Parameter Inference for Computer Models with High-dimensional Spatial Output

Murali Haran

Department of Statistics, Pennsylvania State University

III Forum Mineiro de Estatistica e Probabilidade
Belo Horizonte, August 2014.

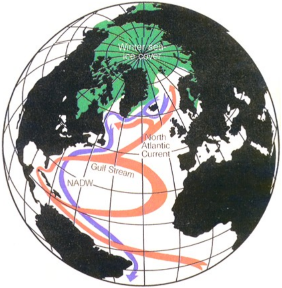
Collaborators:
Won Chang (University of Chicago Statistics)
Patrick Applegate, Klaus Keller, Roman Olson (Penn State Geosciences)

This Talk

- Climate models are often used to make projections about future climate.
- A major source of uncertainty about these projections is due to uncertainty about climate model input parameters.
- We propose a method for learning about climate model parameters from climate model outputs and observations.
- Challenges: Data in the form of high-dimensional spatial fields. Complicated error structures.
- I will describe novel computationally efficient approaches based on principal components (PC) and kernel convolution

Atlantic Meridional Overturning Circulation (AMOC)

Global conveyor belt: Carries warm upper waters into far-northern latitudes and returns cold deep waters southward across the equator



Rahmstorf (1997)

- Important for maintaining equilibrium climate in Europe
- Slowdown in AMOC would have profound implications for climate
- Scientific Goal: Making projections for AMOC using climate model

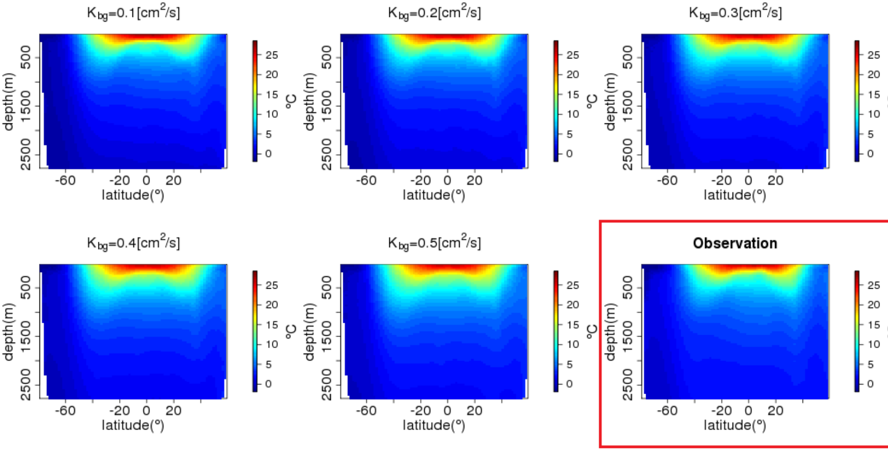
Learning about K_{bg}

- Parametric uncertainty due to unknown vertical diffusivity
  - Vertical mixing is important in AMOC projection.
  - Most of mixing occurs below climate model scale
    $\Rightarrow$ Need "parameterization", that is, parameter K_{bg} is used to represent this mixing

- Background vertical diffusivity (K_{bg}): Model parameter that quantifies intensity of vertical mixing in ocean.

Calibration Problem

Which parameter settings best match observations?



Two-stage Approach to Emulation-Calibration

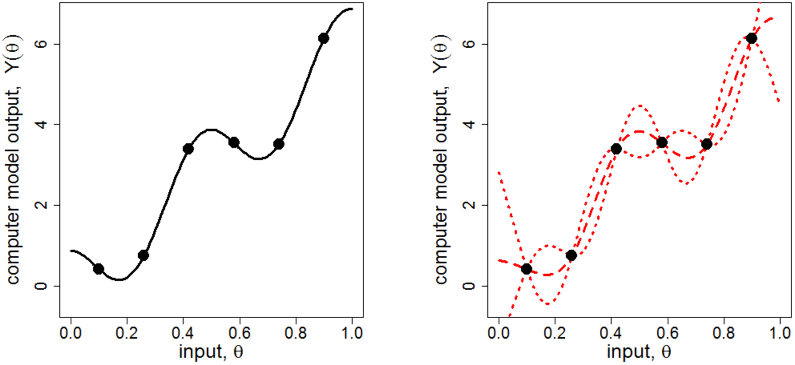
1. Emulation step: Find fast approximation for climate model using Gaussian process (GP)

2. Calibration step: Infer climate parameter using emulator and observations, while accounting for data-model discrepancy

   (Bhat, Haran, Olson, Keller, 2012; Liu, Bayarri and Berger, 2009)

Emulation Step

Toy example: pretend model output is a scalar

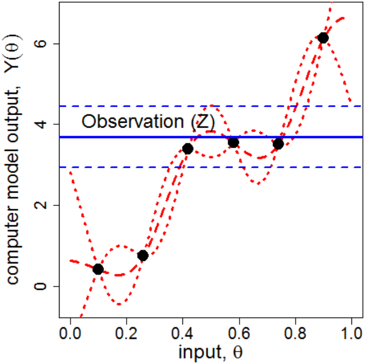


Computer model output (y-axis) vs. input (x-axis)

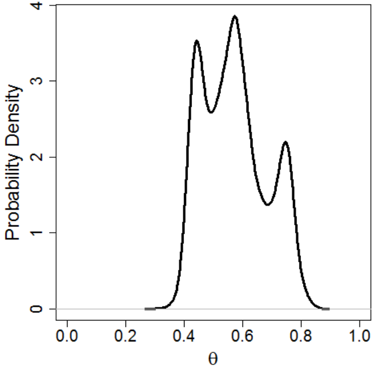
Emulation (approximation) of computer model using GP

Calibration Step

Toy example: pretend model output and observations are scalars



Combining observation and emulator



Posterior PDF of θ given model output and observation

Summary of Statistical Problem

- **Goal**: Learning about $\boldsymbol{\theta}$ based on two sources of information:
  - **Observations***: Mean potential ocean temperature†, $\mathbf{Z} = (Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n))^T$, where $\mathbf{s}_1, \ldots, \mathbf{s}_n$ are 3D locations.
  - **Climate model output**** for mean potential temperature $\mathbf{Y}(\boldsymbol{\theta}_1), \ldots, \mathbf{Y}(\boldsymbol{\theta}_p)$, 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 field (Sriver 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 high-dimensional spatial field. $n = 61,051$ locations, $p = 250$ runs.

*World Ocean Atlas 2009
**University of Victoria (UVic) Earth System Climate Model
†Averaged over 1955-2006

GP for Computer Model Emulation

- Fit GP to *np*-dimensional data $\mathbf{Y} = \left(\mathbf{Y}(\boldsymbol{\theta}_1)^T, \ldots, \mathbf{Y}(\boldsymbol{\theta}_p)^T\right)^T$ for interpolation.

- Covariance used for
  - non-linear relationship between parameter and model output (model output as a function of parameter)
  - non-linear spatial surface (model output as a function of location)

- Covariance function example:

$$\begin{aligned} \mathrm{Cov}\left(Y(\mathbf{s},\boldsymbol{\theta}), Y(\mathbf{s}',\boldsymbol{\theta}'); \boldsymbol{\xi}\right) =& \kappa \exp\left(-\frac{g\left(\mathbf{s},\mathbf{s}'\right)}{\phi_{\mathbf{s}}}\right) \exp\left(-\frac{\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|}{\phi_{\theta}}\right) \\ &+ \zeta I(\boldsymbol{\theta}=\boldsymbol{\theta}') I(\mathbf{s}=\mathbf{s}') \end{aligned}$$

where g is geodesic distance, and $\boldsymbol{\xi} = (\kappa, \phi_s, \phi_\theta, \zeta)$ is covariance parameter.

Step 1: Emulation (Approximating Computer Model)

- Find MLE for covariance parameter ξ, denoted by $\hat{\xi}$
- Get $\eta(\boldsymbol{\theta}_{NEW}, \mathbf{Y})$ for prediction at any $\boldsymbol{\theta}_{NEW} \in \Theta$:
  - GP gives

$$\begin{pmatrix} \mathbf{Y} \\ \mathbf{Y}(\boldsymbol{\theta}_{NEW}) \end{pmatrix} \sim N\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}_{n(p+1)\times 1}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}_{n(p+1)\times n(p+1)} \right)$$

  - Emulator:

$$\eta\left(\boldsymbol{\theta}_{NEW}, \mathbf{Y}\right) = \mathbf{Y}(\boldsymbol{\theta}_{NEW})|\mathbf{Y} \sim N\left(\Sigma_{21}\Sigma_{11}^{-1}\mathbf{Y}, \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right)$$

Step 2: Calibration (Inferring Input Parameter)

- Probability model for **Z** based on

$$\mathbf{Z} = \eta(\boldsymbol{\theta}, \mathbf{Y}) + \boldsymbol{\delta},$$

  where n-dimensional spatial field $\boldsymbol{\delta}$ is model-observation discrepancy with covariance parameter $\boldsymbol{\xi}_\delta$.

- Inference for $\boldsymbol{\theta}$ based on posterior distribution

$$\pi(\boldsymbol{\theta}, \boldsymbol{\xi}_\delta | \mathbf{Z}, \mathbf{Y}, \hat{\boldsymbol{\xi}}) \propto \underbrace{L(\mathbf{Z}|\mathbf{Y}, \boldsymbol{\theta}, \boldsymbol{\xi}_\delta, \hat{\boldsymbol{\xi}})}_{\text{likelihood given by above}} \times \underbrace{p(\boldsymbol{\theta}) \times p(\boldsymbol{\xi}_\delta)}_{\text{priors for } \boldsymbol{\theta} \text{ and } \boldsymbol{\xi}_\delta}$$

  with emulator parameter $\hat{\boldsymbol{\xi}}$ fixed at value estimated in emulation step.

Computational Challenges and Our Approach

- Emulation requires dealing with $np \times np$ covariance matrix of $\mathbf{Y}$ (reminder: $n = 61,051$ $p = 250$):
  - Cholesky decomposition costs $\frac{1}{3}n^3p^3 = 1.185 \times 10^{21}$ flops.
  - Storing covariance matrix requires $8 \times \frac{250^2 \times 61051^2}{1024^3} = 1,735,624$ Gb memory space.

- Calibration faces similar challenges for dealing with $n \times n$ covariance matrix.

Our fast reduced dimension approach: Fast computation using PC and Kernel Convolution

Main Idea

- Consider model outputs at $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p$ as if they were replicates of a multivariate process, thereby obtaining their 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 $\theta_1, \ldots, \theta_p$.

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 $\mathcal{O}(J_y p^3)$ (1.2×10^{21} to 1.0×10^8 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 $\mathbf{a}_1, \ldots, \mathbf{a}_{J_d}$;

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

  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 $\mathbf{Z}^R$:

$$\begin{aligned}\mathbf{Z} &= \underbrace{\mathbf{K}_y \eta(\boldsymbol{\theta}, \mathbf{Y}^R)}_{\text{emulator}} + \underbrace{\mathbf{K}_d \boldsymbol{\nu}}_{\text{discrepancy}} + \underbrace{\boldsymbol{\epsilon}}_{\text{observation error}}, \\ \Rightarrow \mathbf{Z}^R &= (\mathbf{K}^T\mathbf{K})^{-1}\mathbf{K}^T\mathbf{Z} = \begin{pmatrix} \eta(\boldsymbol{\theta}, \mathbf{Y}^R) \\ \boldsymbol{\nu} \end{pmatrix} + (\mathbf{K}^T\mathbf{K})^{-1}\mathbf{K}^T\boldsymbol{\epsilon},\end{aligned}$$

  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

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

Scientific Question: Effect of Data Aggregation

- Common practice: Calibration using aggregated data (e.g. zonal average)
  - Avoiding computational issues
  - Limited skill of climate model in reproducing spatial patterns

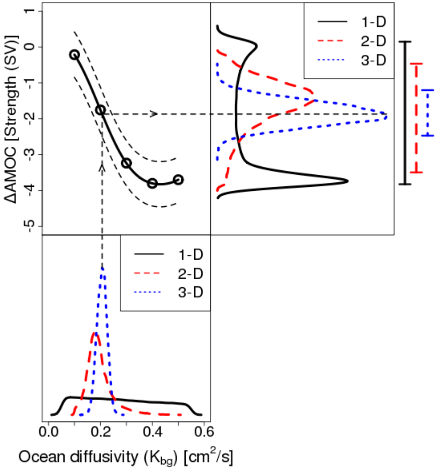
- Using unaggregated data may result in
  - perhaps less uncertainty due to using more data?
  - perhaps more uncertainty due to poor model skill?

- Largely unanswered due to inability to handle unaggregated data

Results

Computational efficiency allows us to calibrate using unaggregated data.

- We compare 1D (depth profile) and 2D (zonal average) with 3D (unaggregated) data.
- Inference with 3D data leads to sharper inference for $\boldsymbol{\theta}$.
- Inference using 3D data is more robust to changes in prior specifications for discrepancy parameters.



Discussion and Ongoing Work

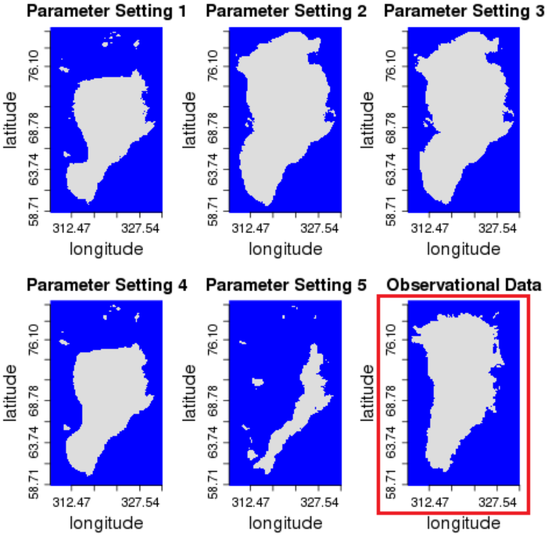
Dimension reduction-based approach:

- Very fast, scales well with n, number of spatial locations
- Very easy to use: Automatic emulation step
- Works for a number of other multivariate settings, e.g. time series, multiple time series, multiple spatial output

How do our methods apply to ice sheet model calibration?

New Challenge: Calibration with Spatial Binary Output

Again, which output best matches the observations?

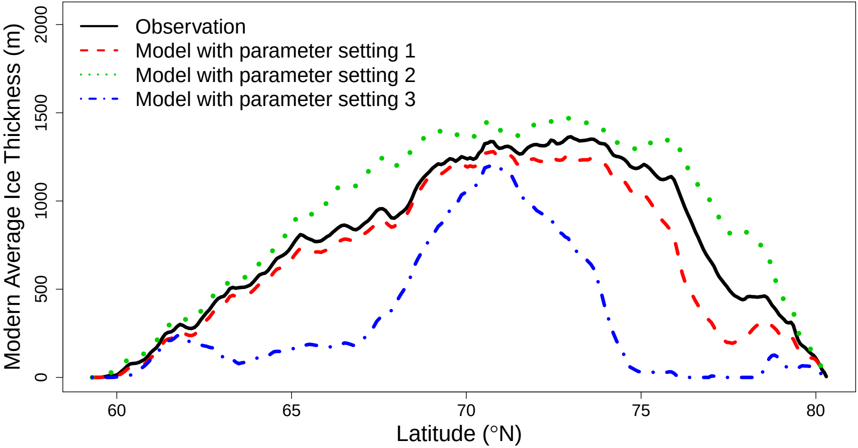


Calibration with Binary Output

- Standard Gaussian process approach does not apply
- Our reduced-dimensional approach also does not apply
- Some options:
  - Aggregation/averaging to obtain “more Gaussian” output, then apply our methods
  - New approach that applies to binary output. Challenging: naive application of spatial generalized mixed model to such data is infeasible

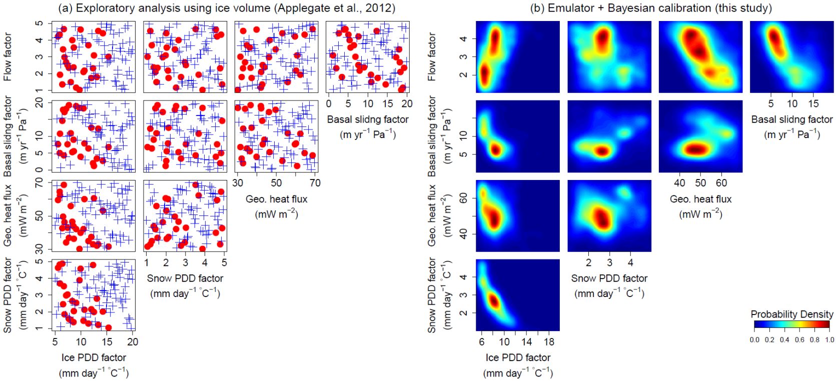
Aggregation Approach

Which parameter settings best match *aggregated* observations?

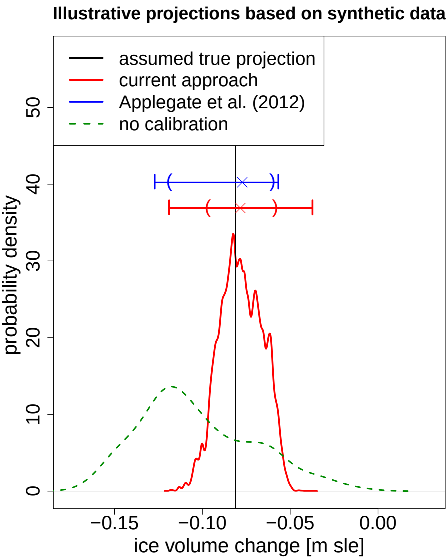


How Does Statistical Rigour Help?

Left: sensible but non-rigourous vs Right: sound statistics
"Underneath the hood": (i) accounting for (epistemic) uncertainties in emulation, (ii) real probability distributions.



Ice Volume Change Projection



Ongoing Work

- Would like to use the original binary data. Hence, reduced-dimensional calibration for binary spatial data.
- Computational issues are even more delicate because a naive latent variable approach would result in severe computational issues
    1. Use binary analogue to regular PCAs.
    2. Discrepancy modeling is tricky...

Acknowledgments

Collaborators:

- Won Chang, University of Chicago
- David Pollard, Earth and Environmental Systems Institute (EESI), Penn State U.
- Patrick Applegate, EESI, Penn State U.
- Klaus Keller, Geosciences, Penn State U.
- Roman Olson, The University of New South Wales

This work was partially supported by the following grants:

- The Network for Sustainable Climate Risk Management (SCRiM), NSF GEO-1240507.
- NSF CDSE/DMS-1418090

Relevant Manuscripts

- Chang, W., M. Haran, R. Olson, and K. Keller (2014): Fast dimension-reduced climate model calibration, *Annals of Applied Statistics*
- Chang, W., Applegate, P., Haran, M. and Keller, K. (2014) Probabilistic calibration of a Greenland Ice Sheet model using spatially-resolved synthetic observations: toward projections of ice mass loss with uncertainties, *Geoscientific Model Development*

Appendix: Cross-Validation for Emulator

- Example of leave-10%-out cross validation result: