Methods for uncertainty quantification: A statistical perspective

Stephan R. Sain

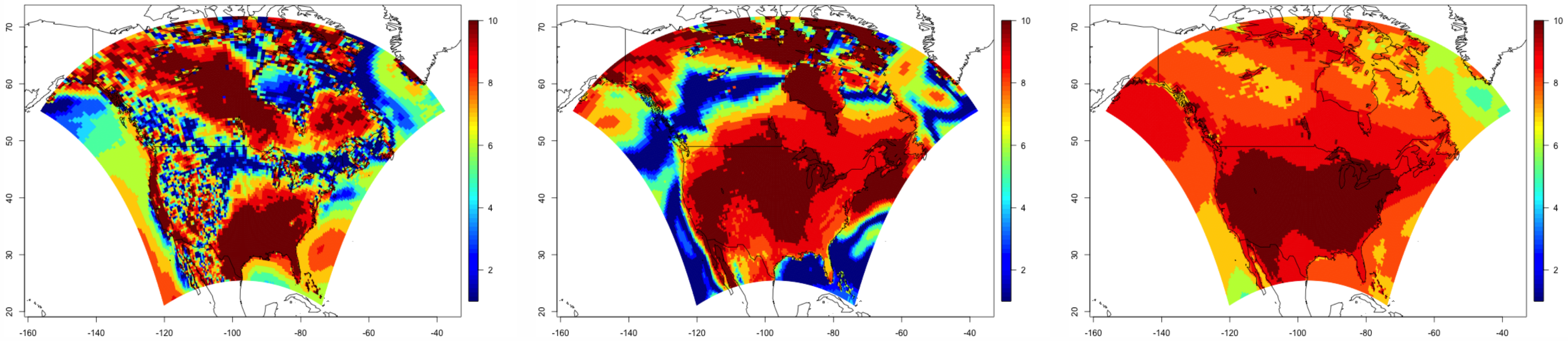
Geophysical Statistics Project
Institute for Mathematics Applied to Geosciences
National Center for Atmospheric Research
Boulder, CO

Uncertainty in Climate Change Research; Boulder CO; August 8, 2012

Computational & Information Systems Laboratory

Outline

- What is UQ?
- Overview of DACE – design, emulators, etc.
- Bayesian analysis of ensembles – pdfs
- ANOVA



What is UQ?

Uncertainty Quantification: The process of quantifying uncertainties associated with model calculations of true, physical quantities of interest (QOIs), with the goals of accounting for all sources of uncertainty and quantifying the contributions of specific sources to the overall uncertainty.

. . . *"quantifying uncertainty"* in a prediction for a QOI means making a quantitative statement about the values that the QOI for the physical system may take, often in a new, unobserved setting. The statement could take the form of a bounding interval, a confidence interval, or a probability distribution, perhaps accompanied by an assessment of confidence in the statement.

- National Research Council (2012), *Assessing the Reliability of Complex Models: Mathematical and Statistical Foundations of Verification, Validation, and Uncertainty Quantification*, Washington, D.C.: The National Academies Press.

Design and analysis of computer experiments (DACE):

- The *forward* problem:
  - Given some input values, run the model to obtain outputs.
  - Goals include exploration of input space or features of outputs, sensitivity analysis, prediction, etc.
  - Statistical issues: choice of inputs (design), propagation of uncertainty, emulators, data assimilation, etc.
- The *inverse* problem:
  - Given some input values, run the model to obtain outputs.
  - Goals focus on finding good values of inputs based on comparing outputs to observations.
  - Statistical issues: choice of inputs (design), emulators, calibration, etc.

Design of Experiments

Five important questions:

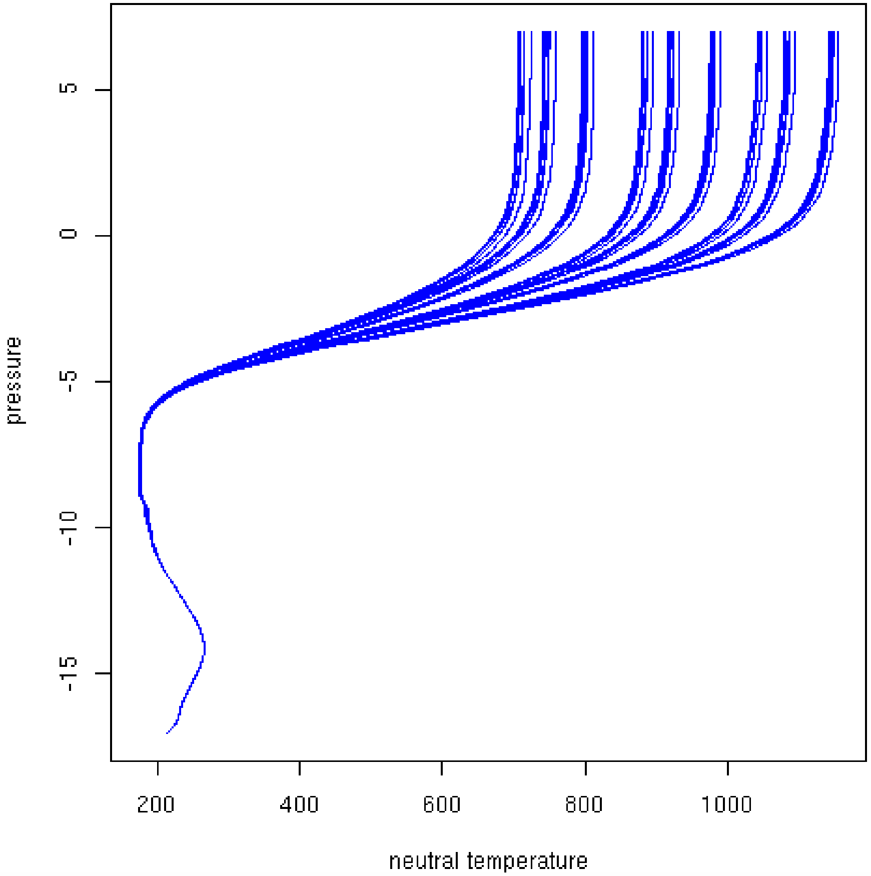
1. How much money (resources) do you have?
2. What do you want to know?
3. How much money (resources) do you have?
4. What do you know?
5. How much money (resources) do you have?

Goal: Given a fixed amount of resources, construct an experiment that maximizes the amount of information that can be obtained.

Example 1: Full factorial experiment

Thermosphere Ionosphere Electrodynamic General Circulation Model (TIEGCM)

- A 1-D global mean model has been used over the years to test parameterizations, chemistry and dynamics.
- Outputs are a function of pressure (focus on neutral temperature).
- A simple experiment with four factors, each at three levels: Solar flux, joule heating, characteristic energy, auroral production factor
- All possible combinations of each factor were run: $3 \times 3 \times 3 \times 3 = 81$ total model runs.



Example 2: Space filling design

Lyon-Fedder-Mobary (LFM) model of the magnetosphere

- LFM uses ideal magnetohydrodynamics (MHD) equations to model the interaction between the solar wind, the magnetosphere, and the ionosphere.
- Outputs are spatial-temporal fields (e.g., energy of electrons).
- Statistical approach to calibration for three unknown parameters.
- A space filling design of 20 runs was used.

(lfmres.mov)

Example 3: A fractional factorial

North American Regional Climate Change Assessment Program (NARCCAP)

	Phase I	Phase II			
	NCEP	GFDL	CGCM3	HADCM3	CCSM
CRCM	finished		finished		finished
ECP2	finished	finished		running	
HRM3	finished	finished		finished	
MM5I	finished			running	finished
RCM3	finished	finished	finished		
WRFG	finished		finished		finished

Emulators

A statistical view of a computer model:

$$\mathbf{Y} = f(\mathbf{x})$$

- $\mathbf{x}$ – model inputs (possibly multivariate and/or functional)
- $\mathbf{Y}$ – model outputs (possibly multivariate and/or functional)
- f – a black-box function that maps the inputs onto the outputs

An emulator is an approximation for f based on a collection of $\{\mathbf{x}_i, \mathbf{Y}_i\}$:

- regression models (i.e., multiple regression, splines, trees, etc.)
- Gaussian processes
- Neural networks and other machine learning methods
- And others...

Gaussian process: let the random function $Y(\mathbf{x})$ be a Gaussian process if, for any finite collection of $\mathbf{x}_1, \ldots, \mathbf{x}_n$, the vector $(y(\mathbf{x}_1), \ldots, y(\mathbf{x}_n))'$ has a multivariate normal distribution.

- Must have a mean function (e.g., $\mu(\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}$) and a (stationary, isotropic) covariance function (e.g., Matérn)

From the assumption of a Gaussian process, the joint distribution of model output at the observed model inputs and a new $\mathbf{x}_0$ is given by

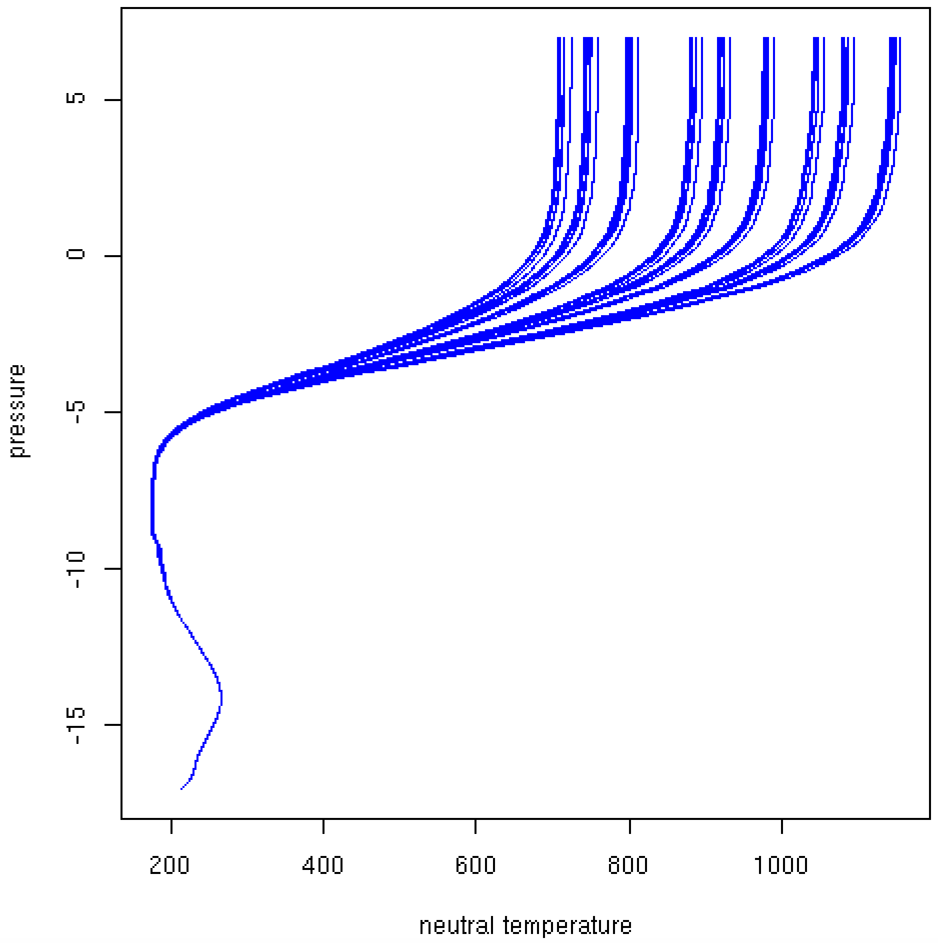
$$\begin{pmatrix} \mathbf{Y} \\ y(\mathbf{x}_0) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{\mu} \\ \mu(\mathbf{x}_0) \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma} & \mathbf{c} \\ \mathbf{c}' & \sigma^2 \end{pmatrix}\right).$$

and the emulator $(\widehat{f})$ is then a conditional Gaussian distribution with mean and variance

$$\begin{aligned} E[y(\mathbf{x}_0)|\mathbf{Y}] &= \mu(\mathbf{x}_0) + \mathbf{c}'\boldsymbol{\Sigma}^{-1}(\mathbf{Y} - \boldsymbol{\mu}) \\ \mathrm{Var}[y(\mathbf{x}_0)|\mathbf{Y}] &= \sigma^2 - \mathbf{c}'\boldsymbol{\Sigma}^{-1}\mathbf{c} \end{aligned}$$

Building an emulator: A case study

TIEGCM Neutral Temperature



Given 80 runs, can we predict the output for the 81st?

A Simple Model

- Expand each curve as a linear combination of basis functions:

$$f_i(x) = \sum_k \beta_{ik} \phi_k(x)$$

  where

  - $\{\phi_k\}$ are known, fixed functions
  - $\{\beta_{ik}\}$ are coefficients related to the ith factor-level combination.

EOFs

- Q: How to choose the basis functions?

- A: Empirical orthogonal functions (principal components)!

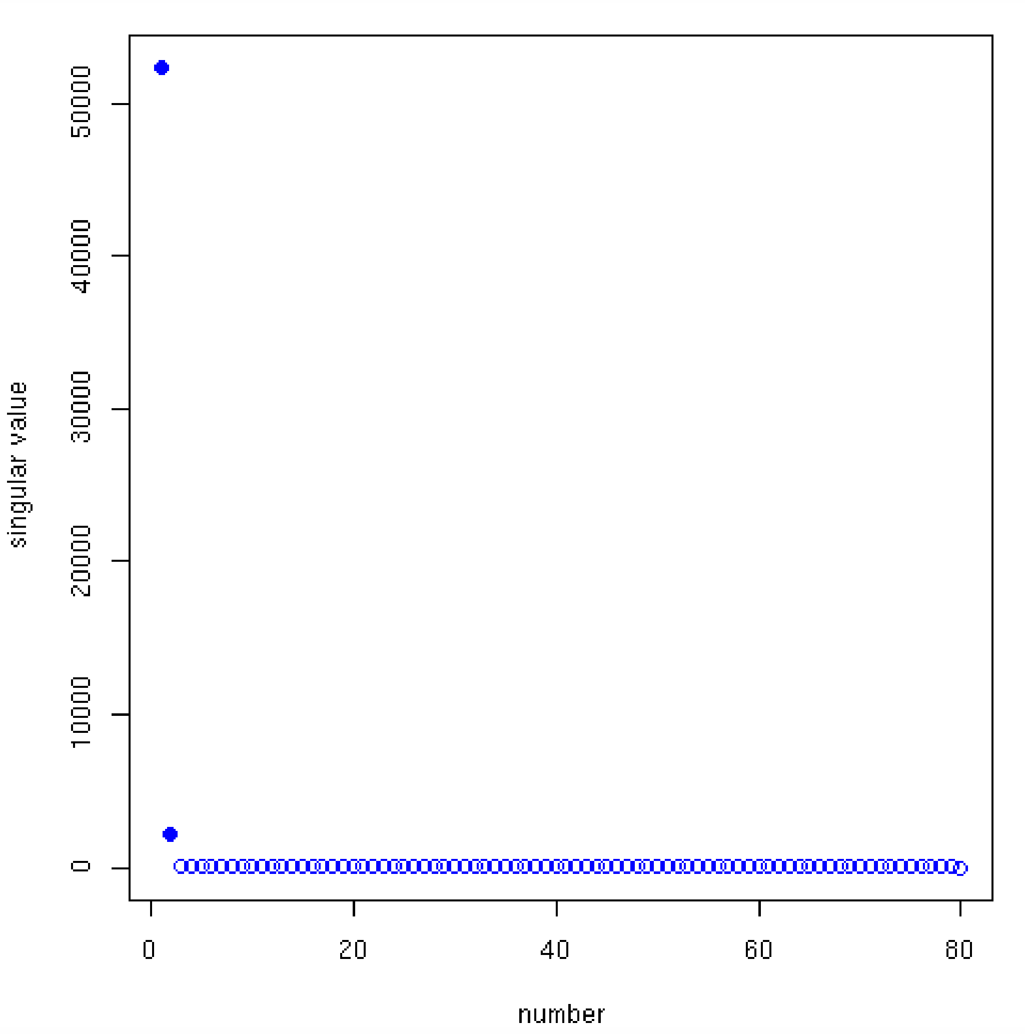
- Let Y denote the 97 × 80 data matrix and write

$$Y = UDV' = \sum_i d_i u_i v_i'$$

  - 80 factor-level combinations and 97 pressure levels

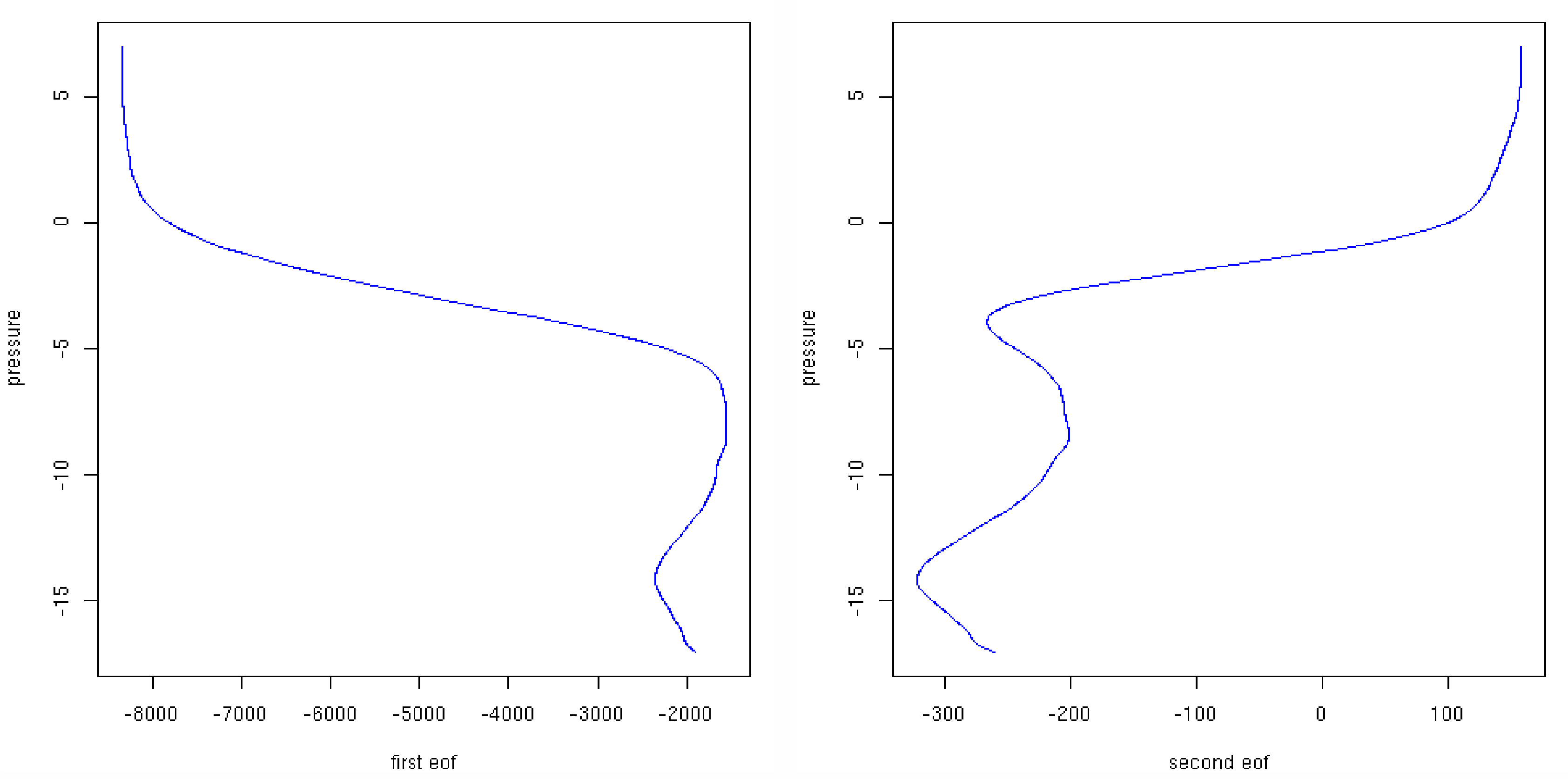
- Choose the first few dominate empirical orthogonal functions ($d_i u_i,\ i = 1, \ldots, m$) and parameterize the coefficients ($v_i,\ i = 1, \ldots, m$) through the experimental conditions.

EOFs – Singular Values



- Almost all of the variation is summarized in the first 1 or 2 principal components.

EOFs


5
0
-5
-10
-15
pressure
-8000
-7000
-6000
-5000
-4000
-3000
-2000
first eof
5
0
-5
-10
-15
pressure
-300
-200
-100
0
100
second eof

The statistical emulator

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \\ \vdots \\ Y_{80} \end{bmatrix} = \left(\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ & & & & \vdots & & & & \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{bmatrix} \otimes Z \right) \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_8 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \\ \vdots \\ \epsilon_{80} \end{bmatrix}$$

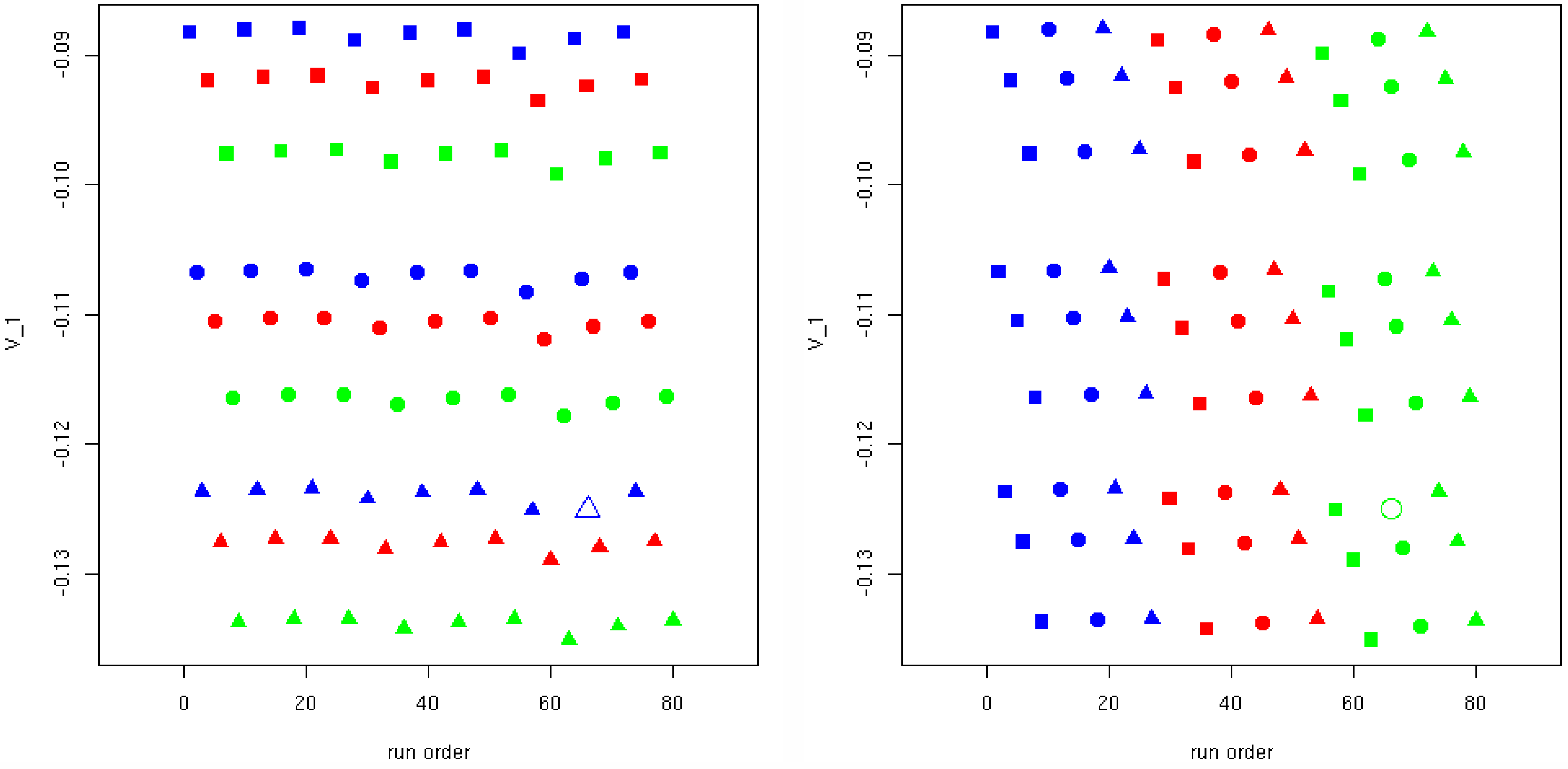
- Y_i is a 97-vector

- Z is a 97×2 matrix containing the EOFs

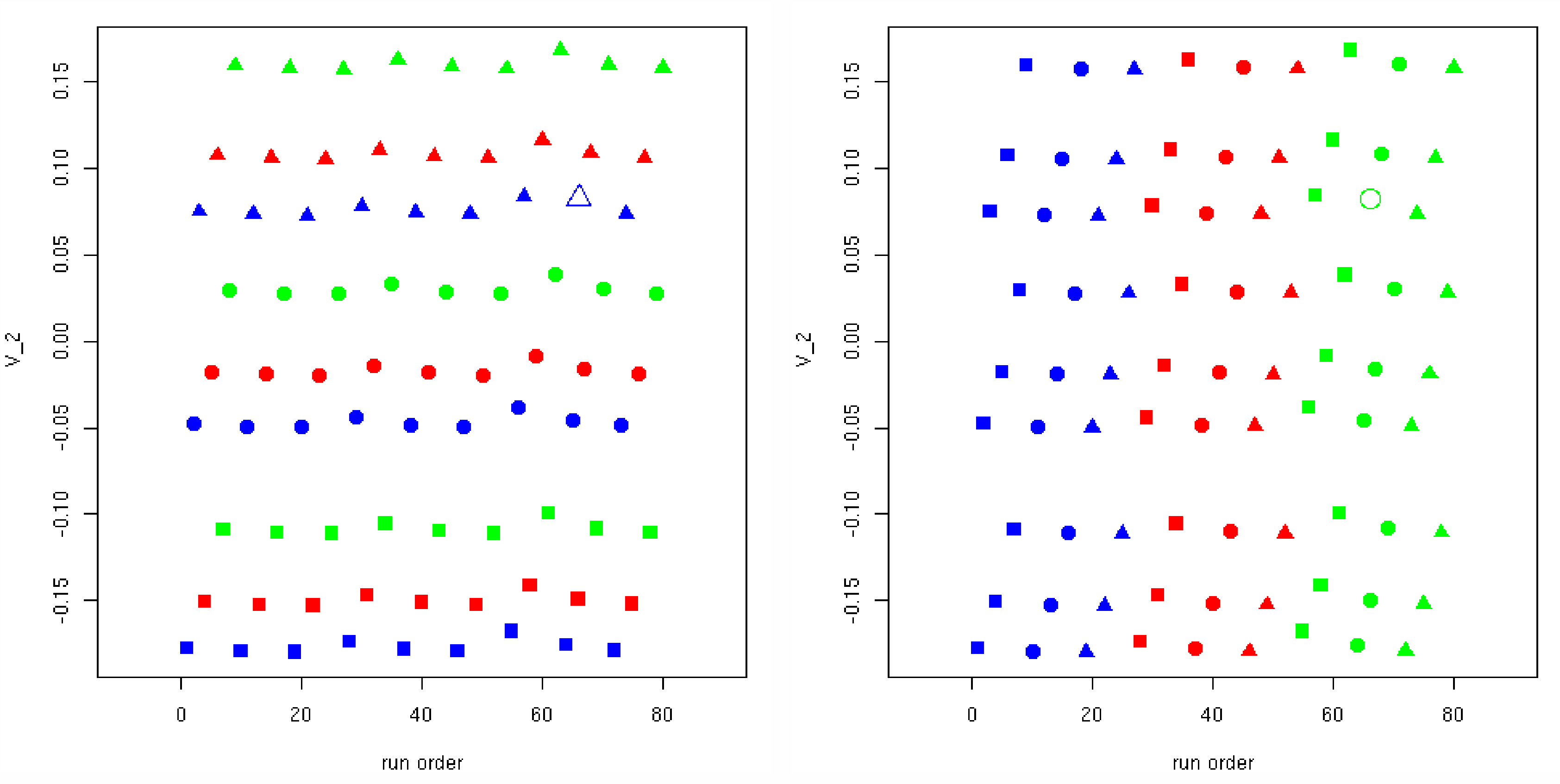
- β_i is a 2-vector

- Additive model (no interactions):
  - All main effects “significant”
  - 1st/2nd EOF: solar flux most important, followed by joule heating, characteristic energy, and auroral production factor
- Second-order interactions:
  - First EOF
    - All main effects important
    - Solar flux/joule heating and characteristic energy/auroral production factor
  - Second EOF
    - Characteristic energy not as important
    - Solar flux/joule heating and characteristic energy/auroral production factor

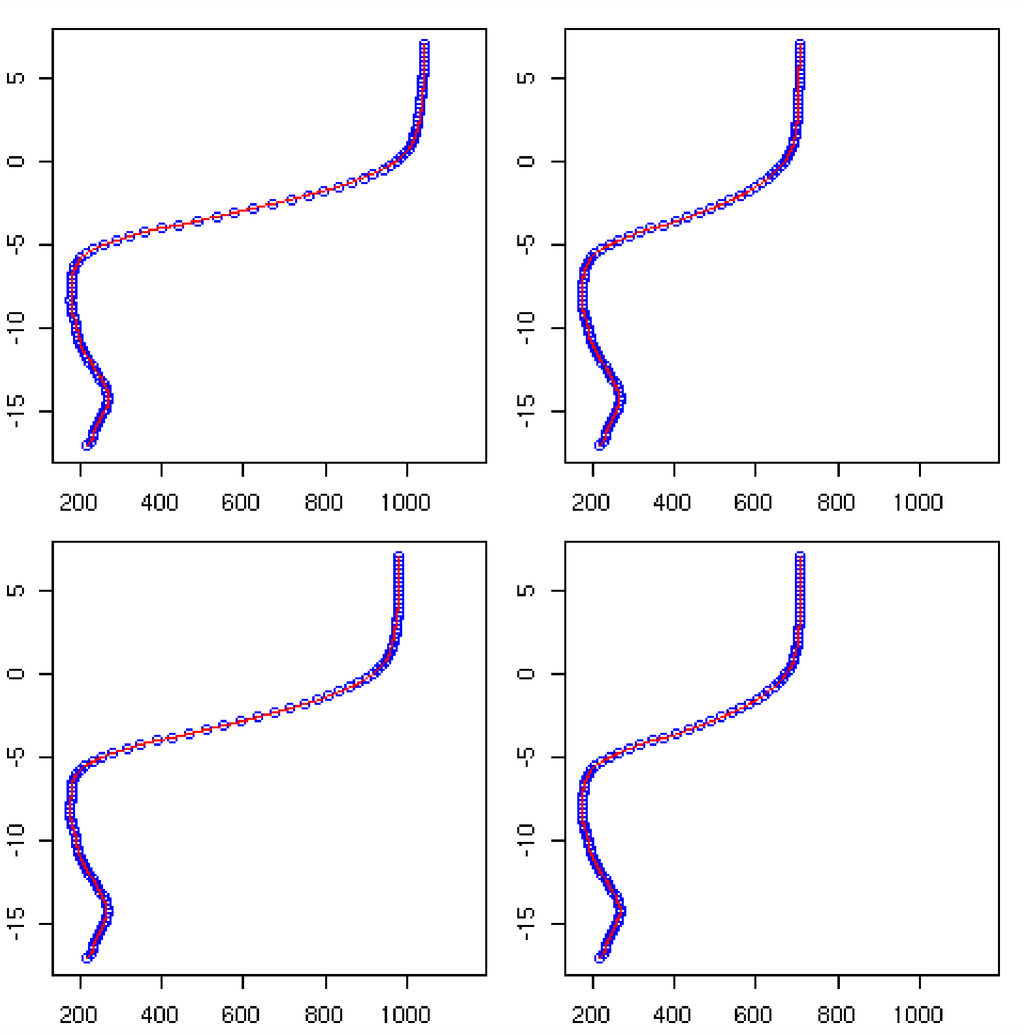
Estimated Coefficients



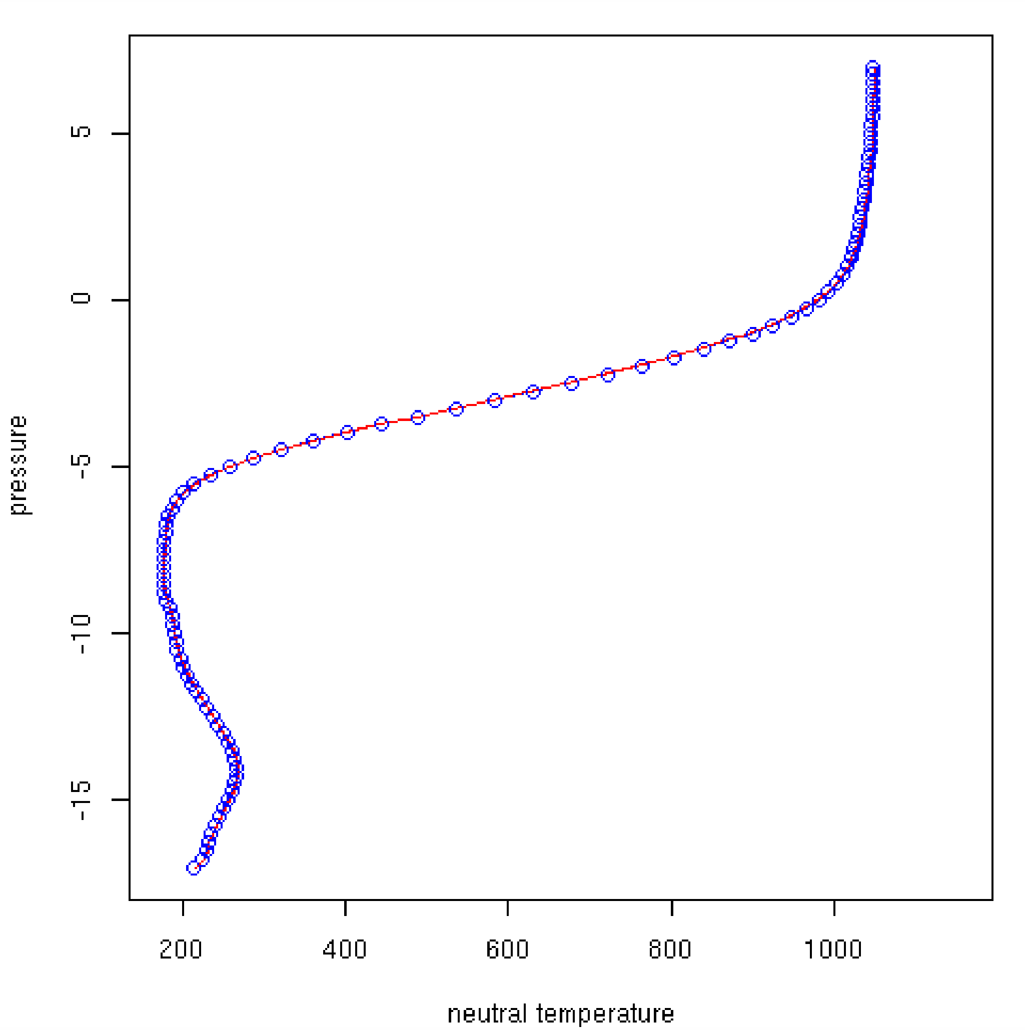
Estimated Coefficients



Cross-validation



The Mystery Run



The Bayesian Paradigm

Postulate a model (pdf) for data that depends on some parameters:

$$Y_1, \ldots, Y_n \sim \pi(Y_1, \ldots, Y_n | \theta).$$

$\Rightarrow$ This forms the *likelihood*.

Postulate a model (pdf) for the parameters:

$$\theta \sim \pi(\theta)$$

$\Rightarrow$ This forms the *prior*.

Inference follows by examining the posterior distribution:

$$\begin{aligned} \pi(\theta | Y_1, \ldots, Y_n) &\propto \pi(Y_1, \ldots, Y_n | \theta) \pi(\theta) \\ \text{posterior} &\propto \text{likelihood} \times \text{prior} \end{aligned}$$

$\Rightarrow$ From *Bayes' Theorem*.

A Simple Model

$$Y_1, \ldots, Y_n \sim \mathcal{N}\left(\mu, \sigma^2\right)$$

$$\begin{aligned} \mu|\sigma^2 &\sim \mathcal{N}\left(\mu_0, \sigma^2/\kappa_0\right) \\ \sigma^2 &\sim \text{Inv} - \chi^2\left(\nu_0, \sigma_0^2\right) \end{aligned}$$

Posterior distribution for μ:

$$p\left(\mu|Y_1, \ldots, Y_n\right) = t_{\nu_n}\left(\mu_n, \sigma_n^2, \kappa_n\right)$$

$$\begin{aligned} \mu_n &= \frac{\kappa_0}{\kappa_0 + n}\mu_0 + \frac{n}{\kappa_0 + n}\bar{Y} \\ \kappa_n &= \kappa_0 + n \\ \nu_n &= \nu_0 + n \\ \nu_n\sigma_n^2 &= \nu_0\sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_0 + n}\left(\bar{Y} - \mu_0\right)^2 \end{aligned}$$

A Simple Model

$$\begin{aligned} Y_1, \ldots, Y_n &\sim \mathcal{N}\left(\mu, \sigma^2\right) \\ \\ \mu|\sigma^2 &\sim \mathcal{N}\left(\mu_0, \sigma^2/\kappa_0\right) \\ \sigma^2 &\sim \mathrm{Inv}-\chi^2\left(\nu_0, \sigma_0^2\right) \end{aligned}$$

Posterior predictive distribution:

1. Sample $\sigma^2|\{Y_i\}$ from $\mathrm{Inv}-\chi^2\left(\nu_n, \sigma_n^2\right)$.

2. Sample $\mu|\sigma^2, \{Y_i\}$ from $\mathcal{N}\left(\mu_n, \sigma^2/\kappa_n\right)$.

3. Sample Y^* from $\mathcal{N}\left(\mu, \sigma^2\right)$.

Hierarchical Models

- A common approach involves a three-level hierarchy:

$$\begin{aligned} \text{Data model:} \quad & [data|process, parameters] \\ \text{Process model:} \quad & [process|parameters] \\ \text{Prior model:} \quad & [parameters] \end{aligned}$$

- Simplifies the problem by factoring a complicated distribution into a series of conditional distributions.

  – `http://oneredpaperclip.blogspot.com/`

- Inference involves sampling the posterior distribution:

$$[process, parameters|data] \propto [data|process, parameters][process|parameters][parameters]$$

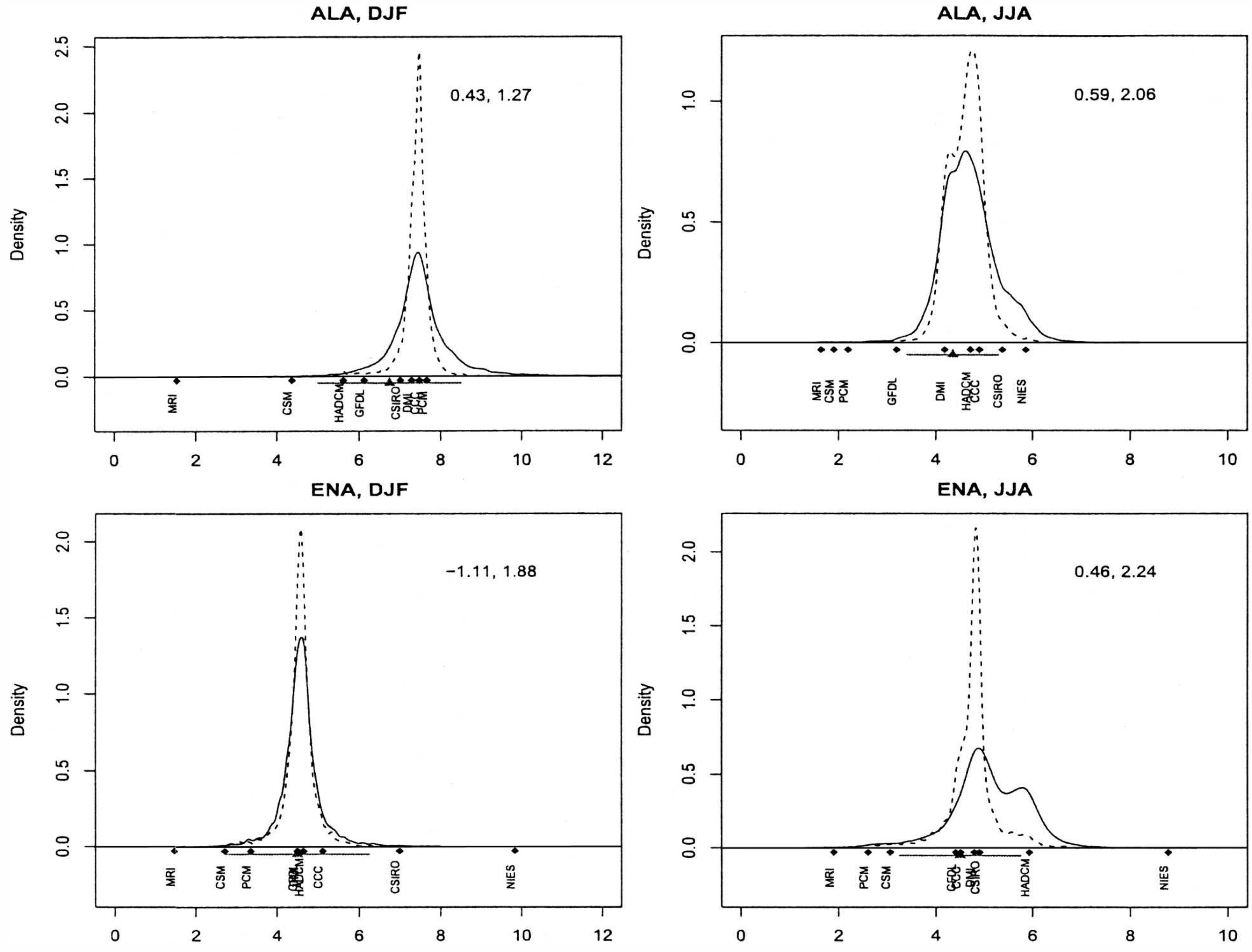
Model weighting and pdfs

Data model:

$$\begin{aligned} X_0 &\sim \mathcal{N}\left(\mu, \lambda_0^{-1}\right) \\ X_j &\sim \mathcal{N}\left(\mu, \lambda_j^{-1}\right) \\ Y_j &\sim \mathcal{N}\left(\nu, \left(\theta\lambda_j\right)^{-1}\right) \end{aligned}$$

- X_0 indicates an observed climate
- X_j, Y_j indicates model output for the current/future time period.
- μ is current mean temperature, ν is future temperature
- λ_j are precision (inverse variance) parameters.
- θ allows the climate model variance to change between time periods.

Turning the Bayesian crank gives inference on $\Delta = \nu - \mu$



Analyzing ensembles: ANOVA

Analysis of variance (ANOVA) is a classical statistical technique for quantifying the impact of known experimental factors (inputs) on the output of an experiment.

North American Regional Climate Change Assessment Program (NARCCAP)

	Phase I	Phase II			
	NCEP	GFDL	CGCM3	HADCM3	CCSM
CRCM	finished		finished		finished
ECP2	finished	finished		running	
HRM3	finished	finished		finished	
MM5I	finished			running	finished
RCM3	finished	finished	finished		
WRFG	finished		finished		finished

What is the relative contribution of the GCM versus the RCM in the NARCCAP experiment?

Northern Rockies...

	GFDL	CGCM3	HADCM3	CCSM	
CRCM		2.66		3.51	3.09
ECP2	1.59		????		1.59
HRM3	3.99		3.44		3.72
MM5I			????	2.15	2.15
RCM3	2.75	2.70			2.72
WRFG		1.74		2.30	2.02
	2.78	2.37	3.44	2.65	2.68

- Is there greater variability across rows or columns?

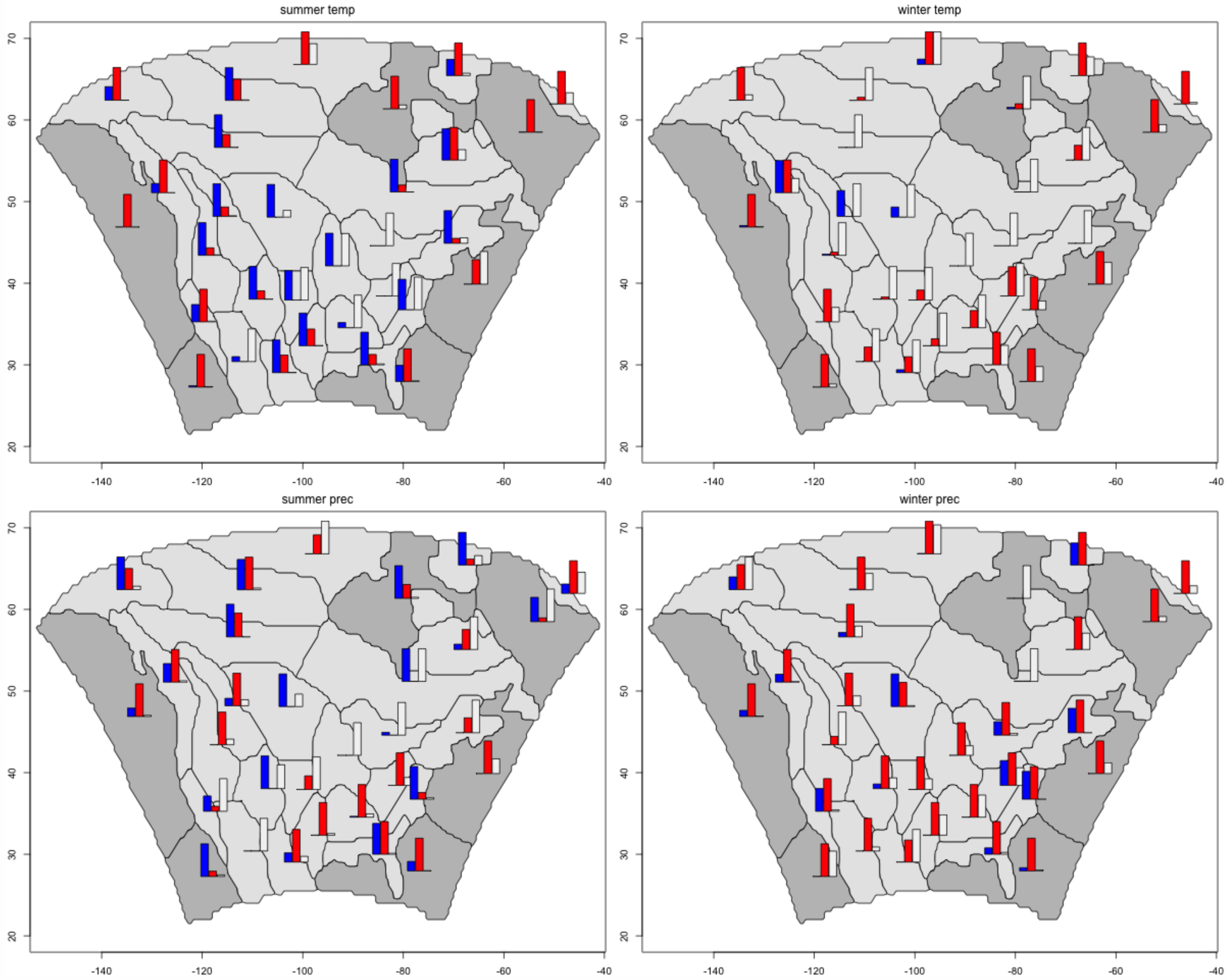
- How well does a model that suggest something systematic across rows and/or columns actually fit?

Consider a random effects ANOVA model:

$$Y_{ij} = \mu + \text{RCM}_i + \text{GCM}_j + \epsilon_{ij}$$

where

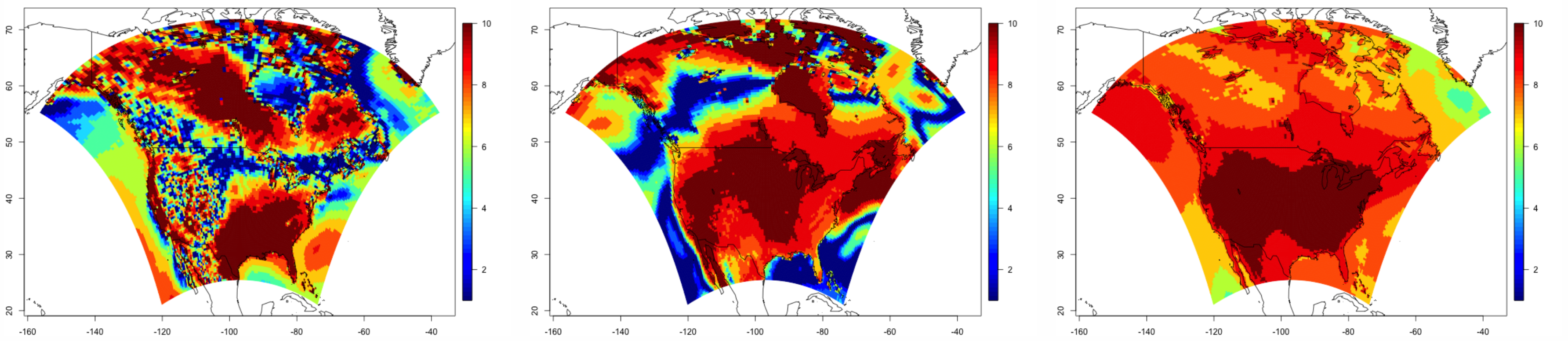
- $\text{Var}[\text{RCM}_i] = \sigma^2_{\text{RCM}}$, $\text{Var}[\text{GCM}_j] = \sigma^2_{\text{GCM}}$
- $\text{Var}[\epsilon_{ij}] = \sigma^2$
- $\text{Cov}[Y_{ij}, Y_{ik}] = \sigma^2_{\text{RCM}}$ – shares an RCM
- $\text{Cov}[Y_{ij}, Y_{kj}] = \sigma^2_{\text{GCM}}$ – shares a GCM


summer temp
winter temp
summer prec
winter prec
70
60
50
40
30
20
-140
-120
-100
-80
-60
-40

Some issues

- Fixed effects, random effects, Bayesian...

- Spatial and spatial-temporal dependence...



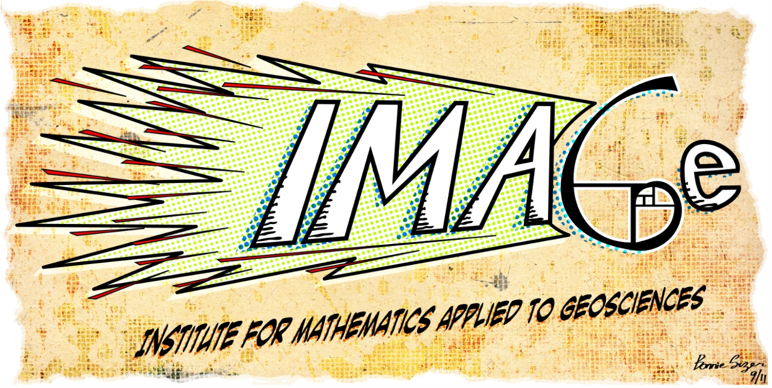
- Multiple testing...

- Non-Gaussian models...

Questions?

Many opportunities for visits and collaboration: ASP, RSVP, SIParCS, GSP, IMAGe, TOY,...

Climate Informatics Workshop, 9/20-21; CIDU, 10/24-26; SAMSI Massive Data, spring 2013.



ssain@ucar.edu
http://www.image.ucar.edu/∼ssain

Thank You!