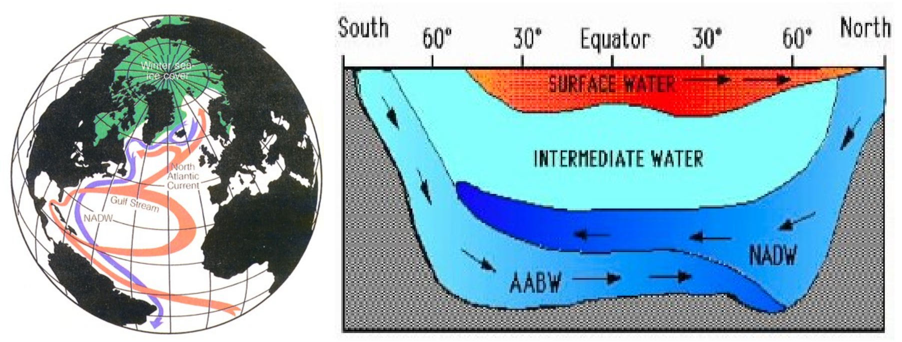
Inferring Likelihoods and Climate System Characteristics from Climate Models and Multiple Tracers

K. Sham Bhat (joint work with Murali Haran, Roman Tonkonojenkov, and Klaus Keller)
Department of Statistics, Pennsylvania State University

PENNSTATE 1855

Motivation

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

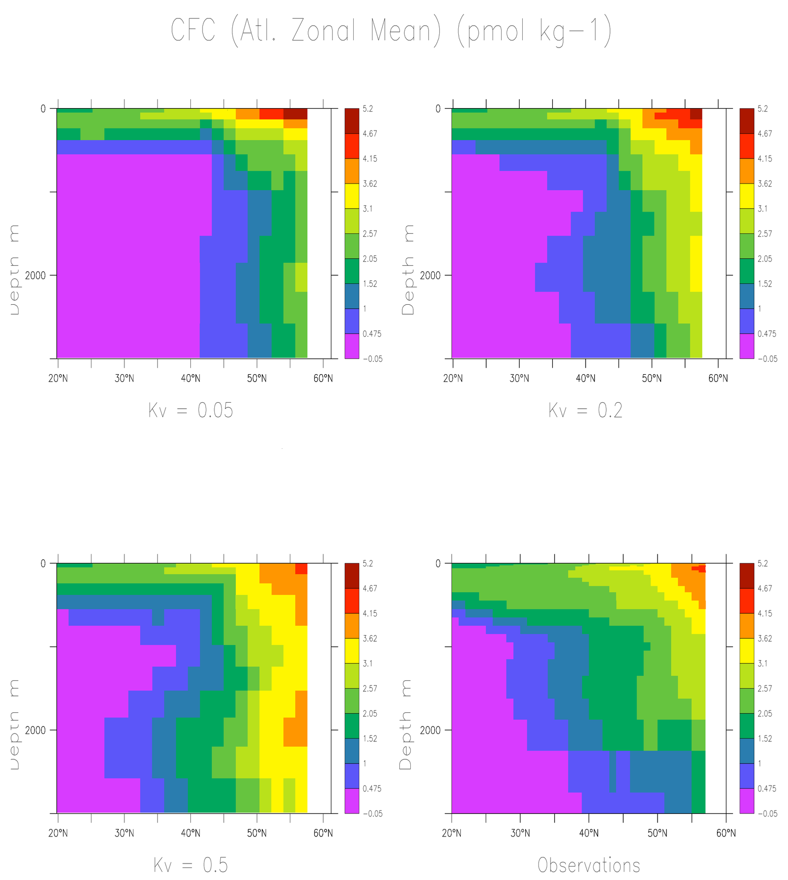


Problem Statement

- **Goal**: Infer important climate characteristics (parameters) that drive major climate systems.
- Sources of information
  - Physical observations of climate system E.g. spatial data on CFC-11 and C-14.
  - Output from complex climate models at several different climate parameters from University of Victoria(UVic) Earth System Climate Model (Weaver et. al. 2001).
- 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 tracers (multivariate spatial data).

Tracers

- Predictions of MOC strength can be made for particular climate parameter settings, such as vertical diffusivity,(K_v).
- K_v values cannot be measured directly.
- Trichlorofluoromethane (CFC11) and Carbon-14 (^{14}C) are considered as stable tracers, used to infer K_v, related to deep ocean behavior such as MOC strength.
- Observations of both tracers collected across globe in the 1990s, locations consist of latitude, longitude, and depth values, aggregated over longitudes.
- Second source of information: climate model output at different values of K_v.
- 3706 observations and 5926 data points from model.
- Latitude between -80 S and 60 N, depths from 0 to 3000m.



Bottom right: observations, other plots are model output.

Statistical Inference: Related Work

- Notation: $\boldsymbol{Z}(\mathrm{s})$ are physical observations at location s=(latitude, depth).
- $\boldsymbol{Y}(\mathrm{s}, \boldsymbol{\theta})$ are climate model output at location s and climate parameter $\boldsymbol{\theta}$.
- Kennedy and O'Hagan (2001) developed a model to solve the calibration problem for computer experiments. Sanso et al. (2007) used a variant for climate models.
  - Assumes that a "true" set of climate parameters $\boldsymbol{\theta}^*$ exists.
  - Combine both $\boldsymbol{Y}$ and $\boldsymbol{Z}$ in one model, treat $\boldsymbol{Y}$ as a Gaussian process, with a mean dependent on climate parameter $\boldsymbol{\theta}$:

$$Z(\mathrm{s_i}) = Y(\mathrm{s_i}, \boldsymbol{\theta}^*) + \delta(\mathrm{s_i}) + \epsilon_i$$

  - where δ is the model error term and ϵ is observation error.
  - Separable covariance function between s and $\boldsymbol{\theta}$.

Our Approach

- Two stage approach to obtain posterior of $\boldsymbol{\theta}$: (1) model relationship between $\boldsymbol{Z}$ and $\boldsymbol{\theta}$ via model output $\boldsymbol{Y}$ and (2) Use observations $\boldsymbol{Z}$ to obtain $\boldsymbol{\theta}^*$.
- Model $\boldsymbol{Y}$ as a Gaussian process: $\boldsymbol{Y} \mid \boldsymbol{\beta}, \boldsymbol{\xi} \sim N(\mu_{\boldsymbol{\beta}}(\boldsymbol{\theta}), \Sigma(\boldsymbol{\xi}))$.
- $\boldsymbol{\beta}$: regression parameters, $\boldsymbol{\xi}$:covariance parameters, covariance function separable among s and $\boldsymbol{\theta}$.
- $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$ estimated using maximum likelihood.
- For locations s at a *given* value of $\boldsymbol{\theta}$, we obtain the predictive distribution $\pi(\boldsymbol{Z}^*(\boldsymbol{\theta}) \mid \boldsymbol{Y})$ which is multivariate normal, a function of $\boldsymbol{\theta}$.

$$\boldsymbol{Z} = \hat{\boldsymbol{\eta}}(\boldsymbol{Z}^* \mid \boldsymbol{\theta}^*, \boldsymbol{Y}) + \delta(\mathbf{s}) + \epsilon$$

- where δ is the model error term and ϵ is observation error.
- $\epsilon \sim N(0, \psi I)$ and δ is modeled as a Gaussian process.
- Two stage approach can be seen as a way of 'cutting feedback' (Best et al. 2006; Rougier, 2008).
- Potential advantages (cf. Rougier, 2008):
  - Protecting emulator from a poor model of climate system.
  - Modeling emulator separately to facilitate careful evaluation of emulator.
  - Computational tractability.
  - Identifiability for δ and ϵ, hope to improve our ability to learn about these errors.
- General framework related to Bayarri et al. (2007), and can be applied to single or multiple tracers.
- Potential drawback: lack of unified model.

Bivariate Conditional Hierarchical Model

- How can we combine information from multiple tracers (CFC-11, C-14) to infer K_v?
- Model $(\boldsymbol{Y}_1, \boldsymbol{Y}_2)$ as a hierarchical model: $\boldsymbol{Y}_1|\boldsymbol{Y}_2$ and $\boldsymbol{Y}_2$ as Gaussian processes. (following Royle and Berliner (1999)).

$$\boldsymbol{Y}_1 \mid \boldsymbol{Y}_2, \boldsymbol{\beta}_1, \boldsymbol{\xi}_1, \gamma \sim N(\mu_{\boldsymbol{\beta}_1}(\boldsymbol{\theta}) + \mathbf{B}(\gamma)\boldsymbol{Y}_2, \Sigma_{1.2}(\boldsymbol{\xi}_1))$$
$$\boldsymbol{Y}_2 \mid \boldsymbol{\beta}_2, \boldsymbol{\xi}_2 \sim N(\mu_{\boldsymbol{\beta}_2}(\boldsymbol{\theta}), \Sigma_2(\boldsymbol{\xi}_2))$$

- $\mathbf{B}(\gamma)$ is a matrix to describe relationship between $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$, We assume a piecewise linear relationship.

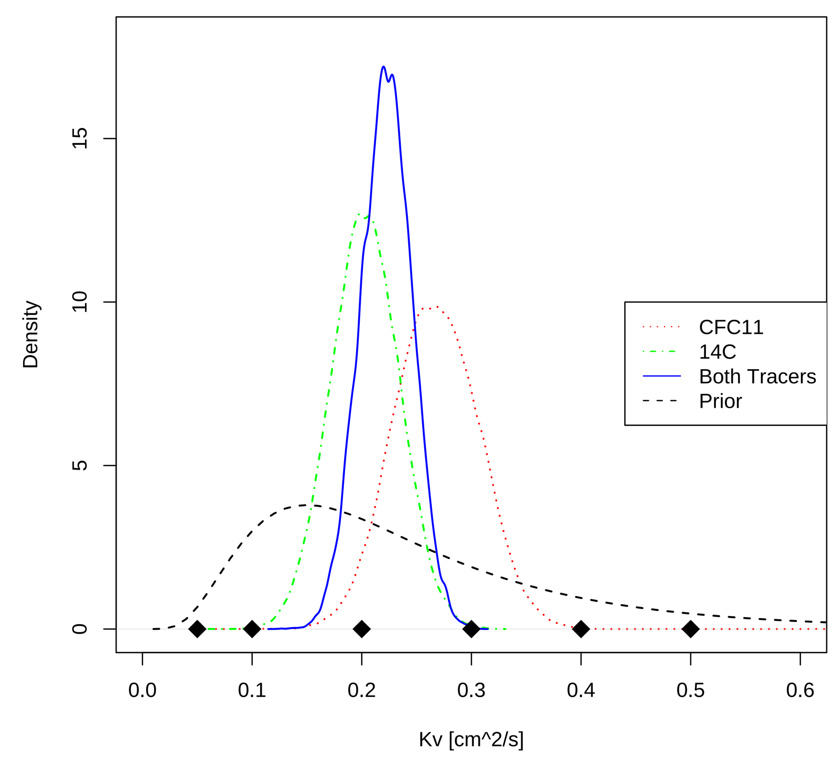
$$\begin{array}{llll}
\mathbf{B}_{ii} = \gamma_1 & \text{if } \ \mathbf{s}_{i2} < 200, \boldsymbol{Y}_{2i} < 2.1 & \mathbf{B}_{ii} = \gamma_2 & \text{if } \ \mathbf{s}_{i2} < 200, \boldsymbol{Y}_{2i} > 2.1 \\
\mathbf{B}_{ii} = \gamma_3 & \text{if } \ 200 < \mathbf{s}_{i2} < 800, \boldsymbol{Y}_{2i} < 1.8 & \mathbf{B}_{ii} = \gamma_4 & \text{if } \ 200 < \mathbf{s}_{i2} < 800, \boldsymbol{Y}_{2i} > 1.8 \\
\mathbf{B}_{ii} = \gamma_5 & \text{if } \ 800 < \mathbf{s}_{i2} < 2500 & \mathbf{B}_{ii} = 0 & \text{if } \ 2500 < \mathbf{s}_{i2} \\
\mathbf{B}_{ij} = 0 & \text{if } \ i \neq j & &
\end{array}$$

$\mathbf{s}_{i2}$ is the depth of the ocean corresponding to the output $(\boldsymbol{Y}_{1i}, \boldsymbol{Y}_{2i})$, and $\boldsymbol{Y}_{2i}$ is the CFC11 value from the climate model.

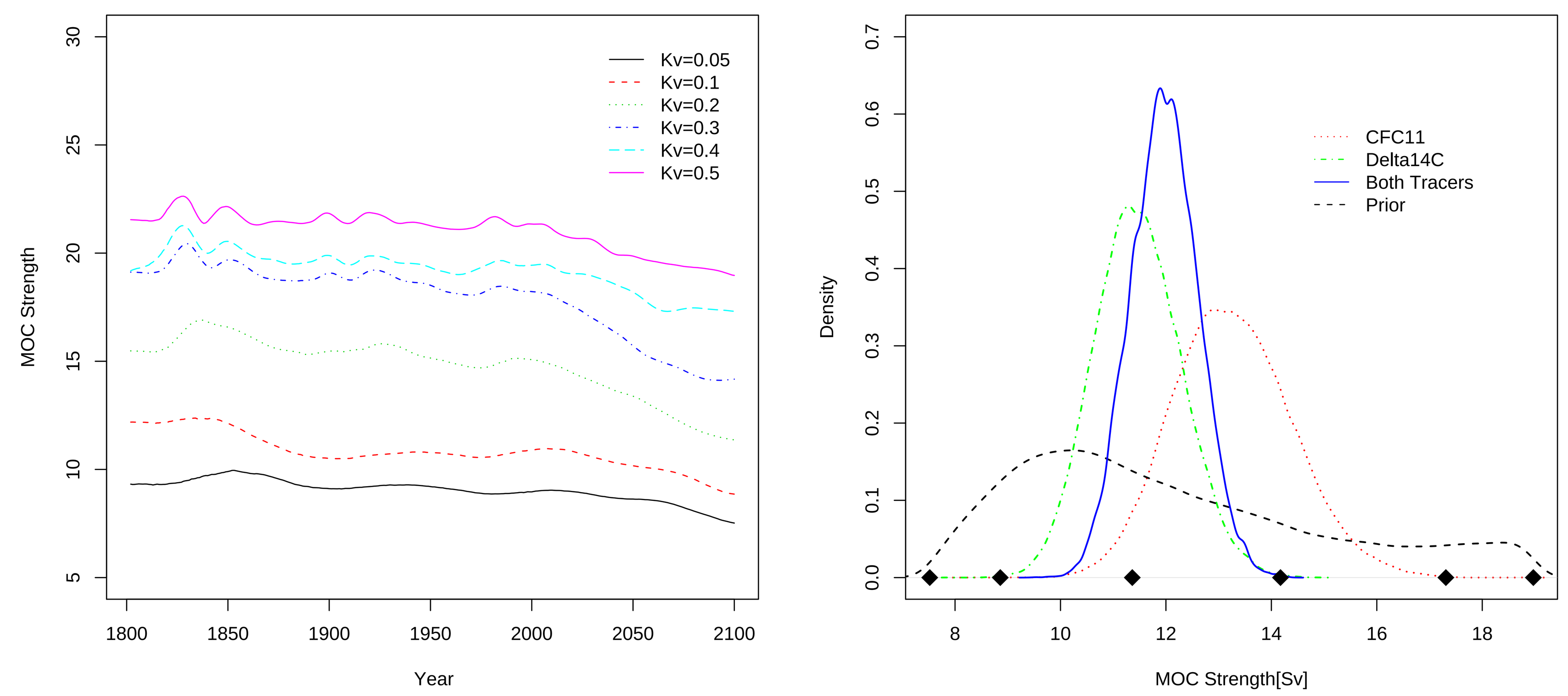
- Likelihoods for $\boldsymbol{Y}_1|\boldsymbol{Y}_2$ and $\boldsymbol{Y}_2$ have no common parameters, and can be maximized separately to obtain MLE.
- Use predictive distribution for $\boldsymbol{Y} = (\boldsymbol{Y}_1\ \boldsymbol{Y}_2)$ to obtain probability model connecting $\boldsymbol{Z} = (\boldsymbol{Z}_1\ \boldsymbol{Z}_2)$ to $\boldsymbol{\theta}$ as before.

Bivariate Model and Computational Issues

- The model for the second stage is $\boldsymbol{Z} = \hat{\boldsymbol{\eta}}(\boldsymbol{Z}^* \mid \boldsymbol{\theta}^*, \boldsymbol{Y}) + \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_1\ \boldsymbol{\epsilon}_2)^T$ is the observation error, $\boldsymbol{\epsilon}_i = (\epsilon_{1i}, \cdots, \epsilon_{Ni})^T \sim N(0, \psi_{mi} I_N)$, $Cov(\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2) = \mathbf{0}$.
- Computations are $\mathcal{O}(N^3)$, where N is the number of observations (could be tens of thousands).
- Used kernel mixing (Higdon (1998)) and Cressie and Johannesson (2008) to write covariance matrix as: $(\psi I_N + \mathbf{K}(I_J)^{-1}\mathbf{K}^T)$: $\mathbf{K}$ kernel matrix with rank J=196.
- Sherman-Woodbury-Morrison identity used to invert large matrices in above form reduces matrix inversions to $J \times J$, resulting in a computationally tractable model.
- Inference on $\boldsymbol{\theta}^*$,ψ_{m1}, and ψ_{m2} is performed using MCMC, computations simplified by using partitioned matrices.



Conclusion: Our model allows for inference based on *both* tracers jointly, while still remaining computationally tractable.



Left: MOC strength projections in sverdrups for several K_v values between 0.05 and 0.5, where MOC strength increasing as K_v increases. Right: Distribution of projected MOC strength(Sv) in 2100 (left) given posterior distributions of K_v for both single tracers and multiple tracers.

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*.