This article can be cited as E. D. Nino-Ruiz, L. Guzman and D. Jabba, Ensemble Driven Shrinkage Covariance Matrix Estimation for Sequential Data Assimilation, International Journal of Artificial Intelligence, vol. 20, no. 2, pp. 48-63, 2022.

Ensemble Driven Shrinkage Covariance Matrix Estimation for Sequential Data Assimilation

Elias D. Nino-Ruiz[1,2], **Luis Guzman**[2], **and Daladier Jabba** [2]

[1]Applied Math and Computer Science Lab
Universidad del Norte
Barranquilla, Colombia
enino@uninorte.edu.co

[2]Department of Computer Science
Universidad del Norte
Barranquilla, Colombia
{enino, lgguzman, djabba}@uninorte.edu.co

ABSTRACT

This paper proposes efficient and practical implementations of the ensemble Kalman filter via shrinkage covariance matrix estimation. Our methods exploit the practical properties of shrinkage covariance matrix estimators such as the Rao-Blackwell Ledoit and Wolf (RBLW), and the Oracle Approximating Shrinkage (OAS) to develop efficient ensemble Kalman filter implementations. These shrinkage-based estimators combine the information brought by the ensemble covariance and (typically) a static one: the resulting estimator is a convex combination of both matrices. As part of our formulations, we dynamically combine the OAS and the RBLW estimators by exploiting the information encapsulated in the eigenvalues of the ensemble covariance matrix; this is, the directions along which forecast errors develop quickly. By imposing a threshold on them, we can target the directions to grow faster forecast errors. Experimental tests are performed by using the Lorenz-96 model and an Atmospheric General Circulation Model. The results reveal that the proposed methods can improve the EnKF based on the RBLW in Root-Mean-Square-Errors. Besides, the computational efforts of our formulations are similar to that of state-of-the-art filter implementations.

Keywords: Ensemble Kalman Filter, Sampling Errors, Shrinkage Covariance Estimation, Spurious Correlations

2020 Mathematics Subject Classification: 37N10, 47A75, 90-08, 62P35, 60G15, 62M20, 86-08 .

ACM Computing Classification System: Mathematics of computing Discrete mathematics Combinatorics Combinatoric problems, Mathematics of computing Discrete mathematics Combinatorics Combinatorial optimization.

1 Introduction

A well-known method in the context of sequential data assimilation is the ensemble Kalman filter (EnKF). The EnKF is a Monte Carlo which is commonly employed to estimate the state of highly non-linear systems (Burgers, Jan van Leeuwen and Evensen, 1998; Evensen, 2003). In the EnKF, an ensemble of model realizations is utilized to estimate the forecast error distribution moments. In practice, ensemble sizes are constrained by the hundreds while model dimensions range in the order of millions (Lahoz, Khattatov and Ménard, 2010). Consequently, the ensemble covariance becomes rank-deficient and spurious correlations can impact the quality of analysis innovations. The effects of sampling errors can be mitigated via localization methods (Houtekamer and Zhang, 2016). In this context, well-known methods are covariance matrix localization (Hamill, Whitaker and Snyder, 2001; Sakov and Bertino, 2011), observation localization (Anderson, 2003), precision matrix estimation (Nino-Ruiz, Sandu and Deng, 2018), and spatial domain localization (Buehner and Charron, 2007). Yet another family of methods is based on shrinkage covariance matrix estimation. In these methods, the rank-deficient ensemble covariance matrix is replaced by a convex combination of a (full rank) target matrix and the ensemble one (Nino-Ruiz and Sandu, 2017). Some covariance matrix estimators in this context are the Ledoit, and Wolf (Ledoit and Wolf, 2004; Chen, Wiesel, Eldar and Hero, 2010), and the Rao-Blackwell Ledoit and Wolf (RBLW), the former improves on the estimation results of the LW under Gaussian assumptions (Nino-Ruiz and Sandu, 2017). Besides, iterative methods such as the Oracle Approximating Shrinkage (OAS) estimator can be employed to improve the results of LW and RBLW estimators iteratively. We think there is an opportunity to combine shrinkage covariance matrix estimators to estimate the forecast error covariance matrix better. The information brought by the ensemble covariance can be exploited to estimate optimal weights for each covariance matrix estimator. For instance, we can target the directions along which forecast errors develop faster via the ensemble covariance matrix's eigenvalues.
This paper is organized as follows: Section 2 discusses theoretical aspects of data assimilation such as ensemble-based methods and covariance matrix estimation, in Section 3, we derive efficient EnKF formulations via the OAS and the RBLW covariance matrix estimators, in this method, weights are dynamically tunned to target directions along which errors overgrow via the eigenvalues of the ensemble covariance matrix, experimental tests are performed in Section 4 by using the Lorenz 96 model and an Atmospheric General Circulation Model, conclusions of this research are stated in Section 5.

2 Preliminaries

The goal of sequential data assimilation is to estimate the state of a dynamical system $\mathbf{x}^* \in \mathbb{R}^{n\times 1}$ which approximately evolves according to some imperfect numerical model:

$$\mathbf{x}_{\text{next}} = \mathcal{M}_{t_{\text{current}}\rightarrow t_{\text{next}}}\left(\mathbf{x}_{\text{current}}\right) , \tag{2.1}$$

where n is the number of model components, and $\mathcal{M} : \mathbb{R}^{n\times 1} \rightarrow \mathbb{R}^{n\times 1}$ is an imperfect model operator which encapsulates our knowledge about the physics and the dynamics of the sys-

tem. Forecasts are typically corrected by using real-noisy observations $\mathbf{y} \in \mathbb{R}^{m\times 1}$, where m is the number of observations. When Gaussian assumptions are employed on forecast and observational errors, posterior states can be computed as follows:

$$\mathbf{x}^a = \mathbf{x}^b + \left[\mathbf{B}^{-1} + \mathbf{H}^T \cdot \mathbf{R}^{-1} \cdot \mathbf{H}\right]^{-1} \cdot \mathbf{H}^T \cdot \mathbf{R}^{-1} \cdot \left[\mathbf{y} - \mathbf{H}\left(\mathbf{x}^b\right)\right] \in \mathbb{R}^n\,, \tag{2.2}$$

where $\mathbf{x}^b \in \mathbb{R}^{n\times 1}$ is the forecast state, $\mathbf{B} \in \mathbb{R}^{n\times n}$ is the background error covariance matrix, $\mathbf{H} : \mathbb{R}^{n\times 1} \to \mathbb{R}^{m\times 1}$ is the observation operator (which maps vector states to observation spaces), $\mathbf{R} \in \mathbb{R}^{m\times m}$ is the data error covariance matrix, and $\mathbf{x}^a \in \mathbb{R}^{n\times 1}$ is known as the analysis state. Henceforth, we consider linear observation operators, non-linear observation operators are well-discussed in (Nino-Ruiz, Ardila, Estrada and Capacho, 2019; Nino-Ruiz, Cheng and Beltran, 2018; Nino-Ruiz and Yang, 2019; van Leeuwen, 2010; Moradkhani, Hsu, Gupta and Sorooshian, 2005). To estimate the moments $\mathbf{x}^b$ and $\mathbf{B}$ of the forecast error distribution ensemble based methods can be employed In this family of methods, an ensemble of model realizations

$$\mathbf{X}^b = \left[\mathbf{x}^{b[1]}, \mathbf{x}^{b[2]}, \ldots, \mathbf{x}^{b[N]}\right] \in \mathbb{R}^{n\times N}\,, \tag{2.3}$$

is employed to estimate the moments of the forecast error distribution (Evensen, 2003), where $\mathbf{x}^{b[e]} \in \mathbb{R}^{n\times 1}$ stands for the e-th ensemble member, for $1 \leq e \leq N$. The estimation process is then performed by using the empirical moments of the ensemble. The forecast state is estimated as follows

$$\overline{\mathbf{x}}^b = \sum_{e=1}^{N} \mathbf{x}^{b[e]} \in \mathbb{R}^{n\times 1}\,, \tag{2.4}$$

and the forecast error covariance matrix:

$$\mathbf{P}^b = \frac{1}{N-1} \cdot \mathbf{\Delta X} \cdot \mathbf{\Delta X}^T \in \mathbb{R}^{n\times n}\,, \tag{2.5}$$

where the matrix of member deviations $\mathbf{\Delta X} \in \mathbb{R}^{n\times N}$ reads:

$$\mathbf{\Delta X} = \mathbf{X}^b - \overline{\mathbf{x}}^b \cdot \mathbf{1}^T\,, \tag{2.6}$$

and $\mathbf{1}$ is a vector of consistent dimension whose components are all ones. The posterior ensemble can be computed as a linear map of the forecast ensemble (2.3) (Houtekamer and Mitchell, 1998). The posterior ensemble reads:

$$\mathbf{X}^a = \mathbf{X}^b + \mathbf{P}^b \cdot \mathbf{H}^T \cdot \left[\mathbf{H} \cdot \mathbf{P}^b \cdot \mathbf{H}^T + \mathbf{R}\right]^{-1} \cdot \mathbf{D} \in \mathbb{R}^{n\times N}\,, \tag{2.7}$$

where the e-th column of matrix $\mathbf{D} \in \mathbb{R}^{m\times N}$ is drawn from the distribution:

$$\mathcal{N}\left(\mathbf{y} - \mathbf{H} \cdot \mathbf{x}^{b[e]}, \mathbf{R}\right)\,, \tag{2.8}$$

Since ensemble sizes are much lesser than model resolutions (in some cases, by orders of

magnitudes), spurious correlations can impact the quality of analysis innovations in (2.7). To counteract the effects of sampling errors, localization methods can be employed to "denoise" the forecast error covariance matrix (2.5) (Anderson, 2007). The main idea behind these approaches is to exploit the spatial distance in model components to dissipate error correlations. Nevertheless, useful information of error dynamics brought by the ensemble members (2.3) can be lost (Lorenc, 2003; Kepert, 2009), and even more, imbalanced information can lead to suboptimal analyses (Houtekamer, Mitchell, Pellerin, Buehner, Charron, Spacek and Hansen, 2005). Yet another family of covariance matrix estimation methods is shrinkage-based ones. These methods aim to improve the estimation of the ensemble covariance matrix instead of forcing its structure; the resulting estimator is a convex combination of some target matrix $\mathbf{T} \in \mathbb{R}^{n \times n}$ and the sample covariance matrix (2.5):

$$\widehat{\mathbf{B}}(\alpha) = \alpha \cdot \mathbf{T} + (1 - \alpha) \cdot \mathbf{P}^b \in \mathbb{R}^{n \times n}\,, \text{ for } \alpha \in [0,\, 1]. \tag{2.9}$$

The shrinkage intensity α rely on the shrinkage method, for instance, in the RBLW α and $\mathbf{T}$ are given as follows:

$$\mathbf{T} = \frac{\mathbf{trace}\left(\mathbf{P}^b\right)}{n} \cdot \mathbf{I}\,, \tag{2.10}$$

and

$$\alpha_{RBLW} = \min\left(\frac{\frac{N-2}{N} \cdot \mathbf{trace}\left(\mathbf{P}^{b2}\right) + \mathbf{trace}^2\left(\mathbf{P}^b\right)}{(N+2) \cdot \left[\mathbf{trace}\left(\mathbf{P}^{b2}\right) - \frac{\mathbf{trace}^2\left(\mathbf{P}^b\right)}{n}\right]},\, 1\right)\,. \tag{2.11}$$

When the number of ensembles is minimal ($N \to 0$), the shrinkage intensity (2.11) becomes 1. Therefore, the estimation falls entirely over the ensemble covariance matrix, which is sensitive to spurious correlations between errors of distant model components. To mitigate the impact of small N, (Chen et al., 2010) proposes an Oracle Approximating Shrinkage (OAS) estimator. For small ensembles sizes, the shrinkage intensity α in the OAS estimator becomes 0, and it is defined as follows:

$$\alpha^*_{OAS} = \min\left(\frac{\frac{1-2}{n} \cdot \mathbf{trace}\left(\mathbf{P}^{b2}\right) + \mathbf{trace}^2\left(\mathbf{P}^b\right)}{(\frac{N+1-2}{n}) \cdot \left[\mathbf{trace}\left(\mathbf{P}^{b2}\right) - \frac{\mathbf{trace}^2\left(\mathbf{P}^b\right)}{n}\right]},\, 1\right)\,. \tag{2.12}$$

In (Chen et al., 2010), it is proven that for small values of N (number of samples), the OAS estimator can provide better results than those of the RBLW estimator.

Estimators of the form (2.9) have been exploited in some EnKF formulations, such as the EnKF based on the RBLW estimator (Nino-Ruiz and Sandu, 2015) and the EnKF based on the LW covariance estimator. In both cases, matrix-free implementations are possible given the particular structure of $\mathbf{P}^b$. An attractive feature of shrinkage-based covariance estimation is that they perform remarkably well for the case $n \gg N$ (Chen et al., 2010) which is very

common in the data assimilation context.

3 Proposed Methods

In this section, we develop efficient and practical implementations of the EnKF based on shrinkage covariance matrix estimation: the first method is based on the OAS estimator, while the second one optimally combines the information brought by the OAS and the RBLW covariance matrix estimators.

3.1 An Ensemble Kalman Filter Based on the Oracle Approximating Shrinkage Covariance Matrix Estimator

To be concise, efficient implementations of the OAS estimator can be derived in the context of EnKF. For instance, we can exploit some properties of traces to avoid the direct computation of $\mathbf{trace}\left(\mathbf{P}^b\right)$ via:

$$\mathbf{trace}\left(\mathbf{P}^b\right) = \sum_{e=1}^{N-1} \sigma_e^2 \,, \tag{3.1}$$

where σ_e is the e-th singular value of $\widehat{\mathbf{\Delta X}} = \sqrt{N-1}^{-1} \cdot \mathbf{\Delta X} \in \mathbb{R}^{n \times N}$ with $\mathbf{T}$ given as in 2.10. Note that, the computational effort for computing (3.1) is linearly bounded with regard to the ensemble size N. In addition, the assimilation process of EnKF can be performed onto the observation space similar to (Nino-Ruiz and Sandu, 2015), this is:

$$\mathbf{X}^a = \mathbf{X}^b + \mathbf{E} \cdot \mathbf{K} \cdot \mathbf{Z} + \alpha_{OAS}^* \cdot \mathbf{H}^T \cdot \mathbf{Z} \in \mathbb{R}^{n \times N} \,, \tag{3.2}$$

where,

$$\mathbf{E} = \sqrt{1 - \alpha_{OAS}^*} \cdot \widehat{\mathbf{\Delta X}} \in \mathbb{R}^{n \times N} \,, \tag{3.3}$$

$$\mathbf{K} = \mathbf{H} \cdot \mathbf{E} \in \mathbb{R}^{m \times N} \,, \tag{3.4}$$

$$\tag{3.5}$$

and $\mathbf{Z} \in \mathbb{R}^{n \times N}$ is given by the solution of the following linear system:

$$\left[\mathbf{Q} + \left[\mathbf{H} \cdot \widehat{\mathbf{\Delta X}}\right] \cdot \left[\mathbf{H} \cdot \widehat{\mathbf{\Delta X}}\right]^T\right] \cdot \mathbf{Z} = \mathbf{D} \,. \tag{3.6}$$

Notice, since $\mathbf{R}$ is a diagonal matrix and $\mathbf{H}$ is a linear operator, the matrix $\mathbf{Q} \in \mathbb{R}^{m \times m}$ is a diagonal matrix with elements $\{\mathbf{R}\}_{jj} + \alpha_{OAS}^*$, for $1 \leq j \leq m$, where $\{\mathbf{R}\}_{jj}$ is the j-th diagonal components of $\mathbf{R}$.

3.2 An Ensemble Kalman Filter Implementation via Dynamic Shrinkage Intensity

For small N, the identity matrix has more weight in the convex combination (2.9), and therefore, the resulting estimator behaves similarly to a B-localization matrix. Consequently, since the resulting estimator is diagonal, analysis innovation will only update forecast components wherein observations are located. This can raise some issues about the dynamical consistency of posterior ensemble members in (3.2). We can choose an appropriated shrinkage intensity α between the shrinkage factors α^*_{RBLW} and α^*_{OAS}. The analysis is simple: for smaller N values $\alpha^*_{RBLW} \approx 0$ while $\alpha^*_{OAS} \approx 1$. The amount of information expressed in the eigenvalues π_e of $\mathbf{P}^b$, for $1 \leq e \leq N-1$, can tell us the directions along which forecast errors grow faster. Recall that, $\pi_e = \sigma_e^2$, where σ_e denotes a singular value of $\widehat{\mathbf{\Delta X}}$. We can impose a threshold on the values π_e. Consider the ratio:

$$\frac{\varphi}{n} \text{ where } \sigma^2_{\varphi} > \frac{\mathbf{trace}\left(\mathbf{P}^b\right)}{N}\,, \text{ and } 1 \leq \varphi \leq N-1\,. \tag{3.7}$$

If the resulting ratio (3.7) is greater than a threshold $\xi \in [0,\, 1]$, we can assume that the ensemble members provide enough information about the error correlations and therefore we can make use of α^*_{RBLW}, we can choose α_{OAS}, otherwise. The resulting covariance estimator is as follows:

$$\widehat{\mathbf{B}}_{DS} = \begin{cases} \widehat{\mathbf{B}}_{OAS} & \text{for } \frac{\varphi}{n} < \xi \\ \widehat{\mathbf{B}}_{RBLW} & \text{otherwise} \end{cases} \tag{3.8}$$

where DS stands for "Dynamical Shrinkage ." Note that:

- we target the directions along which error develops faster by neglecting the smaller singular values of $\widehat{\mathbf{\Delta X}}$ (a low-rank square root approximation of $\mathbf{P}^b$),
- the target matrix $\mathbf{T}$ is similar to that of (2.10),
- the posterior members (3.2) can be easily inflated by multiplying the inflation factor β_{inf} by the matrix of member deviations (2.6), and
- the threshold φ can be estimated using prior information about the numerical model or employing heuristics (i.e., using historical data and different values of φ).

The resulting EnKF implementation is obtained by replacing $\widehat{\mathbf{B}}_{DS}$ in the analysis equations (2.7), we call this implementation the EnKF-DS. Given the particular structure of the covariance matrix estimators $\widehat{\mathbf{B}}_{RBLW}$ and $\widehat{\mathbf{B}}_{OAS}$, similar to the EnKF-OAS (Section 3.1), matrix-free implementations are possible for this filter formulation.

4 Experimental Settings

We test our filter implementations EnKF-OAS and EnKF-DS using two numerical models: the toy model Lorenz-96 and an Atmospheric General Circulation Model. We use the mean of eigenvalues of $\mathbf{P}^b$ to set our dynamical threshold ξ in all experiments.

4.1 Experimental Results with the Lorenz-96 Model

The Lorenz-96 model is a toy model which mimics the dynamics of the atmosphere (Lorenz, 2005). The model is widely employed in the literature to try emergent data assimilation methods. The model is given by the following set of ordinary differential equations:

$$\frac{dx_i}{dt} = \begin{cases} (x_2 - x_{n-1}) \cdot x_n - x_1 + F & \text{for i} = 1 \\ (x_{i+1} - x_{i-2}) \cdot x_{i-1} - x_i + F & \text{for } 2 \leq i \leq n-1 \\ (x_1 - x_{n-2}) \cdot x_{n-1} - x_n + F & \text{for i} = n \end{cases} \tag{4.1}$$

where x_i, for $1 \leq i \leq n$, are the spatial coordinates. In the Lorenz-96 model, a one-time unit represents 120 hours in the atmosphere. This model exhibits extended chaos when the external force equals $F = 8$.
We try two different values for the ensemble size $N \in \{10\,, 20\}$ while the number of observations m is a function of the model dimension n:

$$m = p \cdot n\,, \tag{4.2}$$

where $p = 0.7$ this is, $m = 28$ observations are available during assimilation steps. The observations are randomly placed at each assimilation cycle. The number of assimilation steps reads $M = 300$. Observations are collected every 2.5 days; their error statistics are described by the following Gaussian distribution:

$$\mathbf{y}_\ell \sim \mathcal{N}\left(\mathbf{H}_\ell \cdot \mathbf{x}_\ell^*\,, 0.01^2 \cdot \mathbf{I}\right)\,,\ \text{for } 1 \leq \ell \leq M\,. \tag{4.3}$$

where $\mathbf{x}_\ell^*$ is the reference state at time ℓ, since observations are randomly placed, $\mathbf{H}_\ell$ is randomly formed at each assimilation cycle. Experiments are performed under perfect model assumptions. The inflation factor reads $\beta_{\text{inf}} = 1.1$. As a metric of performance, we consider the L–2 norms of errors, at the assimilation step ℓ, it is defined as follows:

$$\lambda_\ell = \|\mathbf{x}_\ell^* - \overline{\mathbf{x}}_\ell^a\|_2 = \sqrt{\left[\mathbf{x}_\ell^* - \overline{\mathbf{x}}_\ell^a\right]^T \cdot \left[\mathbf{x}_\ell^* - \overline{\mathbf{x}}_\ell^a\right]}\,, \tag{4.4}$$

where $\mathbf{x}_\ell^*$ and $\overline{\mathbf{x}}_\ell^a$ are the reference and the analysis solutions (mean of the analysis ensemble), respectively. The Root-Mean-Square-Error (RMSE) measures, in average, the performance of a filter for a given number of assimilation steps M:

$$\Pi = \sqrt{\frac{1}{M} \cdot \sum_{\ell=1}^{M} \lambda_\ell^2}\,. \tag{4.5}$$

We consider 25 repetitions varying the initial ensemble (and the reference solution) for each configuration to estimate error statistics. As a reference, we make use of the EnKF-RBLW.

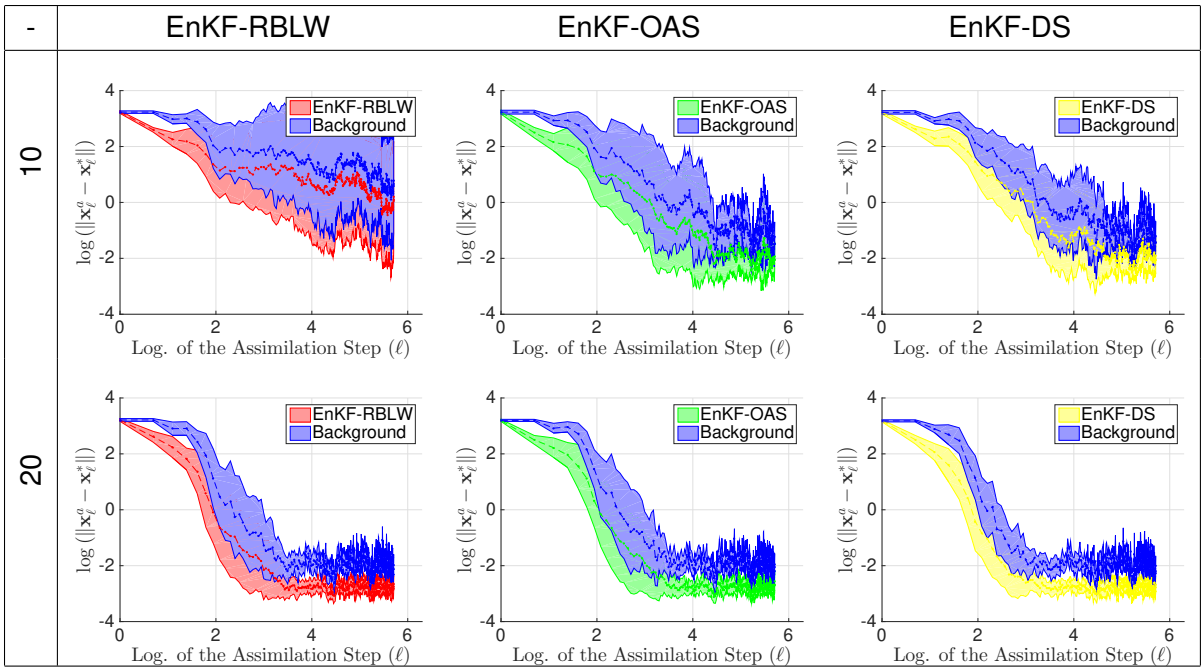


Figure 1: Mean (dashed lines) and standard deviations (shaded regions) of L–2 error norms in the log scale for EnKF-RBLW, EnKF-OAS and EnKF-DS for $p = 0.7$, and $N \in \{10, 20\}$.

Figure 1 shows the results for the EnKF-RBLW, EnKF-OAS, and EnKF-DS. The results are presented in the log scale for easiness in reading. As can be seen, for small ensemble sizes, the proposed filter implementations' accuracy is better than that of the EnKF-RBLW. Besides, our proposals' standard deviations of errors (dispersion of errors) remain small during the entire assimilation window. Note that the performance of the EnKF-RBLW is highly impacted: this can be evidenced as the standard deviation of errors tend to grow. All compared filter implementations exhibit similar performance for large ensemble sizes ($n = 40$). However, note that, in all cases, errors in the EnKF-DS execution have lower variance among the compared filter implementations.

In Table 1, results are reported in terms of RMSE values with ensembles sizes of 10 and 20 for all compared filter implementations. The spin-up period (corresponding to the first 100 assimilation steps) is removed to compute analysis errors. This is done to compare the filter implementations in a stable phase of the numerical model. As should be expected, the OAS estimator becomes better for small ensemble sizes while EnKF-DS improves all compared results for ensemble sizes of 10 and 20.

N	Method	$\overline{\Pi}$	S_{Π}
10	EnKF-RBLW	8.1079	4.5587
	EnKF-OAS	1.7229	1.9715
	EnKF-DS	1.6425	2.0592
20	EnKF-RBLW	0.7971	2.4024
	EnKF-OAS	0.0952	0.0721
	EnKF-DS	0.0679	0.0038

Table 1: Mean ($\overline{\Pi}$) and standard deviations (S_{Π}) of RMSE values in the log scale for $\beta_{\text{inf}} = 1.1$, and $p = 0.7$.

4.2 Experimental Results with an Atmospheric General Circulation Model

The Simplified Parameterizations, privitivE-Equation DYnamics (SPEEDY), is an Atmospheric General Circulation model which is based on a spectral primitive equation dynamical core and a set of simplified physical parameterization schemes(Bracco, Kucharski, Kallummal and Molteni, 2004; Miyoshi, 2011). The number of layers in the SPEEDY model is 7, and the T-30 model resolution (96×48 grid components) is used for the space discretization of each layer (Molteni, 2003; Kucharski, Molteni and Bracco, 2006). The total number of model components is $n = 133,632$. Five model variables are part of the assimilation process: the temperature (K), the zonal (u) and the meridional (v) wind components (m/s), the specific humidity (g/kg), and the pressure (Ps). The general settings are detailed below:

- we try two different ensemble sizes N, 10 and 30,
- the number of observation is given by a function of n, this is,

$$m = p \cdot n\,,$$

  where $p \in [0,\, 1]$ is the percentage of observed components from the model state,

- The standard deviation of observational errors is detailed in the Table 2.

Model Variable	Observational Error Standard Deviation
Zonal Wind Component (u)	1 m/s
Meridional Wind Component (v)	1 m/s
Temperature (T)	1 (K)
Specific humidity (q)	0.0001 (kg/kg)
Surface pressure (p)	100 (Pa)

Table 2: Observational error standard deviation.

- the assimilation window consists of $M = 25$ observations, and
- the inflation factor reads $\beta_{\text{inf}} = 1.5$.

4.2.1 Results with Full Observational Networks $p = 1$

Figure 2 presents the performance of the compared filter implementations for whole observational networks. Again, the log scale shows the results for easiness in the reading. As can be seen, the proposed methods' performance is better for small ensemble sizes than that of the EnKF-RBLW. This obeys the dynamical tunning of the shrinkage factor. This also can be explained as follows: forecast errors grow differently along with different directions, this information is encapsulated into the eigenvalues of $\mathbf{P}^b$, and therefore, by targeting approaches along which errors grow faster (by imposing thresholds on such eigenvalues), we can improve the accuracy of EnKF shrinkage based method. Recall that the threshold ξ in our EnKF-DS filter formulation is given by the mean eigenvalues of $\mathbf{P}^b$. For instance, for small ensemble sizes, we can see in Table 3 that RMSE values of the EnKF-DS are lower than those of the EnKF-RBWL formulation. Similar behaviors can be seen for all filter formulations as the ensemble size increases, and therefore, the ensemble covariance matrix becomes informative.

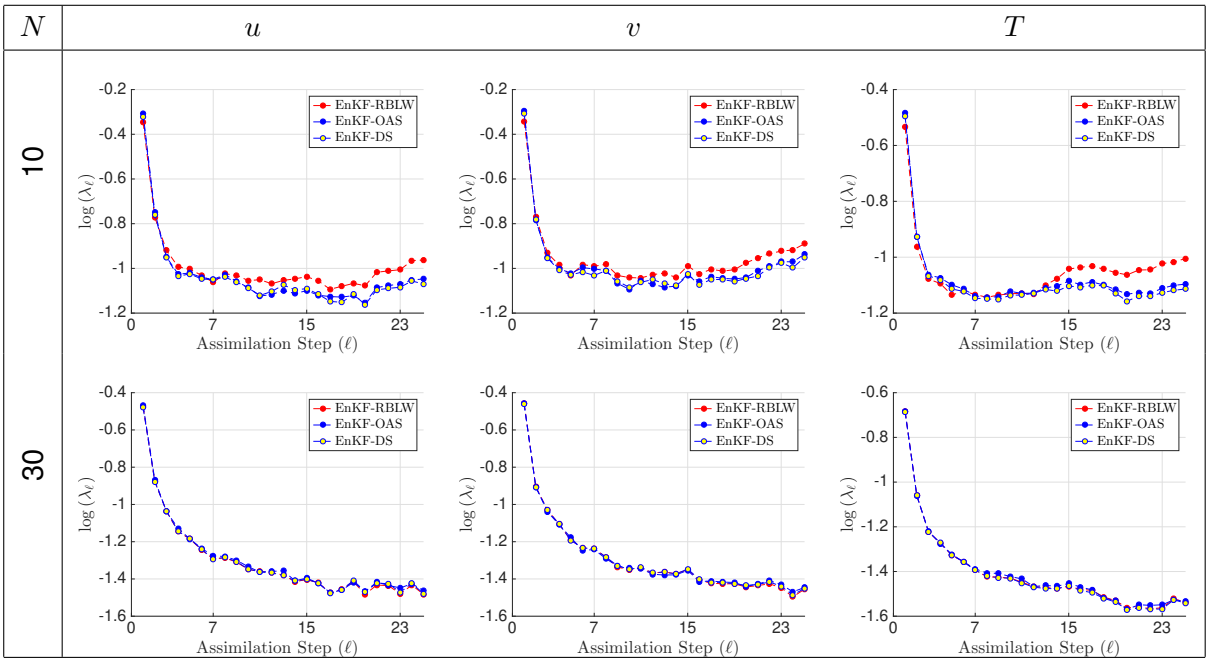


Figure 2: Mean of L–2 error norms in the log scale for EnKF-RBLW, EnKF-OAS and EnKF-DS for $p = 1$ and $N \in \{10, 30\}$.

N	Method	u	v	T
10	EnKF-RBLW	0.34882	0.36132	0.32446
	EnKF-OAS	0.33862	0.35288	0.31796
	EnKF-DS	0.33671	0.35052	0.3146
30	EnKF-RBLW	0.26579	0.26892	0.23425
	EnKF-OAS	0.26786	0.26916	0.23532
	EnKF-DS	0.26590	0.26900	0.23367

Table 3: Mean of RMSE values for $\beta_{\text{inf}} = 1.5$ and $p = 1$.

4.2.2 Results with Sparse Observational Networks $p = 0.09$

We study the performance of the compared filter implementation under sparse observational networks. In this context, just 9% of model components are observed. The irregular observational network is shown in figure 3, it has been employed for other studies in the context of data assimilation (Miyoshi, 2011).

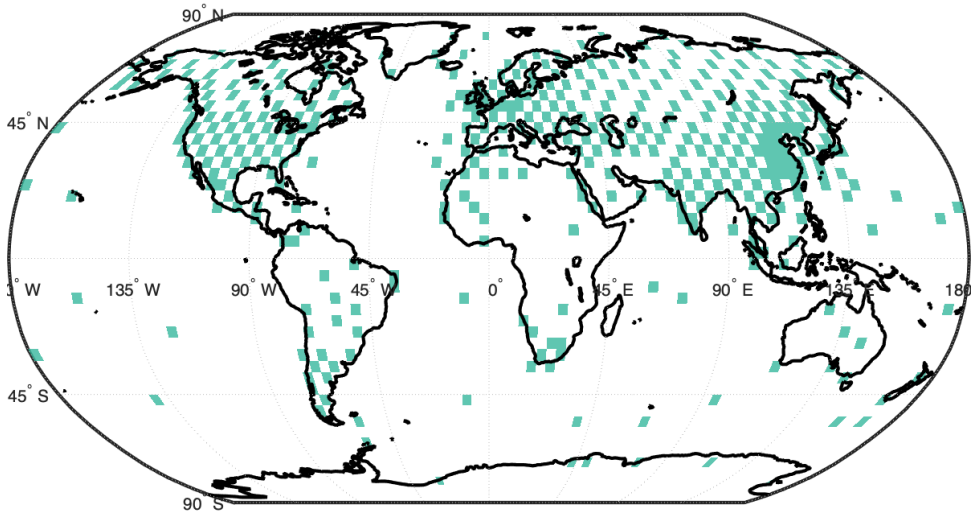


Figure 3: An irregularly distributed realistic observational network. 415 stations (9 % of all grid points) are located mostly over continents in the northern hemisphere.

Figure 4 shows L–2 errors in the log scale for the compared filter implementations and only 9% of observed components from the model state. We can see a similar trend in the performance of the EnKF implementations. For small values of N, the proposed EnKF-DS provides more accurate results than those obtained by the EnKF-RBLW and the EnKF-OAS implementations. Recall that, in the EnKF-DS, the information brought by the observations targets the directions along which errors grow faster. As the dimension of the ensemble increases, all methods tend to have the same performance. Again, this is a direct consequence of having more information in the ensemble covariance. This shows the importance of using EnKF-DS: we can get similar results to state-of-the-art methods, and even more, we can improve their results as the ensemble size becomes smaller (in some cases, N is lesser than n by order of magnitudes). This potentially can make our proposed filter implementations attractive to be employed under operational data assimilation scenarios. For all compared methods, a general overview of their performance can be seen in Table 4.

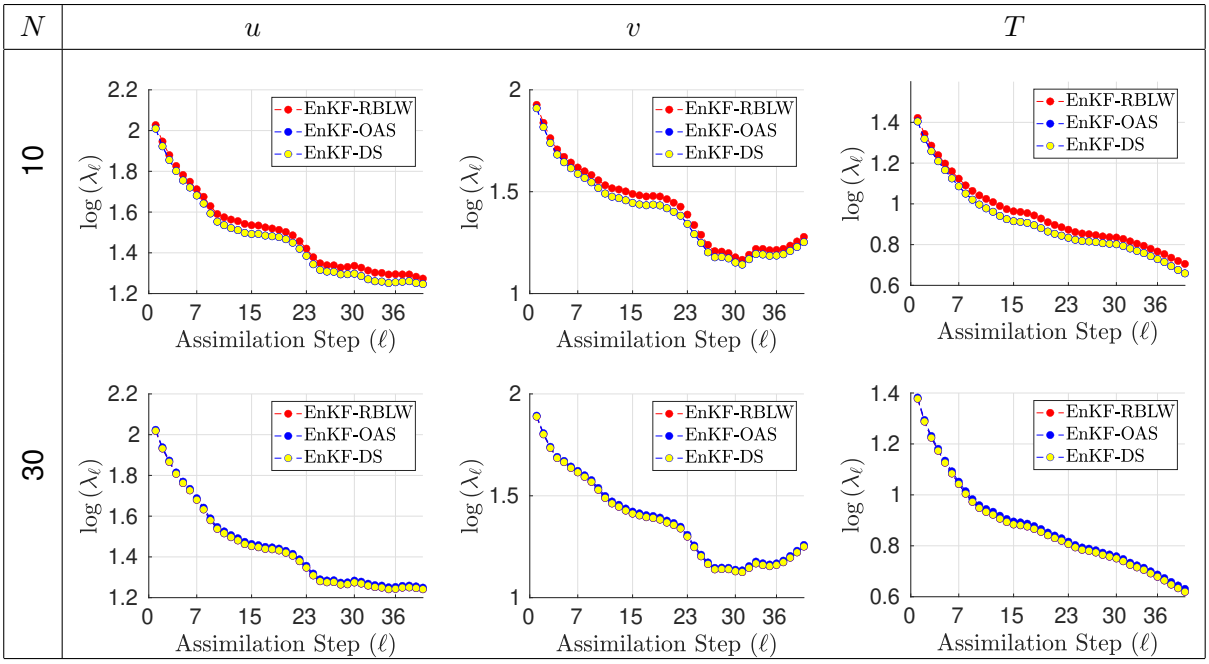


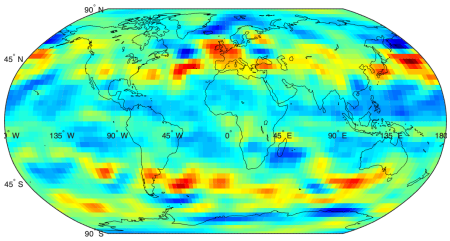
Figure 4: Mean of L–2 error norms in the log scale for EnKF-RBLW, EnKF-OAS and EnKF-DS. Here, $p = 0.09$, and $N \in \{10, 30\}$.

N	Method	u	v	T
10	EnKF-RBLW	4.21191	3.96075	2.34704
	EnKF-OAS	4.05499	3.81652	2.34704
	EnKF-DS	4.05499	3.81652	2.34704
30	EnKF-RBLW	3.94119	3.70926	2.25829
	EnKF-OAS	3.97995	3.74373	2.28144
	EnKF-DS	3.94119	3.70926	2.25829

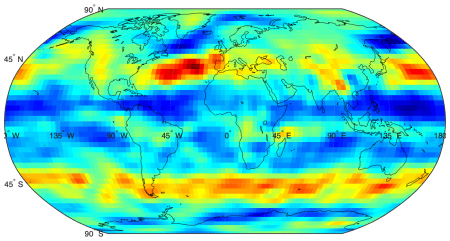
Table 4: RMSE values for $\beta_{\text{inf}} = 1.5$ and $p = 0.09$.

4.2.3 Further Comments

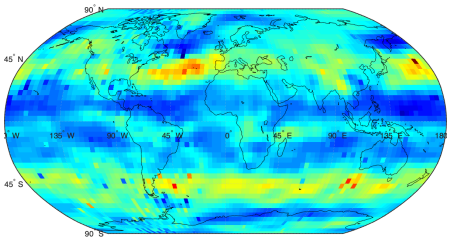
Figures 5 and 6 show snapshots of the initial assimilation step for the model variables u and v, respectively. This initial step is of high interest to us since no previous data has been injected into the numerical forecast (forecast ensemble members). As can be seen in all cases, the EnKF-RBLW and the EnKF-OAS can dissipate superiors waves in zonal and meridional wind components. This implies that the quality of correlations on each estimated forecast error covariance matrix is a reasonable approximation of the actual spatial correlation of forecast errors for the dynamical system SPEEDY.



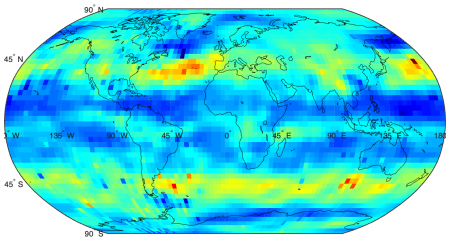
(a) Actual State $\mathbf{x}_0^*$



(b) Background State $\mathbf{x}_0^b$

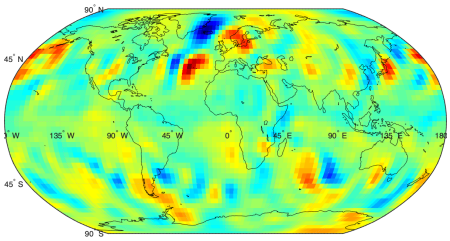


(c) EnKF-RBLW Analysis

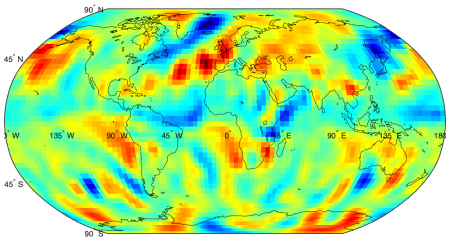


(d) EnKF-DS Analysis

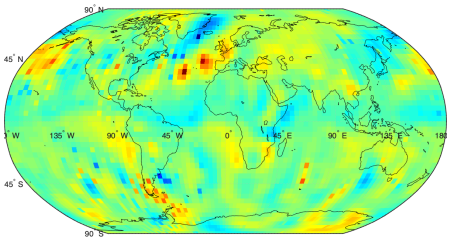
Figure 5: Snapshots of the initial zonal wind component (u) for the actual state of the system, the background state, the analysis state via the EnKF-RBLW implementation, and the analysis state obtained by the EnKF-DS.



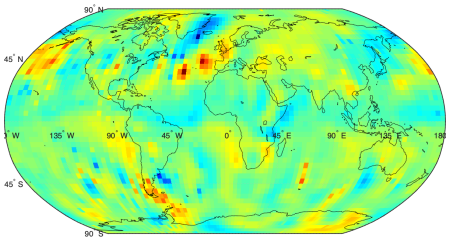
(a) Actual State $\mathbf{x}_0^*$



(b) Background State $\mathbf{x}_0^b$



(c) EnKF-RBLW Analysis



(d) EnKF-DS Analysis

Figure 6: Snapshots of the initial meridional wind component (v) for the actual state of the system, the background state, the analysis state via the EnKF-RBLW implementation, and the analysis state obtained by the EnKF-DS.

5 Conclusions

This paper proposes two efficient and practical implementations of the EnKF: the EnKF based on the OAS covariance matrix estimator (EnKF-OAS) and the EnKF based on the Dynamical Shrinkage factor (EnKF-DS). For small ensemble sizes, the EnKF-OAS method improves the results of shrinkage-based formulations such as the EnKF based on the RBLW estimator. As the ensemble member increases, the resulting estimator in the EnKF-OAS is diagonal, raising issues in the numerical model, for instance, posterior states with inconsistent dynamics. To counteract this, we can dynamically choose between the shrinkage factors of RBLW and OAS. This can be done by targeting directions along which forecast errors develop faster. This is equivalent to imposing a threshold on the eigenvalues of the forecast error covariance matrix. Experimental tests are performed by using the Lorenz-96 model and an Atmospheric General Circulation Model. The results reveal that, for small ensemble sizes, our proposed methods can improve on the results of filters such as the EnKF-RBLW. Besides, the computational efforts of our proposed methods are similar to those of efficient filter formulation.

6 Acknowledgement

This work was supported in part by grant number Colciencias-757 and the Applied Math and Computer Science Laboratory at Universidad del Norte, Colombia.

References

Anderson, J. L. 2003. A local least squares framework for ensemble filtering, *Monthly Weather Review* **131**(4): 634–642.

Anderson, J. L. 2007. Exploring the need for localization in ensemble data assimilation using a hierarchical ensemble filter, *Physica D: Nonlinear Phenomena* **230**(1-2): 99–111.

Bracco, A., Kucharski, F., Kallummal, R. and Molteni, F. 2004. Internal variability, external forcing and climate trends in multi-decadal agcm ensembles, *Climate Dynamics* **23**(6): 659–678.

Buehner, M. and Charron, M. 2007. Spectral and spatial localization of background-error correlations for data assimilation, *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography* **133**(624): 615–630.

Burgers, G., Jan van Leeuwen, P. and Evensen, G. 1998. Analysis scheme in the ensemble kalman filter, *Monthly weather review* **126**(6): 1719–1724.

Chen, Y., Wiesel, A., Eldar, Y. C. and Hero, A. O. 2010. Shrinkage algorithms for mmse covariance estimation, *IEEE Transactions on Signal Processing* **58**(10): 5016–5029.

Evensen, G. 2003. The ensemble kalman filter: Theoretical formulation and practical implementation, *Ocean dynamics* **53**(4): 343–367.

Hamill, T. M., Whitaker, J. S. and Snyder, C. 2001. Distance-dependent filtering of background error covariance estimates in an ensemble kalman filter, *Monthly Weather Review* **129**(11): 2776–2790.

Houtekamer, P. L. and Mitchell, H. L. 1998. Data assimilation using an ensemble kalman filter technique, *Monthly Weather Review* **126**(3): 796–811.

Houtekamer, P. L., Mitchell, H. L., Pellerin, G., Buehner, M., Charron, M., Spacek, L. and Hansen, B. 2005. Atmospheric data assimilation with an ensemble kalman filter: Results with real observations, *Monthly weather review* **133**(3): 604–620.

Houtekamer, P. and Zhang, F. 2016. Review of the ensemble kalman filter for atmospheric data assimilation, *Monthly Weather Review* **144**(12): 4489–4532.

Kepert, J. D. 2009. Covariance localisation and balance in an ensemble kalman filter, *Quarterly Journal of the Royal Meteorological Society* **135**(642): 1157–1176.

Kucharski, F., Molteni, F. and Bracco, A. 2006. Decadal interactions between the western tropical pacific and the north atlantic oscillation, *Climate dynamics* **26**(1): 79–91.

Lahoz, W., Khattatov, B. and Ménard, R. 2010. Data assimilation and information, *Data Assimilation*, Springer, pp. 3–12.

Ledoit, O. and Wolf, M. 2004. A well-conditioned estimator for large-dimensional covariance matrices, *Journal of multivariate analysis* **88**(2): 365–411.

Lorenc, A. C. 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.

Lorenz, E. N. 2005. Designing chaotic models, *Journal of the Atmospheric Sciences* **62**(5): 1574–1587.

Miyoshi, T. 2011. The gaussian approach to adaptive covariance inflation and its implementation with the local ensemble transform kalman filter, *Monthly Weather Review* **139**(5): 1519–1535.

Molteni, F. 2003. Atmospheric simulations using a gcm with simplified physical parametrizations. i: Model climatology and variability in multi-decadal experiments, *Climate Dynamics* **20**(2-3): 175–191.

Moradkhani, H., Hsu, K.-L., Gupta, H. and Sorooshian, S. 2005. Uncertainty assessment of hydrologic model states and parameters: Sequential data assimilation using the particle filter, *Water resources research* **41**(5).

Nino-Ruiz, E. D., Ardila, C., Estrada, J. and Capacho, J. 2019. A reduced-space line-search method for unconstrained optimization via random descent directions, *Applied Mathematics and Computation* **341**: 15–30.

Nino-Ruiz, E. D., Cheng, H. and Beltran, R. 2018. A robust non-gaussian data assimilation method for highly non-linear models, *Atmosphere* **9**(4): 126.

Nino-Ruiz, E. D. and Sandu, A. 2015. Ensemble kalman filter implementations based on shrinkage covariance matrix estimation, *Ocean Dynamics* **65**(11): 1423–1439.

Nino-Ruiz, E. D. and Sandu, A. 2017. Efficient parallel implementation of dddas inference using an ensemble kalman filter with shrinkage covariance matrix estimation, *Cluster Computing* pp. 1–11.

Nino-Ruiz, E. D., Sandu, A. and Deng, X. 2018. An ensemble kalman filter implementation based on modified cholesky decomposition for inverse covariance matrix estimation, *SIAM Journal on Scientific Computing* **40**(2): A867–A886.

Nino-Ruiz, E. D. and Yang, X.-S. 2019. Improved tabu search and simulated annealing methods for nonlinear data assimilation, *Applied Soft Computing* **83**: 105624.

Precup, R.-E., David, R.-C. and Petriu, E. M. 2016. Grey wolf optimizer algorithm-based tuning of fuzzy control systems with reduced parametric sensitivity, *IEEE Transactions on Industrial Electronics* **64**(1): 527–534.

Precup, R.-E. and Hellendoorn, H. 2011. A survey on industrial applications of fuzzy control, *Computers in industry* **62**(3): 213–226.

Preitl, S. and Precup, R.-E. 1999. An extension of tuning relations after symmetrical optimum method for pi and pid controllers, *Automatica* **35**(10): 1731–1736.

Sakov, P. and Bertino, L. 2011. Relation between two common localisation methods for the enkf, *Computational Geosciences* **15**(2): 225–237.

van Leeuwen, P. J. 2010. Nonlinear data assimilation in geosciences: an extremely efficient particle filter, *Quarterly Journal of the Royal Meteorological Society* **136**(653): 1991–1999.