Matching Heterogeneous Observations and Computational Models

Dorin Drignei
Department of Mathematics and Statistics
Oakland University

Outline

- MIT 2D climate model, used as a case study
- Climate model calibration: posterior pdf interpolation approach
- Climate model calibration: multidimensional emulator approach
- Connection between computer model calibration and sensitivity analysis, via multidimensional emulator
- Conclusions

Case study: MIT 2D climate model

- Developed at the Massachusetts Institute of Technology (MIT) Joint Program on the Science and Policy of Global Change.

- Two dimensional (latitude and vertical) atmospheric model coupled with a diffusive ocean model.

- Reproduces many of the nonlinear interactions occurring in simulations with 3D climate models but it requires less computational effort.

- Runs at about 4 hours computational time per model half-century on a 3GHz Pentium4 Linux workstation.

MIT 2D climate model: inputs and output

Inputs (unknown parameters):

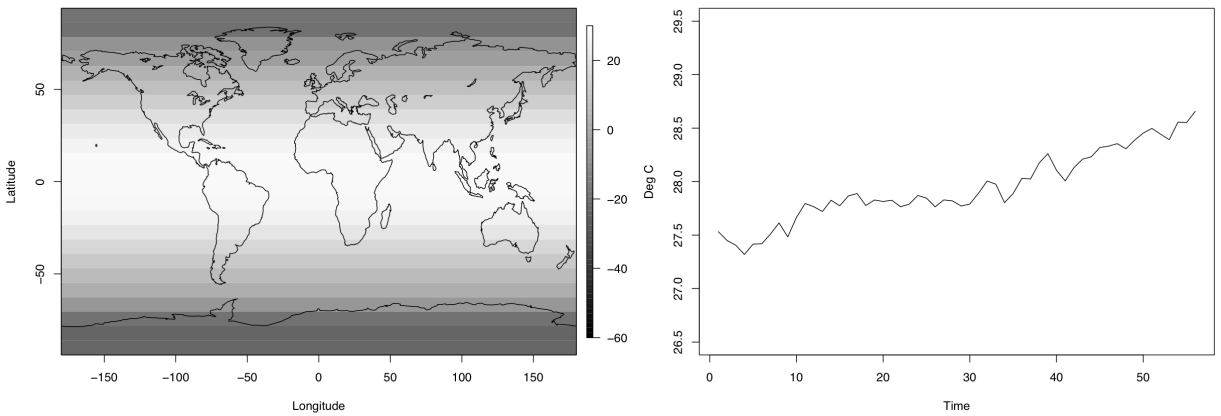
- $\theta = [S, K_v, F_{aer}]$
  - S: **Equilibrium climate sensitivity: global-mean surface temperature change if doubling CO_2 (^oC)**
  - K_v: **Global-mean vertical thermal diffusivity for the mixing of thermal anomalies into the deep ocean (cm^2/sec)**
  - F_{aer}: **Net aerosol forcing (W/m^2)**
- **Model is run at $D = 306$ inputs θ, sampled according to a space-filling design.**

Output at any input vector θ:

- **Surface temperature**
- **Ocean temperature**
- **Upper air temperature**

Example of climate model output data set

Model surface temperature at input $\theta = (10.5000, 0.3952, -0.2580)$



Left: **model surface temperature across 24 latitudes, at year 1950.**
Right: **time series of model surface temperature at Equator.**

Climate observations

- *Surface temperature:*
  - Longest, most spatially complete and documented, due primarily to the existence of meteorological stations throughout the world for a relatively long time.
  - Averaged, corresponding to 4 latitude bands by 5 decades.

- *Deep ocean temperature trend:*
  - Subsurface ocean temperature records are sparse and contain the most uncertainty due primarily to the difficulty in obtaining such temperature data sets.
  - Linear trend (a scalar) in the observed temperature record is retained for analysis.

- *Upper air temperature change:*
  - Data recorded by radiosondes, somewhat more more reliable than satellite-based MSU data because they provide a longer record and a better vertical resolution.
  - Difference in the 1986-1995 and 1961-1980 mean temperatures, recorded at 28 latitudes and at 8 pressure levels.

Statistical inference for climate models

- **In order to use the climate model to predict future climate, we need to calibrate it: estimate its parameters that give the "best" match between output and observations.**
- **Estimation of parameters θ in the nonlinear regression model**

$$\mathbf{Z}(t,s) \quad = \quad Y_\theta(t,s) + \epsilon(t,s)$$

$$\textit{Observed Climate} = \textit{Modeled Climate } (\theta) + \epsilon.$$

where Y_θ is the computationally intensive climate model, depending on unknown parameters θ.

POTENTIAL DIFFICULTY :

Iterative LS/likelihood optimization requires computing the expensive Y_θ possibly a large number of times !!

Aside: Univariate emulator outline

$$\textit{Input } \theta \rightarrow \boxed{\textbf{computer model}} \rightarrow \textit{Output } Y_\theta$$

- **Sample a small, limited number n of inputs (i.e. parameters) $\theta^{(i)}$, run the computer model and obtain the output data $Y_{\theta^{(i)}}$.**

- **Construct a statistical model for the sampled output data.**

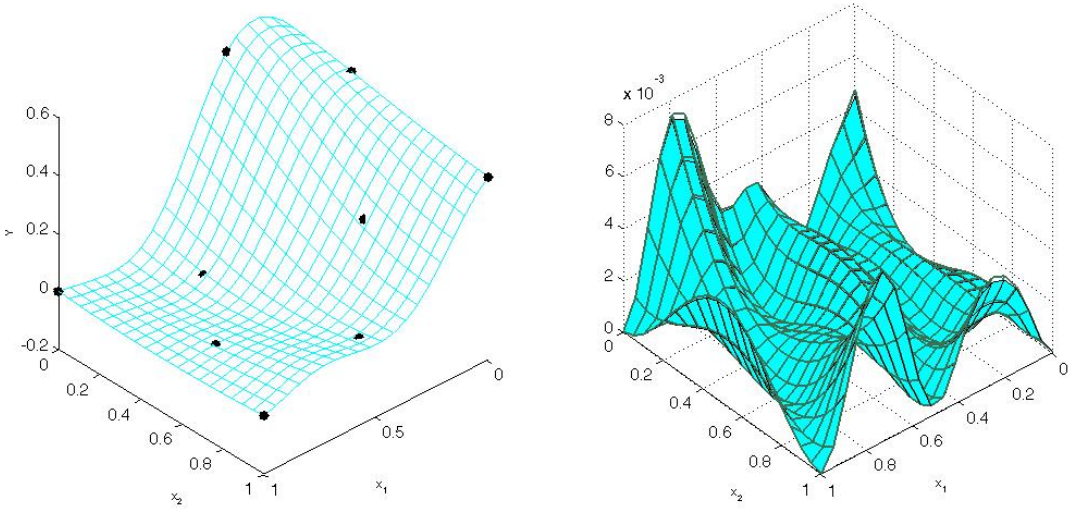
- **Use kriging methods to predict the computer model output data at new, not-sampled θ parameters.**

- **The predictor (or statistical surrogate) $\tilde{Y}_\theta$ will act as the emulator of the slow computer model. It is a linear function of data at the sampled inputs.**

Aside: Univariate emulator - a toy example

Study the effect of melting temperature x_1 and layer thickness x_2 on a utility index y.

- $Y \sim N(\mu 1, \Sigma)$ **(Distance-based covariance Σ).**
- **Simple kriging prediction at new input $\mathbf{x}_0$:**
  $Y_S = \mu + \Sigma_0' \Sigma^{-1}(Y - \mu)$, $\quad V_S = \Sigma_{0,0} - \Sigma_0' \Sigma^{-1} \Sigma_0$
  (the statistical approximation, or surrogate, and its error)



Approach 1: Interpolating the posterior pdf

- **Consider the initial model**

$$\mathbf{Z}(t,s) \quad = \quad Y_\theta(t,s) + \epsilon(t,s)$$

  with $\epsilon \sim MVN(\mathbf{0}, \nu^2 R_s \otimes R_t)$.

- **Compute this probability density function** $f(Z|\theta)$ **at a sample of** θ **values, by running the computationally intensive climate model. Denote these values** $f(Z|\theta^{(1)}), ..., f(Z|\theta^{(D)})$.

- **Assuming a prior distribution exists, compute the posterior joint pdf** $p(\theta^{(1)}|Z), ..., p(\theta^{(D)}|Z)$ **at the sampled** θ **values**

- **Find a emulator model for these posterior pdf scalar values, then obtain the predicted values** $\widehat{p}(\theta|Z)$ **of the posterior pdf at any** θ **value.**

- **Sample form** $\widehat{p}(\theta|Z)$ **to obtain posterior inference for** θ.

Posterior inference for $\theta = [S, K_v, F_{aer}]$

[10] **Probability density for observed data Z,**

$$f(Z|\theta) \propto \frac{1}{\sqrt{det(\nu^2 R_s \otimes R_t)}} exp\{-\frac{1}{2\nu^2}(Z - Y_\theta)'(R_s \otimes R_t)^{-1}(Z - Y_\theta)\}$$

R_s, R_t matrices of exponential correlations.

Z_S observed surface temperature change
Z_O observed deep ocean temperature trend
Z_U observed upper air temperature change

Overall joint pdf (conditional independence)

$$f(Z_S, Z_O, Z_U|\theta) = f(Z_S|\theta) f(Z_O|\theta) f(Z_U|\theta).$$

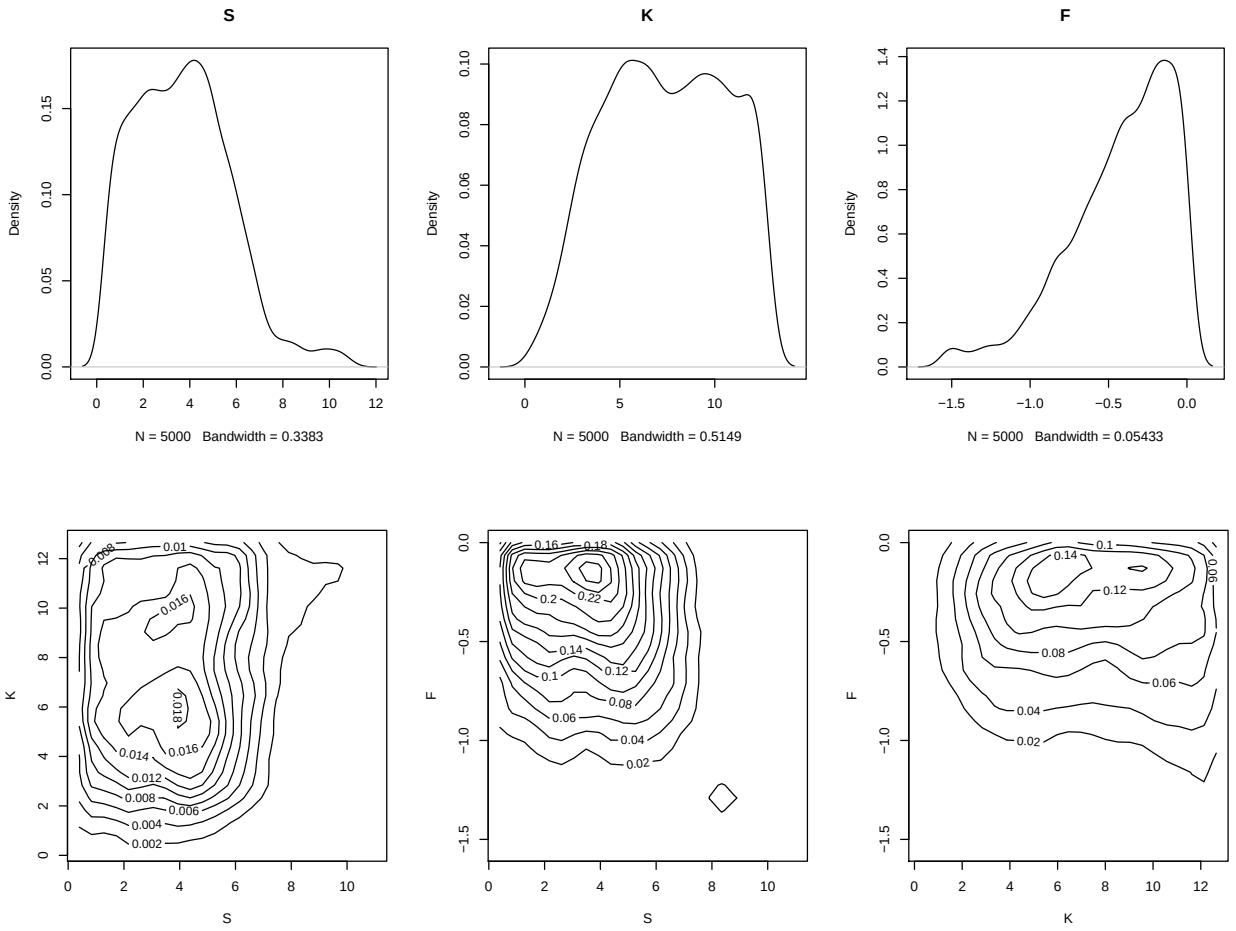
Prior distribution for θ (Forest et al, *Science*, 2000):

S : $beta(\alpha = 2.85, \beta = 14.0)$ scaled over the range 0.0 to 15.0 to represent the distribution of climate sensitivity.
K_v, F_{aer}: uniform

Results Approach 1

11 **Posterior intervals:**

S: (0.4000, 8.4040) K: (1.8070, 12.6491) F: (-1.2640, 0.0000)



Approach 2: Multidimensional emulator

- **Replace the computationally intensive computer model Y_θ with a computationally faster approximation $\tilde{Y}_\theta$, and account for the approximation error E.**

- **Computationally efficient nonlinear mixed model:**

$$z = \tilde{Y}_\theta + (Y_\theta - \tilde{Y}_\theta) + \epsilon$$

$$z = \tilde{Y}_\theta + E + \epsilon$$

- E **and** ϵ **independent errors**
- $E \sim MVN(\mathbf{0}, V_\theta)$
- $\epsilon \sim MVN(\mathbf{0}, \nu^2 R_s \otimes R_t)$

Multidimensional emulator: Statistical model

Surface temperature output data of size $N_S \times N_T \times D = 4 \times 5 \times 306$ **reshaped as vector.**

$$\mathbf{Y} \sim N(\mu(\mathbf{Y}), \mathbf{\Gamma})$$

- **A regression model** $\mu(\mathbf{Y}) = 1_D \otimes \mathbf{X}\beta,$
  where $\mathbf{X} = [1, \mathbf{S}, \mathbf{S^2}, \mathbf{T}, \mathbf{ST}, \mathbf{S^2T}]$

- **Covariance matrix** $\mathbf{\Gamma} = \mathbf{\Omega_\Theta} \otimes \mathbf{C_T} \otimes \mathbf{C_S}$
  where $\mathbf{\Omega_\Theta} = \sigma^2 \mathbf{C_\Theta} + \tau^2 \mathbf{I}$ **(climate signal + noise),**

$$\mathbf{C_\Theta} = [exp(-\eta_1(\theta_{i,1} - \theta_{j,1})^2 - \eta_2(\theta_{i,2} - \theta_{j,2})^2 - \eta_3(\theta_{i,3} - \theta_{j,3})^2)]_{i,j}$$

$$\mathbf{C_T} = [exp(-\eta_t|t(i) - t(j)|)]_{i,j}, \qquad \mathbf{C_S} = [exp(-\eta_s|s(i) - s(j)|)]_{i,j}$$

Multidimensional emulator: Estimation

- $\mathbf{Y}^r_{.,.}$ **is** $\mathbf{Y}$ **reorganized as a** $N_S N_T \times D$ **matrix**
- **Denote** $w_j = \frac{\sum_{i=1}^{D} (\mathbf{\Omega_\Theta}^{-1})_{i,j}}{\sum_{i,j=1}^{D} (\mathbf{\Omega_\Theta}^{-1})_{i,j}}$
- **Regression coefficients estimator**

$$\hat{\beta} = [\mathbf{X}'(\mathbf{C_T}^{-1} \otimes \mathbf{C_S}^{-1})\mathbf{X}]^{-1}\mathbf{X}'(\mathbf{C_T}^{-1} \otimes \mathbf{C_S}^{-1}) \frac{\sum_{i,j=1}^{D} (\mathbf{\Omega_\Theta}^{-1})_{i,j} \mathbf{Y}^r_{.,i}}{\sum_{i,j=1}^{D} (\mathbf{\Omega_\Theta}^{-1})_{i,j}}$$

- **Its variance**

$$var(\hat{\beta}) = (\sum_{i,j=1}^{D} w_i w_j (\mathbf{\Omega_\Theta})_{i,j})[\mathbf{X}'(\mathbf{C_T}^{-1} \otimes \mathbf{C_S}^{-1})\mathbf{X}]^{-1}$$

Multidimensional emulator: Estimation

Covariance parameters: minimize -2 *Log* **(***Likelihood***)/**DN_TN_S

$$\frac{log(Det(\mathbf{\Omega_\Theta}))}{D} + \frac{log(Det(\mathbf{C_T}))}{N_T} + \frac{log(Det(\mathbf{C_S}))}{N_S} + \frac{(\mathbf{Y} - \mathbf{1} \otimes \mathbf{X}\hat{\beta})'\mathbf{\Gamma}^{-1}(\mathbf{Y} - \mathbf{1} \otimes \mathbf{X}\hat{\beta})}{DN_TN_S}$$

Regression coefficients estimates and standard errors

	β_0	β_1	β_2	β_3	β_4	β_5
Estimate	-42.60698	10.41124	-0.40582	0.02066	-0.00177	0.00008
Std. Err.	0.58082	0.10706	0.00415	0.00689	0.00127	0.00005

Covariance parameters estimates

Model	η_1	η_2	η_3	σ^2	τ^2	η_s	η_t
Regression	4.9626	11.9720	4.5164	3.5331	0.0168	28.7669	0.2416

Multidimensional emulator: Simple kriging prediction

Conditional multivariate normal distribution = prediction distribution at a set Π of P new inputs.

$$Mean: \quad \tilde{\mathbf{Y}}_{\Pi} = \mathbf{1}_P \otimes \mathbf{X}\beta + \mathbf{\Gamma}_{\Pi\Theta}\mathbf{\Gamma}^{-1}(\mathbf{Y} - \mathbf{1}_D \otimes \mathbf{X}\beta)$$

$$Covariance \quad Matrix: \quad \mathbf{V}_{\Pi}{}^{s} = \mathbf{\Gamma}_{\Pi} - \mathbf{\Gamma}_{\Pi\Theta}\mathbf{\Gamma}^{-1}\mathbf{\Gamma}'_{\Pi\Theta}$$

where

$$\mathbf{\Gamma}_{\Pi} = (\sigma^2\mathbf{C}_{\Pi} + \tau^2\mathbf{I}) \otimes \mathbf{C}_T \otimes \mathbf{C}_S$$

and

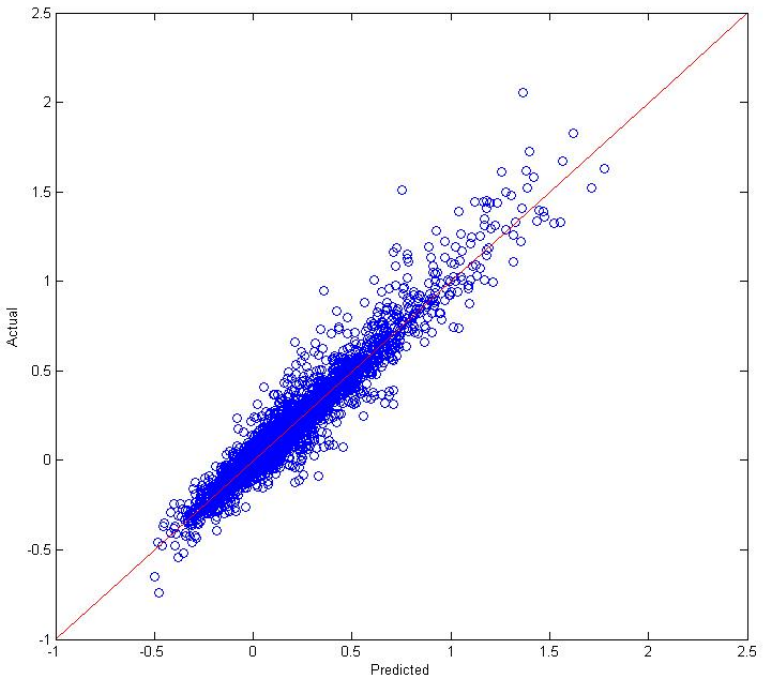
$$\mathbf{\Gamma}_{\Pi\Theta} = \sigma^2(\mathbf{C}_{\Pi\Theta} \otimes \mathbf{C}_T \otimes \mathbf{C}_S).$$

Here

$$\mathbf{C}_{\Pi} = [exp(-\eta_1(\pi_{i,1} - \pi_{j,1})^2 - \eta_2(\pi_{i,2} - \pi_{j,2})^2 - \eta_3(\pi_{i,3} - \pi_{j,3})^2)]_{i,j}$$

$$\mathbf{C}_{\Pi\Theta} = [exp(-\eta_1(\theta_{i,1} - \pi_{j,1})^2 - \eta_2(\theta_{i,2} - \pi_{j,2})^2 - \eta_3(\theta_{i,3} - \pi_{j,3})^2)]_{i,j}$$

Multidimensional emulator: Cross Validation



Computational times at a new input:

- **original climate model: 4 hours**
- **statistical surrogate: less than one minute.**

Approach 2: Multidimensional emulator

- **Replace the computationally intensive nonlinear function Y_θ with the previous computationally faster approximation $\tilde{Y}_\theta$, and account for the approximation error E.**

- **Computationally efficient nonlinear mixed model:**

$$z = \tilde{Y}_\theta + (Y_\theta - \tilde{Y}_\theta) + \epsilon$$

$$z = \tilde{Y}_\theta + E + \epsilon$$

- E **and** ϵ **independent errors**
- $E \sim MVN(\mathbf{0}, V_\theta)$
- $\epsilon \sim MVN(\mathbf{0}, \nu^2 R_s \otimes R_t)$

Posterior inference for $\theta = [S, K_v, F_{aer}]$

[19] **Probability density for observed data Z,**

$$f(Z|\theta) \propto \frac{1}{\sqrt{det(V_\theta + \nu^2 R_s \otimes R_t)}} exp\{-\frac{1}{2}(Z - \tilde{Y}_\theta)'(V_\theta + \nu^2 R_s \otimes R_t)^{-1}(Z - \tilde{Y}_\theta)\}$$

R_z, R_t matrices of exponential correlations.

Z_S observed surface temperature change
Z_O observed deep ocean temperature trend
Z_U observed upper air temperature change

Overall joint pdf (conditional independence)

$$f(Z_S, Z_O, Z_U|\theta) = f(Z_S|\theta) f(Z_O|\theta) f(Z_U|\theta).$$

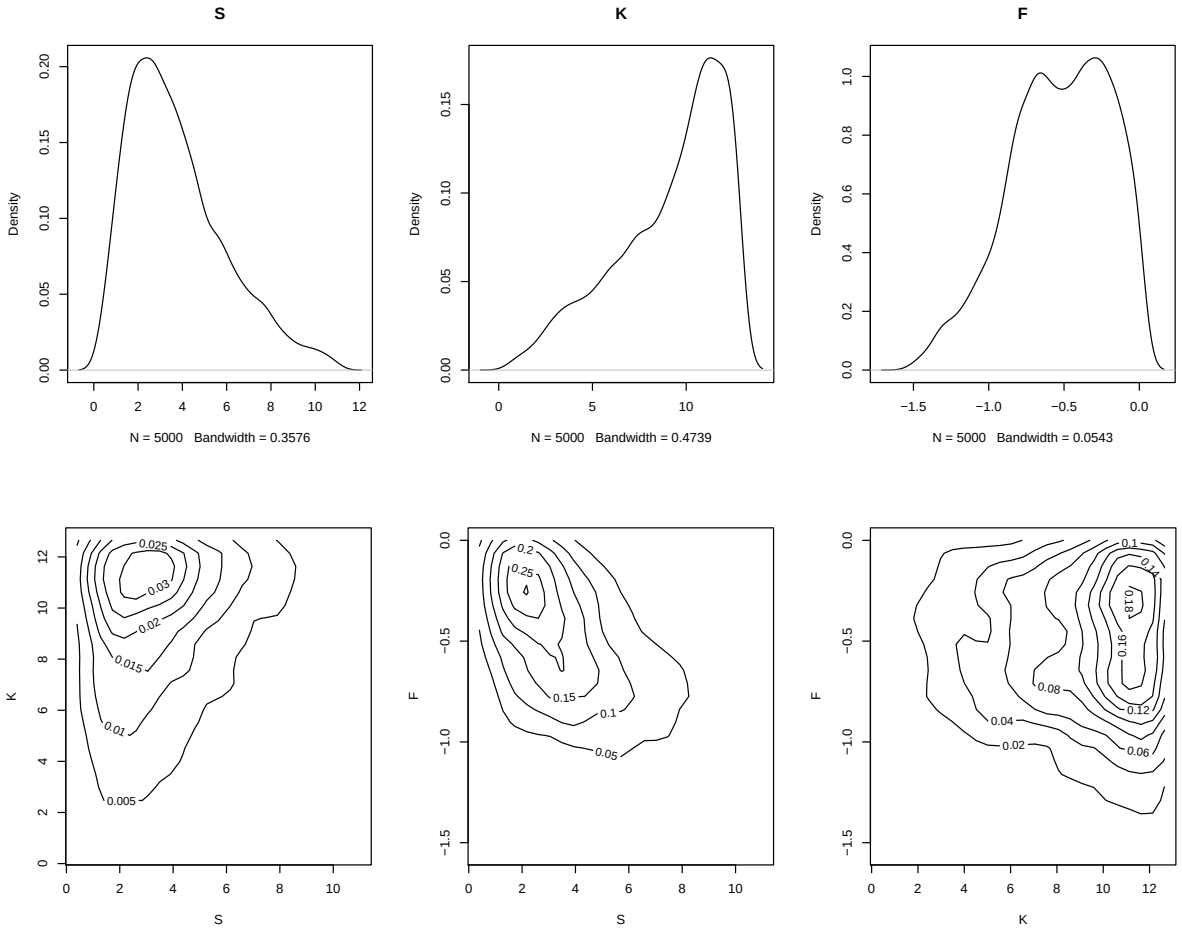
Prior distribution for θ (Forest et al, *Science*, 2000):

S : $beta(\alpha = 2.85, \beta = 14.0)$ scaled over the range 0.0 to 15.0 to represent the distribution of climate sensitivity.
K_v, F_{aer}: uniform

Results Approach 2

20 **Posterior intervals:**

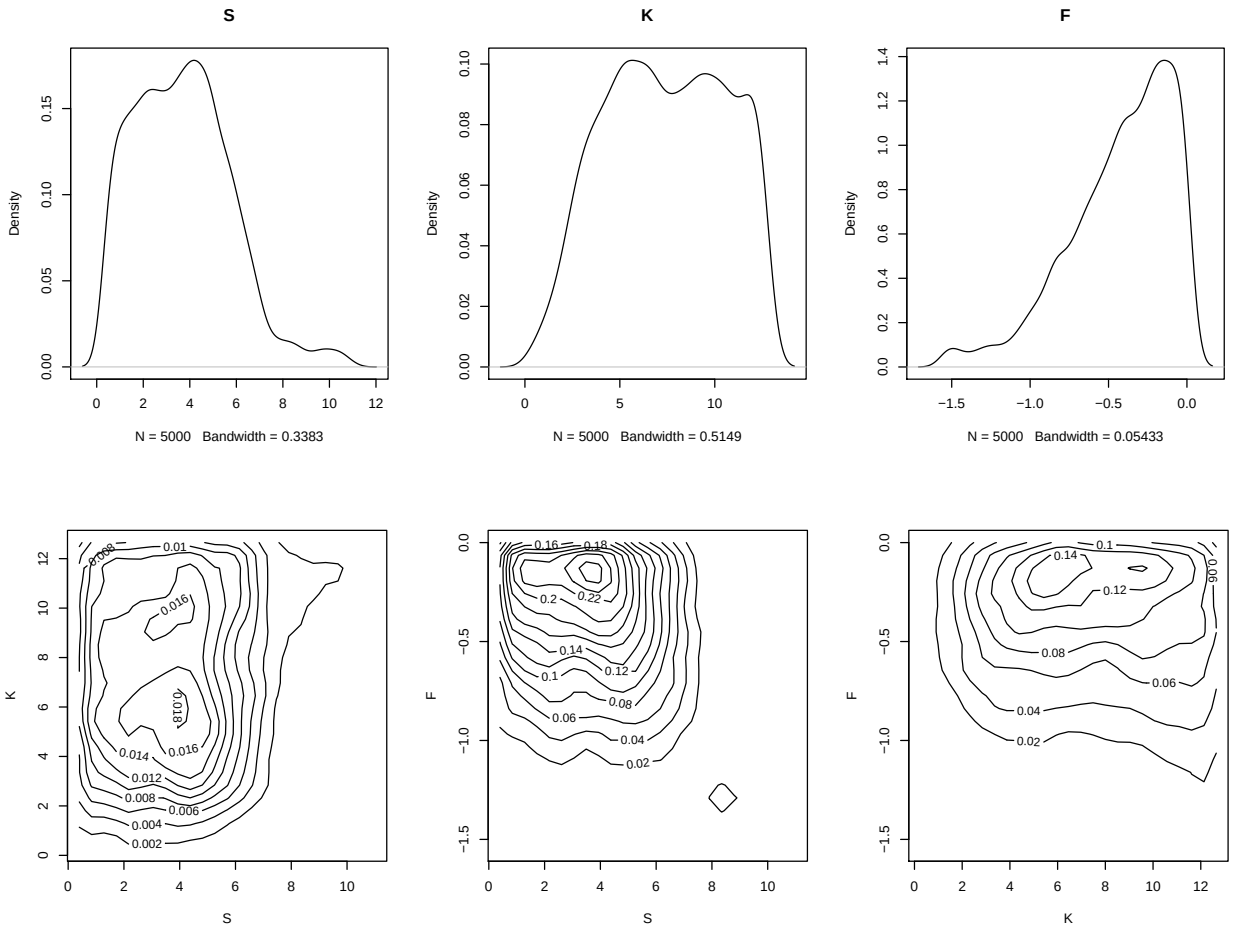
S: (0.7655, 9.1724) K: (2.6170, 12.6491) F: (-1.2814, 0.0000)



Results Approach 1

21 **Posterior intervals:**

S: (0.4000, 8.4040) K: (1.8070, 12.6491) F: (-1.2640, 0.0000)



Sensitivity analysis for multidimensional output

- **Determine the most influential parameters on output indexed at $(t_k)_k$. Input vector $\theta = (\theta_1, ..., \theta_r)$, Output $Y_k = f(t_k, \theta_1, ..., \theta_r)$**
- **Main effect of i^{th} factor at t_k is**

$$M_{i,t_k}(\theta_i) = \int Y_k d\theta_{-i} - \int Y_k$$

- **Variance for i^{th} factor at t_k is defined as**

$$V_i(t_k) = \int M^2_{i,t_k}(\theta_i) d\theta_i.$$

- **Second order interactions at t_k**

$$M_{i,j,t_k}(\theta_i, \theta_j) = \int Y_k d\theta_{-i} d\theta_{-j} - \int Y_k d\theta_{-i} - \int Y_k d\theta_{-j} + \int Y_k$$

- **Output variance component at t_k**

$$V_{ij}(t_k) = \int M^2_{i,j,t_k}(\theta_i, \theta_j) d\theta_i d\theta_j.$$

Sensitivity analysis for multidimensional output

- **Functional ANOVA decomposition of the output**

$$\sum_k V_Y(t_k) = \sum_k \sum_i V_i(t_k) + \sum_k \sum_{i<j} V_{ij}(t_k) + ... + \sum_k V_{12...r}(t_k)$$

- **Sensitivity indices:**

$$S_i = \frac{\sum_k V_i(t_k)}{\sum_k V_Y(t_k)} \tag{1}$$

for the main effects, and

$$S_{ij} = \frac{\sum_k V_{ij}(t_k)}{\sum_k V_Y(t_k)} \tag{2}$$

for second order interactions (similar definitions for higher order interactions).

- **Total sensitivity index of i^{th} factor**

$$TS_i = S_i + S_{i1} + ... + S_{ir} + ... + S_{12...r}.$$

The closer to zero the value of TS_i, the less influential θ_i is

Screening calibration parameters

- Parameters associated with inactive (or unimportant) inputs may be more difficult to constrain by confidence/posterior intervals than the active ones.
- *Reason:* Output "almost" constant in an inactive input parameter has a "very small" partial derivative at any of its values.
- In turn, the Fisher information matrix is "nearly" singular, so its inverse (appearing in the asymptotic covariance of MLEs) is numerically unstable.
- One can better constrain an "almost" inactive input by using more strongly informative priors on that parameter.
- MIT 2D climate model example:
  $TS_S = 0.5409$, $TS_K = 0.2195$, $TS_F = 0.3596$
  so K_v appears to be the least sensitive, it is expected to be harder to constrain (unless an informative prior is placed on K_v).

Approach 1: Interpolating the posterior pdf

- *Advantages:*
  - Easier to set up the scalar emulator than the multidimensional emulator
  - Fewer data to emulate (scalar posterior pdf values at sampled inputs), requiring less computing time.
- *Disadvantages:*
  - It is specifically tailored for calibration; e.g. not designed for other analyses, such as sensitivity analysis.
  - Posterior pdf emulator may create several false probability modes, making the inference more difficult.

Approach 2: Multidimensional emulator

- *Advantages:*
  - Can be used for a more comprehensive analysis, which includes surrogate-based input space exploration, surrogate-based sensitivity analysis (in addition to just calibration).
  - Provides a better understanding of e.g. space-time correlations of model output.

- *Disadvantages:*
  - Requires more complex modeling.
  - For large and/or nonstationary data sets it may be difficult to implement the direct, Kronecker-based surrogate method. As an alternative, basis decomposition methods (e.g. wavelets, principal components) for such data sets have allowed for dimension reduction and/or simplified modeling.

Conclusions

- This talk has presented a comparison of two major computer model calibration methods: interpolating the posterior pdf and using a multidimensional emulator.

- The simplified MIT 2D climate model has been used to illustrate this comparison.

- Sensitivity analysis can be helpful in identifying parameters "difficult" to calibrate *before* the actual calibration is performed.