Tunable Testbed for Detection and Attribution

IDAG Workshop 2018

Nathan Lenssen

Columbia University, Department of Earth and Environmental Sciences
Lamont-Doherty Earth Observatory

March 14, 2018

Lamont-Doherty Earth Observatory
Columbia University | Earth Institute

Statistical Formulation of Detection and Attribution

Ordinary Least Squares (OLS)

Observed Quantities:

- $\boldsymbol{y}$: The observed climate response of interest
- $\boldsymbol{X}^*$ The model-simulated forcing responses $\boldsymbol{X}^* = (\boldsymbol{x}_1^*, \ldots, \boldsymbol{x}_m^*, \ldots, \boldsymbol{x}_M^*)$

$$\boldsymbol{y} = \boldsymbol{X}^*\beta + \boldsymbol{u}$$

Statistical Parameter of Interest:

- β Estimation provides detection, inference (CIs) gives us attribution

Climate Variability:

- $\boldsymbol{u}$ The error due to climate variability where $\boldsymbol{u} \sim \mathcal{N}(0, \mathbf{C})$
- $\mathbf{C}$ Estimated through model control runs

Statistical Formulation of Detection and Attribution

Error-in-Variable (EIV)

Observed Quantities:

- $\boldsymbol{y}$ The climate response of interest
- $\boldsymbol{X}$ The noisy responses to forcings $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m, \ldots, \boldsymbol{x}_M)$

$$\boldsymbol{y} = \boldsymbol{X}^*\beta + \boldsymbol{u}_y$$
$$\boldsymbol{X} = \boldsymbol{X}^* + \boldsymbol{U}$$

Latent Quantities:

- $\boldsymbol{y}^*$ The 'true' climate response where $\boldsymbol{y}^* = \boldsymbol{X}^*\beta$
- $\boldsymbol{X}^*$ The 'true' responses to forcings $\boldsymbol{X}^* = (\boldsymbol{x}_1^*, \ldots, \boldsymbol{x}_M^*)$

Climate Variability:

- $\boldsymbol{u}_y$ As OLS formulation with $\boldsymbol{u}_y \sim \mathcal{N}(0, \mathbf{C})$
- $\boldsymbol{U}$ The error on the forcing responses due to climate variability

$$\boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_M) \overset{\text{iid}}{\sim} \mathcal{N}(0, \mathbf{C})$$

Statistical Formulation of Detection and Attribution

Error-in-Variable (EIV): Multimember Ensembles

For a given forcing m, we run ensemble of size L_m

$$\boldsymbol{x}_m = \left(\boldsymbol{x}_m^{(1)}, \ldots \boldsymbol{x}_m^{(\ell)}, \ldots, \boldsymbol{x}_m^{(L_m)}\right), \qquad \boldsymbol{x}_m^{(\ell)} \overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{x}_m^*, \mathbf{C})$$

The ensemble mean of the m^{th} forced response is

$$\overline{\boldsymbol{x}_m} = \frac{1}{L_m} \sum_{\ell=1}^{L_m} \boldsymbol{x}_m^{(\ell)}, \qquad \overline{\boldsymbol{x}_m} \sim \mathcal{N}\left(\boldsymbol{x}_m^*, L_m^{-1}\,\mathbf{C}\right)$$

Rewriting in the error-in-variable formulation

$$\overline{\boldsymbol{x}_m} = \boldsymbol{x}_m^* + L_m^{-1/2} \boldsymbol{u}_m$$

Or for all forcing responses with $\boldsymbol{L} = \text{diag}(L_1, \ldots, L_M)$

$$\overline{\boldsymbol{X}} = \boldsymbol{X}^* + \boldsymbol{L}^{-1/2} \boldsymbol{U}$$

Statistical Formulation of Detection and Attribution

Error-in-Variable (EIV): Multimember Ensembles

$$\overline{\boldsymbol{X}} = \boldsymbol{X}^* + \boldsymbol{L}^{-1/2}\boldsymbol{U}$$

Plugging into our full error in variable formulation:

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X}^*\beta + \boldsymbol{u}_y & \qquad \boldsymbol{u}_y &\sim \mathcal{N}(0, \mathbf{C}) \\ \overline{\boldsymbol{X}} &= \boldsymbol{X}^* + \boldsymbol{L}^{-1/2}\boldsymbol{U} & \qquad \boldsymbol{u}_m &\overset{\text{iid}}{\sim} \mathcal{N}(0, \mathbf{C}) \end{aligned}$$

- $\overline{\boldsymbol{X}}$ The ensemble means $\overline{\boldsymbol{X}} = (\overline{\boldsymbol{x}_1}, \ldots, \overline{\boldsymbol{x}_M})$
- $\boldsymbol{L}$ The ensemble sizes $\boldsymbol{L} = (L_1, \ldots, L_M)$
- $\boldsymbol{U}$ The forcing variability matrix $\boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_M)$

Statistical Formulation of Detection and Attribution

Observed Response Variability

The incomplete expression for the climate response $\boldsymbol{y}$ is

$$\boldsymbol{y} = \boldsymbol{X}^* \beta + \boldsymbol{u}_y \,, \qquad \boldsymbol{u}_y \sim \mathcal{N}(0, \mathbf{C})$$

We propose three different $\boldsymbol{y}$ states to fully incorporate all of the sources of variability.

- $\boldsymbol{y}^*$ The true climate response (latent) $\quad \boldsymbol{y}^* = \boldsymbol{X}^* \beta$
- $\boldsymbol{y}_{rel}$ The realized climate response (latent) $\quad \boldsymbol{y}_{rel} = \boldsymbol{y}^* + \boldsymbol{u}_Y$
- $\boldsymbol{y}_{obs}$ The observed climate response $\quad \boldsymbol{y}_{obs} = \boldsymbol{y}_{rel} + \varepsilon_Y$
  - With observational error $\varepsilon_Y \sim \mathcal{N}(0, \mathbf{W})$

$$\boldsymbol{y}_{rel} = \boldsymbol{X}^* \beta + \boldsymbol{u}_y \qquad \boldsymbol{u}_y \sim \mathcal{N}(0, \mathbf{C})$$
$$\boldsymbol{y}_{obs} = \boldsymbol{y}_{rel} + \varepsilon_y \qquad \varepsilon_y \sim \mathcal{N}(0, \mathbf{W})$$

Statistical Formulation of Detection and Attribution

Observed Response Variability

$$\boldsymbol{y}_{rel} = \boldsymbol{X}^*\beta + \boldsymbol{u}_y \qquad \boldsymbol{u}_y \sim \mathcal{N}(0, \mathbf{C})$$
$$\boldsymbol{y}_{obs} = \boldsymbol{y}_{rel} + \varepsilon_y \qquad \varepsilon_y \sim \mathcal{N}(0, \mathbf{W})$$

Total Observation Variability: Since the observational and climate variability errors are independent, condense in terms of $\boldsymbol{\nu} = \boldsymbol{u}_y + \varepsilon_y$, the total climate variability

$$\begin{aligned} \boldsymbol{y}_{obs} &= \boldsymbol{y}_{rel} + \varepsilon_y \\ &= \boldsymbol{X}^*\beta + \boldsymbol{\nu}\,, \qquad \boldsymbol{\nu} \sim \mathcal{N}(0, \mathbf{C} + \mathbf{W}) \end{aligned}$$

Note: Flexibility to add additional variability terms to $\boldsymbol{\nu}$

- Linear approximation error (from statistical model)
- Climate model error

Statistical Formulation of Detection and Attribution

Observed Response Variability

$$\boldsymbol{y}_{rel} = \boldsymbol{X}^* \beta + \boldsymbol{u}_y \qquad \boldsymbol{u}_y \sim \mathcal{N}(0, \mathbf{C})$$
$$\boldsymbol{y}_{obs} = \boldsymbol{y}_{rel} + \varepsilon_y \qquad \varepsilon_y \sim \mathcal{N}(0, \mathbf{W})$$

Observational Ensembles: Following the notation of the multimember ensembles with L_y as the size of the observational ensemble

$$\begin{aligned} \boldsymbol{Y}_{obs} &= (\boldsymbol{y}_{obs}^{(1)}, \cdots, \boldsymbol{y}_{obs}^{(L_y)}) \\ &= \boldsymbol{Y}_{rel} + (\varepsilon_y^{(1)}, \cdots, \varepsilon_y^{(L_y)}), \qquad \varepsilon_y^{(\ell)} \overset{\text{iid}}{\sim} \mathcal{N}(0, \mathbf{W}) \end{aligned}$$

Information about **W** may be from gained from multiple observations, but information about **C** does not increase!

Statistical Formulation of Detection and Attribution

Full Model

Full Error-in-Variable Model:

$$
\begin{aligned}
\boldsymbol{y}_{obs} &= \boldsymbol{y}_{rel} + \varepsilon_y & \varepsilon_y &\sim \mathcal{N}(0, \mathbf{W}) \\
\boldsymbol{y}_{rel} &= \boldsymbol{X}^*\beta + \boldsymbol{u}_y & \boldsymbol{u}_y &\sim \mathcal{N}(0, \mathbf{C}) \\
\overline{\boldsymbol{X}} &= \boldsymbol{X}^* + \boldsymbol{L}^{-1/2}\boldsymbol{U} & \boldsymbol{u}_m &\overset{\text{iid}}{\sim} \mathcal{N}(0, \mathbf{C})
\end{aligned}
$$

Scale-Variant Error-in-Variable Model:

$$
\begin{aligned}
\boldsymbol{y}_{obs} &= \boldsymbol{y}_{rel} + \varepsilon_y & \varepsilon_y &\sim \mathcal{N}(0, \mathbf{W}) \\
\boldsymbol{y}_{rel} &= \boldsymbol{X}^*\beta + \boldsymbol{u}_y & \boldsymbol{u}_y &\sim \mathcal{N}(0, \alpha^{-1}\,\mathbf{C}) \\
\overline{\boldsymbol{X}} &= \boldsymbol{X}^* + \boldsymbol{L}^{-1/2}\boldsymbol{U} & \boldsymbol{u}_m &\overset{\text{ind}}{\sim} \mathcal{N}(0, \gamma_m^{-1}\,\mathbf{C})
\end{aligned}
$$

Dorit will talk about fitting this model!

Testbed Motivation

Goal: Determine the contribution of forcings to the observed climate

$$\boldsymbol{y}_{obs} = \boldsymbol{X}\beta + \boldsymbol{u} \qquad u \sim \mathcal{N}(0, \mathbf{C})$$

Reality: The data and resulting relationships between the forced responses and observations are complicated

$$\begin{aligned} \boldsymbol{y}_{obs} &= \boldsymbol{y}_{rel} + \varepsilon_y & \varepsilon_y &\sim \mathcal{N}(0, \mathbf{W}) \\ \boldsymbol{y}_{rel} &= \boldsymbol{X}^*\beta + \boldsymbol{u}_y & \boldsymbol{u}_y &\sim \mathcal{N}(0, \alpha^{-1}\mathbf{C}) \\ \overline{\boldsymbol{X}} &= \boldsymbol{X}^* + \boldsymbol{L}^{-1/2}\boldsymbol{U} & \boldsymbol{u}_m &\overset{\text{ind}}{\sim} \mathcal{N}(0, \gamma_m^{-1}\mathbf{C}) \end{aligned}$$

Crux: Fitting requires estimating $\hat{\mathbf{C}}$, a full-rank $n \times n$ matrix with the number of control runs $L_0 \ll n$

Testbed Motivation

A flexible and tunable testbed will allow researchers working on detection and attribution methods to:

- Evaluate methods by comparing estimated and true parameter values
  - Performance scaling as a function of sample size/dimensionality
- Simulate real-world scenarios to enable testbed results to represent applications
  - Tunable variety of climate response patterns and climate variability covariances
- Determine robustness of methods through perturbations of testbed parameters
- Compare multiple D+A methods on variety of scenarios

Data and Parameters of Interest

$$\begin{aligned} \boldsymbol{y}_{obs} &= \boldsymbol{y}_{rel} + \varepsilon_y & \varepsilon_y &\sim \mathcal{N}(0, \mathbf{W}) \\ \boldsymbol{y}_{rel} &= \boldsymbol{X}^* \beta + \boldsymbol{u}_y & \boldsymbol{u}_y &\sim \mathcal{N}(0, \alpha^{-1}\,\mathbf{C}) \\ \overline{\boldsymbol{X}} &= \boldsymbol{X}^* + \boldsymbol{L}^{-1/2} \boldsymbol{U} & \boldsymbol{u}_m &\overset{\text{ind}}{\sim} \mathcal{N}(0, \gamma_m^{-1}\,\mathbf{C}) \end{aligned}$$

Observed Objects:

$\boldsymbol{X}$ Observed forcing response ensembles

$$\begin{aligned} \boldsymbol{X} &= (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M) \\ &= ([\boldsymbol{x}_1^{(1)}, \ldots, \boldsymbol{x}_1^{(L_1)}], \ldots, [\boldsymbol{x}_M^{(1)}, \ldots, \boldsymbol{x}_M^{(L_M)}]) \end{aligned}$$

$\boldsymbol{Y}_{obs}$ Observed climate response ensemble

$$\boldsymbol{Y}_{obs} = (\boldsymbol{y}_{obs}^{(1)}, \ldots, \boldsymbol{y}_{obs}^{(L_y)})$$

$\boldsymbol{X}_0$ Control runs from the model used for $\boldsymbol{X}$

$$\boldsymbol{X}_0 = (\boldsymbol{X}_0^{(1)}, \ldots, \boldsymbol{X}_0^{(L_0)})$$

Data and Parameters of Interest

$$\begin{aligned} \boldsymbol{y}_{obs} &= \boldsymbol{y}_{rel} + \boldsymbol{\varepsilon}_y & \boldsymbol{\varepsilon}_y &\sim \mathcal{N}(0, \mathbf{W}) \\ \boldsymbol{y}_{rel} &= \boldsymbol{X}^* \beta + \boldsymbol{u}_y & \boldsymbol{u}_y &\sim \mathcal{N}(0, \alpha^{-1}\,\mathbf{C}) \\ \overline{\boldsymbol{X}} &= \boldsymbol{X}^* + \boldsymbol{L}^{-1/2} \boldsymbol{U} & \boldsymbol{u}_m &\overset{\text{ind}}{\sim} \mathcal{N}(0, \gamma_m^{-1}\,\mathbf{C}) \end{aligned}$$

Latent Objects

θ The statistical parameters in the model

$$\theta = (\beta, \alpha, \gamma, \mathbf{C}, \mathbf{W})$$

$\boldsymbol{X}^*$ True forcing response $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M)$

$\boldsymbol{y}^*$ True climate response $\boldsymbol{y}^* = \boldsymbol{X}\beta$

$\boldsymbol{y}_{rel}$ Realized climate response

Bulleted List

Subtitle

- Structure Color: text
- Alert Color: TEXT

Bulleted List with Sub-bullets

- In general, falls somewhere in between the Hadley and Berkeley methods
- Main Point here
  - Sub-point here
- Has rudimentary uncertainty model that resembles Hadley

Numbered List

1. Structure text: further stuff
2. regular point

Blocks

Text outside of blocks

Block

Regular Block is Here

Example Block

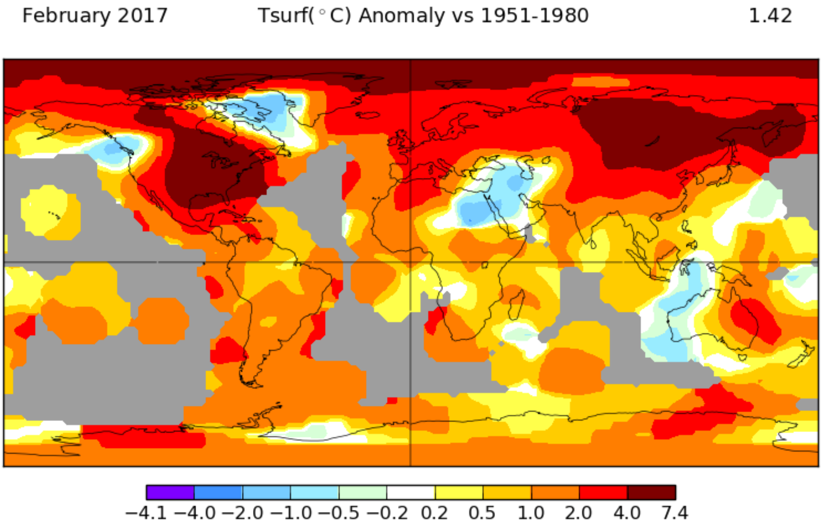
Example is here

Alert Block

Alert block is here

Full Slide Figure

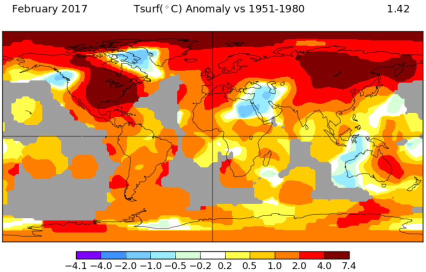
Subtitle Here



Source: Source Here

Two Column Slide (List and Figure)

- Point 1
- Point 2



Source: Source Here