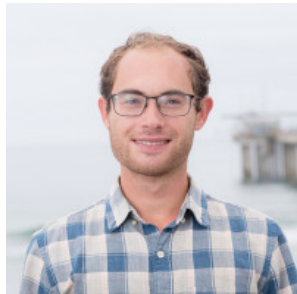


# *High-dimensional covariance estimation*

Matthias Morzfeld

*Cecil H. and Ida M. Green Institute of Geophysics and Planetary Physics  
Scripps Institution of Oceanography  
UC San Diego*

*Powered by*



David Vishny  
(SIO)



Kyle Gwartz  
(NASA)



Oliver Dunbar  
(Caltech)



Eviatar Bach  
(Reading)



Dan Hodyss  
(NRL)

# Problem formulation

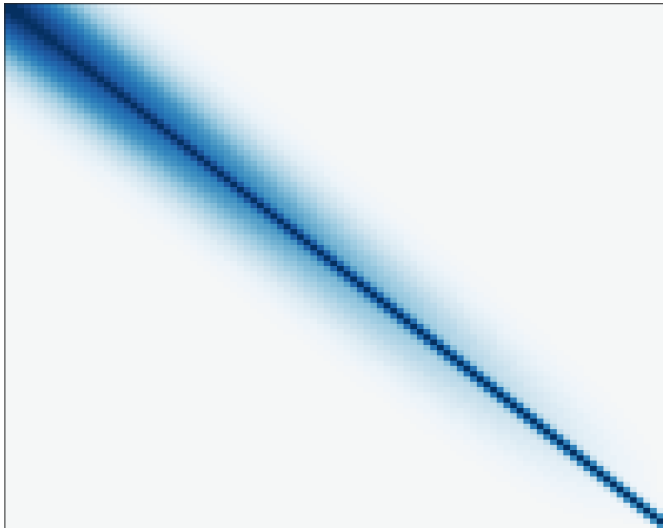
---

$$\hat{\mathbf{P}} = \frac{1}{n_e - 1} \sum_{i=1}^{n_e} (\mathbf{x}_i - \bar{\mathbf{x}}) \otimes (\mathbf{x}_i - \bar{\mathbf{x}}), \quad \bar{\mathbf{x}} = \frac{1}{n_e} \sum_{i=1}^{n_e} \mathbf{x}_i$$

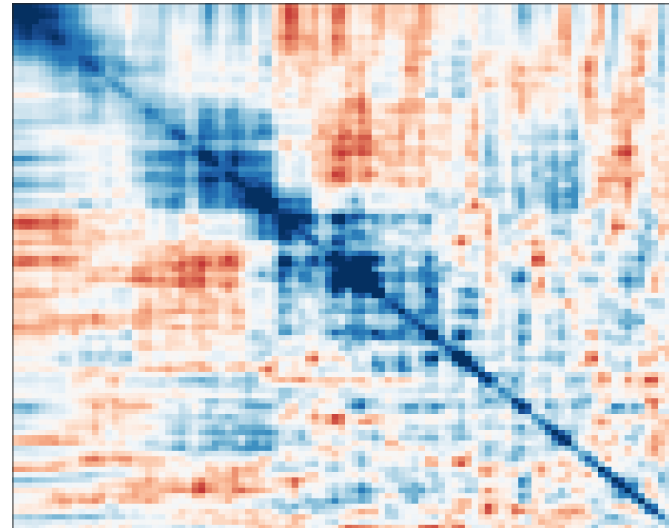
## Goal

- Given  $n_e$  vectors  $\mathbf{x}_i$  of dimension  $n_x$ , compute the covariance matrix of the random variable  $\mathbf{x}$
- **Caveat:** Few samples, huge dimension,  $n_e \ll n_x$

True covariance matrix



Estimate:  $n_x = 100$ ,  $n_e = 20$



# Why do we care about covariance estimation?

## ***Ensemble data assimilation\inversion***

- Global NWP
- Physical oceanography
- Climate modeling

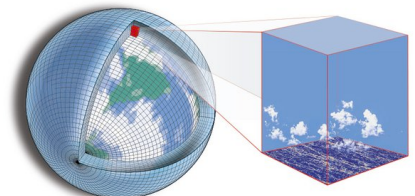
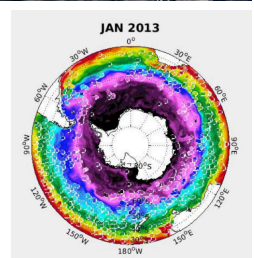
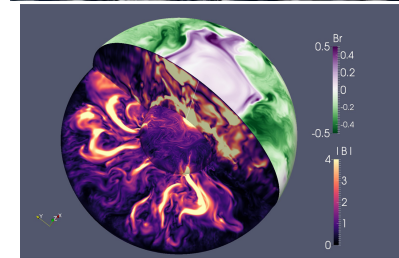
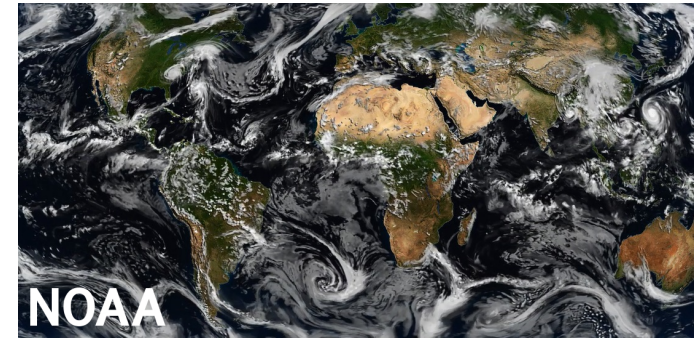
## ***Why is $n_e \ll n_x$ ?***

- Global NWP:  $O(10^8)$  unknowns,  $O(10^6)$  obs.
- PO:  $O(10^8)$  unknowns,  $O(10^7)$  obs.
- Geomagnetic DA:  $O(10^7)$  unknowns,  $O(10^3)$  ops.
- Climate: Train NNs on statistical data (CliMA)

## ***Beyond Earth science***

- Statistics, AI and ML
- “The covariance matrix is, arguably, the second most important object in all of statistics”

Ledoit & Wolf, *The power of (non-) linear shrinking: A review and guide to covariance matrix estimation*, Journal of Financial Econometrics, 20 (2022), pp. 187–218.

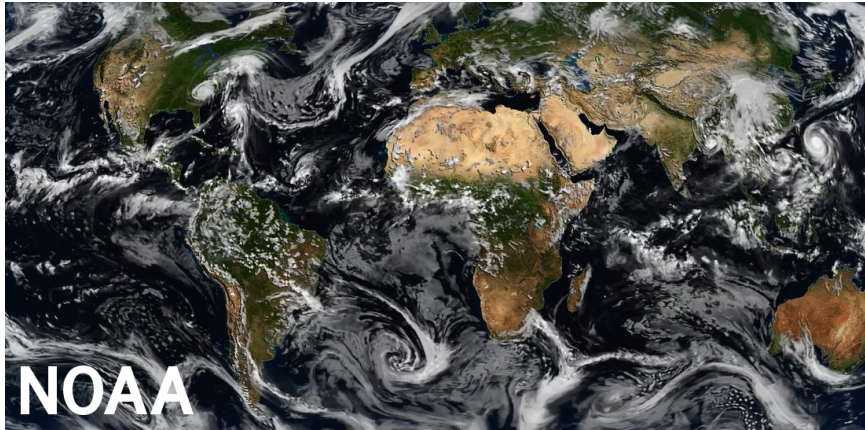


Why is  $n_e \ll n_x$ ?

---

## NWP

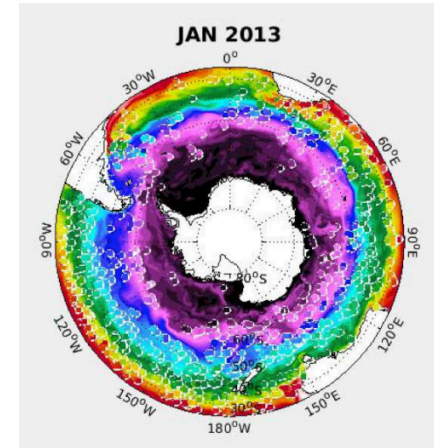
- $O(10^8)$  unknowns
- $O(10^6)$  observations
- Forecast every six hours



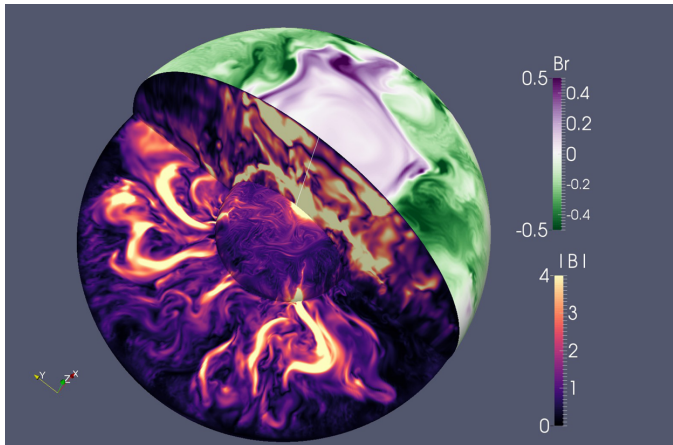
NOAA

## Physical oceanography

- $O(10^8)$  unknowns
- $O(10^7)$  observations
- Assimilate years worth of data



Southern Ocean State Estimate



## Geomagnetic DA

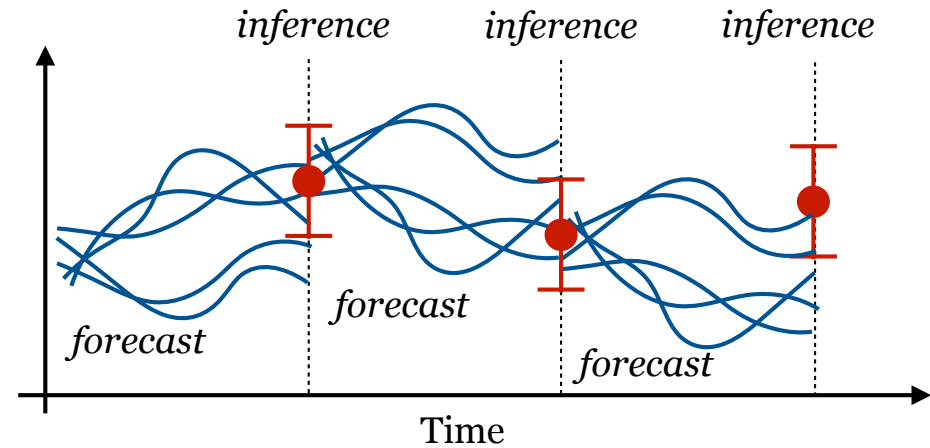
- $O(10^7)$  unknowns
- $O(10^3)$  observations
- Assimilate spherical harmonics



# Ensemble Kalman updates: I/II

## ***Ensemble Kalman filter***

- *Goal:* Update the state of a system with data



- *Forecast ensemble:*  $\mathbf{x}_i, \quad i = 1, \dots, n_e$
- *Observation:*  $\mathbf{y} = \mathbf{H}\mathbf{x} + \boldsymbol{\eta}, \quad \boldsymbol{\eta} \sim \mathcal{N}(0, \mathbf{R})$
- *Update:*  $\mathbf{x}_i^a = \mathbf{x}_i + \mathbf{K} (\mathbf{y} - (\mathbf{H}\mathbf{x}_i + \boldsymbol{\eta}_i))$
- *Kalman gain:*  $\mathbf{K} = \mathbf{P}\mathbf{H}^T (\mathbf{H}\mathbf{P}\mathbf{H}^T + \mathbf{R})^{-1}$

$$\hat{\mathbf{P}} = \frac{1}{n_e - 1} \sum_{i=1}^{n_e} (\mathbf{x}_i - \bar{\mathbf{x}}) \otimes (\mathbf{x}_i - \bar{\mathbf{x}}), \quad \bar{\mathbf{x}} = \frac{1}{n_e} \sum_{i=1}^{n_e} \mathbf{x}_i$$

## ***Ensemble Kalman inversion/I-EnKS***

- *Goal*: Fit a model to data by solving an optimization problem, without derivatives

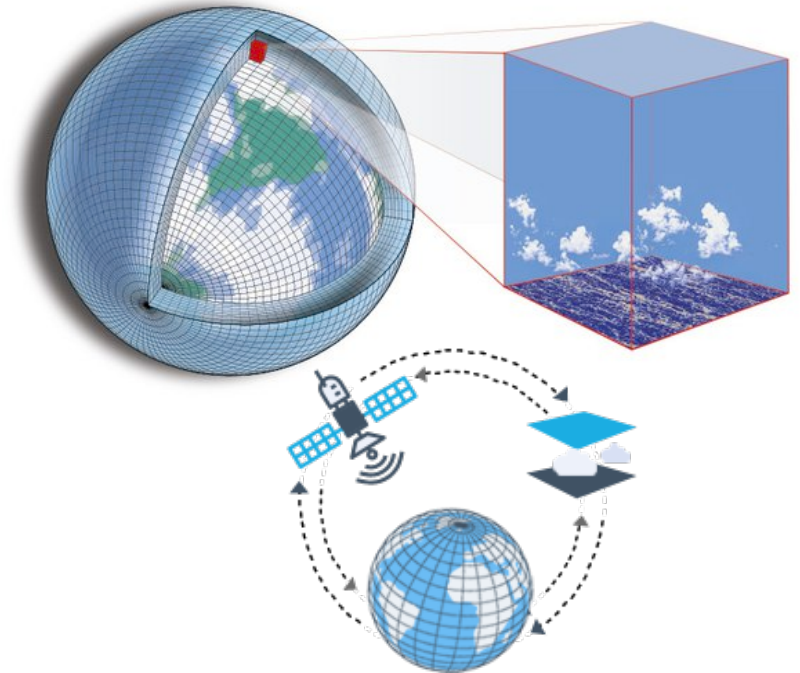
$$\arg \min_{\mathbf{x}} \left\| \underset{\substack{\uparrow \\ \text{Error}}}{\mathbf{R}^{-\frac{1}{2}}} \left( \underset{\substack{\uparrow \\ \text{Data}}}{\mathbf{y}} - \underset{\substack{\uparrow \\ \text{Model}}}{\mathcal{G}(\mathbf{x})} \right) \right\|_2^2$$

- *Why no derivatives?*

*Data*: Statistics of our climate

*Model*: Climate model, run for centuries, algorithms to extract statistics

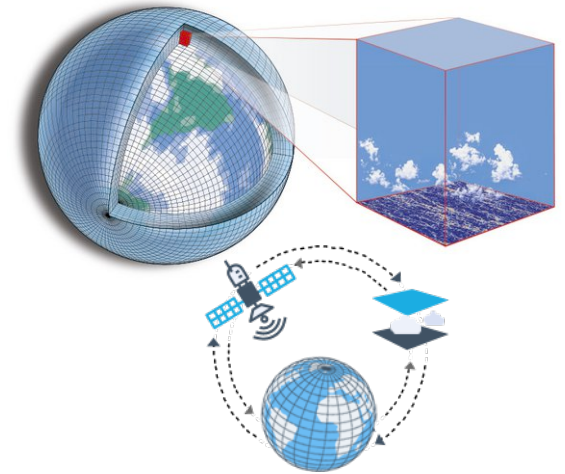
*Unknown*: Cloud microphysics parameterization (Neural Net)



## ***Ensemble Kalman inversion/I-EnKS***

- *Goal*: Fit a model to data by solving an optimization problem, without derivatives

$$\arg \min_{\mathbf{x}} \left\| \underset{\substack{\uparrow \\ \text{Error}}}{\mathbf{R}^{-\frac{1}{2}}} \left( \underset{\substack{\uparrow \\ \text{Data}}}{\mathbf{y}} - \underset{\substack{\uparrow \\ \text{Model}}}{\mathcal{G}(\mathbf{x})} \right) \right\|_2^2$$



### • **EKI iteration**

$$\mathbf{x}_i^{n+1} = \mathbf{x}_i^{n+1} + \underset{\substack{\uparrow \\ \text{Unknown-model} \\ \text{covariance}}}{\mathbf{C}_{\text{xg}}} \underset{\substack{\uparrow \\ \text{Data-data} \\ \text{covariance}}}{\mathbf{C}_{\text{gg}}} (\mathbf{C}_{\text{gg}} + \mathbf{R})^{-1} (\mathbf{y} - (\mathcal{G}(\mathbf{x}_i) + \boldsymbol{\eta}_i))$$

- Convergence guarantees (*with assumptions*)\*
- EKI = *Re-invention* what people have used for a while\*\* without guarantees

Problem formulation

**Partial solution in NWP**

*Inspiration*

NICE

Examples

Summary



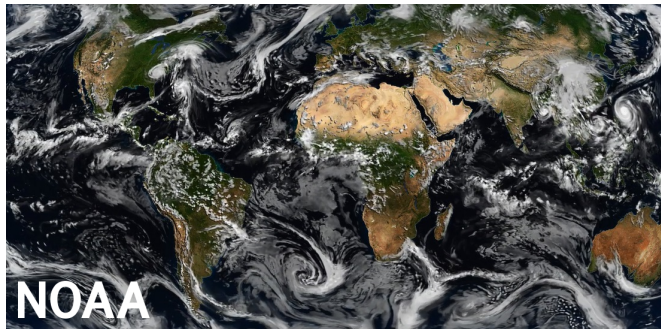
# Is localization magical or rooted in deep mathematics?

---

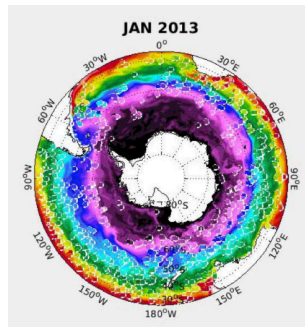
## *Ensemble Kalman filtering*

- $O(10^8)$  unknowns
- $O(10^6)$  observations
- Ensemble size  $\sim 200$

**How can they do this?**



NOAA



SOSE

## How I think the story went (*partially verified*)

---

### ***Ensemble Kalman filtering in NWP***

- They implemented EnKF
- ***It did not work***
- They wondered why
- They looked at some correlations
- They noticed that things are off — weather in La Jolla, California is correlated to weather in Kobe, Japan
- They deleted the correlations that they knew should not be there
- ***EnKF started to work***

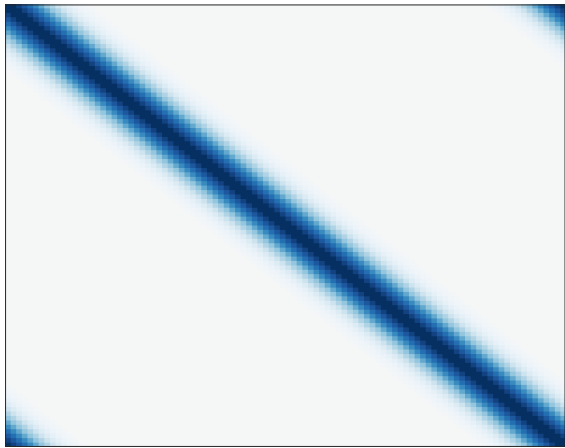
### ***Covariance localization***

- Small ensemble size implies that covariance estimates are noisy
- Noise presents itself as a “spurious correlation”
- Remove spurious correlations

## How it's done.

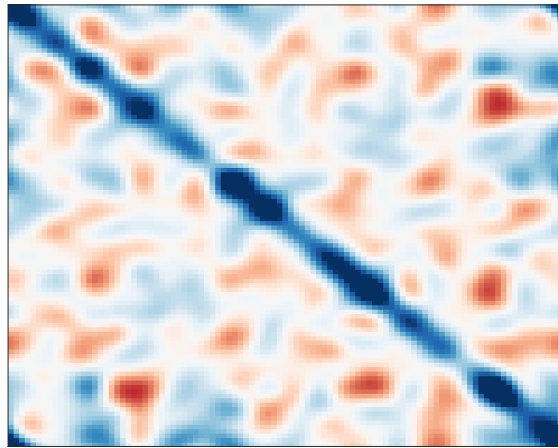
$$\mathbf{L}(\ell) \circ \hat{\mathbf{P}} = \mathbf{P}_{\text{loc}}$$

Localization matrix



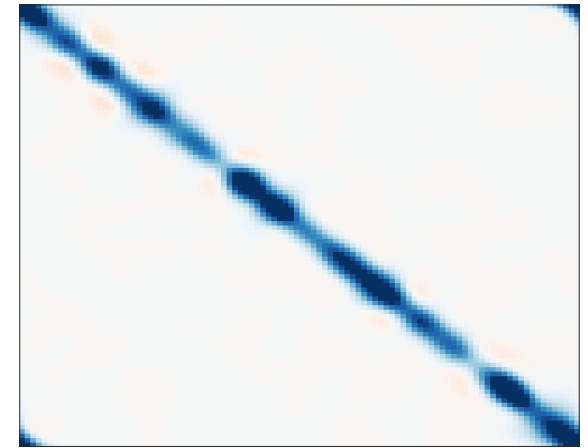
○

Empirical covariance



=

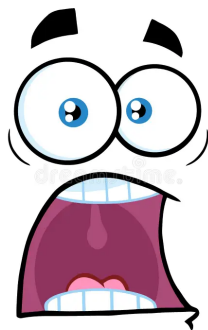
Localized estimate



○

=

True covariance



**What if I can't  
localize?**

Problem formulation

Partial solution in NWP

***Inspiration***

NICE in 3 steps

Examples

Summary

# What if I can't localize?

---

## *Idea*

- Replace spatial decay of correlation with “another” assumption

## *What's the best covariance correction?*

$$\min_{\mathbf{L}} \left| \langle \mathbf{L} \circ \hat{\mathbf{P}} - \mathbf{P}_{n_e \rightarrow \infty} \rangle \right|_{\text{Fro}}^2$$

$$\mathbf{L}_{ij} = \frac{\rho_{ij}^2 n_e (n_e - 1)}{\rho_{iijj} (n_e - 1) + \rho_{ij}^2 (n_e^2 - 2n_e + 2) + 1},$$

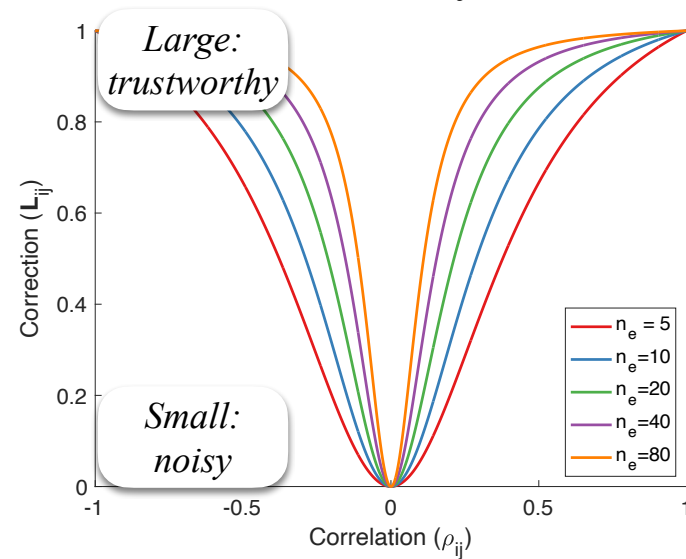
$$\mathbf{L}_{ij} = \frac{\rho_{ij}^2 (n_e - 1)}{1 + \rho_{ij}^2 n_e} \quad (\text{Gaussian})$$



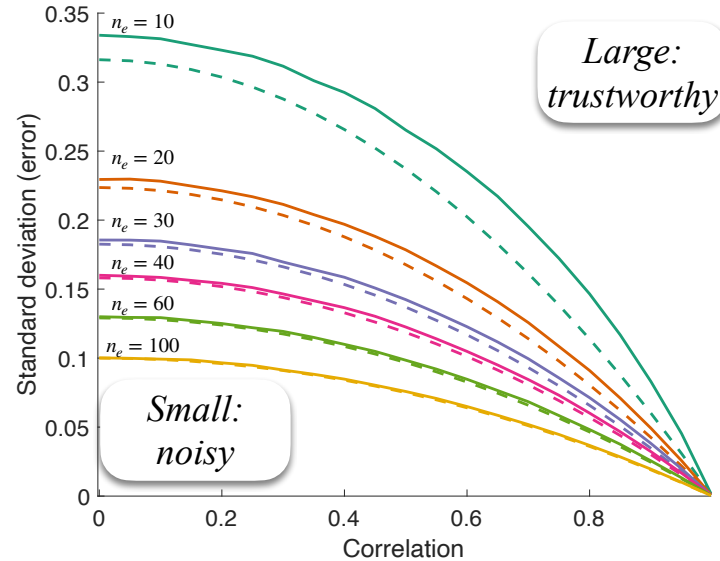
# Accuracy of correlation estimates

## Optimal Localization

$$\mathbf{L}_{ij} = \frac{\rho_{ij}^2 (n_e + 1)}{1 + \rho_{ij}^2 n_e}$$

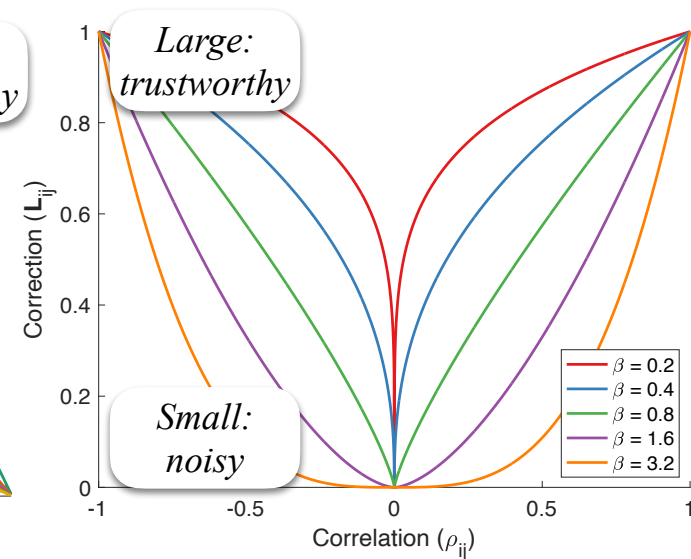


## Standard deviation of ensemble correlation\* (expected error)



## Sampling error correction\*\*

$$\mathbf{L}_{ij} = |\rho_{ij}|^\beta$$



**Small correlations are noisy, large correlations are trustworthy**

\*\* Y. Lee, 2021

\*Anderson, T.W., 2003. *An Introduction to Multivariate Statistical Analysis*.

## ***Idea***

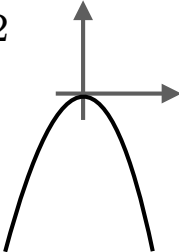
- “*Small correlations are noisy, large correlations are trustworthy*”
- The method should further be:
  - *Adaptive* (no/little tuning)
  - *Inexpensive* (no optimization over matrices)
  - *Guarantee PSD* estimates (stability of EnKFs)

## *Do I really care about PSD?*

- Yes!
- EnKF goes unstable if covariance estimates are not PSD

$$\mathbf{K} = \mathbf{P}\mathbf{H}^T \underbrace{(\mathbf{H}\mathbf{P}\mathbf{H}^T + \mathbf{R})}^{\text{ill-conditioned}}^{-1}$$

- Yes!
- En-Var systems can't minimize if covariance estimates are not PSD

$$J(\mathbf{x}) = \left\| \mathbf{R}^{-1/2}(\mathbf{M}(\mathbf{x}) - \mathbf{y}) \right\|^2 + \left\| \mathbf{B}^{-1/2}(\mathbf{x} - \mathbf{m}) \right\|^2$$


***PSD constraint is not for elegance, it's needed!***

Problem formulation

Partial solution in NWP

Inspiration

***NICE***

Examples

Summary

## ***Idea***

- *“Small correlations are noisy, large correlations are trustworthy”*
- The method should further be:
  - *Adaptive* (no/little tuning)
  - *Inexpensive* (no optimization over matrices)
  - *Guarantee PSD* estimates (stability of EnKFs)

## ***Ingredients***

- *Inexpensive* → *raise correlations to a power*
- *PSD guarantees* → *Choose powers wisely and interpolate*
- *Adaptivity* → *Make corrections based on expected noise level*  
*(Morozov’s discrepancy principle)*

# Morozov's discrepancy principle

---

## *Classical inverse theory*

$$\min_x \|\mathbf{y} - f(\mathbf{x})\|_2^2 + \alpha \|\mathbf{x}\|_2^2$$

- Solution  $\mathbf{x}_\alpha^*$  depends on “regularization parameter”  $\alpha$
- Determine  $\alpha$  such that

$$\|\mathbf{y} - f(\mathbf{x}_\alpha^*)\|_2 \leq S$$

↑  
Noise level

---

## *Application to correlation correction*

- Correct correlation by raising it to a power:

$$\hat{\rho}_\alpha = \rho^\alpha \cdot \rho$$

- Pick largest  $\alpha$  such that:

$$\|\hat{\rho} - \hat{\rho}_\alpha\|_2 \leq S$$

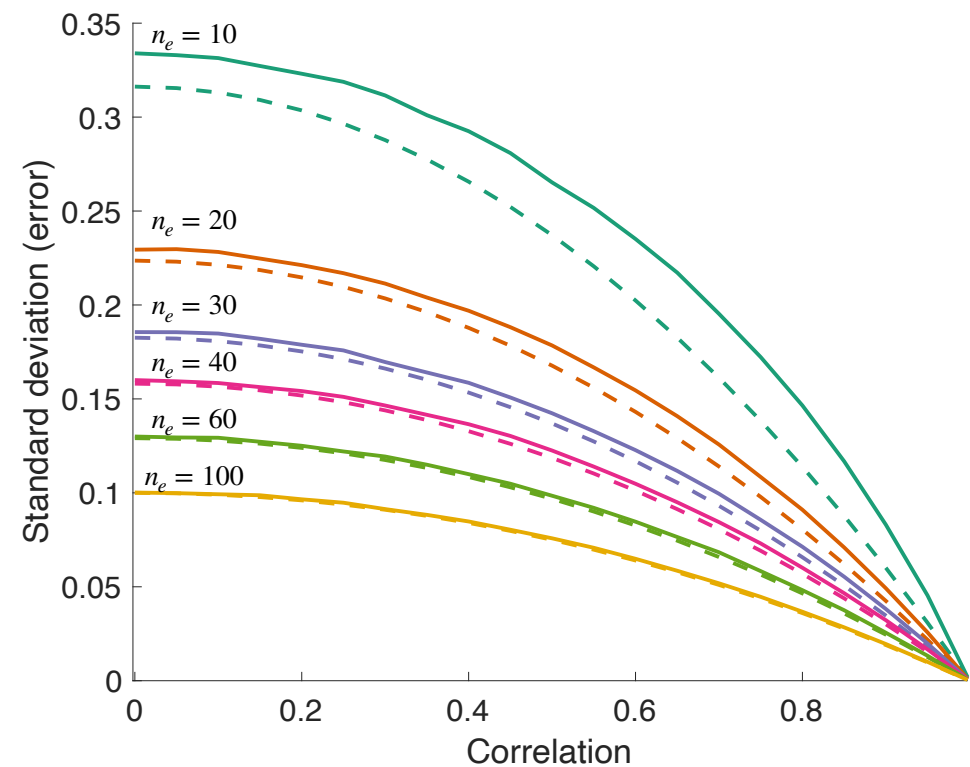
***NICE in 3 steps***



# Noise-informed Covariance Estimation (NICE)

**1. Compute expected noise level**

$$S_\rho = \sqrt{\sum_{i=1}^n \sum_{j=1}^n (\sigma_{\rho_{ij}})^2} \quad \sigma_{\rho_{ij}} = \frac{1 - \rho_{ij}^2}{\sqrt{n_e}}$$



# Noise-informed Covariance Estimation (NICE)

---

1. Compute expected noise level

$$S_\rho = \sqrt{\sum_{i=1}^n \sum_{j=1}^n (\sigma_{\rho_{ij}})^2} \quad \sigma_{\rho_{ij}} = \frac{1 - \rho_{ij}^2}{\sqrt{n_e}}$$

---

## 2. Compute a strong correction

- Raise correlation to an *even* power to preserve PSD property
- Determine power via discrepancy principle - break it to obtain a correction that is “too strong”

$$\hat{\rho}_\gamma = \hat{\rho}^{\circ\gamma} \circ \hat{\rho},$$

$$\|\hat{\rho} - \hat{\rho}_{\gamma^*}\|_{\text{Fro}} \geq \delta S_\rho$$

**Pick the *smallest*  $\gamma$  that breaks the discrepancy principle**

# Noise-informed Covariance Estimation (NICE)

---

1. Compute expected noise level

$$S_\rho = \sqrt{\sum_{i=1}^n \sum_{j=1}^n (\sigma_{\rho_{ij}})^2} \quad \sigma_{\rho_{ij}} = \frac{1 - \rho_{ij}^2}{\sqrt{n_e}}$$

---

2. Compute a strong correction

- Raise correlation to an *even* power to preserve PSD property
- Determine power via discrepancy principle - break it to obtain a correction that is “too strong”

$$\hat{\rho}_\gamma = \hat{\rho}^{\circ\gamma} \circ \hat{\rho},$$
$$\|\hat{\rho} - \hat{\rho}_{\gamma^*}\|_{\text{Fro}} \geq \delta S_\rho$$

**Pick the *smallest*  $\gamma$  that breaks the discrepancy principle**

3. ***Interpolate*** to a lower, even power to get the correction just right

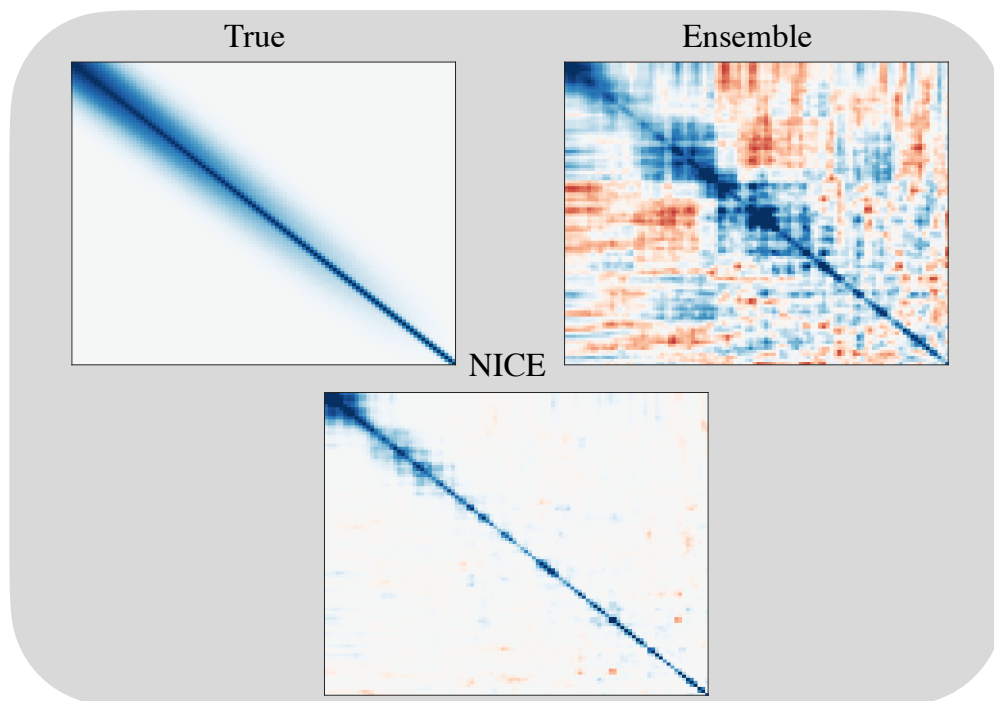
$$\mathbf{L}(\alpha) = \alpha \hat{\rho}^{\circ\gamma^*} + (1 - \alpha) \hat{\rho}^{\circ(\gamma^*-2)}$$

$$\hat{\rho}_\alpha = \mathbf{L}(\alpha) \circ \hat{\rho}$$

$$\|\hat{\rho} - \hat{\rho}_{\alpha^*}\|_{\text{Fro}} \leq \delta S_\rho,$$

# Noise-informed Covariance Estimation (NICE)

1. It's a few lines of code
2. NICE has connections with Jeff Anderson's *sampling error correction* but guarantees PSD estimates
3. Can be used within LETKF or EAKF or En-Var



```
function [Cov_NICER,Corr_NICER] = NICE(X,Y,fac)
Ne = size(X,2) ;
[CorrXY,~] = corr(X',Y');
std_rho = (1-CorrXY.^2)/sqrt(Ne);
sig_rho = sqrt(sum(sum(std_rho.^2)));

expo2 = 2:2:10;
for kk=1:length(expo2)
    L = abs(CorrXY).^expo2(kk);
    Corr_NICER = L.*CorrXY;
    if norm(Corr_NICER - CorrXY,'fro') > fac*sig_rho
        break
    end
end
expo1 = expo2-2;
rho_exp1 = CorrXY.^expo1;
rho_exp2 = CorrXY.^expo2;

a1 = .1:.1:1;
for kk=1:length(a1)
    L = (1-a1(kk))*rho_exp1+a1(kk)*rho_exp2;
    Corr_NICER = L.*CorrXY;
    if kk>1 && norm(Corr_NICER - CorrXY,'fro') > fac*sig_rho
        Corr_NICER = PrevCorr;
        break
    end
    PrevCorr = Corr_NICER;
end

Vy = diag(std(Y,0,2));
Vx = diag(std(X,0,2));
Cov_NICER = Vx*Corr_NICER*Vy;

end
```

# What else can I try?

## Competitors

Statistics

- Graphical lasso (*Tibshirani*)
- Soft thresholding (*Wainwright*)
- Sparse Covariance estimation (*Xue, Ma, Zou*)
- Convex sparse Cholesky selection (*Rajaratnam*)
- Sampling error correction (*Anderson, Lee*)
- Optimal localization (*benchmark*)

adaptive, efficient, PSD

adaptive, efficient, PSD

adaptive, efficient, PSD

adaptive, efficient, PSD

adaptive, efficient, PSD

infeasible

### Stats methods solve optimization problems

Graphical Lasso

$$F_{\text{G-Lasso}}(\Theta) = \text{tr}(\hat{\mathbf{P}}\Theta) - \log \det(\Theta) + \lambda \sum_{j \neq k} |\Theta_{jk}|$$

Convex Sparse  
Cholesky Selection

$$F_{\text{CSCS}}(\mathbf{A}) = \text{tr}(\mathbf{A}^T \mathbf{A} \hat{\mathbf{P}}) - 2 \log \det(\mathbf{A}) + \lambda \sum_{1 \leq j < i} |\mathbf{A}_{ij}|$$

Sparse covariance  
estimation

$$F_{\text{SCE}}(\mathbf{P}) = \frac{1}{2} \|\mathbf{P} - \hat{\mathbf{P}}\|_{\text{Fro}}^2 + \lambda \sum_{j \neq k} |\mathbf{P}_{jk}|$$

Problem formulation

Partial solution in NWP

Inspiration

NICE in 3 steps

***Examples***

Summary

# Does NICE work?

---

## ***Test cases***

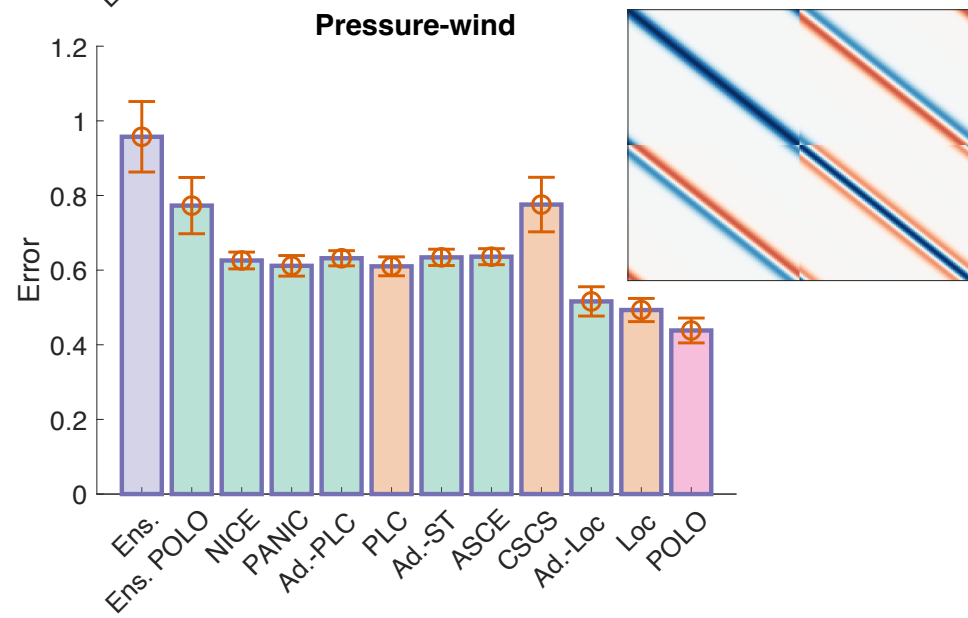
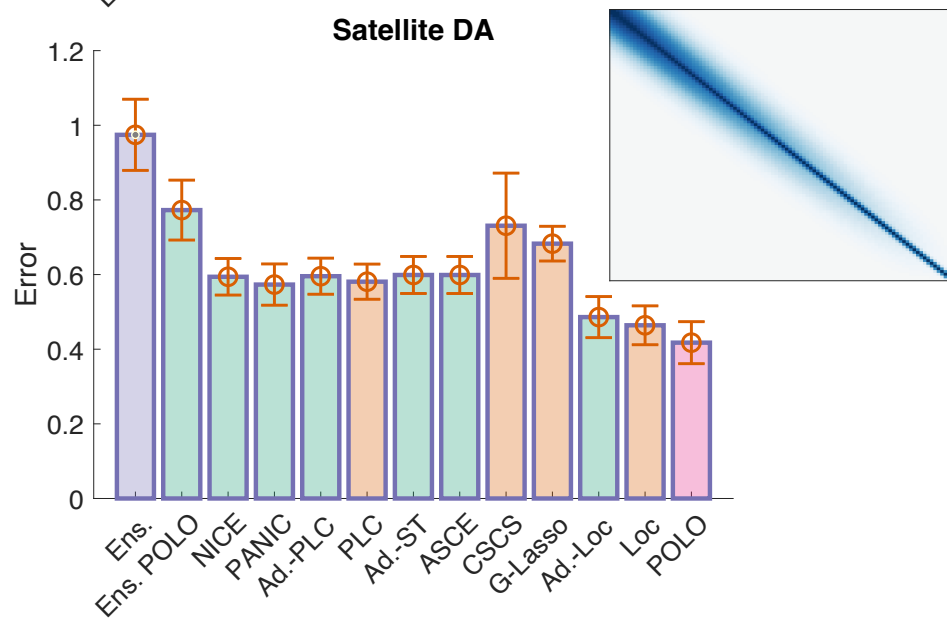
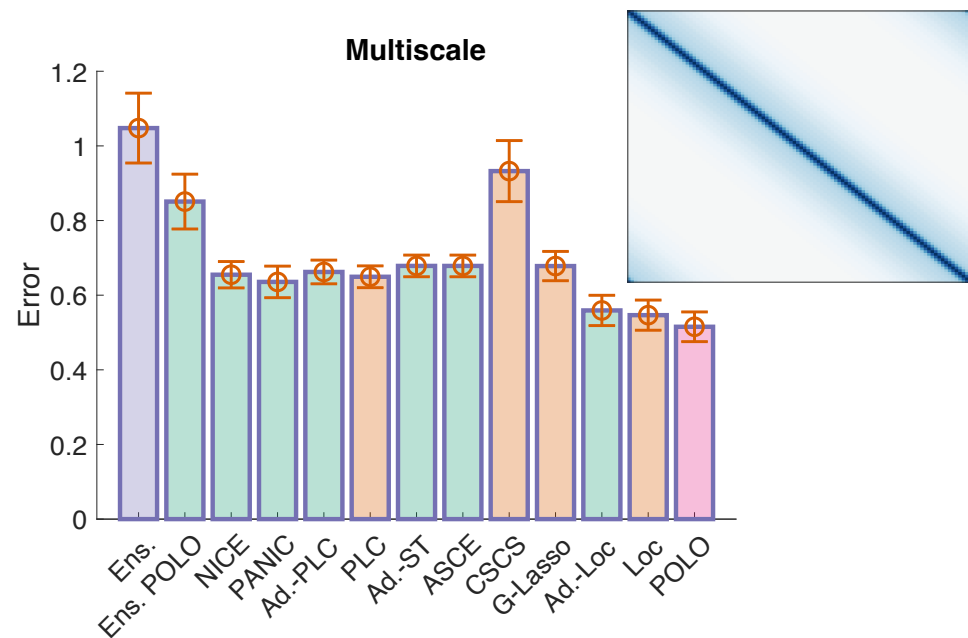
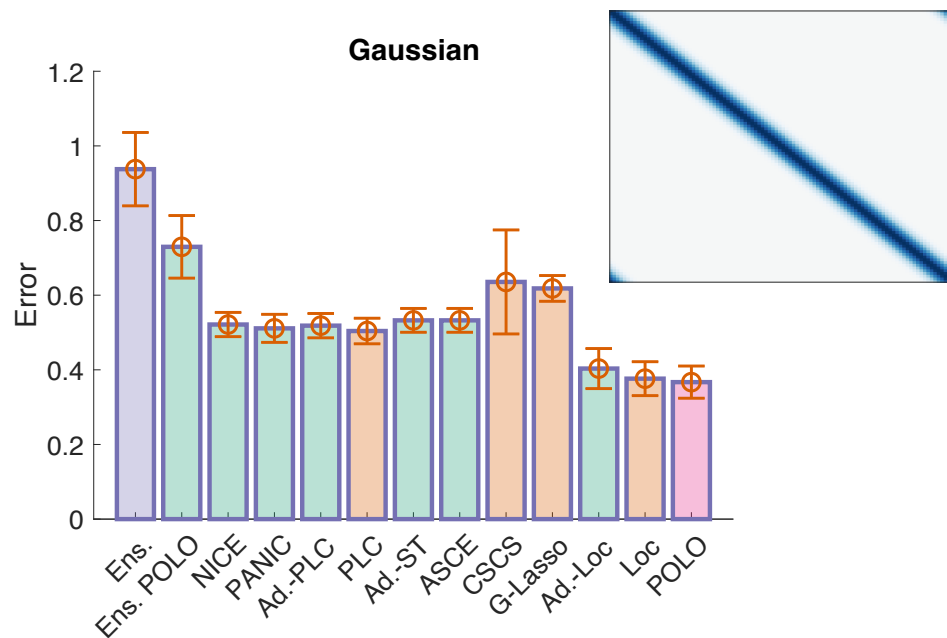
- Synthetic covariance examples (sanity check)
- Cycling DA with EnKF (synthetic and semi-real)
- Inversion of electromagnetic field data
- Training of feed-forward neural net on time averaged data from a chaotic dynamics

***NICE is just as good or better than other, tuned methods, but computationally more efficient***

## ***Bonus:***

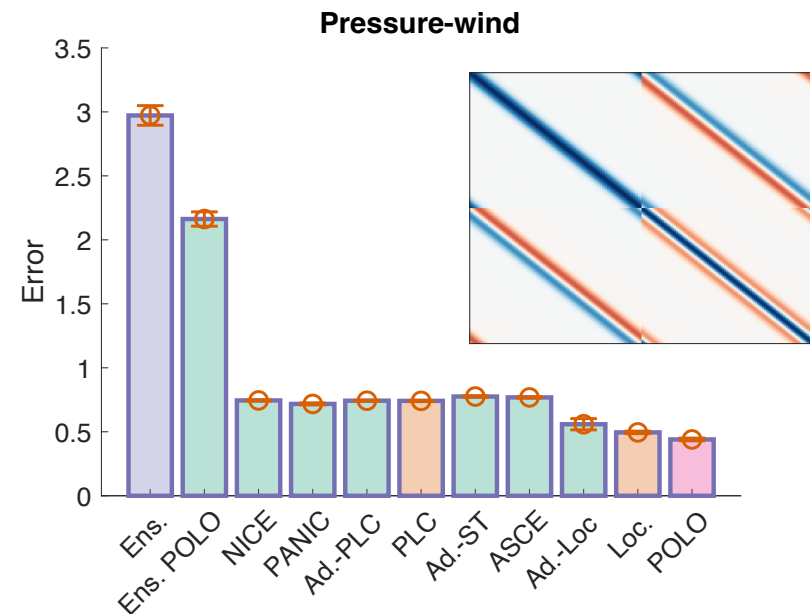
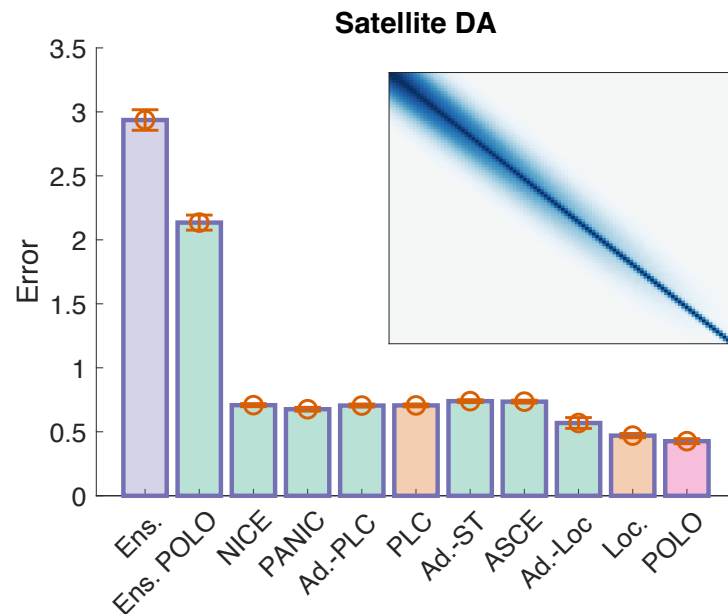
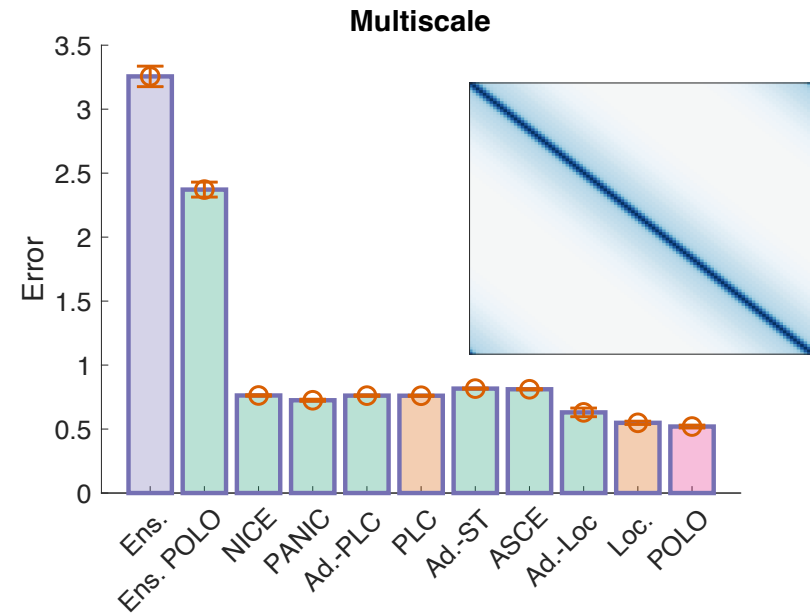
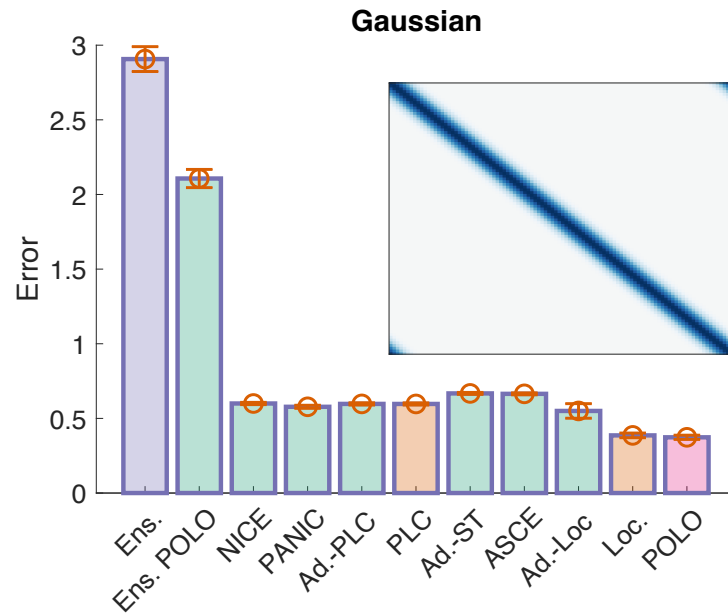
- Discrepancy principle can make other methods adaptive
- We created 5 other adaptive schemes
  1. Partially adaptive NICE (PANIC)
  2. Adaptive localization (Ad.-Loc)
  3. Adaptive power law corrections (Ad.-PLC)
  4. Adaptive soft thresholding (Ad.-ST)
  5. Adaptive sparse covariance estimation (ASCE)

# Sanity checks on Gaussians: $n = 100$ , $n_e = 20$

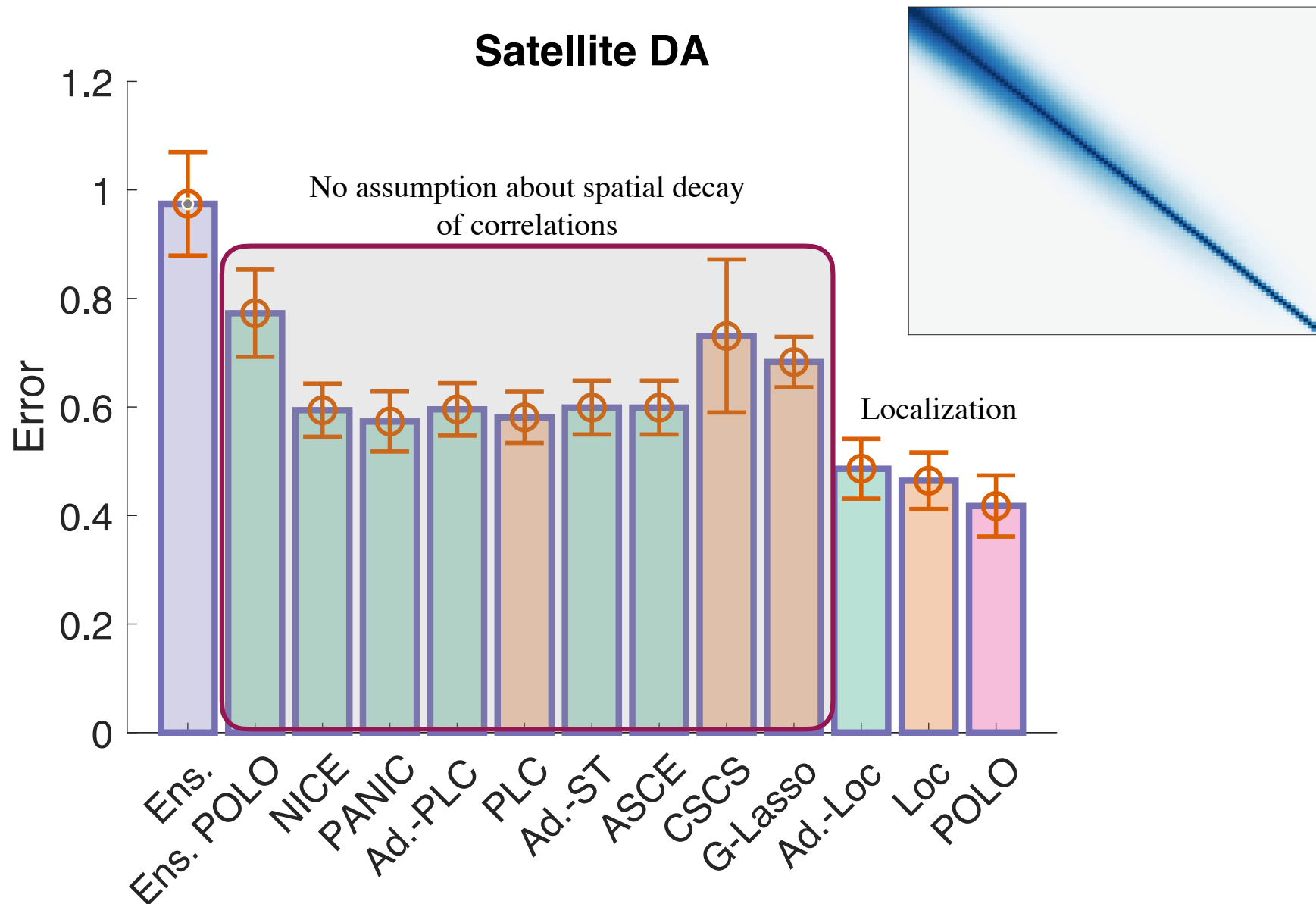




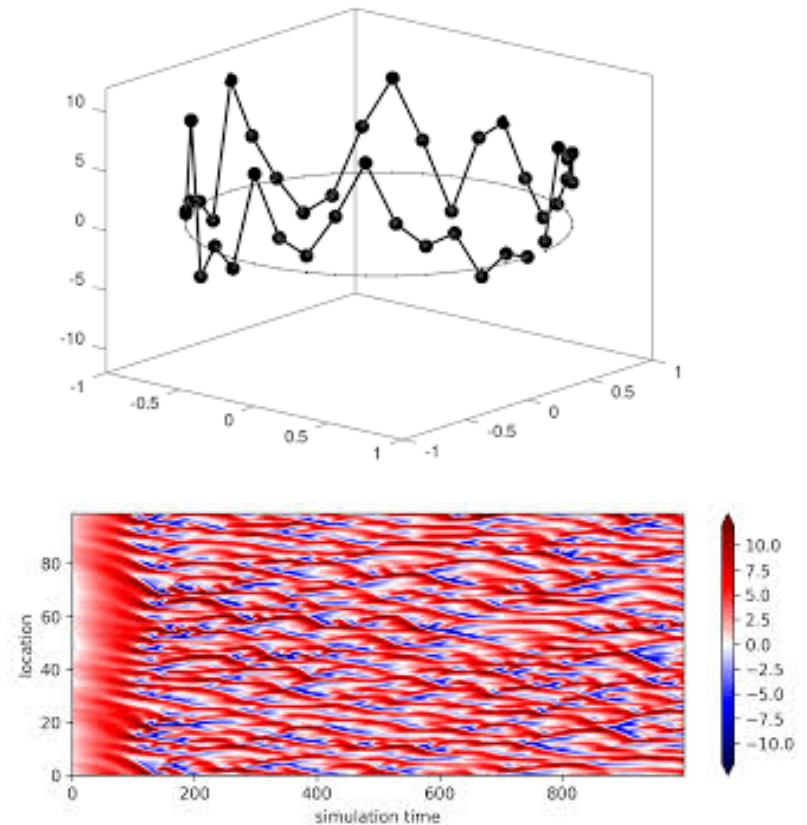
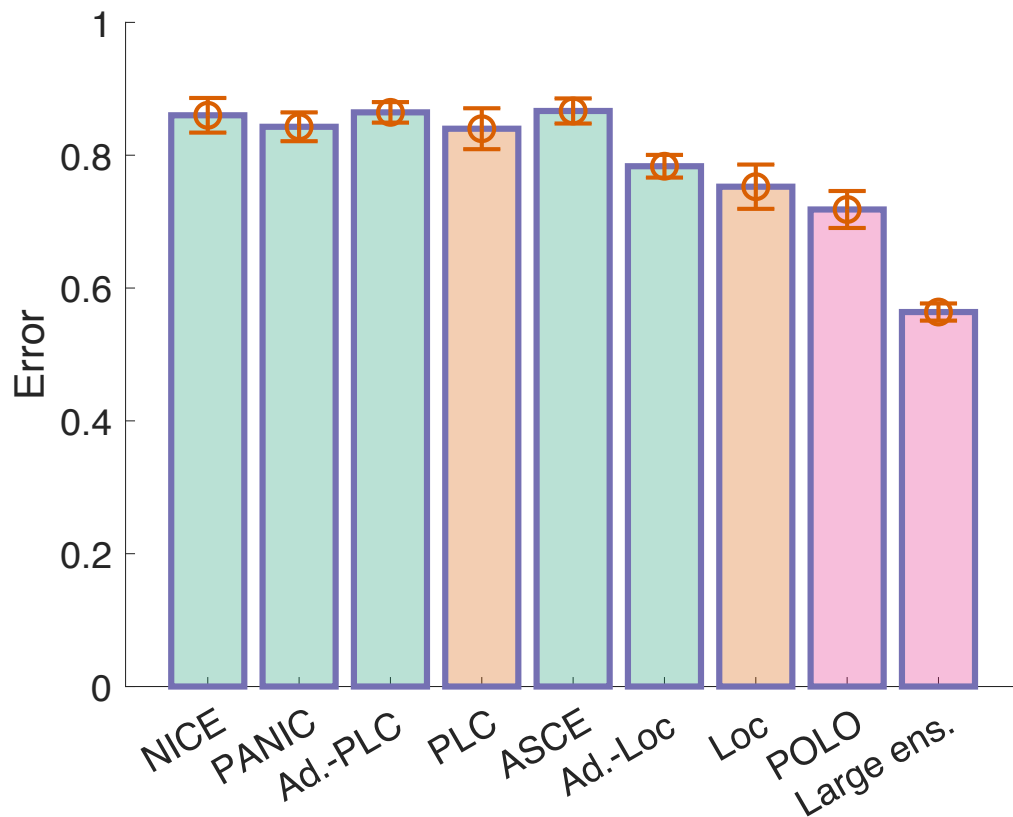
# Sanity checks on Gaussians: $n = 1000$ , $n_e = 20$



# Sanity checks on Gaussians: $n = 100$ , $n_e = 20$



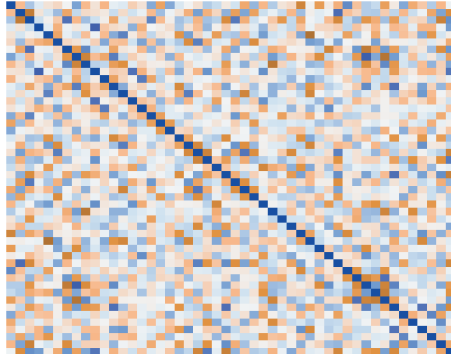
# Sanity checks on Lorenz '95 (*yawn*)



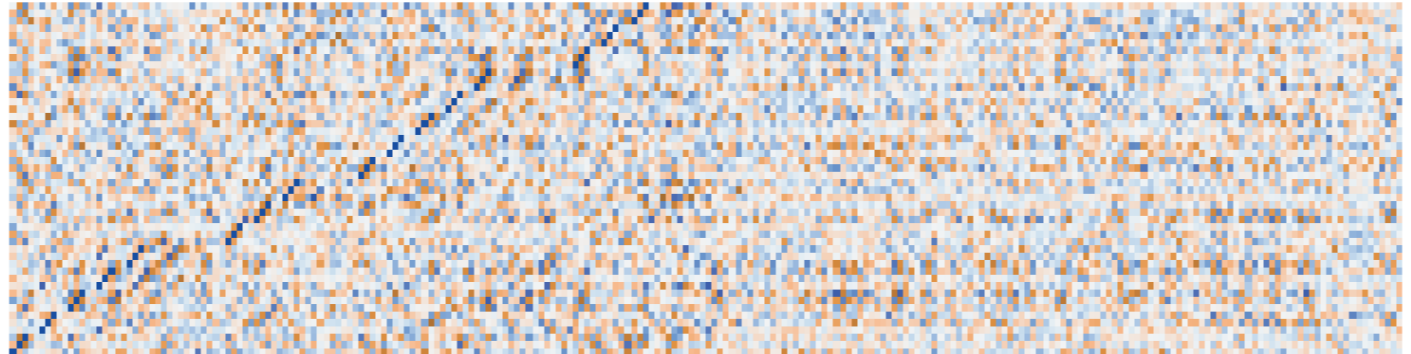
- Typical cycling DA setup from MWR papers
- Reiterates results of static covariance estimation
- Cycling DA does not break our ideas

# Geomagnetic DA

$\mathbf{H}\mathbf{P}\mathbf{H}^T$

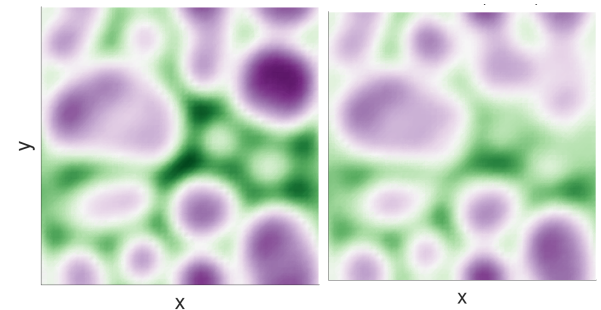


$(\mathbf{P}\mathbf{H}^T)^T$



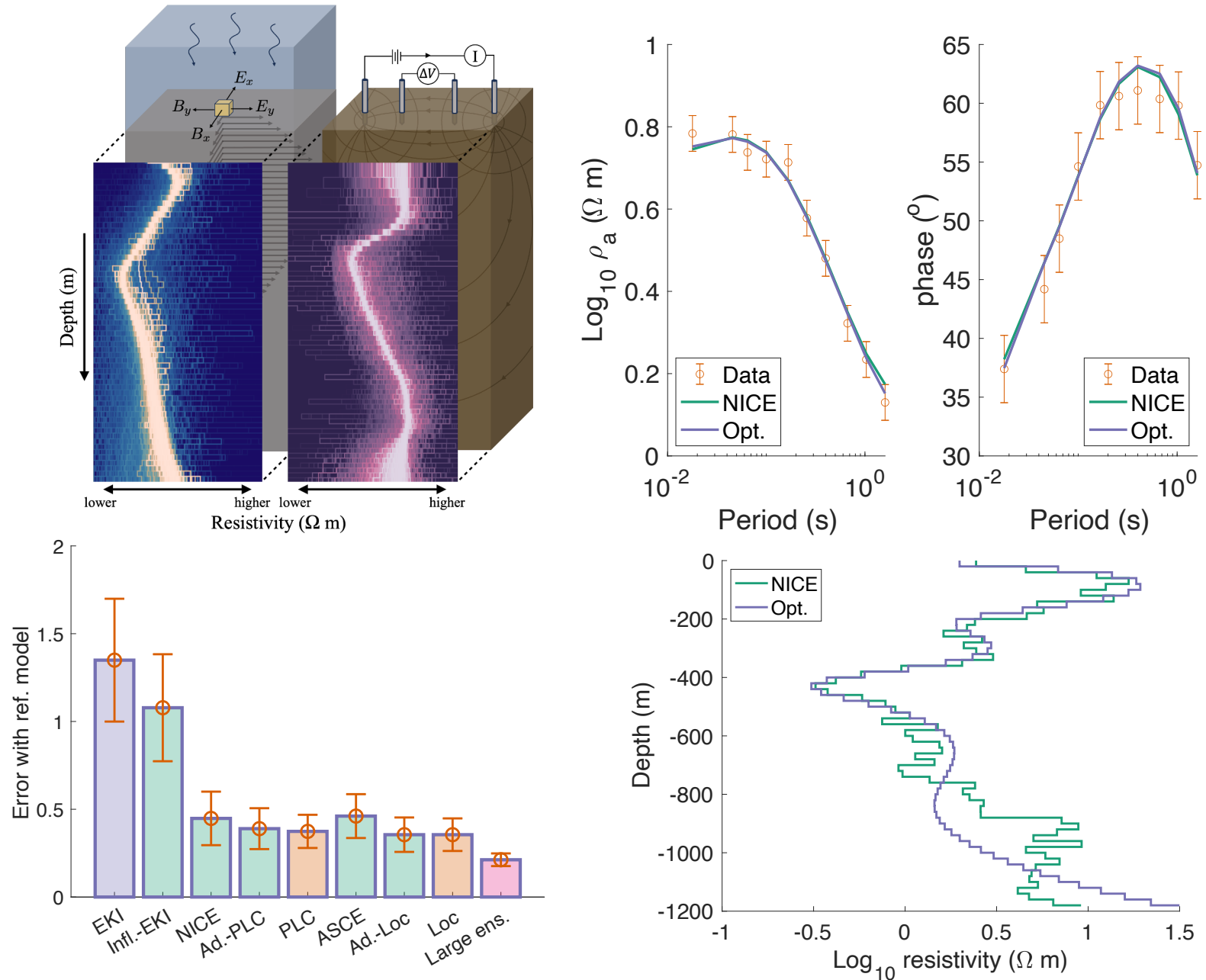
- KS equation coupled to an induction equation
- One quantity (magnetic field) is observed, the other (velocity) unobserved
- Correlation structure is erratic/hard to anticipate
- Localization fails on this example
- NICE outperformed a heavily tuned shrinkage correction
- NICE is not tuned in any way

Vorticity of velocity (Truth)      Vorticity of velocity (NICE)

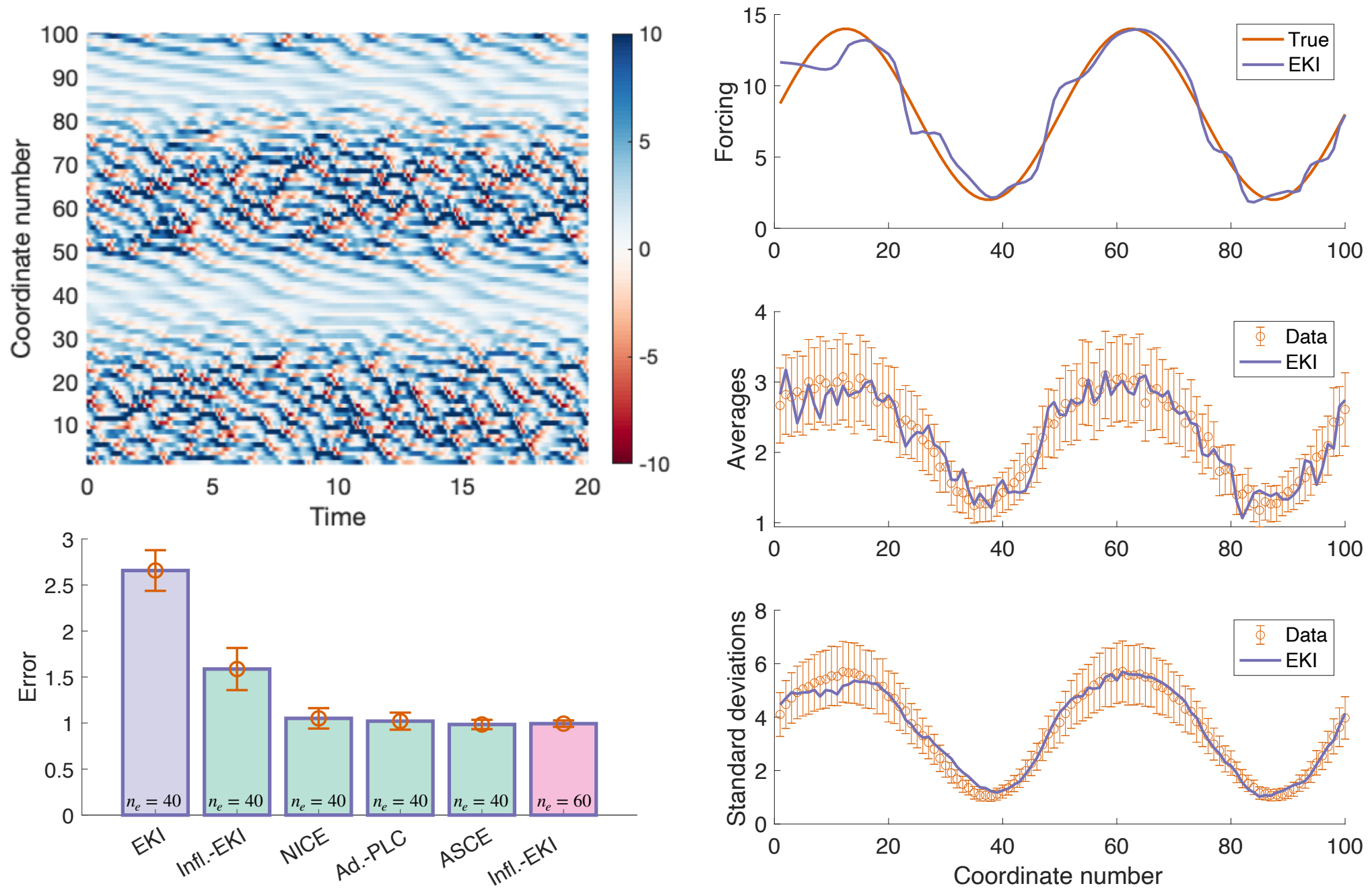


|                   | Error in mag. field | Error in vel. field |
|-------------------|---------------------|---------------------|
| Shrinkage (tuned) | 1.2                 | 2.0                 |
| Ad.-PLC           | 1.2                 | 2.5                 |
| NICE              | 1.0                 | 1.8                 |
| Large ens.        | 0.7                 | 1.1                 |

# Geophysical inversion of marine magnetotelluric data



# Training a NN with time-averaged data of chaotic dynamics



# Where can I try NICE?

---

- Our Matlab code is on github
- **CliMA/EnsembleKalmanProcesses.jl**  
CliMA - Climate Modeling Alliance
- **JEDI/SABER/BUMP**  
JEDI - Joint Effort for Data assimilation Integration  
SABER - System Agnostic Background Error Representation  
BUMP - Background error on Unstructured Mesh Package
- **NEDAS**  
Next-generation Ensemble Data Assimilation System



Problem formulation

Partial solution in NWP

Inspiration

NICE in 3 steps

Examples

***Summary***



# Summary

---

## Facts

- High-dimensional covariance estimation from a small number of samples is (a) important; and (b) difficult.
- NWP has solved this problem under the assumption that correlation decays with distance (*but statisticians don't seem to care*)

## Contribution

- Assumption: *Small/medium correlations are noisy and should be more heavily damped than large ones*
- We turned this idea into a simple, robust, adaptive algorithm (NICE)

## NICE Properties

- Adaptive and tuning-free
- PSD guarantees (as opposed to many other schemes)
- Works just as well or better than more sophisticated alternatives
- Has proven useful in a large number of diverse test problems

## Bonus

- We made 5 other covariance estimation schemes adaptive

## Summary (continued)

---

### Cross-correlations are hard

- NICE works well for  $\mathbf{P}$ ,  $\mathbf{H}\mathbf{P}\mathbf{H}^T$ ,  $\mathbf{C}_{yy}$

$$\mathbf{x}_i^a = \mathbf{x}_i + \mathbf{K} (\mathbf{y} - (\mathbf{H}\mathbf{x}_i + \boldsymbol{\eta}_i))$$

$$\mathbf{K} = \mathbf{C}_{xy} (\mathbf{C}_{yy} + \mathbf{R})^{-1}$$

↑    ↑  
**hard** **easy**

- Current strategy: Turn down the correction with a “fudge factor”
- Replace noise level:  $S_\rho \rightarrow \delta S_\rho$ , where  $\delta \leq 1$

### Scalability: How large can I go?

- NICE requires that we calculate *all* correlations  $\rightarrow$  scalable to  $O(10^4)$  variables
- NICE can be incorporated into scalable frameworks (ETKF, EAKF) and then is as scalable as the framework (requires additional assumptions)
- NICE can also be incorporated into hybrid ensemble-variational schemes (JEDI)



# Thank you.

Email: *matti@ucsd.edu*

*Thanks to the*  
**Office of Naval Research**  
for generous support

- D.N. Vishny, M. Morzfeld, K. Gwartz, E. Bach, O.R.A. Dunbar, D. Hodyss, *High-dimensional covariance estimation from a small number of samples*, Journal of Advances in Modeling Earth Systems, 16, e2024MS004417 (2024)
- M. Morzfeld, D. Hodyss, *A theory for why even simple covariance localization is so useful in ensemble data assimilation*, Monthly Weather Review 151(3), 717-736 (2023)
- D. Hodyss, M. Morzfeld, *How sampling errors in covariance estimates cause bias in the Kalman gain and impact ensemble data assimilation*, Monthly Weather Review 151(9), 2413- 2426 (2023)