An efficient ensemble Kalman Filter implementation via shrinkage covariance matrix estimation

exploiting prior knowledge

Lopez-Restrepo, Santiago; Nino-Ruiz, Elias D.; Guzman-Reyes, Luis G.; Yarce, Andres; Quintero, O. L.; Pinel, Nicolas; Segers, Arjo; Heemink, A. W.

DOI
10.1007/s10596-021-10035-4

Publication date
2021

Document Version
Final published version

Published in
Computational Geosciences

Citation (APA)
Lopez-Restrepo, S., Nino-Ruiz, E. D., Guzman-Reyes, L. G., Yarce, A., Quintero, O. L., Pinel, N., Segers, A., & Heemink, A. W. (2021). An efficient ensemble Kalman Filter implementation via shrinkage covariance matrix estimation: exploiting prior knowledge. *Computational Geosciences*, *25*(3), 985-1003. https://doi.org/10.1007/s10596-021-10035-4

Computational Geosciences
https://doi.org/10.1007/s10596-021-10035-4

ORIGINAL PAPER

An efficient ensemble Kalman Filter implementation via shrinkage covariance matrix estimation: exploiting prior knowledge

Santiago Lopez-Restrepo[1,2] · Elias D. Nino-Ruiz[3] · Luis G. Guzman-Reyes[3] · Andres Yarce[1,2] · O. L. Quintero[1] · Nicolas Pinel[4] · Arjo Segers[5] · A. W. Heemink[2]

Received: 10 June 2020 / Accepted: 14 January 2021

Abstract
In this paper, we propose an efficient and practical implementation of the ensemble Kalman filter via shrinkage covariance matrix estimation. Our filter implementation combines information brought by an ensemble of model realizations, and that based on our prior knowledge about the dynamical system of interest. We perform the combination of both sources of information via optimal shrinkage factors. The method exploits the rank-deficiency of ensemble covariance matrices to provide an efficient and practical implementation of the analysis step in EnKF based formulations. Localization and inflation aspects are discussed, as well. Experimental tests are performed to assess the accuracy of our proposed filter implementation by employing an Advection Diffusion Model and an Atmospheric General Circulation Model. The experimental results reveal that the use of our proposed filter implementation can mitigate the impact of sampling noise, and even more, it can avoid the impact of spurious correlations during assimilation steps.

Keywords Data assimilation · Air quality · Chemical transport model · Ensemble Kalman Filter · Background error covariance matrix

Mathematics Subject Classification (2010) 62L20 · 60J22 · 49K45

1 Introduction

A dynamical system approximately evolves according to some imperfect numerical model:

$$\mathbf{x}_{\text{current}} = \mathcal{M}_{t_{\text{previous}} \rightarrow t_{\text{current}}} \left(\mathbf{x}_{\text{previous}} \right) , \tag{1}$$

where n and m are the model resolution and the number of observations, respectively, and $\mathcal{M} : \mathbb{R}^{n \times 1} \rightarrow \mathbb{R}^{n \times 1}$ is an imperfect model operator which mimics the behavior of a very highly non-linear system such as the ocean and/or the atmosphere.

On the former representation, the model operator maps the state variable into a sequential time steps realization of the behavior of the dynamical system. In most of the cases, there is a control variable included on the operator that related external inputs to the system and allows for the representation of the interactions between the system and the external world. The state variable may or may not be directly measurable and is used as a memory of the system. As seen in Eq. 1, the past behavior of the system affects its future development, but the lack of representation of the state variable may be a pitfall on the full representation of the real world. The relationship between the state space and the real noisy observation $\mathbf{y} \in \mathbb{R}^{m \times 1}$ is sometimes a useful tool for the proper understanding and representation of the full system.

Controllability is a property of the dynamical system that allows measuring the ability of a particular control input to manipulate all the states of the system, taking them from point A to the point B in finite time. On the other hand, observability measures the ability of the particular sensor configuration to supply all the information necessary to estimate all the states of the system. State estimation and Parameter estimation are typically the main concerns in control and systems theory. They are required for the proper control law design and are mandatory for the full observability of the system.

In cases when there is a lack of observability, the problem of state estimation and parameter estimation arose, and it

✉ Santiago Lopez-Restrepo
slopezr2@eafit.edu.co; lopezrestrepo@tudelft.nl

Extended author information available on the last page of the article.

can be solved by means of the solution to the optimal filtering problem. That requires an analytical solution of the Bayes theorem by means of the Kushner or Zakai Equation. These are not feasible for non-linear and non-Gaussian systems. They are approximated most often via particle filters [34, 35]. The linear and Gaussian case is solved by the well known Kalman filter, and its extension to non-linear and Gaussian cases can be found extensively in the literature. For Large scale systems, the solutions to complete the full observability of the system are not straight forward because the course of dimensionality and more sophisticated solutions to the optimal filtering problem were derived.

Sequential Data Assimilation (DA) is a statistical process that optimally combines information brought by an imperfect numerical forecast $\mathbf{x}^b \in \mathbb{R}^{n\times 1}$ and a real noisy observation $\mathbf{y} \in \mathbb{R}^{m\times 1}$ [2, 9] to estimate the actual state $\mathbf{x}^* \in \mathbb{R}^{n\times 1}$ of a dynamic system such as Eq. 1. When Gaussian assumptions are made over prior and observational errors via Bayes' rule, the posterior estimate has the form:

$$\mathbf{x}^a = \mathbf{x}^b + \mathbf{B}\cdot\mathbf{H}^T\cdot\mathbf{A}^{-1}\cdot\mathbf{d} \in \mathbb{R}^{n\times 1}\,, \tag{2}$$

where $\mathbf{B} \in \mathbb{R}^{n\times n}$ is the background error covariance matrix, $\mathbf{d} = \mathbf{y} - \mathcal{H}\left(\mathbf{x}^b\right) \in \mathbb{R}^{m\times 1}$ is the vector of innovations (on the observations), $\mathcal{H}(\mathbf{x}) : \mathbb{R}^{n\times 1} \to \mathbb{R}^{m\times 1}$ is the observation operator (which maps vector states to observations), $\mathcal{H}(\mathbf{x}) \approx \mathcal{H}\left(\mathbf{x}^b\right) + \mathbf{H}\cdot\left[\mathbf{x} - \mathbf{x}^b\right] \in \mathbb{R}^{m\times n}$, $\mathbf{H} \in \mathbb{R}^{m\times n}$ is the Jacobian of $\mathcal{H}(\mathbf{x})$ at $\mathbf{x}^b$, the information matrix reads:

$$\mathbf{A} = \left[\mathbf{R} + \mathbf{H}\cdot\mathbf{B}\cdot\mathbf{H}^T\right] \in \mathbb{R}^{m\times m}\,, \tag{3}$$

and $\mathbf{R} \in \mathbb{R}^{m\times m}$ is the estimated data-error covariance matrix. In practice, an ensemble of model realizations can be employed to estimate the parameters $\mathbf{x}^b$ and $\mathbf{B}$ of prior error distributions. However, given the computational cost of a single model propagation, ensemble sizes are constrained by the hundreds while their underlying error distribution by the millions. Consequently, sampling errors impact the quality of analysis innovations: ensemble covariances are rank-deficient, and even more, they are ill-conditioned [1, 32]. Thus, spurious correlations among distant model components are developed in the ensemble covariance [29]. Localization methods are commonly employed during assimilation steps to mitigate the impact of sampling noise. In this context, well-known methods are covariance matrix localization, precision matrix localization, spatial domain localization, and observation impact localization. The selection of one method over the others relies on computational aspects. Yet another manner to mitigate the impact of spurious correlations is based on Shrinkage Covariance Matrix Estimation. In this family of covariance matrix estimators, the background error covariance matrix is estimated as the convex combination of a target matrix $\mathbf{T} \in \mathbb{R}^{n\times n}$, and the ensemble covariance $\mathbf{P}^b \in \mathbb{R}^{n\times n}$:

$$\widehat{\mathbf{B}} = \gamma\cdot\mathbf{T} + (1-\gamma)\cdot\mathbf{P}^b \in \mathbb{R}^{n\times n}\,,\ \text{for}\ \gamma \in [0,\ 1]\,. \tag{4}$$

The current literature proposes ensemble-based formulations via the covariance estimator (4) in which:

1. the target matrix $\mathbf{T}$ is diagonal (no prior structure is assumed for $\mathbf{B}$), and the weight γ is optimally computed via loss functions [30, 31], or
2. the target matrix $\mathbf{T}$ is static (i.e., it retains climatological information), and the weight γ is ranged in $\gamma \in [0,\ 1]$ [43, 44].

We exploit the opportunity to include our prior knowledge about the structure of $\mathbf{B}$, the information brought by samples from the model dynamics, and the optimal estimation of γ. In this manner, we can obtain a covariance matrix estimator of $\mathbf{B}$ that optimally combines all sources of information. While several techniques have been proposed to reduce spurious correlations, most of them are designed for a specific problem, and it is not possible to generalize them for other DA implementations [10, 24]. We are looking for a robust and generalizable manner to include previous knowledge of the system to a large scale Chemical Transport Model (CTM) for air quality purposes. Data assimilation is not necessarily the most popular technique to incorporate reality into a CTM model. Some practitioners prefer to spend more time developing emissions inventories rather than incorporate ground data, satellite information or vertical measurements. Nevertheless, some applications have been made recently for the CTM LOTOS-EUROS [10, 17, 24], and the particularity of the advection and diffusion dynamics govern and condition the emission and deposition processes. For the north of South America, the highly non-linear and chaotic behavior of weather dynamics mixed with a complex topography and a lack of emissions inventory is indeed a challenge for description and forecast of air pollution. Figure 1 shows an example based on the high-resolution application of the LOTOS-EUROS model to study the behavior of PM_{10} and $PM_{2.5}$ over the Metropolitan Area of the Aburrá Valley in Colombia [23].

Here, it is possible to see how the complex topography that is not well captured by the meteorology and the model conditions the dynamical relations between the states. Traditional localization techniques as covariance localization to avoid spurious correlations are not suitable nor direct applicable to this problem. The idea of design a covariance matrix where the knowledge of the system can be integrated into the DA process comes from this application and its related difficulties using current localization techniques.

This paper is organized as follows: Section 2 discusses well-known issues in ensemble-based data assimilation and how to overcome those. In Section 3, we propose an efficient

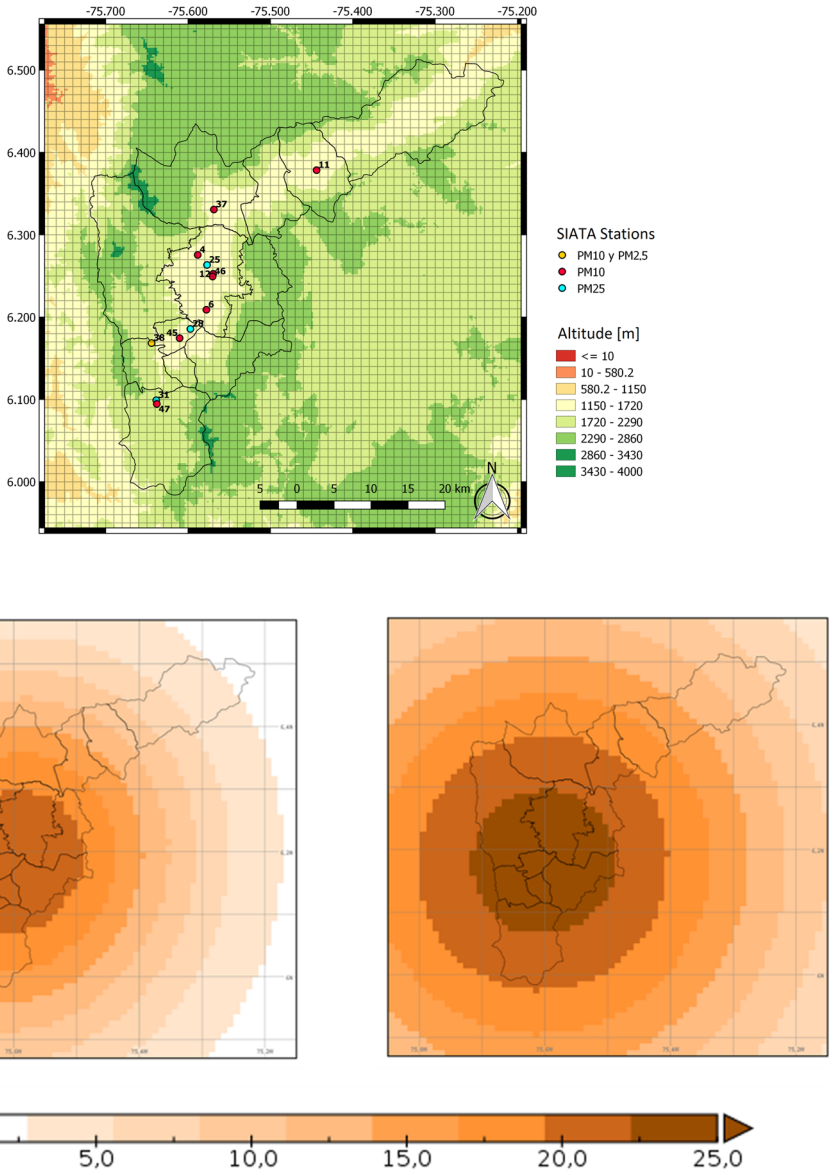

Fig. 1 Topography and correction emission factors for different localization radius using and standard localization technique [23]

ensemble-based method via Shrinkage Covariance Matrix Estimation, which accounts for Prior Knowledge about the background error correlations, localization, and inflation aspects are discussed as well. Experimental tests are performed in Section 4. Two models are employed during the experiments: the Advection Diffusion Model and the high-nonlinear model SPEEDY. Conclusions from this research are stated in Section 5.

2 Preliminaries

In order to state the value of the current contribution, several questions must be solved to demonstrate the feasibility of the new data assimilation technique in an operational fashion [46]: *Does the new method provide guidance that is of higher quality or more use than existing methods? Is the potential benefit of running a new technique cost-effective? Is the new method sufficient with respect to old methods?*. In this section, we discuss ensemble-based data assimilation methods and how those can be implemented in current operational settings. These concepts are necessary to develop our filter formulation.

2.1 Ensemble-based data assimilation

In ensemble-based data assimilation, an ensemble of model realizations

$$\mathbf{X}^b = \left[\mathbf{x}^{b[1]}, \mathbf{x}^{b[2]}, \ldots, \mathbf{x}^{b[N]}\right] \in \mathbb{R}^{n\times N}, \tag{5}$$

is employed to estimate the parameters $\mathbf{x}^b$ and $\mathbf{B}$ of prior error distributions, where $\mathbf{x}^{b[e]} \in \mathbb{R}^{n\times 1}$ is the e-th ensemble

member, for $1 \leq e \leq N$, and N stands for ensemble size. Hence:

$$\mathbf{x}^b \approx \overline{\mathbf{x}}^b = \frac{1}{N} \cdot \sum_{e=1}^{N} \mathbf{x}^{b[e]} \in \mathbb{R}^{n \times 1}, \tag{6}$$

and

$$\mathbf{B} \approx \mathbf{P}^b = \frac{1}{N} \cdot \boldsymbol{\Delta}\mathbf{X} \cdot \boldsymbol{\Delta}\mathbf{X}^T \in \mathbb{R}^{n \times n}, \tag{7}$$

where

$$\boldsymbol{\Delta}\mathbf{X} = \mathbf{X}^b - \overline{\mathbf{x}}^b \cdot \mathbf{1}^T \in \mathbb{R}^{n \times N}, \tag{8}$$

is the matrix of member deviations, $\overline{\mathbf{x}}^b$ is the ensemble mean, $\mathbf{P}^b$ is the ensemble covariance, and $\mathbf{1}$ is a vector of the consistent dimension whose components are all ones. Once an observation is available, the posterior state can be computed via the stochastic Ensemble Kalman Filter (EnKF) [9]:

$$\mathbf{X}^a = \mathbf{X}^b + \mathbf{P}^b \cdot \mathbf{H}^T \cdot \left[\mathbf{R} + \mathbf{H} \cdot \mathbf{P}^b \cdot \mathbf{H}^T\right]^{-1} \cdot \mathbf{D} \in \mathbb{R}^{n \times N}, \tag{9}$$

where the e-th column of the innovation matrix on the synthetic observations $\mathbf{D} \in \mathbb{R}^{n \times N}$ reads $\mathbf{d}^{[e]} = \mathbf{y} + \boldsymbol{\epsilon}^{[e]} - \mathcal{H}\left(\mathbf{x}^{b[e]}\right) \in \mathbb{R}^{m \times 1}$, with $\boldsymbol{\epsilon}^{[e]} \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$. In practice, ensemble sizes are constrained by the hundreds, while model resolutions are bounded by the millions, which mainly obey computational aspects. Consequently, the quality of analysis corrections can be impacted by spurious correlations. Hence, localization methods can be employed to mitigate the impacts of sampling errors. Well-known methods in this context are covariance matrix localization, spatial domain localization, and observation localization.

2.2 Covariance Matrix Localization

For small ensemble sizes, sampling errors can impact the quality of covariances in Eq. 7. As a direct consequence problems such as filter divergence and long range spurious correlations can appear [1]. Localization is based on the assumption that two distant parts of the system are independent for most geophysical systems. The two main localization methods are: domain localization and covariance localization. Domain localization, also called *local analysis*, instructs that instead of performing a global analysis for the complete domain, a local analysis can be applied using just local observations [3]. Covariance localization cuts off longer-range correlations in the error covariances at a specified distance [13, 33]. The localization is performed throw Schur product denoted by $\circ$:

$$[\rho \circ \boldsymbol{P}^b]_{i,j} = [\mathbf{P}^b]_{i,j} \cdot [\rho]_{i,j}. \tag{10}$$

Positive definite covariance matrices can be built with the Gaspari-Cohn function $G(d)$ [11]:

$$G(d) = \begin{cases} \text{if } 0 \leq d < 1: & 1 - \frac{5}{3}r^2 + \frac{5}{8}r^3 + \frac{1}{2}r^4 - \frac{1}{4}r^5 \\ \text{if } 1 \leq d < 2: & 4 - 5r + \frac{5}{3}r^2 + \frac{5}{8}r^3 - \frac{1}{2}r^4 + \frac{1}{12}r^5 - \frac{2}{3r} \\ \text{if } d > 2: & 0 \end{cases}. \tag{11}$$

A cutoff function would be defined by $d \in \mathbb{R}^+ \rightarrow G(d/r)$, where r is a length scaled called the localization radius [3, 37]. The regularized $\rho \circ \mathbf{P}^b$ is used as a replacement for $\mathbf{P}^b$.

2.3 Shrinkage covariance matrix estimation

A more robust family of covariance estimators under the DA case $n \gg N$ are the shrinkage based estimators [8, 42]. This kind of estimators follow the form [22]:

$$\mathbf{B} \approx \widehat{\mathbf{B}}(\alpha) = \alpha \cdot \mathbf{T} + (1 - \alpha) \cdot \mathbf{P}^b \in \mathbb{R}^{n \times n}, \tag{12}$$

where $\alpha \in [0, 1]$, and $\mathbf{T} \in \mathbb{R}^{n \times n}$ is known as the Target matrix. The resulting estimator is a convex combination of the ensemble covariance matrix and the pre-defined $\mathbf{T}$ matrix. When there is not available information about the structure of $\mathbf{B}$, an alternative for $\mathbf{T}$ is [31]:

$$\mathbf{T} = \frac{\mathbf{trace}\left(\mathbf{P}^b\right)}{n} \cdot \mathbf{I}, \tag{13}$$

where $\mathbf{I} \in \mathbb{R}^{n \times n}$ is the identity matrix. The value of α is chosen to minimize the loss function

$$\alpha^* = \arg\min_{\alpha} \mathbb{E}\left[\left\|\mathbf{B} - \widehat{\mathbf{B}}(\alpha)\right\|_F^2\right], \tag{14}$$

where $\|\bullet\|_F$ represents the Frobenius norm. For target matrices of the form Eq. 13, a distribution-free formulation for the optimal α_{LW}^* is proposed by Ledoit and Wolf in [20]:

$$\alpha_{LW}^* = \min\left(\frac{\sum_{e=1}^{N}\left\|\mathbf{P}^b - \boldsymbol{\Delta}\mathbf{x}^{[e]} \cdot \boldsymbol{\Delta}\mathbf{x}^{[e]T}\right\|_F^2}{N^2 \cdot \left[\mathbf{trace}\left(\mathbf{P}^{b2}\right) - \frac{\mathbf{trace}\left(\mathbf{P}^b\right)^2}{n}\right]}, 1\right), \tag{15}$$

where $\boldsymbol{\Delta}\mathbf{x}^{[e]} \in \mathbb{R}^{n \times 1}$ denotes the e-th column of the matrix (8). Based on the LW estimator, for Gaussian samples, the Rao-Blackwell Ledoit and Wolf (RBLW) one is proposed. In the RBLW estimator, the optimal weight is defined by:

$$\alpha_{RBLW}^* = \min\left(\frac{\frac{N-2}{n} \cdot \mathbf{trace}\left(\mathbf{P}^{b2}\right) + \mathbf{trace}^2\left(\mathbf{P}^b\right)}{(N+2) \cdot \left[\mathbf{trace}\left(\mathbf{P}^{b2}\right) - \frac{\mathbf{trace}^2\left(\mathbf{P}^b\right)}{n}\right]}, 1\right). \tag{16}$$

An EnKF implementation which exploits the special structure of this estimator is the EnKF based on the RBLW

estimator (EnKF-RBLW) wherein the posterior ensemble can be built as follows [30, 31]:

$$\widehat{\mathbf{B}}_{RBLW} = \alpha^*_{RBLW} \cdot \mu \cdot \mathbf{I} + (1 - \alpha^*_{RBLW}) \cdot \mathbf{P}^b, \tag{17a}$$

$$\mathbf{X}^a_{RBLW} = \mathbf{X}^b + \widehat{\mathbf{B}}_{RBLW} \cdot \mathbf{H}^T \cdot \left[\mathbf{R} + \mathbf{H} \cdot \widehat{\mathbf{B}}_{RBLW} \cdot \mathbf{H}^T\right]^{-1} \cdot \mathbf{D}, \tag{17b}$$

$$\mu = \frac{\mathbf{trace}\left(\mathbf{P}^b\right)}{n}. \tag{17c}$$

Since numerical models can be highly non-linear, Gaussian assumptions on prior members are commonly broken. This assumption can be relaxed in the EnKF context by employing, for instance, the LW estimator for the estimation of background error covariance matrices during assimilation steps [28]. Besides, different prior structures can be treated in $\mathbf{T}$ to enrich the covariance matrix estimation, this is, to account for prior information about the dynamical system.

3 An ensemble Kalman Filter via shrinkage covariance matrix estimation and prior knowledge

In this Section, a novel EnKF implementation that incorporates prior knowledge of the background error covariance matrix in a practical manner to improve the DA process is presented. The method is based on a shrinkage estimator using a general target matrix. An efficient and totally parallelizable implementation of the method for high-dimensional systems is also proposed.

3.1 Filter derivation

As was mentioned above, shrinkage based covariance matrix estimators which allow the use of a target matrix $\mathbf{T}$ to structure the covariance matrix, are limited to a target matrix with identity matrix structure [31, 39]. Although matrix identity structure can reduce the spurious correlations caused by the ill-conditioned approximation of the error covariance matrix [7, 21, 30], the assumption of a covariance structure without correlation between the states is not always valid or desirable. Using a general target matrix enables the incorporation of prior information about the system into the error covariance matrix. This prior information can be information about the system physics as for instance, parameters, topography, transport phenomena and environmental information, or knowledge about the covariance structure coming from experts or previous experiments. A close formulation to calculate the weight value α using a general target matrix $\mathbf{T}_{KA}$ is proposed in [39, 45],

$$\alpha_{KA} = \min\left(\frac{\frac{1}{N^2} \cdot \sum_{e=1}^{N} \left\|\mathbf{\Delta x}^{[e]}\right\|^4 - \frac{1}{N} \cdot \left\|\mathbf{P^b}\right\|^2}{\left\|\mathbf{P}^b - \mathbf{T}_{KA}\right\|^2}, 1\right). \tag{18a}$$

and the KA (Knowledge-Aided) estimator is obtained using (18a) in

$$\hat{\mathbf{B}}_{KA} = \alpha_{KA} \cdot \mathbf{T}_{KA} + (1 - \alpha_{KA}) \cdot \mathbf{P}^b \in \mathbb{R}^{n \times n}, \tag{18b}$$

It is important to note that no assumptions about the structure of $\mathbf{T}_{KA}$ are made to calculate α_{KA}. This approach can be seen as an extension of that in [6, 21] to a general target matrix and is usable for complex-value data case [1]. Similar to the EnKF-RBLW an implementation of the EnKF can be obtained using the KA shrinkage-based estimator presented in Eq. 18:

$$\mathbf{X}^a = \mathbf{X}^b + \hat{\mathbf{B}}_{KA} \cdot \mathbf{H}^T \cdot [\mathbf{R} + \mathbf{H} \cdot \hat{\mathbf{B}}_{KA} \cdot \mathbf{H}^T] \cdot \mathbf{D},$$

Since the target matrix $\mathbf{T}_{KA}$ in the EnKF-KA is not necessarily a matrix with identity structure, information about the dynamical system can be integrated into the data assimilation process. The prior information is directly related to the error covariance of the model states; this means that it is possible to integrate information of the system and guide the dynamical relationship between the states and the relation between states and observations. Although there are no restrictions in the structure of $\mathbf{T}_{KA}$, it is important to remarks that $\mathbf{T}_{KA}$ is still a covariance matrix, so all the related conditions have to be accomplished. In Section 4 are shown examples of how to select $\mathbf{T}_{KA}$ properly.

3.2 Domain localization

Both most popular concepts of localization can be applied in the EnKF-KA approach: covariance localization [13, 14], and local domain analysis [32]. We explore the implementation of local domain analysis due to the advantages not only in the spurious correlation mitigation but also in the implementations. Since the main idea of the EnKF-KA is to incorporate prior information of the system in the DA framework, it is inherent that this information has to be saved and available in all the DA processes. In high-dimensional applications, it is not convenient and, in some cases, prohibitive to save a matrix of the dimension of $\mathbf{T}_{KA} \in \mathbb{R}^{n \times n}$, and calculate $\mathbf{P}^b \in \mathbb{R}^{n \times n}$ directly. It is here where the concept of local domains is crucial for

[1] The reader can consult [39, 45] for additional information.

the implementations of the EnKF-KA for high-dimensional systems. In local domains, a box of radius r of components around the state of interest is created, and just the states and observations within this box (local domain) are used in the analysis step [16, 32, 38]. This process is repeated for all the state components, doing multiple local analysis (in a smaller dimension) instead of a unique and global analysis (in a higher dimension). Another advantage of this implementation is that it facilitates the parallelization of the analysis since each local analysis can be performed in an independent core [12, 31]. The implementation of the EnKF-KA using local domains analysis is summarized in the next steps:

1. A local domain of radius r is created for any model component. The $k-th$ local domain is formed by n_r ($n_r << n$) and m_r observation. The use of domain decomposition is applied, so that boundary information is shared across neighboring domains. In this manner, we preserve the continuous dynamics of some physical variables such as Temperature, Wind Components, and Pressure. Figure 2 illustrates this strategy. The background ensemble and the analysis ensemble into the box is denoted by $\mathbf{X}_k^b \in \mathbb{R}^{n_r \times N}$ and $\mathbf{X}_k^a \in \mathbb{R}^{n_r \times 1}$ respectively. The covariance model error into the box are denoted by $\mathbf{B}_k \in \mathbb{R}^{n_r \times n_r}$, the local observation is denoted by $\mathbf{y}_k \in \mathbb{R}^{m_r \times 1}$ with observation covariance $\mathbf{R}_k \in \mathbb{R}^{m_r \times m_r}$, and the local innovation matrix is denoted by $\mathbf{D}_k \in \mathbb{R}^{m_r \times N}$.
2. Compute the local sample covariance matrix $\mathbf{P}_k^b \in \mathbb{R}^{n_r \times n_r}$

$$\boldsymbol{\Delta}\mathbf{X}_k^b = \mathbf{X}_k^b - \bar{\mathbf{x}}_k^b \cdot \mathbf{1}_N^T, \tag{19a}$$

$$\mathbf{P}_k^b = \frac{1}{(N-1)} \cdot \boldsymbol{\Delta}\mathbf{X}_k \cdot (\boldsymbol{\Delta}\mathbf{X}_k)^T. \tag{19b}$$

3. Define the local target matrix $\mathbf{T}_k \in \mathbb{R}^{n_r \times n_r}$. On this step, the use of previous knowledge of the model dynamics is required. Knowledge is understood as the human-based experience in front of a large scale model used to represent reality. Large scale models for atmospheric dynamics, weather, water and ocean, reservoir modeling are used normally by experts in their fields. Even if the data to be assimilated is measured, some details and specifications are not captured on the model or included on it. Other possible causes are that due to the spatial-temporal resolution chosen for the numerical solution of the equations, it does not allow to capture intrinsic relationships between the states. We suggest a matrix $\mathbf{T}_k$ built on the basis of that specific knowledge. Although $\mathbf{T}_k$ must meet all requirements of a covariance matrix, the main contribution is that the matrix $\mathbf{T}_k$ must not fulfill any requirement about its structure and also can change dynamically.

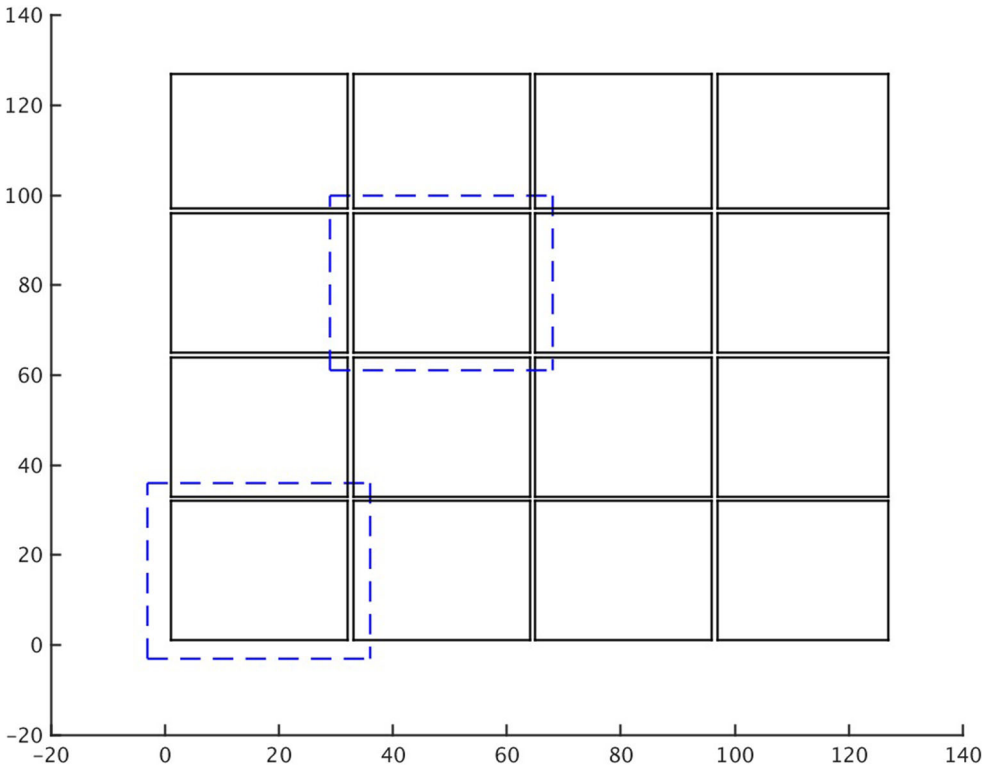

Fig. 2 Domain decomposition is exploited to reduce the computational cost of our proposed method. Dashed regions denote the shared boundary information to be employed during assimilation steps

4. Estimate the local error covariance $\mathbf{B}_k$ throw the KA shrinkage-based estimator $\hat{\mathbf{B}}_k$ using

$$\alpha_k = \min\left(\frac{\frac{1}{N^2}\cdot\sum_{e=1}^{N}\left\|\mathbf{\Delta x}_k^{[e]}\right\|^4 - \frac{1}{N}\cdot\left\|\mathbf{P}_k^b\right\|^2}{\left\|\mathbf{P}_k^b - \mathbf{T}_k\right\|^2}, 1\right), \quad (20a)$$

$$\hat{\mathbf{B}}_k = \alpha_k \cdot \mathbf{T}_k + (1-\alpha_k)\cdot\mathbf{P}_k^b \in \mathbb{R}^{n_r \times n_r}. \quad (20b)$$

5. Perform the local analysis step

$$\mathbf{X}_k^a = \mathbf{X}_k^b + \hat{\mathbf{B}}_k \cdot \mathbf{H}_k^T \cdot [\mathbf{R}_k + \mathbf{H}_k \cdot \hat{\mathbf{B}}_k \cdot \mathbf{H}_k^T] \cdot \mathbf{D}_k. \quad (21)$$

6. Once all the local analyses are performed, map those to the global domain. The global analysis state is then obtained. This does not mean to perform a new global analysis. In [32] two map approaches are proposed. The first one uses only the analysis results at the center point of each local region to form the global analysis vectors. The second one uses the average of all the local analysis where a grid cell is involved in obtaining the global analysis.

Note that with a correct selection of r, the matrix computations in each local domain are inexpensive, so Eq. 20 can be computed efficiently for high-dimensional systems.

3.3 Inflation aspects

In the context of EnKF-KA, the covariance inflation can be efficiently performed increasing the dispersion of matrix (8) by a inflation factor β_{inf} :

$$\widehat{\mathbf{\Delta X}} = \beta_{\text{inf}} \cdot \mathbf{\Delta X} \in \mathbb{R}^{n\times N}, \quad (22)$$

and by noting that:

$$\mathbf{tr}\left(\beta_{\text{inf}}^2 \cdot \mathbf{P}^b\right) = \beta_{\text{inf}}^2 \cdot \mathbf{tr}\left(\mathbf{P}^b\right),$$

where **tr** represent the trace of the matrix. For instance, we can see that covariance inflation on the optimal factor (18a) reads:

$$\alpha_{KA}^{\text{inf}} = \min\left(\frac{\frac{1}{N^2}\cdot\sum_{e=1}^{N}\beta_{\text{inf}}^8\cdot\left\|\mathbf{\Delta x}^{[e]}\right\|^4 - \frac{1}{N}\cdot\beta_{\text{inf}}^2\cdot\left\|\mathbf{P}^b\right\|^2}{\left\|\beta_{\text{inf}}^2\cdot\mathbf{P}^b - \mathbf{T}_{KA}\right\|^2}, 1\right).$$

4 Experimental settings

4.1 Results with an advection diffusion model

This section illustrates the proposed EnKF-KA over simple a advection-diffusion process. The advection-diffusion governs the changes of a conservative property such as the concentration in a fluid environmental [36, 41]. The advection-diffusion equation has been used as a simple model to study the behavior and transport of pollutants in the atmosphere. In two dimensions, the horizontal changes in the concentration of a determinate pollutant **C** in the atmosphere can be approximated as:

$$\frac{\partial C}{\partial t} = D_x\frac{\partial^2 C}{\partial x^2} - v_x\frac{\partial C}{\partial x} + D_y\frac{\partial^2 C}{\partial y^2} - v_y\frac{\partial C}{\partial y} + E(t), \quad (23)$$

where v_x and u_y are the north-south and west-east wind velocities respectively, D_x and D_y are the north-south and west-east diffusion coefficients respectively, and $E(t)$ are the emissions. The experimental settings are:

- The continuous advection-diffusion equation is discretized in a 20×20 domain, obtaining a total of $n = 400$ states representing concentration in each cell.
- The boundary condition used for solving the experiment was the Dirichlet homogeneous zero or null value fixed in the contour.
- Ten emissions points are considered. Additionally, to represent a real scenario where the emissions are the most important uncertainty sources in the atmosphere chemistry modelling [4], uncertainty in every time in the emissions are considered.
- There is no considered uncertainty in initial conditions, boundary conditions, or parameter values.
- With the idea of simulating an imperfect representation of the model environment, an artificial valley is performed in the real scenario, where the true state $\mathbf{x}^*$ and observations $\mathbf{y}$ are taken. The artificial valley is created, increasing the diffusion coefficients and reducing the velocity winds components in a determinate number of cells. This implies that the interchange of pollutants between two locations, one inside and the other outside the valley, is considerably lower than two locations outside or inside the valley. The valley is not included in the model used for assimilation purposes. A graphical representation is shown in Fig. 3.
- A background ensemble is built perturbing the 10 emission points by drawing a sample from the Normal distribution,

$$\mathbf{x}^{b[e]} \sim \mathcal{N}(\overline{\mathbf{x}}^b, \rho_b^2 \cdot \mathbf{I}), \text{ for } 1 \le e \le N, \quad (24)$$

  where $\rho_b = 0.05$
- We propose three ensemble sizes for the experiments $N \in \{10, 50, 100\}$.
- The assimilation window consists of $M = 1000$ time steps. Two observation periods are proposed for the test, each time step and each ten time steps. We denote by $\delta t \in \{1, 10\}$ the elapsed time between two observations.
- The error statistics are associated with the Gaussian distribution,

$$\mathbf{y}_\ell \sim \mathcal{N}(\mathcal{H}_\ell(\mathbf{x}_\ell^*), \rho_o^2 \cdot \mathbf{I}), \text{ for } 1 \le \ell \le M, \quad (25)$$

  where $\rho_o = 0.001$.

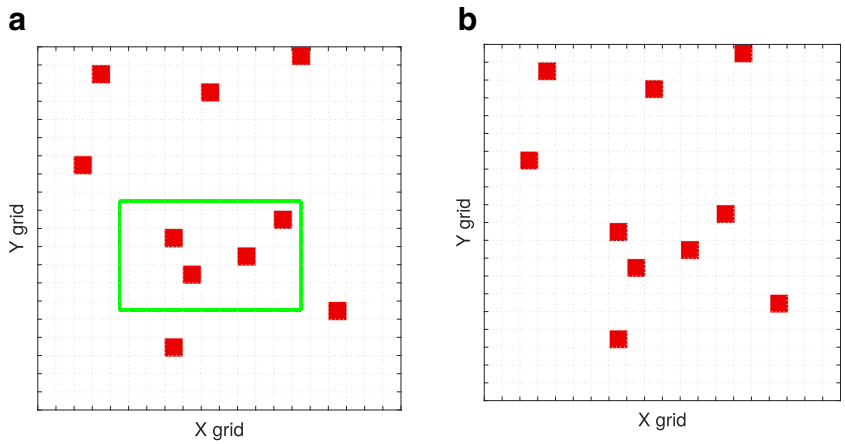

Fig. 3 Comparison of the real scenario vs the model scenario. The green line represents the artificial valley. The red squares represent the emission points. **a** Real scenario. **b** Model for DA purpose

- We consider two fractions of observed components $s \in \{0.12, 0.5\}$. The components are randomly chosen at each assimilation step.
- The L_2 norm of errors are utilized as a measure of accuracy at the assimilation step ℓ,

$$L_\ell = \sqrt{\left[\mathbf{x}_\ell^a - \mathbf{x}_\ell^*\right]^T \cdot \left[\mathbf{x}_\ell^a - \mathbf{x}_\ell^*\right]}, \tag{26}$$

  where $\mathbf{x}_\ell^*$ and $\mathbf{x}_\ell^a$ are the reference and the analysis solution respectively.
- The Root-Mean-Square-Error (RMSE) is used as a measure of performance, in average, on a given assimilation window,

$$\text{RMSE} = \frac{1}{M} \cdot \sum_{\ell=1}^{M} \lambda_\ell^2 , \tag{27a}$$

  where

$$\lambda_\ell = \left\| \mathbf{x}_\ell^a - \mathbf{x}_\ell^* \right\|_2 . \tag{27b}$$

- The percentage of non converge experiments (PNCE) is calculated for all the scenarios.

The idea is to incorporate the physical restrictions that the model does not capture, for this case, the artificial valley, via the EnKF-KA. If we use a standard distance-based localization for a state into the valley to cut the coming information from distant observations, the process will include both observations inside and outside the valley. With the EnKF-KA, we try to cut observations that are outside the valley, even if there are at the same distance, as is represented in Fig. 4.

This is achieved by incorporating the physical restrictions (the topography of the interest domain) into the covariance estimation throw the target matrix $\mathbf{T}_{KA}$. The target matrix is built starting from a Gaspari-Cohn function [11] and reducing to zero the covariance between the states inside and outside the valley. After this process, it is very important to test whether the final $\mathbf{T}_{KA}$ is still a positive semidefinite matrix. Note that the final covariance between the state inside and outside the valley will not be necessary zero because the final covariance matrix is a convex combination of $\mathbf{T}_{KA}$ and $\mathbf{P}^b$. In Fig. 5 is shown an example of a $\mathbf{T}_{KA}$ matrix obtained using the proposed process for an influence radius $r = 4$.

The performance of the EnKF-KA is compared with the shrinkage-based EnKF-RBLW and the standard EnKF using covariance localization EnKF-CL with $r = 1$ (other influence radii were tested, but $r = 1$ presents the best performance) under the experimental setup presented below. A total of 20 experiments are performed for each scenario. The target matrix $\mathbf{T}_{KA}$ is built from a Gaspari-Cohn with $r = 1$ and following the mentioned process including physical restrictions of the valley. The magnitude of $\mathbf{T}_{KA}$ is computed according to the average of the trace of $\mathbf{P}^b$. In Fig. 6 is shown the dynamical evolution of the L_2 norm for different scenarios. Figure 7 presents the values of the average RMSE for all the experiment scenarios and the PNCE for the EnKF-CL for the different ensemble members value. For the EnKF-RBLW and the EnKF-KA the $PNCE = 0\%$ for all the cases.

As is shown in Figs. 6 and 7, the EnKF-KA presents a lower error rates than the EnKF-RBLW and the EnKF-CL in almost all the scenarios. This shows how the integration of the physical restrictions can help the data assimilation process. It is interesting to evaluate the scenarios with a smaller number of ensemble members, where the differences among the three algorithms are more considerable. The RMSE value of the EnKF-KA in these scenarios is much lower than the EnKF-CL, showing that shrinkage-based estimators are more robust than the sample covariance matrix when $n >> N$. Since the ensemble statistics approximate the mean and the covariance of the state, the ensemble spread should describe the system uncertainty [15, 40]. If the filter estimates the state uncertainty correctly, the ensemble spread should matches with the RMSE when there are no model errors

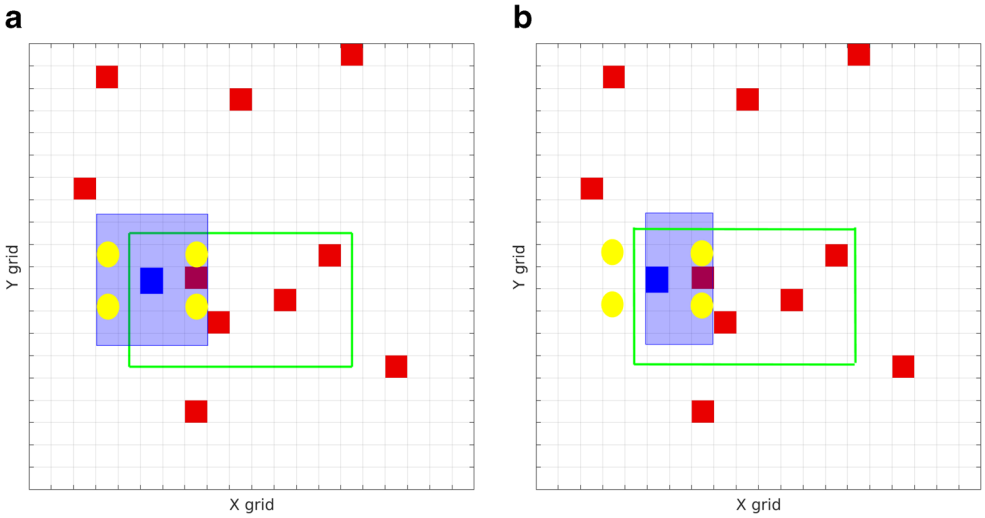

Fig. 4 Comparison of the distance-based localization approach vs the EnKF-KA. In the EnKF-KA the influence region is based on the distance and on the information about the system. The blue square represents the analyzed state, the blue shadow the influence region, and the yellow circles represent the observations. **a** Distance-based localization. **b** EnKF-KA

[27]. Figure 8 shows the ensemble spread of each algorithm among assimilation steps for a specific experiment. It can be seen how all the algorithms reduce the ensemble spread after few assimilation steps, reducing the system uncertainty levels. The Free-Run keep similar uncertainty values among time because no new information is incorporated. Finally, the EnKF-KA presents the lowest spread values matching with the lowest RMSE values, which means that the ENKF-KA can correctly reproduce the system uncertainty and improve estimation accuracy.

In Fig. 9 is presented the time evolution of states in four different spacial location for one experiment scenario. It is evident that the EnKF-KA reproduces more accurately locations in the border of the artificial valley than the other methods, showing the effect of the incorporated information throw $\mathbf{T}_{KA}$.

An aspect that is important to remarks is the value of α_{KA} for different ensemble member values. The mean α_{KA} value for ensemble number of $N = 10$, $N = 50$ and $N = 100$ are $\bar{\alpha_{10}} = 0.698$, $\bar{\alpha_{50}} = 0.591$ and $\bar{\alpha_{100}} = 0.508$. With a small number of ensemble members the assumption of a poor estimation of the covariance throw the sample covariance matrix produces a higher value of α_{KA}, giving more weight to the target matrix than when the number of ensemble, and the quality of the estimation throw the sample covariance matrix, is higher.

4.2 Results with an atmospheric general circulation model

SPEEDY (Simplified Parameterizations, privitivE-Equation DYnamics) is an Atmospheric General Circulation model [5, 25], which help us to study the performance of the EnKF-KA method in a highly non-linear model scenario. The model consists of seven numerical layers, and at each one, a T-30 model resolution is employed (96×48 grid components) [19, 26]. The total number of physical variables at each numerical grid point are five. These are the temperature T (K), the zonal u and the meridional v wind components (m/s), the specific humidity Q (g/kg), and the

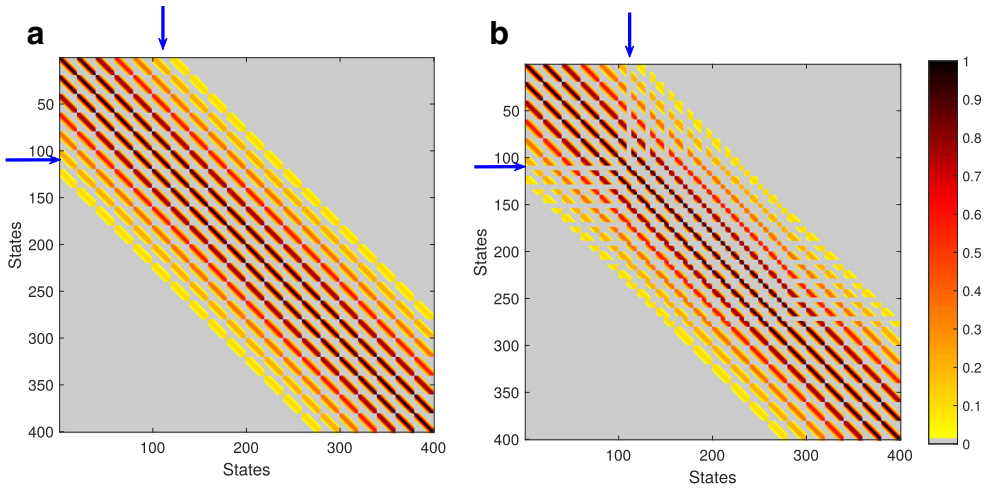

Fig. 5 Graphical representation of the $\mathbf{T}_{KA}$ matrix. The arrows remark the state 110, which is located just in the inside border of the valley (represented as a blue square in Figure 4), and show how the covariance between a state inside and the states outside the valley is fixed in 0. **a** Gaspari-Cohn function. **b** Target matrix TKA

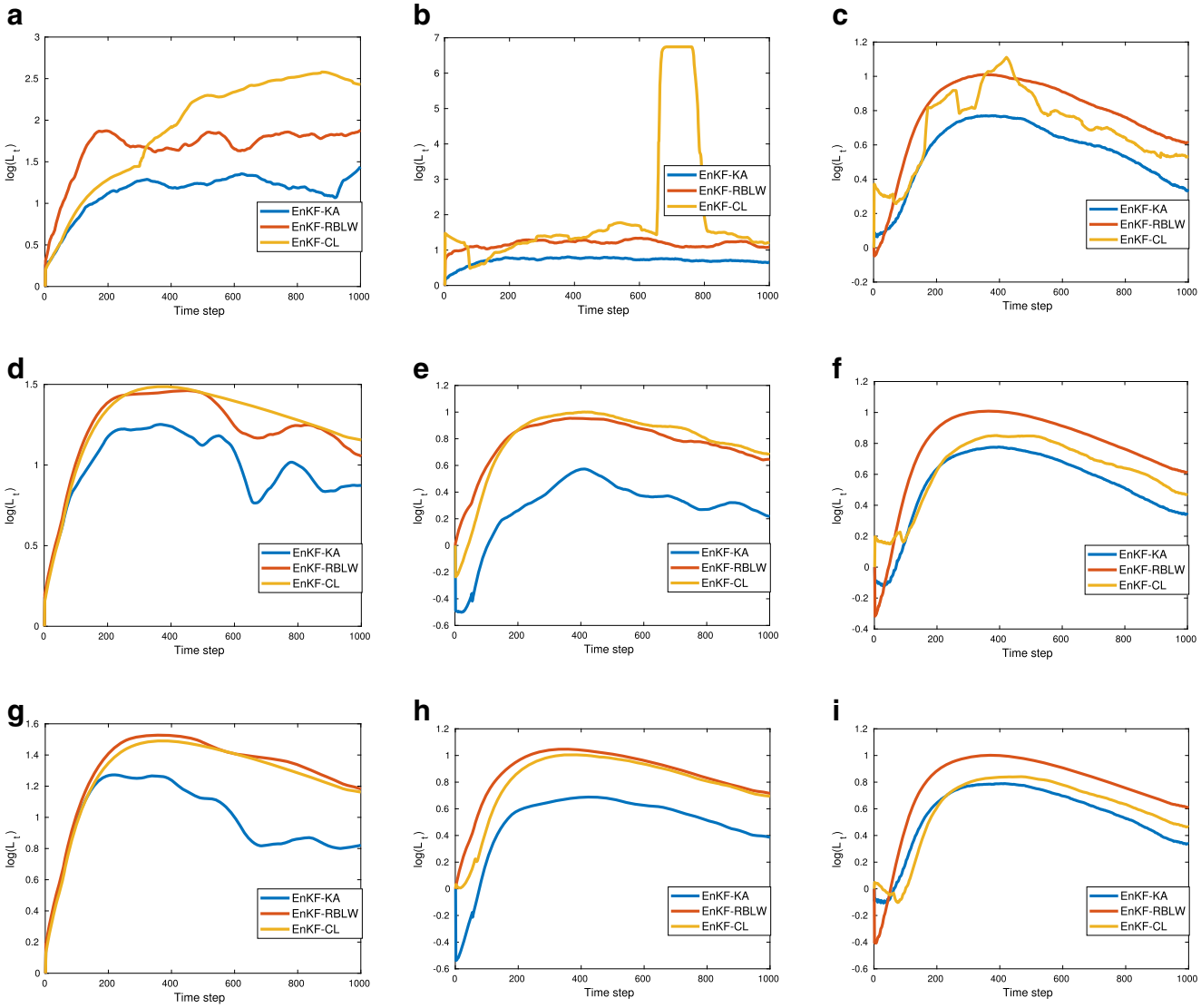

Fig. 6 Comparison of the performance among the EnKF-KA, EnKF-RBLW and EnKF-CL for some scenarios. **a** $N = 10, \delta t = 1, s = 0.12$. **b** $N = 10, \delta t = 1, s = 0.5$. **c** $N = 10, \delta t = 10, s = 0.5$. **d** $N = 50, \delta t = 1, s = 0.12$. **e** $N = 50, \delta t = 1, s = 0.5$. **f** $N = 50, \delta t = 10, s = 0.5$. **g** $N = 100, \delta t = 1, s = 0.12$. **h** $N = 100, \delta t = 1, s = 0.5$. **i** $N = 100, \delta t = 10, s = 0.5$

pressure ρ (hPa). We employ all physical variables into our data assimilation process. Note that, the model dimension in our settings reads $n = 133,632$. During our experiments, we consider ensemble sizes of $N = 10$ and $N = 20$, this applies for all numerical scenarios. Note that, model resolutions are 13,632 and 6,685 times larger than ensemble sizes ($n \gg N$), which takes to current DA operational settings. We follow the experimental settings presented in [18, 28]:

- Long term numerical integrations are applied to build the reference solution as well as the initial background ensemble (two years of a numerical simulation). We start with a system in equilibrium, and after adding a small perturbation, the numerical integration is performed.
- The experiments do not account for model errors.
- Standard deviations of observational errors are detailed in Table 1.
- We employ a highly sparse observational network. The observation coverage is 9% of the spatial resolution. This linear observation operator is shown in Fig. 10. Note that this is an irregularly distributed, realistic observational network.
- The inflation factor is $\beta_{\text{inf}} = 1.3$ for all experiments.
- We set up a total simulation time of two months with observations frequencies about 6 and 12 hours. We expect the non-linear dynamics of the SPEEDY model to impact the quality of analysis states as the observation frequency decreases.

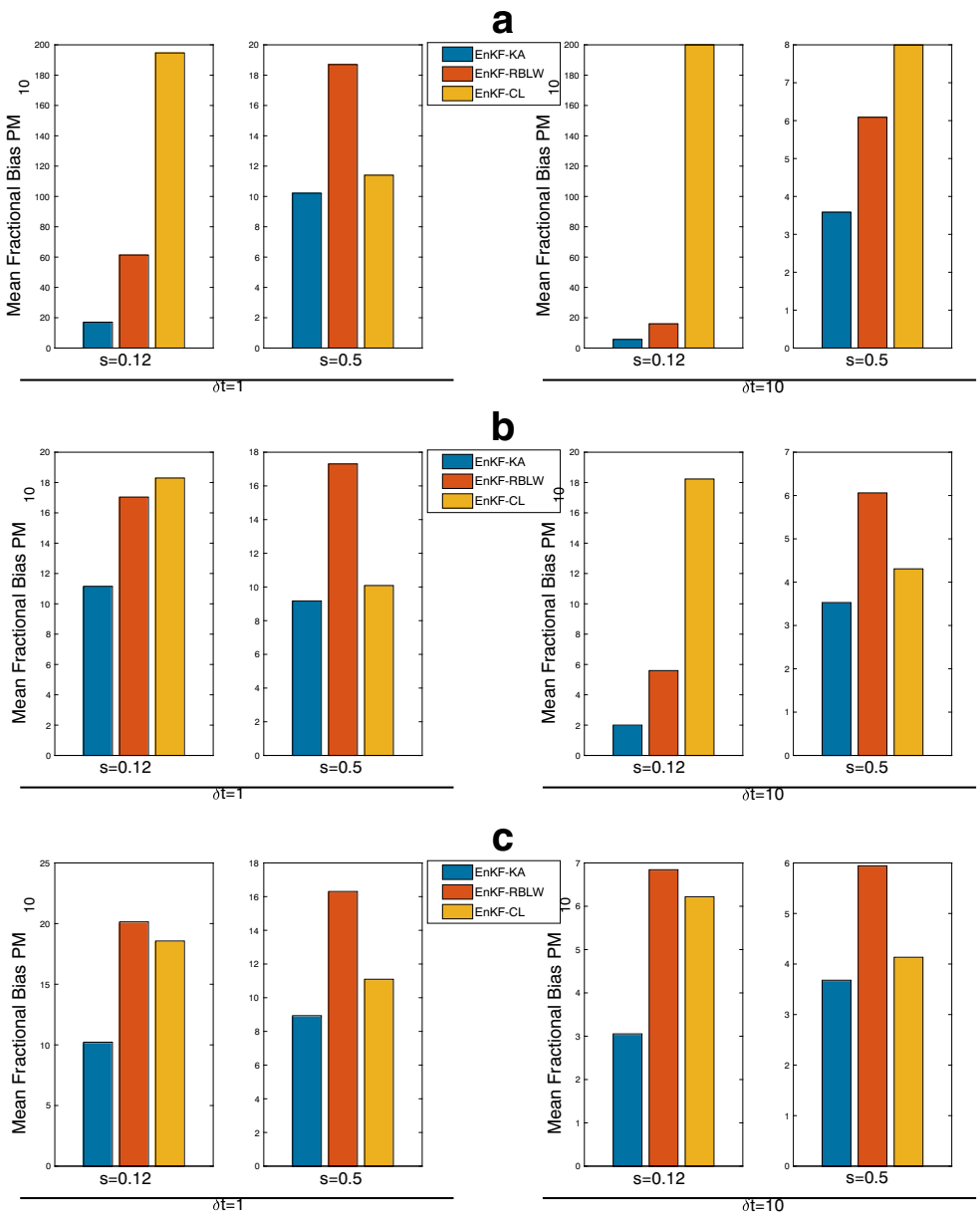

Fig. 7 Comparison performance for the different algorithms. **a** N=10. PNCE EnKF-CL=40%. **b** N=50. PNCE EnKF-CL=13:75% c N=100. PNCE EnKF-CL=0%

– The Root-Mean-Square-Error (RMSE) is employed as a metric of accuracy for a given analysis $\mathbf{x}_\ell^a$ and a reference solution $\mathbf{x}_\ell^*$ (see Eq. 27).

4.3 Analysis errors across pressure levels

Figures 11 and 12 show us the behavior of the proposed method against to EnKF-RBLW. The analysis was made using the RMSE metric for observation frequencies of 6 and 12 hours for u, v, and T model variables in different Pressure Levels. The numerical results show that EnKF-KA can be more accurate than EnKF-RBLW, this obeys the fact that the error correlations are driven by the physics and the numerical model's non-linear dynamics. Therefore, the underlying error distribution of wind components can be non-Gaussian as the frequency of observations decreases (long-term forecasts). This can apply to temperature fields as well. On the other hand, Gaussian assumptions can be valid for model variables such as the specific humidity. For this model variable, slight differences between analysis RMSE can be evidenced for the compared filter implementations. This can be expected since the RBLW covariance matrix estimator can perform well as the underlying error distribution of ensemble members is nearly Gaussian. Nevertheless, these small differences favor the proposed EnKF-KA formulation under the current experimental settings. In general, errors can grow faster across all pressure levels in model variables such as u, v, and T than those in variables that tend to preserve Gaussianity among assimilation steps (i.e., q).

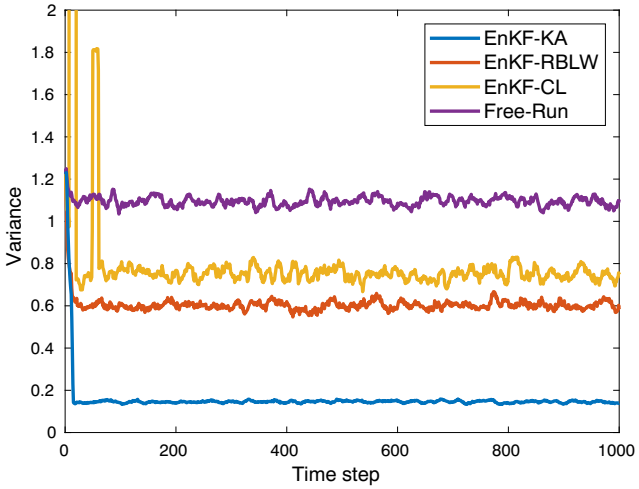

Fig. 8 Ensemble spread for different algorithms. The graph corresponds with one experiment wit $N = 50$, $\delta t = 10$, and $s = 0.5$

Table 1 Observational error standard deviation

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 (ρ)	100 (Pa)

4.4 Evolution of analysis errors among assimilation steps

As we can see in Figs. 13 and 14, the initial errors decrease as observations are assimilated in each analysis step using the proposed method, for observation frequencies of 6 and 12 hours. It should be noted that the observation frequency affects the estimation quality but not the convergence of the EnKF-KA with the configuration of this experiment. On the other hand, the proposed method can outperform the EnKF-RBLW formulation as shown in Figs. 13 and 14. The fact that accurate analysis states can be estimated despite a highly sparse observational network shows that the dynamic system's background error correlations have been captured into the covariance matrix estimators.

4.5 Analysis RMSE for the assimilation window

Tables 2 and 3 shows the analysis RMSE of the EnKF-KA and the EnKF-RBLW using 6 and 12 hours for observation frequencies and ensemble sizes of $N = 10$ and $N = 20$. The RMSE values are computed for 60 days with an initial spin-up period of ten days. As can be seen, the analysis states of the EnKF-KA can improve on the results proposed by the EnKF-RBLW. This can be possible

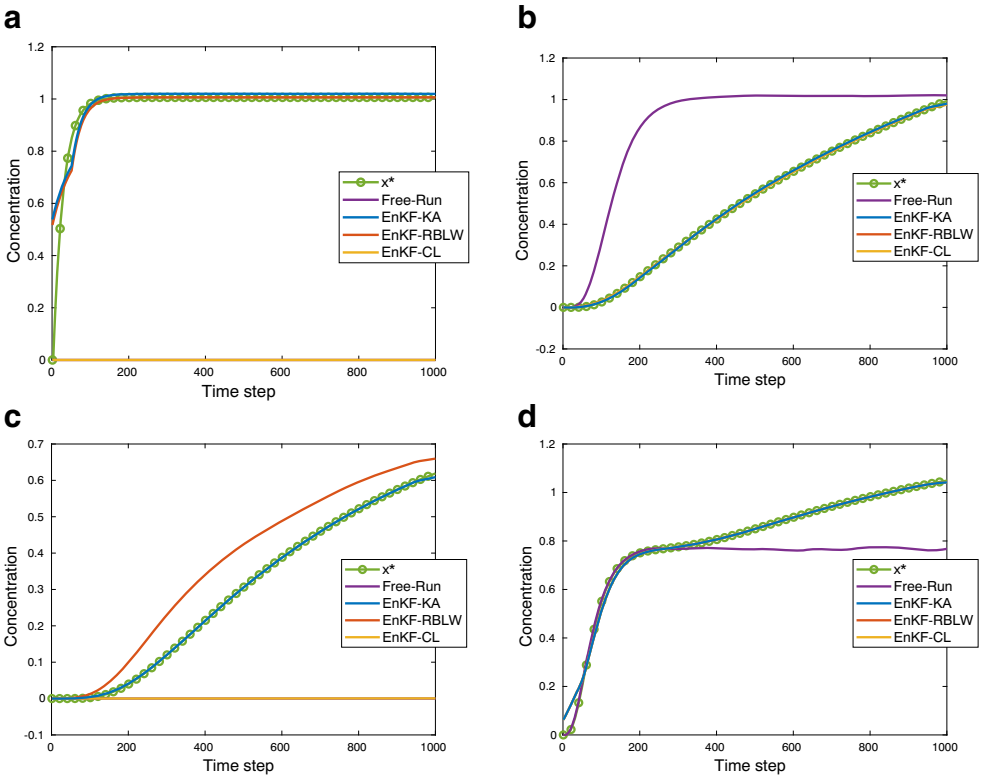

Fig. 9 Time evolution of concentration for different locations. The graph corresponds with one experiment wit $N = 50$, $\delta t = 10$, and $s = 0.5$. **a** Outside the valley. **b** External border of the valley. **c** Internal border of the valley. **d** Inside the valley

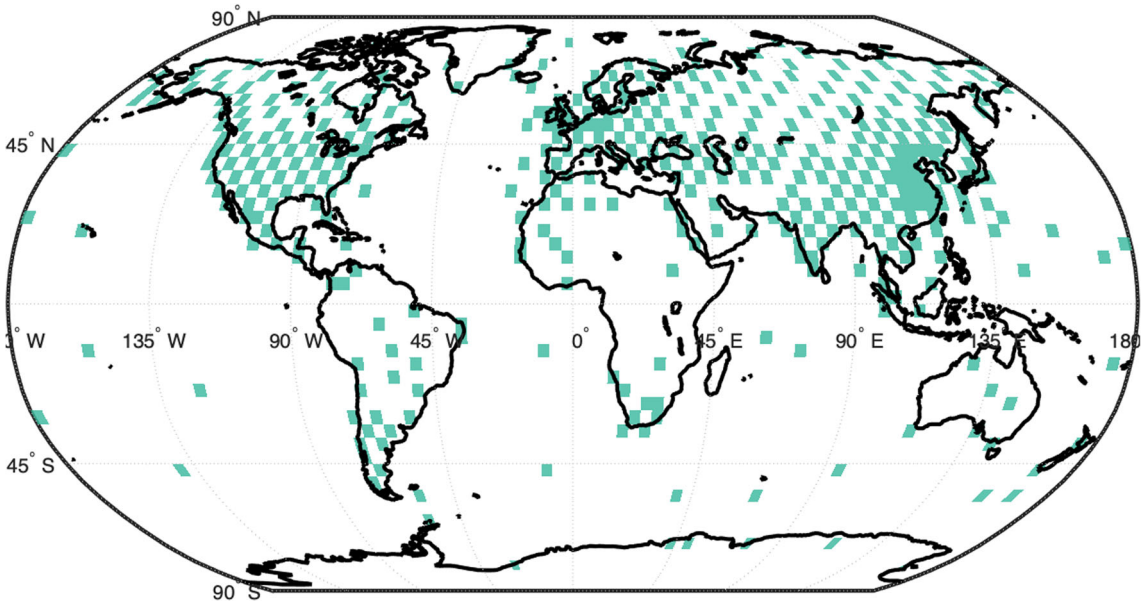

Fig. 10 An irregularly distributed realistic observational network. 415 stations (9% of all grid points) are located mostly over continents in the northern hemisphere

due to EnKF-KA uses a target matrix different from the identity matrix (used in EnKF-RBLW), and the EnKF-RBLW is performed under Gaussian assumptions over prior ensemble members. However, Gaussian assumptions on background errors can be broken by the numerical model's non-linear dynamics. The observational network in the experiment is sparse, about 9% of observations, which means that posterior estimates' quality relies on background error correlations. The proposed method then improves the quality of the analysis results over the compared filter for sparse observational network and very small ensemble sizes in the experiment.

4.6 Uncertainty analysis

For sequential data assimilation based on Ensemble Kalman Filter is known that if the ensemble spread becomes very small or becomes very large, the filter falls into divergence, but also, the ensemble spread can be used to explore the uncertainty associated with the initial condition and the

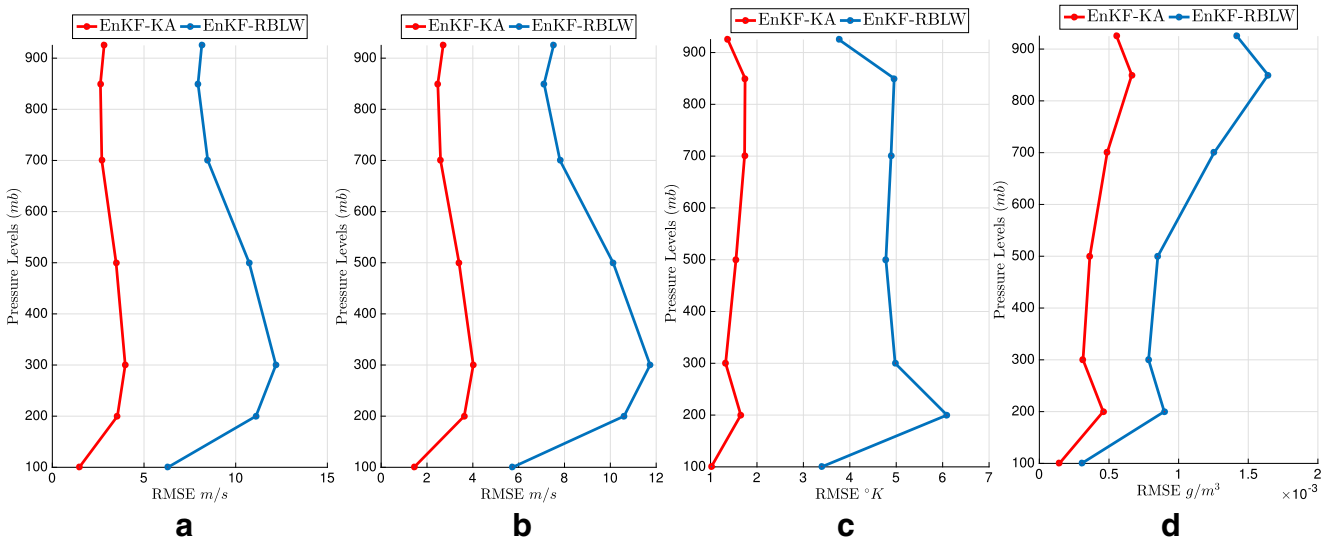

Fig. 11 Analysis RMSE at the all pressure levels temporally averaged for one month and a half after the initial spin-up period of two weeks. The number of ensemble members, reads $N = 10$. The errors per layer are shown for observation frequencies of 6 h. **a** u. **b** v. **c** T. **d** q

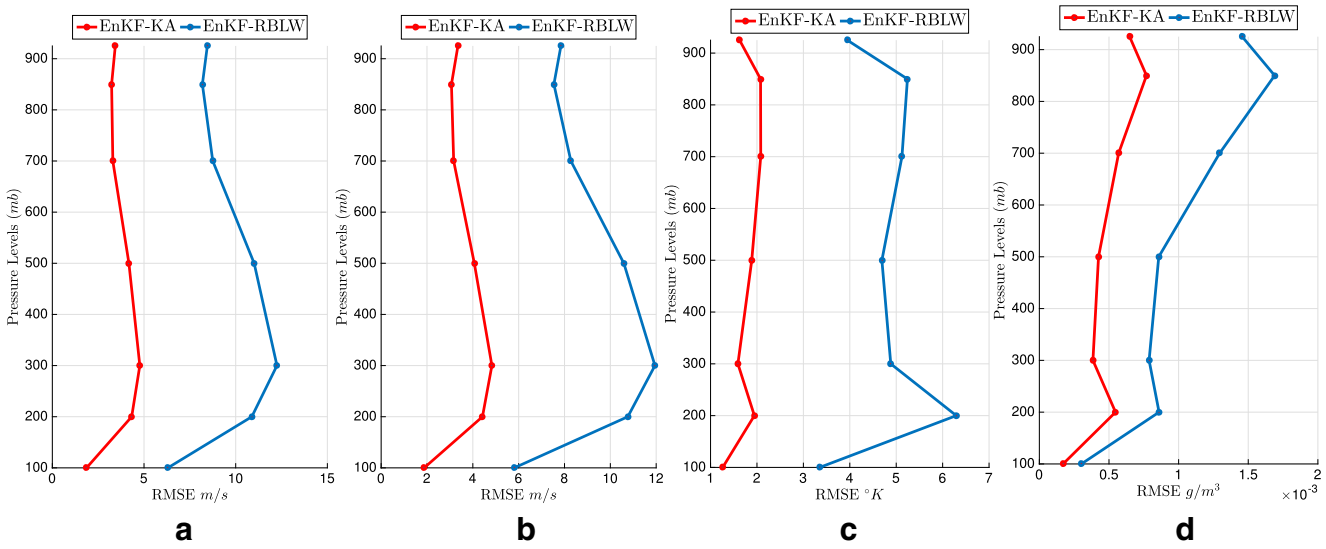

Fig. 12 Analysis RMSE at the all pressure levels temporally averaged for one month and a half after the initial spin-up period of two weeks. The number of ensemble members reads $N = 10$. The errors per layer are shown for observation frequencies of 12 h. **a** u. **b** v. **c** T. **d** q

uncertainty associated to the formulation of the prediction model. Figure 15 shows the mean of ensemble variance among assimilation steps for u, v, T and Q variables in pressure level of 500 Pa. As expected, the ensemble variance decreases as EnK-KA is used for the analysis step. This means that the uncertainty decreases as the observations are assimilated. It should be noted that a covariance inflation factor of 1.3 was used in the experiment. In the same way, Fig. 16 shows samples of the components taken for each of the model's physical variables. It is possible to see how the differences between ensemble members decrease through the assimilation steps.

4.7 CPU-time of analysis steps

Statistics of CPU-Times are computed across all analysis steps for both filters. The reported times are shown in Table 4, where the average and the variance of elapsed time for the analysis step computations are in seconds. The forecast step was realized using parallelism in a CPU

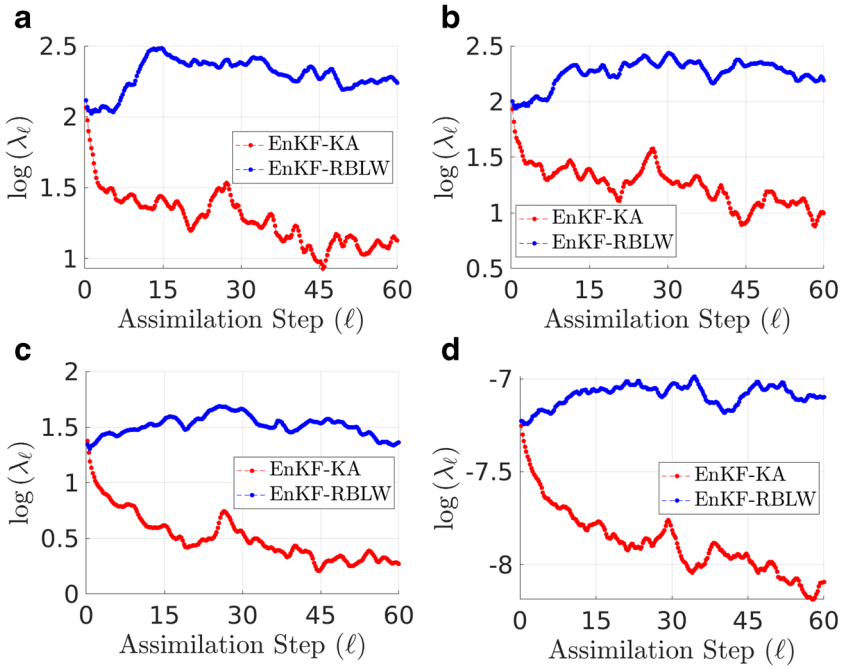

Fig. 13 Evolution of analysis errors among assimilation steps for $N = 10$ and an observation frequency of 6 h. The l_2-norm of errors is displayed in the log-scale for ease of reading. **a** u. **b** v. **c** T. **d** q

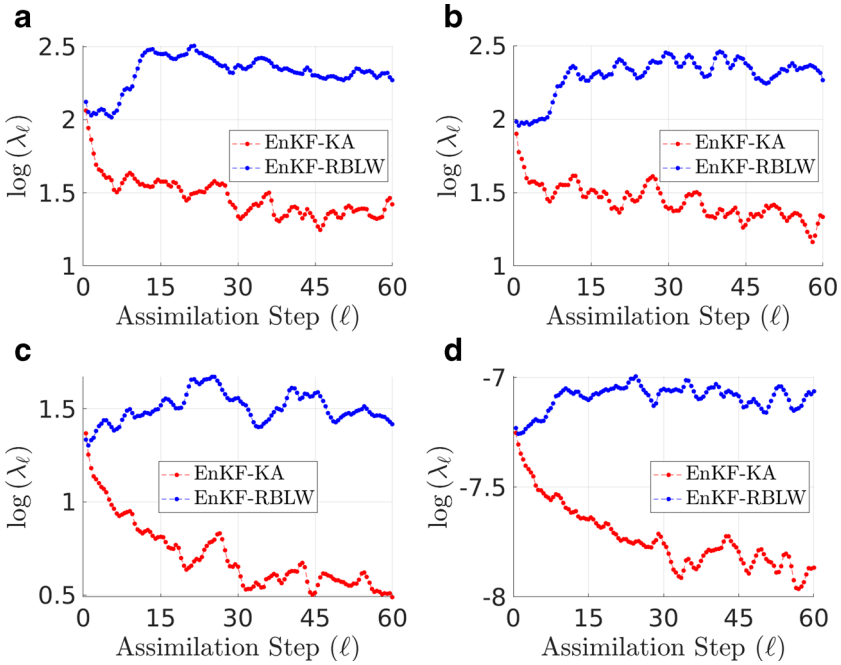

Fig. 14 Evolution of analysis errors among assimilation steps for $N = 10$ and an observation frequency of 12 h. The l_2-norm of errors is displayed in the log-scale for ease of reading. **a** u. **b** v. **c** T. **d** q

Table 2 RMSE values in time for observation frequencies of 6 h and 12 h

Variable	Method	6 hours	12 hours
$u\ (m/s)$	EnKF-KA	3.68249519	4.35949404
	EnKF-RBLW	10.73446620	11.00722997
$v\ (m/s)$	EnKF-KA	3.61925501	4.28412956
	EnKF-RBLW	10.12226756	10.58916871
$T\ (K)$	EnKF-KA	1.65550668	2.00101474
	EnKF-RBLW	4.77765718	4.69134891
$Q\ (kg/kg)$	EnKF-KA	0.00037801	0.00043576
	EnKF-RBLW	0.00085067	0.00085742
$\rho\ (hPa)$	EnKF-KA	3.18603080	3.87572849
	EnKF-RBLW	10.33055067	11.11026910

As the frequency of observations is decreased, the EnKF-KA formulation can improve on the results of the EnKF-RBLW method. The number of ensemble members reads $N = 10$

Table 3 RMSE values in time for observation frequencies of 6 h and 12 h

Variable	Method	6 hours	12 hours
$u\ (m/s)$	EnKF-KA	3.33453262	4.32448462
	EnKF-RBLW	10.23218824	10.59578129
$v\ (m/s)$	EnKF-KA	3.26725648	4.20452900
	EnKF-RBLW	10.27842723	10.08512826
$T\ (K)$	EnKF-KA	1.52252647	1.97627077
	EnKF-RBLW	4.39158134	4.43188562
$Q\ (kg/kg)$	EnKF-KA	0.00034380	0.00042017
	EnKF-RBLW	0.00082716	0.00083432
$\rho\ (hPa)$	EnKF-KA	2.86353932	3.84863434
	EnKF-RBLW	9.86403469	10.10293843

As the frequency of observations is decreased, the EnKF-KA formulation can improve on the results of the EnKF-RBWL method. The number of ensemble members reads $N = 20$

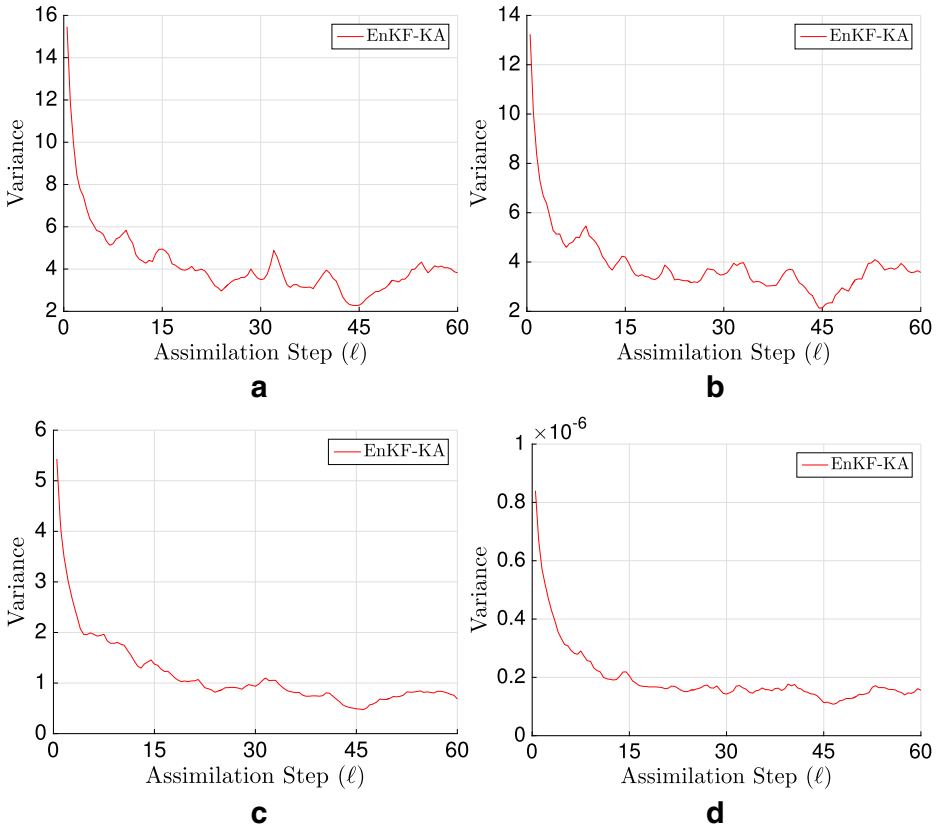

Fig. 15 Mean of variance among assimilation steps for $N = 10$, an observation frequency of 12 h, and a pressure level of 500 Pa. **a** u. **b** v. **c** T. **d** q

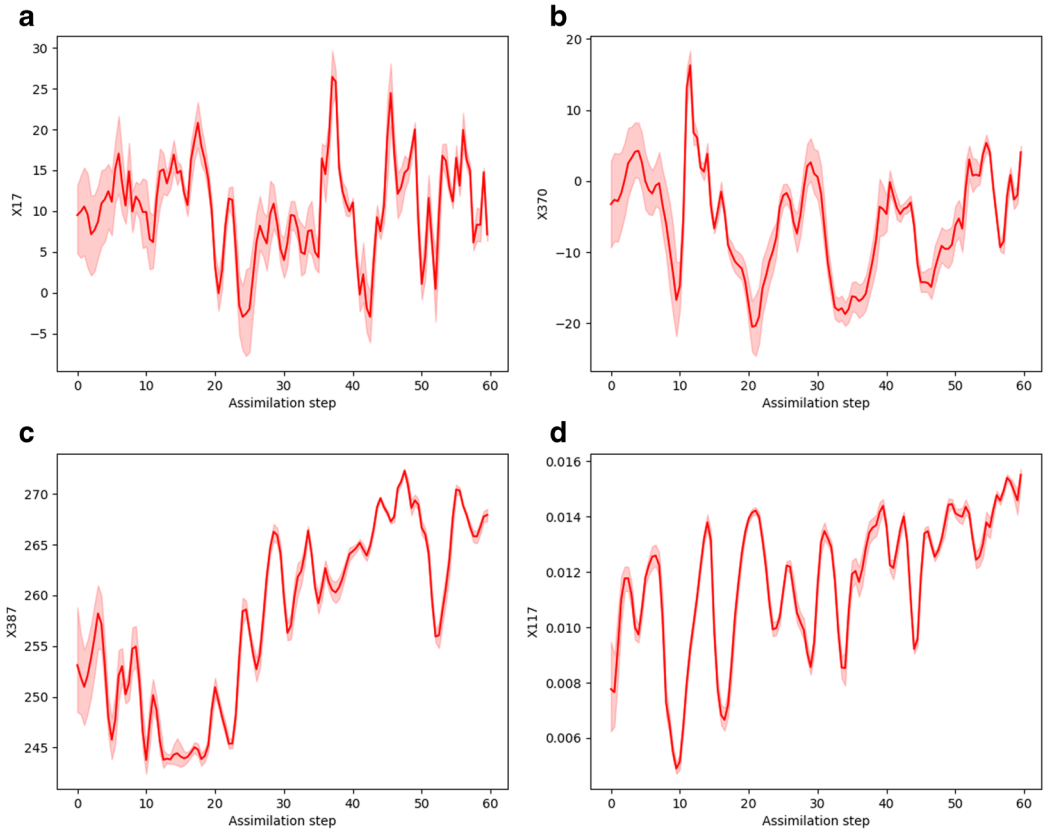

Fig. 16 Ensembles of a component sample among assimilation steps for $N = 10$, an observation frequency of 12 h, and a pressure level of 500 Pa. **a** u. **b** v. **c** T. **d** q

Table 4 Statistics of CPU-Time in seconds for the analysis steps of the compared filters and the forecast step

Method	Average CPU-Time	Stand. Dev. CPU-Time
Analysis EnKF-KA	6.4688	0.1723
Analysis EnKF-RBLW	5.562	0.157
Forecast Step	4.3272	0.209

The number of ensemble members reads 10

with four cores; this means that up to four ensembles were forecast simultaneously.

5 Conclusions

An efficient and practical implementation of the EnKF based on shrinkage covariance matrix estimation (EnKF-KA) was proposed in the present document. The proposed filter implementation exploits the information brought by an ensemble of model realization (numerical model dynamics) and our prior knowledge about the actual dynamical system (i.e., the prior structure of background error correlations). The EnKF-KA uses a target matrix with a general structure, representing a novel approach compared with the current shrinkage-based estimators that use an identity matrix as a target matrix. An efficient implementation for large systems is presented, taking advantage of the local domain decomposition. Experimental tests are performed by using an advection-diffusion model and an Atmospheric General Circulation Model. In both cases, the proposed method can outperform EnKF based on shrinkage covariance estimation where there is no prior information about error correlations, and the standard EnKF using covariance localization. The results support the idea that it is possible to use the information and prior knowledge of the system to improve the current ensemble-based DA method.

Declarations

Conflict of Interests The authors declare that they have no conflict of interest.

References

1. Anderson, J.L.: An Ensemble Adjustment Kalman Filter for Data Assimilation. Monthly Weather Review. https://doi.org/10.1175/1520-0493(2001)129<2884:AEAKFF>2.0.CO;2 (2001)
2. Anderson, J.L., Anderson, S.L.: A Monte Carlo Implementation of the Nonlinear Filtering Problem to Produce Ensemble Assimilations and Forecasts. Monthly Weather Review. https://doi.org/10.1175/1520-0493(1999)127<2741:AMCIOT>2.0.CO;2 (1999)
3. Asch, M., Bocquet, M., Nodet, M.: Data assimilation. Society for industrial and applied mathematics, Philadelphia. https://doi.org/10.1137/1.9781611974546 (2016)
4. Barbu, A.L., Segers, A.J., Schaap, M., Heemink, A.W., Builtjes, P.J.H.: A multi-component data assimilation experiment directed to sulphur dioxide and sulphate over Europe. Atmos. Environ. **43**(9), 1622–1631 (2009). https://doi.org/10.1016/j.atmosenv.2008.12.005
5. Bracco, A., Kucharski, F., Kallummal, R., Molteni, F.: Internal variability, external forcing and climate trends in multi-decadal AGCM ensembles. Clim. Dynam. **23**(6), 659–678 (2004)
6. Chen, Y., Wiesel, A., Eldar, Y.C., Hero, A.O.: Shrinkage algorithms for MMSE covariance estimation. IEEE Trans. Signal Process. **58**(10), 5016–5029 (2010). https://doi.org/10.1109/TSP.2010.2053029
7. Chen, Y.H., Prinn, R.G.: Estimation of atmospheric methane emissions between 1996 and 2001 using a three-dimensional global chemical transport model. J. Geophys. Res. Atmosph. **111**(10), 1–25 (2006). https://doi.org/10.1029/2005JD006058
8. Couillet, R., McKay, M.: Large dimensional analysis and optimization of robust shrinkage covariance matrix estimators. J. Multivar. Anal. **131**, 99–120 (2014)
9. Evensen, G.: The Ensemble Kalman Filter: Theoretical formulation and practical implementation. Ocean Dyn. **53**(4), 343–367 (2003). https://doi.org/10.1007/s10236-003-0036-9
10. Fu, G., Prata, F., Xiang lin, H., Heemink, A., Segers, A., Lu, S.: Data assimilation for volcanic ash plumes using a satellite observational operator: A case study on the 2010 Eyjafjallajökull volcanic eruption. Atmos. Chem. Phys. **17**(2), 1187–1205 (2017). https://doi.org/10.5194/acp-17-1187-2017
11. Gaspari, G., Cohn, S.E.: Construction of correlation functions in two and three dimensions. Q. J. Roy. Meteorol. Soc. **125**(554), 723–757 (1999). https://doi.org/10.1002/qj.49712555417
12. Greybush, S.J., Kalnay, E., Miyoshi, T., Ide, K., Hunt, B.R.: Balance and Ensemble Kalman Filter Localization Techniques. Mon. Weather. Rev. **139**(2), 511–522 (2011). https://doi.org/10.1175/2010MWR3328.1
13. Hamill, T.M., Whitaker, J.S., Snyder, C.: Distance-Dependent Filtering of Background Error Covariance Estimates in an Ensemble Kalman Filter. Monthly Weather Review. https://doi.org/10.1175/1520-0493(2001)129<2776:DDFOBE>2.0.CO;2 (2001)
14. Houtekamer, P., Mitchell, H.: A Sequential Ensemble Kalman Filter for Atmospheric Data Assimilation. Amer. Meteorol. Soc. **129**(ii), 123–137. https://doi.org/10.1175/1520-0493(2001)129<0123:ASEKFF>2.0.CO;2 (2001)
15. Houtekamer, P.L., Zhang, F.: Review of the Ensemble Kalman Filter for Atmospheric Data Assimilation. monthly weather review. https://doi.org/10.1175/MWR-D-15-0440.1 (2016)
16. Hunt, B.R., Kostelich, E.J., Szunyogh, I.: Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter. Physica D: Nonlinear Phenom. **230**(1-2), 112–126 (2007). https://doi.org/10.1016/j.physd.2006.11.008

17. Jin, J., Lin, H.X., Heemink, A., Segers, A.: Spatially varying parameter estimation for dust emissions using reduced-tangent-linearization 4DVar. Atmos. Environ. **187**, 358–373 (2018). https://doi.org/10.1016/j.atmosenv.2018.05.060
18. Kalnay, E., Li, H., Miyoshi, T., Yang, S.C., Ballabrera-poy, J.: 4dvar or ensemble kalman filter? Tellus A: Dyn. Meteorol. Oceanogr. **59**(5), 758–773 (2007)
19. Kucharski, F., Molteni, F., Bracco, A.: Decadal interactions between the western tropical pacific and the north atlantic oscillation. Clim. Dyn. **26**(1), 79–91 (2006)
20. Ledoit, O., Wolf, M.: A well-conditioned estimator for large-dimensional covariance matrices. J. Multivar. Anal. **88**(2), 365–411 (2004)
21. Ledoit, O., Wolf, M.: A well-conditioned estimator for large-dimensional covariance matrices. J. Multivar. Anal. **88**(2), 365–411. https://doi.org/10.1016/S0047-259X(03)00096-4 (2004)
22. Ledoit, O., Wolf, M., et al.: Optimal estimation of a large-dimensional covariance matrix under stein's loss. Bernoulli **24**(4B), 3791–3832 (2018)
23. Lopez-Restrepo, S., Yarce, A., Pinel, N., Quintero, O., Segers, A., Heemink, A.: Forecasting PM_{10} and $PM_{2.5}$ in the Aburrá Valley (Medellín, Colombia) via EnKF based Data Assimilation. Atmospheric Environment. Accepted (2020)
24. Lu, S., Lin, H.X., Heemink, A.W., Fu, G., Segers, A.J.: estimation of volcanic ash emissions using Trajectory-Based 4D-Var data assimilation. Mon. Weather. Rev. **144**(2), 575–589 (2016). https://doi.org/10.1175/MWR-D-15-0194.1
25. Miyoshi, T.: The gaussian approach to adaptive covariance inflation and its implementation with the local ensemble transform kalman filter. Mon. Weather. Rev. **139**(5), 1519–1535 (2011)
26. Molteni, F.: Atmospheric simulations using a gcm with simplified physical parametrizations. i: Model climatology and variability in multi-decadal experiments. Climate Dynam. **20**(2-3), 175–191 (2003)
27. Nan, T., Wu, J.: Groundwater parameter estimation using the ensemble Kalman filter with localization. Hydrogeol. J. **19**(3), 547–561 (2011). https://doi.org/10.1007/s10040-010-0679-9
28. Nino-Ruiz, E.D., Guzman, L., Jabba, D.: An ensemble kalman filter implementation based on the ledoit and wolf covariance matrix estimator. J. Comput. Appl. Math. **384**, 113163. https://doi.org/10.1016/j.cam.2020.113163 (2021)
29. Nino-Ruiz, E.D., Guzman-Reyes, L.G., Beltran-Arrieta, R.: An adjoint-free four-dimensional variational data assimilation method via a modified cholesky decomposition and an iterative woodbury matrix formula. Nonlinear Dyn. **99**(3), 2441–2457 (2020)
30. Nino-Ruiz, E.D., Sandu, A.: Ensemble kalman filter implementations based on shrinkage covariance matrix estimation. Ocean Dyn. **65**(11), 1423–1439 (2015)
31. Nino-Ruiz, E.D., Sandu, A.: Efficient parallel implementation of dddas inference using an ensemble kalman filter with shrinkage covariance matrix estimation. Clust. Comput., 1–11 (2017)
32. Ott, E., Hunt, B.R., Szunyogh, I., Zimin, A.V., Kostelich, E., Corazza, M., Kalnay, E., Patil, D., Yorke, J.A.: A local ensemble Kalman filter for atmospheric data assimilation. Tellus **56**, 415–428 (2004)
33. Petrie, R.E.: Localization in the Ensemble Kalman Filter. Master, University of Reading (2008)
34. Quintero, M.O.L., Amicarelli, A.A., Scaglia, G., di Sciascio, F.: Control based on numerical methods and recursive bayesian estimation in a continuous alcoholic fermentation process. BioResources **4**(4), 1372–1395 (2009)
35. Quintero, M.O.L., Scaglia, G., Di Sciascio, F., Mut, V.: Numerical Methods Based Strategy and Particle Filter State Estimation for Bio Process Control. In: 2008 IEEE International Conference on Industrial Technology, pp. 1–6. IEEE (2008)
36. Richardson, P.L., Mooney, K.: The mediterranean outflow—a simple advection-diffusion model. J. Phys. Oceanogr. **5**(3), 476–482. https://doi.org/10.1175/1520-0485(1975)005<0476:TMOSAD>2.0.CO;2 (1975)
37. Sakov, P., Bertino, L.: Relation between two common localisation methods for the enKF. Comput. Geosci. **15**(2), 225–237 (2011). https://doi.org/10.1007/s10596-010-9202-6
38. Sakov, P., Evensen, G., Bertino, L.: Asynchronous data assimilation with the enKF. Tellus, Ser. A: Dyn. Meteorol. Oceanogr. **62**(1), 24–29 (2010). https://doi.org/10.1111/j.1600-0870.2009.00417.x
39. Stoica, P., Li, J., Zhu, X., Guerci, J.R.: On using a priori knowledge in space-time adaptive processing. IEEE Trans. Signal Process. **56**(6), 2598–2602 (2008). https://doi.org/10.1109/TSP.2007.914347
40. Timmermans, R., Segers, A., Curier, L., Abida, R., Attiė, J.L., El Amraoui, L., Eskes, H., De Haan, J., Kujanpàà, J., Lahoz, W., Oude Nijhuis, A., Quesada-Ruiz, S., Ricaud, P., Veefkind, P., Schaap, M.: Impact of synthetic space-borne NO2 observations from the Sentinel-4 and Sentinel-5P missions on tropospheric NO2 analyses. Atmos. Chem. Phys. **19**(19), 12811–12833 (2019). https://doi.org/10.5194/acp-19-12811-2019
41. Tirabassi, T.: Analytical air pollution advection and diffusion models. Water Air Soil Pollut. **47**(1), 19–24 (1989). https://doi.org/10.1007/BF00468993
42. Touloumis, A.: Nonparametric stein-type shrinkage covariance matrix estimators in high-dimensional settings. Comput. Stat. Data Anal. **83**, 251–261 (2015)
43. Wang, X., Barker, D.M., Snyder, C., Hamill, T.M.: A hybrid etkf–3dvar data assimilation scheme for the wrf model. part i: Observing system simulation experiment. Mon. Weather. Rev. **136**(12), 5116–5131 (2008)
44. Wang, X., Snyder, C., Hamill, T.M.: On the theoretical equivalence of differently proposed ensemble–3dvar hybrid analysis schemes. Mon. Weather. Rev. **135**(1), 222–227 (2007)
45. Zhu, X., Li, J., Stoica, P.: Knowledge-Aided Space-Time adaptive processing. IEEE Trans. Aerosp. Electron. Syst. **47**(2), 1325–1336 (2011)
46. Zhu, Y., Toth, Z., Wobus, R., Richardson, D., Mylne, K.: The Economic Value of Ensemble-Based Weather Forecasts. Bullet. Amer. Meteorol. Soci. 83(1), 73–84. https://doi.org/10.1175/1520-0477(2002)083<0073:TEVOEB>2.3.CO;2 (2002)

Affiliations

Santiago Lopez-Restrepo[1,2] · Elias D. Nino-Ruiz[3] · Luis G. Guzman-Reyes[3] · Andres Yarce[1,2] · O. L. Quintero[1] · Nicolas Pinel[4] · Arjo Segers[5] · A. W. Heemink[2]

Elias D. Nino-Ruiz
enino@uninorte.edu.co

Luis G. Guzman-Reyes
lgguzman@uninorte.edu.co

Andres Yarce
ayarceb@eafit.edu.co; a.yarcebotero@tudelft.nl

O. L. Quintero
oquinte1@eafit.edu.co

Nicolas Pinel
npinelp@eafit.edu.co

Arjo Segers
arjo.segers@tno.nl

A. W. Heemink
a.w.heemink@tudelft.nl

1 Mathematical Modelling Research Group, Universidad EAFIT, Antioquia, Colombia

2 Department of Applied Mathematics, TU Delft, The Netherlands

3 Department of Computer Science, Applied Math and Computer Science Laboratory, Universidad del Norte, Barranquilla, Colombia

4 Biodiversity, Evolution and Conservation Research Group at Universidad EAFIT, Medellí, Colombia

5 Department of Climate, Air and Sustainability, TNO, Utrecht, The Netherlands