Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle *Since 1965*

Uncertainty Quantification and Reduction in Complex Earth System Models through Exploratory Data Analyses and Calibration

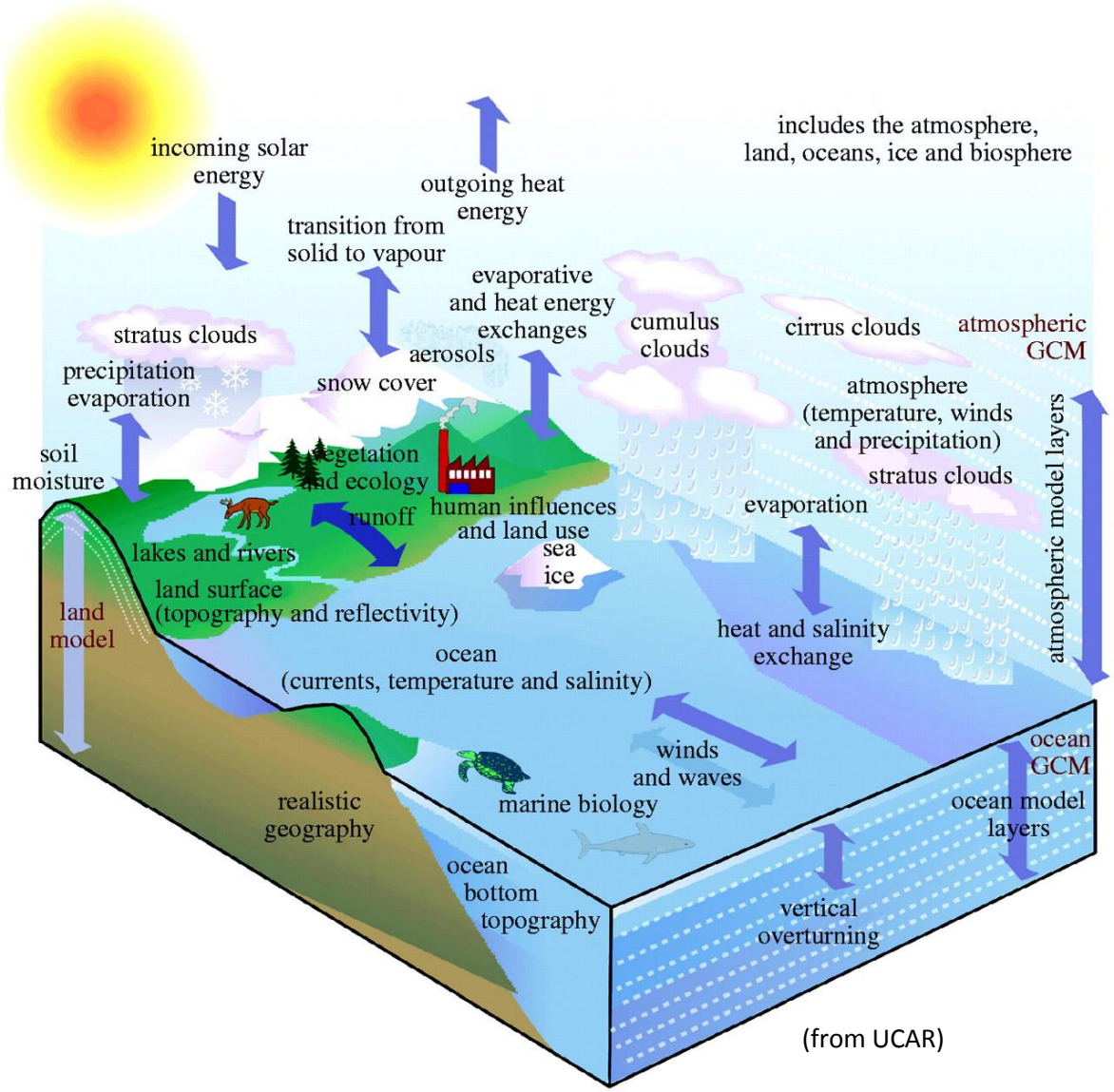
Z. JASON HOU

Pacific Northwest National Laboratory, Richland WA, USA

ICTP Workshop on UQ in Climate Modeling and Projection, Trieste, Italy
July 16 2015

Sources of Uncertainty

- Complex system (e.g., climate): multi-phase, multi-component, involves multiple discipline processes at multiple scales
- High-dimensional parameter space, inadequacy of knowledge, spatiotemporal variability and heterogeneity, resolution and scaling issues



(from UCAR)

Sources of Uncertainty (2)

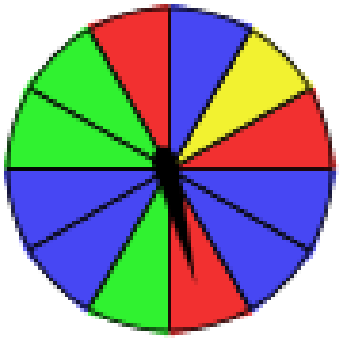
- Model (structural) uncertainty
  - Model inadequacy, bias, discrepancy, simplifications, approximations, lack of knowledge of the physical processes and/or model initial/boundary conditions (exist even if the model parameters are perfectly known)
- Parameter uncertainty
  - Non-informative prior knowledge, non-measurable, measurement errors, under- or down-sampling, non-uniqueness (ill-posed problem), inaccurate calibration (sometimes ‘right answer for wrong reason’)

Sources of Uncertainty (3)

- Data/forcing uncertainty
  - Instrumental errors, consistency, gaps, resolution, scaling
- Natural uncertainty/variability/heterogeneity
  - Intrinsic quantities vary over time, over space, or across individuals in a population
  - Physical processes/mechanisms/features vary over space, time, and individuals

Uncertainty Quantification

- Exercise1: how likely is it that the spinner will land on a blue space?



- certain
- probable
- unlikely
- impossible

Expressions and Measures of Uncertainty

- Given prior information, the probability mass function (pmf) is:

  P(color=red) = 3/12

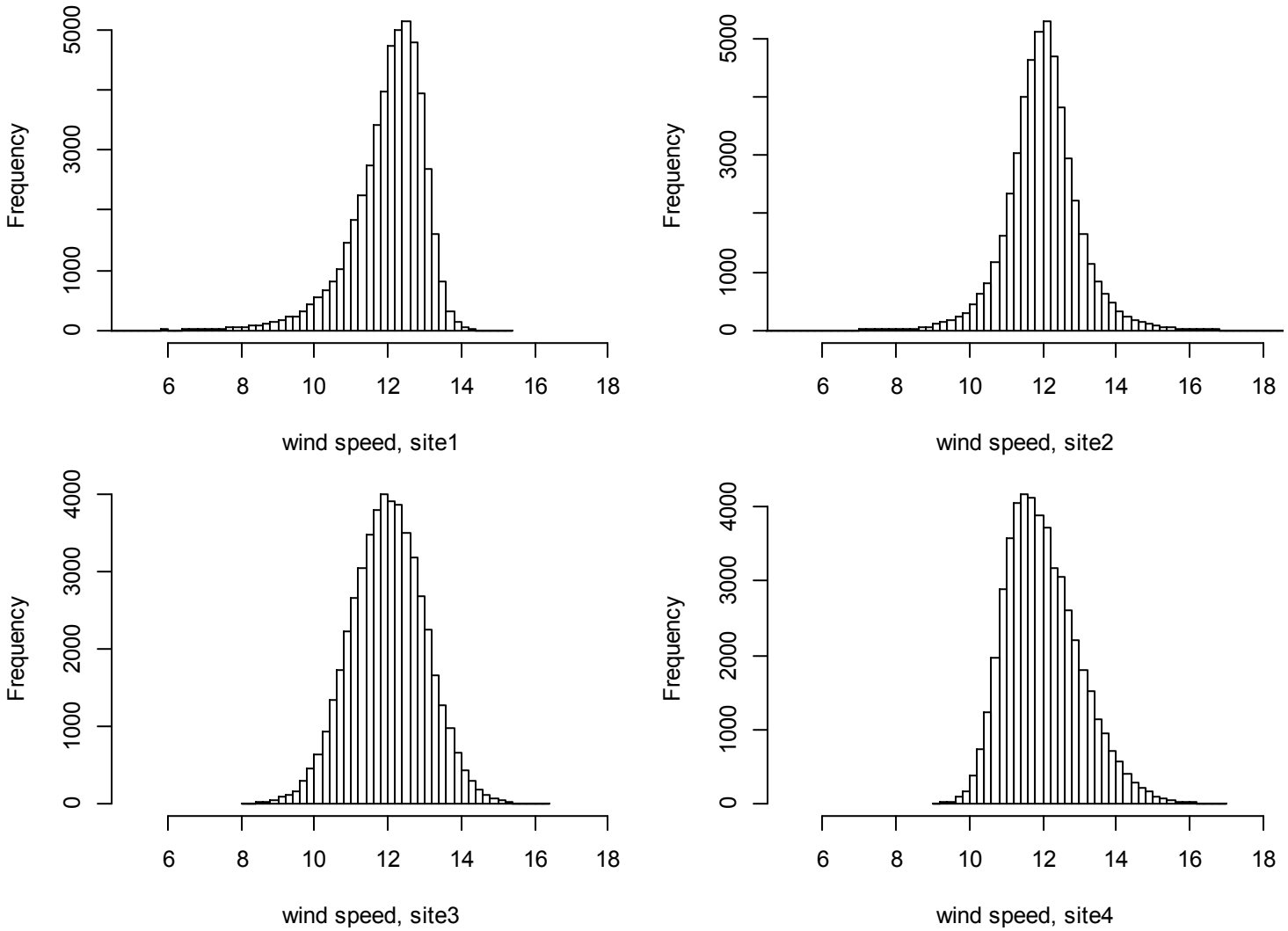
  P(color=yellow) = 1/12

  P(color=green) = 3/12

  P(color=blue) = 5/12

Uncertainty Quantification

- Exercise2: selection of wind farm sites given hourly averaged wind speed

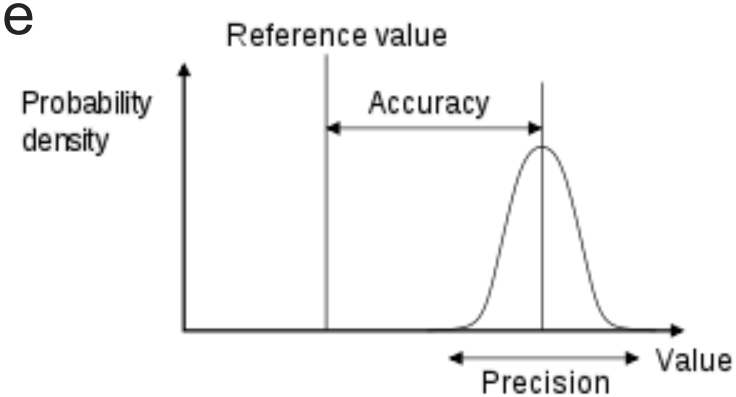


Uncertainty Quantification

- To replace the subjective notion of confidence with a mathematically rigorous measure, honoring and committed to:
  - Hard/direct information:
    - Experimental observations
    - Theoretical arguments
    - Expert opinions
  - Soft/indirect information
    - Inverse methods (Inference, Calibration)

Expressions and Measures of Uncertainty

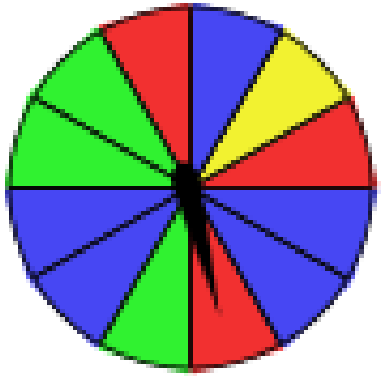
- Summary
  - Mean (bias/accuracy) and confidence intervals (precision)
  - Possible summary statistics (skewness, kurtosis, median, mode, percentiles)
  - Probability density/mass function (for continuous / discrete random variables, respectively), can full describe accuracy/precision of a variable



  - Entropy $\mathrm{H}(\mathcal{S}) = -\sum p_i \log_2 p_i$
  - Kullback–Leibler divergence = relative entropy = information gain $D_{\mathrm{KL}}(P\|Q) = \sum_i P(i) \ln \frac{P(i)}{Q(i)}.$

An example UQ for Decision Making

- Well quantified uncertainty → reliable decision making
- Example:
  - Input variable (parameter): the color of the block that the spinner will land on (uncertain outcome)
  - Formula (forward model): Bet 1€, get 0€ on blue; 1€ on green; 2€ on red; 4€ on yellow.
  - Return (model output): the return after 60 plays
  - Approach:
    - Most likely estimate?
    - Perturbation-based scenarios?
    - Probability mass function?
      - Yes, go for it.
      - adequate number of plays required for validation

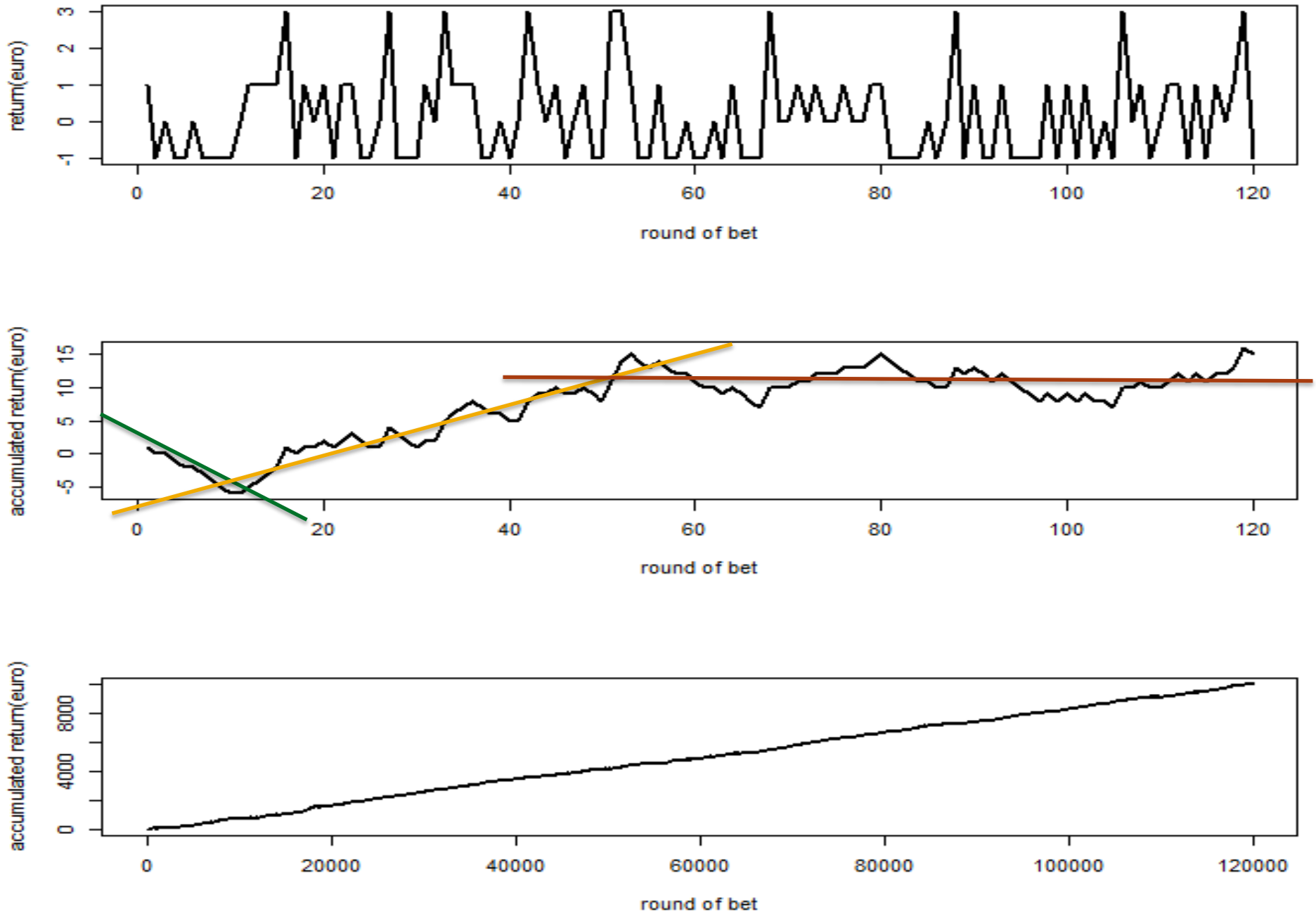


Sampling of pmf

- R code simulating the return for decision making verification (mathematical)

```
N=120
return=sample(c(-1,0,1,3), N, replace = TRUE, prob = c(5/12,3/12,3/12,1/12))
cum.return=cumsum(return)
par(mfrow=c(2,1))
plot(return,type='l',lwd=2,xlab='round of bet',ylab='return(euro)')
plot(cum.return,type='l',lwd=2,xlab='round of bet',ylab='accumulated
    return(euro)')
```

- Observe adequate number of plays for validation (actual)
- Update the pmf based on the observations for more accurate prediction of return (e.g., the bottom blocks have higher odds, due to gravity effect?)


return(euro)
-1
0
1
2
3
0
20
40
60
80
100
120
round of bet
accumulated return(euro)
-5
0
5
10
15
0
20
40
60
80
100
120
round of bet
accumulated return(euro)
0
4000
8000
0
20000
40000
60000
80000
100000
120000
round of bet

UQ components

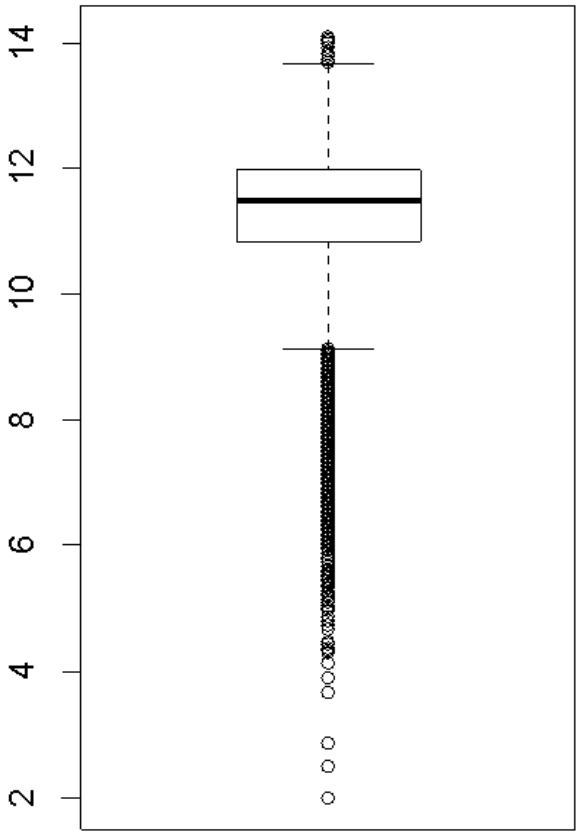
- A convincing UQ study might involve, but is not limited to:
  - Reliable quantification of a input uncertainty (parametric, forcing, natural variability, etc.)
  - Exploratory experimental design (e.g., efficient sampling)
  - Accurate forward model
  - Updating the prior knowledge (e.g., pdfs) given observations
  - Quantification and reduction of output uncertainty (accuracy & precision)
- Focuses of this presentation
  - Derivation of pdfs for exploratory experimental design (EED)
  - Efficient sampling for EED
  - Bayesian inversion
  - Computational challenges and solutions

Derivation of pdfs

- Summary statistics:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
2.0	10.8	11.5	11.3	12.0	14.1

- [Q1 – 1.5×IQR Q3 + 1.5×IQR]: [9.1, 13.7]
- Mean: 11.3
- stdev: 1.0
- 99% CI: [8.8, 13.9], actual 97.8%
- Discussion
  - Normal distribution? Uniform?
  - Most likely estimate for system risk analysis?
  - Perturbation-based scenarios?
  - Probability density function?
    - Yes, fully represented possibilities.
    - But how to derive the pdfs for sensitivity analysis and/or model calibration?



Derivation of pdfs

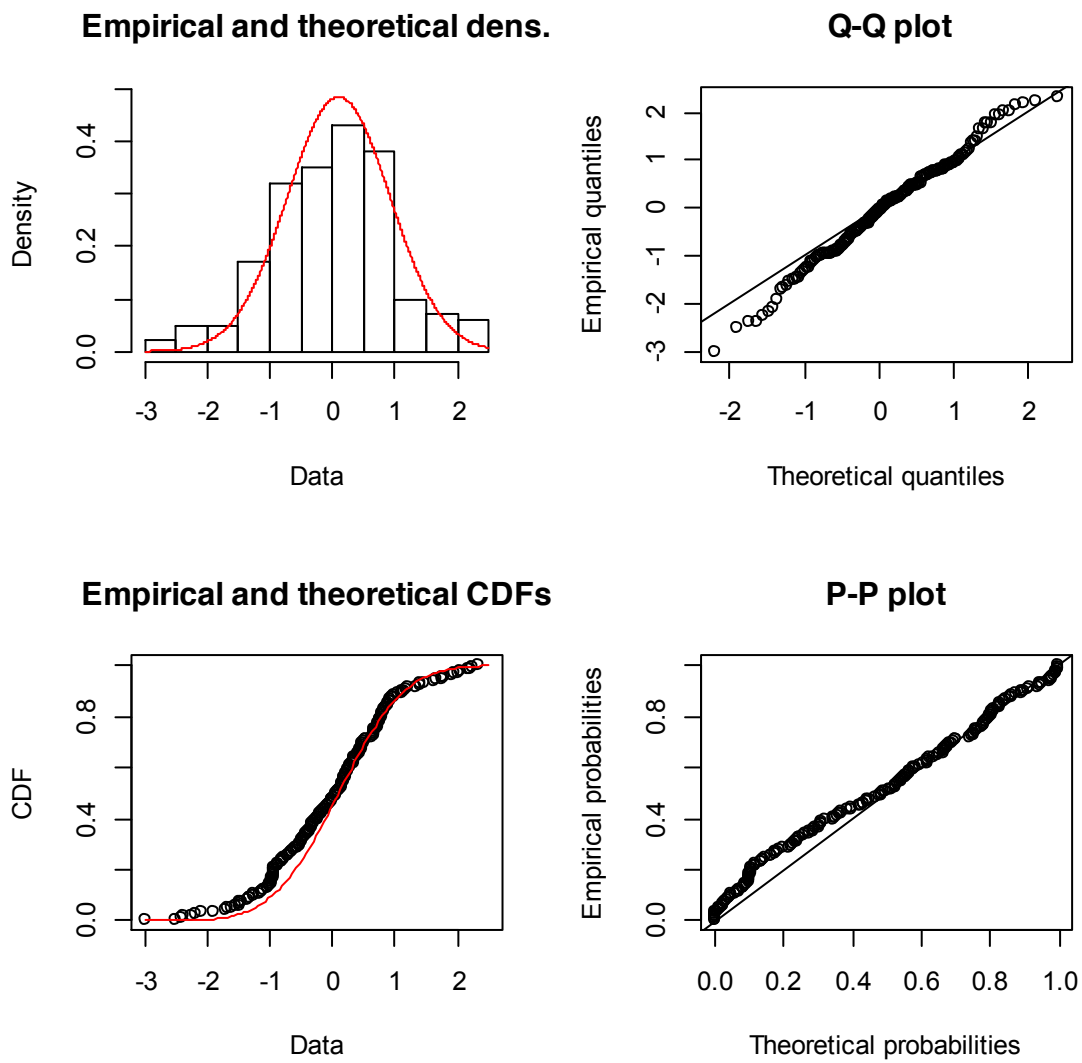
- Other issues:
  - Truncated distributions due to physical bounds
  - Bimodal distributions
  - Tailed distributions
  - Parameter spanning several orders of magnitude
  - Significant amount of zeroes in measurements
  - No observations/measurements

Pdf training/fitting

- R fitdistrplus:
  - require(fitdistrplus)
  - set.seed(1)
  - dat <- rnorm(50,0,1)
  - f1 <- fitdist(dat,"norm")
  - plotdist(dat,"norm",para=list(mean=f1$estimate[1],sd=f1$estimate[2]))

- Provides a closed formula for each of the following distributions: "normal", "lognormal", "Poisson", "exponential", "gamma", “nbinomial", "geometric", "beta", "uniform“, and "logistic“, and so on.
- Tailed (e.g., packages heavy/spd)
- Bimodal (e.g., mixtools)
- Truncated (e.g., gamlss.tr)

Pdf training/fitting

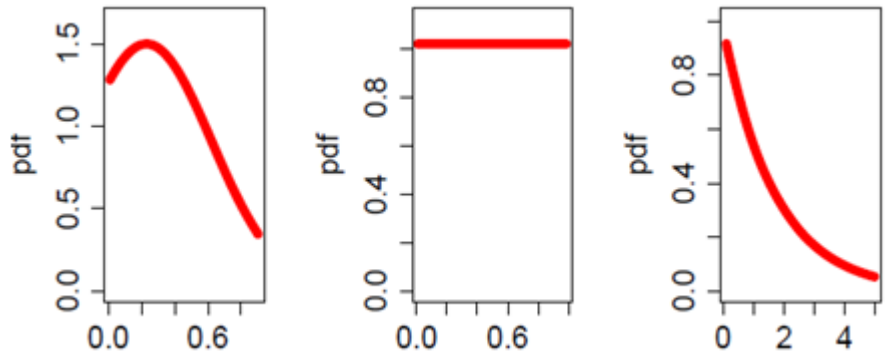


Pdf derivation based on entropy theory

- In practice, measurements might be missing
- Given statistical knowledge from literature/databases/experiences, close-form pdfs can be derived using minimum-relative-entropy (MRE) concept (Hou and Rubin 2005):

$$f(x) = \frac{\sqrt{\frac{\gamma}{\pi}} \exp\left[-\gamma\left(x + \frac{\beta}{2\gamma}\right)^2\right]}{\Phi\left[\sqrt{2\gamma}\left(U + \frac{\beta}{2\gamma}\right)\right] - \Phi\left[\sqrt{2\gamma}\left(L + \frac{\beta}{2\gamma}\right)\right]} \qquad L \le x \le U$$

$$f(x) = \frac{\beta e^{-\beta x}}{e^{-\beta L} - e^{-\beta U}} \qquad L \le x \le U$$



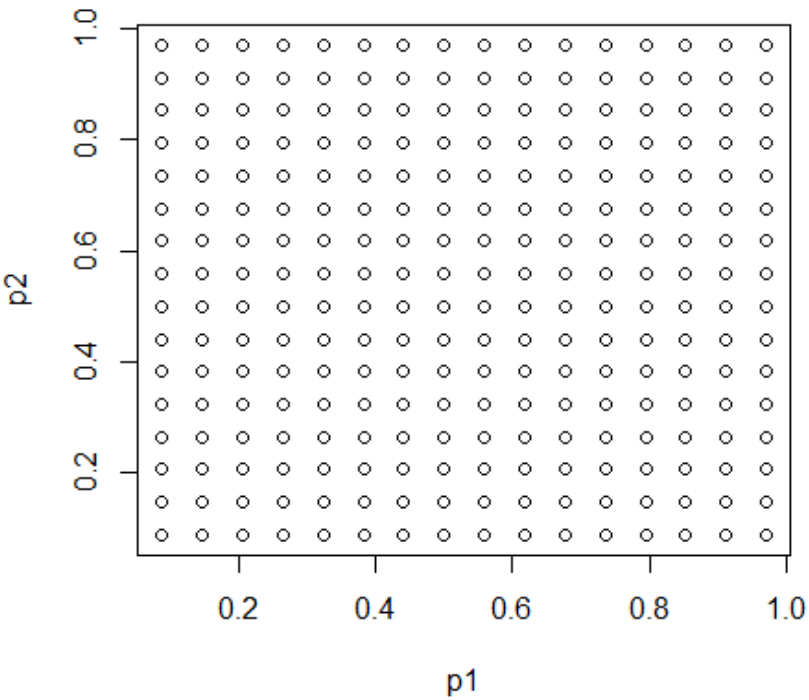
What's next?

- Well quantified uncertainty → systematic ensemble design (with efficient sampling)
- Sampling the prior pdfs using Latin Hypercube Sampling (LHS) or Quasi Monte Carlo (QMC)
  - Applications:
    - Sensitivity analysis
    - Parameter ranking and screening
    - Parameter dimensionality reduction
    - Development of predictive models
    - Development of surrogates for model calibration
- Sampling the posterior (e.g., MCMC) with observational data available – stochastic calibration
  - Improved parameter values → improve model predictive capability and reduced uncertainty → reliable risk assessment and decision making
  - Sampling with direct numerical simulator vs with surrogates

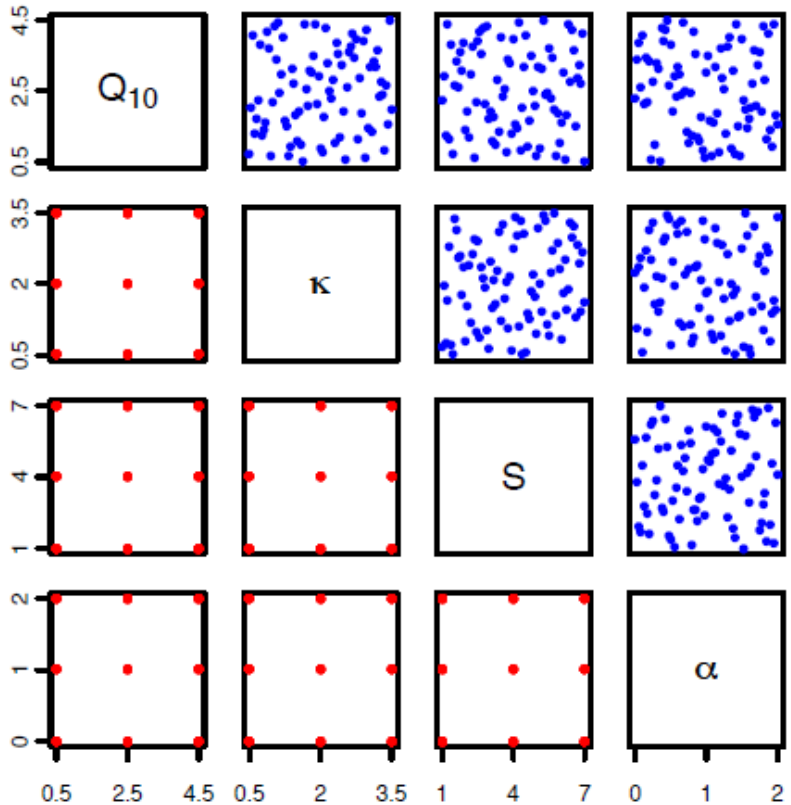
Ensemble design by sampling the pdf

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Grid design



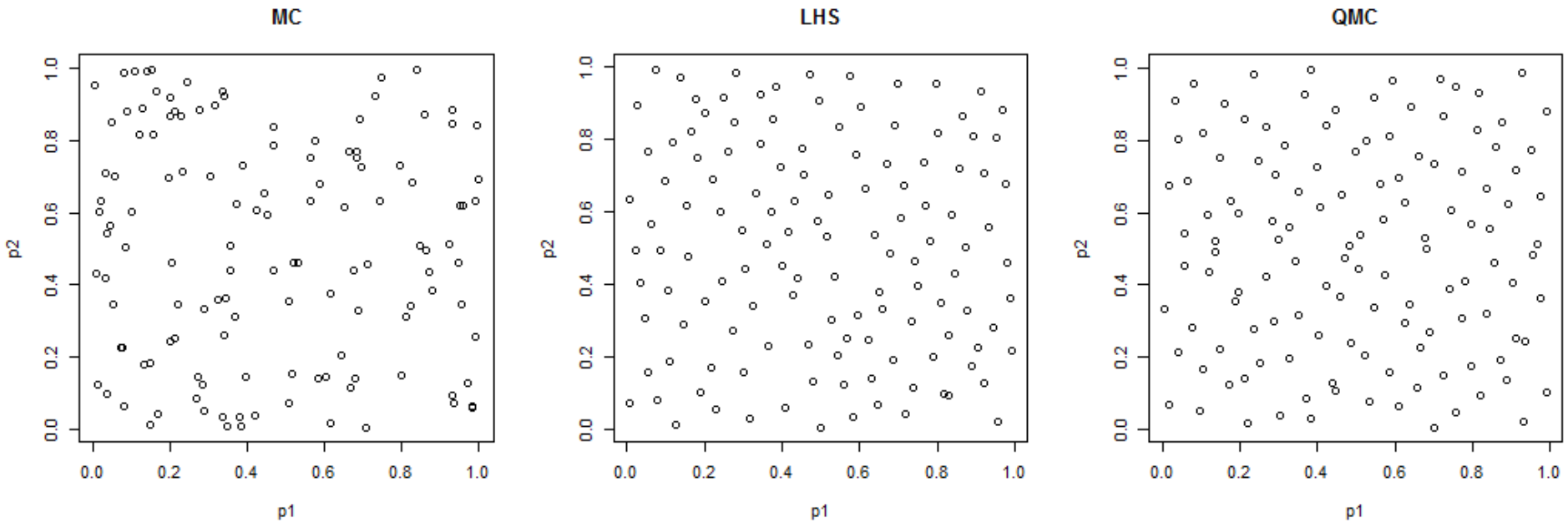
Grid design vs LHS design



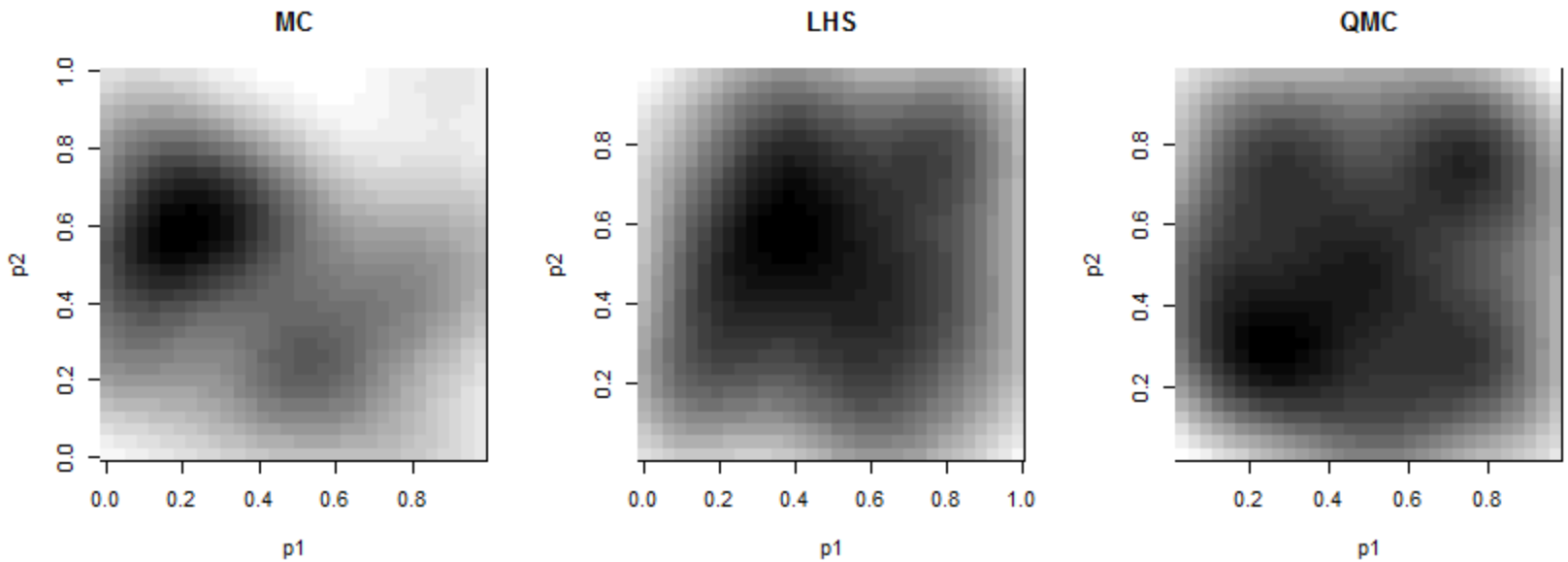
Designs of *p = 4 dimensional, N = 81 member ensembles.* Depicted are scatterplots of the designs projected onto two-parameter subspaces. *Lower left: Grid design. Upper right: Latin hypercube* design. Note each point in grid scatterplots represents 32 = 9 different ensemble points: due to a grid design's collapsing property, unique points in 4-dimensional parameter space can project onto identical points in a 2-dimensional subspace. (Urban et al 2010)

Ensemble design by sampling the pdf

- Effective selection of a subset of individuals from within a statistical population to estimate characteristics of the whole population
  - Representative of the parameter space
  - Avoid clumping and gaps (space filling without redundancy)
  - Avoid extrapolations

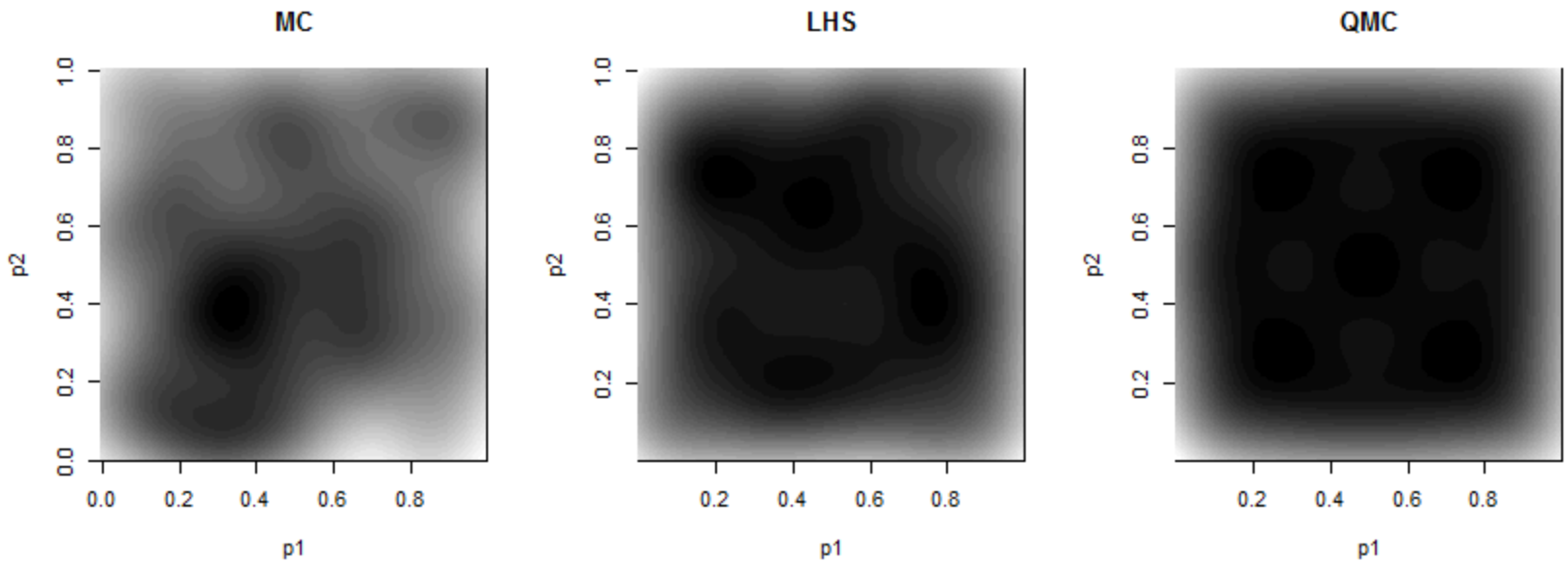


Sampling to fill the probability space



32 samples for a 2D probability space

Sampling to fill the probability space



256 samples for a 2D probability space

Sampling with hierarchical mixed parameters

- Hierarchical parameters: hyper-parameter or scenarios (e.g., RCP for CMIP5)
- Mixed types (discrete vs continuous, Normal vs Uniform vs other distribution types, independent vs correlated subset parameters)
- Sampling with hierarchical mixed parameter types can be done? Sure thing!
  - Generate samples in the probability space (uniform)
  - Projection to parameter space via inverse CDF (quantile) function
  - Apply Cholesky decomposition onto subset of samples where appropriate for correlated subset parameters

Uncertainty quantification/reduction framework – land surface modeling

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Parameterization
- candidate parameters of interest

Symbol	Definition
f_{max}	Max fractional saturated area, from DEM
C_s	Shape parameter of the topographic index distribution
f_{over}	Decay factor (m^{-1}) for fsat
b	Clapp and Hornberger exponent
K_s	Hydraulic conductivity ($mm\ s^{-1}$)
θ_s	porosity
Ψ_s	Saturated soil matrix potential (mm)
f_{drai}	Decay factor (m^{-1}) drainage
$q_{drai,max}$	Max drainage ($kg\ m^{-2}\ s^{-1}$)
S_y	Average specific yield

Prior information (mean, variance, bounds)

Entropy method
Minimum Relative Entropy (MRE)

$$f_{M|I}(m|I) = \prod_{j=1}^{p} \frac{\beta_j e^{-\beta_j m}}{e^{-\beta_j L_j} - e^{-\beta_j U_j}}$$

$$f_{M|I}(m|I) = \prod_{j=1}^{p} \frac{\sqrt{\frac{\gamma_j}{\pi}} \exp\left[-\gamma_j \left(m_j + \frac{\beta_j}{2\gamma_j}\right)^2\right]}{\Phi\left[\sqrt{2\gamma_j}\left(U_j + \frac{\beta_j}{2\gamma_j}\right)\right] - \Phi\left[\sqrt{2\gamma_j}\left(L_j + \frac{\beta_j}{2\gamma_j}\right)\right]}$$

Closed-form Prior probability density functions (pdfs)

Quasi Monte Carlo sampling

Realizations of parameter sets

CLM forward modeling

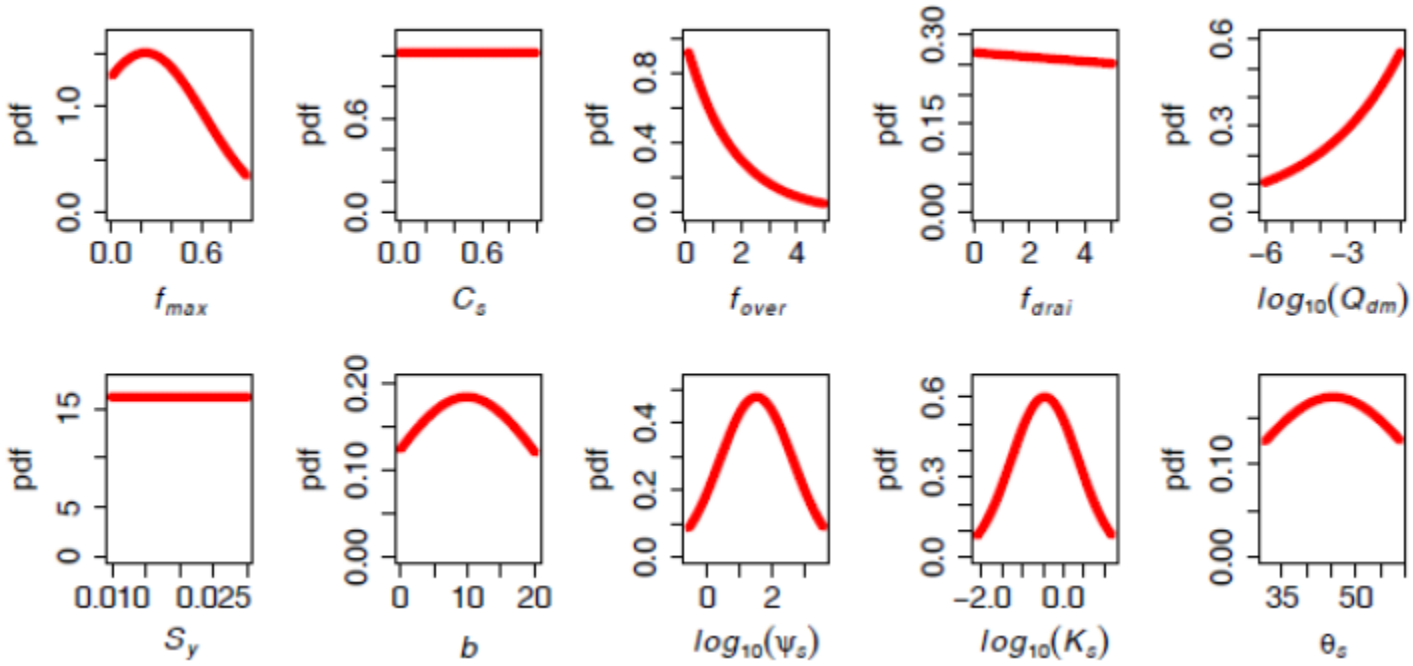
Output responses:
Latent heat fluxes (LH)
Sensible heat fluxes (SH)
Total runoff

Output statistics:
1. Natural variability of LH/SH/runoff
2. Propagation of uncertainty of input parameters

Multi-variate generalized linear model analysis and significance statistical tests:
1. Rank of significance of input parameters
2. Relationships between LH/SH/runoff and inputs

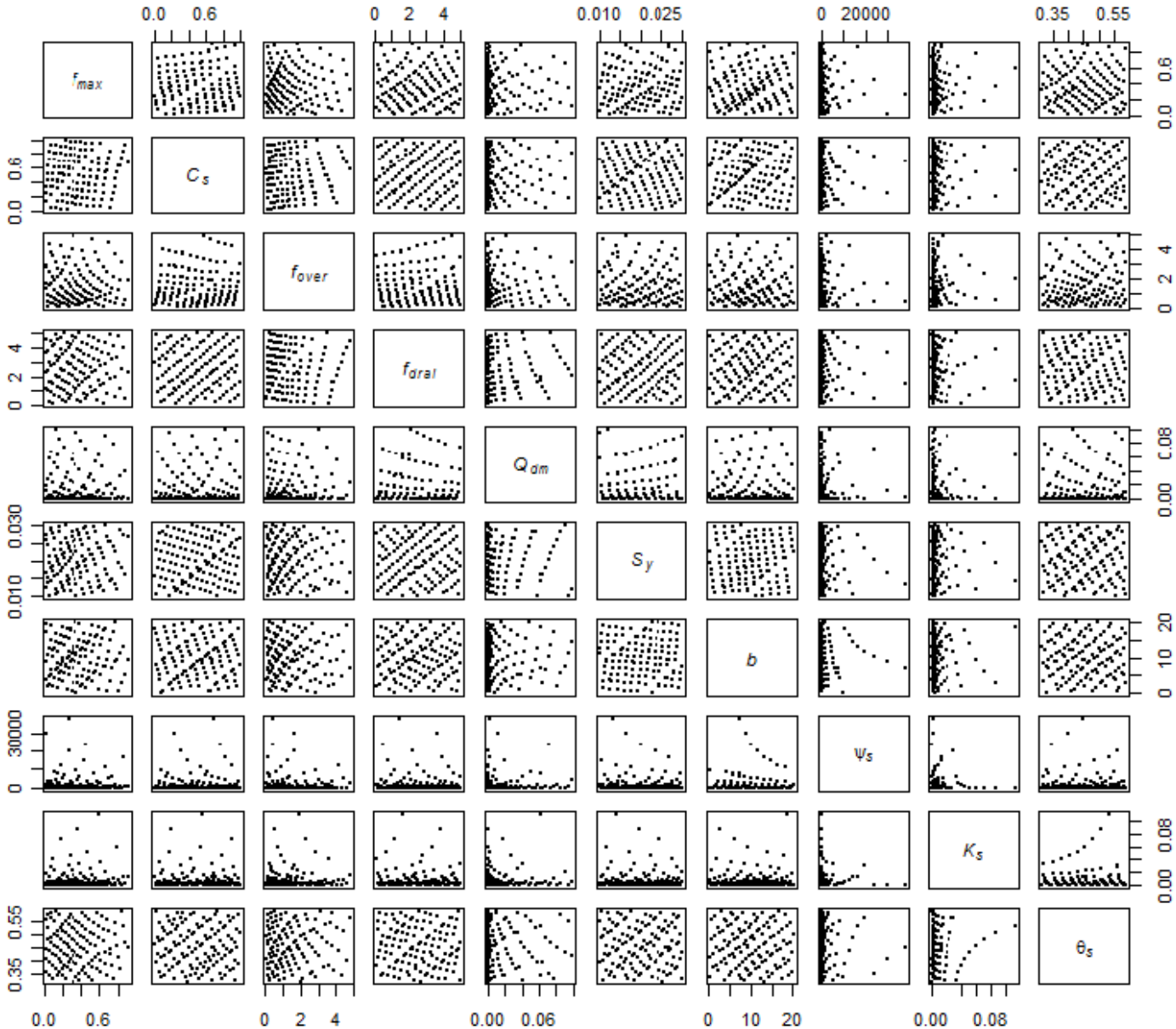
Model reduction
and
adaptive sampling

Application: UQ in Land Surface Modeling

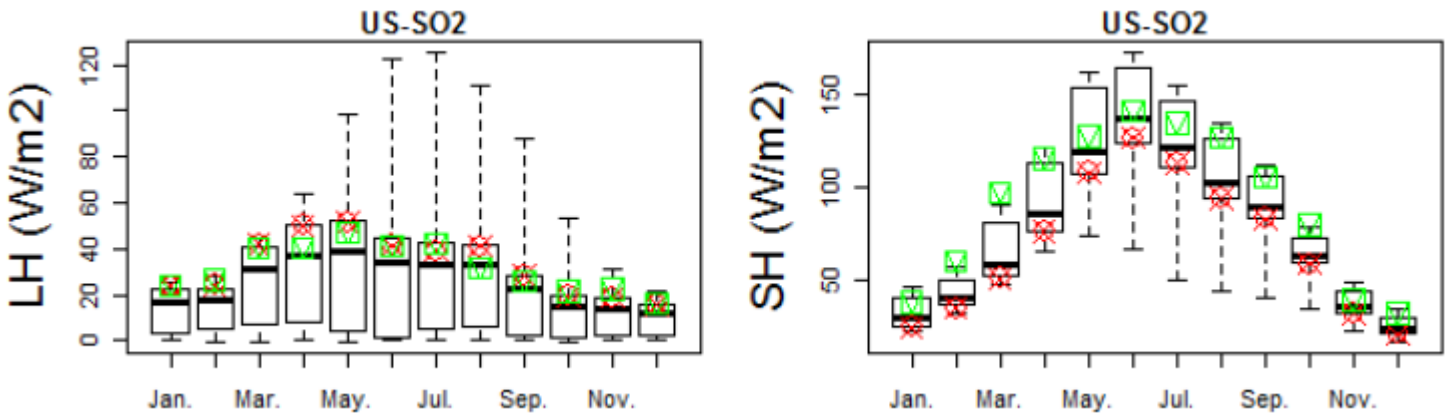


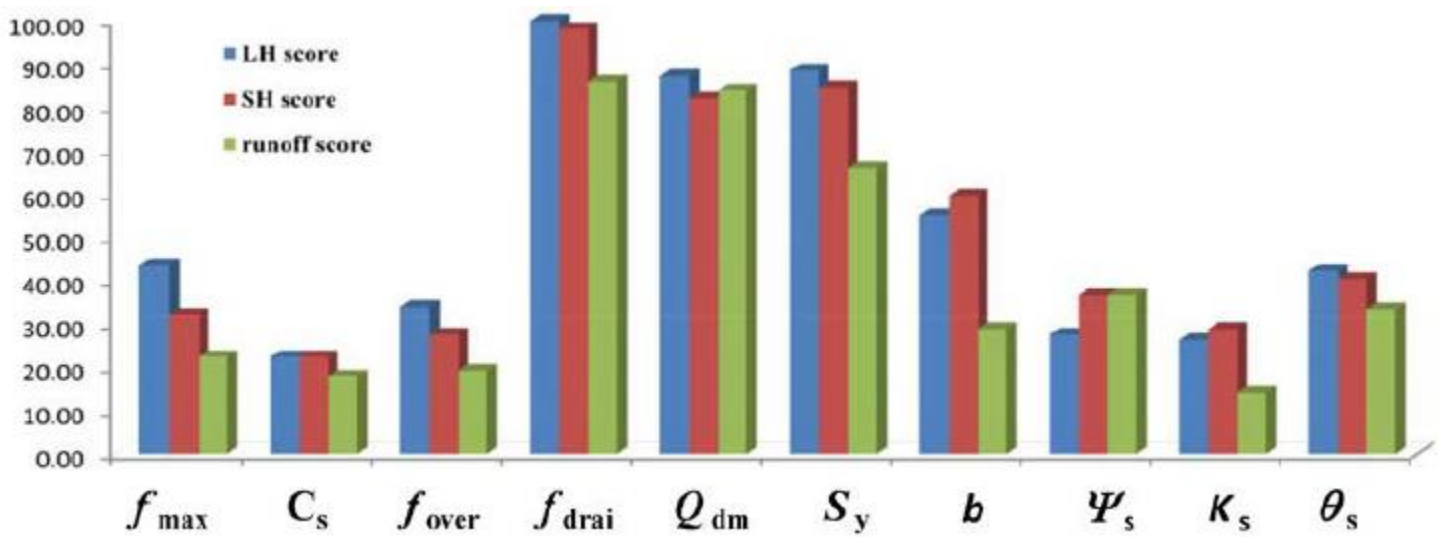
Application of UQ in Land Modeling

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle *Since 1965*



Application of UQ in Land Modeling





Sensitivity analyses

- ANOVA/MARS/Sobol's sensitivity index
- Model selection of nonlinear regression models
  - R^2 = FSS/TSS = 1 - RSS/TSS
  - Akaike Information Criterion

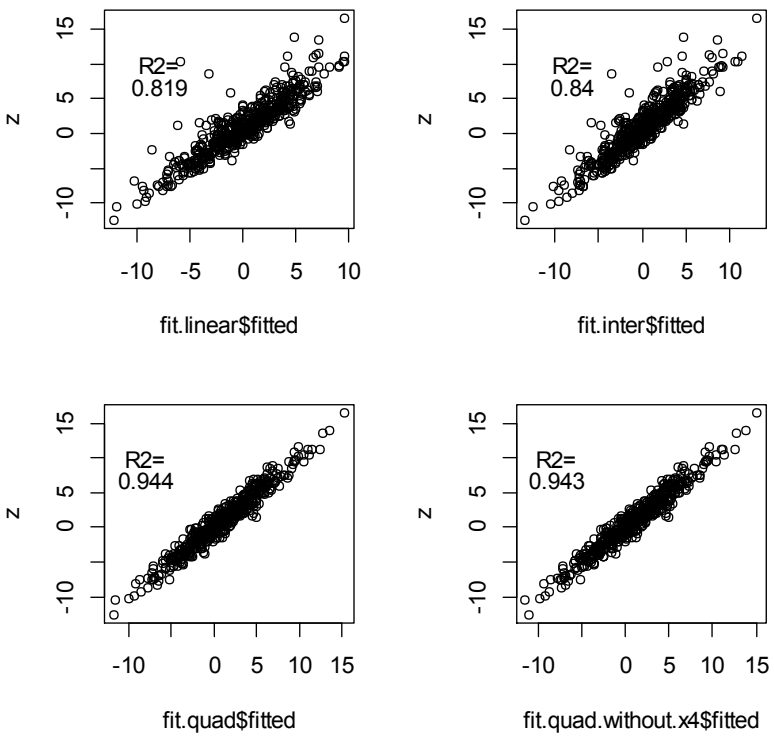
$$AIC = 2k - 2 * \ln L$$

AIC not only rewards goodness of fit but also includes a penalty that is an increasing function of the number of estimated parameters

  - Bayesian Information Criterion

$$BIC = -2 * \ln L + k \ln (n)$$

Prefers a model with fewer explanatory variables, or better fits, or both



Parameter Dimensionality Reduction

- Parameter dimensionality reduction can help reduce over-fitting and make inverse problems less ill-posed.
- Parameter dimensionality reduction by evaluating parameter identifiability, which may include
  - Observability

$$R_{o,i} = \frac{FSS_i}{TSS}$$

  - Differentiability

$$R_{d,i} = \frac{FSS_i}{FSS_i + \sum_j^{j \neq i} FSS_{ij}}$$

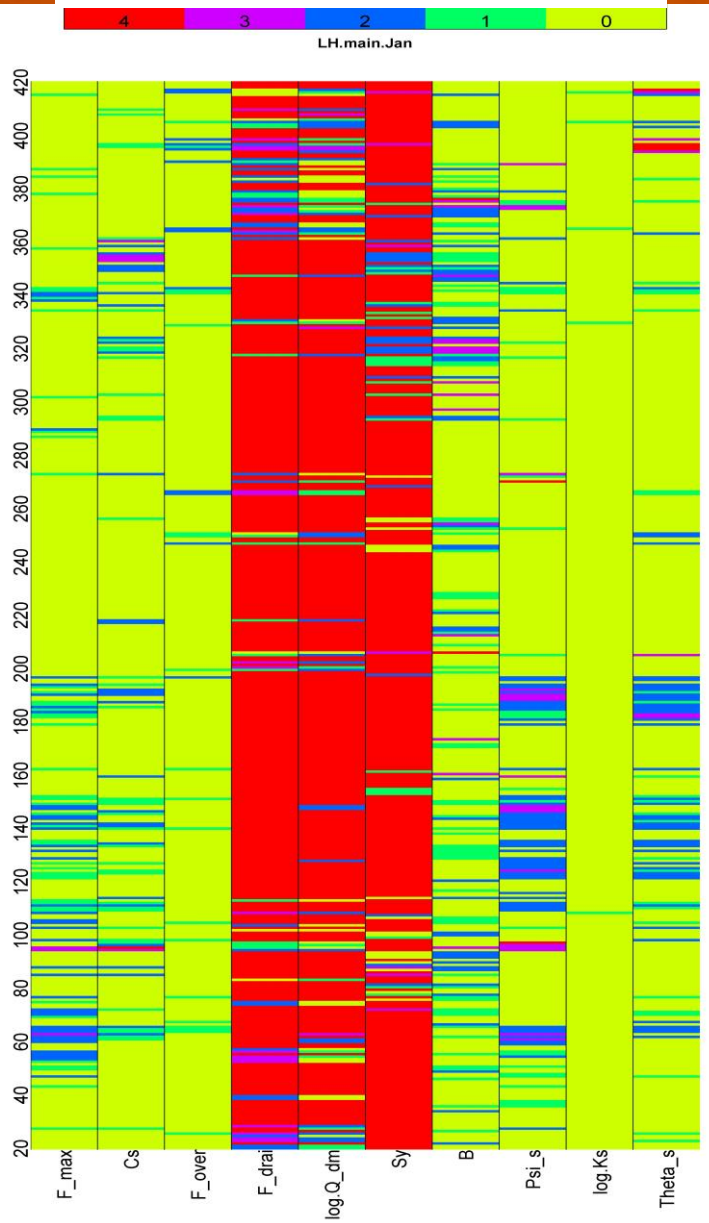
  - (non)-linearity/simplicity

$$R_{s,i} = \frac{FSS^{(1)}{}_i}{\sum_k FSS^{(k)}{}_i}$$

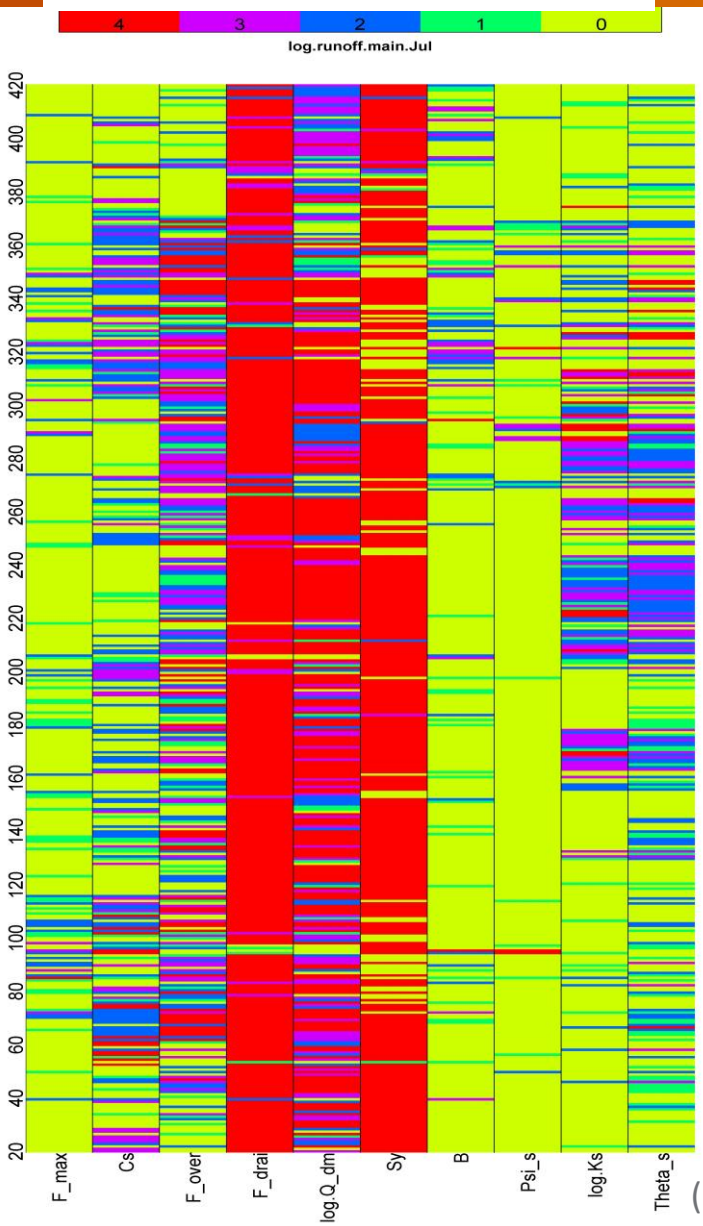
$FSS_i = \sum_k FSS^{(k)}{}_i$, k is the order of the explanatory variable p_i.

System classification

Sensitivity of LH

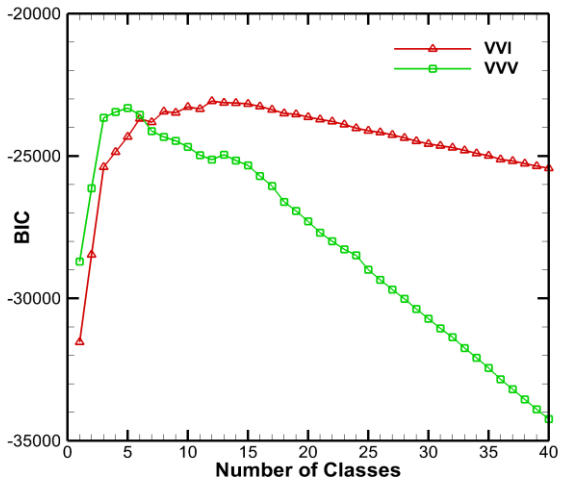


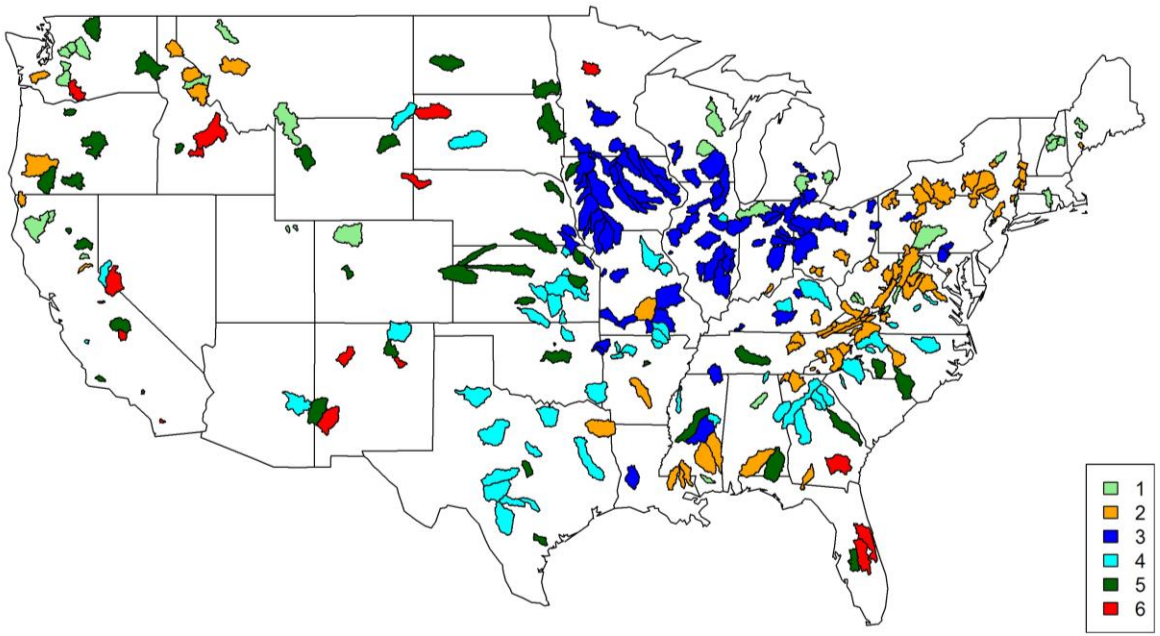
Sensitivity of runoff



(Ren et al 2015)

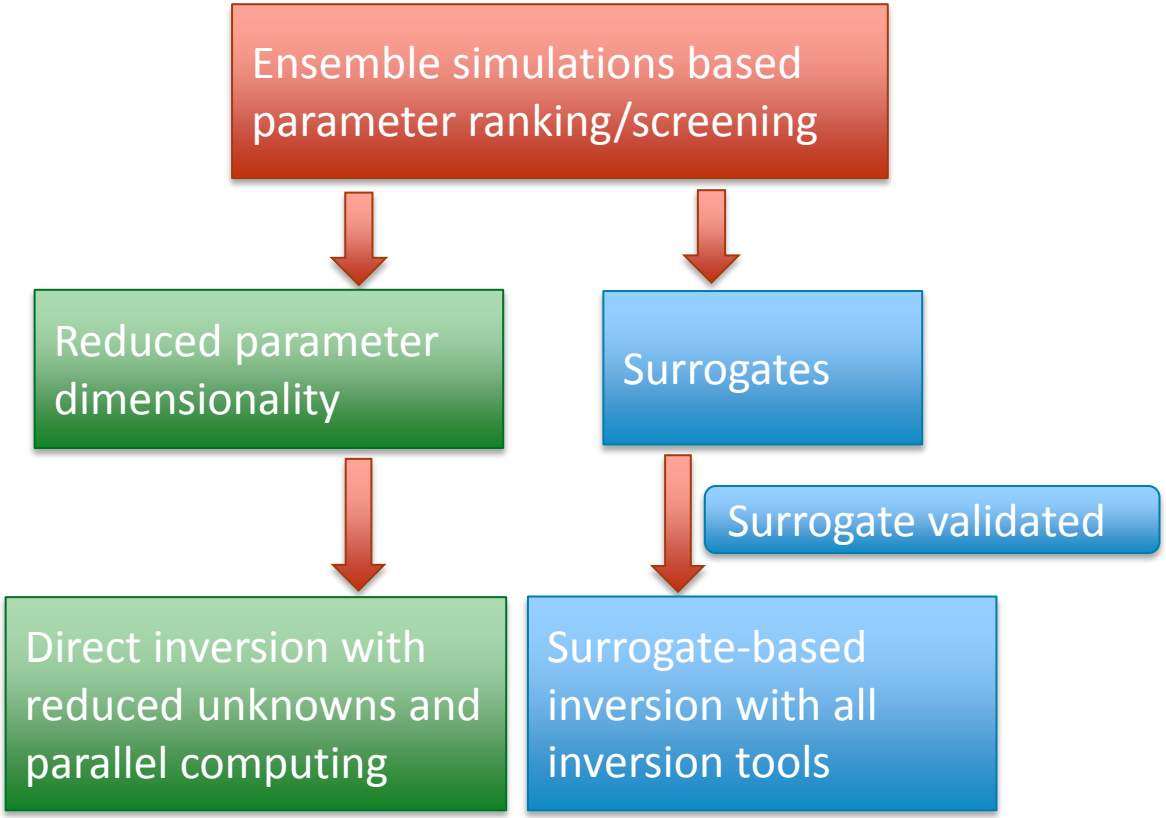
System complexity reduction





Parameter sensitivity analysis and classification show that we can group the watersheds into different classes, within each the parameter dimensionality can be reduced in the same way. And the reduced dimensionality makes the inverse problems less ill-posed.

Sampling the posterior -- inversion



Entropy-Bayesian Inversion

- Bayesian updating: $f(m|d,I) \propto f(d|m,I) * f(m|I)$
- Prior pdf

$$f_{\boldsymbol{M}|\,\boldsymbol{I}(\boldsymbol{m}|\boldsymbol{I})} = \prod_{j=1}^{p} \frac{\sqrt{\frac{\gamma_j}{\pi}} \exp\left[-\gamma_j (m_j + \frac{\beta_j}{2\gamma_j})^2\right]}{\Phi\left[\sqrt{2\gamma_j}\left(U_{j+} \frac{\beta_j}{2\gamma_j}\right)\right] - \Phi\left[\sqrt{2\gamma_j}\left(L_{j+} \frac{\beta_j}{2\gamma_j}\right)\right]}$$

- Likelihood function

$$f_{D|M,\Sigma,I(d^*|m,\sigma,I)} = \prod_{i=1}^{K} \prod_{j=1}^{N_i} \left[\frac{1}{\sqrt{2\pi}\sigma_{ij}} \exp\left\{ -\frac{1}{2\sigma_{ij}^2} \left[d_{ij}^* - g_{ij}(m) \right]^2 \right\} \right]$$

- Posterior pdf from entropy-Bayesian inversion

$$f_{\boldsymbol{M}|\boldsymbol{D},\boldsymbol{I}(\boldsymbol{m}|\boldsymbol{d}^*,\boldsymbol{I})} \propto \frac{f_{\boldsymbol{M}(\mathbf{m})} \prod_{i=1}^{K} \prod_{j=1}^{N_i} \left[\exp(-\frac{1}{2\sigma_{ij^2}} \left(d_{ij}^* - g_{ij}(m) \right)^2) \right]}{\int_m f_{\boldsymbol{M}}(\mathbf{m}) \prod_{i=1}^{K} \prod_{j=1}^{N_i} \left[\exp(-\frac{1}{2\sigma_{ij^2}} \left(d_{ij}^* - g_{ij}(m) \right)^2) \right] d^p \boldsymbol{m}}$$

MCMC-Bayesian Inversion

- Metropolis-Hasting sampling

a) Initialize a random vector $\boldsymbol{m}$ from the prior distributions $\{m_i^{(0)}, i = 1, \ldots, p\}$, where p is the number of parameters;

b) Generate a random variable $m_i^*, i = 1, \ldots, p$ from the proposal distributions, and calculate the following ratio (note the probabilities in the formula are calculated using equation 2):

$$\alpha = \min\left(p_{ra}, \frac{prob(m_i^* \mid m_1^{(1)}, m_2^{(1)}, \ldots, m_{i-1}^{(1)}, m_{i+1}^{(0)}, m_{i+2}^{(0)}, \ldots, m_p^{(0)})}{prob(m_i^{(0)} \mid m_1^{(1)}, m_2^{(1)}, \ldots, m_{i-1}^{(1)}, m_{i+1}^{(0)}, m_{i+2}^{(0)}, \ldots, m_p^{(0)})} \right),$$

where p_{ra} is reference acceptance probability;

c) Generate a random value u uniformly from interval (0,1);

d) If $\alpha > u$, let $m_i^{(1)} = m_i^*$; otherwise, let $m_i^{(1)} = m_i^{(0)}$.

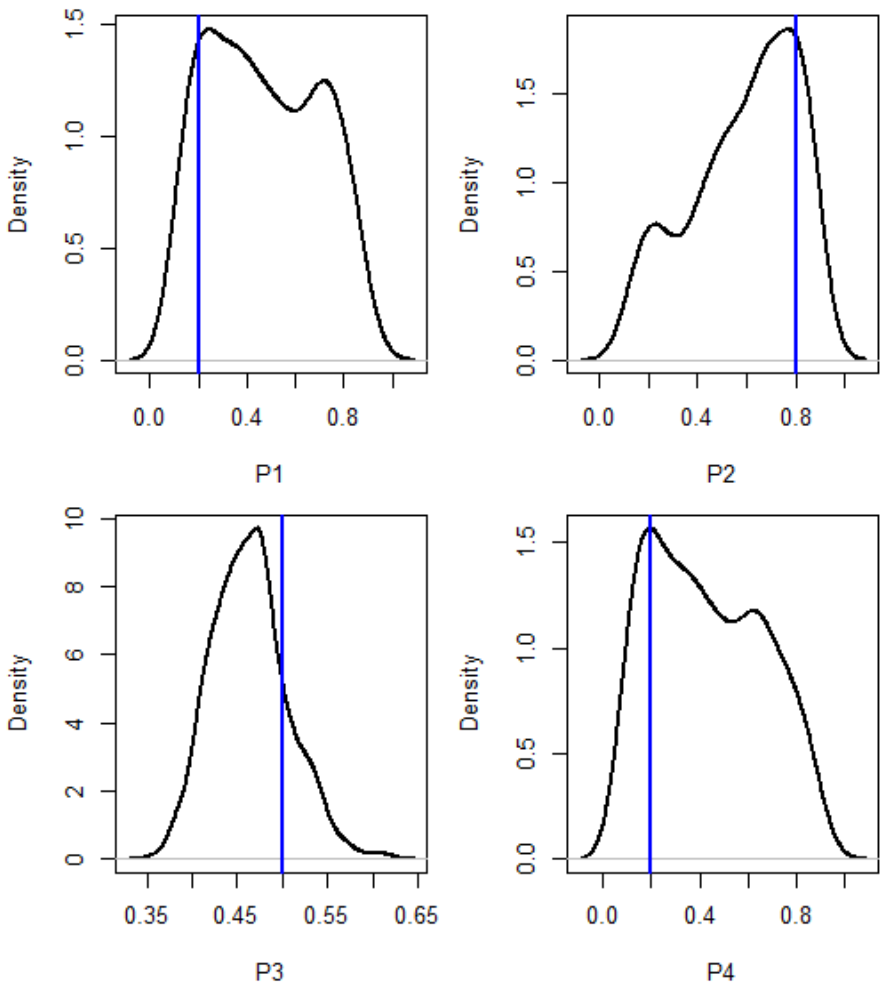
Repeating steps (b) to (d) by replacing index (k) with index (k+1), we can obtain many samples as follows: $\left\{(m_i^{(k)}) : i = 1, \ldots, p, \quad k = 0, 1, \cdots, n\right\}$, where n is the number of sample sets.

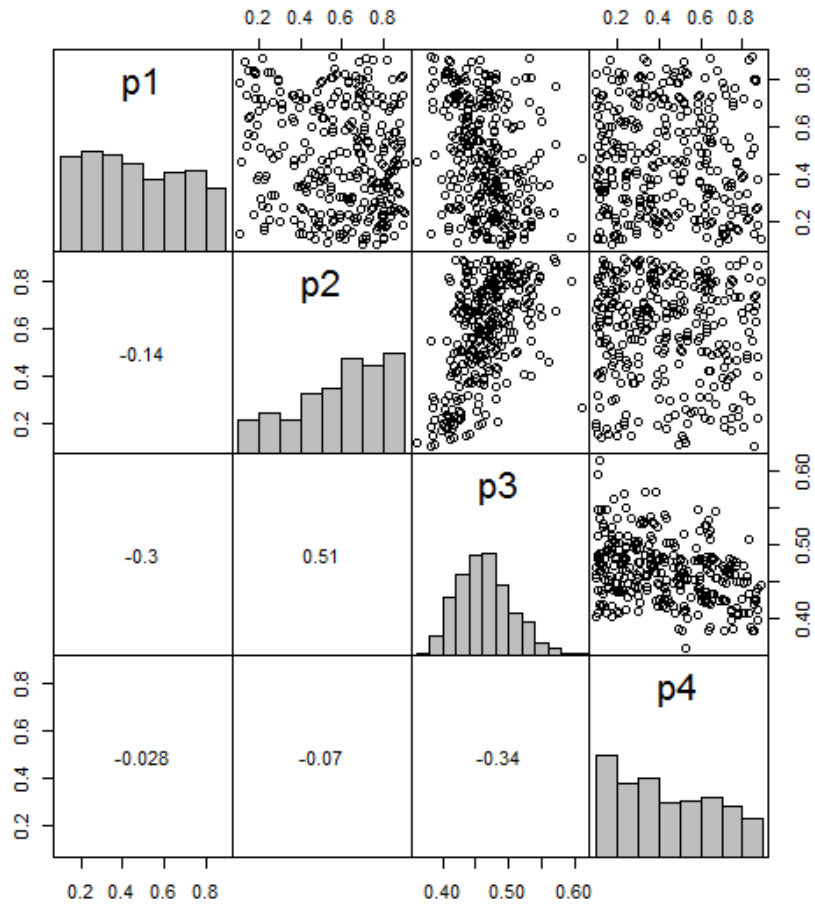
DR and AM

- Delaying rejection (DR)
  - a way of modifying the standard Metropolis-Hastings algorithm (MH) to improve efficiency of the resulting MCMC estimates
  - Upon rejection in a MH, a second stage move is proposed and therefore allows partial adaptation of the proposal within each step of the Markov chain.
- Adaptive Metropolis (AM)
  - Adaptively tuning the proposal distribution in a MH based on the past sample path of the chain.
- DRAM: combining the above two.

MCMC-Bayesian example

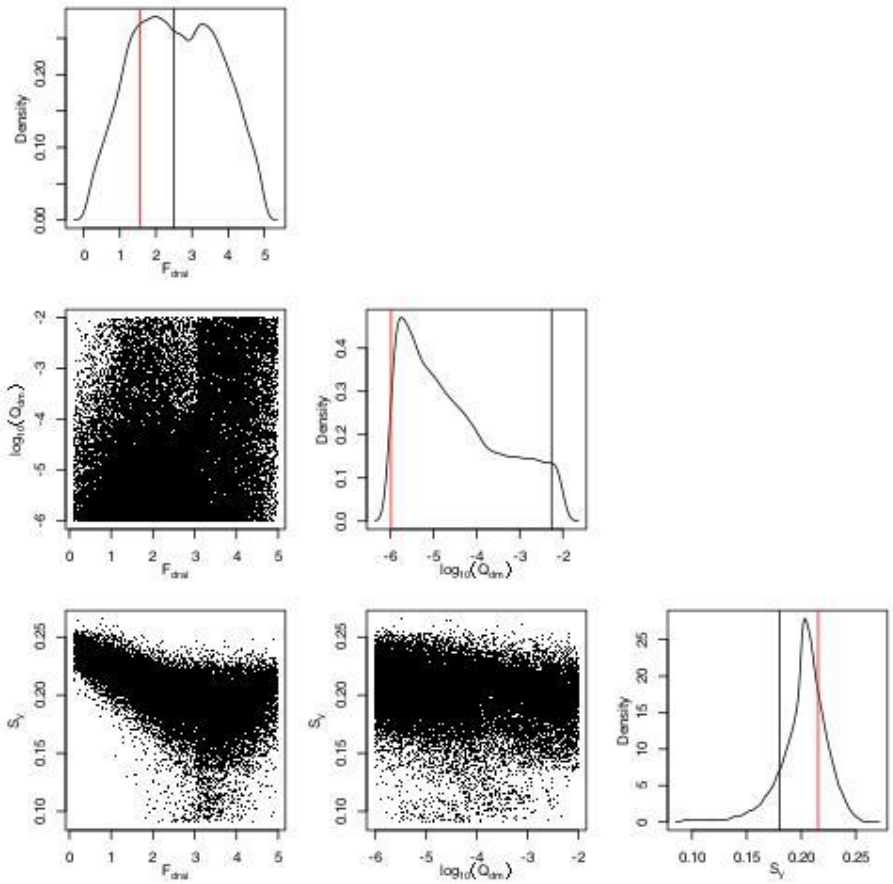
- An example problem using MCMC DRAM inversion





MCMC-Bayesian example

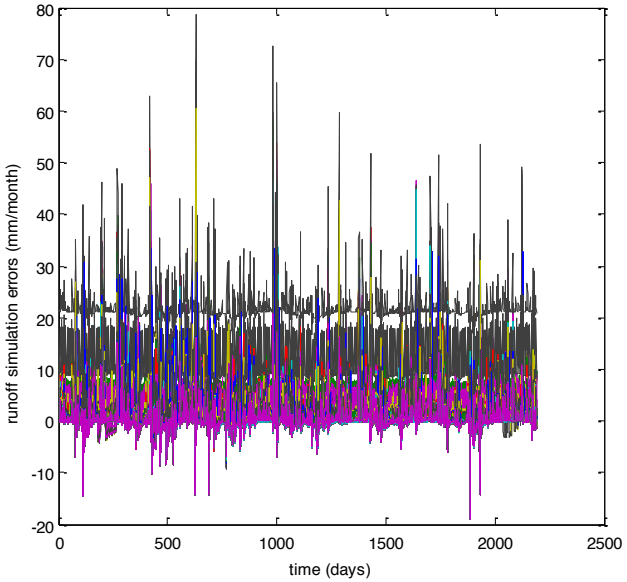
- MCMC-Bayesian inversion for estimating hydrological parameters using latent heat fluxes, runoff, and other metrics
- Ray et al 2015; Hou et al 2015; Huang et al 2015



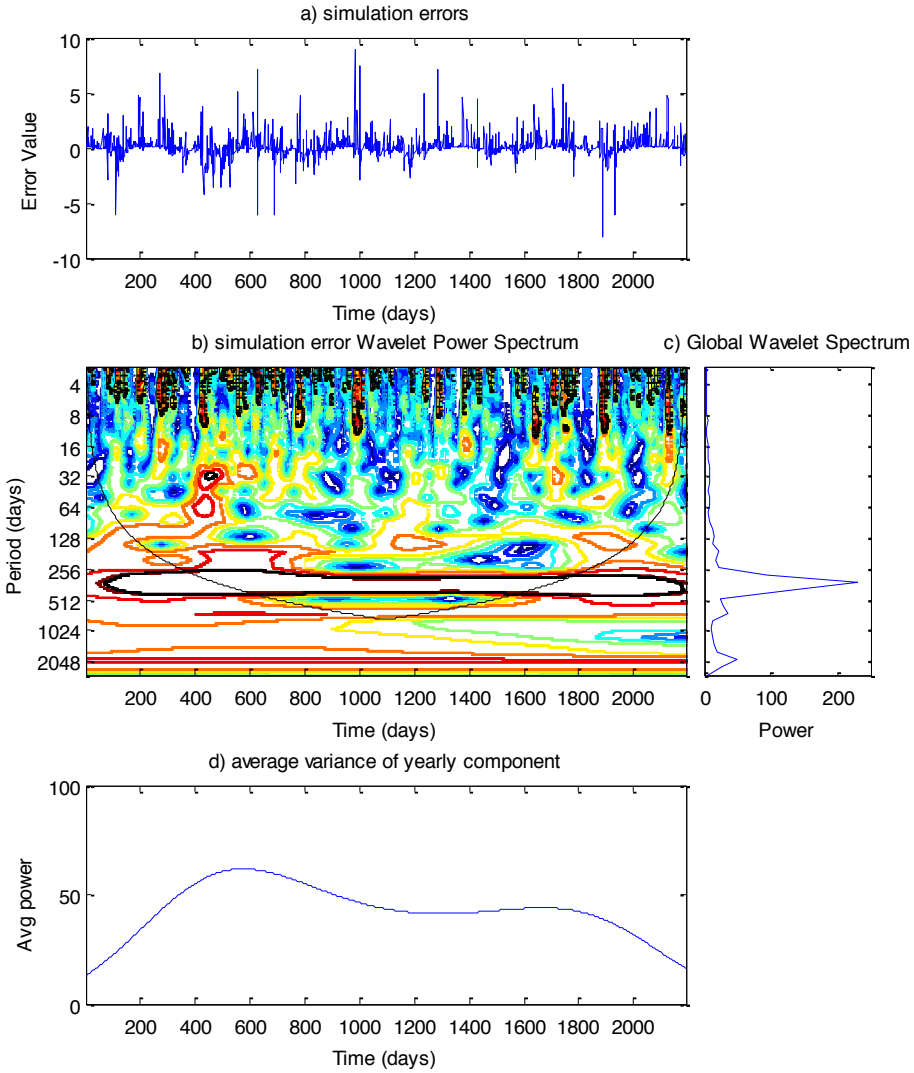
Time-frequency decomposition of simulation errors

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

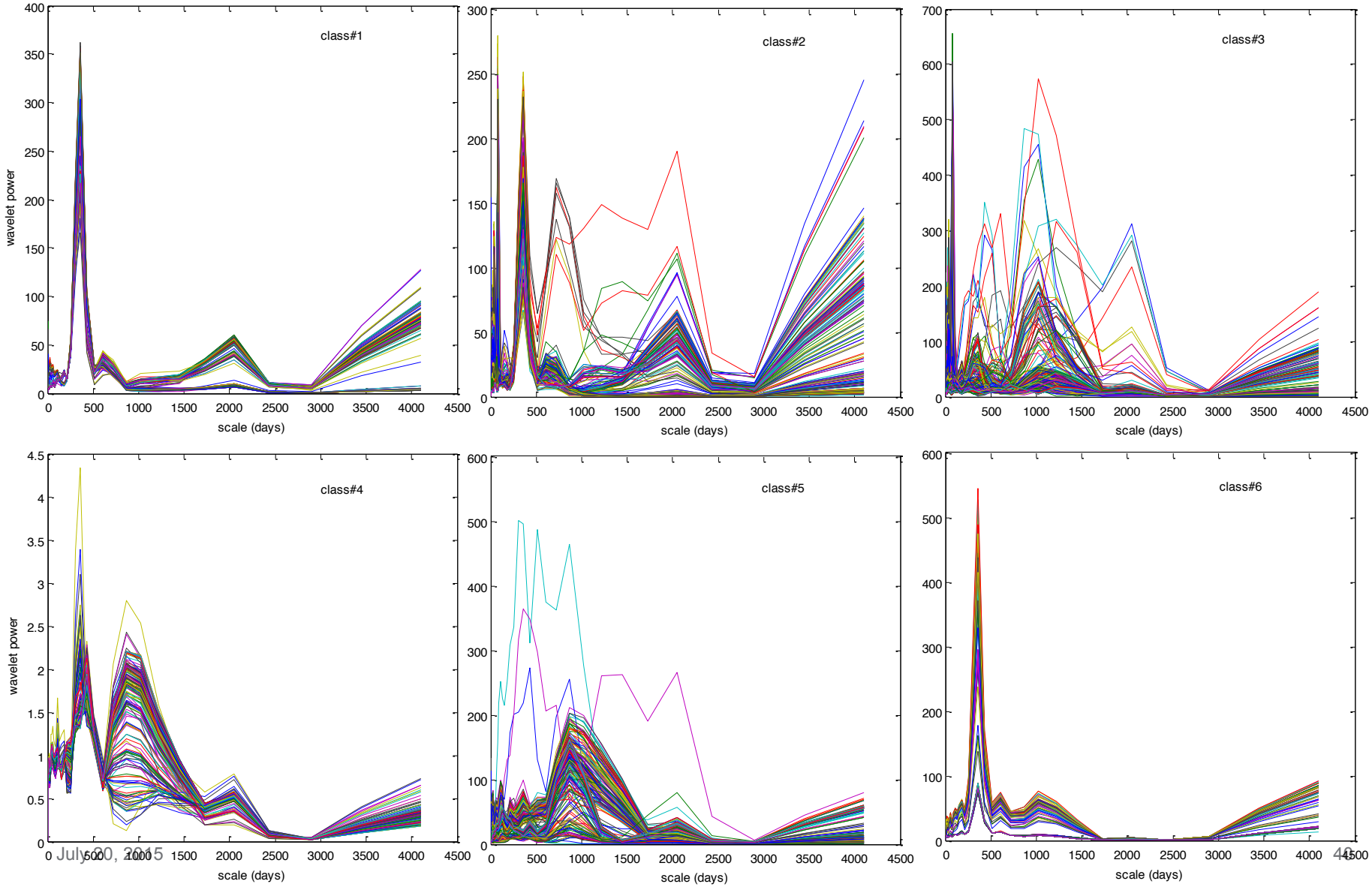
Model simulation errors



Wavelet analysis of simulation errors



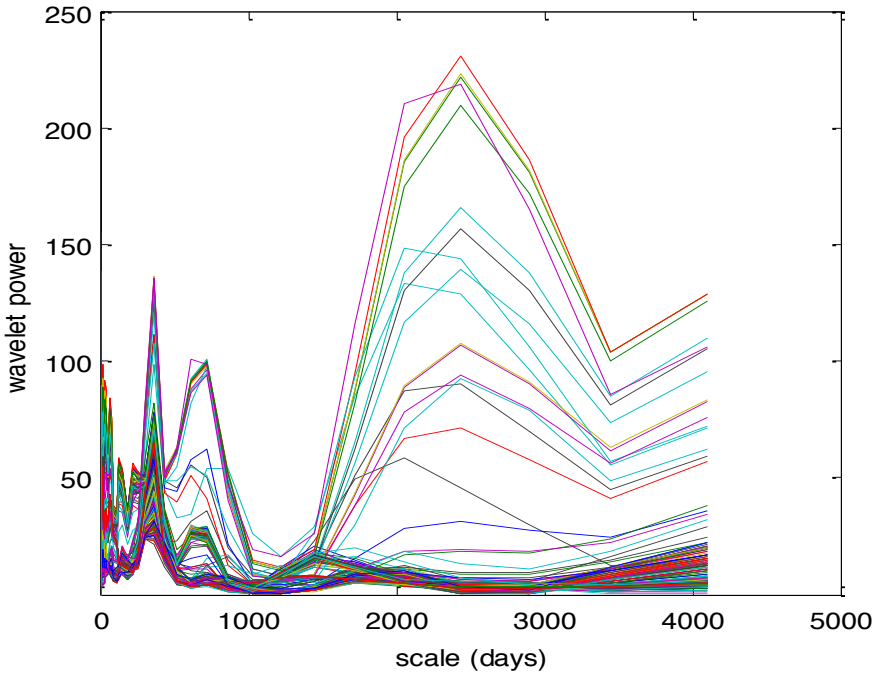
Time-frequency decomposition of ensemble errors



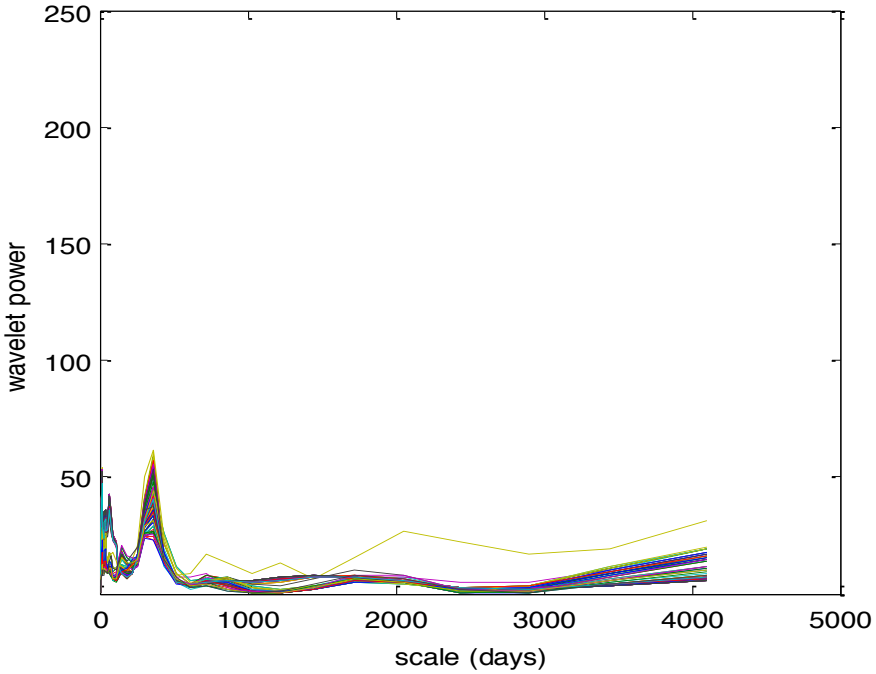
Decomposition of posterior ensemble errors

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Power spectrum of simulation errors before parameter inversion



Power spectrum of simulation errors after parameter inversion



Summary of time-frequency analysis of posterior errors

- It is possible to identify model structural errors by quantifying input parameter uncertainty and fully exploring the input parameter space.
- (Ensemble) model simulation errors provide information about the processes (and/or parameters) with major contributions to the errors.
- Assuming no systematic errors in the conceptual models and observational data, the model structural errors can possibly be separated after parametric uncertainty is reduced.
- The remaining errors would provide guidance on further model improvement, e.g., by modifying the physical models or parameterizations that numerically affect the errors at the major spatial-temporal scales.

MCMC-Bayesian Inversion

- Practical implementations of MCMC-Bayesian to complex earth system modeling (e.g., CLM)
  - Surrogate model validation
  - Filtering of unreasonable parameter sets with support vector machine (SVM)
  - Kriged model for discrepancy
  - Correlated errors/misfits
  - Scalable adaptive chain ensemble sampling (SAChES)

MCMC-Bayesian Inversion

- Demonstrated using various classic inverse problems as well as CLM, oil/gas exploration, and environmental remediation problems
- Evaluated the sensitivity of inversion results to initial values of parameter, data resolution, data worth/redundancy, climate and soil conditions, acceptance probability, width of proposal distributions, multi-chain calculations, and Bayesian model averaging

Software to be released

- Other than the standard packages such as LHS/QMC/ANOVA/MARS, our unique capabilities include:
  - Prior pdf derivation and sampling software (MREQMC)
  - Scalable calibration enabling multi-chain parallel computing (SAChES)

Summary

- Parameter dimensionality can be reduced through parameter screening based on efficient sampling, response surface analysis
- Complex system can be classified into simpler ones by clustering based on parameter identifiability/transferability.
- Reduced dimensionality helps to evaluate impacts of most significant parameters/factors, and make inverse/calibration problems less ill-posed.
- Scalable chain-ensemble sampling has the potential to calibrate complex climate models that are computationally demanding

Discussion topics

- Importance of and approaches for stochastically representing uncertainties
- Fitting probability density functions (pdfs) with direct measurements
- Derivation of probabilistic distributions in explicit forms
- Statistical tests on checking necessity of parameter transformations
- Effects of reliable pdfs on surrogate development, SA, and calibration
- Efficient sampling of joint pdfs of hierarchical mixed parameters or factors

Discussion topics cont'd

- Pdf provides a better measure of uncertainty compared to other summary statistics
- Perturbation method or one-factor-a-time sensitivity analysis usually is not meaningful
- Compared to ANOVA/MARS, random forest approach can be used to develop localized surrogates, which might require fewer numerical simulations to achieve reliable response surfaces
- Adaptive sampling (response surface gradient –based) is another way to reduce required numerical simulations

Questions? mailto:zhangshuan.hou@pnnl.gov