Statistical Methods for Complex Models in Climate Science

Murali Haran

Department of Statistics, Pennsylvania State University

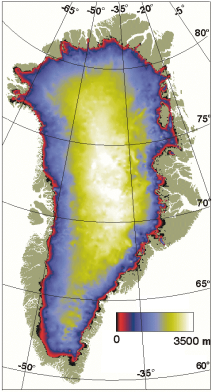
Network for Sustainable Climate Risk Management (SCRiM)
Summer School Lunch Talk. July 2014

This Talk

- Complex models are often used to make projections about future climate. E.g. Ice models are often used to make projections about future ice sheet behavior.
- Model input parameters are like knobs/dials on the climate model. They greatly influence how the model behaves.
- What values should these parameters be set to? How sure/uncertain are we about their values?
- Use data! E.g. recent data on the ice sheet
- Lots of challenges: large spatial data, complicated errors.
- A SCRiM research group is developing statistical methods for parameter inference for climate models.

Greenland Ice Sheet

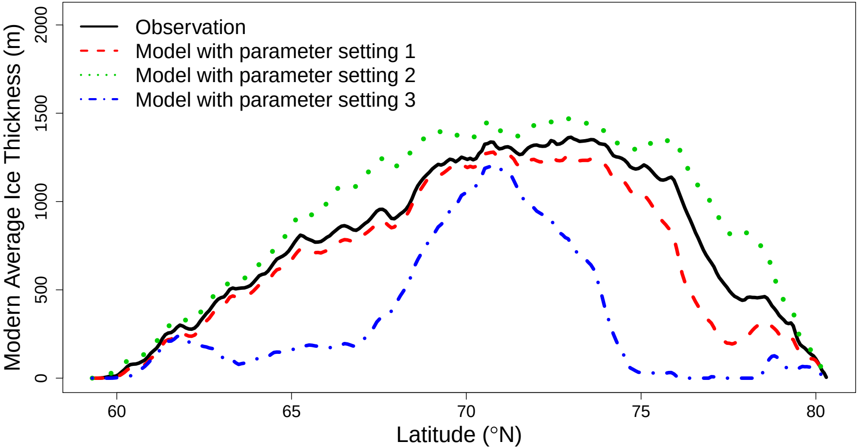
Important contributor to sea level rise: Total melting results in sea level rise of 7m.



Bamber et al. (2001)

Calibration Problem

Which parameter settings best match observations?

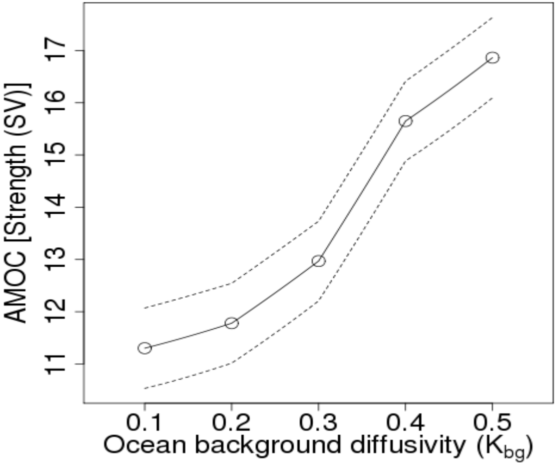


The AMOC and Climate Change

- Atlantic Meridional Overturning Circulation (AMOC): AMOC heat transport makes a substantial contribution to the moderate climate of Europe (cf. Bryden et al., 2005)
- Any slowdown in the overturning circulation may have major implications for climate change
- AMOC projections from climate models.

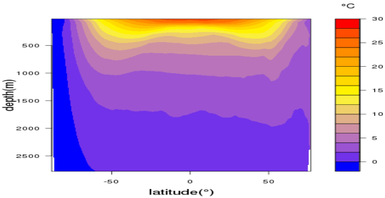
A major source of uncertainty about the AMOC is due to uncertainty about $\mathbf{K_{bg}}$: model parameter that quantifies the intensity of vertical mixing in the ocean.

AMOC and Model Parameter K_{bg}

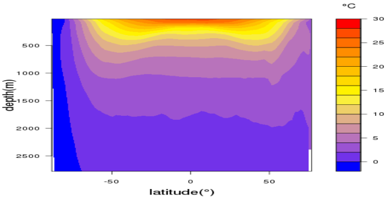


Ocean Temperatures

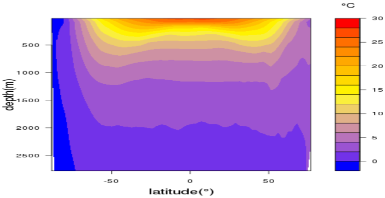
K_{bg} of 0.1



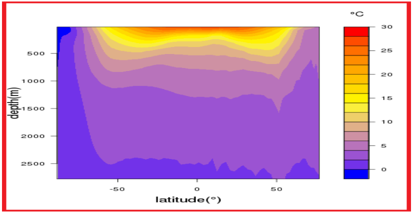
K_{bg} of 0.2



K_{bg} of 0.3



Observations



(2D versions of 3D data)

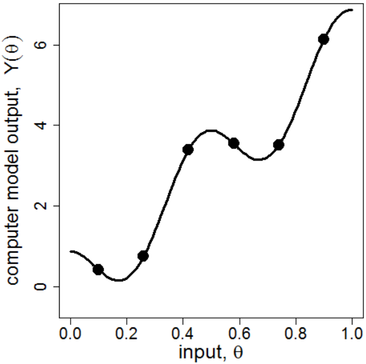
Two-stage Approach to Emulation-Calibration

1. Emulation step: Find fast approximation for climate model using Gaussian process (GP).
   Information used: climate model runs at various parameter settings.
2. Calibration step: Infer climate parameter using emulator and observations. Important: account for errors in the data, and data-model discrepancy (model is an imperfect representation of reality).
   Information used: climate observations and emulator.

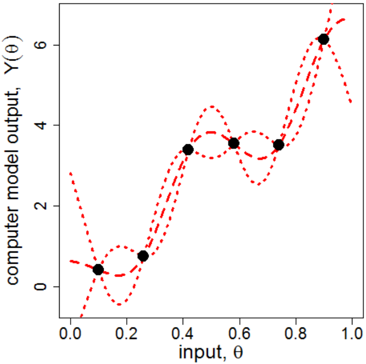
References: Bhat, Haran, Olson, Keller (2012); Chang, Haran, Olson, Keller (2014); Liu, Bayarri and Berger (2009)

Emulation Step: A Simple Example

We use a statistical model called a **Gaussian process.** This model is a fast emulator (approximation) of the computer model.



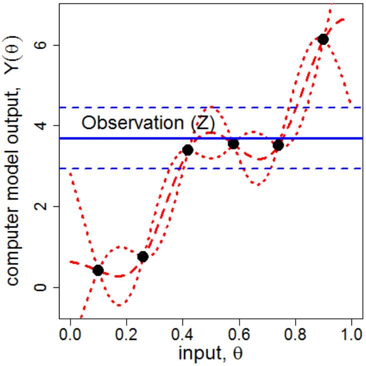
Computer model output (y-axis) vs. input (x-axis)



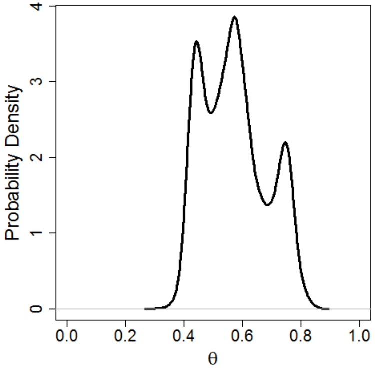
Emulation (approximation) of computer model using GP

Calibration Step: A Simple Example

We use statistical methods called **Bayesian inference and Markov chain Monte Carlo**: Use emulator (from before) and observations to learn about parameters.

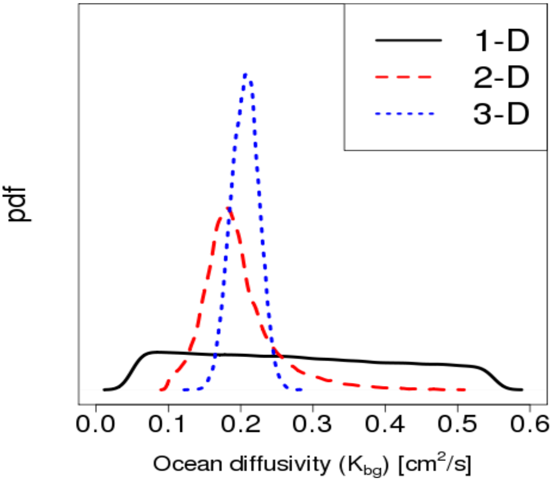


Combining observation and emulator



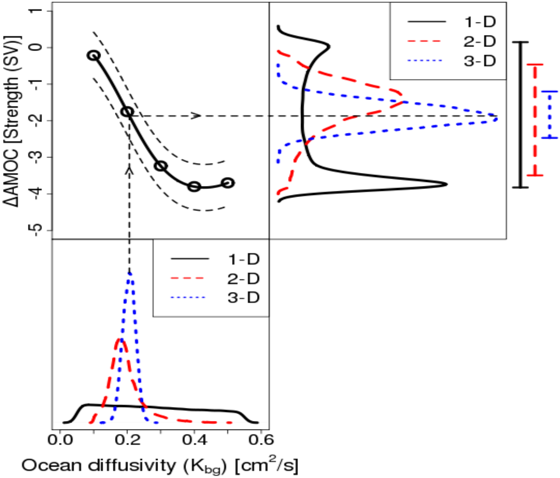
Posterior PDF of θ given model output and observation

Results for K_{bg} Inference



(from Chang, Haran, Olson and Keller, 2014)

MOC Projections for 2100 Using Inferred K_{bg}



(from Chang, Haran, Olson and Keller, 2014)

Concluding Thoughts

- Without probability and statistics, it is not possible to quantify risk.
- Advanced statistical methods allow us to
  - Utilize all (*large*) data sets which can often help reduce uncertainties about projections.
  - Account for errors, uncertainties carefully.
  - Learn about various sources of error, e.g. discrepancy between data and model.
  - Learn about complicated interactions among model parameters.
- This work requires expertise (people with M.S. and Ph.D.s!) in statistics and geosciences. We work together very closely.

Statistical Methods: Details

BEGIN: FANCY STATS...

Summary of Statistical Problem

- **Goal**: Learning about $\boldsymbol{\theta}$ based on two sources of information:
  - **Observations**: Mean ice thickness profile† $\mathbf{Z} = (Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n))^T$, where $\mathbf{s}_1, \ldots, \mathbf{s}_n$ are latitude points.
  - **Ice model output*** for mean ice thickness $\mathbf{Y}(\boldsymbol{\theta}_1), \ldots, \mathbf{Y}(\boldsymbol{\theta}_p)$, where each $\mathbf{Y}(\boldsymbol{\theta}_i) = (Y(\mathbf{s}_1, \boldsymbol{\theta}_i), \ldots, Y(\mathbf{s}_n, \boldsymbol{\theta}_i))^T$ is spatial process (Applegate et al 2012).

  $\mathbf{Z}$ and $\mathbf{Y}(\boldsymbol{\theta}_i)$'s are n-dimensional vectors

- Important: output at each $\boldsymbol{\theta}_i$ is a spatial process. $n = 264$ locations, $p = 100$ runs.

†Averaged over longitude

*SICOPOLIS (Greve, 1997; Greve et al., 2011)

Step 1: Dimension Reduction

- Consider model outputs at $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p$ as replicates and obtain PCs

$$\begin{pmatrix} Y(\mathbf{s}_1, \boldsymbol{\theta}_1) & \ldots & Y(\mathbf{s}_n, \boldsymbol{\theta}_1) \\ \vdots & \ddots & \vdots \\ Y(\mathbf{s}_1, \boldsymbol{\theta}_p) & \ldots & Y(\mathbf{s}_n, \boldsymbol{\theta}_p) \end{pmatrix}_{p \times n} \Rightarrow \begin{pmatrix} Y_1^R(\boldsymbol{\theta}_1) & \ldots & Y_{J_y}^R(\boldsymbol{\theta}_1) \\ \vdots & \ddots & \vdots \\ Y_1^R(\boldsymbol{\theta}_p) & \ldots & Y_{J_y}^R(\boldsymbol{\theta}_p) \end{pmatrix}_{p \times J_y}$$

- PCs pick up characteristics of model output that vary most across input parameters $\theta_1, \ldots, \theta_p$.

Step 2: Emulation Using PCs

- Fit 1-dimensional GP for each series $Y_j^R(\boldsymbol{\theta}_1), \ldots, Y_j^R(\boldsymbol{\theta}_p)$
- $\boldsymbol{\eta}(\boldsymbol{\theta}, \mathbf{Y}^R)$: J_y-dimensional emulation process for PCs, $\mathbf{Y}^R$ is collection of PCs
- Computation reduces from $\mathcal{O}(n^3p^3)$ to $\mathcal{O}(J_yp^3)$ (6.13×10^{12} to 3.33×10^6 flops).
- Emulation for original output: compute $\mathbf{K}_y\boldsymbol{\eta}(\boldsymbol{\theta}, \mathbf{Y}^R)$ where $\mathbf{K}_y$ is matrix of scaled eignvectors

Dimension Reduction for Discrepancy Process

- Kernel convolution: Specifying n-dimensional discrepancy process δ using J_d-dimensional knot process $\boldsymbol{\nu}$ ($J_d < n$) and kernel functions

- Kernel basis matrix $\mathbf{K}_d$ links grid locations $\mathbf{s}_1, \ldots, \mathbf{s}_n$ to knot locations $\mathbf{a}_1, \ldots, \mathbf{a}_{J_d}$;

$$\{\mathbf{K}_d\}_{ij} = \exp\left(-\frac{\|\mathbf{s}_i - \mathbf{a}_j\|}{\phi_d}\right)$$

  with $\phi_d > 0$. Fix ϕ_d at large value determined by expert judgment

- Results in better identifiability: Overly flexible discrepancy process may be confounded with emulator

Calibration in Reduced Dimensions

- Probability model for dimension-reduced observation $\mathbf{Z}^R$:

$$\mathbf{Z} = \underbrace{\mathbf{K}_y \boldsymbol{\eta}(\boldsymbol{\theta}, \mathbf{Y}^R)}_{\text{emulator}} + \underbrace{\mathbf{K}_d \boldsymbol{\nu}}_{\text{discrepancy}} + \underbrace{\boldsymbol{\epsilon}}_{\text{observation error}},$$

$$\Rightarrow \mathbf{Z}^R = (\mathbf{K}^T\mathbf{K})^{-1}\mathbf{K}^T\mathbf{Z} = \begin{pmatrix} \boldsymbol{\eta}(\boldsymbol{\theta}, \mathbf{Y}^R) \\ \boldsymbol{\nu} \end{pmatrix} + (\mathbf{K}^T\mathbf{K})^{-1}\mathbf{K}^T\boldsymbol{\epsilon},$$

  with combined basis $[\mathbf{K}_y\ \mathbf{K}_d]$, knot process $\boldsymbol{\nu} \sim N(\mathbf{0}, \kappa_d \mathbf{I})$, and observational error $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$.

- Infer $\boldsymbol{\theta}$ through posterior distribution

$$\pi(\boldsymbol{\theta}, \kappa_d, \sigma^2 | \mathbf{Z}^R, \mathbf{Y}^R) \propto \underbrace{L(\mathbf{Z}^R | \mathbf{Y}^R, \boldsymbol{\theta}, \kappa_d, \sigma^2)}_{\text{likelihood given by above}} \underbrace{p(\boldsymbol{\theta})p(\kappa_d)p(\sigma^2)}_{\text{priors}}$$