Inferring Likelihoods and Climate System Characteristics from Climate Models and Multiple Tracers

Murali Haran [1]

(joint work with K. Sham Bhat[1], Roman Tonkonojenkov[2], and Klaus Keller[2])

[1] Department of Statistics, The Pennsylvania State University

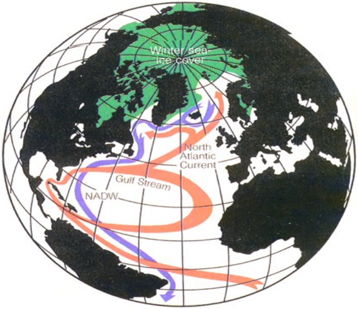
[2] Department of Geosciences,The Pennsylvania State University

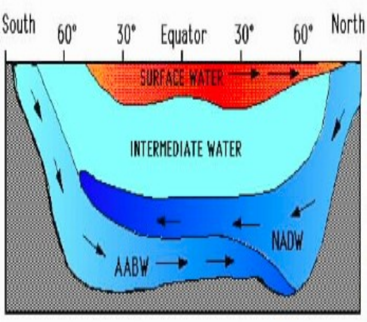
Ninth Valencia International Meeting on Bayesian Statistics

June 2010

Motivation

- Example of climate change: potential collapse of the Atlantic meridional overturning circulation (AMOC), results in disruptions in the equilibrium state of the climate.
- An AMOC collapse may result in drastic changes in temperatures and precipitation patterns.



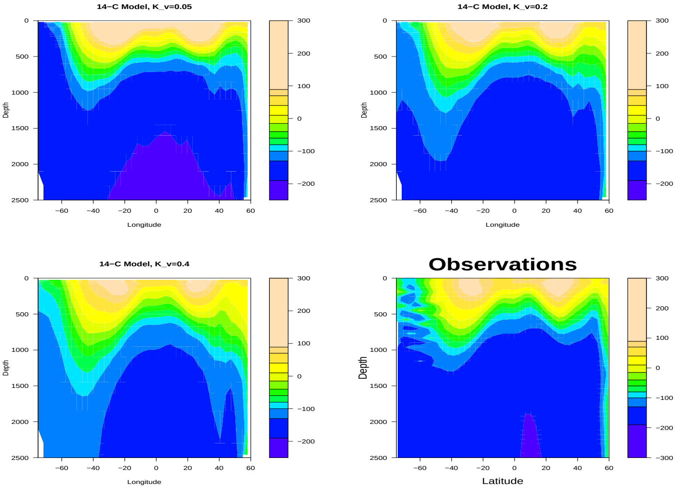


(plots: Rahmstorf (Nature, 1997) and Behl and Hovan)

Overview of Statistical Problem

- **Goal**: Infer important climate characteristics (parameters) that drive major climate systems, and hence AMOC.
- AMOC affected by vertical diffusivity (K_v), latter cannot be measured directly.
- Two sources of (indirect) information about K_v
  - Observations: ^{14}C and CFC11 collected in the 1990s (latitude, longitude, depth), zonally averaged.
  - Output from complex climate models at 6 values of K_v.
- Challenges
  - No direct connection between observations and climate parameter, need to rely on sparse climate model runs.
  - Large data sets: both observations and climate model output.
  - Combining information from multiple spatial fields in a flexible manner (multivariate spatial data).

^{14}C Observations and Model Output



Bottom right: observations, Other plots: model output at different K_v values. Latitude: 80 S-60 N, depths:0-3000m.

Statistical Inference

- Notation: $Z(\mathbf{s})$: physical observations, $Y(\mathbf{s}, \boldsymbol{\theta})$: model output at location $\mathbf{s}$, and calibration parameter $\boldsymbol{\theta}$. $\mathbf{Y}$ and $\mathbf{Z}$ are spatial fields.
- *Emulation* of climate model (Sacks et al, 1989): replace complex model with simple stochastic model.
- Computer model calibration : Kennedy & O'Hagan (2001), Sanso et al. (2008) in the context of climate models, etc.
- **Data Sources**: Observations for ^{14}C/CFC11: $\mathbf{Z}_1, \mathbf{Z}_2$ (locations at $\mathbf{S}$).
- Climate model runs at several values of θ : $\mathbf{Y}_1, \mathbf{Y}_2$.
- **Goal**: Inference for climate parameter θ.

Our Approach

- Two stage approach to obtain posterior of $\boldsymbol{\theta}$: (i) model relationship between **Z** and $\boldsymbol{\theta}$ via model output **Y** and (ii) Use observations **Z** to infer $\boldsymbol{\theta}$.
- Model **Y** (model output) as a Gaussian process (emulator): $\mathbf{Y} \mid \boldsymbol{\beta}, \boldsymbol{\xi} \sim N(\mu_{\boldsymbol{\beta}}(\boldsymbol{\theta}), \Sigma(\boldsymbol{\xi}))$.
- $\boldsymbol{\beta}$: regression parameters, $\boldsymbol{\xi}$: covariance parameters.
- $\boldsymbol{\eta}(\mathbf{Y}, \theta)$ (a random variable) is the prediction at a new θ and locations **S**; obtained by using the standard kriging framework with a multivariate normal distribution.
- $\mathbf{Z} = \boldsymbol{\eta}(\mathbf{Y}, \theta) + \delta(\mathbf{S}) + \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon}$ is observation error and $\delta(\boldsymbol{S})$ is model discrepancy.
- Inference on $\boldsymbol{\theta}$ performed using MCMC.

Bivariate Conditional Hierarchical Model

- How can we combine information from multiple tracers (^{14}C, CFC11) in a flexible manner to infer K_v?
- Model $(\mathbf{Y}_1, \mathbf{Y}_2)$ as a hierarchical model: $\mathbf{Y}_1|\mathbf{Y}_2$ and $\mathbf{Y}_2$ as Gaussian processes. (following Royle and Berliner (1999)).

$$\mathbf{Y}_1 \mid \mathbf{Y}_2, \boldsymbol{\beta}_1, \boldsymbol{\xi}_1, \gamma \sim N(\mu_{\boldsymbol{\beta}_1}(\boldsymbol{\theta}) + \mathbf{B}(\gamma)\mathbf{Y}_2, \Sigma_{1.2}(\boldsymbol{\xi}_1))$$

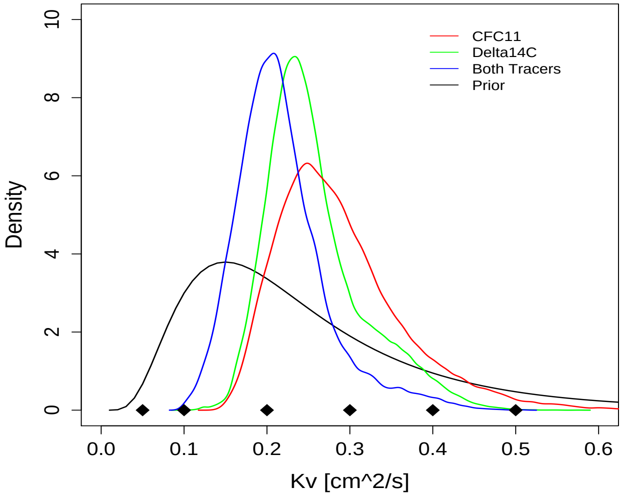
$$\mathbf{Y}_2 \mid \boldsymbol{\beta}_2, \boldsymbol{\xi}_2 \sim N(\mu_{\boldsymbol{\beta}_2}(\boldsymbol{\theta}), \Sigma_2(\boldsymbol{\xi}_2))$$

- $\mathbf{B}(\gamma)$ is a matrix to describe relationship between $\mathbf{Y}_1$ and $\mathbf{Y}_2$, we assume a piecewise linear relationship.
- Use predictive distribution $\boldsymbol{\eta}(\mathbf{Y}, \theta)$ (for $\mathbf{Y} = (\mathbf{Y}_1 \ \mathbf{Y}_2)$) to obtain probability model connecting $\mathbf{Z} = (\mathbf{Z}_1 \ \mathbf{Z}_2)$ to $\boldsymbol{\theta}$.
- $\mathbf{Z} = \boldsymbol{\eta}(\mathbf{Y}, \theta) + \boldsymbol{\delta}(\mathbf{S}) + \boldsymbol{\epsilon}$, confounding between $\boldsymbol{\delta}$ and θ.
- We need to model $\boldsymbol{\delta}(\mathbf{S})$ appropriately, not doing so may result in overfitting θ. Use GP model for unknown $\boldsymbol{\delta}$.

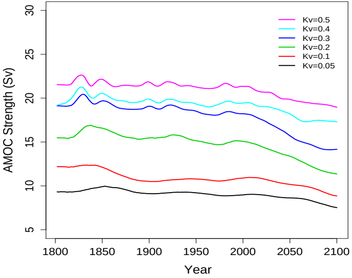
Computational Issues

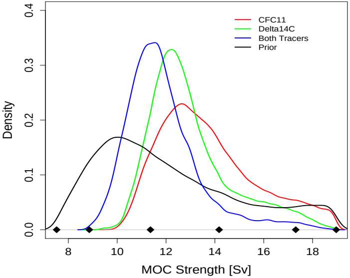
- Matrix computations are $\mathcal{O}(N^3)$, where N is the number of observations (Data size: 3706(observations); 5926(model) per tracer, could be much larger).
- Used reduced rank approach based on kernel mixing (Higdon, 1998): continuous process created by convolving a discrete white noise process with a kernel function.
- Write covariance matrix as: $(\mathbf{A} + \mathbf{KCK}^T)$, **K** kernel matrix with rank J=196 (Cressie and Johannesson, 2008).
- Special structure + Sherman-Woodbury-Morrison identity used to reduce matrix computations to dimension $J \times J$.
- Unlike Higdon (1998), we do not estimate the knot process. Parameter space is therefore greatly reduced, and computations are simplified.

Inference Based on Both Tracers



AMOC Strength Projections for K_v Values





Left: AMOC strength projections in Sv, increasing with K_v.

Right: Distribution of projected AMOC strength(Sv) in 2100 (left) given posterior distributions of K_v for both single tracers and multiple tracers.

Discussion

- Advantages of our approach:
  - Statistical inferential approach to solve an important climate science problem.
  - Flexible bivariate model to combine information from multiple tracers.
  - Improve computational tractability enabling analysis of larger data sets.
  - Modularization (e.g. Liu et al., 2009): reduce computational issues and identifiability problems.
- Potential issues:
  - Possible over-fitting in first stage (ML estimation).
  - Confounding between calibration parameter and model discrepancy.
  - May be difficult to extend to more than two spatial fields.

Key References

- Royle, J. A. and Berliner, L.M. (1999), A Hierarchical Approach to Multivariate Spatial Modeling and Prediction, *Journal of Agricultural, Biological, and Environmental Statistics.*
- Kennedy, M.C. and O'Hagan, A.(2001), Bayesian calibration of computer models, *Journal of the Royal Statistical Society. Series B (Statistical Methodology).*
- Sanso, B. and Forest, C.E. and Zantedeschi, D. (2008) , Inferring Climate System Properties Using a Computer Model, *Bayesian Analysis*.
- Higdon, D. (1998) A process-convolution approach to modelling temperatures in the North Atlantic Ocean, *Environmental and Ecological Statistics*.