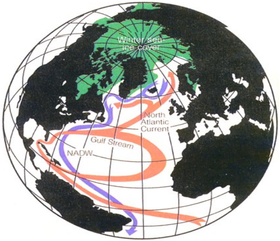
Gaussian Processes for Learning about Climate Model Parameters

Murali Haran

Department of Statistics
Pennsylvania State University

Center for Statistics and the Social Sciences
University of Washington
January 2012

The Atlantic Meridional Overturning Circulation



The Atlantic meridional overturning circulation (MOC) carries warm upper waters into far-northern latitudes and returns cold deep waters southward across the Equator.

Rahmstorf (Nature, 1997)

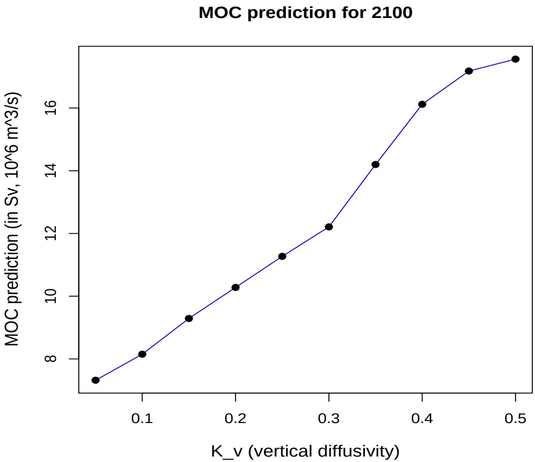
The MOC and Climate Change

- Its heat transport makes a substantial contribution to the moderate climate of maritime and continental Europe (cf. Bryden et al., 2005.)
- Any slowdown in the overturning circulation would have profound implications for climate change.

The MOC and K_v

- Collapse of MOC: example of potentially catastrophic climate change.
- Climate scientists use sophisticated climate models to make projections about the MOC.
- These models have many unknown parameters (inputs).
- A key source of uncertainty in MOC projections is uncertainty about *background ocean vertical diffusivity*, K_v.

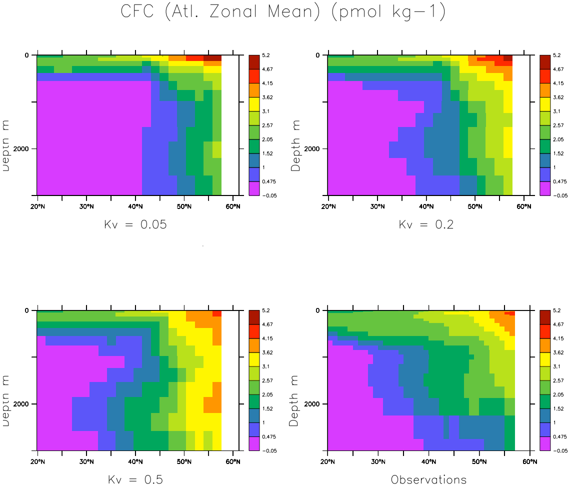
MOC predictions versus K_v



Learning about K_v

- K_v is a model parameter that quantifies the intensity of vertical mixing in the ocean. Cannot be measured directly.
- We work with two sources of indirect information:
  - **Observations** of two ocean 'tracers', both provide information about K_v: $\Delta^{14}C$ and trichlorofluoromethane (CFC11) collected in the 1990s.
  - **Climate model output** at different values of K_v from the University of Victoria (UVic) Earth System Climate Model (Weaver et. al. 2001).
- Data are in the form of spatial fields.
- Data size: for each of two tracers, 3706 observations and 5926 model output per input.

CFC Example



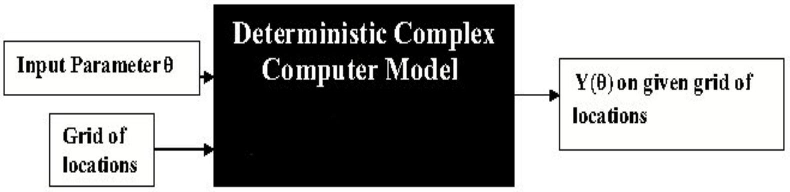
Bottom right corner: observations

Other plots: climate model output at 3 settings of K_v.

Challenges

1. No direct connection between observations and climate parameter. Connection provided by climate model runs.
2. The climate model is very computationally intensive. Hence, can only be run at a few different settings.
3. Large spatial data: pose computational challenges for inference.
4. Combining information from multiple tracers, CFC-11, $\Delta^{14}C$: need a computationally tractable model for flexible relationships between the spatial fields.

Computer Model Emulation



- **Emulation** replace a complicated computer model with a simpler approximation.
- Sacks et. al. (1989) introduced a linear Gaussian process model as an emulator. Also, Currin, Mitchell, Morris, Ylvisaker (1991), Bayarri et al (2007;2008), Sanso et al. (2008) and many others.

Gaussian Processes: Basics

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

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

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

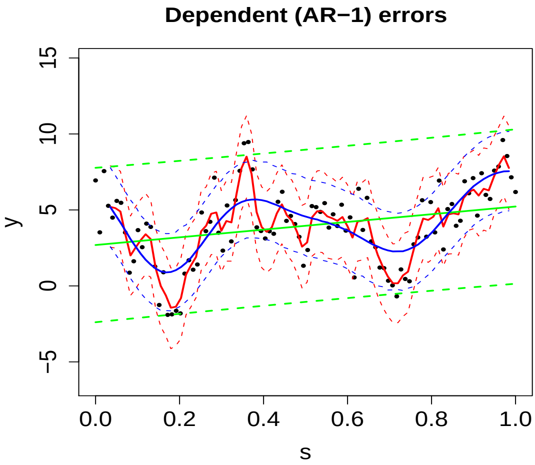
$$\mathbf{w} \mid \xi \sim N(0, \Sigma(\xi)), \;\; \xi \text{ are covariance parameters}$$

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

$$E(\mathbf{Z}^* \mid \mathbf{Z}, \boldsymbol{\beta}, \xi) = \boldsymbol{\mu}_2 + \Sigma_{21}\Sigma_{11}^{-1}(\mathbf{Z} - \boldsymbol{\mu}_1)$$

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

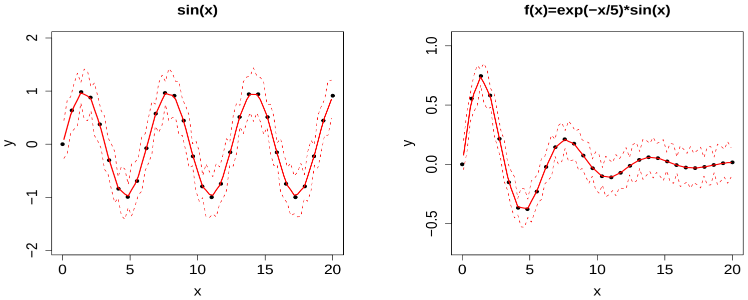
GP Model for Dependence: Toy 1-D Example



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

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

GP Model for Emulation: Toy 1-D Example



Functions: $f(x) = \sin(x)$ and $f(x) = \exp(-x/5)\sin(x)$.

Same model for both, $f(x) = \alpha + \epsilon(x)$, where

$\{\epsilon(x),\ x \in (0, 20)\}$ is a GP.

GPs: model dependence and complicated functions

Notation

- $Z_1(\mathbf{s}), Z_2(\mathbf{s})$: physical observations of tracer 1 and 2 at location $\mathbf{s}$=(latitude, depth).
  Let $\mathbf{Z}_1, \mathbf{Z}_2$ be the two spatial fields.
- $Y_1(\mathbf{s}, \boldsymbol{\theta})$, $Y_2(\mathbf{s}, \boldsymbol{\theta})$: model output for tracer 1 and 2 at location $\mathbf{s}$=(latitude, depth), and climate parameter $\boldsymbol{\theta}$.
  Let $\mathbf{Y}_1, \mathbf{Y}_2$ be the model output for the two tracers, spatial fields across multiple parameter settings.

Goal: Inference for climate parameter $\boldsymbol{\theta}$ using $\mathbf{Z}_1, \mathbf{Z}_2, \mathbf{Y}_1, \mathbf{Y}_2$.

Bayesian Computer Model Calibration

- Kennedy and O'Hagan (2001) developed a fully Bayes approach for computer model calibration. Sanso et al. (2008) develop a model for climate parameters.
- Assumption:

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

  Can think of θ^* as a fitted value (Bayarri, Berger et al. 2007).

Two-Stage Computer Model Calibration

Our approach

1. **Emulation**: Model relationship between $\mathbf{Z} = (\mathbf{Z}_1, \mathbf{Z}_2)$ and $\boldsymbol{\theta}$ via emulation of model output.
   - i An approximation to the computer model using $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2)$: $f(\mathbf{Y} \mid \boldsymbol{\theta})$
   - ii Take above approximation + systematic model-data discrepancy + error. This gives a model for the observations $\mathbf{Z}$: $f(\mathbf{Z} \mid \boldsymbol{\theta})$
2. **Calibration**: Infer $\boldsymbol{\theta}$ using $\mathbf{Z}$. Prior for $\boldsymbol{\theta}$ (p($\boldsymbol{\theta}$)) and likelihood from above gives posterior distribution

$$\pi(\boldsymbol{\theta} \mid \mathbf{Z}) \propto f(\mathbf{Z} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})$$

Emulation with Multiple Spatial Fields

- 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 relating $\mathbf{Y}_1$ and $\mathbf{Y}_2$, with parameters γ.
- The covariances of the Gaussian processes depend on both $\mathbf{s}$ (spatial distance) and $\boldsymbol{\theta}$ (distance in parameter space).
- $\boldsymbol{\beta}$s, $\boldsymbol{\xi}$s are regression, covariance parameters.

Very flexible relationship between $\mathbf{Y}_1$ and $\mathbf{Y}_2$.

Calibration with Multiple Spatial Fields

- Fit GP via maximum likelihood, then obtain predictive distribution at locations of observations.
- Model observations by adding measurement error and a model discrepancy term to the GP emulator:

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

  where $\boldsymbol{\delta}(\mathbf{Y}) = (\boldsymbol{\delta}_1 \; \boldsymbol{\delta}_2)^T$ is the model discrepancy, $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_1 \; \boldsymbol{\epsilon}_2)^T$ is the observation error.
- Model discrepancy term can make crucial adjustment to $\boldsymbol{\theta}$ estimates (Bayarri et al. 2007; Bhat et al., 2010).

Inference for $\boldsymbol{\theta}$

- Use Markov chain Monte Carlo (MCMC) to estimate $\pi(\boldsymbol{\theta} \mid \mathbf{Z}, \mathbf{Y})$, integrating 'out' remaining parameters.
- Separating emulation and calibration: 'modularization' (e.g. Liu et al., 2009).

Computational issues

- Need long MCMC runs.
- Matrix computations (every likelihood evaluation) are $\mathcal{O}(N^3)$ where N is matrix dimension.
- Create covariances with special structure + Sherman-Woodbury-Morrison identity used to reduce matrix computations.
- In MLE step: take advantage of structure of hierarchical model to reduce computations.

Computational Details: Kernel Mixing

- Model spatial dependence terms ($w(\mathbf{s})$) via kernel mixing (Higdon, 1998, 2001).
- New process created by convolving a continuous white noise process with a kernel, k (here: "circular normal")

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

- Replace original GP 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: allows non-stationarity and nonseparability.

Matrix Identities

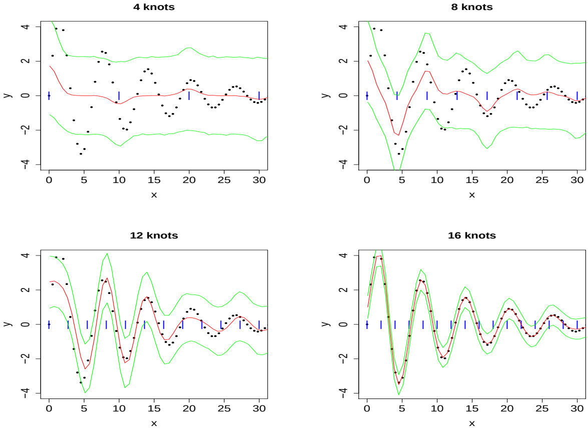
- Kernel mixing can be used to induce special matrix forms.
- Sherman-Woodbury-Morrison identity: useful if matrix is of form $A + UCV$, where A is easy to invert $N \times N$; U is $N \times J$; V is $J \times N$; C is $J \times J$. Inverse:

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}.$$

  Invert $J \times J$ rather than $N \times N$ matrices.

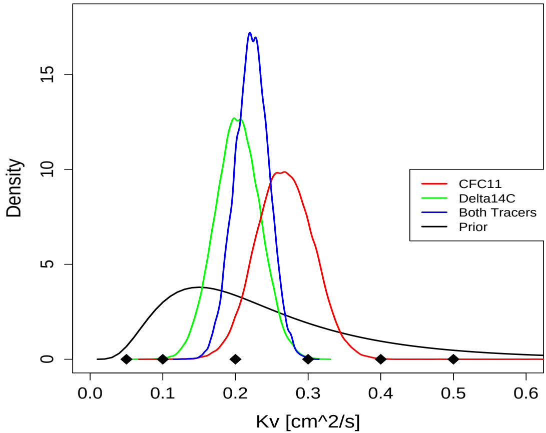
  For us: $J = 190$ versus $N =$4,500.

Kernel Mixing: Dimension Reduction



- 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.

Results for K_v Inference



posteriors: only CFC-11, only $\Delta^{14}C$, both CFC-11 & $\Delta^{14}C$.

Result: $\mathbf{K_v}$ pdf suggests weakening of MOC in the future.

Caveats

- Other parameters besides K_v also influence the MOC. We have fixed these at their current best estimates, thereby not accounting for uncertainties associated with them.
- This work is predicated on assumptions underlying the particular climate model. e.g. UVic has a simplified representation of internal climate variability.
- While we believe the data set used (Key et al., 2004) is best available, may be useful to consider uncertainties in data product which is based on relatively sparse observations.

Summary

1. Our approach:
   - Obtain a model connecting CFC-11, $\Delta^{14}C$ tracer observations to $\mathbf{K_v}$: fit a Gaussian process to climate model runs + account for other uncertainties, biases.
   - Using this model, infer a posterior density for $\mathbf{K_v}$ from data.
2. We model multivariate spatial data via a flexible hierarchical structure.
3. We use kernel mixing to obtain patterned covariances, making computations tractable for large data sets.

Inferred $\mathbf{K_v}$ can be used to project the MOC.

Some References

- Kennedy, M.C. and O'Hagan, A.(2001), Bayesian calibration of computer models, *JRSS(B)*.
- Sanso, B,. Forest, C.E., Zantedeschi, D (2008) , Inferring Climate System Properties Using a Computer Model, *BA*.
- Higdon (1998) A process-convolution approach to modelling temperatures in the North Atlantic Ocean, *Envir. Ecol. Statistics*.
- Bhat, K.S., Haran, M., Olson, R., Keller, K. (2012) "Inferring likelihoods and climate system characteristics using climate models and multiple tracers," under revision for *Environmetrics*
- Bhat, K.S., Haran, M., Goes, M. (2010) "Computer model calibration with multivariate spatial output"', *Frontiers of Statistical Decision Making and Bayesian Analysis*

Collaborators

- Sham Bhat, Los Alamos National Laboratories
- Roman Olson, Department of Geosciences, Penn State University
- Klaus Keller, Department of Geosciences, Penn State University

SUNDRIES

Joint modeling approach: pros and cons

- Bayesian machinery and MCMC makes it relatively easy to write down a reasonable joint model.
- Modelers (especially Bayesians) often argue that having a joint model is critical. Pragmatic argument: propogation of uncertainty through the model.
- However, joint model adds computational burdens. Also leads to identifiability issues. Hence, in order to build a joint model: have to resort to unrealistic covariance assumptions and heavy spatial and temporal aggregation of both observations and model output.

Alternative: Two stage approach

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

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

- $\boldsymbol{\xi}$ is the set of covariance parameters, covariance function assumed to be separable among **s**, t, and $\boldsymbol{\theta}$.
- Covariance parameters:
  - Maximum likelihood estimates by optimization.
  - Bayesian approach: obtain posterior via MCMC.

Two stage approach (cont'd)

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

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

- where $\boldsymbol{\delta}$ is the model error term and $\boldsymbol{\epsilon}$ is observation error.
- $\boldsymbol{\epsilon} \sim N(0, \psi I)$ and $\boldsymbol{\delta}$ is modeled as a Gaussian process, $\boldsymbol{\epsilon}$ and $\boldsymbol{\delta}$ are assumed to be independent. Strong prior information for $\boldsymbol{\epsilon}$ can help identify the errors.
- We can now perform inference on $\boldsymbol{\theta}^*$.

Observations

- Our approach is perhaps counter to standard Bayesian modeling philosophy: instead of a coherent joint model, we are fitting models stagewise.
- 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.])
- 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).

More advantages

- Computational advantages allow 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? Data sets/problems considered so far: not obvious that this is the case. (Also, cannot compare results for the large multivariate spatial data since cannot fit the joint model.)

Kernel mixing for climate model output

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

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

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

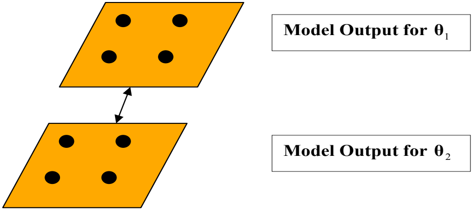
$$\sim N\left(\mathbf{X}(\boldsymbol{\theta}_i)\boldsymbol{\beta} + \sum_{j=1}^{J} K_{ij}(\phi_s, \boldsymbol{\phi}_c) w(\mathbf{u}_j, v_j, \boldsymbol{\ell}_j), \psi\right)$$

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

Bayesian model calibration (cont'd)

- Let observation error, $\boldsymbol{\epsilon} = (\epsilon_1, \cdots, \epsilon_n)^T$. Modeled as Normal $(0, \psi\Sigma)$, where Σ is estimated from other model runs (different runs from the ones used here; for e.g. 'control' runs that exclude human intervention/forcings.)
- Cov $(Y(\mathbf{s}_i, \boldsymbol{\theta}_{i'}), Y(\mathbf{s}_j, \boldsymbol{\theta}_{j'})) = \kappa \, \Sigma_{ij} \, r(\boldsymbol{\theta}_{i'}, \boldsymbol{\theta}_{j'})$.
- $\boldsymbol{\phi}_c = (\phi_{c1} \cdots \phi_{ck})$ are the climate covariance parameters.

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



Bayesian model calibration: inference

- Hence the joint distribution of **Z** and **Y** is a multivariate normal, and

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

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

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

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

Computation

- $\pi(\boldsymbol{\theta}^*, \boldsymbol{\xi}_s, \phi_c, \boldsymbol{\beta}|\mathbf{Z}, \mathbf{Y})$ is intractable, so rely on sample-based inference: Markov Chain Monte Carlo (MCMC).
- Computational bottleneck: matrix computations (e.g. Choleski factors) are of order N^3, where N is the dimension of the matrix.
- Kronecker products greatly reduce the computational burden. *Important*: This is brought about by assuming the same covariance Σ in modeling dependence among observations (**Z**), computer model output (**Y**) and in the block cross-covariance.

Future work

- Many open problems, research avenues including:
  - Combining information from multiple climate models: Multiresolution/multiscale modeling ideas, Bayesian model averaging.
  - Flexible covariance functions, non-stationarity.
  - Combining information from several tracers (e.g. 5–10).
- Other projects that can potentially borrow some of this methodology:
  - Atmospheric Science: Estimating mean temperature fields over the past millenia using proxies and climate models.
  - Infectious disease: inferring infectious disease dynamics from sparse observations and dynamic models.