High-dimensional covariance estimation from a small number of samples

Article

Published Version

Creative Commons: Attribution-Noncommercial-No Derivative Works 4.0

Open Access

Vishny, D., Morzfeld, M., Gwirtz, K., Bach, E. ORCID: https://orcid.org/0000-0002-9725-0203, Dunbar, O. R. A. and Hodyss, D. (2024) High-dimensional covariance estimation from a small number of samples. Journal of Advances in Modeling Earth Systems, 16 (9). e2024MS004417. ISSN 1942-2466 doi: https://doi.org/10.1029/2024MS004417 Available at https://centaur.reading.ac.uk/117933/

It is advisable to refer to the publisher's version if you intend to cite from the work. See Guidance on citing.

Published version at: https://doi.org/10.1029/2024MS004417

To link to this article DOI: http://dx.doi.org/10.1029/2024MS004417

Publisher: American Geophysical Union

All outputs in CentAUR are protected by Intellectual Property Rights law, including copyright law. Copyright and IPR is retained by the creators or other copyright holders. Terms and conditions for use of this material are defined in the End User Agreement.

www.reading.ac.uk/centaur

CentAUR

Central Archive at the University of Reading

Reading's research outputs online

JAMES | Journal of Advances in Modeling Earth Systems®

AGU ADVANCING EARTH AND SPACE SCIENCES

RESEARCH ARTICLE
10.1029/2024MS004417

High-Dimensional Covariance Estimation From a Small Number of Samples

David Vishny[1], **Matthias Morzfeld**[1], **Kyle Gwirtz**[2], **Eviatar Bach**[3,4], **Oliver R. A. Dunbar**[3], **and Daniel Hodyss**[5]

[1]Scripps Institution of Oceanography, University of California, San Diego, CA, USA, [2]NASA Goddard Space Flight Center, Greenbelt, MD, USA, [3]California Institute of Technology, Pasadena, CA, USA, [4]University of Reading, Reading, UK, [5]Remote Sensing Division, Naval Research Laboratory, Washington, DC, USA

Key Points:

- We introduce several methods of covariance matrix estimation that adaptively select regularization parameters based on estimates of sampling error
- One method, Noise-Informed Covariance Estimation, stands out because it guarantees a positive semi-definite estimator at a low computational cost
- All new covariance estimation methods perform well on a large variety of test problems

Correspondence to:

M. Morzfeld,
matti@ucsd.edu

Citation:

Vishny, D., Morzfeld, M., Gwirtz, K., Bach, E., Dunbar, O. R. A., & Hodyss, D. (2024). High-dimensional covariance estimation from a small number of samples. *Journal of Advances in Modeling Earth Systems*, *16*, e2024MS004417. https://doi.org/10.1029/2024MS004417

Received 29 APR 2024
Accepted 11 AUG 2024

Author Contributions:

Conceptualization: Matthias Morzfeld, Daniel Hodyss
Formal analysis: David Vishny, Matthias Morzfeld, Kyle Gwirtz, Eviatar Bach, Oliver R. A. Dunbar, Daniel Hodyss
Funding acquisition: Matthias Morzfeld, Daniel Hodyss
Investigation: David Vishny, Matthias Morzfeld, Kyle Gwirtz, Eviatar Bach, Oliver R. A. Dunbar, Daniel Hodyss
Methodology: David Vishny, Matthias Morzfeld, Kyle Gwirtz, Eviatar Bach, Oliver R. A. Dunbar, Daniel Hodyss

© 2024 The Author(s). Journal of Advances in Modeling Earth Systems published by Wiley Periodicals LLC on behalf of American Geophysical Union. This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

Abstract We synthesize knowledge from numerical weather prediction, inverse theory, and statistics to address the problem of estimating a high-dimensional covariance matrix from a small number of samples. This problem is fundamental in statistics, machine learning/artificial intelligence, and in modern Earth science. We create several new adaptive methods for high-dimensional covariance estimation, but one method, which we call Noise-Informed Covariance Estimation (NICE), stands out because it has three important properties: (a) NICE is conceptually simple and computationally efficient; (b) NICE guarantees symmetric positive semi-definite covariance estimates; and (c) NICE is largely tuning-free. We illustrate the use of NICE on a large set of Earth science–inspired numerical examples, including cycling data assimilation, inversion of geophysical field data, and training of feed-forward neural networks with time-averaged data from a chaotic dynamical system. Our theory, heuristics and numerical tests suggest that NICE may indeed be a viable option for high-dimensional covariance estimation in many Earth science problems.

Plain Language Summary Models of physical processes must be fitted to real-world data before they are useful for prediction. In some cases, the most practical way to fit models to data is to run a set—or *ensemble*—of simulations with different physics or initial conditions. One then uses the covariances among the inputs and outputs to modify the simulations so that they fit the data better. To reduce noise in the covariances, one ideally uses an ensemble size that is larger than the number of unknown variables, but this becomes impractical when the number of unknowns is large. To improve the performance of this fitting process when the ensemble size is small, one can discount covariances between variables that are likely due to noise. We introduce several methods of covariance estimation that determine the degree to which covariances are discounted based on expected levels of noise. All new methods perform well on a series of Earth science–inspired problems, but we highlight one method that preserves a key property of covariance matrices at a low computational cost.

1. Introduction

We consider the problem of estimating the covariance matrix $\mathbf{P}$ of an n_x-dimensional random variable $\mathbf{x}$, based on a set of $n_e \ll n_x$ independent samples $\mathbf{x}_i$, $i = 1, \ldots, n_e$. Estimating a covariance matrix from scarce samples is a fundamental challenge in science, engineering, statistics, and in the sub-fields of machine learning and artificial intelligence (Wainwright, 2019). Our interest in covariance estimation is motivated by the problem of fitting models of Earth processes to data. As an example, consider numerical weather prediction (NWP), where the $\mathbf{x}_i$ represent an ensemble of global weather forecasts. The dimension n_x corresponds to the number of unknowns in a global atmospheric model, and it is on the order of 10^8. The number of forecasts (the ensemble size n_e) is small because each forecast requires a simulation of Earth's atmosphere, which is expensive. A commonly used ensemble size in NWP is on the order of 10^2—six orders of magnitude smaller than the number of unknowns. A common approach to update the forecast with atmospheric data is the ensemble Kalman filter (EnKF, Evensen, 1994, 2009). The EnKF updates rely on the covariance matrix associated with the ensemble, but the *empirical covariance matrix*

$$\hat{\mathbf{P}} = \frac{1}{n_e - 1}\sum_{i=1}^{n_e}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^T, \quad \hat{\boldsymbol{\mu}} = \frac{1}{n_e}\sum_{i=1}^{n_e}\mathbf{x}_i, \tag{1}$$

Project administration: Matthias Morzfeld, Daniel Hodyss
Resources: Matthias Morzfeld
Software: David Vishny, Matthias Morzfeld, Kyle Gwirtz, Eviatar Bach, Oliver R. A. Dunbar
Supervision: Matthias Morzfeld
Validation: David Vishny, Matthias Morzfeld, Kyle Gwirtz, Eviatar Bach, Oliver R. A. Dunbar
Visualization: David Vishny, Matthias Morzfeld
Writing – original draft: David Vishny, Matthias Morzfeld
Writing – review & editing: David Vishny, Matthias Morzfeld, Kyle Gwirtz, Eviatar Bach, Oliver R. A. Dunbar, Daniel Hodyss

is generally inaccurate if $n_e \ll n_x$ (Bickel & Levina, 2008; Wainwright, 2019). Various strategies for improving the accuracy of the empirical estimate have been developed over the years, and we review the relevant literature below. The prevailing method of covariance estimation in NWP is called *localization* (Houtekamer & Mitchell, 1998, 2001; Ott et al., 2004). Localization enforces on the empirical covariance matrix the assumption that covariances decay with spatial distance (although the terminology has also been used in other contexts and to refer to covariance corrections that are not spatial, see, e.g., Morzfeld et al., 2019). To execute the localization, one defines an $n_x \times n_x$, symmetric PSD *localization matrix* $\mathbf{L}$, which encodes the spatial decay pattern of correlations (Gaspari & Cohn, 1999; Gilpin et al., 2023). One then obtains the localized covariance estimator

$$\hat{\mathbf{P}}_{\text{loc}} = \mathbf{L} \circ \hat{\mathbf{P}}, \tag{2}$$

where the open circle denotes the Hadamard (element-wise) product. Localization has proven successful for estimating high-dimensional covariance matrices from a small set of samples in NWP and, for that reason, localization is a standard component in operational weather forecasting systems (Bannister, 2017; Hamill et al., 2009). We present a new covariance estimation method that is more broadly applicable than classical localization because it does not require a priori assumptions about the correlation structure (e.g., the spatial decay in covariance localization). We call our method **N**oise-**I**nformed **C**ovariance **E**stimation (NICE). NICE replaces assumptions about the correlation structure with the assumption that *small to medium correlations are likely caused by sampling error* and, therefore, should be damped or deleted. This assumption is not universally true (it is easy to come up with counter examples), but it is rooted in rigorous sampling error theory (Anderson, 2012; Flowerdew, 2015; Lee, 2021b; Ménétrier et al., 2015; Morzfeld & Hodyss, 2023). NICE achieves three main objectives:

1. *Adaptivity.* NICE ensures that differences between sampled and corrected correlations are within an expected noise level. The noise level is determined by the sample size and the distribution of empirical correlations so that the entire covariance estimation process is adaptive and largely tuning-free.
2. *Positive semi-definiteness.* NICE guarantees a symmetric PSD covariance estimator. Symmetry and positive semi-definiteness are defining properties of covariance matrices, but some competing methods do not guarantee PSD estimates.
3. *Computational efficiency.* NICE is computationally efficient and easy to implement because it avoids solving optimization problems over PSD matrices.

We put NICE to the test in a variety of problems with different and unknown correlation structures: (a) estimation of covariance matrices from Gaussian samples; (b) cycling data assimilation (DA) problems with ensemble Kalman filters (Evensen, 2009); (c) inversion of geophysical data with regularized ensemble Kalman inversion (EKI, Chada et al., 2020); and (d) training of a feed-forward neural network with EKI (Cleary et al., 2021; Iglesias et al., 2013; Kovachki & Stuart, 2019). Various error metrics are used to evaluate performance in these problems. Across all examples and all error metrics, we find that NICE works out-of-the-box with minimal tuning. Estimated noise levels can also be used to make other covariance estimation methods adaptive and largely tuning-free. We introduce *new* adaptive versions of Power law corrections (PLC) (Ad.-PLC, see Lee (2021b) and Section 3.4.1), adaptive (spatial) localization (Ad.-Loc., Section 3.4.2), adaptive soft-thresholding (Ad.-ST, see Wainwright (2019) and Section 3.4.3) and adaptive sparse covariance estimation (ASCE, see Xue et al. (2012) and Section 3.4.4). All new methods fall under the umbrella of NICE because all of them leverage an understanding of noise in empirical correlations. However, some do not guarantee a PSD estimator and others are more computationally involved. The specific method we refer to as NICE is the *only* method that satisfies all three of our objectives: adaptivity, PSD guarantees and computational efficiency.

It is important to be specific about the terms "high-dimensional" and about computational efficiency. In this paper, we focus on covariance estimation methods that construct the entire covariance matrix. As such, the methods are limited in their use to matrices of dimension $10^4 \times 10^4$ or smaller. Higher-dimensional problems, for example, of the extreme size of NWP (10^8 or more unknowns), require that we perform computations without constructing the whole covariance matrix. The methods we describe here could potentially be adapted to such problems, but these adaptations are beyond the scope of this paper. The computational efficiency of covariance estimation depends on the algorithms used. We focus on algorithms that perform simple element-wise operations

19422466, 2024, 9, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024MS004417 by Test, Wiley Online Library on [30/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

19422466, 2024, 9, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024MS004417 by Test, Wiley Online Library on [30/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

on the empirical covariance matrix. Many methods in the statistical literature, however, perform covariance estimation by solving optimization problems over PSD matrices, which is computationally expensive.

The rest of this paper is organized as follows. Section 2 reviews background materials. We first explain why covariance estimation from a small number of samples is important in Earth science, specifically in EnKF and in EKI. We further emphasize the importance of PSD covariance estimates in the context of EnKF or EKI. We then review covariance localization in NWP and several covariance estimation methods from the statistical literature. Finally, we briefly describe Morozov's discrepancy principle, a classical concept in inverse theory. The discrepancy principle is the tool we use to make covariance estimation methods adaptive. Section 3 describes our new methodology (NICE), and other new adaptive covariance estimation methods. We apply NICE and a large number of competing methods (new and old) in a wide variety of problems in Section 4, before ending the paper with a summary and conclusions in Section 5.

2. Background

2.1. Ensemble Kalman Filters and Their Localization

The goal of ensemble Kalman filtering (EnKF) is to use data to update a forecast generated by a computational model. An important example is NWP, where the forecast describes atmospheric states in the form of n_e vectors $\mathbf{x}_i$, $i = 1, \ldots, n_e$, each of dimension n_x. The vectors $\mathbf{x}_i$ are referred to as "ensemble members." Typically, the ensemble size n_e is smaller than the dimension of the ensemble members ($n_e \ll n_x$). The reason is that each ensemble member is the result of a simulation with a computationally expensive atmospheric model, so that n_e must be small, or else the computations are infeasible. In NWP, n_e is usually a few hundred, and n_x is in the billions.

The forecast is updated by an observation (data), which is an n_y-dimensional vector $\mathbf{y}$, where, often but not always, $n_e \ll n_y \ll n_x$. For ease of presentation, we assume that the observation is a linear function of the forecasted variables so that

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \boldsymbol{\varepsilon}, \tag{3}$$

where $\mathbf{H}$ is an $n_y \times n_x$ matrix and $\boldsymbol{\varepsilon}$ is a Gaussian random variable with mean zero and covariance matrix $\mathbf{R}$, which we write as $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$ The assumption of a linear observation is commonly violated; nevertheless, we demonstrate in our numerical experiments that the intuition and conclusions from the linear analysis extend to the nonlinear case.

Ensemble Kalman filtering (EnKF) is a catch-all term for a whole suite of methods that merge the observation and the forecast within a Bayesian framework. The update step of a stochastic EnKF (Burgers et al., 1998; Evensen, 1994, 2009) is

$$\mathbf{x}_i^a = \mathbf{x}_i + \hat{\mathbf{K}}(\mathbf{y} - (\mathbf{H}\mathbf{x}_i + \boldsymbol{\varepsilon}_i)), \tag{4}$$

where $\boldsymbol{\varepsilon}_i$ is a sample drawn from $\mathcal{N}(\mathbf{0}, \mathbf{R})$. The Kalman gain $\hat{\mathbf{K}}$ is computed from the ensemble as

$$\hat{\mathbf{K}} = \hat{\mathbf{P}}\mathbf{H}^T\left(\mathbf{H}\hat{\mathbf{P}}\mathbf{H}^T + \mathbf{R}\right)^{-1}, \tag{5}$$

where $\hat{\mathbf{P}}$ is the empirical covariance in (1). The Kalman gain defines how to update each ensemble member in view of the observation. Since the Kalman gain depends critically on the forecast covariance $\hat{\mathbf{P}}$, the EnKF update is only useful if the covariance estimate is accurate, which usually requires that n_e is larger than n_x (although the situation can be more complex, e.g., with n_e directly depending on the number of observations and their independence (Agapiou et al., 2017; Al Ghattas & Sanz-Alonso, 2022; Chorin & Morzfeld, 2013; Hodyss & Morzfeld, 2023)).

Localization is a technique that enables the use of EnKF when $n_e \ll n_x$. A common version of localization in the EnKF is to use Hadamard products as in Equation 2 and to define the localization matrix by the Gaspari–Cohn covariance function (Gaspari & Cohn, 1999) or its anisotropic extensions (Gilpin et al., 2023). The localization matrix implements a spatial decay of correlation and the rate of decay can be controlled via a length scale. Different methods for adaptively selecting this length scale, or localizing in a flow-dependent manner to account for temporal variations in the correlation structure, have been proposed (Anderson, 2012; Bishop & Hodyss, 2007, 2009a, 2009b, 2011; Chevrotiére & Harlim, 2017; Luk et al., 2024; Zhen & Zhang, 2014).

Other implementations of the EnKF include the ensemble adjustment Kalman filter (EAKF) (Anderson, 2001) and ensemble transform filters (ETKF) (Ott et al., 2004; Tippett et al., 2003). Localization in an EAKF is achieved by working directly with the Kalman gain, reducing the effects of an observation on elements of the Kalman gain that are far from the observation (Hodyss & Morzfeld, 2023; Morzfeld & Hodyss, 2023). Localization in an ETKF is implemented by performing a "local" analysis, so that each grid point is updated by a set of nearby observations (domain localization). Variational/hybrid DA algorithms combine a classical minimization (variational) approach (Talagrand & Courtier, 1987) with an ensemble to approximate uncertainties (Buehner et al., 2013; Hamill & Snyder, 2000; Kuhl et al., 2013; Lorenc, 2003; Poterjoy & Zhang, 2015; Zhang et al., 2009). Hybrid DA also requires localization, which is usually applied using Hadamard products, but without explicitly forming the covariance matrix (Buehner, 2005). Multi-scale extensions of localization are available for hybrid DA and/or EnKFs (Buehner, 2012; Buehner & Shlyaeva, 2015; Harty et al., 2021; Lorenc, 2017; Miyoshi & Kondo, 2013).

Finally, we note that all conventional localization methods require tuning. The tuning process usually amounts to picking a length scale that defines the localization and then running a cycling EnKF over a set of training observations. This process is repeated with various length scales until one encounters a length scale that leads to an acceptable error metric.

2.2. Ensemble Kalman Inversion

The goal in EKI (Iglesias et al., 2013) is to minimize the cost function

$$J(\mathbf{x}) = \left\| \mathbf{R}^{-1/2}(\mathbf{y} - \mathcal{G}(\mathbf{x})) \right\|_2^2, \tag{6}$$

where vertical bars denote the two-norm (i.e., $||\mathbf{b}||_2 = \sqrt{\mathbf{b}^T\mathbf{b}}$), $\mathbf{y}$ are data, $\mathbf{x}$ are unknown model parameters, and $\mathcal{G}(\cdot)$ is a nonlinear model that maps the model parameters to the data; the symmetric positive definite matrix $\mathbf{R}$ defines expected errors in the data, represented by a mean-zero Gaussian random variable with covariance matrix $\mathbf{R}$; $\mathbf{R}^{-1/2}$ is the inverse of a matrix square root of $\mathbf{R} = \mathbf{R}^{1/2}(\mathbf{R}^{1/2})^T$.

EKI performs the optimization by iteratively updating an ensemble as follows. The ensemble at iteration k are the n_e vectors $\mathbf{x}_i^k$ and we define n_e corresponding vectors $\mathbf{g}_i^k = \mathcal{G}(\mathbf{x}_i^k)$. Each ensemble member is updated according to

$$\mathbf{x}_i^{k+1} = \mathbf{x}_i^k + \hat{\mathbf{C}}_{xg}^k(\hat{\mathbf{C}}_{gg}^k + \mathbf{R})^{-1}(\mathbf{y} - (\mathbf{g}_i^k + \boldsymbol{\eta}_i)), \tag{7}$$

where $\hat{\mathbf{C}}_{gg}^k$ is the covariance matrix associated with the vectors $\mathbf{g}_i^k$, $\hat{\mathbf{C}}_{xg}^k$ is the covariance between the vectors $\mathbf{x}_i^k$ and $\mathbf{g}_i^k$ and where $\boldsymbol{\eta}_i$ is a draw from the Gaussian with mean zero and covariance matrix $\mathbf{R}$. More specifically, if we define the matrices (ensemble perturbations).

$$\mathbf{X}^k = \frac{1}{\sqrt{n_e - 1}}\begin{pmatrix}\mathbf{x}_1^n - \bar{\mathbf{x}}^k & \mathbf{x}_2^k - \bar{\mathbf{x}}^k & \cdots & \mathbf{x}_{n_e}^k - \bar{\mathbf{x}}^k\end{pmatrix}, \quad \bar{\mathbf{x}}^k = \frac{1}{n_e}\sum_{j=1}^{n_e}\mathbf{x}_j^k, \tag{8}$$

$$\mathbf{G}^k = \frac{1}{\sqrt{n_e - 1}}\begin{pmatrix}\mathbf{g}_1^n - \bar{\mathbf{g}}^k & \mathbf{g}_2^k - \bar{\mathbf{g}}^k & \cdots & \mathbf{g}_{n_e}^k - \bar{\mathbf{g}}^k\end{pmatrix}, \quad \bar{\mathbf{g}}^k = \frac{1}{n_e}\sum_{j=1}^{n_e}\mathbf{g}_j^k, \tag{9}$$

then the covariances are

$$\hat{\mathbf{C}}_{xg}^k = \mathbf{X}^k \otimes \mathbf{G}^k, \tag{10}$$

$$\hat{\mathbf{C}}_{gg}^k = \mathbf{G}^k \otimes \mathbf{G}^k, \tag{11}$$

19422466, 2024, 9, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024MS004417 by Test, Wiley Online Library on [30/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

where the symbol $\otimes$ denotes the outer product $\mathbf{A} \otimes \mathbf{B} = \mathbf{A}\mathbf{B}^T$, where $\mathbf{A}$ and $\mathbf{B}$ are vectors or matrices of compatible sizes. Note that the EKI update Equation 7 is equivalent to an EnKF update in (4) because $\hat{\mathbf{C}}^k_{xg} = \hat{\mathbf{P}}\mathbf{H}^T$ and $\hat{\mathbf{C}}^k_{gg} = \mathbf{H}\hat{\mathbf{P}}\mathbf{H}^T$ when $\mathcal{G}(\mathbf{x}) = \mathbf{H}\mathbf{x}$ is linear. The theory around EKI tells us that the iteration (7) converges, in the sense that the ensemble collapses onto the minimizer of the cost function, under typical assumptions (Chada & Tong, 2022; Schillings & Stuart, 2017, 2018). As with EnKF, there are several variants of EKI (Huang et al., 2022; Lee, 2021a).

Convergence of the EKI iteration requires that the covariance estimates $\hat{\mathbf{C}}^k_{xg}$ and $\hat{\mathbf{C}}^k_{gg}$ are sufficiently accurate, which usually means that the ensemble size is large. Localization can be used within an EKI to keep the ensemble size small (Al Ghattas & Sanz-Alonso, 2022; Lee, 2021b; Tong & Morzfeld, 2023).

EKI has found application in climate sciences (Bieli et al., 2022; Cleary et al., 2021; Dunbar et al., 2022; Schneider et al., 2021), and Julia code for it is available (Dunbar et al., 2022). In a climate science context, model parameters appear in sub-gridscale closures of climate models (e.g., physical constants or weights of neural networks (NN)). A promising approach to optimizing sub-gridscale closures is to formulate the cost function based on the misfit between modeled and observed climate statistics (Schneider et al., 2024). In this scenario, derivatives of the cost function with respect to the model parameters are difficult or impossible to compute, making derivative-free optimization via EKI attractive.

Ensemble algorithms that are related to EKI, and which in fact pre-date EKI, are known as iterative ensemble Kalman filters/smoothers (Bocquet, 2016; Bocquet & Sakov, 2014; Chen & Oliver, 2010, 2013, 2017; Emerick & Reynolds, 2011; Hodyss, Bishop, & Morzfeld, 2016; Luo et al., 2018) or multiple DA (Emerick & Reynolds, 2013). These methods are popular in reservoir modeling, but also find applications in atmospheric sciences. A recent, mathematical overview of how some of the methods are related is given by Chada et al. (2021) and an NWP-focused overview is provided by Hodyss, Bishop, and Morzfeld (2016).

2.3. Positive Semi-Definite Covariance Estimators

A fundamental property of covariance matrices is that they are symmetric PSD (Horn & Johnson, 1991). The practical relevance of PSD estimates of $\hat{\mathbf{P}}$ is apparent in the EnKF, where the Kalman gain (Equation 5) requires that the matrix

$$\mathbf{H}\hat{\mathbf{P}}\mathbf{H}^T + \mathbf{R}, \tag{12}$$

is well-conditioned. Since the observation error covariance matrix $\mathbf{R}$ is usually positive definite, a PSD estimate $\hat{\mathbf{P}}$ guarantees that Equation 12 is positive definite, invertible and well-conditioned. One can run into numerical trouble if $\hat{\mathbf{P}}$ is not PSD, because the matrix in Equation 12 may be singular or ill-conditioned. Localization via Hadamard products, as used in NWP, guarantees a PSD covariance estimate by the Schur product theorem when the localization matrix $\mathbf{L}$ is PSD (Schur, 1911).

In general, however, the PSD constraint is not easy to satisfy during covariance estimation, and many covariance estimation methods do not guarantee a PSD estimate (Khare et al., 2019; Xue et al., 2012). When we review covariance estimation methods, we comment on their PSD guarantees.

2.4. Beyond Localization

It has long been recognized that the assumption of a spatial decay of correlation, which is at the core of localization, is not universally applicable. Adaptive localization methods (Anderson, 2012; Bishop & Hodyss, 2007, 2009a, 2009b; Lee, 2021b) are well established in Earth science, and recent theoretical works (Flowerdew, 2015; Ménétrier et al., 2015; Morzfeld & Hodyss, 2023) address this issue as well.

Covariance estimation is also a fundamental problem in statistics. Theoretical aspects of localization in NWP, for example, are described by Furrer and Bengtsson (2007) and Bickel and Levina (2008), and a review of various covariance estimation methods is provided by Pourahmadi (2011). The textbook by Wainwright (2019) emphasizes the difficulty of estimating a covariance matrix when the ensemble size is small. As representatives of the many statistical techniques that have been created over the years, we consider a soft-thresholding method

19422466, 2024, 9, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024MS004417 by Test, Wiley Online Library on [30/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

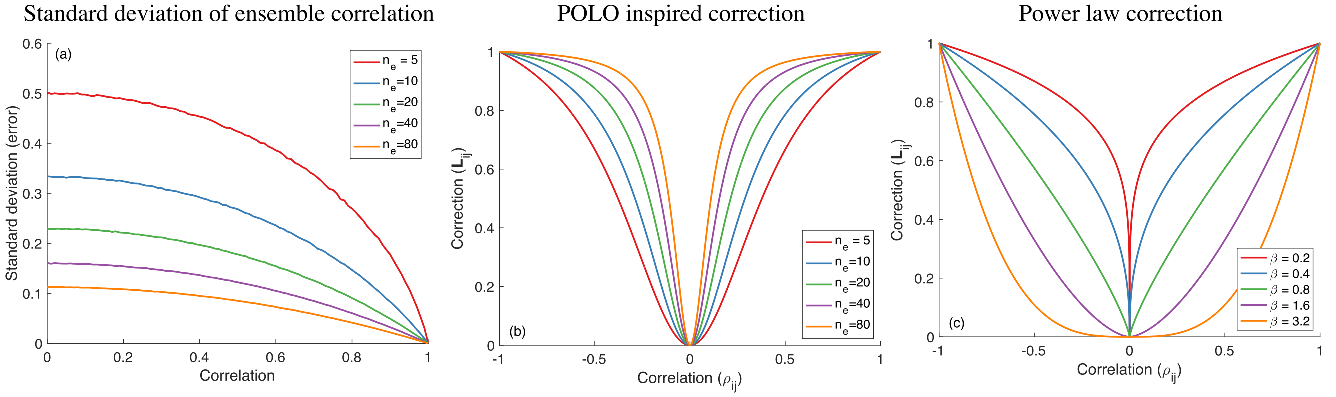

Figure 1. (a) Standard deviation of ensemble correlation as a function of correlation (adapted from Figure 1 of Lee (2021b)). (b) POLO inspired correction factor, shown as a function of correlation for different ensemble sizes. (c) Power law correction factor, as proposed by Lee (2021b), shown as a function of correlation for different choices of the exponent β.

(Wainwright, 2019), the graphical Lasso (G-Lasso, Friedman et al., 2007), convex sparse Cholesky selection (CSCS, Khare et al., 2019), and sparse covariance estimation (Xue et al., 2012).

2.4.1. Prior Optimal Localization

The idea of optimal localization is to find a Hadamard product estimator, defined by the matrix **L**, that minimizes the cost function

$$F_{\text{POLO}}(\mathbf{L}) = \left\| \langle \mathbf{L} \circ \hat{\mathbf{P}} - \mathbf{P}_{n_e\to\infty} \rangle \right\|_{\text{Fro}}, \tag{13}$$

where $\mathbf{P}_{n_e\to\infty}$ is the "true" covariance matrix one would obtain from an infinite ensemble and where the brackets $\langle\cdot\rangle$ denote an expected value over ensemble draws (Flowerdew, 2015; Ménétrier et al., 2015; Morzfeld & Hodyss, 2023); $\|\cdot\|_{\text{Fro}}$ is the Frobenius norm, that is, the square root of the sum of the squares of all elements of a matrix. Under Gaussian assumptions, one can solve this optimization analytically to obtain

$$[\mathbf{L}]_{ij} = \frac{\rho_{ij}^2(n_e - 1)}{1 + \rho_{ij}^2 n_e}, \tag{14}$$

which we refer to as prior optimal localization (POLO). Here, ρ_{ij} is the true correlation between the variables with indices i and j. While POLO does *not* rely on a spatial decay of correlations, it assumes that the correlations are known. POLO is, therefore, not a viable algorithm but it can be used as a benchmark for practical algorithms. Empirical localization functions are closely related to optimal localization and are implemented based on the idea of learning a localization matrix from training/simulation data (Anderson & Lei, 2013).

POLO does *not* guarantee a PSD estimator. To see why, consider a theorem in linear algebra: If one applies a function element-wise to a PSD matrix whose elements are in (0, 1), the only functions that always preserve semi-definiteness have a power series representation with non-negative coefficients (Guillot & Rajaratnam, 2015; Schoenberg, 1942). The POLO matrix in Equation 14 does not satisfy this theorem and, hence, the matrix **L** is not guaranteed to be PSD, which in turn implies that the POLO covariance estimator is not guaranteed to be PSD. Indeed, we routinely observe non-PSD POLO estimates in the numerical examples in Section 4.

2.4.2. Sampling Error Corrections and Power Law Corrections

Anderson (2012) introduces the terminology and methodology of *sampling error correction* (SEC). SEC constructs covariance corrections quite similarly to POLO, but the SEC corrections are based on numerical experiments with "training data" and groups of ensembles, so that the correction depends on the sample correlation, rather than on the true correlation (compare Figure 1b of this paper with Figure 1 of Anderson (2012)).

19422466, 2024, 9, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024MS004417 by Test, Wiley Online Library on [30/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Lee (2021b) subsequently noticed that the corrections may be efficiently approximated by a power law. Specifically, let $\hat{\boldsymbol{\rho}}$ be the empirical estimate of the ensemble *correlations* and define the PLC estimator of the correlations by

$$\hat{\boldsymbol{\rho}}_{\mathrm{PLC}} = \mathbf{L}(\beta) \circ \hat{\boldsymbol{\rho}}, \tag{15}$$

where the elements of the matrix $\mathbf{L}(\beta)$ are given by

$$[\mathbf{L}(\beta)]_{ij} = |[\hat{\boldsymbol{\rho}}]_{ij}|^{\beta}, \tag{16}$$

that is, we raise the absolute values of the empirical correlations *element-wise* to the power β. The exponent β is a tunable parameter. Once we have selected a suitable β, we obtain the covariance estimator

$$\hat{\mathbf{P}}_{\mathrm{PLC}} = \hat{\mathbf{V}}\hat{\boldsymbol{\rho}}_{\mathrm{PLC}}\hat{\mathbf{V}}, \tag{17}$$

where $\hat{\mathbf{V}}$ is a $n \times n$ diagonal matrix whose diagonal elements are the ensemble standard deviations. For the rest of this paper, we refer to this algorithm as "power law corrections" (PLC).

PLC does not guarantee a PSD covariance estimator: one can apply the same theorems and logic as outlined above when discussing the PSD property in the context of POLO. The PLC correlation estimate, however, *is* PSD if the exponent is "large enough." To understand why, we derive lower bounds for the eigenvalues of $\mathbf{L}(\beta)$ using Gershgorin's circle theorem. The theorem implies that an eigenvalue, λ, of $\mathbf{L}(\beta)$ satisfies the inequalities

$$1 - Z_i \leq \lambda \leq 1 + Z_i, \tag{18}$$

where Z_i is the sum of the absolute values of the off-diagonal elements of a row (or column) of $\mathbf{L}(\beta)$:

$$Z_i = \sum_{i \neq j} |\hat{\rho}_{ij}|^{\beta}. \tag{19}$$

If we pick the exponent β to guarantee that $Z_i \leq 1$ for all i (all rows of $\mathbf{L}(\beta)$), then Gershgorin's theorem implies positive semi-definiteness of the matrix $\mathbf{L}(\beta)$ and, via the Schur product theorem, positive semi-definiteness of the PLC estimator. In our examples, and with our adaptive strategy for choosing the exponent β (see Section 3.4.1), we never ran into trouble with definiteness of the estimators, but we cannot guarantee that this is generally the case.

Covariance estimation using powers of ensemble correlations is also at the core of a method called ECO-RAP (ensemble correlations raised to a power, Bishop & Hodyss 2009a, 2009b, 2007). In ECO-RAP, only positive, even exponents are considered, which ensures that the ECO-RAP estimator is PSD, and that ECO-RAP, embedded within an ensemble transform approach, is scalable to high-dimensional problems.

2.4.3. Soft-Thresholding

The idea of thresholding is to set small covariances to zero. This can be achieved by applying the soft-thresholding function

$$T_{\lambda}(s) = \begin{cases} s - \lambda\ \mathrm{sign}(s) & \text{if } |s| > \lambda \\ 0 & \text{otherwise} \end{cases}, \tag{20}$$

element-wise to the sampling covariance matrix, so that the soft-thresholding covariance estimate is

$$[\hat{\mathbf{P}}_{\mathrm{ST}}]_{ij} = T_{\lambda}\left([\hat{\mathbf{P}}]_{ij}\right). \tag{21}$$

19422466, 2024, 9, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024MS004417 by Test, Wiley Online Library on [30/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

19422466, 2024, 9, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024MS004417 by Test, Wiley Online Library on [30/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Here, λ is a positive scalar. Soft-thresholding has favorable asymptotic properties (Wainwright, 2019) and is computationally simple to implement, but the soft-thresholding covariance estimator is not always PSD (Khare et al., 2019). The parameter λ is usually determined via a tuning process. In Section 3.4.3, we describe how to find this parameter adaptively.

2.4.4. Sparse Covariance Estimation

Xue et al. (2012) note that soft-thresholding corresponds to the minimizer of the cost function

$$\mathcal{F}_{\text{Soft Thres.}}(\mathbf{P}) = \frac{1}{2}||\mathbf{P} - \hat{\mathbf{P}}||^2_{\text{Fro}} + \lambda \sum_{j \neq k} |\mathbf{P}_{jk}|, \tag{22}$$

where $\hat{\mathbf{P}}$ is the empirical covariance matrix. The authors then describe an algorithm to minimize the cost function (22) subject to the constraint that $\mathbf{P} \geq \epsilon \mathbf{I}$ (i.e., the matrix $\mathbf{P} - \epsilon \mathbf{I}$ is PSD), where $\mathbf{I}$ is the identity matrix and where $\epsilon > 0$ is a nuisance parameter that can be set to a small number (10^{-5} is suggested). The constraint guarantees that the covariance estimator is symmetric positive definite. Moreover, the estimator is sparse because large off-diagonal elements are penalized and the 1-norm drives small covariances to zero. This means that this technique, which we call *sparse covariance estimation*, is most applicable in situations where one expects that most covariances should be zero. We note that sparse covariance estimation requires tuning to find an appropriate regularization strength λ. In Section 3.4.4, we explain how to find the regularization strength adaptively.

2.4.5. Graphical Lasso

Soft-thresholding and sparse covariance estimation find sparse estimates of the covariance matrix, that is, the underlying assumption is that the majority of the covariances are equal to zero. One can also search for a covariance matrix whose *inverse* is sparse. The inverse of the covariance matrix is called the precision matrix, $\mathbf{\Theta} = \mathbf{P}^{-1}$. The graphical Lasso (G-Lasso, Friedman et al., 2007) finds an estimator of the precision matrix $\mathbf{\Theta}$ by minimizing the cost function

$$\mathcal{F}_{\text{G-Lasso}}(\mathbf{\Theta}) = \text{tr}\left(\hat{\mathbf{P}}\mathbf{\Theta}\right) - \log\,\det(\mathbf{\Theta}) + \lambda \sum_{j,k} |\mathbf{\Theta}_{jk}|, \tag{23}$$

over all PSD matrices $\mathbf{\Theta}$. Here, $\hat{\mathbf{P}}$ is the empirical covariance matrix and λ is a regularization strength, so that large λ promote sparsity of the precision matrix estimate. Note that minimizing Equation 23 over all PSD matrices guarantees that the precision matrix estimate is PSD, which in turn guarantees that the covariance matrix estimate is PSD. On the other hand, a sparse precision matrix does not, in general, guarantee a sparse covariance matrix, so the underlying assumptions of the G-Lasso and sparse covariance estimation or soft-thresholding are quite different (Bickel & Lindner, 2012; Morzfeld et al., 2019). The G-Lasso can be computationally expensive because (a) the optimization problem Equation 23 is non-trivial; (b) the method requires tuning to find an appropriate λ.

2.4.6. Convex Sparse Cholesky Selection

Khare et al. (2019) describe a method called *CSCS*, which works with the triangular Cholesky factor $\mathbf{A}$ of the precision matrix $\mathbf{\Theta} = \mathbf{A}^T\mathbf{A}$. Specifically, the goal is to find a sparse Cholesky factor by minimizing the cost function

$$\mathcal{F}_{\text{CSCS}}(\mathbf{A}) = \text{tr}\left(\mathbf{A}^T\mathbf{A}\hat{\mathbf{P}}\right) - 2\,\log\,\det(\mathbf{A}) + \lambda \sum_{1 \leq j < i} |\mathbf{A}_{ij}|, \tag{24}$$

where $\lambda > 0$. Due to the Cholesky factorization, the CSCS method guarantees that the resulting estimators of the precision or covariance matrices are PSD.

19422466, 2024, 9, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024MS004417 by Test, Wiley Online Library on [30/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

2.5. Morozov's Discrepancy Principle

Morozov's discrepancy principle is a technique to adjust regularization parameters in inverse problems (Anzengruber & Ramlau, 2009; Morozov, 1984). Suppose that we are interested in solving the inverse problem whose cost function is

$$F_\alpha(\mathbf{x}) = \frac{1}{2}\| \mathbf{y} - f(\mathbf{x})\|_2^2 + \frac{\alpha}{2}\|\mathbf{x}\|_2^2, \tag{25}$$

where $\mathbf{y}$ are the data, $\mathbf{x}$ is a vector of unknowns, $f(\cdot)$ is a nonlinear function (forward model) and α is a regularization parameter. Solving the inverse problems amounts to minimizing the cost function. We denote the solution of the inverse problem for a given α as $\mathbf{x}_\alpha^*$. The discrepancy principle determines the regularization parameter to be the largest value of α such that

$$\| \mathbf{y} - f(\mathbf{x}_\alpha^*)\|_2 \le S, \tag{26}$$

where the scalar S describes the "noise level" in the problem. For example, if the errors in the data are described by Gaussian noise, then S is derived from the variances of that noise. Application of Morozov's discrepancy principle in practice requires that we solve a sequence of inverse problems, parameterized by α, to find the regularization parameter that leads to a solution that is compatible with the assumed noise level. These ideas can also be used to obtain "regularized" covariance estimates, as we will explain below.

3. New Methods for Noise-Informed Covariance Estimation

Our goal is to design a Hadamard product estimator as in Equation 2, which means that we must build a correction matrix $\mathbf{L}$. Our design must go beyond assuming a spatial decay of correlations, because this assumption is not reasonable in many cases. The design must also adapt itself to diverse situations in order to minimize tuning. We focus on correcting correlations, and we estimate variances directly from the ensemble. This is common in NWP (Gharamti et al., 2019; Hodyss, Campbell, & Whitaker, 2016; Whitaker & Hamill, 2012) and perhaps intuitive because correlations are naturally scaled to the interval [−1, 1].

3.1. Motivation: Damp Small Correlations More Heavily Than Larger Ones

We base the design of our new method on a basic fact about estimating correlations: estimating small correlations is notoriously difficult, and estimating large correlations is, by comparison, easy. One way to understand this fact is to generate ensembles of bivariate Gaussian random variables with varying degrees of correlation and then compute the ensemble correlation. Repeating this process many times allows us to compute the standard deviation in the correlation estimate as a proxy for the error we should expect in the correlation estimate (Anderson, 2012; Lee, 2021b). The average standard deviation (averaged over independent ensemble draws) as a function of the "true" correlation is shown in Figure 1a, for several ensemble sizes.

We note that the standard deviation, or expected error, in the correlation estimate is large if the "true" correlation is small. This means that small correlations are usually not trustworthy, unless the ensemble size is huge. Consequently, it is natural to damp small correlations because it is nearly impossible to distinguish "true" small correlations from sampling error. Large correlations, on the other hand, are usually trustworthy, even if the ensemble size is small. In fact a correlation equal to one should *always* be trusted—the standard deviation goes to zero as the correlation goes to one. This simple numerical experiment thus tells us that a reasonable correlation correction should damp small correlations more heavily than large correlations. The larger error in estimating small correlations is a known feature of the sampling distribution of the correlation coefficient between Gaussian random variables (Flowerdew, 2015).

POLO reiterates the idea that one can usually "trust" large correlations and that small correlations should be damped. To see why, note that if we re-scale the POLO correction Equation 14 so that correlations equal to one are uncorrected, we obtain

$$[\mathbf{L}]_{ij} = \frac{(n_e + 1)\rho_{ij}^2}{1 + \rho_{ij}^2 n_e}. \quad (27)$$

This re-scaled correction factor is shown as a function of correlation in Figure 1b, and we see that it mimics the ideas described just above. At any ensemble size, small correlations are subject to a stronger correction than large ones.

Power law corrections (Lee, 2021b) and "ensemble correlation raised to a power" (ECO-RAP, Bishop & Hodyss 2009a, 2009b, 2007) are also based on the simple fact that one should damp small correlations more severely than larger ones. This is illustrated in Figure 1c, where we show PLC correction factors ($|\rho|^{\beta}$) for a few choices of β. Moreover, PLC nicely resembles the SEC of Anderson (2012) (compare Figures 1b and 1c with Figure 1 of Anderson (2012)).

3.2. Noise-Informed Covariance Estimation

Our new covariance estimator is based on the simple idea that small correlations should be reduced more heavily than large correlations, which we implement by adapting ideas from PLC. Additionally, we make use of an understanding of sampling error (noise) in empirical correlations to make the method adaptive. The use of uncertainties leads to the name of the method, "noise-informed covariance estimation" (NICE).

NICE requires some a priori work that will be used to define the noise level within the correlation estimates. Following the ideas described in Figure 1a, we use (offline) numerical experiments to determine a standard deviation associated with a "grid" of empirical correlations (using bivariate Gaussian random variables, see Section 3.1). We then form a lookup table so that we can assign a standard deviation to *any* empirical correlation via interpolation.

After the offline work, the first actual step of NICE is to compute the n empirical ensemble standard deviations, and the $n(n-1)/2$ empirical ensemble correlations, which we compile in a symmetric $n \times n$ correlation matrix $\hat{\boldsymbol{\rho}}$ (with ones on the diagonal). The sum total noise level, which we call S_{ρ}, is defined as follows. Using the lookup table, we can assign a standard deviation $\sigma_{\rho_{ij}}$ to each correlation $\hat{\rho}_{ij}$ in the matrix $\hat{\boldsymbol{\rho}}$, with the understanding that the standard deviation is zero if the correlation is one. The noise level S_{ρ} is a sum of all noises in the empirical estimate of the correlations:

$$S_{\rho} = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} (\sigma_{\rho_{ij}})^2}. \quad (28)$$

In the second step, we use Morozov's discrepancy principle, applied to the estimation of correlation matrices. The "data" are the empirical estimates of the correlations $\hat{\boldsymbol{\rho}}$, and the preliminary correlation estimate is

$$\hat{\boldsymbol{\rho}}_{\gamma} = \hat{\boldsymbol{\rho}}^{\circ\gamma} \circ \hat{\boldsymbol{\rho}}, \quad (29)$$

where γ is a positive, even integer. The elements of the matrix $\hat{\boldsymbol{\rho}}^{\circ\gamma}$ are $[\hat{\boldsymbol{\rho}}^{\circ\gamma}]_{ij} = \left([\hat{\boldsymbol{\rho}}]_{ij}\right)^{\gamma}$, that is, we raise the empirical correlations *element-wise* to an even, positive power γ. Morozov's discrepancy principle suggests to pick γ such that

$$\| \hat{\boldsymbol{\rho}} - \hat{\boldsymbol{\rho}}_{\gamma} \|_{\text{Fro}} \leq \delta S_{\rho}, \quad (30)$$

where the scalar δ is a tunable factor which we usually set to be equal to one (see numerical examples in Section 4, for cross-covariances in EKI we set $\delta = 0.5$). Specifically, we pick the smallest even, positive integer γ^* that violates the discrepancy principle so that

$$\| \hat{\boldsymbol{\rho}} - \hat{\boldsymbol{\rho}}_{\gamma^*} \|_{\text{Fro}} \geq \delta S_{\rho}, \quad (31)$$

19422466, 2024, 9, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024MS004417 by Test, Wiley Online Library on [30/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

19422466, 2024, 9, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024MS004417 by Test, Wiley Online Library on [30/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

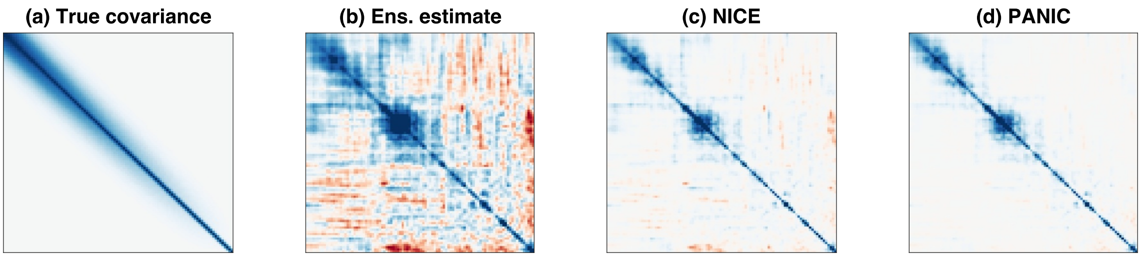

Figure 2. (a) The true covariance matrix. (b) Empirical estimate of the covariance matrix. (c) Noise-Informed Covariance Estimation approximation of the covariance matrix (Section 3.2). (d) Partially Adaptive Noise-Informed Covariance approximation of the covariance matrix (Section 3.3). All estimation methods use $n_e = 20$ samples. The colormap is red for -1, white for 0 and blue for 1.

This procedure determines an exponent γ^* that leads to a correlation matrix estimate that is PSD (γ^* is positive and even) and too strongly regularized according to the discrepancy principle.

The third and final step linearly interpolates between a correction matrix that is too strong (power γ^*) and a correction with a smaller even integer (power $\gamma^* - 2$), which is ostensibly "too weak":

$$\mathbf{L}(\alpha) = \alpha \hat{\boldsymbol{\rho}}^{\circ\gamma^*} + (1-\alpha)\hat{\boldsymbol{\rho}}^{\circ(\gamma^*-2)}. \tag{32}$$

The associated correlation estimate is

$$\hat{\boldsymbol{\rho}}_\alpha = \mathbf{L}(\alpha) \circ \hat{\boldsymbol{\rho}}. \tag{33}$$

The discrepancy principle then determines the interpolation factor α. Specifically, we find α^* to be the largest $\alpha \in [0, 1]$ such that

$$\left\| \hat{\boldsymbol{\rho}} - \hat{\boldsymbol{\rho}}_{\alpha^*} \right\|_{\text{Fro}} \leq \delta S_\rho, \tag{34}$$

that is, we determine the largest PSD correction that satisfies the discrepancy principle. The resulting, corrected correlation estimate is

$$\hat{\boldsymbol{\rho}}_{\text{nice}} = \mathbf{L}(\alpha^*) \circ \hat{\boldsymbol{\rho}}, \tag{35}$$

which in turn leads to the covariance estimate

$$\hat{\mathbf{P}}_{\text{nice}} = \hat{\mathbf{V}} \hat{\boldsymbol{\rho}}_{\text{nice}} \hat{\mathbf{V}}, \tag{36}$$

where $\hat{\mathbf{V}}$ is a $n \times n$ diagonal matrix whose diagonal elements are the ensemble standard deviations.

We can summarize NICE in the following steps.

1. Compute the empirical correlations $\hat{\boldsymbol{\rho}}$ and empirical standard deviations.
2. Determine the noise level S_ρ via a lookup table and Equation 28.
3. Determine the smallest positive, even integer γ^* that violates the discrepancy principle (Equation 31).
4. Determine the largest interpolation factor α^* that satisfies the discrepancy principle (Equation 34).
5. Perform the element-wise correction of the correlation matrix in Equation 35.
6. Use the corrected correlation matrix along with the empirical variances to compute the covariance estimate via Equation 36

The effects of NICE are illustrated in Figure 2, where it is applied to estimate a 100 × 100 covariance matrix, used by Bishop et al. (2017) to study localization in the context of satellite DA (compare our Figure 2a to Figure 2 in Bishop et al. (2017)).

We show the true covariance in Figure 2a, the empirical estimate in Figure 2b, and the NICE estimator in Figure 2c. All approximations use the same ensemble of size $n_e = 20$. The empirical estimate is noisy (large off-

19422466, 2024, 9, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024MS004417 by Test, Wiley Online Library on [30/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

diagonal elements represent spurious covariances) and NICE improves on the empirical estimate by damping small correlations.

3.2.1. Implementation Details and Positive Semi-Definiteness

Step 1 limits the applicability of NICE in extremely high dimensions because we assume that *all* empirical correlations can be computed. As is, NICE can be applied to problems with thousands of unknowns (which we demonstrate in numerical experiments), but it may be computationally expensive if the dimension is 10^5 or larger. When used in EnKF or EKI, one may be able to push these limitations further if the number of observations is relatively small (see Section 4.3), or if NICE is incorporated within an ensemble transform framework (as in ECORAP, Bishop & Hodyss, 2009a, 2009b), or serial filters (Anderson, 2001), or hybrid DA.

Step 2 is trivial, unless the dimension is huge (see comments above about Step 1). For Step 3, we first try $\gamma = 2$ and check the discrepancy principle. If it is violated, we have found γ^* and move to Step 3. If not, we try $\gamma = 4$ and so on. In the examples below, a correction with $\gamma^* = 6$ (or less) was always sufficient, meaning that we need about three (or less) simple iterations to determine γ^*. Moreover, note that if $\gamma^* = 2$ is selected, then step four interpolates between the element-wise power two and the power zero (no correction). For Step 4, we try a small α and gradually increase it (line search) until we violate the discrepancy principle, which then defines the "optimal" α^* to be the previous α we just tried. Alternatively, a root-finding algorithm (e.g., the bisection method) could be used.

We note that instead of a lookup table, one can also directly estimate the noise level S_ρ in (28) using the Fisher transformation. The distribution of the sample correlation coefficient $\hat{\rho}_{ij}$ between normally distributed variables is such that, when the Fisher transformation is taken,

$$z_{ij} = \operatorname{arctanh}\left(\hat{\rho}_{ij}\right) = \frac{1}{2}\log\left(\frac{1+\hat{\rho}_{ij}}{1-\hat{\rho}_{ij}}\right), \tag{37}$$

we have that for $n_e > 3$,

$$z_{ij} \overset{\text{approx.}}{\sim} \mathcal{N}\left(\operatorname{arctanh}\left(\rho_{ij}\right), \frac{1}{n_e - 3}\right), \tag{38}$$

where ρ_{ij} is the true correlation (see, e.g., Flowerdew, 2015). Thus, we can estimate the standard deviation of $\hat{\rho}_{ij}$ as follows. Taking $\hat{\rho}_{ij}$ as an estimate of ρ_{ij}, we draw m samples z_{ij} from the above Gaussian distribution, but replacing $\operatorname{arctanh}(\rho_{ij})$ with $\operatorname{arctanh}\left(\hat{\rho}_{ij}\right)$ in the mean. Second, we apply the inverse Fisher transformation $\tanh(z_{ij})$ to each of the samples and compute their standard deviation. This strategy of computing the noise level in the correlations is attractive because it is (a) easy; and (b) it avoids having to pre-compute lookup tables. The lookup tables, however, have a slight edge over the Fisher transformation approach in terms of their online cost.

Finally, the positive semi-definiteness of the correlation estimator, $\hat{\boldsymbol{\rho}}_{\text{nice}}$, follows from basic facts about Hadamard products. Specifically, raising the elements of a PSD matrix to an even power preserves definiteness, and the sum of two PSD matrices is PSD. The positive semi-definiteness of the covariance estimator $\mathbf{P}_{\text{nice}}$ follows from the fact that a PSD correlation matrix leads to a PSD covariance matrix.

3.3. Partially Adaptive Noise-Informed Covariance (PANIC)

In some problems, for example, in NWP, one may be in the situation where details of the correlation structure are not well-understood, but one may be quite certain that correlations should decay at far distances. For example, Bishop and Hodyss (2011) use a "partially adaptive" method which combines an adaptive localization matrix with a tuned (non-adaptive) localization matrix that eliminates correlations in the far-field. If the problem indeed has this structure (far-field being uncorrelated), then adding this information should increase the accuracy of the estimator because small sampling errors in the far-field accumulate to large errors in high-dimensions (Hodyss & Morzfeld, 2023; Morzfeld & Hodyss, 2023).

One can easily combine these ideas with NICE. Since the resulting method requires some tuning, it is "partially adaptive" (using the language in Bishop and Hodyss (2011)) and we call the method Partially Adaptive Noise-Informed Covariance (PANIC). PANIC amounts to localizing the NICE estimator. Specifically, we use a localization matrix $\mathbf{L}(\ell)$, that depends on a length scale ℓ, to obtain

$$\hat{\boldsymbol{\rho}}_{\text{panic}} = \mathbf{L}(\ell) \circ \hat{\boldsymbol{\rho}}_{\text{nice}}. \tag{39}$$

Here, the length scale ℓ is chosen a priori to be "large enough" to be certain that correlations beyond that length scale are physically implausible. With the correlation estimate we obtain the covariance matrix in the usual way via

$$\hat{\mathbf{P}}_{\text{panic}} = \hat{\mathbf{V}}\hat{\boldsymbol{\rho}}_{\text{panic}}\hat{\mathbf{V}}, \tag{40}$$

where $\hat{\mathbf{V}}$ is a $n \times n$ diagonal matrix whose diagonal elements are the ensemble standard deviations. Figure 2d illustrates PANIC and compares it to NICE. We note that the PANIC estimator reduces spurious correlations in the far field, but in the near field, PANIC and NICE are quite similar by construction. Moreover, the PANIC estimator is PSD because NICE generates a PSD covariance estimate which is subsequently localized (Schur product with a PSD localization matrix); both steps preserve symmetry and definiteness.

3.4. Other New Adaptive Covariance Estimation Methods

Within NICE, we combine an understanding of the noise in empirical correlations with Morozov's discrepancy principle and, for that reason, the method is adaptive and tuning-free. This idea extends to other covariance estimation methods as well, and we now describe how to make some existing covariance estimation methods adaptive.

3.4.1. Adaptive Power Law Correction

PLC requires that one determines the exponent β. In adaptive PLC (Ad.-PLC), we use the largest (but not necessarily integer) β that satisfies the discrepancy principle

$$||\hat{\boldsymbol{\rho}} - \mathbf{L}(\beta) \circ \hat{\boldsymbol{\rho}}||_{\text{Fro}} \leq S_\rho. \tag{41}$$

Recall that $\mathbf{L}(\beta)$ is a matrix whose elements are the absolute values of the empirical correlations raised to the power β: $[\mathbf{L}(\beta)]_{ij} = |[\hat{\boldsymbol{\rho}}]_{ij}|^\beta$. For that reason, Ad.-PLC does *not* guarantee positive semi-definiteness of the covariance estimator (just as PLC). In our numerical examples, however Ad.-PLC always leads to PSD covariance estimators, because the adaptive strategy picks out exponents that are large enough to ensure that the matrix is PSD (see Section 2.4.2).

3.4.2. Adaptive Localization

In "traditional" localization, we define a localization matrix by a length scale ℓ that controls the decay of correlations. In adaptive localization (Ad.-Loc), we determine ℓ to be the largest length scale that satisfies the discrepancy principle

$$||\hat{\boldsymbol{\rho}} - \mathbf{L}(\ell) \circ \hat{\boldsymbol{\rho}}||_{\text{Fro}} \leq S_\rho. \tag{42}$$

In our numerical experiments below we use a simple line search over the length scale ℓ to find an optimal length scale.

3.4.3. Adaptive Soft-Thresholding

Soft-thresholding requires that we determine the thresholding parameter λ in Equation 20. In adaptive soft-thresholding (Ad.-ST), we correct *correlations* and determine the thresholding parameter λ^* to be the largest λ that satisfies the discrepancy principle

19422466, 2024, 9, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024MS004417 by Test, Wiley Online Library on [30/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

$$\|\hat{\boldsymbol{\rho}} - \hat{\boldsymbol{\rho}}_\lambda\|_{\text{Fro}} \leq S_\rho, \tag{43}$$

where $\hat{\boldsymbol{\rho}}_\lambda$ is the empirical correlation matrix thresholded with parameter λ, that is,

$$[\hat{\boldsymbol{\rho}}_\lambda]_{ij} = T_\lambda([\hat{\boldsymbol{\rho}}]_{ij}), \tag{44}$$

where $T_\lambda(\cdot)$ is the soft-thresholding function in (20). With λ^* defined in this way, we obtain the Ad.-ST covariance estimator by

$$\mathbf{P}_{\text{Ad.-ST}} = \hat{\mathbf{V}}\hat{\boldsymbol{\rho}}(\lambda^*)\,\hat{\mathbf{V}}, \tag{45}$$

where $\hat{\mathbf{V}}$ is a diagonal matrix whose diagonal elements are the ensemble standard deviations (as in NICE). Note that Ad.-ST, just like soft-thresholding, does not guarantee a PSD estimate.

3.4.4. Adaptive Sparse Covariance Estimation

The sparse covariance estimation algorithm (Xue et al., 2012), which we briefly describe in Section 2.4, finds a covariance estimate by minimizing the cost function (Equation 22) subject to the constraint that the estimator satisfies $\mathbf{P}_{\text{ASC}} \geq \epsilon\mathbf{I}$, which guarantees that the covariance estimator is PSD. The optimization problem can be solved efficiently, but the optimization problem depends on the regularization parameter λ, which defines the amount of sparsity in the estimate.

Adaptive sparse covariance estimation (ASCE) determines the regularization parameter automatically. As noted by Xue et al. (2012), sparse covariance estimation and soft-thresholding are closely related, because sparse covariance estimation solves the same optimization problem as soft thresholding does, except with an added PSD constraint. Thus, we first perform adaptive soft-thresholding to find an optimal λ^*, and then perform a single optimization with this λ^* to find a sparse correlation estimator $\hat{\boldsymbol{\rho}}_{\text{ASCE}}$ (note that we work exclusively with correlations, not covariances). The ASCE correlation estimator defines the ASCE covariance estimator by

$$\mathbf{P}_{\text{ASCE}} = \hat{\mathbf{V}}\hat{\boldsymbol{\rho}}_{\text{ASCE}}\hat{\mathbf{V}}, \tag{46}$$

where $\hat{\mathbf{V}}$ is, as before, a diagonal matrix whose diagonal elements are the ensemble standard deviations.

4. Numerical Illustrations

We compare NICE to a variety of competing methods, some new and some old. Specifically, we consider the following 13 methods for covariance estimation. We introduce abbreviations for all methods that will be used in the numerical illustrations and in the Figures.

New adaptive methods

1. Noise informed covariance estimation (**NICE**, Section 3.2)
2. Partially adaptive noise informed covariance (**PANIC**, Section 3.3).
3. Adaptive power law corrections (**Ad.-PLC**, Section 3.4.1).
4. Adaptive localization (**Ad.-Loc**, Section 3.4.2).
5. Adaptive soft-thresholding (**Ad.-ST**, Section 3.4.3).
6. Adaptive sparse covariance estimation (**ASCE**, Section 3.4.4).

Methods for comparision

7. The uncorrected, empirical estimate (**Ens.**) serves as the baseline for the improvement a more sophisticated covariance estimation can achieve.
8. **POLO** uses the correction matrix defined in Equation 14 with the “true” correlations (see Section 2.4.1). Using POLO in this way describes a best-case scenario, but we remind the reader that POLO is not a practical algorithm because the true correlations are typically unknown (except in some of our synthetic numerical illustrations).

19422466, 2024, 9, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024MS004417 by Test, Wiley Online Library on [30/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 1
Summary of Covariance Estimation Methods and Their Properties

Method	Adaptivity	Assumptions	PSD guarantees	Computational cost
NICE	Yes	Small corr. noisy	Yes	Low
PANIC	Yes	Small corr. noisy + spatial decay of correlation	Yes	Low
Ad.-PLC	Yes	Small corr. noisy	No	Low
PLC	No	Small corr. noisy	No	Low
Ad.-Loc	Yes	Spatial decay of correlation	Yes	Low
Loc	No	Spatial decay of correlation	Yes	Low
Ad.-ST	Yes	Small corr. noisy	No	Low
ASCE	Yes	Sparse cov.	Yes	Medium
POLO	–	Known corr.	No	Low
Ens.-POLO	–	None	No	Low
Ens.	–	None	Yes	Low
G-Lasso	No	Sparse inv. cov.	Yes	High
CSCS	No	Sparse Cholesky	Yes	High

Note. In the assumptions column, "small corr. noisy" stands for the assumption that small correlations are noisy and, therefore reduced; "sparse cov." stands for the assumption of a sparse covariance matrix; "known corr." stands for the assumption that all correlations are known; "sparse inv. cov." stands for the assumption that the inverse of the covariance matrix is sparse; "sparse Cholesky" stands for the assumption that a Cholesky factor of the inverse covariance matrix is sparse. We assign a "low" computational cost if the technique performs simple operations on the elements of a covariance or correlation matrix. We assign a "high" computational cost if the technique solves optimization problems over (PSD) matrices. The computational cost is labeled "medium" for ASCE, which does perform an optimization over matrices but does so in a particularly speedy way (Xue et al., 2012).

9. POLO with ensemble correlations (**Ens.-POLO**) uses the correction matrix **L** in (14), but the correlations ρ_{ij} are uncorrected *empirical* correlations. This is perhaps the simplest method of increasing the accuracy of the empirical covariance matrix, but we will see that NICE and other methods are superior.
10. Localization (**Loc**) is implemented via a Gaussian localization whose elements are

$$[\mathbf{L}]_{ij} = \exp\left(-(d_{ij}/\ell)^2\right), \tag{47}$$

where d_{ij} is the distance between grid points i and j and where the length scale ℓ is tuned (see below for details). This is an example of the commonly used Hadamard product localization in NWP, which relies on the assumption of a spatial decay of correlation.

11. PLC (Section 2.4.2), with tuned (non-integer) exponent β.
12. The Graphical Lasso (**G-Lasso**, Section 2.4) is implemented in Matlab code that is available on GitHub (we downloaded the code at https://gist.github.com/samwhitehall/6422598). The code yields the G-Lasso estimate of the precision matrix and we subsequently compute its inverse to obtain an estimate of the covariance matrix. We tune the regularization parameter of the G-Lasso in the same way as we tune localization and PLC.
13. Convex sparse Cholesky selection (**CSCS**, Khare et al., 2019, see also Section 2.4) gives a Cholesky factor of the inverse of the covariance matrix. As in the G-Lasso, we use matrix inversion to find the covariance matrix. We tune the regularization parameter in CSCS in the same way as we tune localization, PLC or G-Lasso.

The various covariance estimation techniques and some of their properties are summarized in Table 1. All techniques, except G-Lasso, CSCS and ASCE can be used on non-square correlation matrices (and cross-covariance matrices), which will become important in examples with EKI and in the geomagnetic DA example.

We tune the localization (length scale ℓ), PLC (exponent β), G-Lasso and CSCS (regularization parameter) as follows. We perform a (large) number of training experiments in which we vary the tunable parameter (line search). We then compute an average error and declare the parameter that leads to the smallest error as optimal.

19422466, 2024, 9, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024MS004417 by Test, Wiley Online Library on [30/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

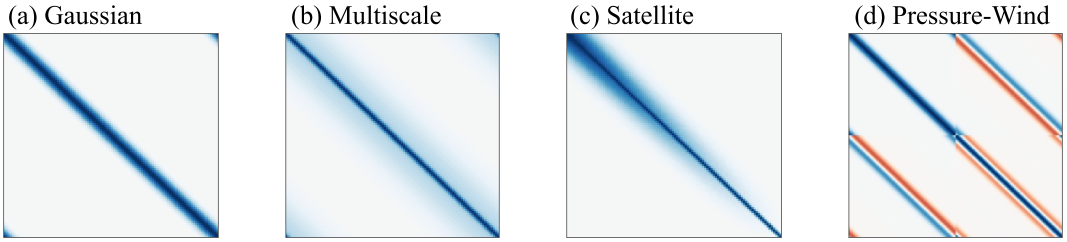

Figure 3. The covariance matrices used in Section 4.1. (a) Gaussian kernel. (b) Multi-scale kernel. (c) Covariance inspired by satellite data assimilation. (d) Covariance of two spatial fields (pressure and wind). Color indicates the matrix elements with blue corresponding to one, white to zero, and red to minus one.

The optimal parameter is used in subsequent experiments. We repeat the tuning for each numerical example because the optimal tunable parameters are problem-dependent.

We use all 13 methods in our first set of numerical experiments with simple Gaussians. Subsequently, we do not use methods that are computationally expensive and that do not yield good results in simple experiments. G-Lasso and CSCS, for example, are quite slow and do not perform well on our first set of simple tests. Other methods, for example, PANIC, may not be applicable in subsequent examples because they presume a spatial decay of correlation. Finally, POLO (with true correlations) can only be used in synthetic scenarios where the correlations are known a priori, which is only true for our first set of very simple experiments.

4.1. Simple Gaussian Tests

We define a 100 × 100 covariance matrix **P** and draw $n_e = 20$ ensemble members from the corresponding Gaussian with mean zero. We then use NICE to estimate the covariance matrix and measure the error in the estimate by

$$\text{Error} = \frac{\|\hat{\mathbf{P}}_{\text{nice}} - \mathbf{P}\|_{\text{Fro}}}{\|\mathbf{P}\|_{\text{Fro}}}. \tag{48}$$

Since the error is random, we average over ensemble draws, and the average error indicates an error we should typically expect. We use the same procedure to compute the error of other covariance estimation methods.

We consider four different covariance matrices, illustrated in Figure 3.

1. *Gaussian kernel.* A covariance matrix **P** with a Gaussian kernel is defined by the elements

$$[\mathbf{P}]_{ij} = \exp\left(-\frac{1}{2}\left(\frac{d_{ij}}{\ell}\right)^2\right),$$

where the length scale is $\ell = 5$ and where d_{ij} is a periodic distance between the grid points i and j. Note that this covariance has the same kernel function as the localization matrix used during classical covariance localization (Loc).

2. *Multi-scale kernel.* A multi-scale covariance **P** is defined by the superposition of two covariance matrices with Gaussian kernels and different length scales:

$$[\mathbf{P}]_{ij} = 0.7\exp\left(-\frac{1}{2}\left(\frac{d_{ij}}{\ell_1}\right)^2\right) + 0.3\exp\left(-\frac{1}{2}\left(\frac{d_{ij}}{\ell_2}\right)^2\right).$$

We chose the length scales to be $\ell_1 = 2$ and $\ell_2 = 20$ (Flowerdew, 2015; Morzfeld & Hodyss, 2023).

3. *Satellite DA covariance.* Bishop et al. (2017) consider the covariance matrix

19422466, 2024, 9, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024MS004417 by Test, Wiley Online Library on [30/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

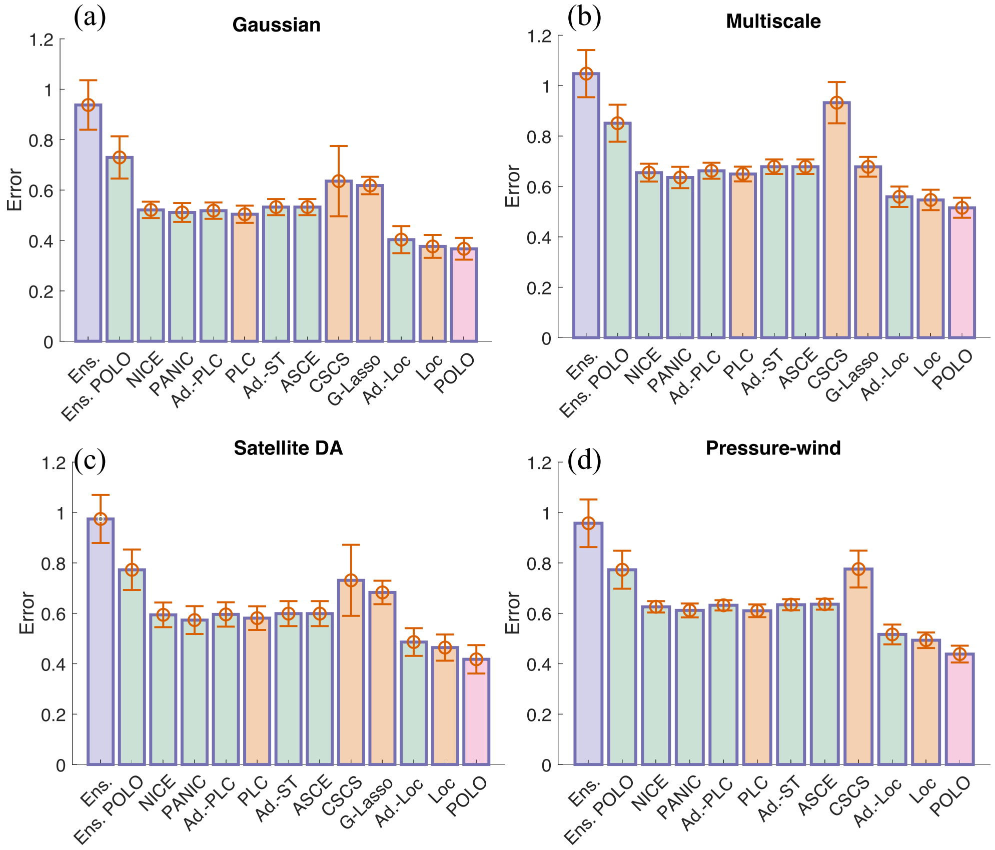

Figure 4. Error (mean and one standard deviation error bars) in covariance matrix estimates for various covariance types with dimension $n_x = 100$. The ensemble size is $n_e = 20$. (a) Gaussian covariance kernel. (b) Multi-scale covariance kernel. (c) Satellite data assimilation covariance matrix. (d) Pressure-wind covariance matrix. The bar chart is color coded so that the vanilla method (Ens.) appears in blue, tuning-free/adaptive methods (Ens.-POLO, Noise-Informed Covariance Estimation, Partially Adaptive Noise-Informed Covariance, Ad.-PLC, Ad.-ST, adaptive sparse covariance estimation, Ad.-Loc) appear in green, tuned methods (Power law corrections, convex sparse Cholesky selection, G-Lasso, Loc) appear in orange, and the infeasible method (POLO) appears in pink (rightmost bar in each panel).

$$[\mathbf{P}]_{ij} = \sqrt{\frac{ij}{n^2}}\exp\left(-\frac{1}{2}\left(\frac{i-j}{\ell_1}\right)^2\right) + \sqrt{\left(1-\frac{i}{n}\right)\left(1-\frac{j}{n}\right)}\exp\left(-\frac{1}{2}\left(\frac{i-j}{\ell_2}\right)^2\right).$$

as a toy problem for satellite DA. Following Bishop et al. (2017), we chose the length scales to be $\ell_1 = 1$ and $\ell_2 = 8$. Note that this covariance matrix is "nonstationary" covariance because the elements of $\mathbf{P}$ depend on i and j, not just of $i - j$.

4. *Pressure-wind covariance.* We consider two spatially extended fields u (pressure) and w (wind), related by a derivative such that $w = \mathrm{d}u/\mathrm{d}x$. We assume that the pressure variable has a Gaussian covariance kernel with length scale $\ell = 5$ and we construct the covariance of w, as well as the cross covariances between u and w, using a finite difference operator (see Morzfeld and Hodyss (2023) for more details). We note that if both u and w have 100 components, the overall dimension of the problem is $n_x = 200$.

We apply all 13 covariance estimation techniques listed above for all but the pressure-wind covariance, for which we do not apply G-Lasso because the code runs very slowly on this 200-dimensional problem. Note that all four covariance matrices we consider here have exponentially small correlations in the far field (away from the diagonal), so that the use of a localization and PANIC are appropriate. Results are summarized in Figure 4, which shows the average error (10^3 trials) for each method and covariance type along with one standard deviation error bars.

19422466, 2024, 9, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024MS004417 by Test, Wiley Online Library on [30/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

The numerical experiments support the following conclusions.

1. For all four covariance types, *all* covariance estimation techniques are more accurate than the sample covariance matrix, which always has the largest error.
2. POLO with ensemble correlations (Ens.-POLO) improves the covariance estimates in all four cases, but not to the extent of the other methods we tried.
3. NICE, Ad.-PLC, Ad.-ST and ASCE lead to similar errors which are in turn comparable to the errors of a finely tuned PLC. The fact that all four adaptive methods perform as well as a related finely tuned method suggests that the discrepancy principle and the pre-computed noise level are robustly applicable to adaptive covariance estimation.
4. The adaptive localization (Ad.-Loc) leads to errors almost as small as the errors obtained by a finely tuned localization (Loc). This reiterates our previous point, that is, that adapting localization/covariance estimation parameters via a discrepancy principle is a robust idea.
5. The errors of PANIC are slightly smaller than the errors of NICE, which suggests that reducing the (non-adaptive) far-field correlations has a positive effect.
6. Localization (Loc) comes close to the optimal errors obtained by POLO and Loc and POLO lead to the smallest errors in all four examples.
7. G-Lasso and CSCS lead to smaller errors than Ens.-POLO, but the errors are larger than for the new adaptive methods. G-Lasso and CSCS also require significantly more computations than the competing methods, and we conclude that G-Lasso and CSCS are not competitive in these examples. Recall, however that G-Lasso and CSCS are designed to estimate the precision matrix (not the covariance matrix as we do here). CSCS further targets applications with a natural ordering of the data.

During the trials of our experiments we monitored if a covariance matrix estimate was PSD or not. When the exponent in PLC was chosen adaptively (Ad.-PLC) or via tuning, we encountered *no* negative eigenvalues, while POLO, Ens. POLO, and Ad.-ST often produced non-PSD estimates. This is an interesting result because POLO is the estimator with the lowest errors and yet it is not always PSD. Our error metric here, however, does not account for this deficiency, violating the PSD property may cause instability within EnKFs or EKI (see Section 2.3).

When we increase the dimension of the problem, the decrease in errors is more dramatic (Hodyss & Morzfeld, 2023). Figure 5 summarizes results obtained for problems of dimension $n = 1{,}000$.

We note qualitatively the same results as in the $n = 100$ dimensional example: NICE, Ad.-PLC, Ad.-ST and ASCE are comparable and, even though these methods do *not* require tuning, they are as good as a tuned PLC. These four methods, however, do not lead to errors as small as those obtained by localization (tuned or adaptive) or an optimal correction (POLO).

Finally, note that the correlations decay with distance in all above examples, which is exploited by classical (or adaptive) localization, but this correlation structure is *discovered* by the adaptive methods (NICE, Ad.-PLC, Ad.-ST and ASCE). Our first set of simple tests thus suggests that NICE, Ad.-PLC, Ad.-ST and ASCE can be viable options in problems where assumptions about the underlying correlation structure are unavailable or in problems where one wishes to reduce the tuning costs.

4.2. Cycling Data Assimilation Experiments With the Lorenz '96 Model

We perform cycling DA experiments with the Lorenz'96 model (L'96, Lorenz, 1996) and an EnKF (stochastic EnKF implementation, Burgers et al., 1998; Evensen, 2009, 1994). Specifically, we apply, within the EnKF, the covariance estimation methods NICE, PANIC, Ad.-PLC, ASCE, PLC, Ad.-Loc, localization, and a version of POLO that indicates a best-case scenario at the expense of requiring a very large ensemble (hence being infeasible in practice). As is common in DA, we apply the covariance estimation (NICE, etc.) in conjunction with a covariance *inflation*. For the inflation, we simply set

$$\mathbf{P} \leftarrow (1+\kappa)\mathbf{P}, \tag{49}$$

where $\kappa > 0$ is an inflation parameter (tuned, see below).

The tuning of covariance estimation and/or the inflation is as follows. For the adaptive methods (NICE, Ad.-PLC, ASCE and Ad.-Loc), we only need to tune the inflation parameter κ. For PANIC, we also only tune the inflation

19422466, 2024, 9, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024MS004417 by Test, Wiley Online Library on [30/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

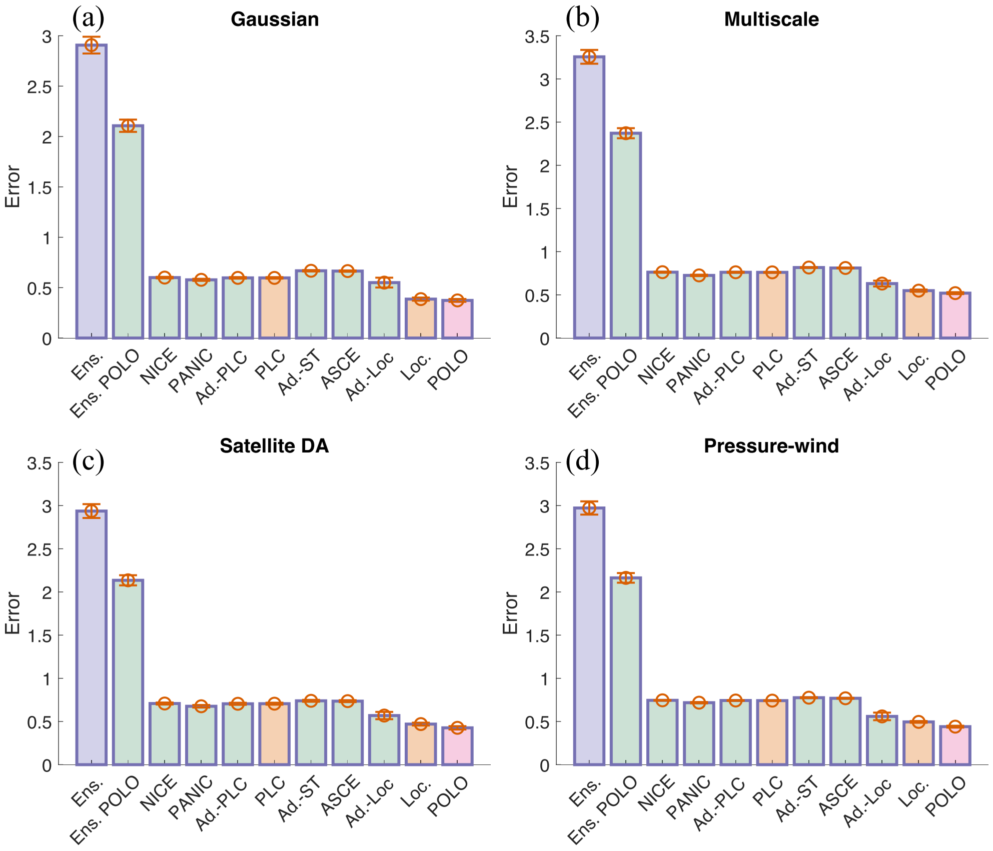

Figure 5. Same as Figure 4, but with $n = 1{,}000$. Error (mean and one standard deviation error bars) in covariance matrix estimates for various covariance types with dimension $n = 1{,}000$. (a) Gaussian covariance kernel. (b) Multi-scale covariance kernel. (c) Satellite data assimilation covariance matrix. (d) Pressure-wind covariance matrix. The bar chart is color coded so that the vanilla method (Ens.) appears in blue, tuning-free methods (Ens.-POLO, Noise-Informed Covariance Estimation, Partially Adaptive Noise-Informed Covariance, Ad.-PLC, Ad.-ST, adaptive sparse covariance estimation, Ad.-Loc) appear in green, tuned methods (Power law corrections, Loc) appear in orange, and the infeasible method (POLO) appears in pink (rightmost bars of each panel).

and set the length scale for the spatial localization to $\ell = 10$, which is wide enough to expect that correlations beyond that length scale are unreasonable. For localization, we tune the length scale jointly with the inflation parameter κ. Similarly, for PLC we tune the exponent β jointly with the inflation parameter κ. In all cases, the tuning is done by running 2,000 DA cycles, disregarding the first 200 cycles as "spin-up," and recording the associated, time-averaged root mean square error (RMSE) for each inflation parameter and, if needed, additional covariance estimation parameters. The parameters that lead to the smallest time-averaged RMSE in the training experiment are subsequently used in another, independent experiment in which we perform 1,000 DA cycles, disregard the first 100 cycles as spin-up, and average RMSE after the spin-up period. Throughout the experiments, we hold the ensemble size constant at $n_e = 20$. The state dimension is $n = 40$ and we observe every other variable, that is, the number of observations is equal to 20. All observation error variances are equal to one. Observations are collected every 0.4 (dimensionless) time units and the time step of the numerical integrator (a fourth order forward Runge-Kutta method) is set to $\Delta t = 0.05$ (this is the same setup as in Hodyss and Morzfeld (2023)).

To establish a best-case scenario, we use an EnKF *without* inflation or localization/covariance estimation but with a large ensemble size $n_e = 500$. We further apply POLO to an EnKF with $n_e = 20$ (with tuned inflation), but run, in parallel, the large ensemble size EnKF ($n_e = 500$) to obtain the correlation information. These latter experiments can indicate what a near-optimal localization may achieve (assuming the large ensemble size EnKF reveals the main features of the "true" correlation).

19422466, 2024, 9, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024MS004417 by Test, Wiley Online Library on [30/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

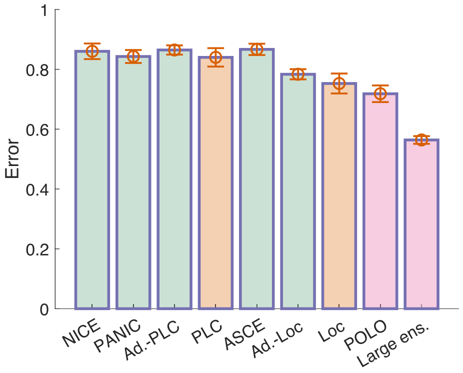

Figure 6. Time-averaged analysis root mean square error after spin up along with one standard deviation error bars in ensemble Kalman filter (EnKF) with various covariance estimation techniques. The bar chart is color coded so that tuning-free methods (Noise-Informed Covariance Estimation, Partially Adaptive Noise-Informed Covariance, Ad.-PLC, adaptive sparse covariance estimation, Ad.-Loc) appear in green, tuned methods (Power law corrections, Loc) appear in orange, and infeasible methods (POLO and the large ensemble EnKF) appear in pink (two bars on the right).

The results of our numerical experiments are summarized in Figure 6, which shows the time average of the analysis RMSE of EnKFs with various covariance estimation/localization techniques.

The results of the cycling DA experiments follow a similar pattern as the simpler tests with "static" covariances of the previous section.

1. NICE, Ad.-PLC, ASCE and the tuned PLC lead to nearly identical errors, and the adaptive localization (Ad.-Loc) comes fairly close to the tuned localization (Loc), reiterating that the discrepancy principle is robustly applicable to adaptive covariance estimation.
2. PANIC reduces the error as compared to NICE, because the assumption of zero (or near-zero) correlations in the far-field is valid for L'96. The additional error reduction that PANIC achieves over NICE, however, is minor (as in the previous, non-cycling examples).
3. Localization leads to smaller errors than NICE, PANIC, Ad.-PLC or PLC, but the errors are still larger than what can be achieved with a large ensemble size or a nearly optimal localization (POLO).
4. POLO based on correlations extracted from a large ensemble leads to smaller errors than all other techniques, but still cannot reach the low error achieved by a large ensemble size. This could be due to the Gaussian assumption underpinning POLO, which is not satisfied in cycling DA experiments with L'96, or it could indicate more general limitations of correlation-based covariance corrections.
5. We encountered no negative eigenvalues during the cycling DA experiments with PLC or Ad.-PLC.

We further note that all covariance estimation methods (NICE, PANIC, Ad.-PLC, PLC, ASCE, Ad.-Loc, Loc, POLO) lead to much smaller errors than a vanilla EnKF without inflation or localization/covariance estimation. The vanilla EnKF diverges and, therefore, leads to macroscopic error.

Our test with the L'96 model re-iterates our conclusions based on the simpler experiments of the previous section: NICE, Ad.-PLC and ASCE reduce the error in covariance estimates and, therefore, in a cycling EnKF *without* making any assumptions about the underlying correlation structure and *without* tuning (only inflation is tuned for these methods). The fact that localization leads to smaller errors than NICE, Ad.-PLC or ASCE stems from the heavy tuning and, perhaps more importantly, from the fact that the underlying correlation structure here is consistent with the assumptions of classical localization.

4.3. Cycling Data Assimilation Experiments With a Geomagnetic Proxy Model

We consider cycling DA experiments with an EnKF on a proxy model for geomagnetic DA, described in detail by Gwirtz et al. (2021). The model consists of a (chaotic) Kuramoto-Sivashinsky (KS) equation coupled to an induction equation, and describes the spatial and temporal variations of a velocity field coupled, via induction, to a magnetic field. We consider the model in a 2D configuration on a square and discretize the partial differential equations by a spectral method (Fourier series), which leads to a state dimension of $n = 1{,}920$ Fourier coefficients. Following Gwirtz et al. (2021), we collect observations of Fourier modes of the magnetic field with wavenumbers in the x- and y-directions that are less than or equal to three (for a sum total of 48 Fourier coefficients). The time interval between two consecutive observations is about 7% of the model's e-folding time. Note that the velocity field is entirely unobserved. This setup is somewhat indicative of what to expect in a larger numerical dynamo model for decadal-scale forecasts of the geomagnetic field (Gwirtz et al., 2021).

We assimilate the spectral observations using a stochastic EnKF with ensemble size $n_e = 100$, essentially repeating the DA experiment reported in Section 4.2 of Gwirtz et al. (2021). Since we observe Fourier coefficients, we have no natural notion of a "spatial" distance, and we therefore resort to NICE and Ad.-PLC to correct the covariances within the EnKF. We have tried hard, but failed to find a localization based on a spatial decay of correlation that reduces errors, see also Gwirtz et al. (2021). Note that the state dimension is large ($n_x = 1{,}920$), but the number of observations is small ($n_y = 48$), so that it is natural to estimate the matrices $\mathbf{H}\hat{\mathbf{P}}\mathbf{H}^T$ (48×48) and $\hat{\mathbf{P}}\mathbf{H}^T$ ($1{,}920 \times 48$), rather than the ensemble covariance $\hat{\mathbf{P}}$ ($1{,}920 \times 1{,}920$). The results reported below, however, do not change much if we estimate the ensemble covariance $\hat{\mathbf{P}}$ using the same methods. A more

19422466, 2024, 9, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024MS004417 by Test, Wiley Online Library on [30/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

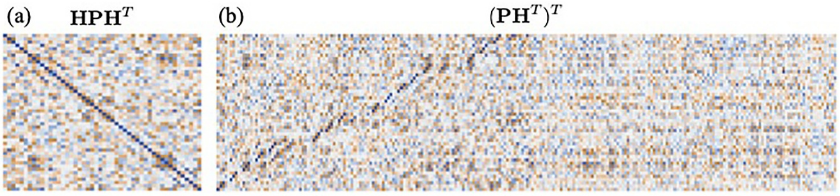

Figure 7. Correlations in a cycling ensemble Kalman filter (EnKF) for the geomagnetic model (large ensemble size). (a) The 48 × 48 correlation matrix associated with $\mathbf{HPH}^T$ during one cycle (post spin-up) of an EnKF with ensemble size $n_e = 1{,}000$. (b) The correlation matrix associated with $\mathbf{PH}^T$, during one cycle (post spin-up) of an EnKF with ensemble size $n_e = 1{,}000$. The correlation matrix associated with $\mathbf{PH}^T$ is truncated at wave number five so that its size is 240 × 48. To make the figure easier to read, we transpose the correlation matrix associated with $\mathbf{PH}^T$. What is important to note from this figure is that (i) there are strong correlations across various variables represented in $\mathbf{HPH}^T$ and $\mathbf{PH}^T$; and (ii) the correlations follow no discernible pattern and, what's worse, large and small correlations switch places across various assimilation cycles (not shown, but see Gwirtz et al. (2021)). Color indicates the matrix elements with blue corresponding to one, white to zero, and red to minus one.

detailed discussion of the differences between these two approaches in the context of localization can be found in Campbell et al. (2010).

The apparent absence of correlation structure in covariance matrices within an EnKF is described in detail in Gwirtz et al. (2021) (see, e.g., Figures 5a, 5b, and 10 of Gwirtz et al. (2021)), and is also illustrated in Figure 7, where we plot correlation matrices associated with $\mathbf{PH}^T$ and $\mathbf{HPH}^T$ during one cycle of an EnKF with a large ensemble size $n_e = 1{,}000$. It is clear from the figure that correlations are strong throughout the system, but also that there is no coherent pattern. The lack of discernible correlation patterns makes estimating covariances from a small ensemble difficult, but, as we will see, NICE and Ad.-PLC handle this problem well. Moreover, the correlations change from one DA cycle to the next (see Figure 10 in Gwirtz et al. (2021)), but since NICE and Ad.-PLC are adaptive, these methods can capture the time-varying correlation structure within this cycling EnKF.

For our numerical tests, we set the ensemble size to $n_e = 100$ and use NICE and Ad.-PLC, along with a 6% covariance inflation of both $\mathbf{HPH}^T$ and $\mathbf{PH}^T$ (Gwirtz et al., 2021). For each EnKF, we perform 600 DA cycles, with the first 300 cycles being discarded as "spin-up."

Figure 8 illustrates the results of our numerical experiments for NICE (results for Ad.-PLC are similar). Panels (a) and (b) show errors (truth minus a one-cycle forecast) as a function of the DA cycle for the velocity field and magnetic field for EnKFs with NICE (green). We note the spin-up period and the subsequent stable DA phase. The errors in the figure are normalized by the macroscopic error, which is the error one would expect without any DA. Panels (c)–(e) illustrate a forecast based on an EnKF using NICE for covariance estimation. Shown is the vorticity of the velocity field approximately 4.7 e-folding times *after* the last assimilation cycle (panel (c)), along with the NICE-EnKF forecast (panel (d)) and the difference of the two (panel (e)). It is notable that the EnKF with a NICE covariance estimation can be used to create forecasts that are accurate on practically relevant time scales.

We compare the performance of an EnKF with a small ensemble size ($n_e = 100$) using NICE and Ad.-PLC, to an EnKF with a large ensemble ($n_e = 1,\ 000$) but *without* covariance corrections. In this context, it is important to note that Gwirtz et al. (2021) showed that an EnKF without covariance corrections stabilized on this problem with an ensemble size of $n_e = 800$. We further consider an EnKF with $n_e = 100$, and with covariance estimation based on a shrinkage estimator, which decreases the magnitude of all off-diagonal elements of a covariance matrix. The shrinkage estimator is taken from Gwirtz et al. (2021), where it was heavily tuned, and was found to be "the best" covariance estimation method for this problem.

The results of our numerical experiments and relevant results reported in Gwirtz et al. (2021) are summarized in Table 2, which lists errors in magnetic (observed) and velocity (unobserved) fields.

We note a similar pattern as in our earlier experiments: NICE and Ad.-PLC are as good or better than a finely tuned estimator (Shrinkage) and the adaptive covariance estimation methods indeed come quite close to the performance of an EnKF with a much larger ensemble size. Moreover, both methods succeed in propagating information from the observed magnetic field to the unobserved velocity field, as indicated by the small errors in the unobserved velocity field. In this example, NICE leads to smaller errors than Ad.-PLC (in both fields).

19422466, 2024, 9, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024MS004417 by Test, Wiley Online Library on [30/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

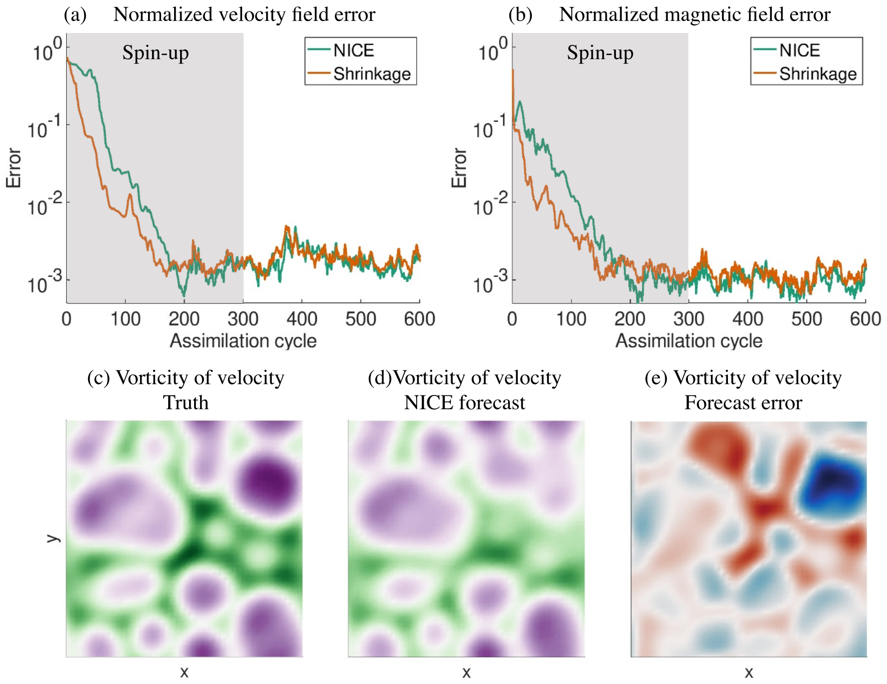

Figure 8. Illustration of covariance estimation within a cycling ensemble Kalman filter for a geomagnetic proxy model. (a) Normalized error in the unobserved velocity field as a function of assimilation cycle for two covariance estimation methods (Noise-Informed Covariance Estimation (NICE) and shrinkage). (b) Normalized error in the partially observed magnetic field as a function of assimilation cycle for two covariance estimation methods (NICE and shrinkage). (c) Vorticity of the velocity field. (d) Forecast of the vorticity of the velocity field based on data assimilation with NICE. (e) Forecast error (difference between panels (c) and (d)).

Nonetheless, the fact that both adaptive methods succeed with essentially no tuning on a problem that is much harder and much more high-dimensional than the previous test problems is reassuring and speaks to the robustness of the proposed techniques. Moreover, NICE leads to smaller errors than the best method thus far reported in the literature (the heavily tuned shrinkage estimator of Gwirtz et al. (2021)).

Finally, we report (again) that even though Ad.-PLC does not guarantee PSD covariance estimates, all covariances ($\mathbf{HPH}^T$) that were estimated with this method in this example turned out to be PSD.

4.4. Inversion of Electromagnetic Data

We now apply EKI to a marine electromagnetic (EM) inverse problem. The goal of the inversion is to compute resistivity as a function of depth from measurements of apparent resistivity and phase, both as a function of period. The seafloor magnetotelluric (MT) data (10 apparent resistivities along with 10 phases, see Figure 10d) are collected off-shore of New Jersey (Blatter et al., 2019; Gustafson et al., 2019). The data are equipped with error estimates in the form of standard deviations. The MT model uses a standard recursion relationship (Ward and Hohmann (2012), see also Blatter et al. (2022a, 2022b)), and is discretized with 60 layers, each 20 m thick,

As is common in geophysical inversion, we use a quadratic regularization, that is, we minimize the cost function

$$F(\mathbf{x}) = \left\| \mathbf{R}_d^{-\frac{1}{2}}(\mathbf{d} - \mathcal{M}(\mathbf{x})) \right\|_2^2 + \mu \left\| \mathbf{B}^{-\frac{1}{2}}\mathbf{x} \right\|_2^2, \tag{50}$$

where $\mathbf{d}$ are the data, $\mathbf{x}$ are the unknown resistivities, $\mathcal{M}$ is the MT model, $\mathbf{R}_d$ is a diagonal matrix that contains the variances associated with the data on its diagonal, and where $\mathbf{B}$ is a regularization matrix, which we chose to be a

Table 2
Normalized Errors Scaled by the Respective Macroscopic Errors and Multiplied by 10^3 for Three Covariance Estimation Methods and for an Ensemble Kalman Filter With a Large Ensemble Applied to a Geomagnetic Proxy Model

	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

19422466, 2024, 9, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024MS004417 by Test, Wiley Online Library on [30/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

covariance matrix with a Gaussian kernel and length scale $\ell = 200$ m. The regularization parameter μ was obtained via an Occam inversion (Constable et al., 1987). We note that discovering an appropriate regularization strength μ in EM inversions is an interesting subject in itself, but for the purposes of this numerical demonstration, it is sufficient to think of μ as being given. A similar EM inverse problem was considered by Tong and Morzfeld (2023), also in the context of localizing EKI.

To apply EKI to this regularized problem, we recast the cost function as

$$f(x) = \left\| \mathbf{R}^{-\frac{1}{2}}(\mathbf{y} - \mathcal{G}(\mathbf{x})) \right\|_2^2, \tag{51}$$

where

$$\mathbf{R}^{-\frac{1}{2}} = \begin{pmatrix} \mathbf{R}_d^{-\frac{1}{2}} & \mathbf{0} \\ \mathbf{0} & \sqrt{\mu}\ \mathbf{B}^{-\frac{1}{2}} \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} \mathbf{d} \\ \mathbf{0} \end{pmatrix}, \quad \mathcal{G}(\mathbf{x}) = \begin{pmatrix} \mathcal{M}(\mathbf{x}) \\ \mathbf{x} \end{pmatrix}. \tag{52}$$

This “trick” is explained in detail in Chada et al. (2020) where the resulting method is called Tikhonov regularized EKI (TEKI). Note that the EKI framework (see Section 2.2) can now be directly applied, but at the expense that the “data-data” correlations in $\hat{\mathbf{C}}_{gg}$ are stacks of an ensemble of model outputs and the ensemble itself.

Recall that the EKI iteration requires that we repeatedly estimate the covariances $\hat{\mathbf{C}}_{gg}$ and $\hat{\mathbf{C}}_{xg}$ from the ensemble. We correct these covariances using NICE, Ad.-PLC and ASCE. For all three methods, our numerical experiments indicate that the tunable parameter δ in the discrepancy principle needs to be decreased when we correct the data-to-unknown covariances $\hat{\mathbf{C}}_{xg}$. A factor of $\delta = 0.5$ leads to good results, whereas $\delta = 1$ leads to TEKI iterations that do not reduce the error as low as with $\delta = 0.5$. The reason for reducing δ is that a smaller δ leads to a softer correction, which is needed because several of the “true” data-to-unknown covariances are small, and it is advantageous to keep them, rather than to remove them, in order to propagate information from the data to the unknown variables. This effect is illustrated in Figure 9: NICE with a “strong” correction ($\delta = 1$) is adequate for the data-data correlations (top row), but inadequate for the cross correlations (bottom row).

Implementing a spatial localization is neither intuitive nor easy in this example, but we tried it nonetheless. First, we apply a localization to $\hat{\mathbf{C}}_{gg}$, although this has little physical motivation. We chose a localization matrix with a Gaussian kernel and a length scale $\ell = 200$ m after some initial tries (no careful tuning). Performing a localization on $\hat{\mathbf{C}}_{xg}$ (60×80) is more tricky. We apply *no* localization to the first 20 columns of $\hat{\mathbf{C}}_{xg}$, which corresponds to the covariances computed from the ensemble of model outputs. We apply a Gaussian localization with length scale $\ell = 200$ m to the remaining 60 columns. With this same setup, we can also apply an adaptive localization (Ad.-Loc).

We now run TEKIs with various covariance estimation schemes and covariance inflation (see Equation 49). The inflation depends on the RMSE, defined by

$$\text{RMSE} = \sqrt{\frac{1}{n_d}\sum_{i=1}^{n_d}\left(\frac{\mathbf{d}_i - \hat{\mathbf{d}}_i}{\sigma_i}\right)^2}, \tag{53}$$

where $\mathbf{d}_i$, $i = 1, \ldots, n_d$, are the n_d data points, σ_i are the corresponding observation error standard deviations (given as part of the MT data set as the diagonal elements of $\mathbf{R}_d^{1/2}$); $\hat{\mathbf{d}} = \mathcal{M}(\hat{\mathbf{x}})$ are model predictions based on the mean of the TEKI ensemble, $\hat{\mathbf{x}}$. The inflation is $\kappa = 15\%$ when $\text{RMSE} > 1.2$, $\kappa = 10\%$ when $1.1 \leq \text{RMSE} \leq 1.2$, and we turn the inflation off ($\kappa = 0$) when $\text{RMSE} < 1.1$. We did not tune the inflation systematically.

We use TEKIs with ensemble size $n_e = 30$ and 200 iterations. For each TEKI, we perform 100 independent experiments, each with a different random initial ensemble and then average the results. Our findings are summarized in Figure 10.

19422466, 2024, 9, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024MS004417 by Test, Wiley Online Library on [30/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

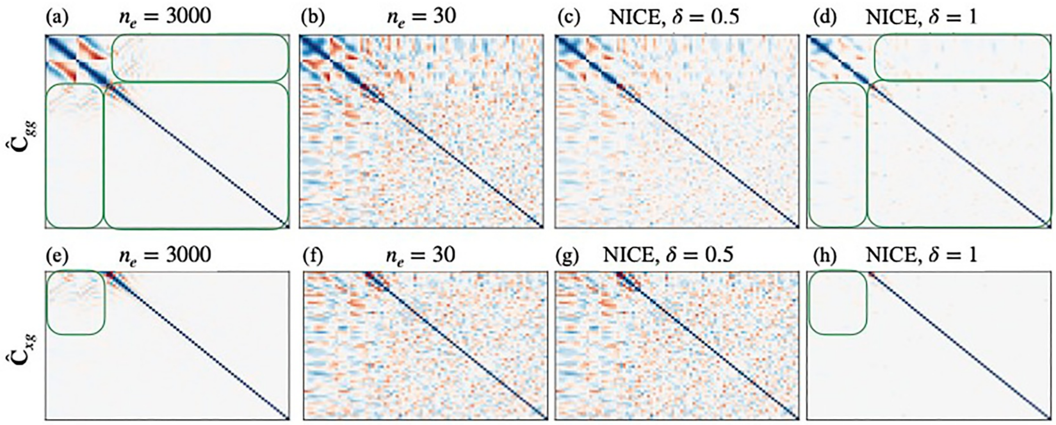

Figure 9. Correlation matrices corresponding to $\mathbf{C}_{gg}$ (top row) and $\mathbf{C}_{xg}$ (bottom row) during one step of a TEKI. Panels (a) and (e) show estimates of the correlations for a large ensemble size. Panels (b) and (f) show estimates of the correlations for a small ensemble size. Panels (c) and (g) show the Noise-Informed Covariance Estimation (NICE) estimator with $\delta = 0.5$. Panels (d) and (h) show the NICE estimator with $\delta = 1$. In panel (a), green squares highlight areas in which correlations are weak, which NICE with $\delta = 1$ (panel (d)) dampens, but NICE with $\delta = 0.5$ keeps. In panel (e), a green square highlights an area in which correlations are present, but which are dampened too strongly by NICE with $\delta = 1$, as in panel (h), whereas NICE with $\delta = 0.5$ keeps these correlations. Color indicates the matrix elements with blue corresponding to one, white to zero, and red to minus one.

Panel (a) shows the averaged RMSE for each TEKI. Note that an RMSE of approximately one is good because then the TEKI estimate fits the data to within the assumed error level (standard deviation of the data). First, we note as before that all covariance estimation methods (NICE, Ad.-PLC, PLC, ASCE, Ad.-Loc, Loc) lead to TEKIs which can achieve an acceptably low RMSE and that the adaptive methods are nearly as good as the tuned methods or a TEKI with a larger ensemble ($n_e = 200$). Second, we note that the inflation already has a large effect on the RMSE: An inflated TEKI reaches an RMSE that is lower than a "vanilla" TEKI without any covariance estimation or inflation.

We further assess the "quality" of our TEKI inversions by comparing the TEKI results to a gradient-based optimization (Gauss-Newton). We measure the difference between the TEKI result and the Gauss-Newton result by the error

$$\text{Ref. Error} = \frac{||\text{res.}_{\text{GN}} - \text{res.}_{\text{teki}}||_2^2}{||\text{res.}_{\text{GN}}||_2^2} \tag{54}$$

where res. refers to the (log) resistivity and the subscript GN refers to the Gauss-Newton method and subscript teki refers to a TEKI result. The error with respect to a reference model is shown in Figure 10b. We see that the reference error behaves very similarly to the RMSE (not surprisingly): The covariance estimation methods all lead to a small reference error and all methods perform similarly. The inflated TEKI (no additional covariance estimation) leads to a significantly larger reference error than the other TEKIs, although the RMSE is comparable. The two errors can be different here because many different models fit the MT data similarly well. The large reference error indicates that the model obtained with an inflated TEKI is quite different from the reference model. Thus, the covariance estimation is helpful here to "smooth" the models so that they are similar to the reference model, obtained by Gauss-Newton (see also Figure 10c).

Finally, Figures 10c and 10d show a typical result obtained with TEKI and NICE. Panel (c) shows the (log) resistivity as a function of depth and panel (d) shows the associated fit to the data. For comparison, we also show the resistivity and data fit we obtain via Gauss-Newton. The TEKI approximation with NICE is very similar to the Gauss-Newton result for depths up to about 600 m, where the data are most informative (small error with respect to the reference model in Figure 10b) and the fit to the data for TEKI and Gauss-Newton is nearly identical (small RMSE in Figure 10a).

19422466, 2024, 9, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024MS004417 by Test, Wiley Online Library on [30/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

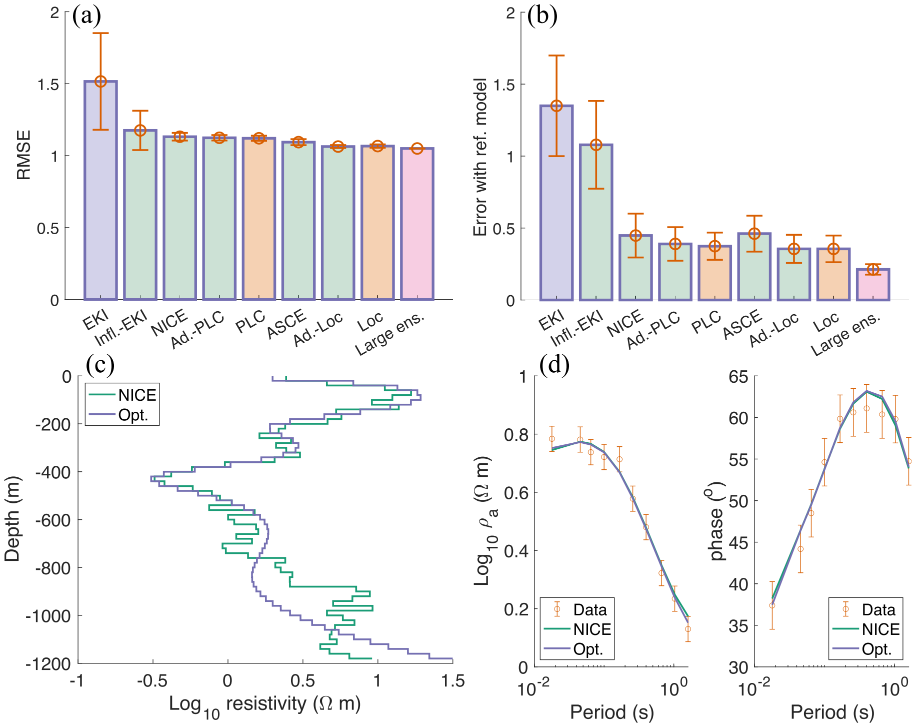

Figure 10. Summary of results for the electromagnetic (EM) inversion. (a) root mean square error (see (53)) of various TEKI implementations. (b) Error with respect to a reference model, obtained via gradient-based optimization (see (54)) of various TEKI implementations. In panels (a) and (b), the bars are averages over results obtained by randomizing the initial TEKI ensemble and the error bars denote one standard deviation. The bars are color coded so that blue labels a "vanilla" TEKI, green labels a TEKI with an adaptive covariance estimation, orange labels a TEKI with a tuned covariance estimation, and pink (furthest to the right) labels a large ensemble result. Panel (c) shows (log) resistivity as a function of depth obtained via Gauss-Newton optimization (blue) and TEKI with Noise-Informed Covariance Estimation (NICE) (green). Panel (d) shows the EM data and the model output resulting from TEKI with NICE (green) and Gauss-Newton method (purple). Averages and standard deviations are computed from 100 independent numerical experiments.

4.5. Training Feed-Forward Neural Networks With Time-Averaged Data

Our last example is a simplification of a climate sciences problem in which sub-grid parameterizations of a climate model are represented by a NN. The training strategy for the neural network is to define a loss function in terms of time-averaged data of the climate model and to adjust the weights and biases of the NN to minimize the loss function. The usual back propagation (gradient descent) cannot be used in this context because the "map" from the NN weights and biases to the time-averaged data of the climate model may not be differentiable, or derivatives may be difficult to obtain (Schneider et al., 2024).

As a "cartoon" for this difficult problem, we consider a modified Lorenz model (mL'96) as a stand-in for a climate model and we parameterize the forcing of mL'96 by a simple feed-forward neural network. Specifically, the mL'96 model is

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = (x_{i+1} - x_{i-2})\,x_{i-1} - x_i + F_i, \tag{55}$$

where $x_{-1} = x_{n_x-1}$, $x_0 = x_{n_x}$, $x_{n_x+1} = x_1$ (periodicity) and where

$$F_i = 8 + 6\sin\left(\frac{4\pi}{n_x}i\right), \tag{56}$$

is a coordinate-dependent forcing. Note that the forcing is the only modification we make, and our modification is inspired by the storm-track model of Bishop et al. (2017). We choose the state dimension to be $n_x = 100$.

19422466, 2024, 9, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024MS004417 by Test, Wiley Online Library on [30/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

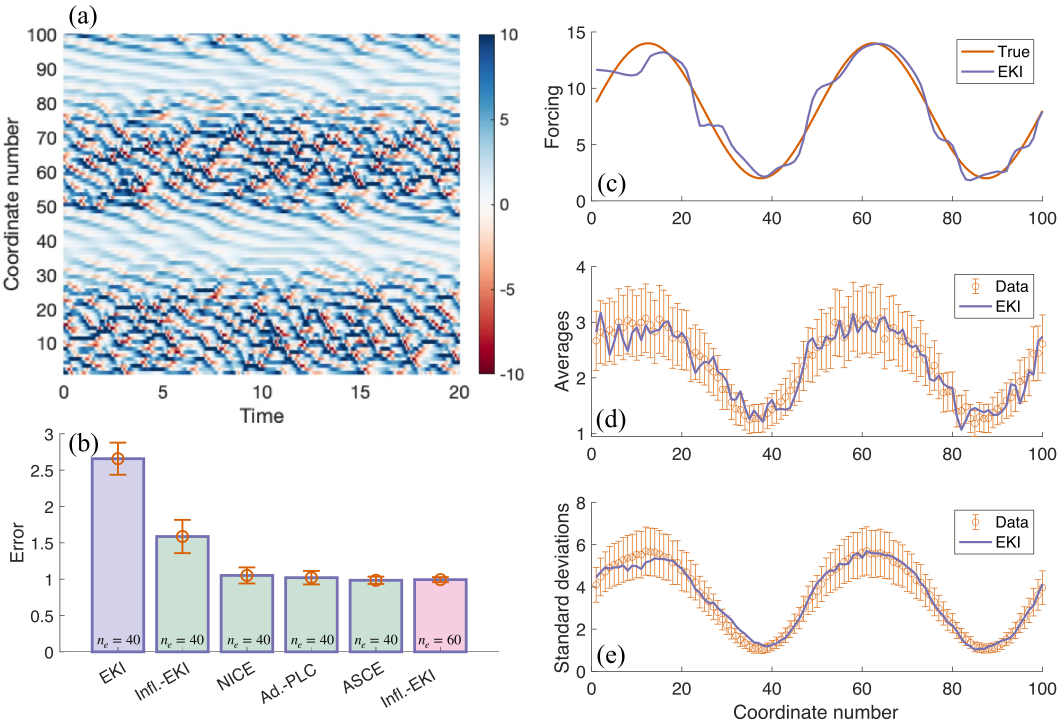

Figure 11. (a) Hovmöller diagram of the mL'96 model, showing the time evolution of all n_x coordinates as a function of time. (b) Average root mean square error (bars) and standard deviations (error bars) of several ensemble Kalman inversion (EKI) variants, computed over 10 independent experiments, each using a different set of perturbations within the various EKIs. (c)–(e) Results of a typical EKI inversion with Noise-Informed Covariance Estimation covariance estimation. (c) Recovered forcing, parameterized by an neural networks, trained with EKI (purple) and true forcing (orange). (d) Averages of the $n_x = 100$ mL'96 coordinates (error bars) and EKI-NN reconstructions (purple). (e) Standard deviations of the $n_x = 100$ mL'96 coordinates (error bars) and EKI-NN reconstructions (purple).

Figure 11a shows a Hovmöller diagram of the mL'96 model and illustrates the time evolution of all $n_x = 100$ coordinates as a function of time.

Due to the sinusoidal forcing, we can identify regions of chaotic dynamics (larger F_i) and regions with more predictable characteristics (smaller F_i).

Our goal is to recover the forcing F_i, $i = 1, \ldots, n_x$ from time-averaged data, which are the means and standard deviations of all n_x coordinates over a period of $T = 500$ time units ($2n_x = 200$ data points). The noise in the data are independent mean-zero Gaussians with standard deviations equal to 10% of that of the data points. The neural network that parameterizes the forcing is a feed-forward neural net with one input layer, one hidden layer and one output layer (Goodfellow et al., 2016). The total number of weights and biases in the network is 91, largely due to the size of the hidden layer, which we adjusted so that the neural network is expressive enough to capture the sinusoidal forcing.

EKI requires an initial ensemble which we generate using ideas from transfer learning. We draw n_e realizations of a smooth Gaussian process (Gaussian kernel, length scale is $\ell = 5$ (Rasmussen & Williams, 2005)) and then train a NN on each random function draw. Here, we use back-propagation, as is standard in simple function approximation tasks, because the NN is differentiable—the time-averaged data are not. The weights and biases of the NNs we obtain from training on random smooth functions represent the initial ensemble for our EKIs. This simple strategy works well for small ensembles (up to $n_e = 60$), but it leads to instabilities with EKIs with larger ensemble sizes. More sophisticated initialization may make it possible to run EKI with large n_e on this problem, but since our focus is on EKI and small ensemble sizes, we do not pursue initialization of NNs in EKI further.

A typical result we obtain with EKI and NICE is illustrated in Figures 11c–11e, which shows the recovered forcing (panel (c)) and data fits (panels (d) and (e)). The EKI can train the NN so that the mL'96 model with the

19422466, 2024, 9, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024MS004417 by Test, Wiley Online Library on [30/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

NN parameterization fits the data to within the assumed errors. Moreover, the recovered NN captures the sinusoidal variation of the forcing.

We now follow our usual procedure and compare EKIs of various flavors: (a) EKI with NICE; (b) EKI with ASCE; and (c) EKI with Ad.-PLC. All EKIs apply covariance estimation to $\hat{\mathbf{C}}_{gg}$ and $\hat{\mathbf{C}}_{xg}$, and we again adjust the tuning factor δ to be equal to 0.5 when estimating $\hat{\mathbf{C}}_{xg}$. The EKIs with NICE, ASCE or Ad.-PLC further inflate the covariance matrices with the same strategy as described in Section 4.4. We compare the above EKIs to a vanilla EKI, as well as to an EKI with inflation.

The results of our comparison are illustrated in Figure 11b, which shows the average RMSE of the various EKIs, computed over 10 independent experiments, each using different random perturbations during 30 iterations. The EKIs with NICE, ASCE or Ad.-PLC use an ensemble size $n_e = 40$ and we compare their performance to EKIs with or without inflation and ensemble sizes 40 and 60. The results we obtain in this example are in line with our earlier findings: NICE, ASCE and Ad.-PLC perform very similarly, reduce the error compared to inflated or vanilla EKIs and lead to a good fit to the data. Moreover, NICE, ASCE or Ad.-PLC result in similar errors as an inflated EKI with a larger ensemble size ($n_e = 60$). In summary, we can apply EKI to train a neural network that parameterizes a chaotic dynamical system, and covariance estimation methods such as NICE, ASCE or Ad.-PLC help with the computational efficiency of the inversion because they enable us to run the EKI with a small ensemble size.

5. Summary and Conclusions

We consider the problem of estimating a covariance matrix from a small number of samples in the context of Earth science applications. Our focus is on problems in which the correlation structure is *unknown*, because the problem of high-dimensional covariance estimation with a priori assumptions about the correlation structure is essentially solved (i.e., covariance localization in NWP).

A new method for covariance estimation, called NICE, is built on a single fundamental fact we know about estimating correlations: Small correlations are notoriously hard to compute, while it is relatively easy to compute large correlations. We translate this simple idea into an efficient and adaptive covariance estimation method that guarantees a symmetric PSD covariance estimate.

Adaptivity of NICE is achieved by (a) estimating a noise level for the correlation matrix; and (b) adjusting the correlation corrections so that the resulting correlation estimate is compatible with the noise level. We also used these ideas to design a few other adaptive covariance estimation methods: adaptive PLC (Ad.-PLC), adaptive localization (Ad.-Loc), adaptive soft-thresholding (Ad.-ST), and ASCE.

We compared our new covariance estimation methods to several other methods on a large set of numerical experiments with correlation structures that are not easy to anticipate or decipher. Our tests include cycling DA with a geomagnetic proxy model, geophysical inversion of field data, and the training of a feed-forward neural network with time-averaged data from a chaotic dynamical system. *All* new covariance estimation methods we created perform well on this diverse set of numerical tests and are similar in accuracy to related tuned methods, which speaks for the robustness of our approach to adaptive covariance estimation. NICE, however, has the advantage of guaranteeing a PSD covariance estimator at a low computational cost.

Data Availability Statement

The code and data used in this manuscript are available at (Vishny et al., 2024).

Acknowledgments
We thank Jeff Anderson of NCAR and an anonymous reviewer for their careful reviews that helped us improve the manuscript. DV was supported by the NSF through Grant AGS-2202991. MM is supported by the US Office of Naval Research (ONR) Grant N00014-21-1-2309. KG is supported by an appointment to the NASA Postdoctoral Program at Goddard Space Flight Center, administered by Oak Ridge Associated Universities under contract with NASA. EB is supported by the Foster and Coco Stanback Postdoctoral Fellowship, Office of Naval Research through Grant N00014-23-1-2654 and through a Department of Defense Vannevar Bush Faculty Fellowship awarded to Andrew Stuart. ORAD was supported by the generosity of Eric and Wendy Schmidt by recommendation of the Schmidt Futures program, National Science Foundation Grant AGS-1835860, the Defense Advanced Research Projects Agency (Agreement No. HR00112290030), the Heising-Simons Foundation, Audi Environmental Foundation, the Cisco Foundation and by the US Office of Naval Research (ONR) Grant N00014-23-1-2654. DH is supported by the US Office of Naval Research (ONR) Grant N0001422WX00451. We thank Dr. Dani Blatter of Las Positas College for help and advice with the MT data, the MT forward modeling code, and the geophysical inversion.

References

Agapiou, S., Papaspiliopoulos, O., Sanz-Alonso, D., & Stuart, A. M. (2017). Importance sampling: Intrinsic dimension and computational cost. *Statistical Science*, *32*(3), 405–431. https://doi.org/10.1214/17-sts611

Al Ghattas, O., & Sanz-Alonso, D. (2022). Non-asymptotic analysis of ensemble Kalman updates: Effective dimension and localization. arXiv:2208.03246.

Anderson, J. L. (2001). An ensemble adjustment Kalman filter for data assimilation. *Monthly Weather Review*, *129*(12), 2884–2903. https://doi.org/10.1175/1520-0493(2001)129<2884:aeakff>2.0.co;2

Anderson, J. L. (2012). Localization and sampling error correction in ensemble Kalman filter data assimilation. *Monthly Weather Review*, *140*(7), 2359–2371. https://doi.org/10.1175/mwr-d-11-00013.1

Anderson, J. L., & Lei, L. (2013). Empirical localization of observation impact in ensemble Kalman filters. *Monthly Weather Review*, *141*(11), 4140–4153. https://doi.org/10.1175/mwr-d-12-00330.1

19422466, 2024, 9, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024MS004417 by Test, Wiley Online Library on [30/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

19422466, 2024, 9, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024MS004417 by Test, Wiley Online Library on [30/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Anzengruber, S. W., & Ramlau, R. (2009). Morozov's discrepancy principle for Tikhonov-type functionals with nonlinear operators. *Inverse Problems*, *26*(2), 025001. https://doi.org/10.1088/0266-5611/26/2/025001

Bannister, R. N. (2017). A review of operational methods of variational and ensemble-variational data assimilation. *Quarterly Journal of the Royal Meteorological Society*, *143*(703), 607–633. https://doi.org/10.1002/qj.2982

Bickel, P., & Levina, E. (2008). Covariance regularization by thresholding. *Annals of Statistics*, *36*(6), 2577–2604. https://doi.org/10.1214/08-AOS600

Bickel, P., & Lindner, M. (2012). Approximating the inverse of banded matrices by banded matrices with applications to probability and statistics. *Theory of Probability and Its Applications*, *56*(1), 1–20. https://doi.org/10.1137/S0040585X97985224

Bieli, M., Dunbar, O. R. A., de Jong, E. K., Jaruga, A., Schneider, T., & Bischoff, T. (2022). An efficient Bayesian approach to learning droplet collision kernels: Proof of concept using "Cloudy," a new n-moment bulk microphysics scheme. *Journal of Advances in Modeling Earth Systems*, *14*(8), e2022MS002994. https://doi.org/10.1029/2022ms002994

Bishop, C. H., & Hodyss, D. (2007). Flow-adaptive moderation of spurious ensemble correlations and its use in ensemble-based data assimilation. *Quarterly Journal of the Royal Meteorological Society*, *133*(629), 2029–2044. https://doi.org/10.1002/qj.169

Bishop, C. H., & Hodyss, D. (2009a). Ensemble covariances adaptively localized with ECO-RAP. Part 1: Tests on simple error models. *Tellus A: Dynamic Meteorology and Oceanography*, *61*(1), 84–96. https://doi.org/10.1111/j.1600-0870.2007.00371.x

Bishop, C. H., & Hodyss, D. (2009b). Ensemble covariances adaptively localized with ECO-RAP. Part 2: A strategy for the atmosphere. *Tellus A: Dynamic Meteorology and Oceanography*, *61*(1), 97–111. https://doi.org/10.1111/j.1600-0870.2007.00372.x

Bishop, C. H., & Hodyss, D. (2011). Adaptive ensemble covariance localization in ensemble 4D-VAR state estimation. *Monthly Weather Review*, *139*(4), 1241–1255. https://doi.org/10.1175/2010MWR3403.1

Bishop, C. H., Whitaker, J. S., & Lei, L. (2017). Gain form of the ensemble transform Kalman filter and its relevance to satellite data assimilation with model space ensemble covariance localization. *Monthly Weather Review*, *145*(11), 4575–4592. https://doi.org/10.1175/mwr-d-17-0102.1

Blatter, D., Key, K., Ray, A., Gustafson, C., & Evans, R. (2019). Bayesian joint inversion of controlled source electromagnetic and magnetotelluric data to image freshwater aquifer offshore New Jersey. *Geophysical Journal International*, *218*(3), 1822–1837. https://doi.org/10.1093/gji/ggz253

Blatter, D., Morzfeld, M., Key, K., & Constable, S. (2022a). Uncertainty quantification for regularized inversion of electromagnetic geophysical data – Part II: Application in 1-D and 2-D problems. *Geophysical Journal International*, *231*(2), 1075–1095. https://doi.org/10.1093/gji/ggac242

Blatter, D., Morzfeld, M., Key, K., & Constable, S. (2022b). Uncertainty quantification for regularized inversion of electromagnetic geophysical data-Part I: Motivation and theory. *Geophysical Journal International*, *231*(2), 1057–1074. https://doi.org/10.1093/gji/ggac241

Bocquet, M. (2016). Localization and the iterative ensemble Kalman smoother. *Quarterly Journal of the Royal Meteorological Society*, *142*(695), 1075–1089. https://doi.org/10.1002/qj.2711

Bocquet, M., & Sakov, P. (2014). An iterative ensemble Kalman smoother. *Quarterly Journal of the Royal Meteorological Society*, *140*(682), 1521–1535. https://doi.org/10.1002/qj.2236

Buehner, M. (2005). Ensemble-derived stationary and flow-dependent background-error covariances: Evaluation in a quasi-operational NWP setting. *Quarterly Journal of the Royal Meteorological Society*, *131*(607), 1013–1043. https://doi.org/10.1256/qj.04.15

Buehner, M. (2012). Evaluation of a spatial/spectral covariance localization approach for atmospheric data assimilation. *Monthly Weather Review*, *140*(2), 617–636. https://doi.org/10.1175/MWR-D-10-05052.1

Buehner, M., Mourneau, J., & Charette, C. (2013). Four-dimensional ensemble-variational data assimilation for global deterministic weather prediction. *Nonlinear Processes in Geophysics*, *20*(5), 669–682. https://doi.org/10.5194/npg-20-669-2013

Buehner, M., & Shlyaeva, A. (2015). Scale-dependent background-error covariance localisation. *Tellus A: Dynamic Meteorology and Oceanography*, *67*(1), 28027. https://doi.org/10.3402/tellusa.v67.28027

Burgers, G., Leeuwen, P. V., & Evensen, G. (1998). Analysis scheme in the ensemble Kalman filter. *Monthly Weather Review*, *126*(6), 1719–1724. https://doi.org/10.1175/1520-0493(1998)126<1719:asitek>2.0.co;2

Campbell, W., Bishop, C., & Hodyss, D. (2010). Vertical covariance localization for satellite radiances in ensemble Kalman filters. *Monthly Weather Review*, *138*(1), 282–290. https://doi.org/10.1175/2009mwr3017.1

Chada, N. K., Chen, Y., & Sanz-Alonso, D. (2021). Iterative ensemble Kalman methods: A unified perspective with some new variants. *Foundations of Data Science*, *3*(3), 331–369. https://doi.org/10.3934/fods.2021011

Chada, N. K., Stuart, A. M., & Tong, X. T. (2020). Tikhonov regularization within ensemble Kalman inversion. *SIAM Journal on Numerical Analysis*, *58*(2), 1263–1294. https://doi.org/10.1137/19m1242331

Chada, N. K., & Tong, X. T. (2022). Convergence acceleration of ensemble Kalman inversion in nonlinear settings. *Mathematics of Computation*, *91*(335), 1247–1280.

Chen, Y., & Oliver, D. (2010). Cross-covariances and localization for EnKF in multiphase flow data assimilation. *Computational Geosciences*, *14*(4), 579–601. https://doi.org/10.1007/s10596-009-9174-6

Chen, Y., & Oliver, D. (2013). Levenberg-Marquardt forms of the iterative ensemble smoother for efficient history matching and uncertainty quantification. *Computational Geosciences*, *17*(4), 689–703. https://doi.org/10.1007/s10596-013-9351-5

Chen, Y., & Oliver, D. (2017). Localization and regularization for iterative ensemble smoothers. *Computational Geosciences*, *21*(1), 13–30. https://doi.org/10.1007/s10596-016-9599-7

Chevrotiére, M. D. L., & Harlim, J. (2017). A data-driven method for improving the correlation estimation in serial ensemble Kalman filters. *Monthly Weather Review*, *145*(3), 985–1001. https://doi.org/10.1175/MWR-D-16-0109.1

Chorin, A. J., & Morzfeld, M. (2013). Conditions for successful data assimilation. *Journal of Geophysical Research: Atmospheres*, *118*(20), 11522–11533. https://doi.org/10.1002/2013JD019838

Cleary, E., Garbuno-Inigo, A., Lan, S., Schneider, T., & Stuart, A. M. (2021). Calibrate, emulate, sample. *Journal of Computational Physics*, *424*, 109716. https://doi.org/10.1016/j.jcp.2020.109716

Constable, S., Parker, R., & Constable, C. (1987). Occam's inversion: A practical algorithm for generating smooth models from electromagnetic sounding data. *Geophysics*, *3*(52), 289–300. https://doi.org/10.1190/1.1442303

Dunbar, O. R. A., Lopez-Gomez, I., Garbuno-Iñigo, A., Huang, D. Z., Bach, E., & Wu, J.-L. (2022). EnsembleKalmanProcesses.jl: Derivative-free ensemble-based model calibration. *Journal of Open Source Software*, *7*(80), 4869. https://doi.org/10.21105/joss.04869

Emerick, A. A., & Reynolds, A. (2011). Combining sensitivities and prior information for covariance localization in the ensemble Kalman filter for petroleum reservoir applications. *Computational Geosciences*, *15*(2), 251–269. https://doi.org/10.1007/s10596-010-9198-y

Emerick, A. A., & Reynolds, A. C. (2013). Ensemble smoother with multiple data assimilation. *Computers & Geosciences*, *55*, 3–15. https://doi.org/10.1016/j.cageo.2012.03.011

Evensen, G. (1994). Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research*, *99*(C5), 10143–10162. https://doi.org/10.1029/94JC00572

Evensen, G. (2009). *Data assimilation: The ensemble Kalman filter* (2nd ed.). Springer.

Flowerdew, J. (2015). Towards a theory of optimal localisation. *Tellus A: Dynamic Meteorology and Oceanography*, *67*(1), 25257. https://doi.org/10.3402/tellusa.v67.25257

Friedman, J., Hastie, T., & Tibshirani, R. (2007). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, *9*(3), 432–441. https://doi.org/10.1093/biostatistics/kxm045

Furrer, R., & Bengtsson, T. (2007). Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants. *Journal of Multivariate Analysis*, *98*(2), 227–255. https://doi.org/10.1016/j.jmva.2006.08.003

Gaspari, G., & Cohn, S. (1999). Construction of correlation functions in two and three dimensions. *Quarterly Journal of the Royal Meteorological Society*, *125*(554), 723–757. https://doi.org/10.1002/qj.49712555417

Gharamti, M. E., Reader, K., & Anderson, J. L. (2019). Comparing adaptive prior and posterior inflation for ensemble filters using an atmospheric general circulation model. *Monthly Weather Review*, *147*, 2535–2553.

Gilpin, S., Matsuo, T., & Cohn, S. E. (2023). A generalized, compactly supported correlation function for data assimilation applications. *Quarterly Journal of the Royal Meteorological Society*, *149*(754), 1953–1989. https://doi.org/10.1002/qj.4490

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Guillot, D., & Rajaratnam, B. (2015). Functions preserving positive definiteness for sparse matrices. *Transactions of the American Mathematical Society*, *367*(1), 627–649. https://doi.org/10.1090/s0002-9947-2014-06183-7

Gustafson, C., Key, K., & Evans, R. (2019). Aquifer systems extending far offshore on the U.S. Atlantic margin. *Scientific Reports*, *9*(1), 8709. https://doi.org/10.1038/s41598-019-44611-7

Gwirtz, K., Morzfeld, M., Kuang, W., & Tangborn, A. (2021). A testbed for geomagnetic data assimilation. *Geophysical Journal International*, *227*(3), 2180–2203. https://doi.org/10.1093/gji/ggab327

Hamill, T. M., & Snyder, C. (2000). A hybrid ensemble Kalman filter–3D variational analysis scheme. *Monthly Weather Review*, *128*(8), 2905–2919. https://doi.org/10.1175/1520-0493(2000)128<2905:ahekfv>2.0.co;2

Hamill, T. M., Whitaker, J. S., Anderson, J. L., & Snyder, C. (2009). Comments on "sigma-point Kalman filter data assimilation methods for strongly nonlinear systems". *Journal of the Atmospheric Sciences*, *66*(11), 3498–3500. https://doi.org/10.1175/2009jas3245.1

Harty, T., Morzfeld, M., & Snyder, C. (2021). Eigenvector-spatial localisation. *Tellus A: Dynamic Meteorology and Oceanography*, *73*(1), 1–18. https://doi.org/10.1080/16000870.2021.1903692

Hodyss, D., Bishop, C. H., & Morzfeld, M. (2016). To what extent is your data assimilation scheme designed to find the posterior mean, the posterior mode or something else? *Tellus*, *68*, 1–17. https://doi.org/10.3402/tellusa.v68.30625

Hodyss, D., Campbell, W. F., & Whitaker, J. S. (2016). Observation-dependent posterior inflation for the ensemble Kalman filter. *Monthly Weather Review*, *144*(7), 2667–2684. https://doi.org/10.1175/mwr-d-15-0329.1

Hodyss, D., & Morzfeld, M. (2023). How sampling errors in covariance estimates cause bias in the Kalman gain and impact ensemble data assimilation. *Monthly Weather Review*, *151*(9), 2413–2426. https://doi.org/10.1175/mwr-d-23-0029.1

Horn, R. A., & Johnson, C. R. (1991). *Topics in matrix analysis*. Cambridge University Press. https://doi.org/10.1017/CBO9780511840371

Houtekamer, P. L., & Mitchell, H. L. (1998). Data assimilation using an ensemble Kalman filter technique. *Monthly Weather Review*, *126*(3), 796–811. https://doi.org/10.1175/1520-0493(1998)126<0796:dauaek>2.0.co;2

Houtekamer, P. L., & Mitchell, H. L. (2001). A sequential ensemble Kalman filter for atmospheric data assimilation. *Monthly Weather Review*, *129*(1), 123–137. https://doi.org/10.1175/1520-0493(2001)129<0123:asekff>2.0.co;2

Huang, D., Huang, J., Reich, S., & Stuart, A. (2022). Efficient derivative-free Bayesian inference for large-scale inverse problems. *Inverse Problems*, *38*(12), 125006. https://doi.org/10.1088/1361-6420/ac99fa

Iglesias, M. A., Law, K. J., & Stuart, A. M. (2013). Ensemble Kalman methods for inverse problems. *Inverse Problems*, *29*(4), 045001. https://doi.org/10.1088/0266-5611/29/4/045001

Khare, K., Oh, S.-Y., Rahman, S., & Rajaratnam, B. (2019). A scalable sparse Cholesky based approach for learning high-dimensional covariance matrices in ordered data. *Machine Learning*, *108*(12), 2061–2086. https://doi.org/10.1007/s10994-019-05810-5

Kovachki, N., & Stuart, A. (2019). Ensemble Kalman inversion: A derivative-free technique for machine learning tasks. *Inverse Problems*, *35*(9), 095005. https://doi.org/10.1088/1361-6420/ab1c3a

Kuhl, D. D., Rosmond, T. E., Bishop, C. H., McLay, J., & Baker, N. L. (2013). Comparison of hybrid ensemble/4DVar and 4DVar within the NAVDAS-AR data assimilation framework. *Monthly Weather Review*, *141*(8), 2740–2758. https://doi.org/10.1175/mwr-d-12-00182.1

Lee, Y. (2021a). l_p regularization for ensemble Kalman inversion. *SIAM Journal on Scientific Computing*, *43*(5), A3417–A3437. https://doi.org/10.1137/20M1365168

Lee, Y. (2021b). Sampling error correction in ensemble Kalman inversion. *arXiv*. https://doi.org/10.48550/ARXIV.2105.11341

Lorenc, A. (2003). The potential of the ensemble Kalman filter for NWP – A comparison with 4D-Var. *Quarterly Journal of the Royal Meteorological Society*, *129*(595), 3183–3203. https://doi.org/10.1256/qj.02.132

Lorenc, A. (2017). Improving ensemble covariances in hybrid variational data assimilation without increasing ensemble size. *Quarterly Journal of the Royal Meteorological Society*, *143*(703), 1062–1072. https://doi.org/10.1002/qj.2990

Lorenz, E. (1996). Predictability: A problem partly solved. *Proceedings of the ECMWF Seminar on Predictability*, *1*, 1–18.

Luk, E., Bach, E., Baptista, R., & Stuart, A. (2024). Learning optimal filters using variational inference (No. arXiv:2406.18066). *arXiv*.

Luo, X., Bhakta, T., & Nævdal, G. (2018). Correlation-based adaptive localization with applications to ensemble-based 4D-seismic history matching. *SPE Journal*, *23*(02), 396–427. https://doi.org/10.2118/185936-pa

Ménétrier, B., Montmerle, T., Michel, Y., & Berre, L. (2015). Linear filtering of sample covariances for ensemble-based data assimilation. Part I: Optimality criteria and application to variance filtering and covariance localization. *Monthly Weather Review*, *143*(5), 1622–1643. https://doi.org/10.1175/MWR-D-14-00157.1

Miyoshi, T., & Kondo, K. (2013). A multi-scale localization approach to an ensemble Kalman filter. *SOLA (Scientific Online Letters on the Atmosphere)*, *9*(0), 170–173. https://doi.org/10.2151/sola.2013-038

Morozov, V. (1984). *Methods for solving incorrectly posed problems*. Springer.

Morzfeld, M., & Hodyss, D. (2023). A theory for why even simple covariance localization is so useful in ensemble data assimilation. *Monthly Weather Review*, *151*(3), 717–736. https://doi.org/10.1175/mwr-d-22-0255.1

Morzfeld, M., Tong, X. T., & Marzouk, Y. M. (2019). Localization for MCMC: Sampling high-dimensional posterior distributions with local structure. *Journal of Computational Physics*, *380*, 1–28. https://doi.org/10.1016/j.jcp.2018.12.008

Ott, E., Hunt, B., Szunyogh, I., Zimin, A., Kostelich, E., Corazza, M., et al. (2004). A local ensemble Kalman filter for atmospheric data assimilation. *Tellus*, *56*(5), 415–428. https://doi.org/10.1111/j.1600-0870.2004.00076.x

19422466, 2024, 9, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024MS004417 by Test, Wiley Online Library on [30/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

19422466, 2024, 9, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024MS004417 by Test, Wiley Online Library on [30/08/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Poterjoy, J., & Zhang, F. (2015). Systematic comparison of four-dimensional data assimilation methods with and without the tangent linear model using hybrid background error covariance: E4DVar versus 4DEnVar. *Monthly Weather Review*, *143*(5), 1601–1621. https://doi.org/10.1175/mwr-d-14-00224.1

Pourahmadi, M. (2011). Covariance estimation: The GLM and regularization perspectives. *Statistical Science*, *26*(3), 369–387. https://doi.org/10.1214/11-STS358

Rasmussen, C. E., & Williams, C. (2005). *Gaussian processes for machine learning (adaptive computation and machine learning)*. The MIT Press.

Schillings, C., & Stuart, A. M. (2017). Analysis of the ensemble Kalman filter for inverse problems. *SIAM Journal on Numerical Analysis*, *55*(3), 1264–1290. https://doi.org/10.1137/16m105959x

Schillings, C., & Stuart, A. M. (2018). Convergence analysis of ensemble Kalman inversion: The linear, noisy case. *Applicable Analysis*, *97*(1), 107–123. https://doi.org/10.1080/00036811.2017.1386784

Schneider, T., Leung, L. R., & Wills, R. C. J. (2024). Opinion: Optimizing climate models with process-knowledge, resolution, and AI. *EGUsphere*, *2024*, 1–26. https://doi.org/10.5194/egusphere-2024-20

Schneider, T., Stuart, A. M., & Wu, J.-L. (2021). Learning stochastic closures using ensemble Kalman inversion. *Transactions of Mathematics and Its Applications*, *5*(1), tnab003. https://doi.org/10.1093/imatrm/tnab003

Schoenberg, I. J. (1942). Positive definite functions on spheres. *Duke Mathematical Journal*, *9*(1), 96–108. https://doi.org/10.1215/S0012-7094-42-00908-6

Schur, J. (1911). Bemerkungen zur Theorie der beschränkten Bilinearformen mit unendlich vielen Veränderlichen. *Journal für die Reine und Angewandte Mathematik*, *140*, 1–28. https://doi.org/10.1515/crll.1911.140.1

Talagrand, O., & Courtier, P. (1987). Variational assimilation of meteorological observations with the adjoint vorticity equation. I: Theory. *Quarterly Journal of the Royal Meteorological Society*, *113*(478), 1311–1328. https://doi.org/10.1002/qj.49711347812

Tippett, M., Anderson, J., Bishop, C., Hamill, T., & Whitaker, J. (2003). Ensemble square root filters. *Monthly Weather Review*, *131*(7), 1485–1490. https://doi.org/10.1175/1520-0493(2003)131⟨1485:ESRF⟩2.0.CO;2

Tong, X., & Morzfeld, M. (2023). Localized ensemble Kalman inversion. *Inverse Problems*, *39*(6), 064002. https://doi.org/10.1088/1361-6420/accb08

Vishny, D., Morzfeld, M., & Gwirtz, K. (2024). Matlab code & data for "High-dimensional covariance estimation from a small number of samples". *Zenodo*. https://doi.org/10.5281/zenodo.12701351

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press. https://doi.org/10.1017/9781108627771

Ward, S. H., & Hohmann, G. W. (2012). Electromagnetic theory for geophysical applications. In *Electromagnetic methods in applied geophysics: Volume 1, theory* (pp. 130–311). Society of Exploration Geophysicists. https://doi.org/10.1190/1.9781560802631.ch4

Whitaker, J., & Hamill, T. (2012). Evaluating methods to account for system errors in ensemble data assimilation. *Monthly Weather Review*, *140*(9), 3078–3089. https://doi.org/10.1175/mwr-d-11-00276.1

Xue, L., Ma, S., & Zou, H. (2012). Positive-definite ℓ_1-penalized estimation of large covariance matrices. *Journal of the American Statistical Association*, *107*(500), 1480–1491. https://doi.org/10.1080/01621459.2012.725386

Zhang, F., Zhang, M., & Hansen, J. A. (2009). Coupling ensemble Kalman filter with four-dimensional variational data assimilation. *Advances in Atmospheric Sciences*, *26*, 1–8. https://doi.org/10.1007/s00376-009-0001-8

Zhen, Y., & Zhang, F. (2014). A probabilistic approach to adaptive covariance localization for serial ensemble square root filters. *Monthly Weather Review*, *142*(12), 4499–4518. https://doi.org/10.1175/MWR-D-13-00390.1