Testbed for Detection and Attribution Methods

Nathan Lenssen

December 2017

0.1 Description of MATLAB Files in finalCode

- `fullProcedure.m` is a master function that runs the generation procedure, has the option to plot, and saves the relevant output. It takes no more than 30 seconds on my laptop at the current settings.

- `parameterList.m` User-set parameters for the Detection and Attribution testbed generation matlab program. Generally the only .m file that should be edited for basic use of the testbed.

- `generateSimulation.m` is a script that handles the entire simulation. It returns all of the objects of interest as laid out in Section 0.3, plus a few bonus objects used in plotting. It depends on a few user functions to handle the spatial covariances. It currently needs major attention with respect to the relative scales of $\boldsymbol{x}^*$, $\boldsymbol{y}^*$, $\boldsymbol{\nu}$, and $\boldsymbol{\eta}$.

- `makePlots.m` Creates all of the plots found in this report if plotting is turned on when `fullProcedure.m` is run. Will be expanded as the project progresses.

- `writeOutput.m` Saves the relevant objects to disk. Will eventually save as netCDF to facilitate sharing between programming languages.

- `generateMatern.m` is an evaluation of the Matérn radial basis function as done in `fields`. It is used to generate the spatial covariance of the forced responses.

- `generateRecBasis.m` returns the eigenvectors and eigenvalues of an exponential covariance matrix on a $[0, 1] \times [0, 1]$ grid for a given scale parameter.

- `generateCovExact.m` returns various covariance objects of the modified exponential to be used as the internal climate variability $\mathbf{C}$. The modification is an in Hannart 2016 in which the eigenvalues of a exponential covariance matrix are modified by multiplicative random factors.

- `runLocal.sh` runs the testbed on a local machine. Note that matlab must be correctly added to the PATH variable

- `runHabanero.sh` runs the testbed on the Habanero cluster. Useful for users that do not have local access to MATLAB

0.2 Model and parameters of interest

Following equation (3) [Hannart, 2016], we work with the classical problem for detection and attribution where $\boldsymbol{y}$ is the observed spatial climate response, $\boldsymbol{x}_j$ with $j = 1, \ldots, J$ are the externally forced signals, and $\boldsymbol{\Sigma}$ is the internal climate variability.

$$
\begin{aligned}
\boldsymbol{y} &= \boldsymbol{x}^*\boldsymbol{\beta} + \boldsymbol{\nu} & \boldsymbol{\nu}_j &\overset{iid}{\sim} \mathcal{N}_n(0, \boldsymbol{\Sigma}) \\
\boldsymbol{x} &= \boldsymbol{x}^* + \boldsymbol{\eta} & \boldsymbol{\eta}_j &\overset{iid}{\sim} \mathcal{N}_n(0, \boldsymbol{\Sigma}/m)
\end{aligned}
\tag{1}
$$

To make explicit that the fingerprints we are regressing in the EIV model above are ensemble means, we let $\bar{\boldsymbol{x}}$ be the matrix of ensemble means for the forced climate responses and rewrite. Letting $\mathbf{C}$ represent the internal climate variability we define

$$
\begin{aligned}
\boldsymbol{y} &= \boldsymbol{x}^*\boldsymbol{\beta} + \boldsymbol{\nu} & \boldsymbol{\nu}_j &\overset{iid}{\sim} \mathcal{N}_n(0, \mathbf{C}) \\
\bar{\boldsymbol{x}} &= \boldsymbol{x}^* + \boldsymbol{\eta} & \boldsymbol{\eta}_j &\overset{iid}{\sim} \mathcal{N}_n(0, \mathbf{C}/m).
\end{aligned}
\tag{2}
$$

Adding in the notation from [Smith et al., 20xx] where we include a superscript $\boldsymbol{x}_j^{(\ell)}$ with $\ell = 1, \ldots, L$ to represent ensemble members for a given forcing, we unpack to see the ensemble behavior of the climate variability in the forced responses

$$
\begin{aligned}
\boldsymbol{y} &= \boldsymbol{x}^*\boldsymbol{\beta} + \boldsymbol{\nu} & \boldsymbol{\nu}_j &\overset{iid}{\sim} \mathcal{N}_n(0, \mathbf{C}) \\
\bar{\boldsymbol{x}} &= \boldsymbol{x}^* + \frac{1}{m}\sum_{\ell=1}^{L} \boldsymbol{\nu}^{(\ell)} & \boldsymbol{\nu}_j &\overset{iid}{\sim} \mathcal{N}_n(0, \mathbf{C}) \\
&= \boldsymbol{x}^* + \boldsymbol{\eta} & \boldsymbol{\eta}_j &\overset{iid}{\sim} \mathcal{N}_n(0, \mathbf{C}/L)
\end{aligned}
\tag{3}
$$

Given the formulation presented in equation 2, the primary goals of inference the $\boldsymbol{\beta}$ and $\mathbf{C}$ parameters. In the development of a testbed, we need to simulate all of these objects (see Section 0.4 for full list) and organize in an intuitive and efficient way.

This notation is still under development and clearly a poor middle ground in places. Possibly discuss today the best way to merge the notation of Alexis, Aurélien, Philippe, et al. with the notation of Richard, Matthias, Dorit et al. for the final code and documentation.

0.3 Assumptions for Simulation

Using the framework proposed by Alexis in Section 4a of [Hannart, 2016] as a starting point, we create a preliminary testbed with the following assumptions:

(I) We assume that the climate variability in the observed climate signal $\boldsymbol{\nu}$ is proportional to the climate variability in the ensemble mean of the forced signal $\boldsymbol{\eta}$ through the ensemble size. This is different from The L^{-1} scaling in the covariance of $\boldsymbol{\eta}$ is assuming the fingerprints used are ensemble means. A clear and relatively simple extension is to incorporate the α and γ covariance scaling factors that are introduced in [Smith et al., 20xx]

(II) We assume that the internal climate variability $\mathbf{C}$ has a form as in [Hannart, 2016]. The generative process is modifying the eigenvalues of a exponential covariance matrix by multiplicative random factors. This procedure is a computationally efficient way to generate a covariance structure that is non-isotropic.

(III) We assume that the unobserved regressors $\boldsymbol{x}^*$, with columns $\boldsymbol{x}_j^*$ a single forced fingerprint, have spatial covariance according to the Matérn $\boldsymbol{\Sigma}_{x_j}$. This covariance structure is not representative of the noise on the simulated $\boldsymbol{x}_j^{(\ell)}$ which is represented in Equation 2, but rather the underlying structure of the true forced climate responses $\boldsymbol{x}_j^*$. Furthermore, we can specify a different spatial correlation for each simulated forced response. We also assume that the columns of $\boldsymbol{x}^*$ are independent.

Note that this is different than the forced responses in [Hannart, 2016] where they were taken to be spatially independent, but correlated across forcings. In conversation with Alexis, we decided to proceed with generating our forced responses with spatial correlation and forcing independence for now.

(IV) Further work is needed thinking about to control the relative scales of the signal from the forcing responses and the noise from the internal climate variability. This could be handled in the simulation by the addition of the covariance scaling parameters suggested by Richard.

0.4 Applicability to Smith et al. 20xx and KHS 2017

The model for $\boldsymbol{y}$ in [Katzfuss et al., 2017] is given in equation (8) of Smith et al. 20xx as

$$\boldsymbol{y}|\boldsymbol{x}_1,\dots\boldsymbol{x}_J,\boldsymbol{\beta},\mathbf{C} \quad \sim \quad \mathcal{N}_n\left[\sum_{j=1}^J \boldsymbol{x}_j\beta_j, \mathbf{C}\right] \tag{4}$$

The extension proposed in Smith et al. 20xx is adding a scaling factor α to control the relationship between the noise in the control run and forced run simulations.

$$\boldsymbol{y}|\boldsymbol{x}_1,\dots\boldsymbol{x}_J,\boldsymbol{\beta},\mathbf{C},\alpha \quad \sim \quad \mathcal{N}_n\left[\sum_{j=1}^J \boldsymbol{x}_j\beta_j, \alpha^{-1}\mathbf{C}\right] \tag{5}$$

The model error for a single forcing j for a single realization ℓ allowing for forcing specific scaling factors γ_j is

$$\boldsymbol{x}_j^{(\ell)}|\boldsymbol{x}_j,\mathbf{C},\gamma \quad \sim \quad \mathcal{N}_n\left[\boldsymbol{x}_j, \gamma_j^{-1}\mathbf{C}\right], \ \ell = 1,\dots,L, \ j = 1,\dots,J. \tag{6}$$

The Bayesian formulations are particularly useful in the context of a testbed simulation as they directly prescribe a procedure for generating data.

The major assumptions made in Smith et al. 20xx are

(I) We no longer require as strict of an assumption as (I) in the Hannart formulation. We allow scaling between the various models through α and between the various forcing scenarios through γ.

(II) The dimension reduction of the climate variability $\mathbf{C}$ is handled by informative priors

(III) Additional flexibility in handling ensembles of $\boldsymbol{y}$ by allowing another variance term $\boldsymbol{W}$ to model the behavior of the observational noise.

0.5 Objects of Interest

Relevant sizes:

- n the number of observations
- r the number of control runs used in the estimation of $\mathbf{C}$ $(r \ll n)$
- k the number of fingerprints/PCs/eigenvectors to be used in the analysis $(k \leq r)$
- L the size of the ensemble for each forcing (will eventually be L_j)
- J the number of externally forced signals (regressors)

Objects that simulate the real problem we are facing:

- $\boldsymbol{y}$ the observed climate change field $[n \times 1]$
- $\bar{\boldsymbol{x}}$ the model-mean fingerprints from the p forcing response runs $[n \times J]$
- $\boldsymbol{x}$ the m ensemble member fingerprints from the p forcing response runs $[n \times (L \cdot J)]$
- $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_r)$, the control runs used to estimate $\mathbf{C}$ $[n \times r]$

Underlying objects from the simulation that we return to evaluate our performance:

- $\boldsymbol{y}^*$ the 'true' climate change pattern $[n \times 1]$
- $\boldsymbol{x}^*$ The 'true' fingerprints for each forcing $[n \times J]$
- $\boldsymbol{\Sigma}_x$ the spatial covariance used to generate the pattern of the $\boldsymbol{x}^*$. Note that this covariance is not necessarily equal for all patterns. $[n \times n \times J]$
- $\boldsymbol{\beta}$ the 'true' coefficients $[J \times 1]$
- $\mathbf{C}$ the 'true' internal variability $[n \times n]$

Objects that we are evaluating our inference of:

- $\hat{\boldsymbol{\beta}}$ the coefficients of attribution $[J \times 1]$
- $\hat{\mathbf{C}}$ the internal climate variability $[n \times n]$

0.6 Initial Simulation Outline

In this initial simulation we set the number of observations $n = 2500 = 50^2$ on the unit square, the number of forced signals $J = 2$ with $\boldsymbol{\beta} = (\beta_1, \beta_2) = (1, 1)$. The rest of the simulation procedure is as follows:

1) Generate the internal variability $\mathbf{C}$ with the modified eigenvalues method

   - We specify the scale of the exponential kernel and the magnitude of the random factors used to modify the eigenvalues

   - Additionally, the climate variability $\mathbf{C}$ is whitened according to the parameter ρ

   - Various visualizations of the resulting $\mathbf{C}$ can be seen in Figures 1 and 2.

2) Calculate $\boldsymbol{\Sigma}_{x_j}$ from the Matérn family for each of the forced responses

   - To demonstrate the capabilities of the testbed, we set one forcing very smooth and long range and the other forcing rough and short range.

3) Generate the true forcing responses $\boldsymbol{x}_j^*$ using $\boldsymbol{\Sigma}_{x_j}$ for the forcing-specific correlation.

   - The fields for $\boldsymbol{x}_1^*$ and $\boldsymbol{x}_2^*$ can be seen in the top left panel of Figures 3 and 4.

4) Simulate the observed forcing responses $\boldsymbol{x}_j^{(\ell)} = \boldsymbol{x}_j^* + \boldsymbol{\nu}$ for each forcing j and for each ensemble member ℓ.

   - We calculate the ensemble means of the forcing responses $\bar{\boldsymbol{x}}_j = \frac{1}{L} \sum_{\ell=1}^{L} \boldsymbol{x}_j^{(\ell)}$

   - The ensemble means are plotted in the top right of Figures 3 and 4.

   - Ensemble members are plotted in the bottom four plots of Figures 3 and 4.

5) Simulate the control runs from $\mathcal{N}(0, \mathbf{C})$

   - We currently simulate 300 members so that the sensitivity of methods to the size of the control run sample can be easily tested without rerunning the testbed simulation.

6) Calculate the true climate response according to $\boldsymbol{y}^* = \boldsymbol{x}^* \boldsymbol{\beta}$

   - Shown in the top left of Figure 5

7) Simulate a climate response $\boldsymbol{y} = \boldsymbol{x}^* \boldsymbol{\beta} + \boldsymbol{\nu}$

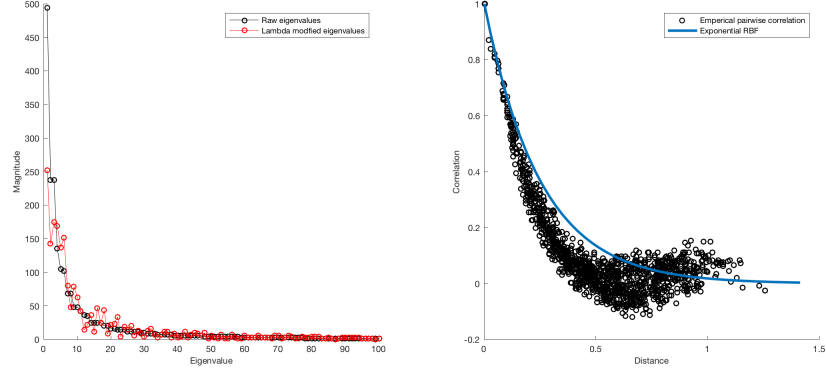   - Shown in the top right of Figure 5

Figure 1: **Left:** Visualization of the modification to the eigenvalues of the exponential covariance matrix for the construction of the climate variability **C Right:** Comparing the observed pairwise correlations in **C** with the exponential (without modified eigenvalues) as in Figure 3(c) of Hannart 2016.

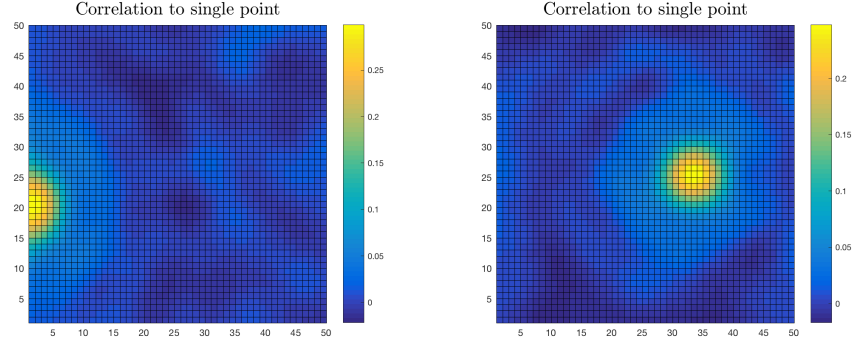Figure 2: Two single points correlation matrices according to the climate internal varability **C**. We are viewing the same covariance matrix from two different points as in Figure 3(d) of Hannart 2016.

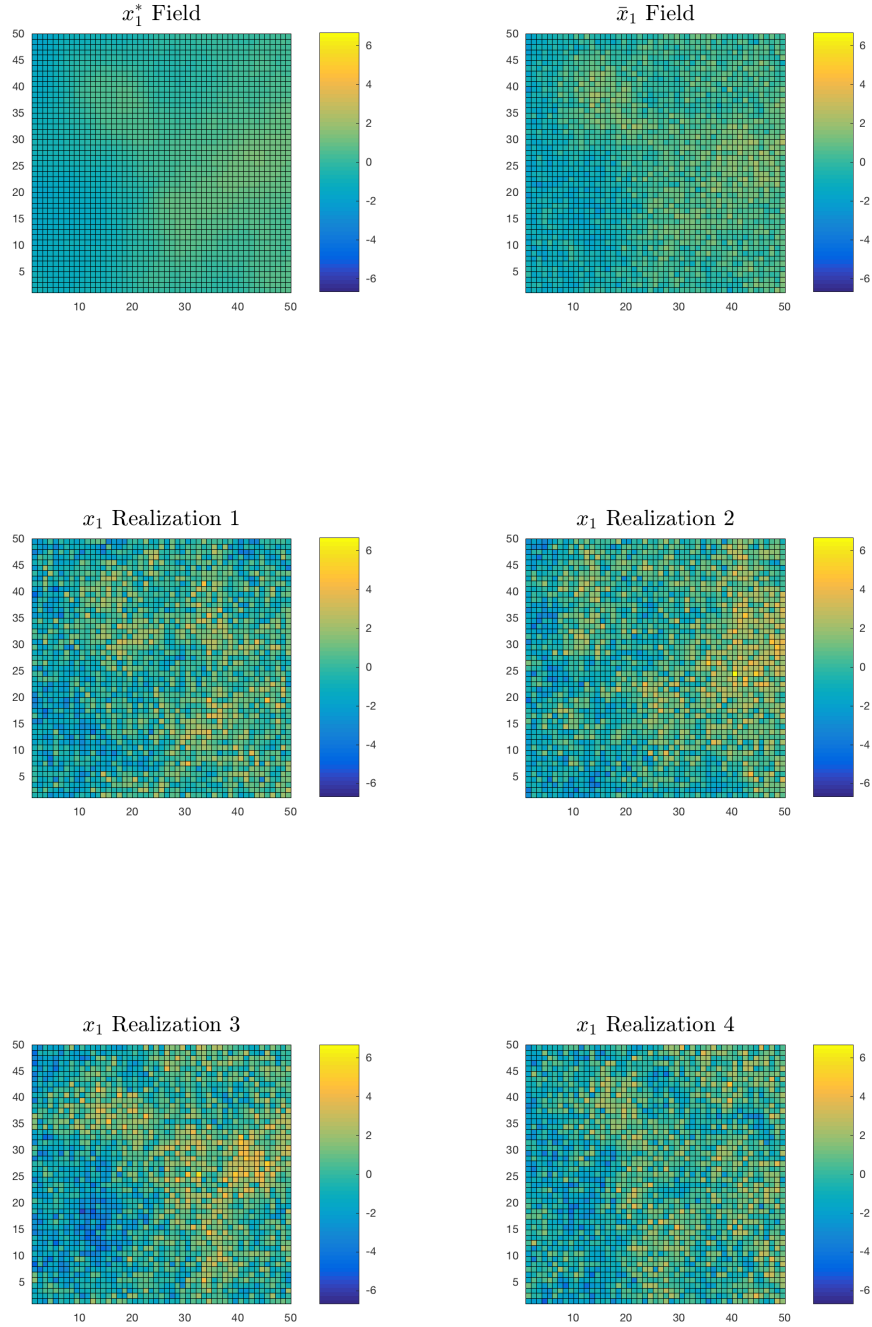Figure 3: The relationship between the underlying true $\boldsymbol{x}_1^*$ and the observed ensemble mean fingerprint $\bar{\boldsymbol{x}}_1$. Individual ensemble members are shown in the bottom four plots.

7

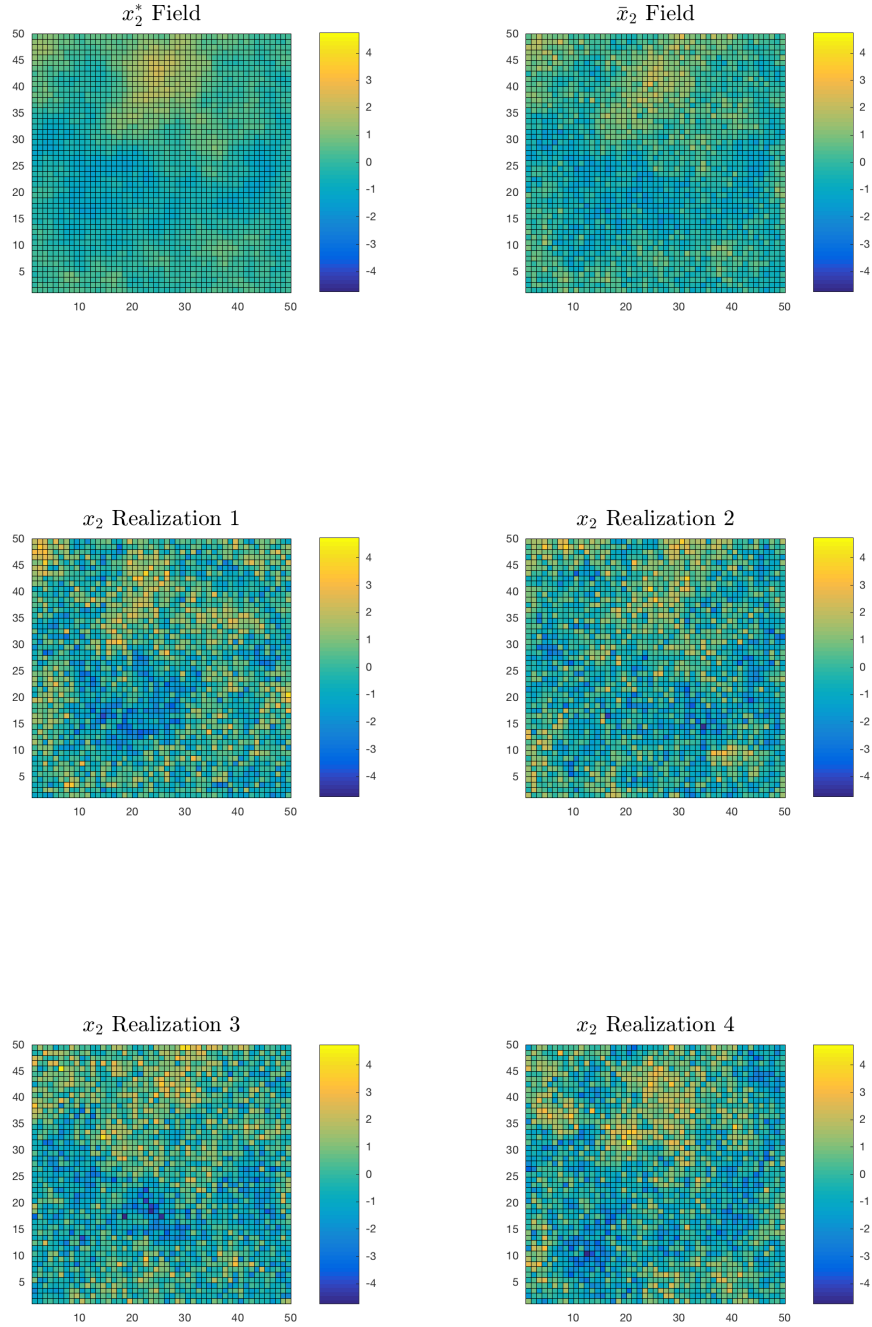Figure 4: The relationship between the underlying true $\boldsymbol{x}_2^*$ and the observed ensemble mean fingerprint $\bar{\boldsymbol{x}}_2$. Individual ensemble members are shown in the bottom four plots.
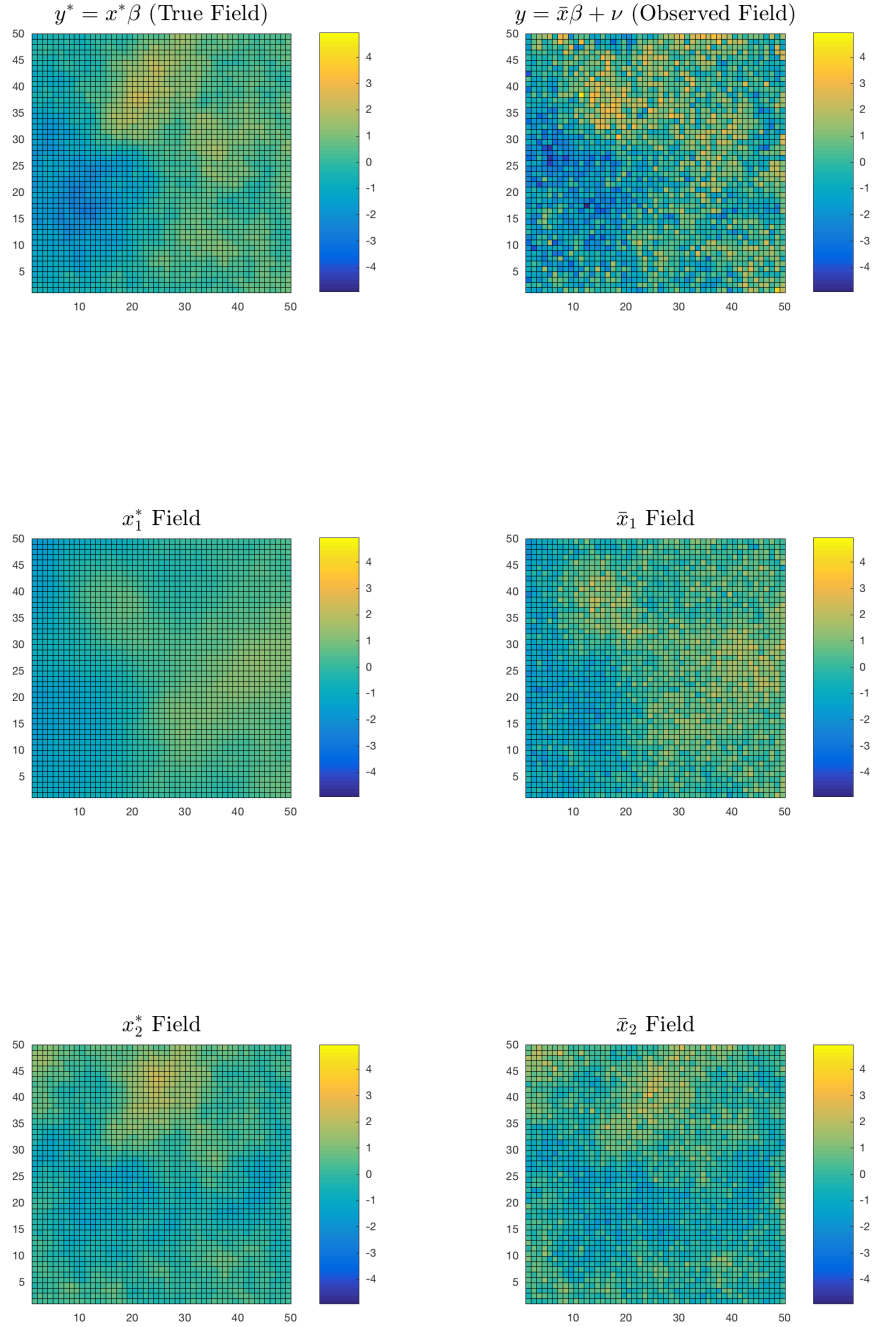
Figure 5: The progression of the $\boldsymbol{y}$ object as it is built. The true latent objects are in the left column and the observed objects are on the right. Since $\boldsymbol{\beta} = (1, 1)$, the top row is the sum of the bottom two rows.

References

[Hannart, 2016] Hannart, A. (2016). Integrated optimal fingerprinting: Method description and illustration. *Journal of Climate*, 29(6):1977–1998.

[Katzfuss et al., 2017] Katzfuss, M., Hammerling, D., and Smith, R. L. (2017). A bayesian hierarchical model for climate change detection and attribution. *Geophysical Research Letters*, 44(11):5720–5728. 2017GL073688.