Non-sequential Ensemble Kalman Filtering using Distributed Arrays

Cedric Travelletti

June 23, 2023

Joint work with: D. Ginsbourger, J. Franke, S. Brönnimann and V. Valler (Oeschger Center for Climate Change Research)
C. Travelletti acknowledges funding from the Swiss National Science Fundation under project nr. 178858.

Motivation

Ensemble Kalman Filtering is one of the most popular data assimilation (DA) technique nowadays.

- Works on large-scale problems (no need to work with large covariance matrix).
- Allows flexible prior specification (simulations, ...).

Motivation

Ensemble Kalman Filtering is one of the most popular data assimilation (DA) technique nowadays.

- Works on large-scale problems (no need to work with large covariance matrix).
- Allows flexible prior specification (simulations, ...).

However ...

Motivation

Ensemble Kalman Filtering is one of the most popular data assimilation (DA) technique nowadays.

- Works on large-scale problems (no need to work with large covariance matrix).
- Allows flexible prior specification (simulations, ...).

However ...

In its most commonly used form (**localized**, **sequential**, **square root** filter), it is **WRONG**.

Motivation

Ensemble Kalman Filtering is one of the most popular data assimilation (DA) technique nowadays.

- Works on large-scale problems (no need to work with large covariance matrix).
- Allows flexible prior specification (simulations, ...).

However ...

In its most commonly used form (**localized**, **sequential**, **square root** filter), it is **WRONG**.

Will show recent computational techniques (distributed arrays) can fix it.

Section 1

Introduction: the Kalman Filter and its Variants

Kalman Filter: Setting

Unknown (true) state vector

$$\boldsymbol{\psi}_t^* \in \mathbb{R}^m$$

Dynamics

$$\boldsymbol{\psi}_{t+1}^* = \mathcal{F}_t \boldsymbol{\psi}_t^* + \boldsymbol{\delta}_t,\ \boldsymbol{\delta}_t \sim \mathcal{N}(0, \boldsymbol{\Delta}_t),$$

where $\mathcal{F}_t : \mathbb{R}^m \to \mathbb{R}^m$ is linear and $\boldsymbol{\delta}$ is random model noise.

Observations

At each time step t, we are given observations:

$$\boldsymbol{y}_t = \boldsymbol{G}_t \boldsymbol{\psi}_t^* + \boldsymbol{\epsilon}_t,\ \boldsymbol{\epsilon}_t \sim \mathcal{N}(0, \boldsymbol{E}_t), \tag{1}$$

where $\boldsymbol{G}_t : \mathbb{R}^m \to \mathbb{R}^{n_t}$ is linear and $\boldsymbol{\epsilon}_t$ is a random observation noise vector.

Kalman Filter: Update Equations

Bayesian approach: Start with prior $\mathbf{\Psi}_0 \sim \mathcal{N}(\boldsymbol{\mu}_0, \mathbf{\Sigma}_0)$ on initial state ψ_0.

- Approximate ψ_t by distribution conditionally on the dynamics and the observations up to t.
- **Forecast Step:**

$$\boldsymbol{\mu}_t^{\boldsymbol{f}} = \mathcal{F}_t \boldsymbol{\mu}_{t-1},\ \mathbf{\Sigma}_t^{\boldsymbol{f}} = \mathcal{F}_t \mathbf{\Sigma}_{t-1} \mathcal{F}_t^T + \mathbf{\Delta}_t.$$

- **Update step:** Conditional distribution of the state is Gaussian with mean and covariance:

$$\boldsymbol{\mu}_t = \boldsymbol{\mu}_t^{\boldsymbol{f}} + \boldsymbol{K}_t \left(\boldsymbol{y}_t - \boldsymbol{G}_t \boldsymbol{\mu}_t^{\boldsymbol{f}} \right),\ \mathbf{\Sigma}_t = \mathbf{\Sigma}_t^{\boldsymbol{f}} - \boldsymbol{K}_t \boldsymbol{G}_t \mathbf{\Sigma}_t^{\boldsymbol{f}},$$

with $\boldsymbol{K}_t = \mathbf{\Sigma}_t^{\boldsymbol{f}} \boldsymbol{G}_t^T \left(\boldsymbol{G}_t \mathbf{\Sigma}_t^{\boldsymbol{f}} \boldsymbol{G}_t^T + \boldsymbol{E}_t \right)^{-1}$

Tedious for high-dimensional state spaces.

Ensemble Kalman Filter [Eve94, Eve03, E+09]

Idea

Replace conditional distribution by **ensemble** of i.i.d. samples of it

$$\boldsymbol{\psi}_t^{(1)}, \dots, \boldsymbol{\psi}_t^{(p)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$$

- Avoids updating large covariance matrices.
- Flexible "prior" specification by providing a starting ensemble.

Original EnKF formulation presents two challenges:

1. Observations have to be randomly perturbed in order for update not to underestimate the variance.
2. Need to estimate the covariance from a limited set of ensemble members.

Ensemble Kalman Filter: Square Root Version [WH02, TAB+03]

Perturbed observations be avoided by using modified deterministic updates.

Algorithm

Update ensemble mean $\bar{\boldsymbol{\psi}}_t$ and deviations $\boldsymbol{\psi}_t^{(i)'} := \boldsymbol{\psi}_t^{(i)} - \bar{\boldsymbol{\psi}}_t$ via:

$$\bar{\boldsymbol{\psi}}_t = \bar{\boldsymbol{\psi}}_t^f + \hat{\boldsymbol{K}} \left(\boldsymbol{y}_t - \boldsymbol{G} \bar{\boldsymbol{\psi}}_t^f \right), \tag{2}$$

$$\boldsymbol{\psi}_t^{(i)} = \boldsymbol{\psi}_t^{f(i)} - \tilde{\boldsymbol{K}}_t \boldsymbol{G} \boldsymbol{\psi}_t^{f(i)}, \tag{3}$$

$$\tilde{\boldsymbol{K}}_t = \hat{\boldsymbol{\Sigma}}_t^{\boldsymbol{f}} \boldsymbol{G}_t^T \left(\sqrt{\boldsymbol{G}_t \hat{\boldsymbol{\Sigma}}_t^{\boldsymbol{f}} \boldsymbol{G}_t^T + \boldsymbol{E}_t} \right)^{-1} \left(\sqrt{\boldsymbol{G}_t \hat{\boldsymbol{\Sigma}}_t^{\boldsymbol{f}} \boldsymbol{G}_t^T + \boldsymbol{E}_t} + \sqrt{\boldsymbol{E}_t} \right)^{-1},$$

where $\hat{\boldsymbol{\Sigma}}^{\boldsymbol{f}}$ is an estimator of $\boldsymbol{\Sigma}^{\boldsymbol{f}}$ and $\tilde{\boldsymbol{K}}$ is the Kalman gain with $\boldsymbol{\Sigma}^{\boldsymbol{f}}$ replaced by $\hat{\boldsymbol{\Sigma}}^{\boldsymbol{f}}$.

Overperforms EnKF with perturbed observations for small ensemble sizes.

Localization

Problem

Bare empirical sample covariance as estimator of $\mathbf{\Sigma}_t^{\boldsymbol{f}}$ suffers from undersampling errors.

- Need to regularize estimate of the covariance.

Localization

Estimate using empirical covariance tapered by an SPD matrix:

$$\hat{\mathbf{\Sigma}}_t^{\boldsymbol{f}} = \mathrm{Cov}\left(\left(\boldsymbol{\psi}_t^{f(i)}\right)_{i=1,\dots,p}\right) \circ \boldsymbol{\rho}.$$

In spatial problems, localization matrix built from some kernel function.

Problem: In practice, data has to be assimilated **sequentially**, but the localized update equations are wrong in that setting.

Section 2

Distributed, Non-Sequential Ensemble Kalman Filtering

Dask Distributed Arrays

Dask library provide easy interface to **distributed arrays** (www.dask.org).

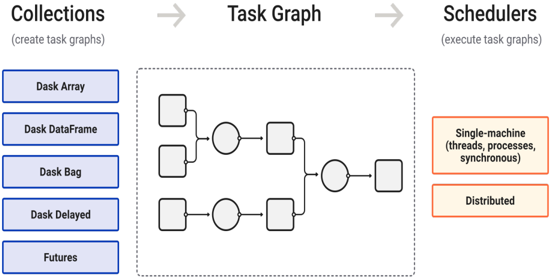


Figure: Dask workflow overview (courtesy *dask.org* under BSD 3-clause licence)

- high-level, numpy-like interface
- arrays distributed in cluster memory under the hood
- lazy evaluation via task graphs

Easy to scale ML workflows (Ensemble Kalman filter, ...).

EnSRF and Dask: Lazy, Distributed SVD

Halko's algorithm [HMT11] provides fast rank k approximate SVD. Lazy, distributed implementation in `dask.array.linalg.compressed_svd`.

Can be used to compute approximate inverses and square roots of covariance estimate:

$$\hat{\boldsymbol{\Sigma}}^{-1} \approx [u_1, \ldots, u_k] \begin{pmatrix} 1/\lambda_1 & & \\ & \ddots & \\ & & 1/\lambda_k \end{pmatrix} [u_1, \ldots, u_k]^T$$

$$\sqrt{\hat{\boldsymbol{\Sigma}}} \approx [u_1, \ldots, u_k] \begin{pmatrix} \sqrt{\lambda_1} & & \\ & \ddots & \\ & & \sqrt{\lambda_k} \end{pmatrix} [u_1, \ldots, u_k]^T,$$

where λ_i and $u_i, i = 1, \ldots, k$ are the k-largest (approximate) singular values and corresponding left singular vectors.

In a lazy setting, this means that one only has to store $2k(m+1)$ values.

Lazy, Distributed SVD and EnSRF

Distributed (approximate) SVD allows to run all-at-once (aao) EnSRF on large datasets.

Algorithm 1 Distributed EnSRF update

Require: Ensemble $\boldsymbol{\psi}_t^{(1)}, \dots, \boldsymbol{\psi}_t^{(p)}$, observation operator $\boldsymbol{G}_t$ and observed data $\boldsymbol{y}_t$ SVD cutoff rank k.

Ensure: Updated ensemble $\boldsymbol{\psi}_t^{(1)}, \dots, \boldsymbol{\psi}_t^{(p)}$.

Build localized estimated covariance $\hat{\boldsymbol{\Sigma}}_t$ in distributed memory.

$(\lambda_i, u_i)_{i=1,\dots,k} \leftarrow \textsc{DistributedSVD}(\hat{\boldsymbol{\Sigma}}_t, \mathrm{rank} = k)$

$\hat{\boldsymbol{\Sigma}}_t^{-1} \leftarrow \textsc{ApproximateInverse}((\lambda_i, u_i)_{i=1,\dots,k})$

$\sqrt{\hat{\boldsymbol{\Sigma}}_t} \leftarrow \textsc{ApproximateSqrt}((\lambda_i, u_i)_{i=1,\dots,k})$

$(\boldsymbol{\psi}_t^{(i)})_{i=1,\dots,p} \leftarrow \textsc{KalmanUpdate}((\boldsymbol{\psi}_t^{(i)})_{i=1,\dots,p}, \hat{\boldsymbol{\Sigma}}_t^{-1}, \sqrt{\hat{\boldsymbol{\Sigma}}_t})$ ▷ Update using Eq. (3)

Section 3

Sequential Vs All-at-once: Experimental Comparison

Experimental Testbed

Compare *seq* and *aao* EnSRF on two offline data assimilation problems:

1. Gaussian process regression problem (controlled environment).
2. Paleoclimate reconstruction problem (real data).

Performances of the assimilation schemes compared using:

- RMSE on reconstruction using updated mean.
- Reduction of error skill score [ME89].
  - widely used in climatology
  - aggregated over time
  - only considers point prediction
- Energy score multivariate scoring rule [GSG^{+}08].
  - proper scoring rule [GR07] for probabilitsic multivariate forecasts
  - considers full updated ensemble

GP Regression Task

Ground truth sampled from a Matérn GP $Z \sim \text{Gp}(0, k_{\nu=3/2,\lambda=0.1})$ with unit variance on unit square $[0,1]^2$.

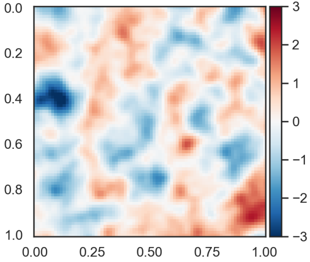
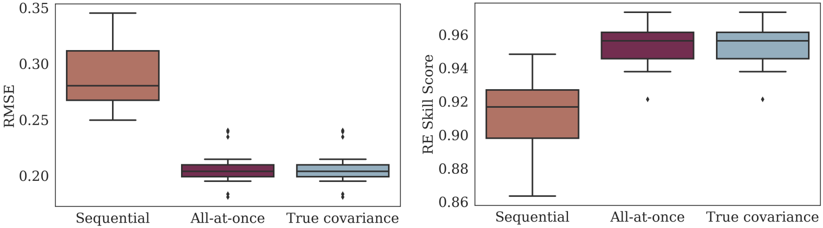


Figure: Example of ground truth sampled from the GP model.

- Starting ensemble sampled from model (well-specified).
- Assimilate data at 500 randomly chosen locations ($\sigma_\epsilon = 0.01$).
- Localize using $k_{\nu=3/2,\lambda=0.2}$ (undersmoothing).

Resample ground truth and repeat 50 times.



(a) RMSE (b) RE skill score (median)

Figure: Comparison of the distributions of the RMSE and RE score.

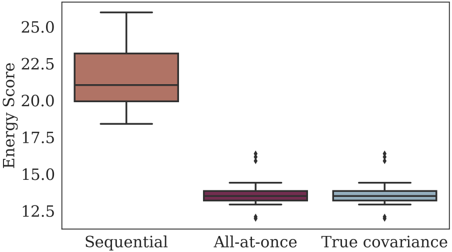
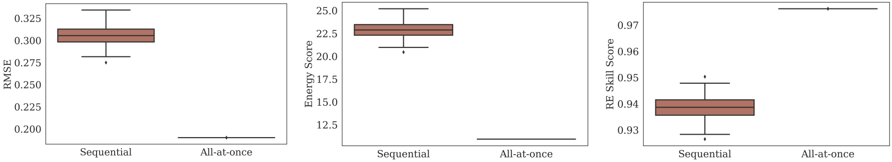


Figure: Comparison of the enery score for the different assimilation methods.

Ordering Dependency

It is well-known in the community that results of **localized** and **sequential** EnSRF depend on **observations ordering**.



(a) RMSE (b) Energy score (c) RE skill score

Figure: Distributions of acuracy metrics for different observation orderings (sequential).

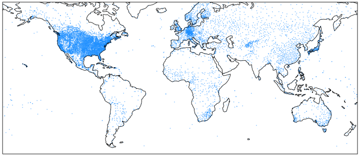
Dependence on observation ordering can have effect of up to 3-5%.

- First study of ordering dependency since small ($n = 40$) study of [Ner15].

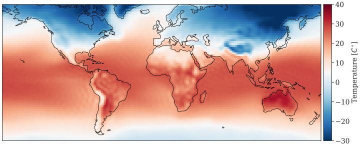
Climate Reconstruction Problem

Task

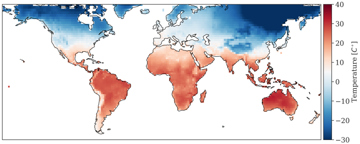
Reconstruct state-of-the art climatology over the 1960-1980 period by blending **sparse station data** with climate **simulations** that include **known external forcings** (solar irradiance, volcanic activity, greenhouse gas concentrations, ...).



(a) Station data



(b) Prior mean



(c) Reference

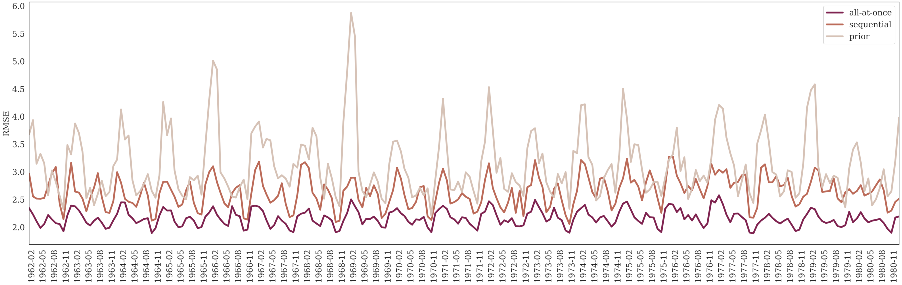
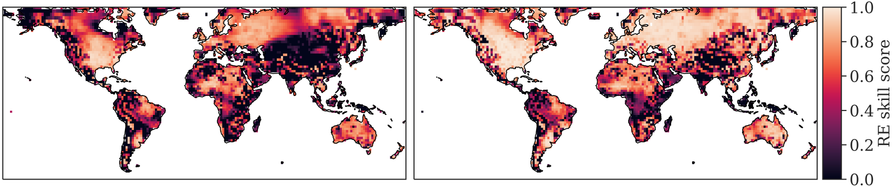


Figure: Root mean square error of the reconstruction for different assimilation schemes.



(a) sequential

(b) all-at-once

Figure: RE skill scores for reconstruction of the reference dataset over the 1960-1980 period.

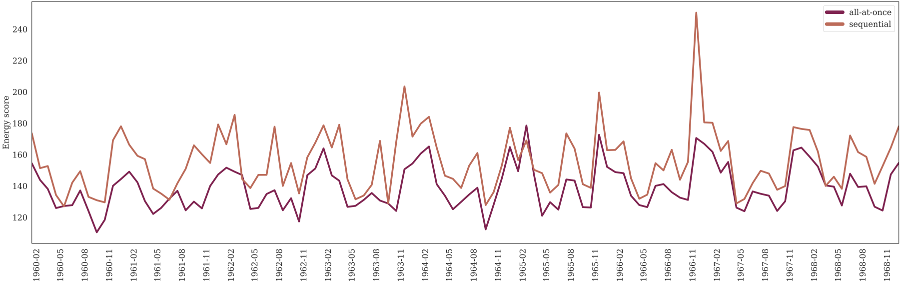


Figure: Energy score for reconstruction of the reference dataset over the 1960-1980 period.

Conclusion

Lazy, distributed arrays (implemented in `Dask`) allow for correct implementation of **sequential**, **localized** ensemble square root Kalman filter.

- Experiments show performance increase of order 5% compared to traditional implementation in synthetic test cases.
- Beats sequential implementation on real climate reconstruction problems.
- Scales to state spaces of sizes $10^5 - 10^6$.
- Opens way to more complex estimation of large covariances.

Note: there are related works [Bop17, FB19], but not distributed, and without detailed study.

Packages

Self-contained package available at
https://github.com/CedricTravelletti/DIESEL

References I

Shaunak D. Bopardikar, *Randomized matrix factorization for kalman filtering*, 2017 American Control Conference (ACC), 2017, pp. 5795–5800.

Geir Evensen et al., *Data assimilation: the ensemble kalman filter*, vol. 2, Springer, 2009.

Geir Evensen, *Sequential data assimilation with a nonlinear quasi-geostrophic model using monte carlo methods to forecast error statistics*, Journal of Geophysical Research: Oceans **99** (1994), no. C5, 10143–10162.

______, *The ensemble kalman filter: Theoretical formulation and practical implementation*, Ocean dynamics **53** (2003), no. 4, 343–367.

Alban Farchi and Marc Bocquet, *On the efficiency of covariance localisation of the ensemble kalman filter using augmented ensembles*, Frontiers in Applied Mathematics and Statistics **5** (2019).

Tilmann Gneiting and Adrian E Raftery, *Strictly proper scoring rules, prediction, and estimation*, Journal of the American Statistical Association **102** (2007), no. 477, 359–378.

References II

Tilmann Gneiting, Larissa I. Stanberry, Eric P. Grimit, Leonhard Held, and Nicholas A. Johnson, *Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds*, TEST **17** (2008), no. 2, 211–235.

Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp, *Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions*, SIAM review **53** (2011), no. 2, 217–288.

Allan H. Murphy and Edward S. Epstein, *Skill scores and correlation coefficients in model verification*, Monthly Weather Review **117** (1989), no. 3, 572 – 582.

Lars Nerger, *On serial observation processing in localized ensemble kalman filters*, Monthly Weather Review **143** (2015), no. 5, 1554 – 1567.

Michael K Tippett, Jeffrey L Anderson, Craig H Bishop, Thomas M Hamill, and Jeffrey S Whitaker, *Ensemble square root filters*, Monthly weather review **131** (2003), no. 7, 1485–1490.

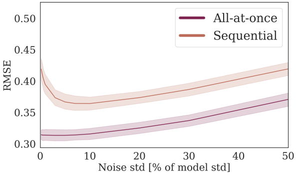
Jeffrey S. Whitaker and Thomas M. Hamill, *Ensemble data assimilation without perturbed observations*, Monthly Weather Review **130** (2002), no. 7, 1913 – 1924.

Section 4

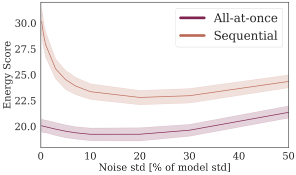
Appendix

Noise Dependence (synthetic case)

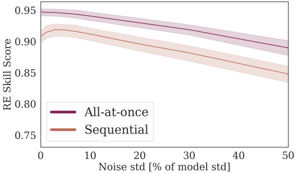
As noted by [Ner15], wrong update equations in **localized** and **sequential** EnSRF should have little effect when observation error is of same order as model errors.



(a) RMSE



(b) Energy score



(c) RE skill score

Figure: Evolution of accuracy metrics as a function of the observational noise standard deviation.