Inferring likelihoods and climate system characteristics from climate models and spatio-temporal tracer data

Murali Haran

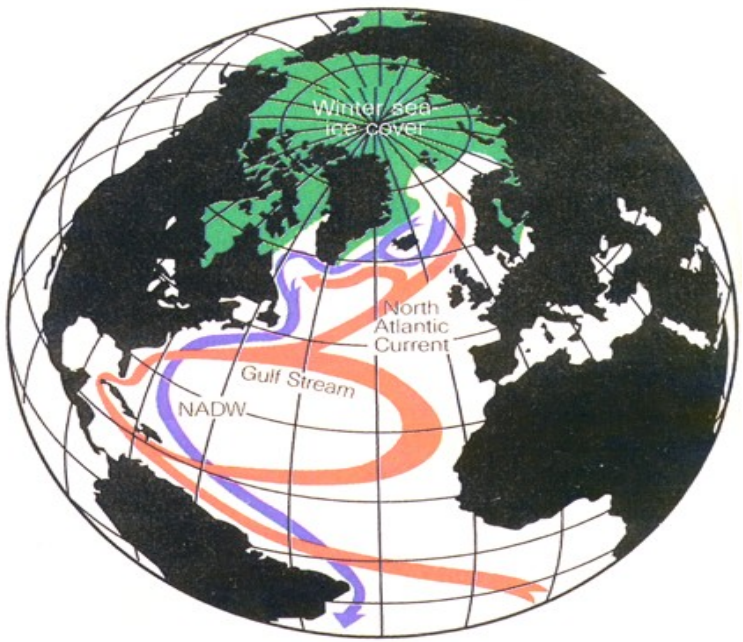
Department of Statistics
Penn State University

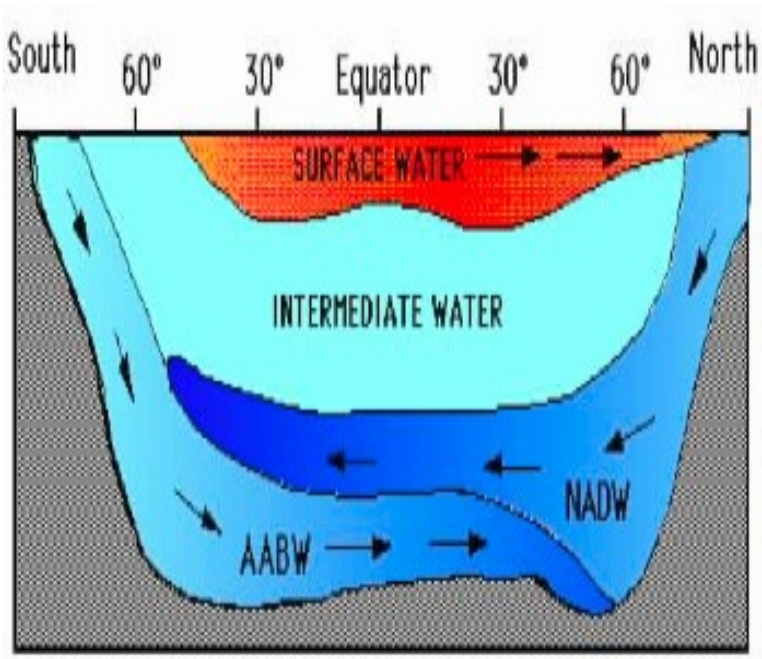
(joint work with Sham Bhat (Statistics), Roman Tonkonojenkov (Geosciences) and Klaus Keller (Geosciences))

September 2008

Motivation

- What is the risk of human induced climate change?
- Example of climate change: potential collapse of meridional overturning circulation (MOC).
- Early and accurate predictions of the risk of MOC collapse would save billions of dollars (Keller and McInerney, 2007)





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

Motivation-MOC

- MOC phenomenon: Movement of water from equator to higher latitudes, deep water masses created by cooling of water in Atlantic, resulting in sea ice formation. Result is denser salt water, which sinks, causing ocean circulation.
- MOC weakening results in disruptions in the equilibrium state in the climate leading to non-trivial changes. (e.g. French climate $\rightarrow$ Algerian climate.)
- Predictions of MOC strength can be made for particular climate parameter settings.
- These climate parameter values are difficult to measure directly. Two sources of indirect information:
  - Climate models: output at different parameter settings.
  - 'Tracers' of climate parameters: spatio-temporal data.

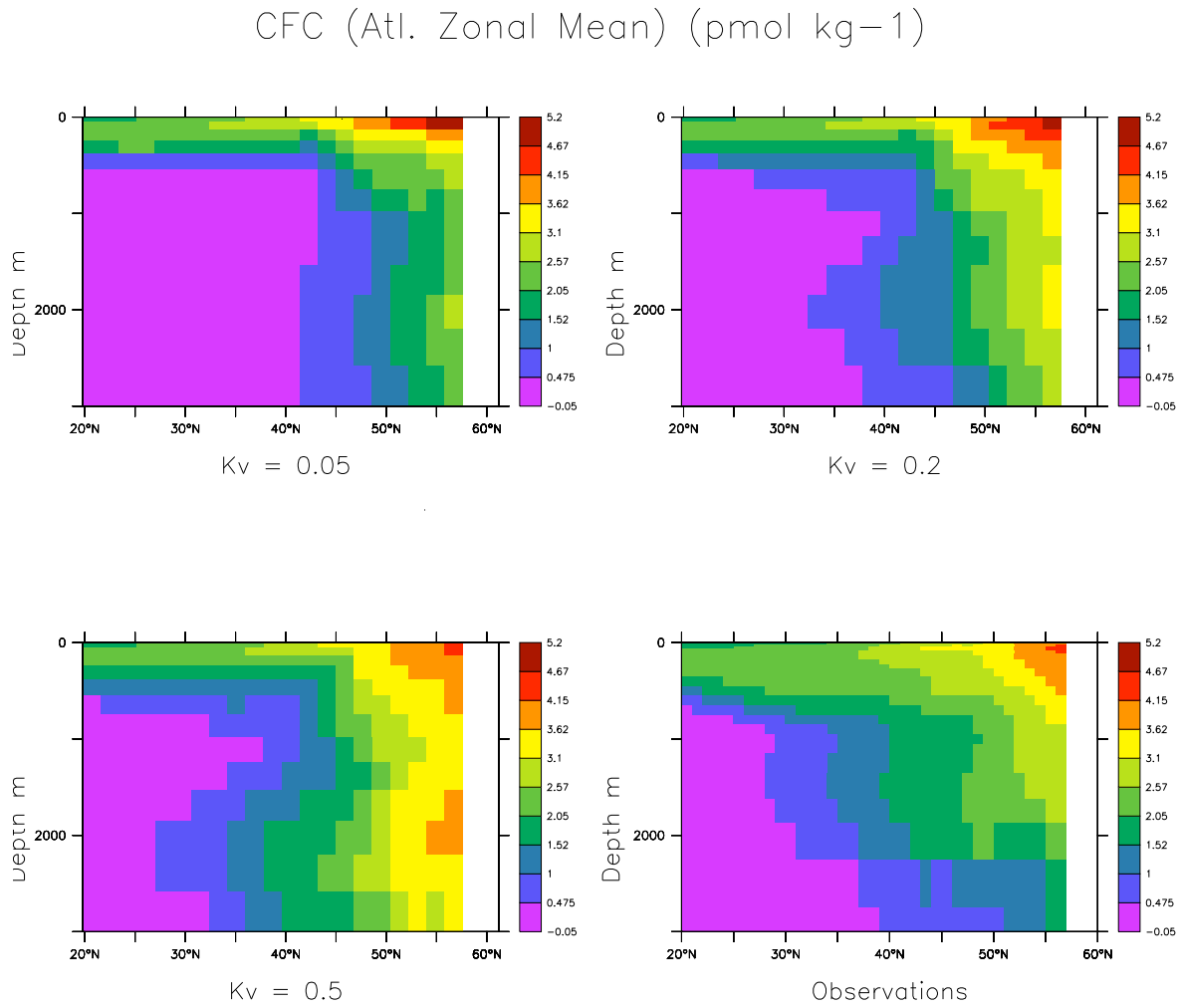
Motivation-CFC-11 Data

- Trichlorofluoromethane (CFC11), is often used as a tracer to understand deep ocean behavior such as MOC strength in the Atlantic Ocean.
- CFC-11 has high signal-to-noise ratio, is considered a stable tracer; unaffected by physical conditions and not produced in nature.
- The CFC11 data set was collected in the 1990s and consists of latitude, longitude, and depth values, averaged over longitudes.
- Prominent articles in Science and Nature (e.g. Knutti et al 2002) use ad-hoc, non-stochastic approaches which do not account for variability.

The Statistical Problem

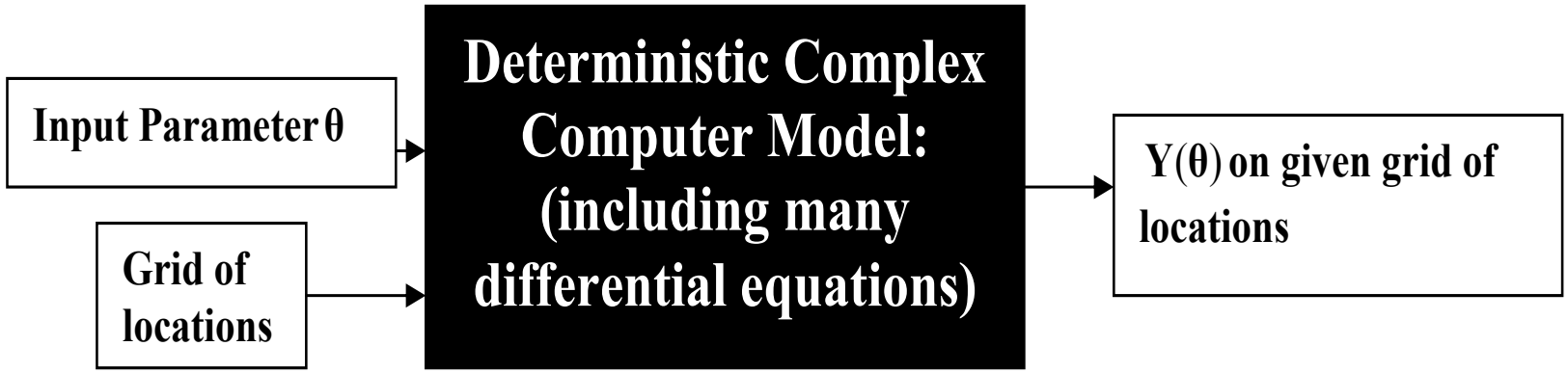
- **Goal**: Infer important climate characteristics (parameters) that drive major climate systems.
- Sources of information
  - Physical observations of climate system e.g. CFC
  - Output from complex climate models at several different climate parameters. e.g. for CFC data, climate model output at different values of vertical diffusivity.
- Challenges: No direct connection between observations and climate parameter. Rely on climate model, which is very complex, take months to run.
- Notation: $\mathbf{s}, t$ be space, time; θ be climate parameter.
  - $Z(\mathbf{s}, t)$: physical observations
  - $Y(\mathbf{s}, t, \theta)$: climate model output

CFC Example



- Climate model is run at several parameter settings of vertical diffusivity (K_v). Above: 3 settings; observations

Computer Model Emulation



- **Emulation** involves replacing a complicated computer model with a simpler (usually stochastic) approximation.
- Sacks et. al. (1989) introduced a linear Gaussian process model as an emulator for a complex nonlinear function. Related work by: Currin, Mitchell, Morris, Ylvisaker (1991), Bayarri et al (2007;2008) and many others.
- Advantage of emulation: obtain an approximate output at any parameter setting relatively easily along with associated uncertainty.

Gaussian Processes

- Model random variable at location $\mathbf{s}$ by

$$Z(\mathbf{s}) = X(\mathbf{s})\beta + w(\mathbf{s}), \text{ for } \mathbf{s} \in D \subset \mathbb{R}^d$$

- $\{w(\mathbf{s}), \mathbf{s} \in D\}$ is (infinite dimensional) Gaussian process.
- Let $\mathbf{w} = (w(\mathbf{s}_1), \ldots, w(\mathbf{s}_n))^T$, $\mathbf{Z} = (Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n))^T$, and $\mathbf{Z}^* = (Z^*(\mathbf{s}_1), \ldots, Z^*(\mathbf{s}_n))^T$.

$$\mathbf{w} \mid \xi \sim N(0, \Sigma(\xi)), \quad \xi \text{ are covariance parameters}$$

- $\mathbf{Z}^* \mid \mathbf{Z}$ is multivariate normal with mean and covariance:

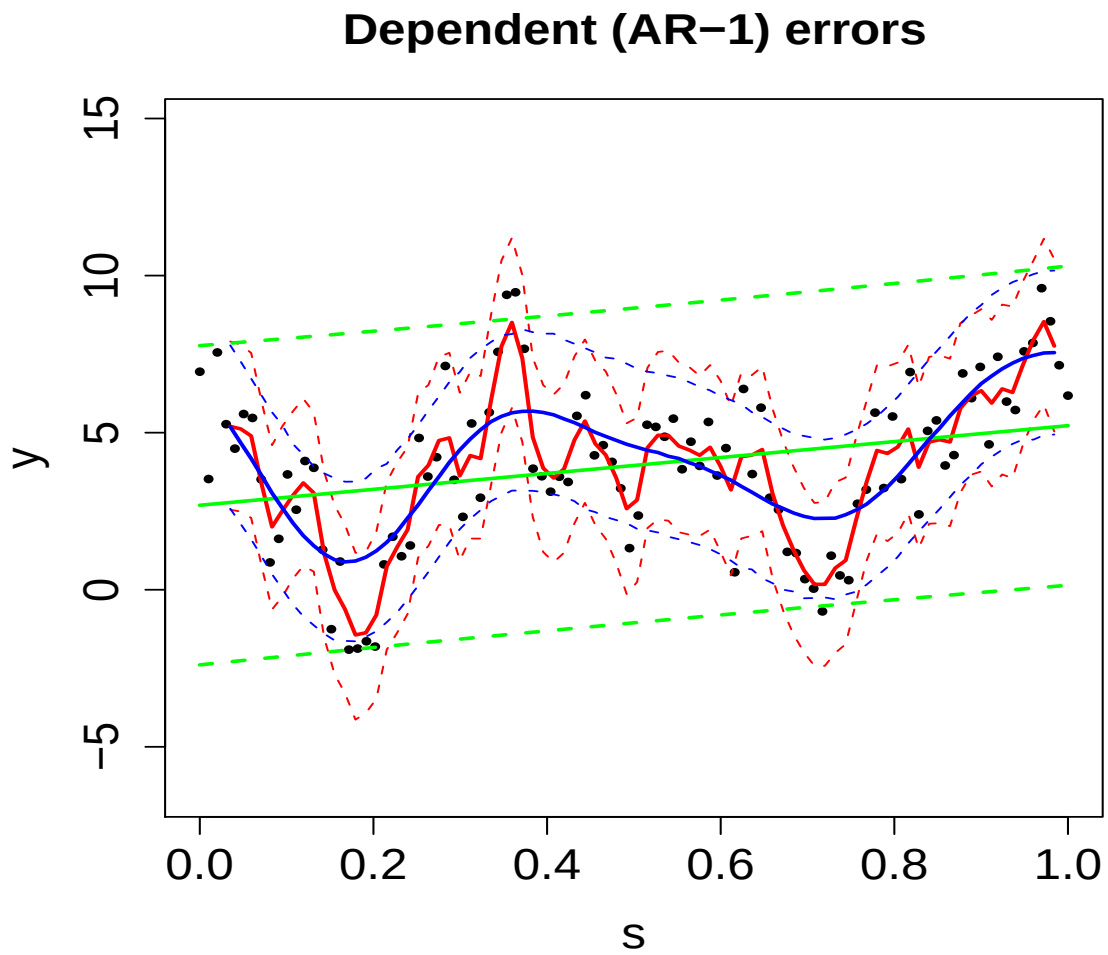
$$E(\mathbf{Z}^* \mid \mathbf{Z}, \beta, \xi) = \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(\mathbf{Z} - \mu_1)$$

$$\text{Cov}(\mathbf{Z}^* \mid \mathbf{Z}, \beta, \xi) = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}.$$

Gaussian Processes (contd)

- Standard assumption: Covariance function that determines $\Sigma(\xi)$ belongs to Matérn family.
- Predictions: obtain estimates $\hat{\xi}, \hat{\beta}$.
  - ML inference: plug $\hat{\xi}, \hat{\beta}$ into conditional distribution $\mathbf{Z}^* \mid \mathbf{Z}$.
  - Bayesian inference: find posterior $\pi(\xi, \beta \mid \mathbf{Z})$ and obtain *posterior predictive distribution* $\pi(\mathbf{Z}^* \mid \mathbf{Z})$, integrating with respect to β, ξ over the posterior distribution $\pi(\xi, \beta \mid \mathbf{Z})$.
  - Very convenient and very flexible models for both spatially dependent processes and complicated functions.

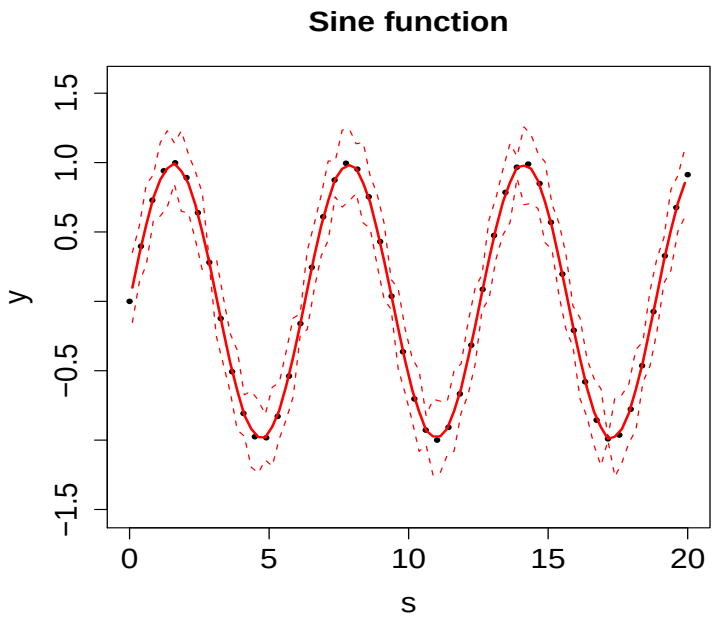
GP Model for Dependence: Toy 1-D example

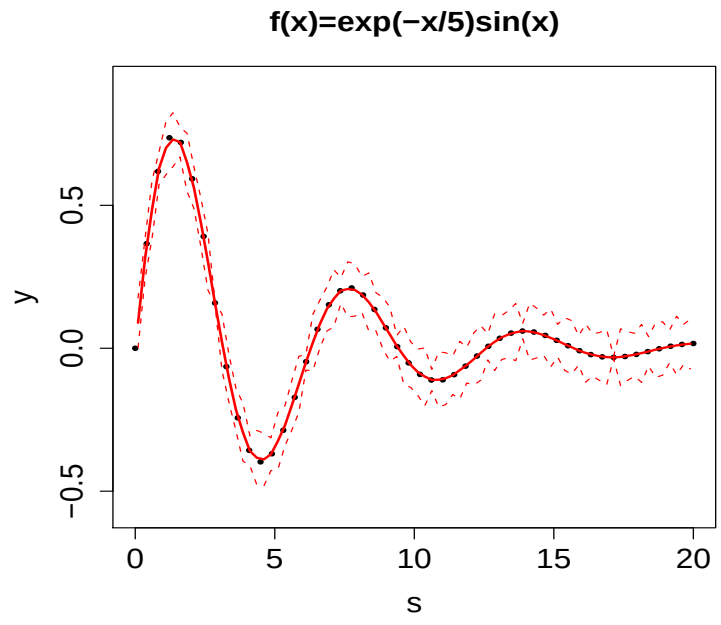


Black: 1-D AR-1 process simulation. Green: independent error.

Red: GP with exponential, Blue: GP with gaussian covariance.

GP Model for Emulation





Functions: $f(x) = \sin(x)$ and $f(x) = \exp(-x/5)\sin(x)$, both fit with stochastic models: linear GP model of simple form $f(x) = \alpha + \epsilon(x)$ where $\{\epsilon(x),\ x \in (0, 20)\}$ is a GP.

Joint Modeling Approaches

- Want to determine parameter settings that are 'most likely' given **Y**, **Z**. (the 'computer model calibration' problem).
- Kennedy and O'Hagan (2001) developed a model for this. Sanso et al. (2007) used a variant for climate model output (denoted as SFZ approach).
- Methods combine data, model output, model error, observational error, and any bias into a single model.
- Assumption of a "true" set of climate parameters θ^* exists.

$$Z(\mathbf{s_i}, t_i) = Y(\mathbf{s_i}, t_i, \theta^*) + \epsilon_i.$$

Note: there is no true θ^*, so perhaps more appropriate to think of it as a fitted value (Bayarri, Berger et al. 2007).

- Model **Y** and **Z** jointly. Model **Y** as a Gaussian process, with a mean dependent on climate parameter θ.

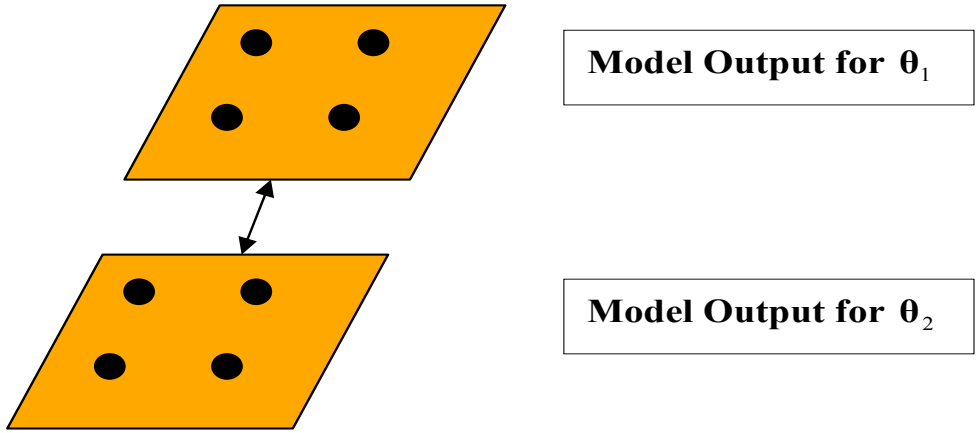
Joint Modeling Approaches

- $\boldsymbol{\epsilon} = (\epsilon_1, \cdots, \epsilon_n)^T$: observation error is modeled as Normal $(0, \psi\Sigma)$.
- Σ is a matrix estimated from control runs.

$$\text{Cov}(Y(\mathbf{s}_i, t_i, \boldsymbol{\theta}_{i'}), Y(\mathbf{s}_j, t_j, \boldsymbol{\theta}_{j'})) = \kappa\, \Sigma_{ij}\, r(\boldsymbol{\theta}_{i'}, \boldsymbol{\theta}_{j'})$$

- $\boldsymbol{\phi}_c = (\phi_{c1} \cdots \phi_{ck})$ are the climate covariance parameters.

$$r(\boldsymbol{\theta}_{i'}, \boldsymbol{\theta}_{j'}) = \prod_{m=1}^{k} \exp\left(-\frac{|\boldsymbol{\theta}_{i'm} - \boldsymbol{\theta}_{j'm}|}{\phi_{cm}}\right)$$



Joint Modeling Approaches

- Hence the joint distribution of $\mathbf{Z}$ and $\mathbf{Y}$ is a multivariate normal, and

$$\begin{bmatrix} \mathbf{Z} \\ \mathbf{Y} \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{M}(\boldsymbol{\theta}^*)^T \\ \mathbf{M} \end{bmatrix} \boldsymbol{\beta}, \begin{bmatrix} (\psi + \kappa) \otimes \Sigma & r(\boldsymbol{\theta}^*)^T \otimes \Sigma \\ r(\boldsymbol{\theta}^*) \otimes \Sigma & \mathbf{R} \otimes \Sigma \end{bmatrix} \right)$$

- Inference for $\boldsymbol{\theta}^*$, $\boldsymbol{\xi}_{st}$, etc is based on the posterior distribution $\pi(\boldsymbol{\theta}^*, \boldsymbol{\xi}_{st}, \boldsymbol{\phi}_c, \boldsymbol{\beta} | \mathbf{Z}, \mathbf{Y})$

$$\pi(\boldsymbol{\theta}^*, \boldsymbol{\xi}_{st}, \boldsymbol{\phi}_c, \boldsymbol{\beta} | \mathbf{Z}, \mathbf{Y}) \propto \mathcal{L}(\mathbf{Z}, \mathbf{Y} \mid \boldsymbol{\theta}^*, \boldsymbol{\xi}_{st}, \boldsymbol{\phi}_c, \boldsymbol{\beta}) \times p(\boldsymbol{\theta}^*) p(\boldsymbol{\xi}_{st}) p(\boldsymbol{\phi}_c) p(\boldsymbol{\beta})$$

  - $\mathcal{L}(\mathbf{Z}, \mathbf{Y} \mid \boldsymbol{\theta}^*, \boldsymbol{\xi}_{st}, \boldsymbol{\phi}_c, \boldsymbol{\beta})$: likelihood(multivariate normal)
  - $\boldsymbol{\xi}_{st} = (\psi, \kappa, \phi_s, \phi_t)$: covariance parameters.
- Priors: $\boldsymbol{\theta}^*$ based on scientific knowledge, other parameters are low precision priors (critical to do sensitivity analysis).

Computation

- $\pi(\boldsymbol{\theta}^*, \boldsymbol{\xi}_{st}, \boldsymbol{\phi}_c, \boldsymbol{\beta}|\mathbf{Z}, \mathbf{Y})$ is intractable, so rely on sample-based inference: Markov Chain Monte Carlo (MCMC).
- Computational bottleneck: matrix computations (e.g. Choleski factors) are of order $\mathcal{O}(N^3)$, where N is the number of observations (thousands).
- Kronecker products greatly reduces the computational burden. *Important*: This is brought about by assuming the same covariance Σ in modeling dependence among observations ($\mathbf{Z}$), computer model output ($\mathbf{Y}$) and in the block cross-covariance. This assumption is primarily due to computational considerations.
- Multimodality issues: used slice sampling (even more expensive).

Joint Modeling Approach: Pros and Cons

- Modelers (especially Bayesians) often argue that having a joint model is critical. Pragmatic argument: propogation of uncertainty through the model.
- Bayesian machinery and MCMC makes it relatively easy to write down a reasonable joint model.
- Identifiability issues and computational issues: unrealistic covariance assumptions and heavy spatial and temporal aggregation of both observations and model output.

Two Stage Approach

- Two stage approach to obtain posterior of θ:
  - Model the **Y**'s stochastically to 'infer a likelihood', connecting θ to **Y**.
  - Model **Z** using fitted model from above, with additional errors, biases, to infer θ (along with errors, biases.)
- Model **Y** as a Gaussian process emulator, with mean a linear function of θ.

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

- ξ is the set of covariance parameters, covariance function assumed to be separable among **s**, t, and θ.
- Covariance parameters:
  - Maximum likelihood estimates by optimization.
  - Bayesian approach: obtain posterior via MCMC.

Two stage approach (contd)

- For locations $(\mathbf{s}, t)$ at a given value of θ, we can then obtain the predictive distribution $\pi(\mathbf{Z}(\theta)^* \mid \mathbf{Y})$, multivariate normal for a *given* $\hat{\xi}, \hat{\beta}$ (MLE or posterior mean/mode). Otherwise this is not in closed form.
- This multivariate normal is our approximate likelihood, written explicitly with mean and variance as functions of θ from conditional distribution.

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

- 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, ϵ and δ are assumed to be independent. Strong prior information for ϵ can help identify the errors.
- We can now perform inference on θ^*.

Observations

- Our approach is counter to standard Bayesian modeling philosophy: instead of a coherent joint model, we are fitting models stagewise.
- Our approach can be seen as a way of 'cutting feedback' (Best et al. 2006; Rougier, 2008). Advantages:
  - Protecting emulator from a poor model of climate system.
  - Modeling emulator separately to facilitate careful evaluation of emulator. (Rougier, 2008).
- Principle: If we had a likelihood, $\mathcal{L}(\mathbf{Z}; \theta)$, we could perform inference for θ based on data $\mathbf{Z}$.
- Here: We are using climate model output ($\mathbf{Y}$) to 'infer' this likelihood and then perform standard likelihood-based inference. Intuitively: separate problems (see "Subjective likelihood" [Rappold, Lavine, Lozier, 2005.])

More advantages

- Computational advantages allows for relaxing unreasonable assumptions, e.g. no need to assume same covariance for both spatiotemporal dependence and observation error.
- Potentially helps with identification of variance/covariance components since not all parameters are being estimated/sampled at once; parameters estimated from first stage are fixed.
- Concern: are we ignoring crucial variability in parameter estimates by not propogating it as in the Bayesian formulation? We suspect this is not an important issue — flatness of likelihood surfaces lead to fairly similar prediction intervals for frequentist and Bayesian inference.

Large data sets

- Computational problems due to large climate model output and tracer observations.
- It is critical to use as much information as possible as this can inform discrepancies between the climate model and reality (based on the observations). Scientists are really interested in learning about these discrepancies; they can also inform decisions about model choice/averaging.
- Many potential alternatives: e.g. approximate likelihood (Vecchia, 1988; Caragea and Smith, 2002; Stein et al., 2004), kernel mixing (Higdon, 1998,2002; Paciorek and Schervish, 2006), frequency domain (Fuentes, 2007), sparse matrix approaches, (Cornford et al. 2005, Rue and Tjelmeland, 2002).

Kernel Mixing for Spatial Processes

- Alternatively, spatial data modeled, $w(\mathbf{s})$ via kernel mixing of white noise process(Higdon, 1998, 2001).
- New process created by convolving a continuous white noise process with a kernel, k, which is a circular normal.

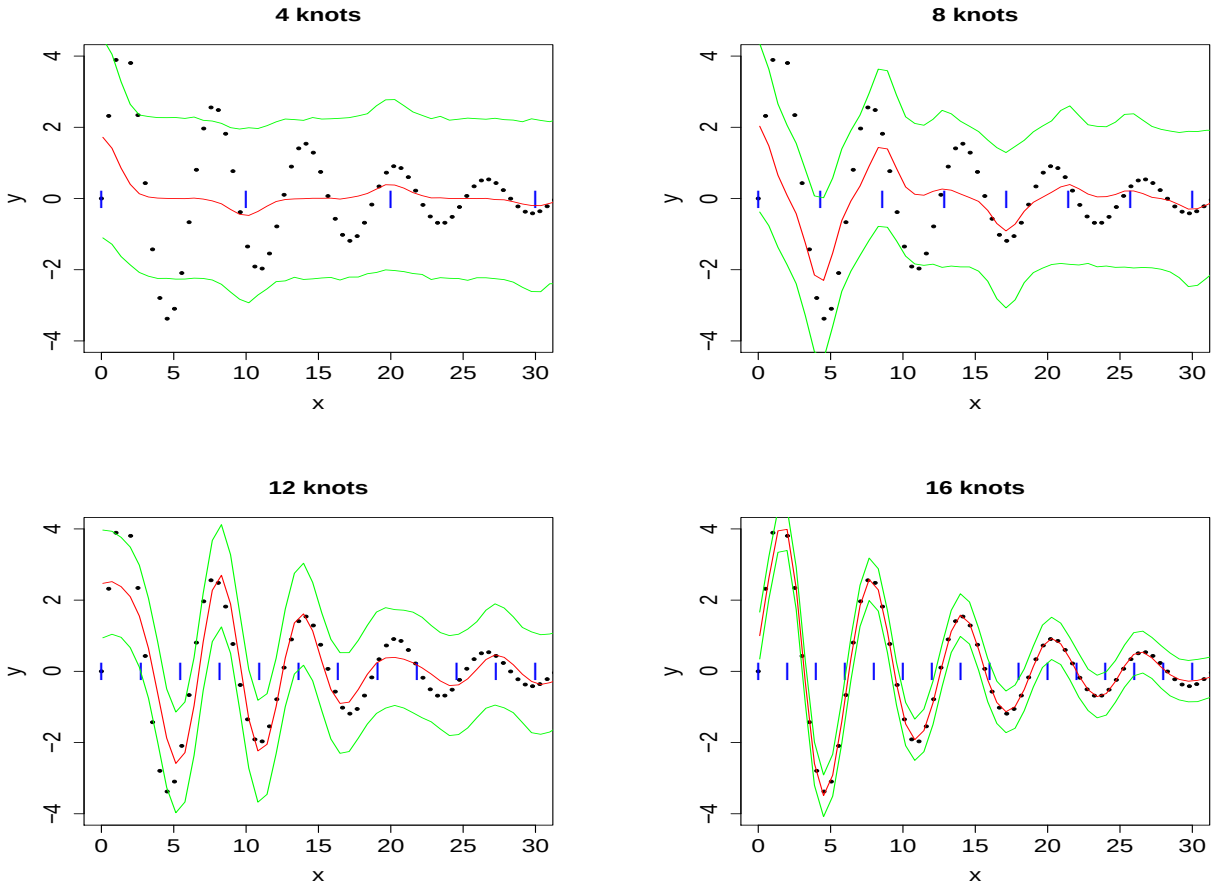
$$w(\mathbf{s}) = \int_D k(\mathbf{u} - \mathbf{s}) z(\mathbf{u}) d\mathbf{u}.$$

- Replace z by a finite sum approximation $\mathbf{z}$ defined on a lattice $\mathbf{u}_1, \ldots, \mathbf{u}_J$ (knot locations).

$$w(\mathbf{s}) = \sum_{j=1}^{J} k(\mathbf{u}_j - s) z(\mathbf{u}_j) + \mu(\mathbf{s}),$$

- Flexible: easily allows for nonstationarity and nonseparability.

Kernel Mixing for Spatial Processes (cont'd)



- Dimension reduction: Computation involves only the J random variables $z_1, \ldots, z_J$ at the locations $\mathbf{u}_1, \ldots, \mathbf{u}_J$.
- Figures are for 4, 8, 12, and 16 knots.

Kernel Mixing for Climate Model Output

- Extend kernel and knot process $\mathbf{z}$ to t and θ dimensions:

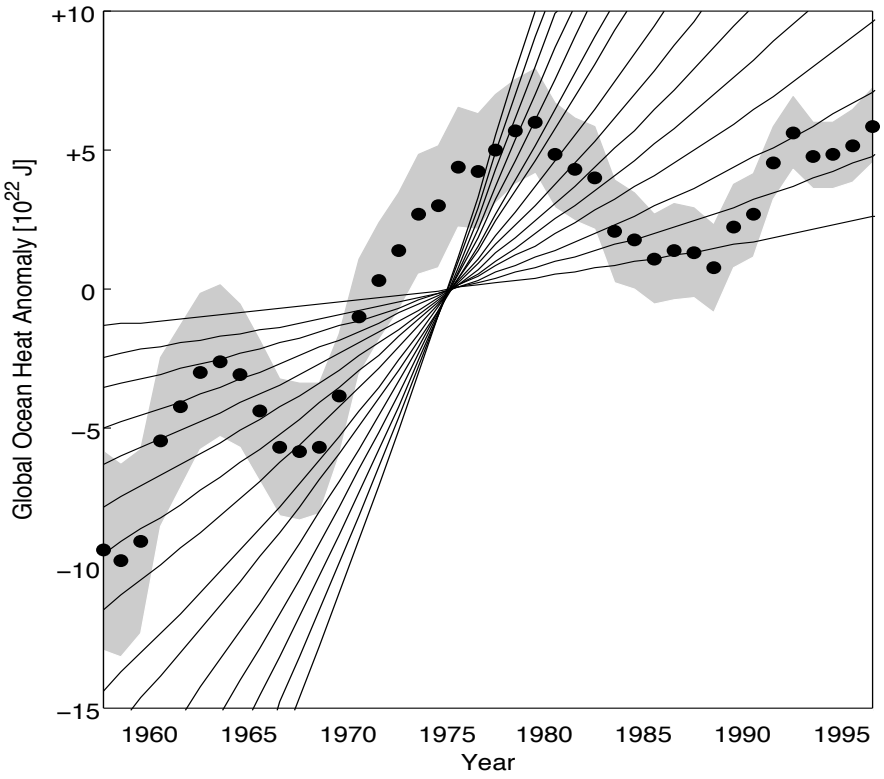
$$Y(\mathbf{s}, t, \theta) = \sum_{j=1}^{J} k(\mathbf{u}_j - \mathbf{s}; v_j - t, \ell_{1j} - \theta_1, \cdots \ell_{kj} - \theta_k) w(\mathbf{u}_j, v_j, \ell_j) + \mu(\theta)$$

- where the set of knots are $\mathbf{u}_j$, v_j, ℓ_j for $j = 1, \ldots, J$.
  $w(\mathbf{u}_j, v_j, \ell_j)$ is the process at the jth knot.
- The random field for $\mathbf{Y}(\mathbf{s}_i, t_i, \theta_i)$ is
  $\mathbf{Y}(\mathbf{s}_i, t_i, \theta_i) \mid \mathbf{w}, \psi, \kappa, \beta, \phi_s, \phi_t, \phi_c$

$$\sim N\left(\mathbf{X}(\theta_i)\beta + \sum_{j=1}^{J} K_{ij}(\phi_s, \phi_t, \phi_c) w(\mathbf{u}_j, v_j, \ell_j), \psi\right)$$

- Linear mean trend on θ and kernel is separable covariance function over $\mathbf{s}$, t, θ.

Example: Ocean Heat Anomalies

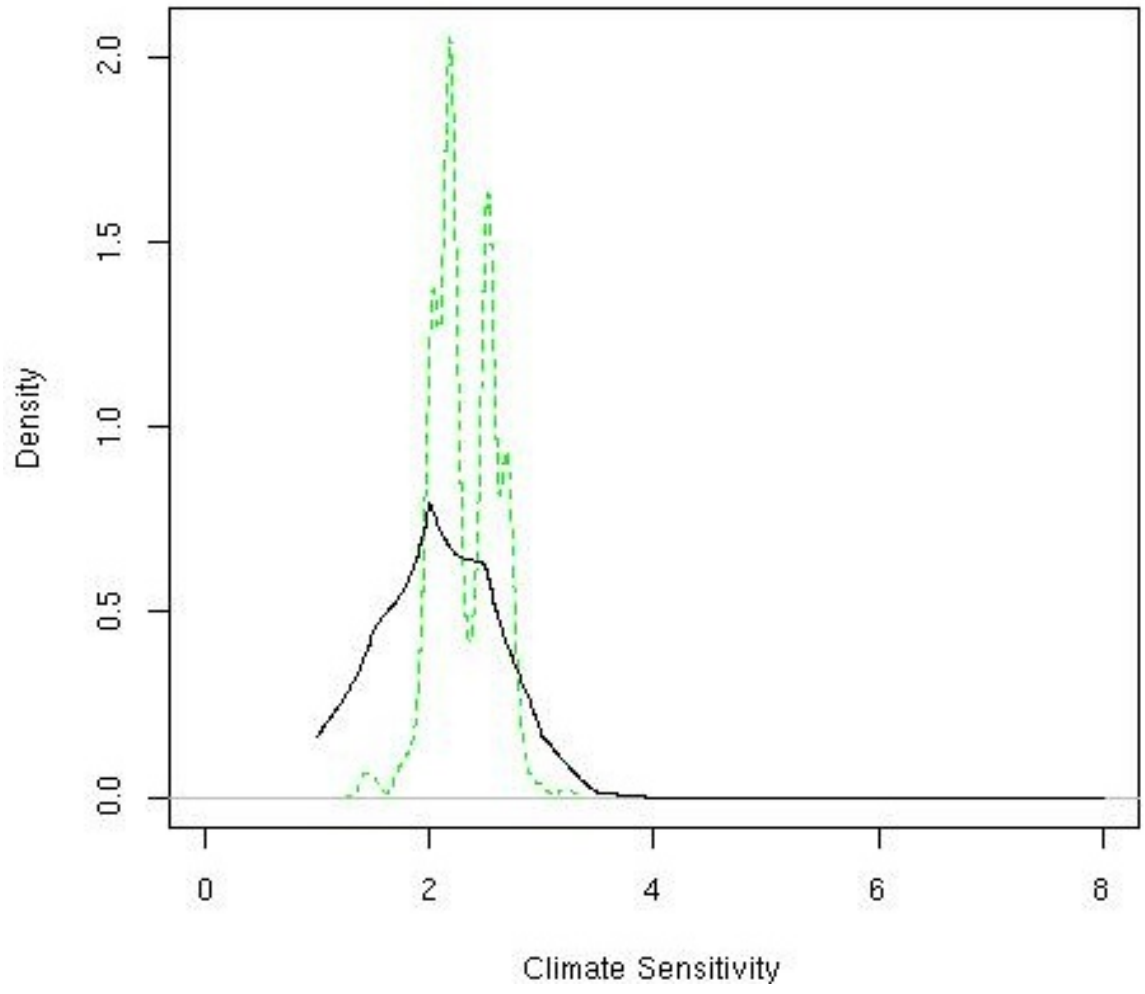


- Small data set from Levitus (2005): ocean heat anomalies over 40 years. Climate sensitivity($\mathcal{S}$) settings between 1 and 8 at intervals of 0.5. Goal: infer the distribution of $\mathcal{S}$.

Ocean Heat Anomalies: Application

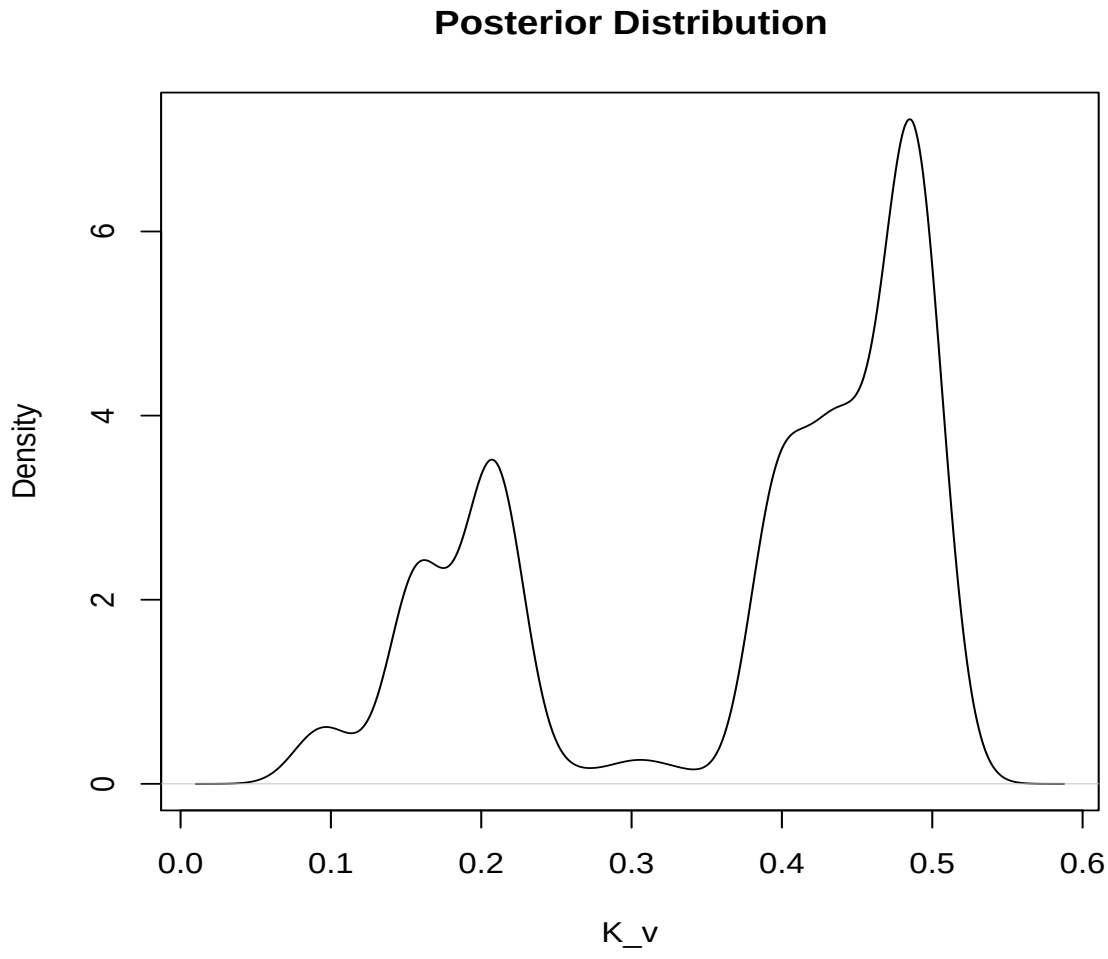
- Inferred likelihood approach: MLE plug in method used.
- In second stage,ϵ term but not δ included, and used continuous uniform prior on θ^* between 1 and 8, and relatively flat (low precision) IG prior on ψ.
- Extended SFZ approach also used a continuous uniform prior of θ^* between 1 and 8, and relatively flat priors on other covariance parameters.
- The MLE plug-in method requires more than 3 minutes to run. The second stage of the inferred likelihood method less than one minute.
- Extended SFZ method requires more than 10 minutes for 20000 iterations. The computational differences are more significant for bigger data sets.

Ocean Heat Anomalies: Posterior Distribution



- Solid black lines: Inferred likelihood method.
- Dotted green lines: SFZ method.

Preliminary Results for CFC Data



- One climate parameter θ, which is K_V ($cm^2 s^{-1}$), seven different settings from 0.05 to 0.5.

Summary

- Two main approaches considered for climate change inference and prediction:
  - Working on a critical extension of the SFZ approach: avoiding simplifying approach for ϵ.
  - Two stage 'inferred likelihood' approach that may result in improved computation and identifiability.
- Many open problems, research avenues including:
  - What is the best approach for resolving computational issues? e.g. kernel mixing, covariance tapering, spectral domain.
  - Bayesian model averaging approaches.
  - Flexible covariance functions, non-stationarity.
  - Multivariate space-time output.

Key References

- Kennedy, M.C. and O'Hagan, A.(2001), Bayesian calibration of computer models, *Journal of the Royal Statistical Society. Series B (with discussion)*
- Bayarri, M.J., Berger, J.O., Higdon, D., Kennedy, M.C., Kottas, A et al. (2007), A Framework for Validation of Computer Models *Technometrics.*
- Sanso, B. and Forest, C.E. and Zantedeschi, D (2007) , Inferring Climate System Properties Using a Computer Model, *Bayesian Analysis (with discussion)*.
- Higdon (1998) A process-convolution approach to modelling temperatures in the North Atlantic Ocean, *Environmental and Ecological Statistics*.