Noise-informed covariance estimation

Matthias Morzfeld

Cecil H. and Ida M. Green Institute of Geophysics and Planetary Physics

Scripps Institution of Oceanography

University of California, San Diego

Powered by

David Vishny
(SIO)

Kyle Gwirtz
(NASA)

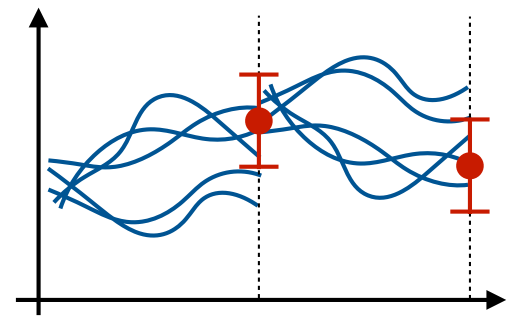
Oliver Dunbar
(Caltech)

Eviatar Bach
(Reading)

Dan Hodyss
(NRL)

Ensemble Kalman filter/smoother

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

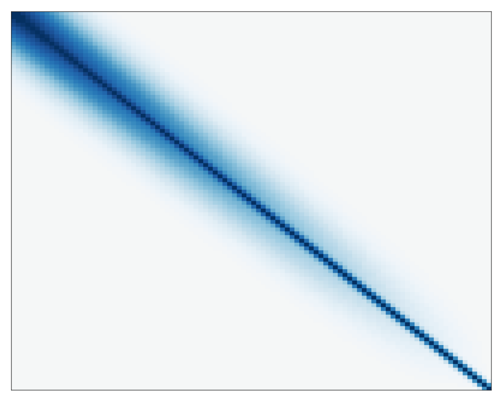


$$\hat{\mathbf{P}} = \frac{1}{n_e - 1} \sum_{i=1}^{n_e} (\mathbf{x}_i - \bar{\mathbf{x}}) \otimes (\mathbf{x}_i - \bar{\mathbf{x}}), \quad \bar{\mathbf{x}} = \frac{1}{n_e} \sum_{i=1}^{n_e} \mathbf{x}_i$$

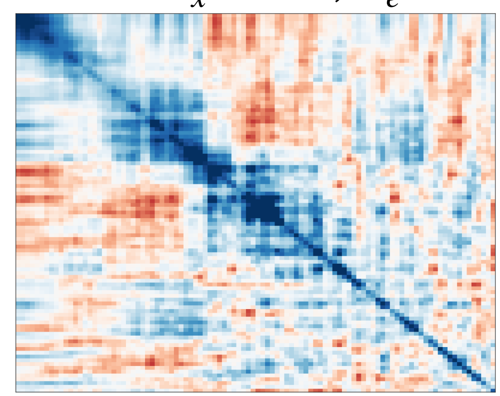
Caveat

- Ensemble size is small
- Dimension is large, $n_e \ll n_x$
- Covariance estimate is inaccurate

True covariance matrix

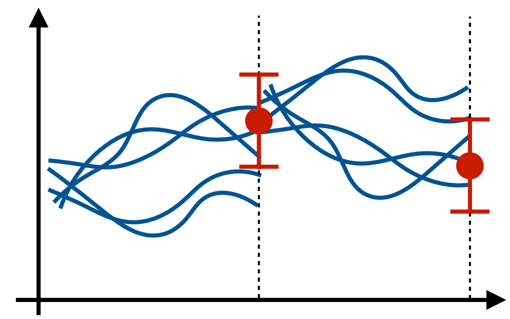


Estimate: $n_x = 100$, $n_e = 20$



Ensemble Kalman filter/smoother

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



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

Caveat

- Ensemble size is small
- Dimension is large, $n_e \ll n_x$
- Covariance estimate is inaccurate

EnKF in NWP

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

How?

How I think the story went (*unverified*)

Ensemble Kalman filtering (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, CA*, is correlated to weather in *Vienna, Austria*
- They deleted the correlations that they knew should not be there
- ***EnKF started to work***

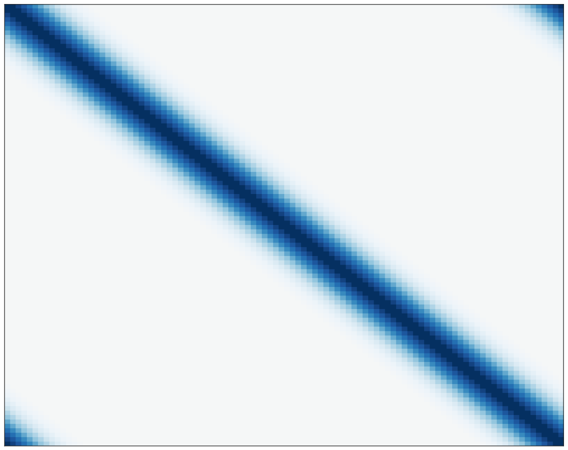
Covariance localization

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

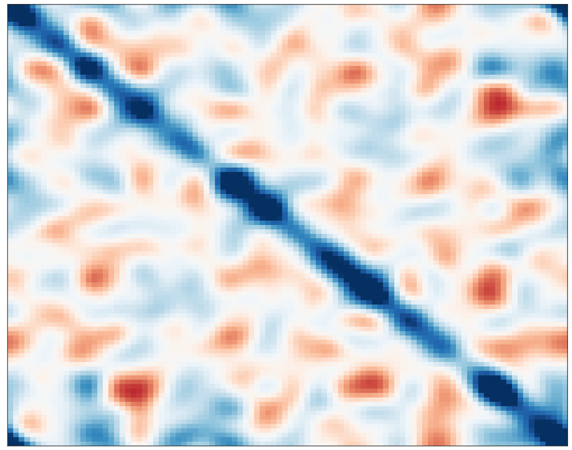
Covariance localization

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

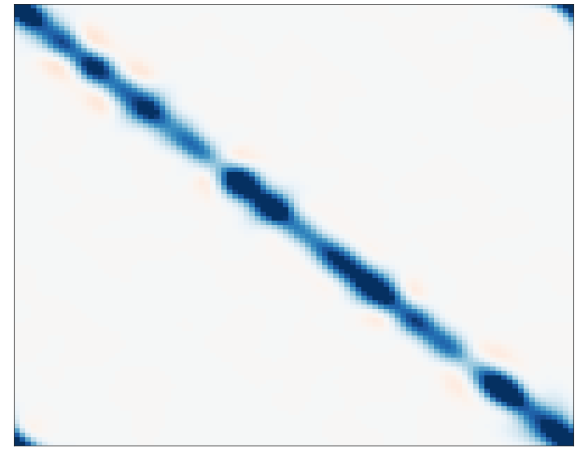
Localization matrix $\circ$ Empirical covariance $=$ Localized estimate



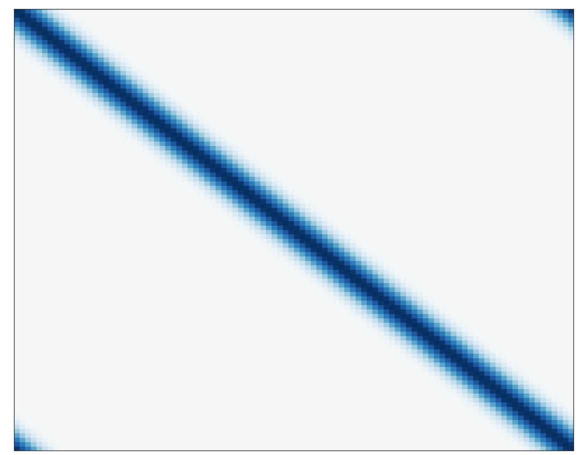
$\circ$

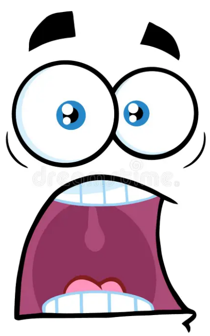


$=$



True covariance





What if I can't localize?

$n_e = 20, \quad n_x = 100$

Agenda

- Problem formulation
- ***Inspiration***
- NICE in 3 steps
- Examples
- Summary

What if I can't localize?

Examples

- Parameter estimation
- Training of neural networks for sub grid parameterizations (CliMA)
- Assimilation of nonlocal/integrated observations

Idea

- Replace spatial decay of correlation with "another" assumption

What's the best localization?

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

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

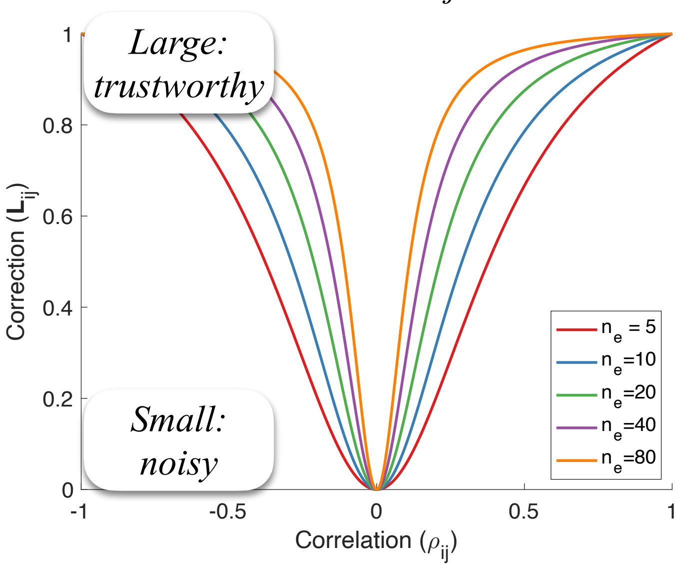
$$\mathbf{L}_{ij} = \frac{\rho_{ij}^2 (n_e - 1)}{1 + \rho_{ij}^2 n_e}$$ (Gaussian)

Morzfeld & Hodyss, 2023
Menetrier et al. 2015, Flowerdew, 2015

Accuracy of correlation estimates

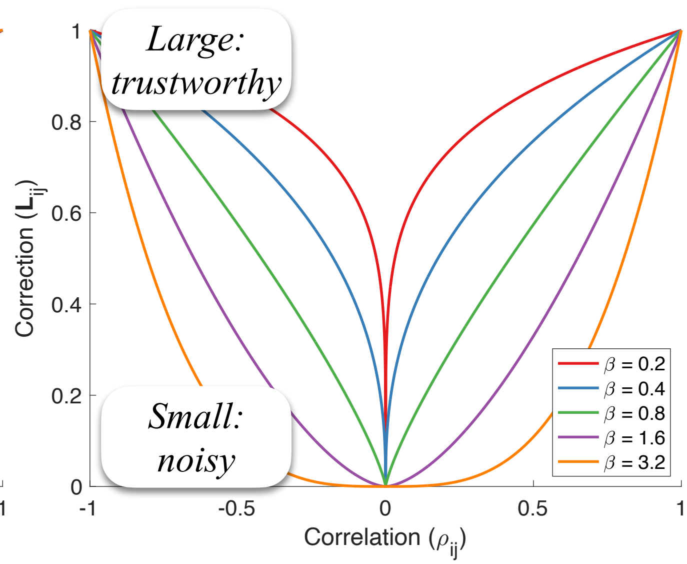
Optimal Localization

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

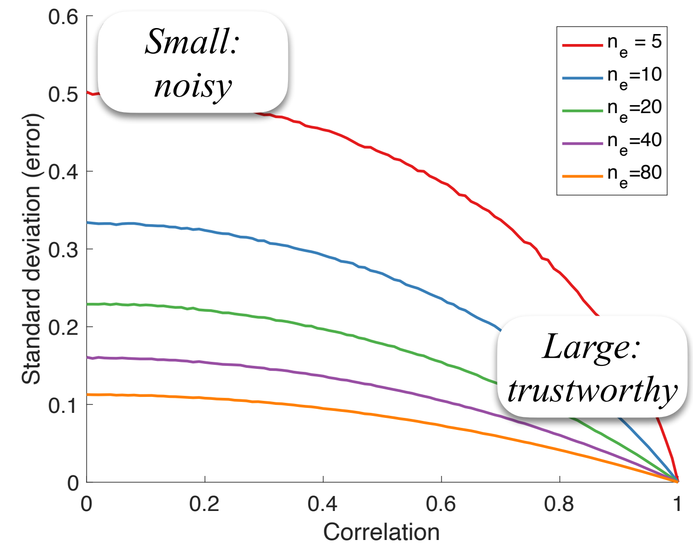


*Sampling error correction**

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



Standard deviation of ensemble correlation (expected error)



Small correlations are noisy, large correlations are trustworthy

* Y. Lee, 2021

Noise-informed Covariance Estimation (NICE)

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)

- Problem formulation
- Inspiration
- ***NICE in 3 steps***
- Examples
- Summary

Noise-informed Covariance Estimation (NICE)

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" α
- Determine α 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 α such that:

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

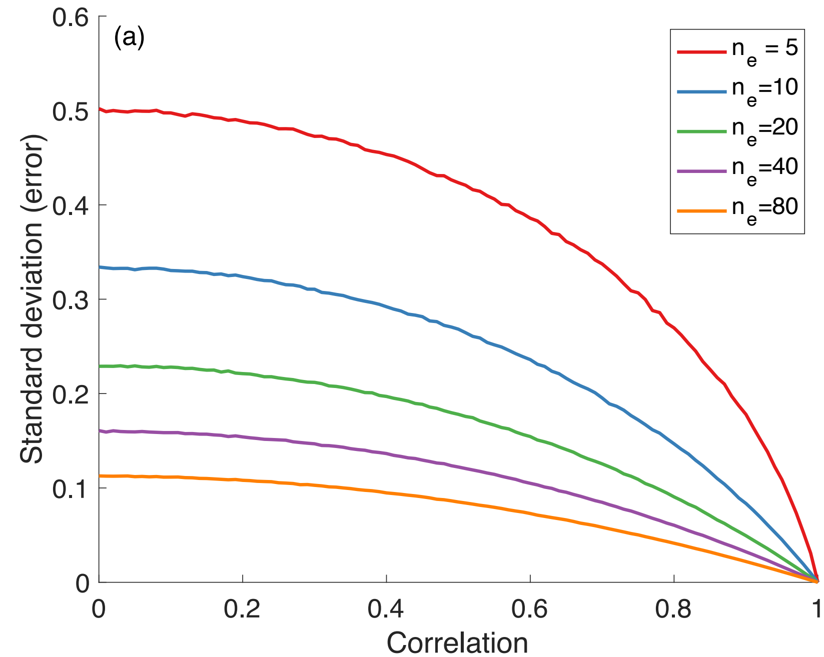
NICE in 3 steps

Noise-informed Covariance Estimation (NICE)

1. Compute expected noise level (look-up table)

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

- Compute the standard deviation of ensemble correlation
- Standard deviation is a proxy for the error
- Add (or average) over all variables



Noise-informed Covariance Estimation (NICE)

1. Compute expected noise level (look-up table)

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

2. Compute a strong correction

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

- 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"

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

Pick the *smallest* γ that breaks the discrepancy principle

Noise-informed Covariance Estimation (NICE)

1. Compute expected noise level (look-up table)

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

2. Compute a strong PSD correction

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

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

3. Interpolate to a lower, even power to get the correction just right (“back-paddle”)

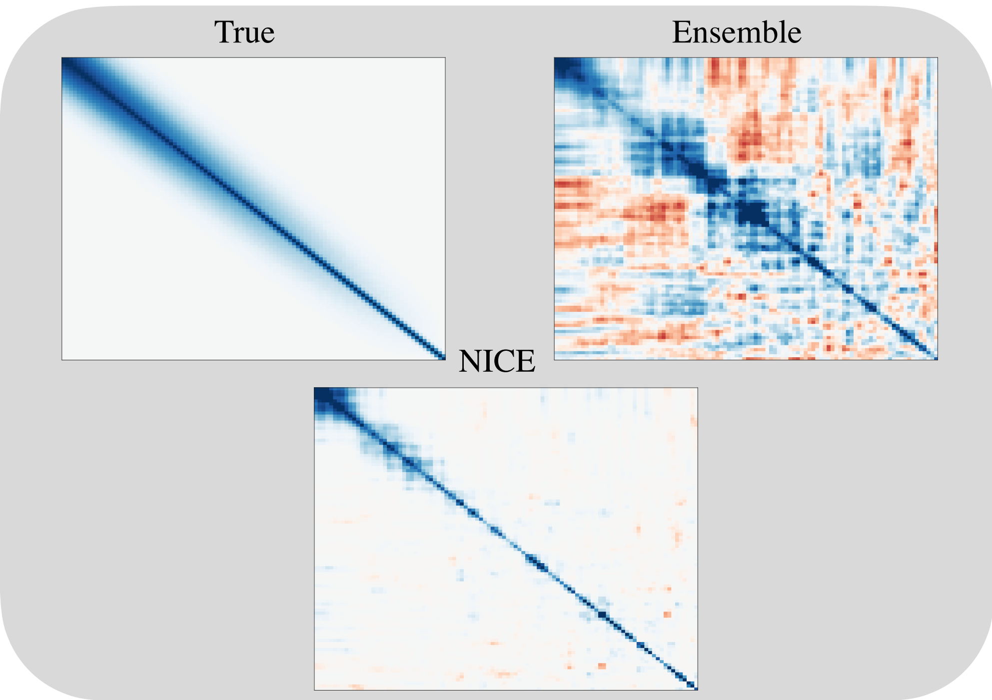
$$\mathbf{L}(\alpha) = \alpha\hat{\boldsymbol{\rho}}^{\circ\gamma^*} + (1-\alpha)\hat{\boldsymbol{\rho}}^{\circ(\gamma^*-2)}$$

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

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

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 does not require training/ensembles of ensembles and guarantees PSD estimates
3. Can be used within LETKF or EAKF




```
function [Cov_NICE,Corr_NICE] = NICE(X,Y,fac)
Ne = size(X,2) ;
FileName = strcat('std_ro_Ne_',num2str(Ne),'.mat');
load(FileName,'r','stdCrs')

[CorrXY,~] = corr(X',Y');
std_rho = interp1(r,stdCrs,CorrXY,'linear','extrap');
std_rho(CorrXY==1) = 0;
sig_rho = sqrt(sum(sum(std_rho.^2)));

go = 1;
expo2 = 0;
while go == 1
    expo2 = expo2+2;
    L = abs(CorrXY).^expo2;
    Corr_NICER = L.*CorrXY;
    if norm(Corr_NICER - CorrXY,'fro') > fac*sig_rho
        go = 0;
    end
end
expo1 = expo2-2;
rho_exp1 = CorrXY.^expo1;
rho_exp2 = CorrXY.^expo2;

al = 0.1:.1:1;
for kk=1:length(al)
    L = (1-al(kk))*rho_exp1+al(kk)*rho_exp2;
    Corr_NICE = L.*CorrXY;
    if kk>1 && norm(Corr_NICER - CorrXY,'fro') > fac*sig_rho
        Corr_NICE = PrevCorr;
        break
    elseif norm(Corr_NICE - CorrXY,'fro') > fac*sig_rho
        break
    end
    PrevCorr = Corr_NICE;
end

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

Cov_NICE = Vx*Corr_NICER*Vy;
```

What else can I try?

Competitors

Statistics

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

Stats methods solve optimization problems

Graphical Lasso

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

Convex Sparse Cholesky Selection

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

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

Inspiration

NICE in 3 steps

Examples

Summary

Does NICE work?

Test cases

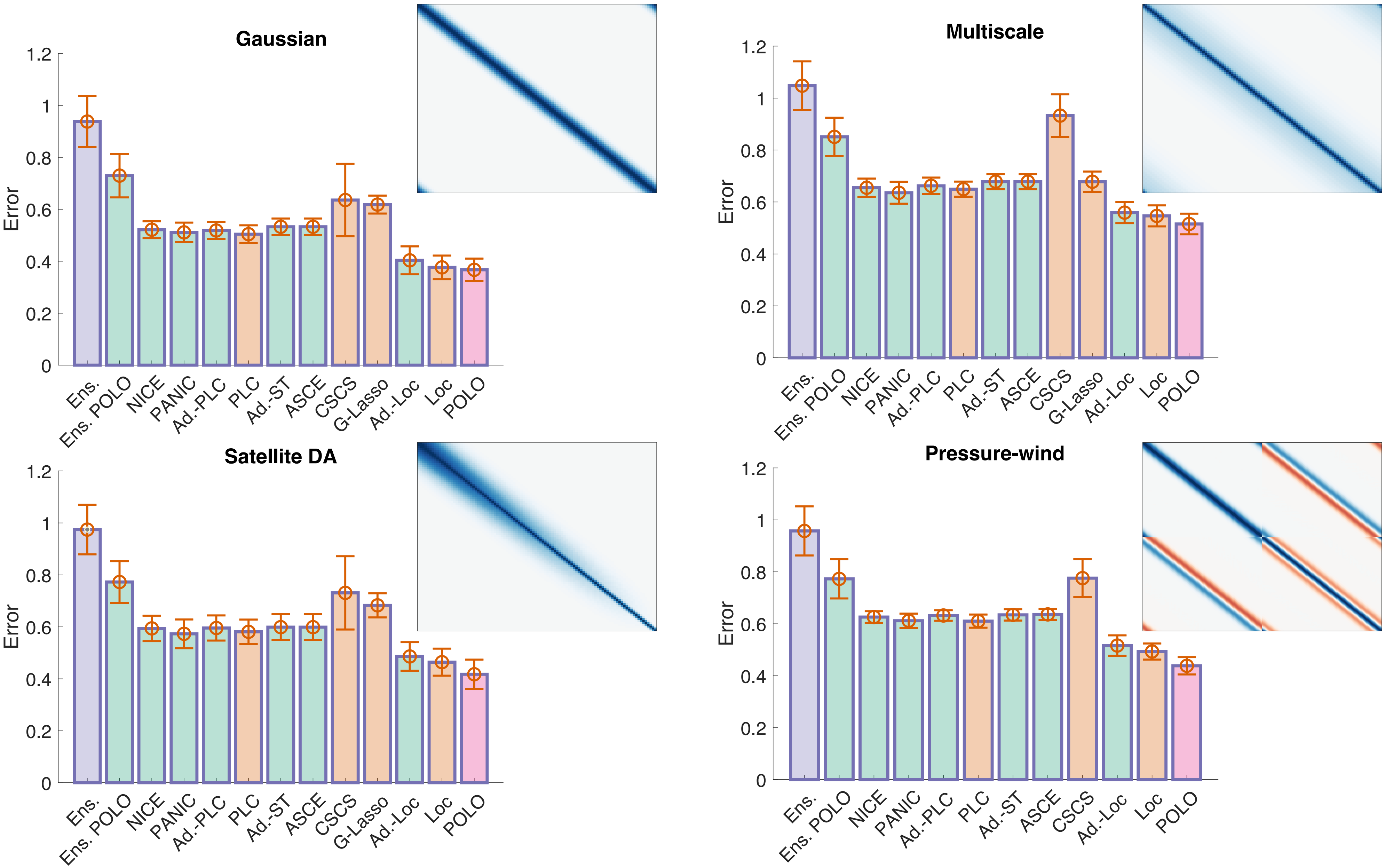
- Synthetic covariance examples (sanity check)
- Cycling DA with EnKF (synthetic and semi-real)
- Inversion of electromagnetic field data
- Training of feed-forward neural net on time averaged data from a chaotic dynamics

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

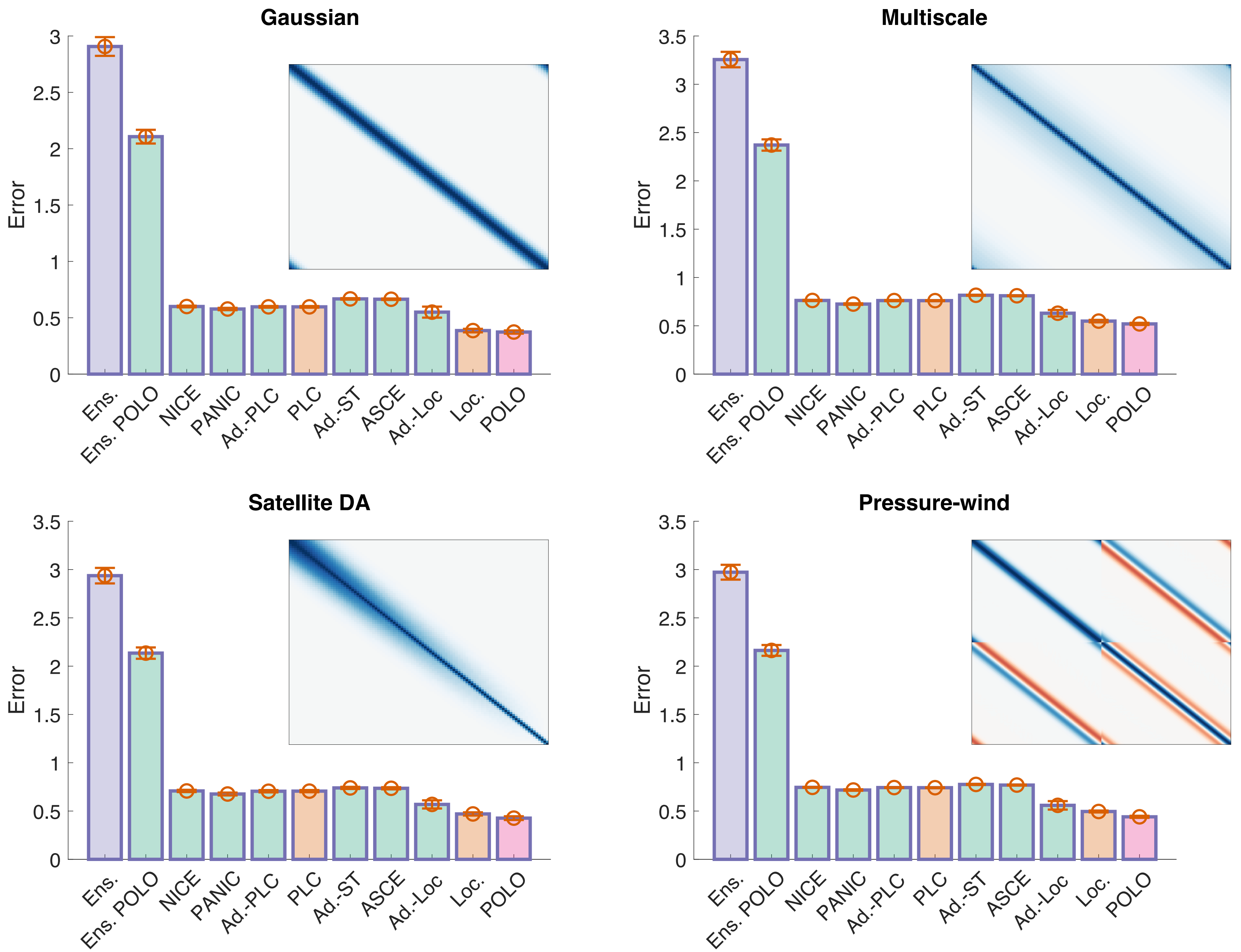
Bonus:

- 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)*

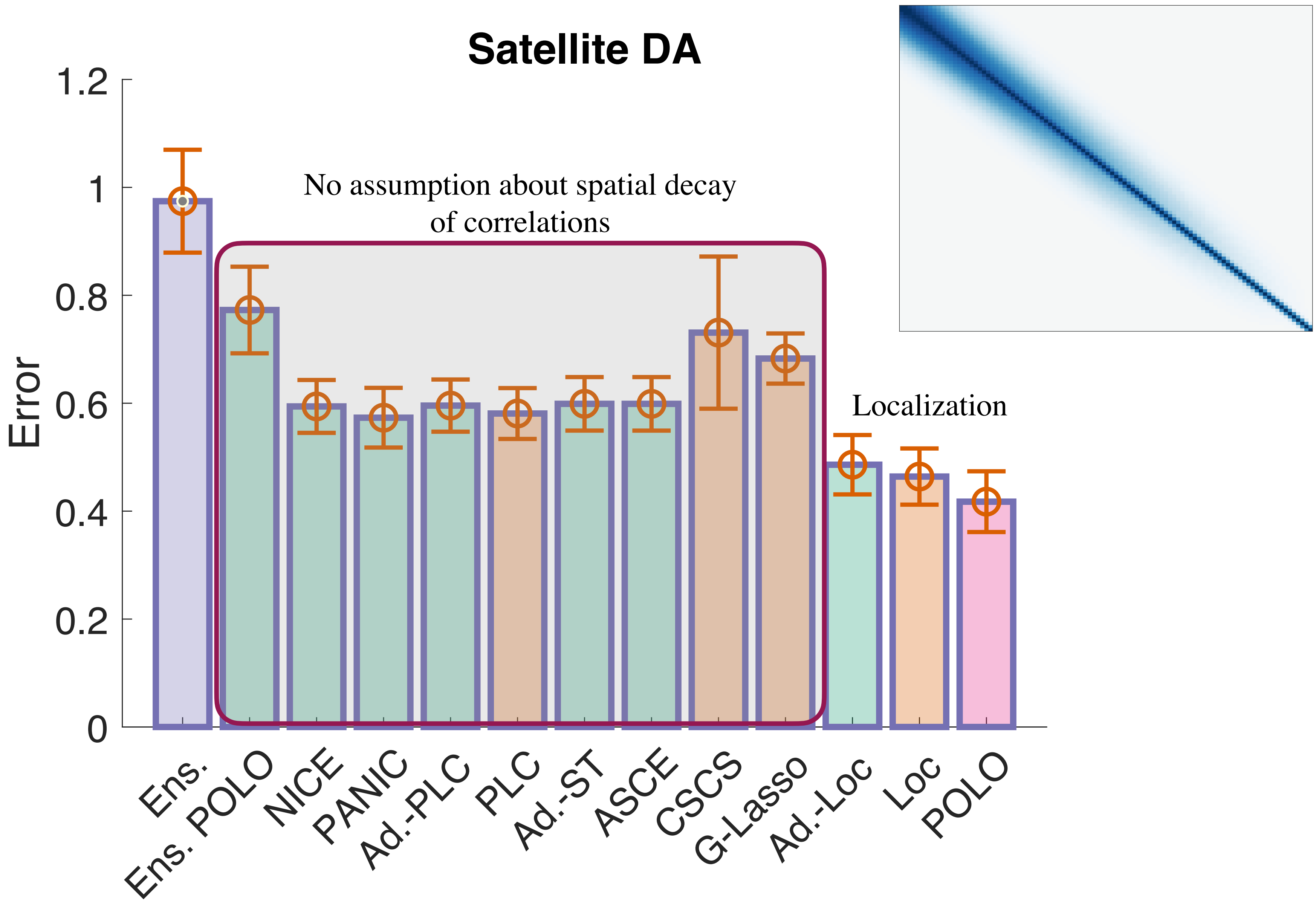
Sanity checks on Gaussians: $n = 100,\ n_e = 20$



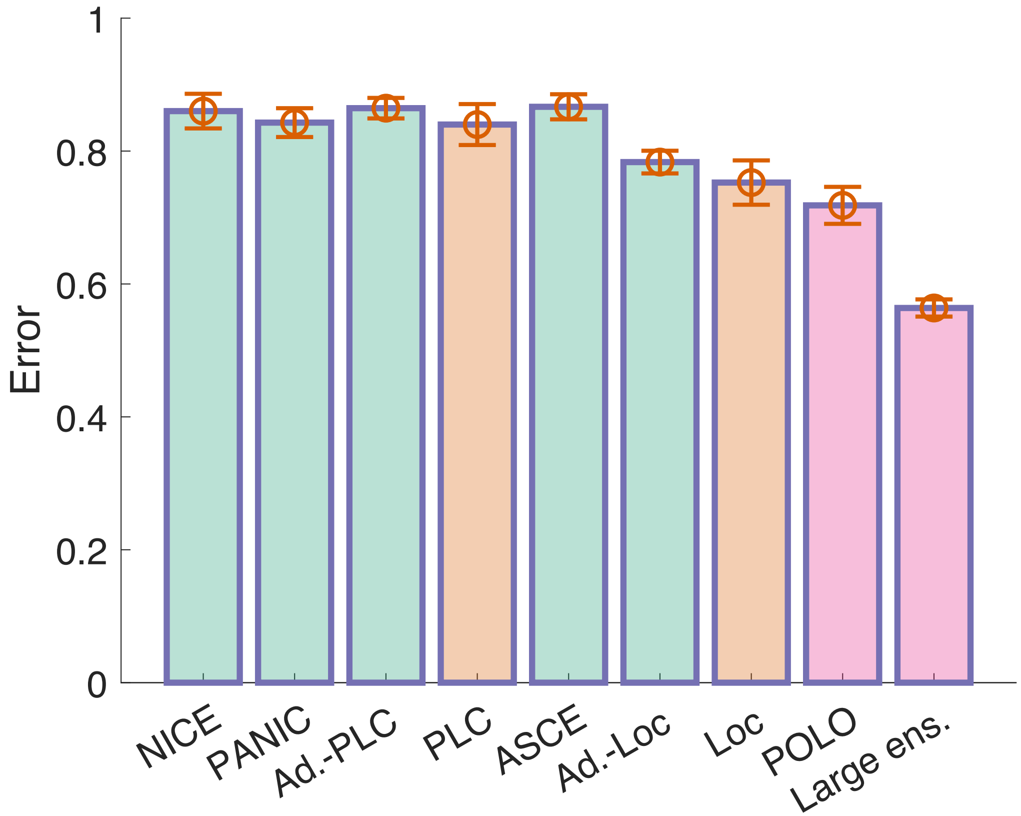
Sanity checks on Gaussians: $n = 1000, \ n_e = 20$



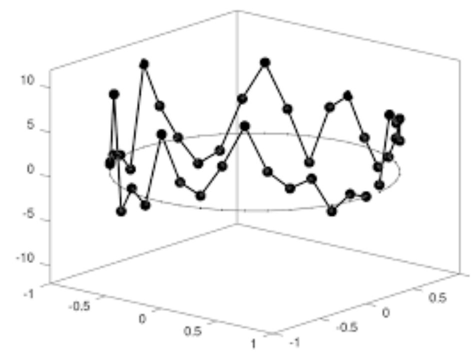
Sanity checks on Gaussians: $n = 100,\ n_e = 20$

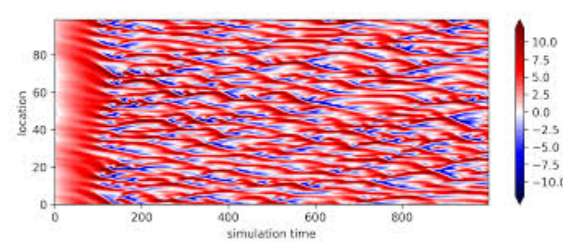


Sanity checks on Lorenz ’95 (*yawn*)

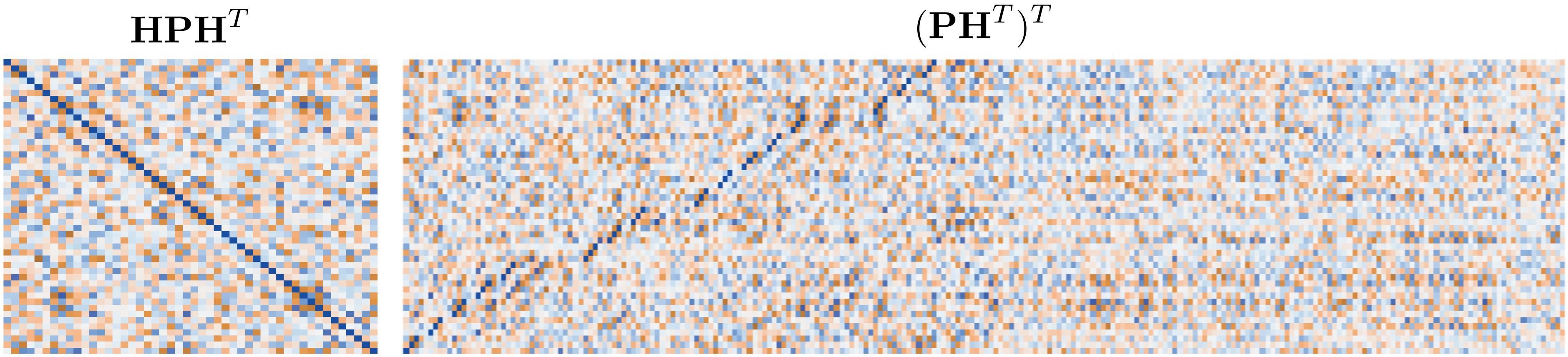


- Typical cycling DA setup from MWR papers
- Reiterates results as with static covariance estimation problems
- Cycling DA does not break our ideas

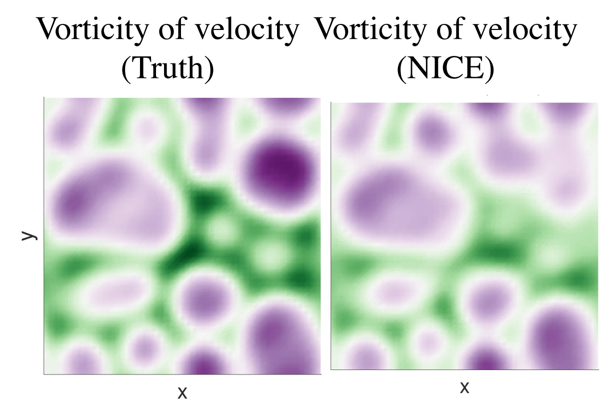




Geomagnetic DA

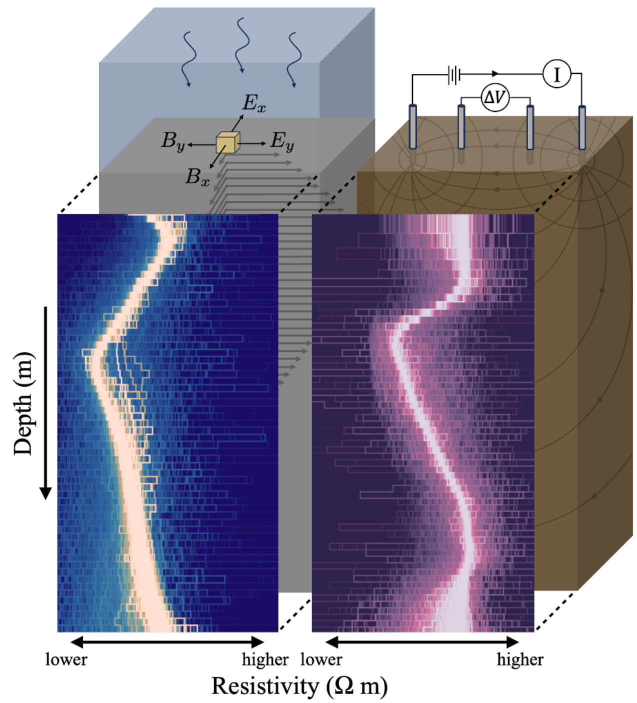


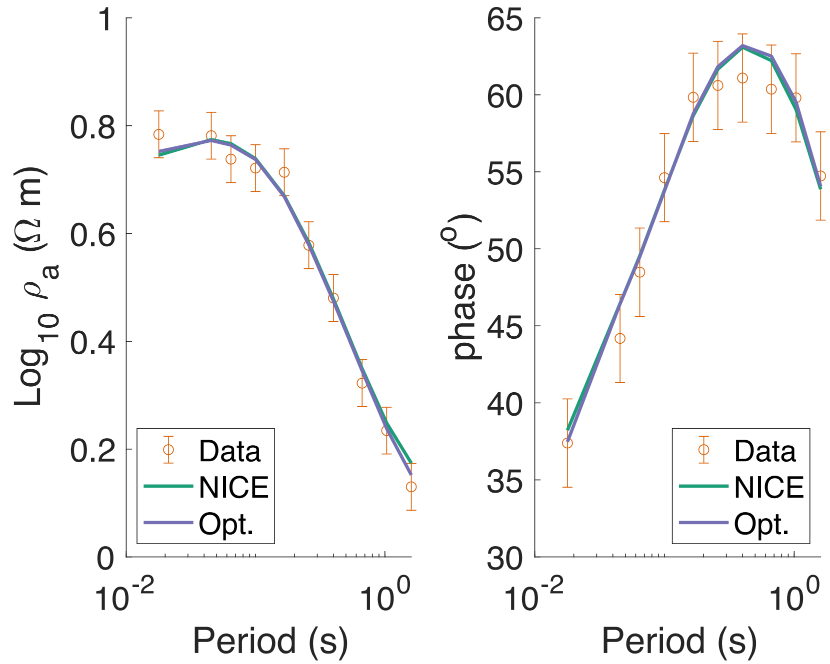
- KS equation coupled to an induction equation
- One quantity (magnetic field) is observed, the other (velocity) unobserved
- Correlation structure is erratic/hard to anticipate
- Localization fails on this example
- NICE outperformed a heavily tuned shrinkage correction
- NICE is not tuned in any way

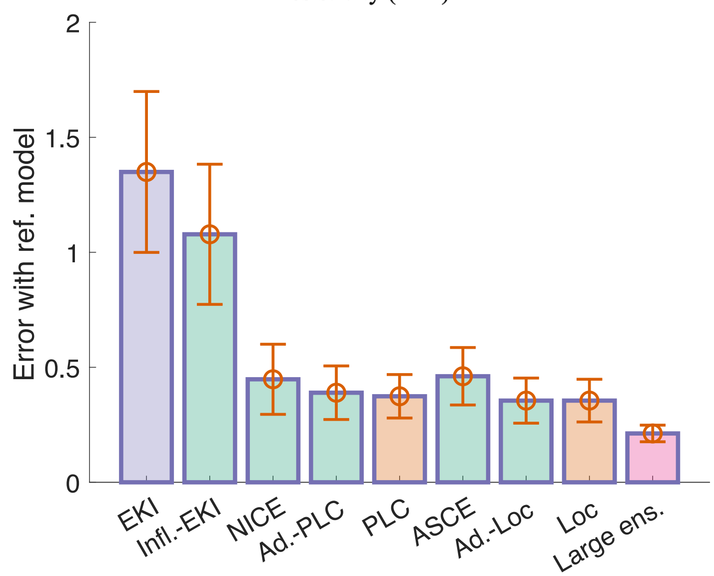


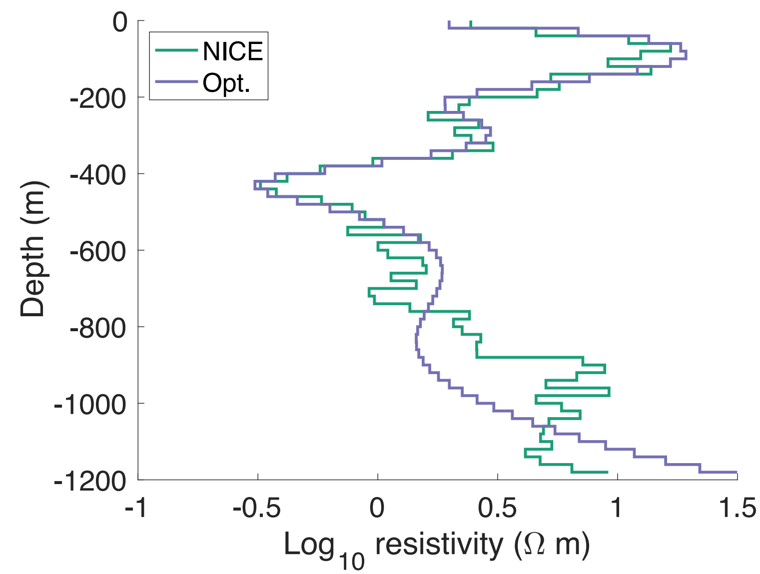
	Error in mag. field	Error in vel. field
Shrinkage (tuned)	1.2	2.0
Ad.-PLC	1.2	2.5
NICE	1.0	1.8
Large ens.	0.7	1.1

Geophysical inversion of marine magentotelluric data

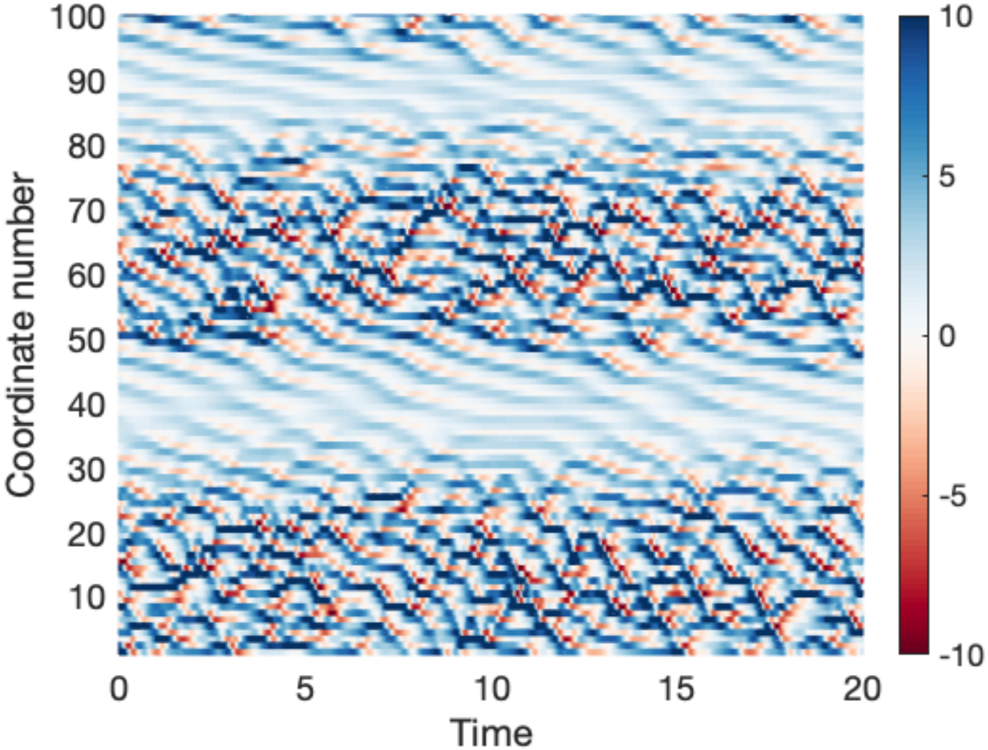


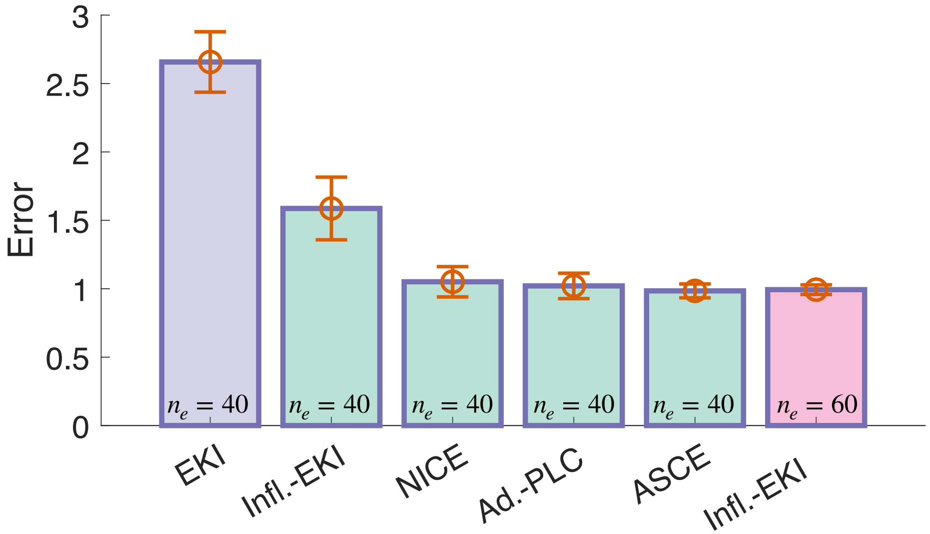


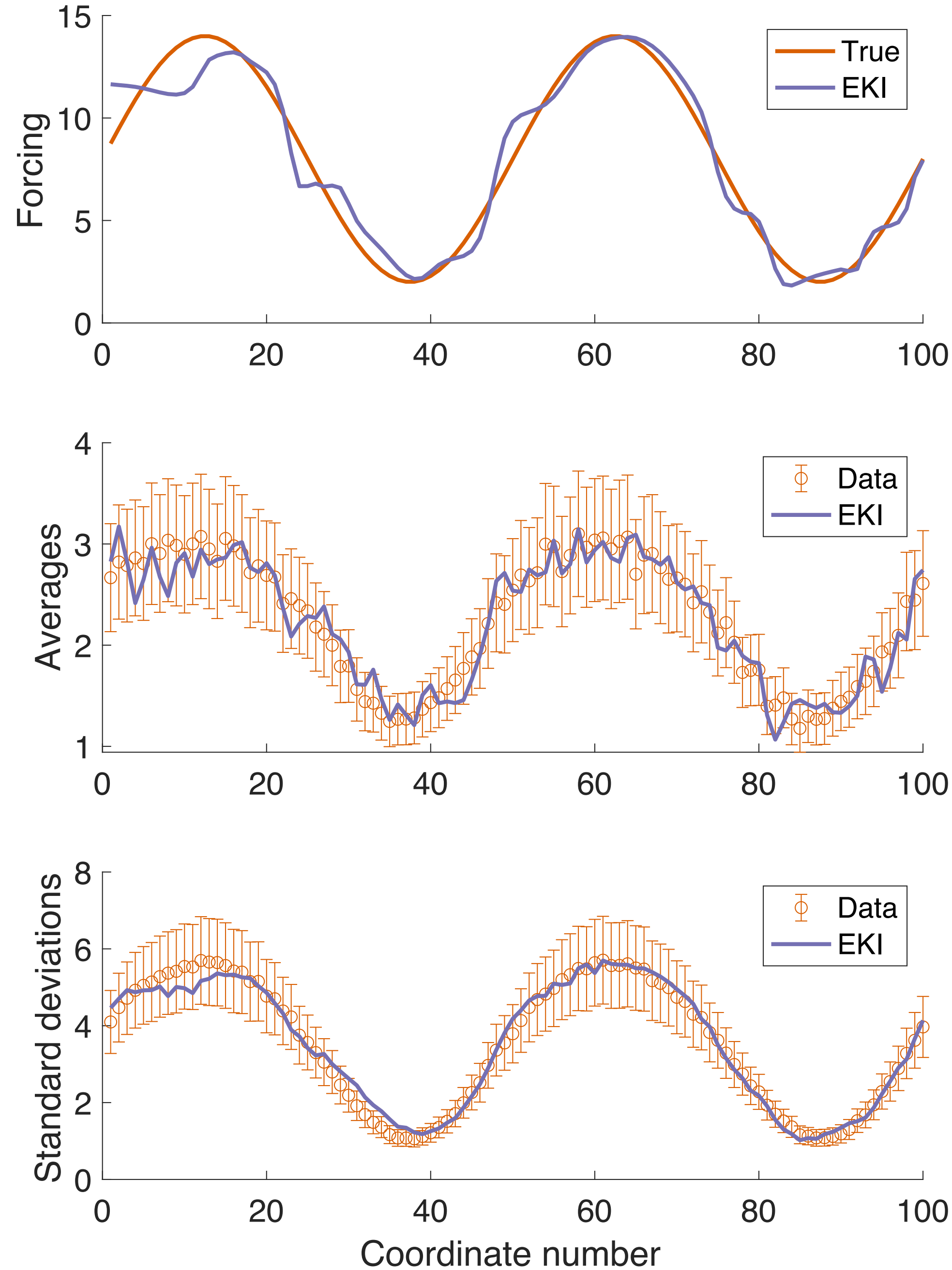




Training a NN with time-averaged data of chaotic dynamics







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

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 Statistician 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

Cross-correlations are hard

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

$$\mathbf{x}_i^a = \mathbf{x}_i + \mathbf{K}\left(\mathbf{y} - (\mathbf{H}\mathbf{x}_i + \boldsymbol{\eta}_i)\right)$$
$$\mathbf{K} = \mathbf{C}_{xy}\left(\mathbf{C}_{yy} + \mathbf{R}\right)^{-1}$$

hard (↑ $\mathbf{C}_{xy}$) **easy** (↑ $\mathbf{C}_{yy}$)

- 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 → 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. Vishny, M. Morzfeld, K. Gwirtz, 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, to appear (2024).