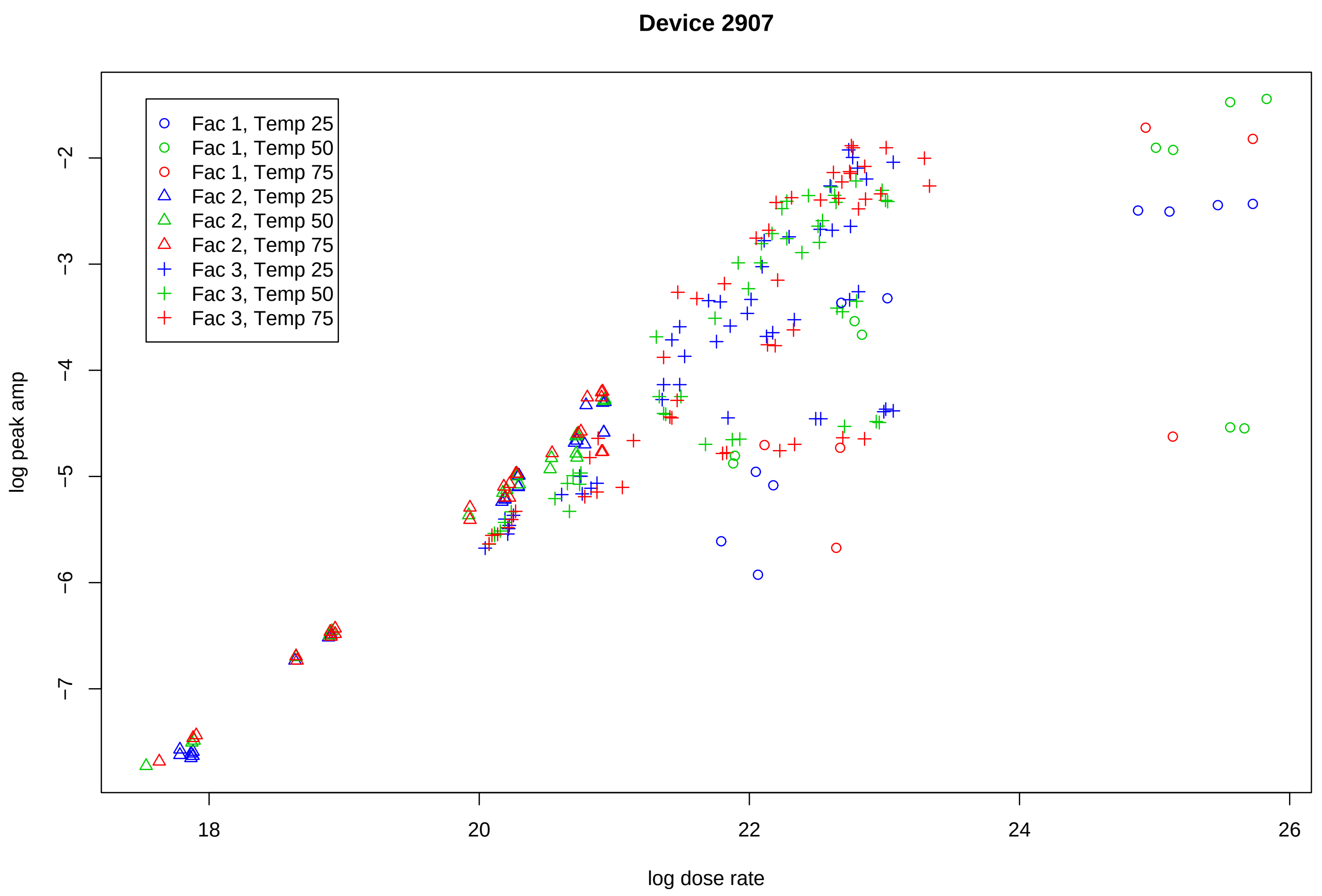
Example Data Plot 1



Example Data Plot 2



Example Data Plot 3



Selection Methodology

- Fit a six-component Gaussian mixture model, with the component centers restricted to observed datapoints
- Fit a fully Bayesian model, with priors that expect the centers to be spread over the space, each "representing" about 1/6 of the dataset
- Choose the representative points to be the MAP center points

Full Probability Model

Likelihood:

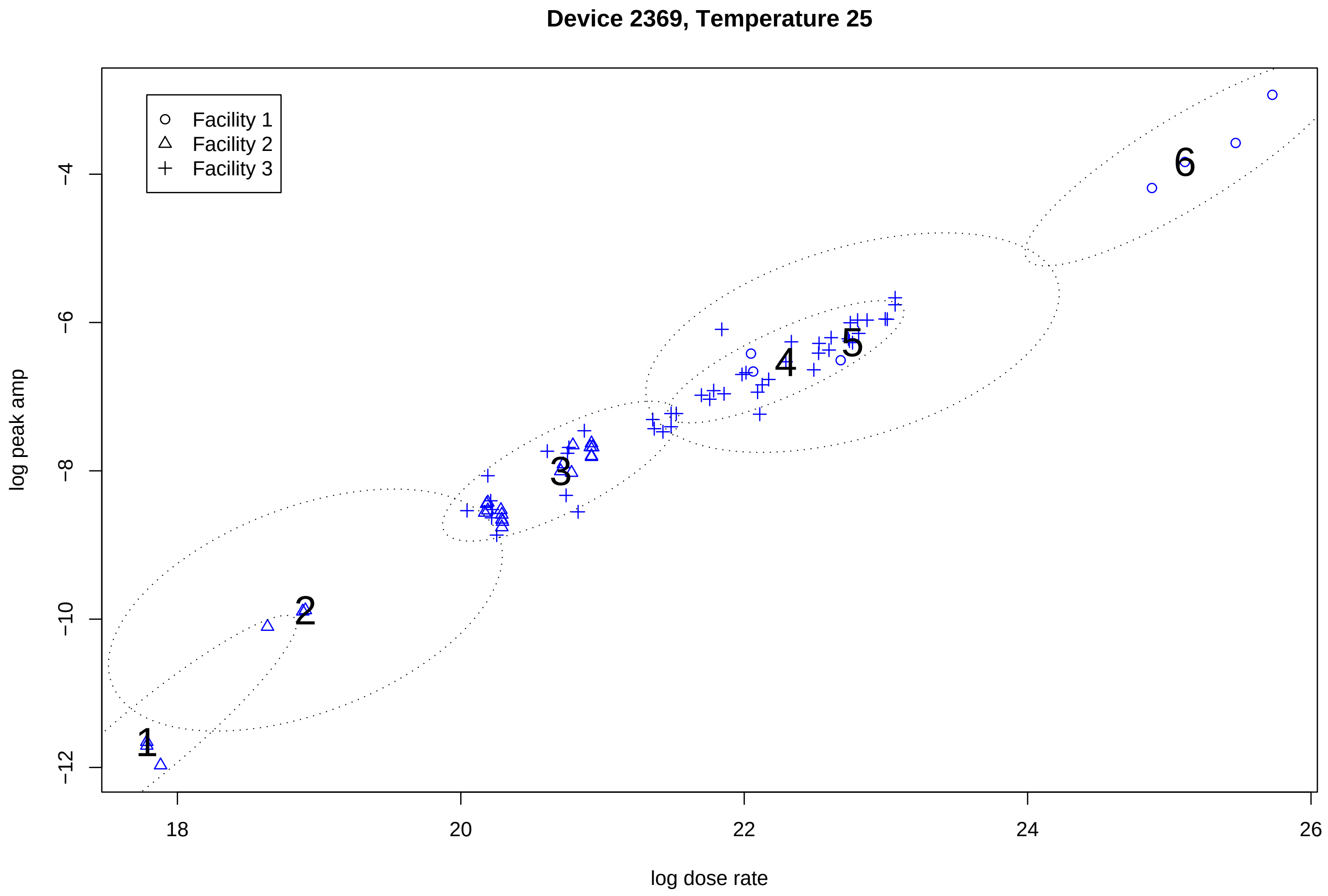
$$\begin{aligned} f(\mathbf{x}, \mathbf{y} | \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{p}) &= \sum_{j=1}^{k} p_j (2\pi)^{-n/2} \left| \boldsymbol{\Sigma_j} \right|^{-1/2} \\ &\quad * \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} \left(\left[\begin{pmatrix} x_i \\ y_i \end{pmatrix} - \begin{pmatrix} \mu_{1j} \\ \mu_{2j} \end{pmatrix} \right]^T \boldsymbol{\Sigma_j}^{-1} \left[\begin{pmatrix} x_i \\ y_i \end{pmatrix} - \begin{pmatrix} \mu_{1j} \\ \mu_{2j} \end{pmatrix} \right] \right) \right\} \end{aligned}$$

Priors:

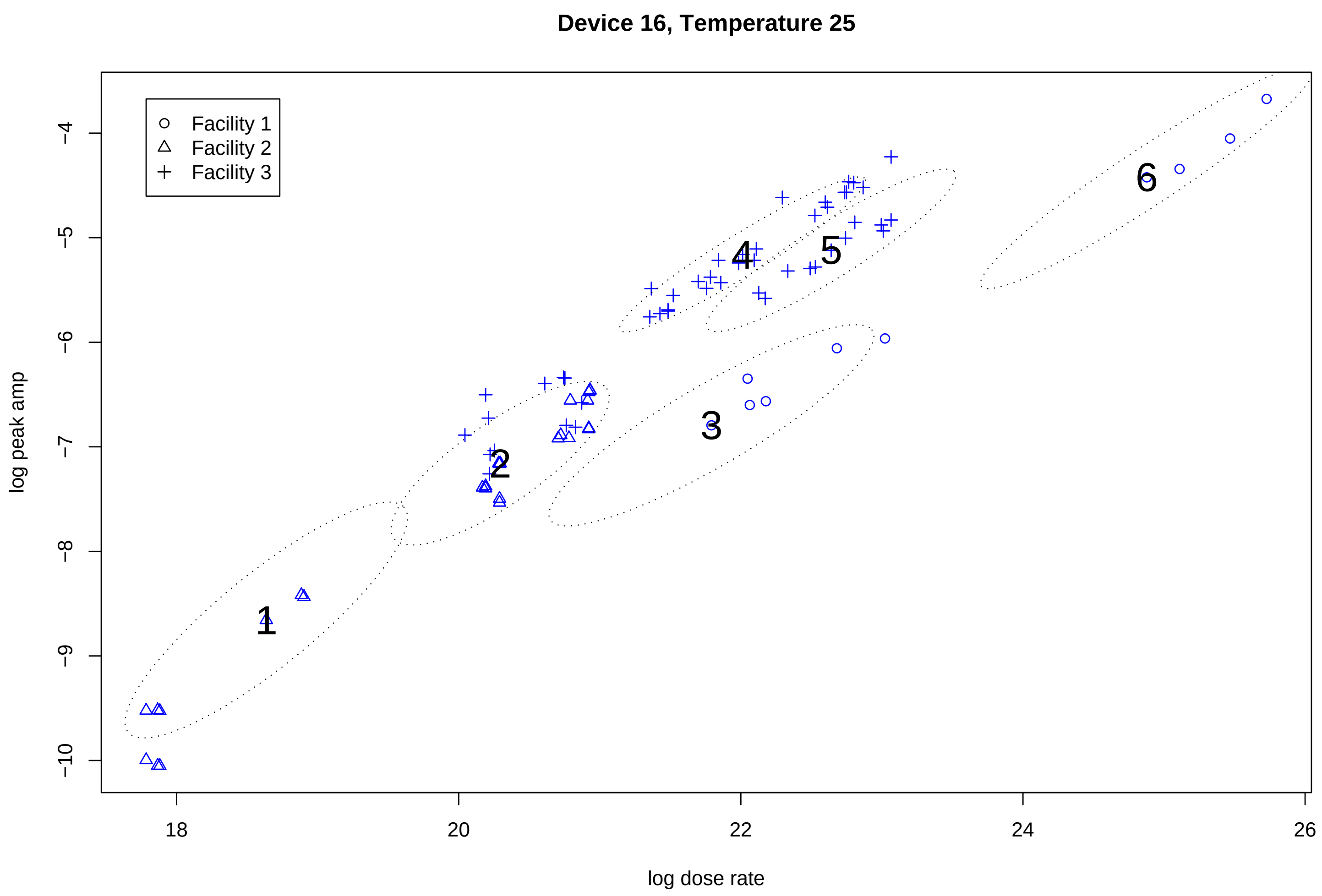
- μ_{hj} is normal centered $j/(k+1)$ of the range
- Σ_j has diagonal elements inverse-gamma, correlations uniform
- p is Dirichlet$(\frac{n}{k}, \ldots, \frac{n}{k})$

Fit with Metropolis-Hastings

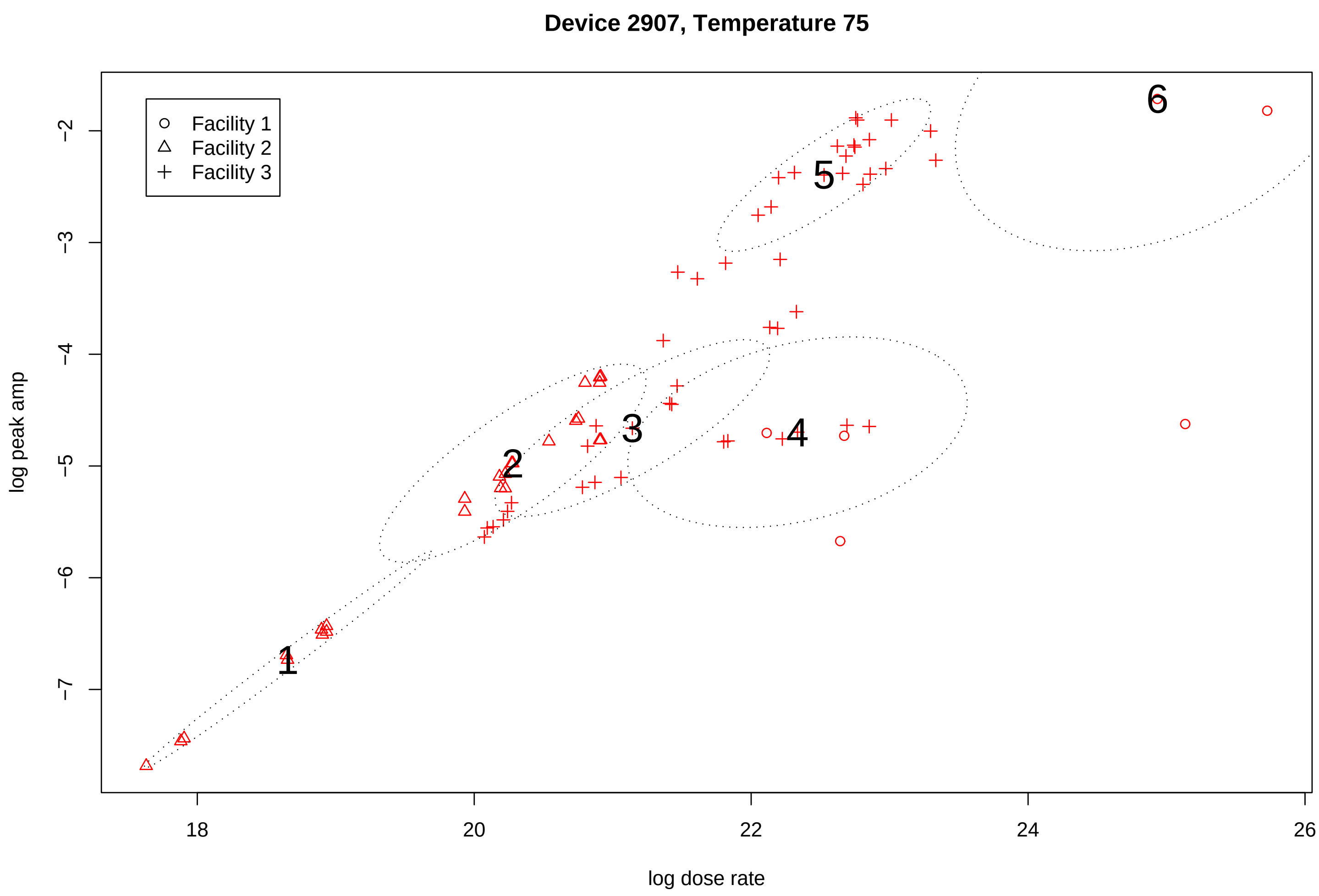
Example Representative Point Plot 1



Example Representative Point Plot 2



Example Representative Point Plot 3



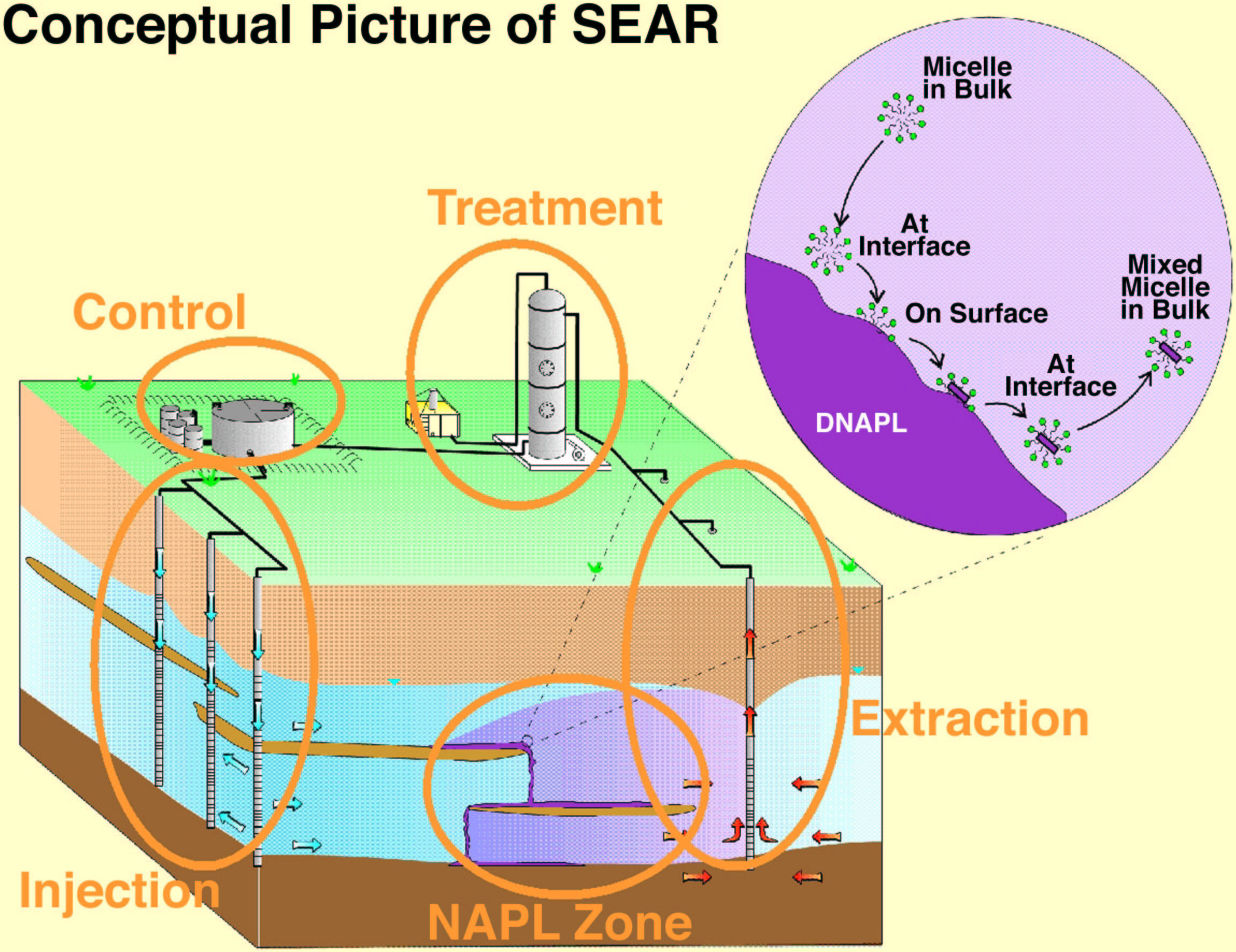
Bayesian Spatial Models for High-Dimensional Parameters in Simulations of Fluid Flow in Porous Media

Collaborators: Dave Higdon, Marco Ferreira, Zhuoxin Bi, Mike West, Chris Holloman, John Trangenstein, Bill Allard, Akhil Datta-Gupta

Project Goals

- Model fluid flow through porous media
- Deal with very high-dimensional parameters (soil permeability) on multiple scales
- Account for uncertainty due to both the high-dimensional model and the deterministic simulation
- Applications to contaminant clean-up and oil reservoir exploration

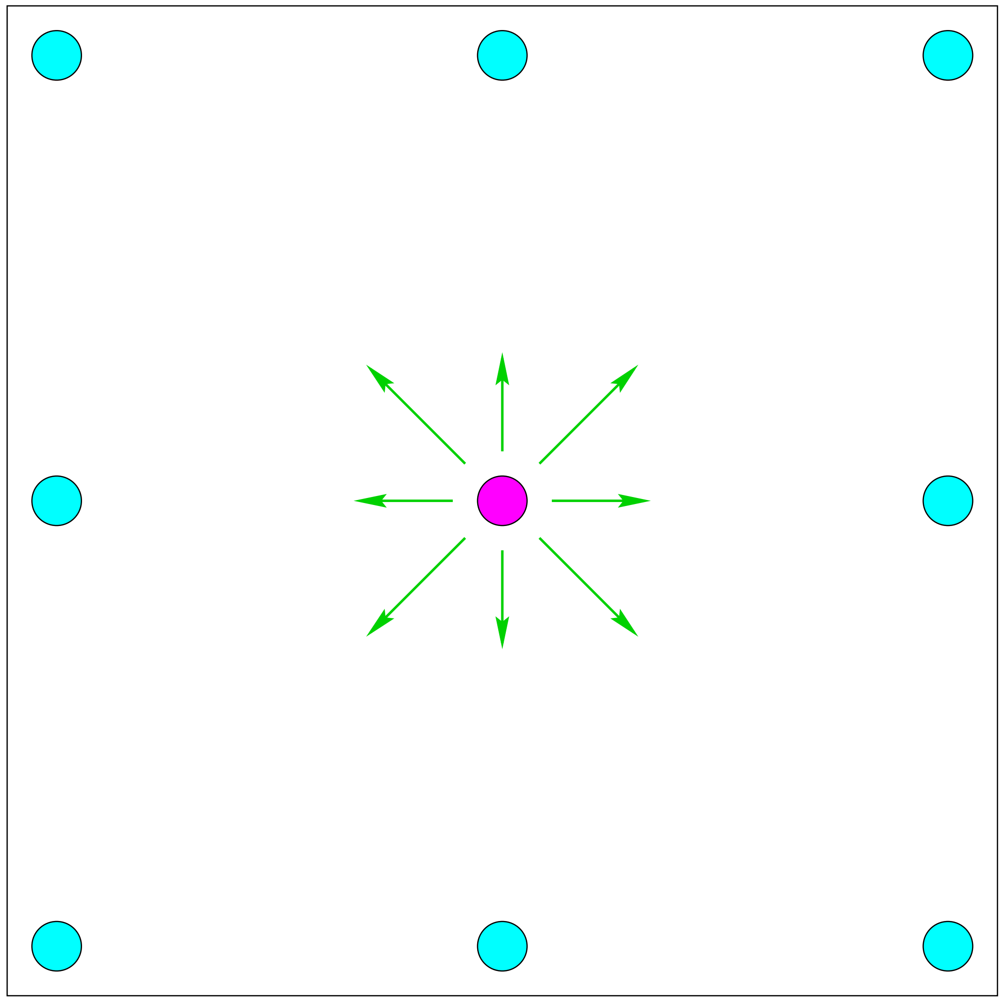
Conceptual Picture of SEAR



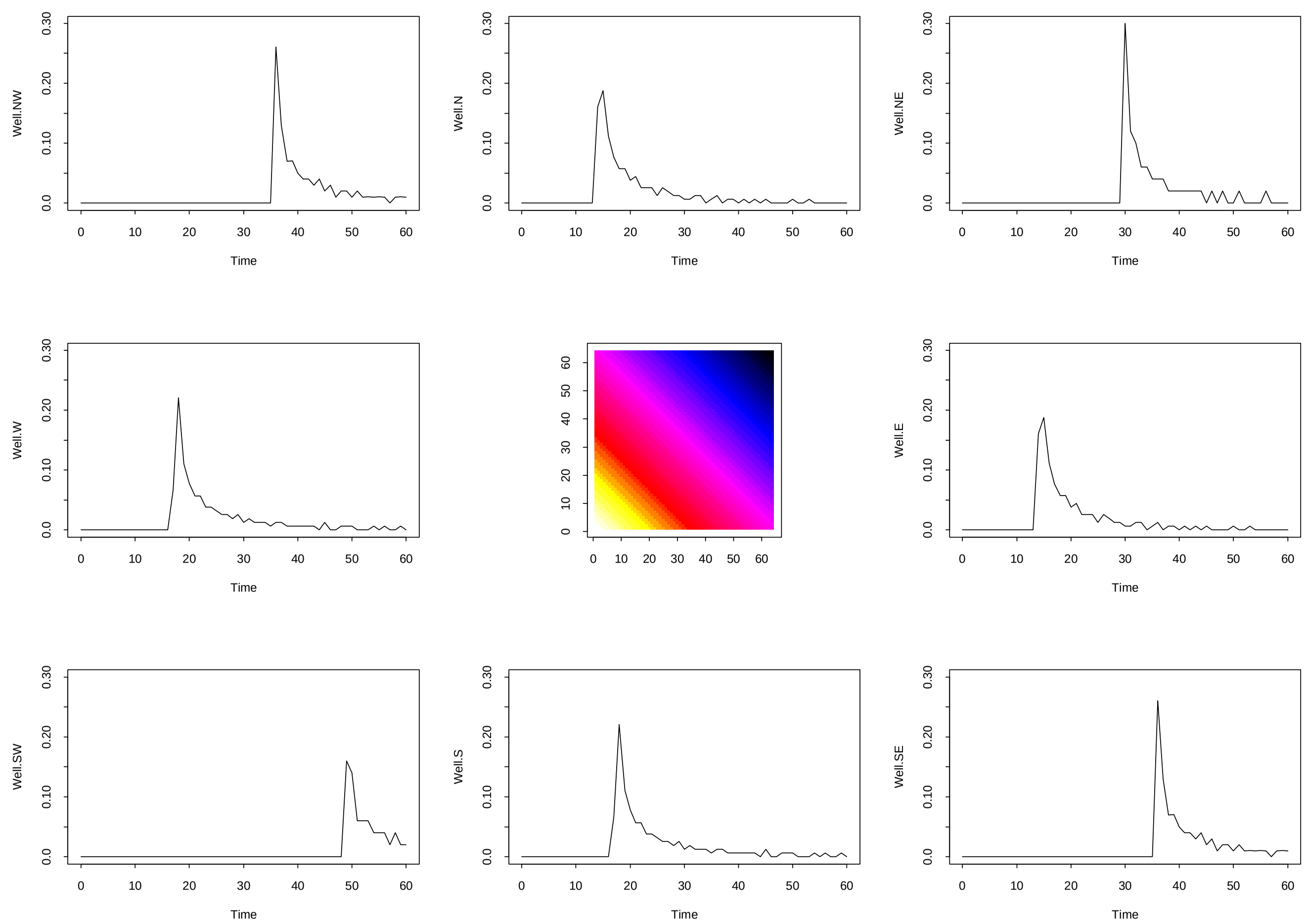
Groundwater Flow

- Forward Problem
  - Solve for flow given soil characteristics
  - Solution from differential equations
- Inverse Problem
  - Infer permeability from flow data
  - Use solutions of the forward problem iteratively

Well Setup



Concentration Curves and Breakthrough Times



Likelihood

$$f(b_h|\psi) \propto \exp\left\{-\frac{1}{2\sigma^2}\sum_h (b_h - \hat{b}_h)^2\right\}$$

b_h is observed breakthrough time

$\hat{b}_h$ is predicted breakthrough time from the flow simulator for a particular value of permeabilities ψ

Clearly, nothing will be conjugate for ψ ...

Spatial Models for the Permeability Field

- Gaussian Processes
- Markov Random Fields (Conditional Autoregressive Processes)
- Combination Convolution Models

Markov Random Fields

Intrinsic Autoregression

$$\pi(\psi) \propto \theta^{\frac{m}{2}} |W|^{\frac{1}{2}} \exp\{-\tfrac{1}{2}\theta\psi^T W\psi\}$$
$$\propto \theta^{\frac{m}{2}} |W|^{\frac{1}{2}} \exp\left\{-\tfrac{1}{2}\theta\sum_{i\sim j}(\psi_i-\psi_j)^2\right\}$$

$$\pi(\psi_i|\psi_{-i}) \propto \exp\left\{-\tfrac{1}{2}\sum_{j\in\mathcal{N}_i}\theta(\psi_i-\psi_j)^2\right\}$$

Parameters of an MRF

- Neighborhood Structure W or $\mathcal{N}_i$
- Precision parameter θ

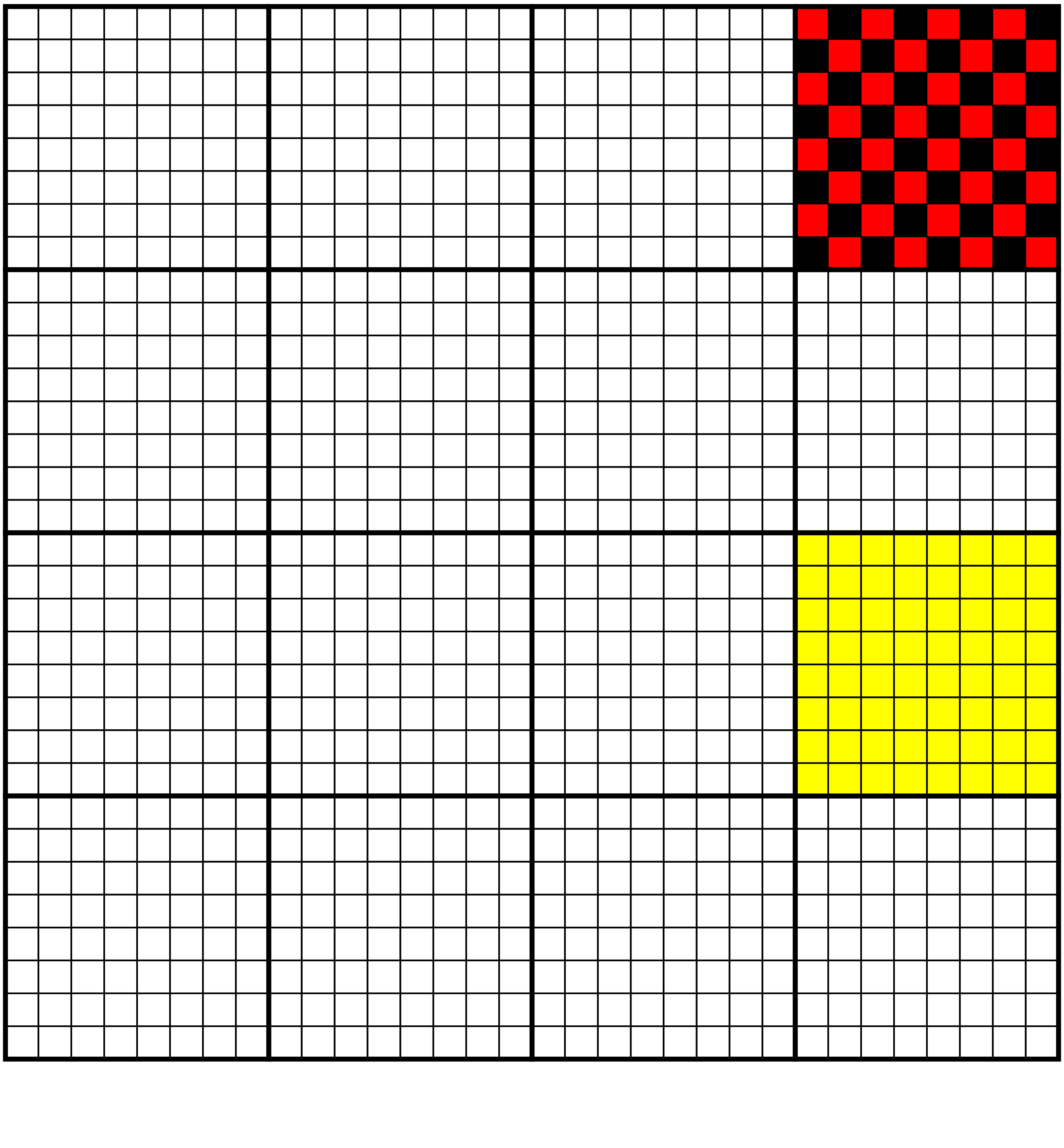
MCMC implementation

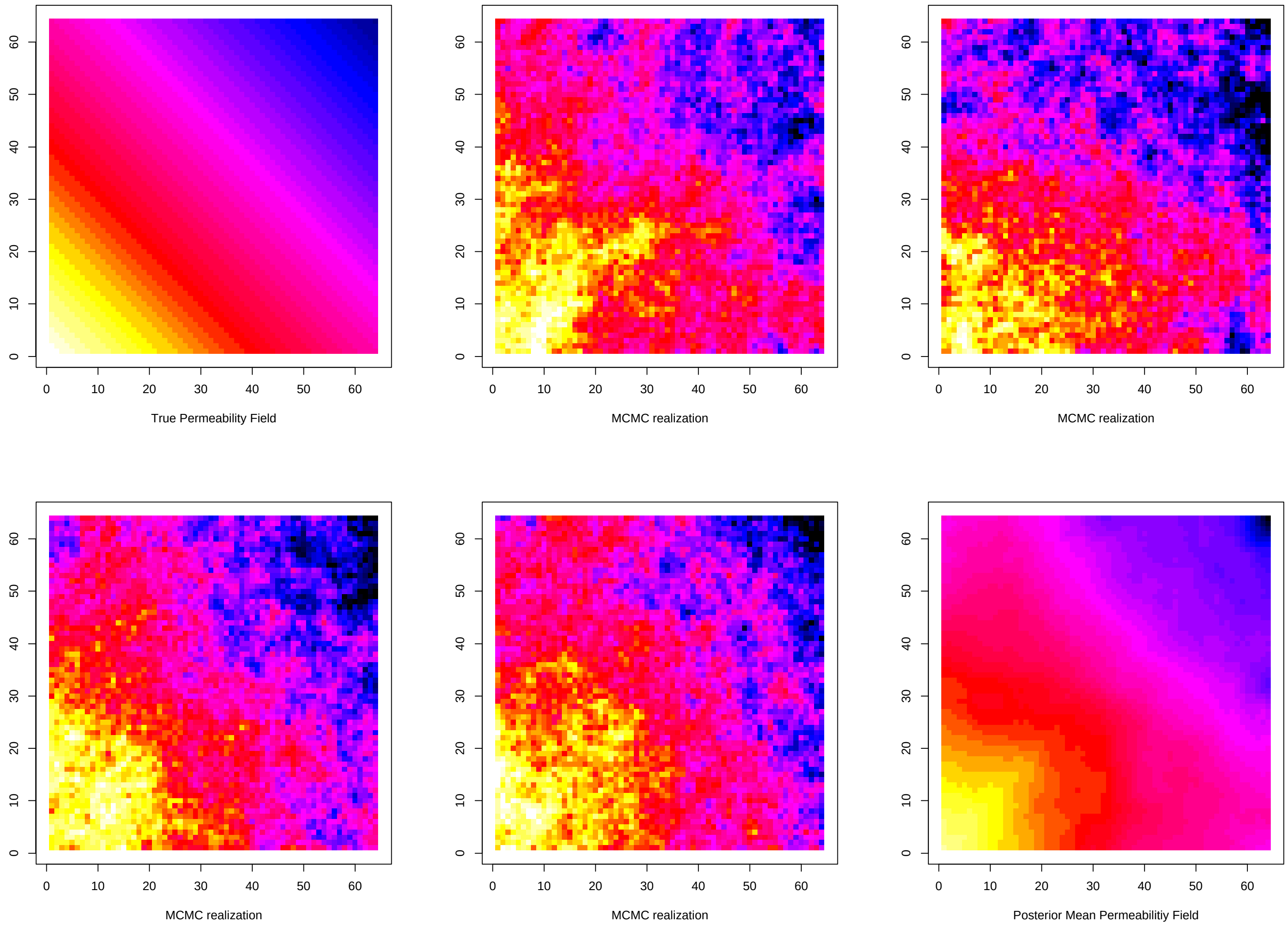
(Lee, Higdon, Bi, Ferreira, and West, 2002)

- Update θ and σ via Gibbs steps
- Update permeabilities via Metropolis-Hastings

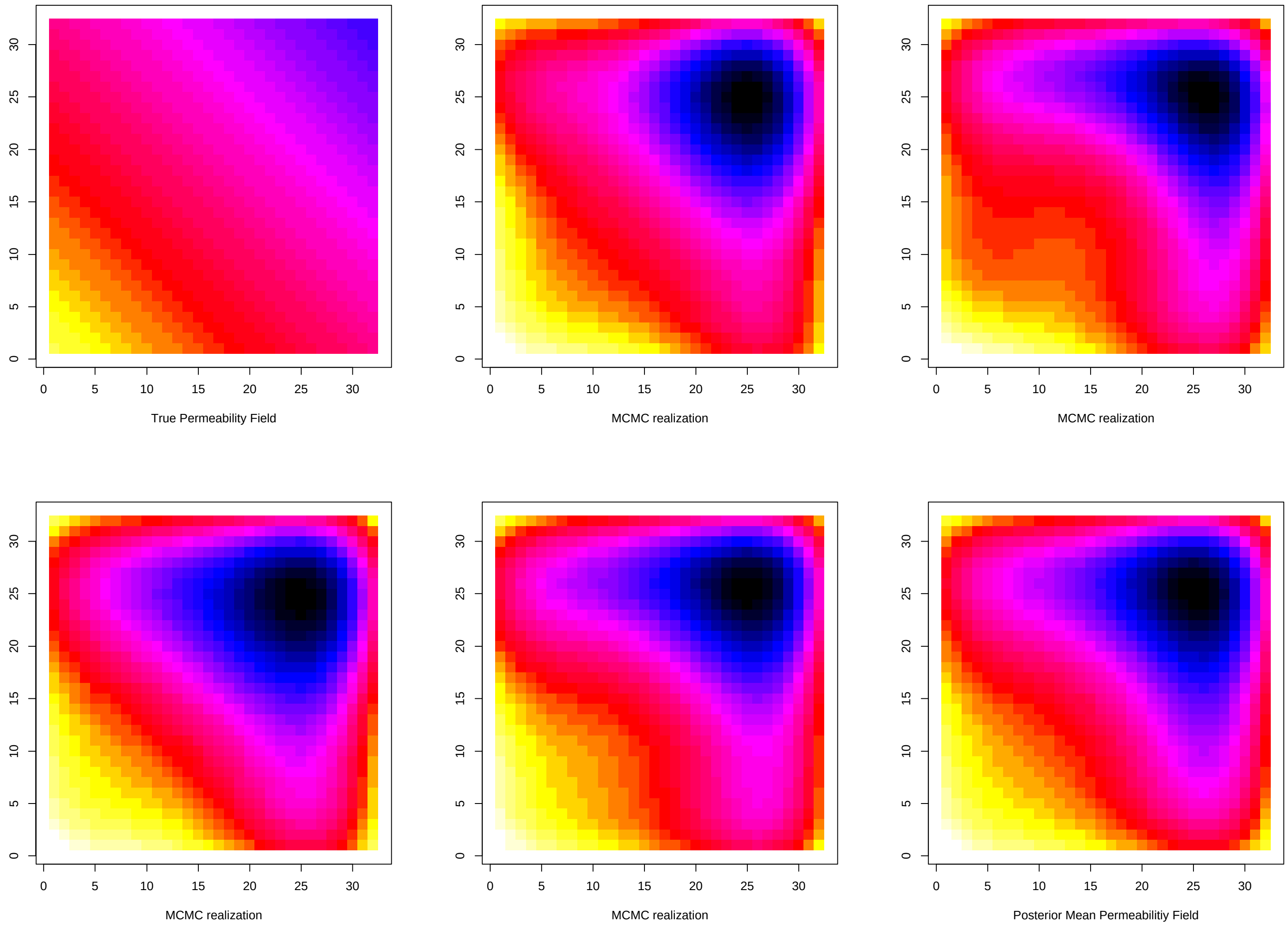
For first-order neighborhoods, we have two types of group candidate proposals on blocks of grid cells:

- Update "red" or "black" cells via draw from prior $\tilde{\psi}_i \sim N(\overline{\psi}_i, \frac{1}{n_i\theta})$ where n_i is the number of neighbors of site s_i, and $\overline{\psi}_i = \frac{1}{n_i}\sum_{j\in\partial s_i}\psi_j$
- Shift whole block by an amount drawn from the prior over the border of the block $\quad \Delta\psi \sim N(\frac{1}{n_{\mathrm{edge}}}\sum(\psi_j - \psi_i), \frac{1}{n_{\mathrm{edge}}\theta})$

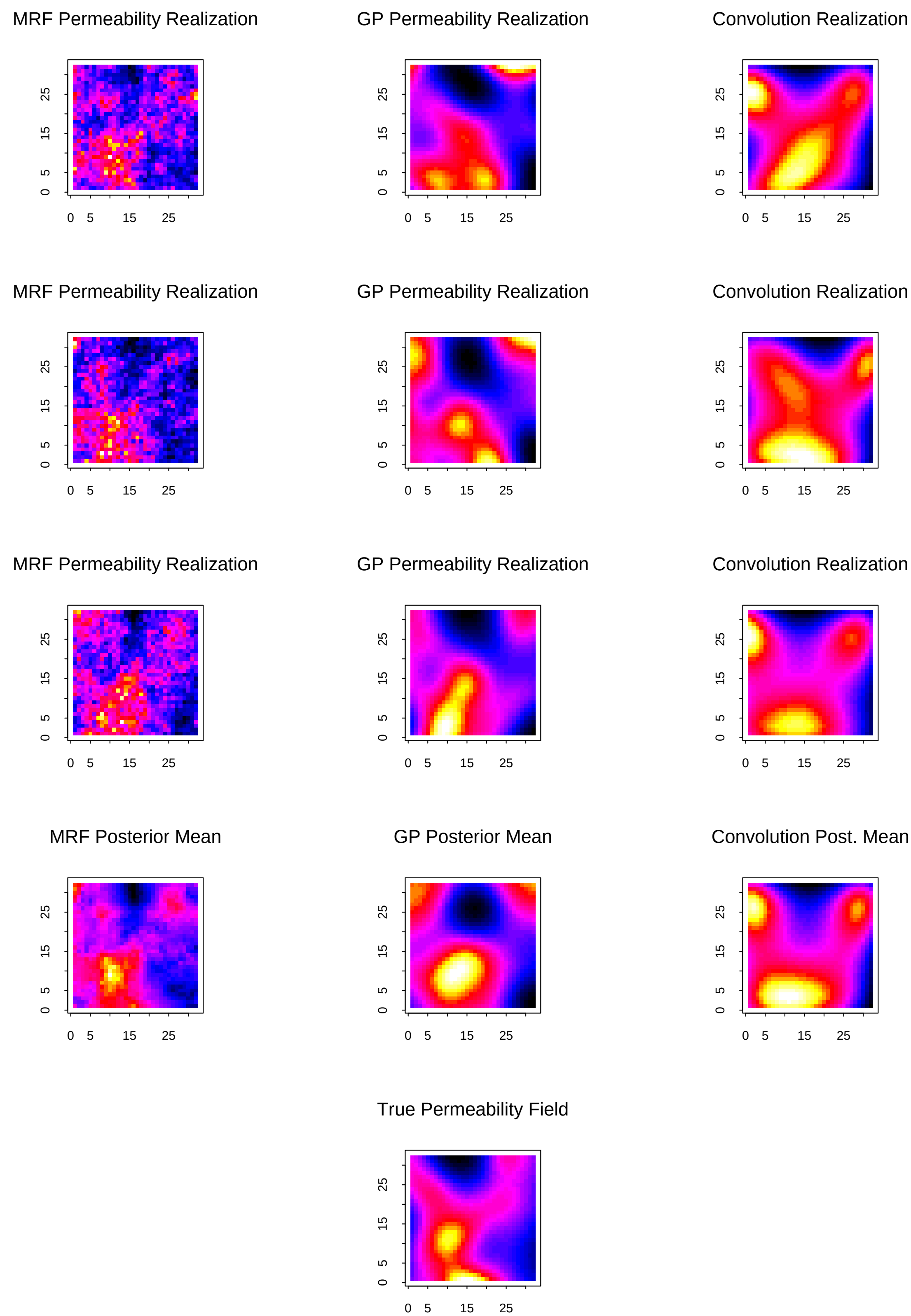




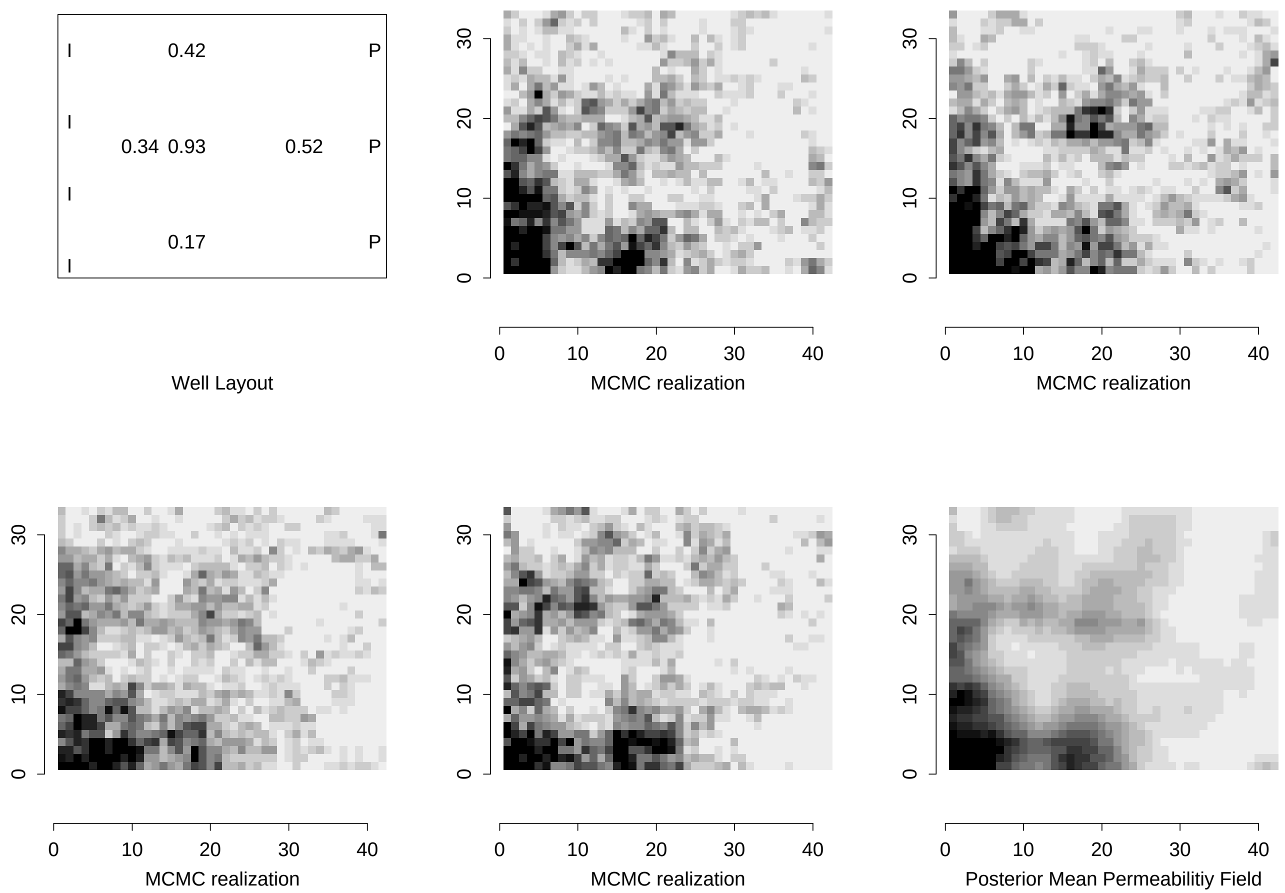
True Permeability Field
MCMC realization
MCMC realization
MCMC realization
MCMC realization
Posterior Mean Permeabilitiy Field
0
10
20
30
40
50
60

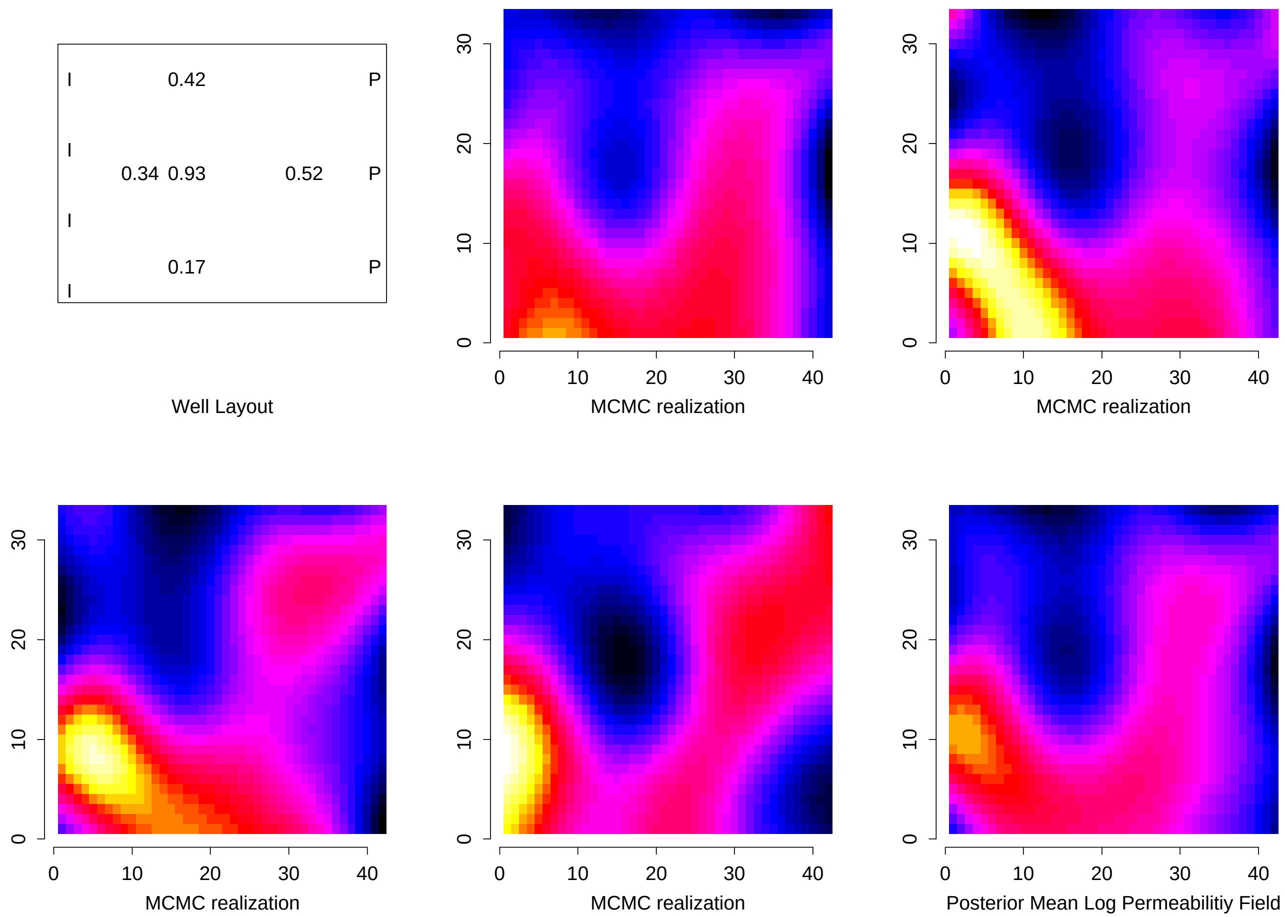


0
5
10
15
20
25
30
True Permeability Field
MCMC realization
MCMC realization
MCMC realization
MCMC realization
Posterior Mean Permeabilitiy Field


MRF Permeability Realization
GP Permeability Realization
Convolution Realization
MRF Permeability Realization
GP Permeability Realization
Convolution Realization
MRF Permeability Realization
GP Permeability Realization
Convolution Realization
MRF Posterior Mean
GP Posterior Mean
Convolution Post. Mean
True Permeability Field
0 5 15 25

A Real Example




I
0.42
P
I
0.34 0.93
0.52
P
I
0.17
P
I
Well Layout
MCMC realization
MCMC realization
MCMC realization
MCMC realization
Posterior Mean Log Permeabilitiy Field

Frequent Computational Difficulties

- Local modes
- Slow mixing
- Likelihood computationally expensive

Coarser grid helps with all three problems

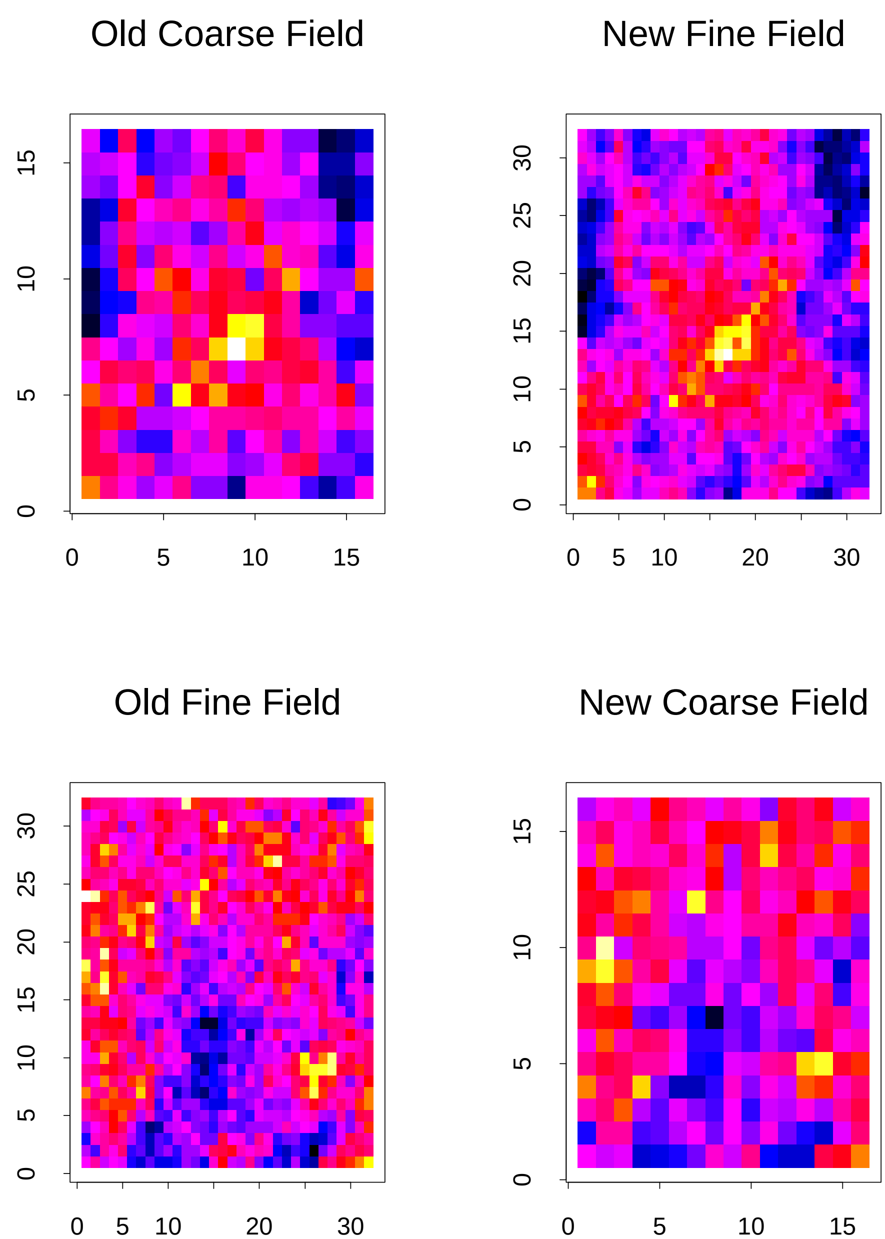
Multi-scale Data

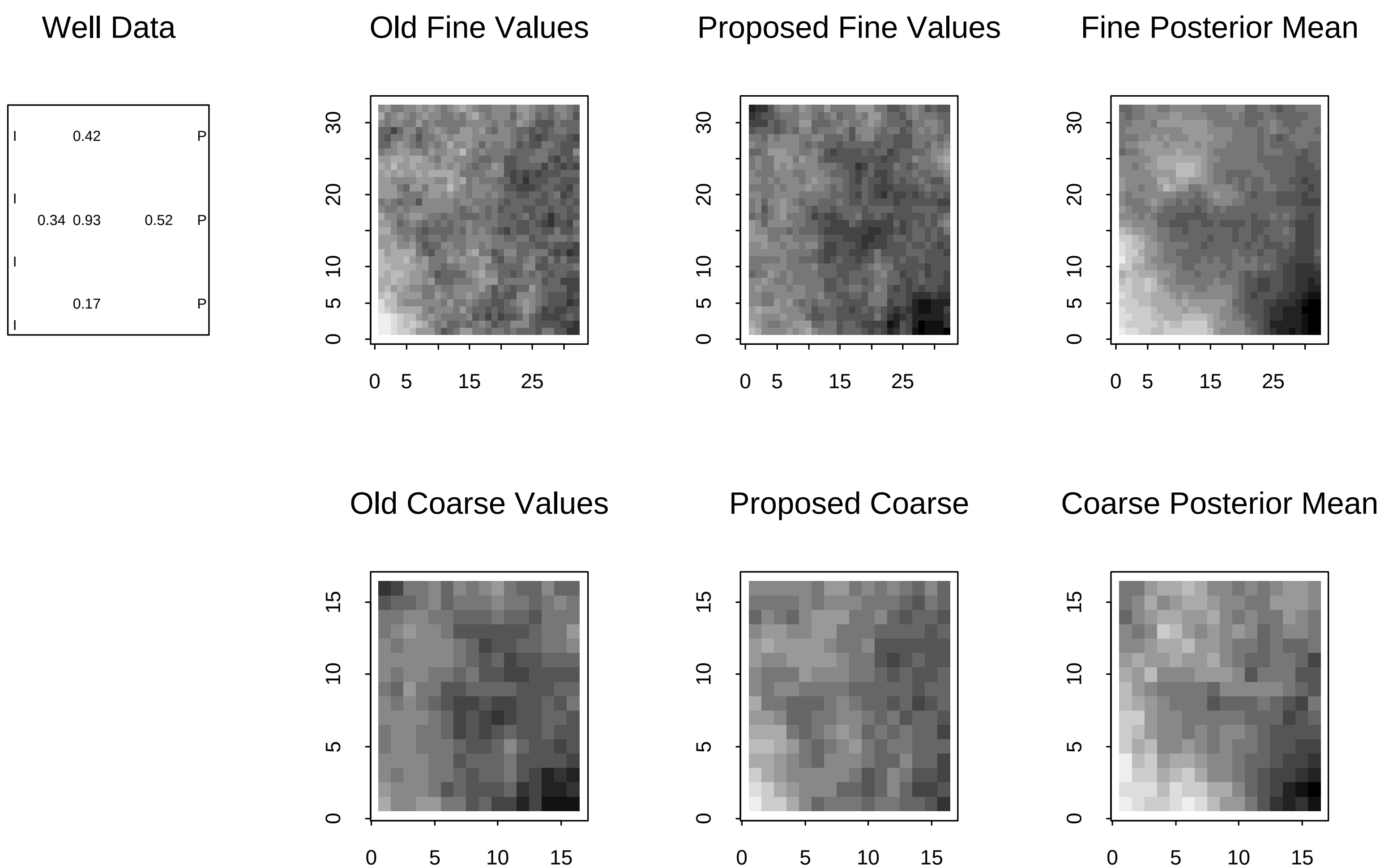
In hydrology, three different sources of data exist on different scales:

- Core Samples
- Flow Data
- Seismic Data

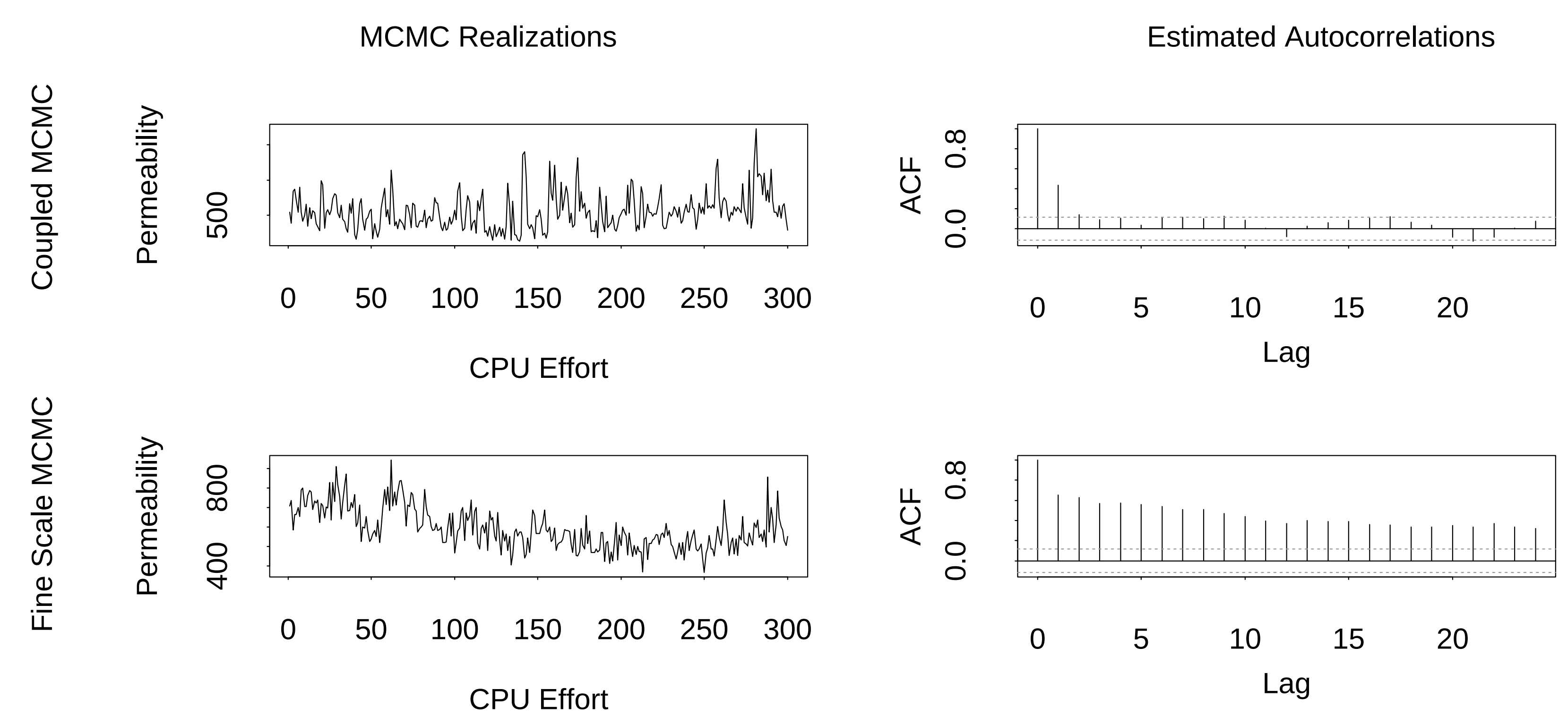
Want a coherent model for multiple scales, link implicitly:

- Metropolis-coupled chains
- Multi-scale Genetic algorithm-based MCMC


Old Coarse Field
New Fine Field
Old Fine Field
New Coarse Field
0
5
10
15
0
5
10
20
30
25



Well Data
I
0.42
P
I
0.34 0.93
0.52
P
I
0.17
P
I
Old Fine Values
Proposed Fine Values
Fine Posterior Mean
Old Coarse Values
Proposed Coarse
Coarse Posterior Mean


MCMC Realizations
Estimated Autocorrelations
Coupled MCMC
Permeability
500
0
50
100
150
200
250
300
CPU Effort
ACF
0.0
0.8
0
5
10
15
20
Lag
Fine Scale MCMC
Permeability
400
800
0
50
100
150
200
250
300
CPU Effort
ACF
0.0
0.8
0
5
10
15
20
Lag

Conclusions

Methodology is quite general — basic components:

- unknown spatial field
- spatial Gaussian prior on the spatial field
- "black box" likelihood
- efficient computer code for evaluating the likelihood