RMetS
Royal Meteorological Society

An iterative ensemble Kalman smoother

M. Bocquet[a,b]* and P. Sakov[c]

[a]*Université Paris-Est, CEREA, Joint Laboratory École des Ponts ParisTech and EDF R&D, Champs-sur-Marne, France*
[b]*INRIA, Paris Rocquencourt Research Centre, Paris, France*
[c]*Bureau of Meteorology, Melbourne, Australia*

*Correspondence to: M. Bocquet, CEREA, École des Ponts ParisTech, 6-8 avenue Blaise Pascal, Cité Descartes, Champs-sur-Marne, 77455 Marne la Vallée Cedex, France. E-mail: bocquet@cerea.enpc.fr

The iterative ensemble Kalman filter (IEnKF) was recently proposed in order to improve the performance of ensemble Kalman filtering with strongly nonlinear geophysical models. The IEnKF can be used as a lag-one smoother and extended to a fixed-lag smoother: the iterative ensemble Kalman smoother (IEnKS). The IEnKS is an ensemble variational method. It does not require the use of the tangent linear of the evolution and observation models, nor the adjoint of these models: the required sensitivities (gradient and Hessian) are obtained from the ensemble. Looking for optimal performance, out of the many possible extensions we consider a quasi-static algorithm. The IEnKS is explored for the Lorenz '95 model and for a two-dimensional turbulence model. As the logical extension of the IEnKF, the IEnKS significantly outperforms standard Kalman filters and smoothers in strongly nonlinear regimes. In mildly nonlinear regimes (typically synoptic-scale meteorology), its filtering performance is marginally but clearly better than the standard ensemble Kalman filter and it keeps improving as the length of the temporal data assimilation window is increased. For long windows, its smoothing performance outranks the standard smoothers very significantly, a result that is believed to stem from the variational but flow-dependent nature of the algorithm. For very long windows, the use of a multiple data assimilation variant of the scheme, where observations are assimilated several times, is advocated. This paves the way for finer reanalysis, freed from the static prior assumption of 4D-Var but also partially freed from the Gaussian assumptions that usually impede standard ensemble Kalman filtering and smoothing.

Key Words: ensemble Kalman filter; iterative ensemble Kalman filter; ensemble Kalman smoother; iterative ensemble Kalman smoother; ensemble variational method

Received 25 January 2013; Revised 11 July 2013; Accepted 11 July 2013; Published online in Wiley Online Library 29 October 2013

1. Introduction

1.1. Filtering and smoothing

One of the most fundamental statistical objects of data assimilation is the probability density function (pdf) of the present system state, conditional on present and past observations. At time t_k, $k \in \{1, \cdots, K\}$, the observation vector is denoted as $\mathbf{y}_k$, of dimension d_k. The system state at the same time is denoted by $\mathbf{x}_k$, of dimension M. This *filtering* pdf is $p(\mathbf{x}_k|\mathbf{y}_1, \cdots, \mathbf{y}_k)$. The argument of the maximum of this pdf gives the most likely state of the system at time t_k, which may be used as the initial condition for a forecast.

Besides present-time analysis and forecast, retrospective analysis (as coined by Cohn *et al.*, 1994) is also of great interest, in particular for reanalysis or for parameter estimation. As a consequence, another statistical object of focus is the pdf of the system state conditional on past, present and future observations, which is called the *smoothing* pdf. If t_K stands for present time, $p(\mathbf{x}_k|\mathbf{y}_1, \cdots, \mathbf{y}_K)$ with $1 \leq k < K$ would designate such a pdf. Smoothing can be performed over the whole data assimilation window $[t_1, t_K]$, leading to the smoothing pdf $p(\mathbf{x}_1, \cdots, \mathbf{x}_K|\mathbf{y}_1, \cdots, \mathbf{y}_K)$. By definition, this smoothing pdf offers a comprehensive statistical description of the true system state trajectory over $[t_1, t_K]$ (Jazwinski, 1970; Evensen and van Leeuwen, 2000).

Because the filtering pdf $p(\mathbf{x}_K|\mathbf{y}_1, \cdots, \mathbf{y}_K)$ is a marginal of the smoothing pdf $p(\mathbf{x}_1, \cdots, \mathbf{x}_K|\mathbf{y}_1, \cdots, \mathbf{y}_K)$ when integrating over $\mathbf{x}_1, \ldots, \mathbf{x}_{K-1}$, there is no advantage in using the smoothing pdf for present-time analysis over the filtering pdf. However, this is only proven to hold for optimal data assimilation methods intended to solve for the exact filtering pdf.

1.2. Suboptimality of Gaussian schemes

One way to portray current data assimilation methods is to understand them as methods meant to estimate the filtering

pdf or smoothing pdf approximately, or some of their statistical moments.

Models of the evolution of the atmosphere or the ocean used in geophysical data assimilation are typically nonlinear, as sometimes are the observation operators used in data assimilation schemes. As a result, the statistics of the errors of such systems are non-Gaussian, even when the initial assumption of the errors that characterize our initial ignorance is Gaussian (Andersson *et al.*, 2005; Bocquet *et al.*, 2010). At the synoptic scale, the nonlinear primitive equations lead to chaos and undermine the performance of data assimilation schemes, often based on Gaussian statistical hypotheses. Further, in geophysical models there often exist other sources of even stronger nonlinearities that do not necessarily lead to chaos, such as the microphysics of atmospheric constituents including water, ice, gas, aerosols, biochemical tracers, etc.

Most of the efficient data assimilation techniques dealing with high-dimensional nonlinear systems rely on Gaussian assumptions: as an approximation, the distribution of the errors is modelled as a Gaussian distribution. In particular, in the sequential context of an operational forecast, popular methods such as the ensemble Kalman filter and 4D-Var rely on a prior that is assumed Gaussian (or, more exactly, assumed to be known up to second-order moments).

Ensemble Kalman filters can be seen as a way to estimate the filtering pdf approximately, while 4D-Var estimates the most likely state of a conditional smoothing pdf. Like ensemble Kalman filters, practical ensemble Kalman smoothers are also based on Gaussian assumptions but estimate the smoothing distribution (Evensen, 2009; Cosme *et al.*, 2012, and references therein). Over a specific data assimilation window, with linear models and identical Gaussian priors, the Kalman smoother and 4D-Var will give the same most likely state trajectory over the time window. Additionally, given these assumptions, the Kalman filter, Kalman smoother and 4D-Var methods must give the same estimate at the end time of the assimilation window.

With realistic models and observations, these equivalences should only be approximately satisfied, depending on how nonlinear the models are.

1.3. The iterative ensemble Kalman filter

When the models are significantly nonlinear and the statistics are non-Gaussian, these equivalences could be broken, leading to suboptimalities in retrospective and present-time analysis as well as in forecasting. One solution would be to abandon the Gaussian methods and opt for fully non-Gaussian ones such as particle filters (van Leeuwen, 2009; Bocquet *et al.*, 2010). Another solution, which is the focus of this article, would be to extend the data assimilation window (DAW) length (Fisher *et al.*, 2005). This way, it is hoped that suboptimality in using the Gaussian priors would not affect the estimation in the bulk of the DAW too much. Extending the DAW comes with its own problems, since nonlinearities lead to an increasing number of local minima in the cost function. This can be alleviated by using a weak-constraint formulation of the variational problem (Fisher *et al.*, 2005) or alternatively using a quasi-static algorithm (Pires *et al.*, 1996). We point out that it is not clear how extending the DAW can help to improve nowcasting or forecasting, since both depend on the end time of the DAW.

The iterative ensemble Kalman filter (IEnKF: Sakov *et al.*, 2012) is an illustration of this strategy, with a clear demonstration of its success for strongly nonlinear low-order models and a circumstantial answer to the last question. The IEnKF performs variational analysis over the time interval between two updates. Regarding the state at the start of the DAW, it can be seen as a lag-one iterative ensemble Kalman smoother.

The IEnKF was shown to perform significantly better in strongly nonlinear conditions for the Lorenz '63 and Lorenz '95 models, when the time interval between updates increases, making the system transition between updates more nonlinear.

1.4. Context, choices of this study and objective

The goal of this article is to extend the iterative ensemble Kalman filter to an iterative ensemble Kalman smoother (IEnKS) using an extended DAW.

The IEnKS will be tested on two low-order models, the one-dimensional Lorenz '95 model and a two-dimensional turbulence model. We are interested in both the final-state analysis and the smoothing performance (retrospective analysis) of this extension, compared with filters and other non-iterative smoothers.

This preliminary study of the IEnKS will rely on several assumptions. First, the model is assumed to be perfect. Although this is a widely shared assumption, it is essential for the method, similarly to currently operational 4D-Var. Considering imperfect models and weak-constraint alternatives would require substantial change to the method and is therefore outside the scope of this article.

Also, we shall not consider issues related to localization. The localization of the IEnKS schemes will be necessary for application to realistic geophysical models. However, the problem of localization in the IEnKF or IEnKS deserves special attention. It will be discussed in the conclusions of this study. The implementation of a local IEnKS will be reported in a subsequent article.

Even in the perfect model case there are suboptimalities, such as sampling errors, that do typically require the use of inflation. In this article, this will be dealt with using finite-size versions of the smoothers (Bocquet, 2011) that do not require the use of inflation in perfect model conditions.

This study is primarily focused on the performance of the smoothers. This performance is measured by the average root-mean-square error (RMSE) between the analysis state and the truth state. We do not look into higher moments of the ensemble or scores for the quality of the ensemble (with one exception). However, as for any sequential data assimilation method, the performance of the scheme depends indirectly but strongly on the uncertainty quantification. Therefore, the RMSE of the analysis is also an indirect and partial measure of the quality of the ensemble. Computational cost is not of primary concern, but will nevertheless be discussed to a limited extent.

1.5. Outline

In section 2, we review the IEnKF following Sakov *et al.* (2012) and Bocquet and Sakov (2012). The derivation of the algorithm is justified from a probabilistic standpoint that was not used in those references. In section 3, we introduce and derive the iterative ensemble Kalman smoother. Two implementations are discussed that either assimilate each observation once (single data assimilation scheme) or assimilate observations several times over (multiple data assimilation scheme). The former is simple and derived rigorously, while the latter is derived heuristically and seems to offer better stability for a long DAW. Along with these schemes, a linearized version of the IEnKS is proposed. We also discuss theoretical differences between the IEnKS and a standard smoother, the solution of which is optimal and efficient in linear and Gaussian conditions. In section 4, tests of the new schemes and the Gaussian smoother are performed on the Lorenz '95 model and a two-dimensional turbulence model. Section 5 summarizes and further discusses the outcomes and next steps of this study (including localization).

Throughout the article, time t_k and its index k do not indicate an absolute date but instead measure time relative to the start of the DAW, which is t_0, using conventions similar to those of 4D-Var.

2. The iterative ensemble Kalman filter (IEnKF)

The iterative ensemble Kalman filter (Sakov *et al.*, 2012) can be seen as a lag-one iterative ensemble Kalman smoother,

even though only its performance as a filter was studied in this reference. Since the iterative ensemble Kalman smoother naturally stems from it, it is important to review the principles of the IEnKF. Also, a Bayesian probabilistic standpoint will be used here to justify the scheme and will be helpful to extend it.

In the following, $n(\mathbf{x}|\overline{\mathbf{x}}, \mathbf{P})$ will stand for the pdf of the multivariate Gaussian distribution of mean $\overline{\mathbf{x}}$ and covariance matrix $\mathbf{P}$.

2.1. Derivation of the scheme

Consider a data assimilation cycle (forecast and update). At time t_0, the system state $\mathbf{x}_0$ is assumed to follow a Gaussian distribution of pdf:

$$p(\mathbf{x}_0) = n\left(\mathbf{x}_0|\mathbf{x}_0^{(0)}, \mathbf{P}_0\right), \tag{1}$$

where $\mathbf{x}_0^{(0)}$ is the first guess and $\mathbf{P}_0$ is the forecast-error covariance matrix at t_0. The upper index (0) means that $\mathbf{x}_0^{(0)}$ relates to the first iteration of updates to be defined later.

The system evolves between t_0 and t_1 according to a possibly nonlinear model $\mathcal{M}_{1\leftarrow 0}$. This model is assumed to be perfect, so that

$$p(\mathbf{x}_1|\mathbf{x}_0) \propto \delta\left[\mathbf{x}_1 - \mathcal{M}_{1\leftarrow 0}(\mathbf{x}_0)\right], \tag{2}$$

where δ is the Dirac distribution defined in $\mathbb{R}^M$.

At t_1, present time, the observation vector $\mathbf{y}_1$ is assimilated. The observation likelihood is assumed to follow a Gaussian distribution:

$$p(\mathbf{y}_1|\mathbf{x}_1) = n\left(\mathbf{y}_1 - H_1(\mathbf{x}_1)|\mathbf{0}, \mathbf{R}_1\right). \tag{3}$$

Note that the method can handle nonlinear observation operators. The smoothing pdf of state $\mathbf{x}_0$ conditional on $\mathbf{y}_1$ is obtained using Bayes' rule and the perfect model assumption:

$$\begin{aligned} p(\mathbf{x}_0|\mathbf{y}_1) &\propto p(\mathbf{y}_1|\mathbf{x}_0)p(\mathbf{x}_0) \\ &\propto p(\mathbf{y}_1|\mathbf{x}_1 = \mathcal{M}_{1\leftarrow 0}(\mathbf{x}_0))p(\mathbf{x}_0). \end{aligned} \tag{4}$$

A cost function $\mathcal{J}$ for the estimation of $\mathbf{x}_0$ can be derived from this smoothing pdf:

$$p(\mathbf{x}_0|\mathbf{y}_1) \propto \exp\left(-\mathcal{J}(\mathbf{x}_0)\right). \tag{5}$$

Using Eqs (1), (2) and (3), one obtains

$$\begin{aligned} \mathcal{J}(\mathbf{x}_0) =& \frac{1}{2}\left[\mathbf{y}_1 - H_1 \circ \mathcal{M}_{1\leftarrow 0}(\mathbf{x}_0)\right]^{\mathrm{T}} \\ &\times \mathbf{R}_1^{-1}\left[\mathbf{y}_1 - H_1 \circ \mathcal{M}_{1\leftarrow 0}(\mathbf{x}_0)\right] \\ &+ \frac{1}{2}\left(\mathbf{x}_0 - \mathbf{x}_0^{(0)}\right)^{\mathrm{T}} \mathbf{P}_0^{-1}\left(\mathbf{x}_0 - \mathbf{x}_0^{(0)}\right). \end{aligned} \tag{6}$$

Then, the filtering distribution is obtained via

$$\begin{aligned} p(\mathbf{x}_1|\mathbf{y}_1) &= \int \mathrm{d}\mathbf{x}_0\, p(\mathbf{x}_1|\mathbf{x}_0, \mathbf{y}_1)p(\mathbf{x}_0|\mathbf{y}_1) \\ &= \int \mathrm{d}\mathbf{x}_0\, \delta\left[\mathbf{x}_1 - \mathcal{M}_{1\leftarrow 0}(\mathbf{x}_0)\right] p(\mathbf{x}_0|\mathbf{y}_1). \end{aligned} \tag{7}$$

Note that this expression accounts for the full data assimilation cycle: forecast from t_0 to t_1 followed by an analysis at t_1. A more orthodox approach would be to use Bayes' rule:

$$p(\mathbf{x}_1|\mathbf{y}_1) \propto p(\mathbf{y}_1|\mathbf{x}_1)p(\mathbf{x}_1), \tag{8}$$

where $p(\mathbf{x}_1)$ would be assumed Gaussian as in the ensemble Kalman filter. Equations (7) and (8) are theoretically equivalent. In practice, however, with nonlinear models and suboptimal Gaussian schemes and hence approximated intermediary pdfs, the two approaches could lead to different analyses.

2.2. The cost function in ensemble space

Because we are interested in implementing an ensemble method, a reduced-rank formulation of cost function Eq. (6) is written in ensemble space, following Hunt *et al.* (2007) and Bocquet and Sakov (2012). The ensemble has N states $\left\{\mathbf{x}_{[n]}\right\}_{n=1,N}$, where the brackets around n mean that it refers to the member index in the ensemble, whereas the absence of brackets means that the index refers to time. $\mathbf{E}$ will denote the ensemble matrix, $\mathbf{E} = [\mathbf{x}_{[1]}, \ldots, \mathbf{x}_{[N]}]$, and the ensemble mean is $\overline{\mathbf{x}} = \mathbf{E}\mathbf{1}/N$, with $\mathbf{1} = (1, \ldots, 1)^{\mathrm{T}}$ a vector of size N. $\mathbf{A}$ will designate the anomaly matrix: $\mathbf{A} = [\mathbf{x}_{[1]} - \overline{\mathbf{x}}, \ldots, \mathbf{x}_{[N]} - \overline{\mathbf{x}}]$, and the corresponding ensemble covariance matrix is $\mathbf{A}\mathbf{A}^{\mathrm{T}}/(N-1)$. We look for an increment $\mathbf{x} - \overline{\mathbf{x}}$ in the space spanned by the ensemble anomalies, i.e. the columns of $\mathbf{A}$, so that $\mathbf{x}$ can be written $\mathbf{x} = \overline{\mathbf{x}} + \mathbf{A}\mathbf{w}$. Note that the vector $\mathbf{w}$ is not unique, since $\mathbf{x}$ does not depend on the translation $\mathbf{w} \leftarrow \mathbf{w} + \eta\mathbf{1}$, where η is a scalar constant. A restriction of cost function $\mathcal{J}$ to the ensemble space, where $\mathbf{P}_0^{-1}$ becomes the Moore–Penrose inverse of $\mathbf{A}_0\mathbf{A}_0^{\mathrm{T}}/(N-1)$, is

$$\widetilde{\mathcal{J}}(\mathbf{w}) = \mathcal{J}(\overline{\mathbf{x}}_0 + \mathbf{A}_0\mathbf{w}) + \mathcal{G}(\mathbf{w}), \tag{9}$$

where $\mathbf{A}_0$ is the anomaly matrix at t_0 and $\mathcal{G}(\mathbf{w})$ is a so-called gauge-fixing term meant to freeze the degree of freedom η but of no consequence to the optimal $\mathbf{x}_0$. With a proper choice of $\mathcal{G}$ (Bocquet, 2011) and identifying $\mathbf{x}_0^{(0)} = \overline{\mathbf{x}}_0$, one obtains

$$\begin{aligned} \widetilde{\mathcal{J}}(\mathbf{w}) =& \frac{1}{2}\left[\mathbf{y}_1 - H_1 \circ \mathcal{M}_{1\leftarrow 0}\left(\mathbf{x}_0^{(0)} + \mathbf{A}_0\mathbf{w}\right)\right]^{\mathrm{T}} \\ &\times \mathbf{R}_1^{-1}\left[\mathbf{y}_1 - H_1 \circ \mathcal{M}_{1\leftarrow 0}\left(\mathbf{x}_0^{(0)} + \mathbf{A}_0\mathbf{w}\right)\right] \\ &+ \frac{1}{2}(N-1)\mathbf{w}^{\mathrm{T}}\mathbf{w}. \end{aligned} \tag{10}$$

The analysis on $\mathbf{w}$ is obtained by the iterative minimization of this cost function. One has a choice of minimization scheme: for instance, Sakov *et al.* (2012) used a Gauss–Newton scheme whereas Bocquet and Sakov (2012) advocated the use of the Levenberg–Marquardt scheme (Levenberg, 1944; Marquardt, 1963) for strongly nonlinear systems. In this article we shall use a Gauss–Newton scheme, because the emphasis is not specifically on strongly nonlinear systems and the number of iterations for convergence in the experiments below is rather limited for most experiments.

2.3. Minimization in ensemble space

The iterative minimization of the cost function following the Gauss-Newton algorithm reads

$$\mathbf{w}^{(j+1)} = \mathbf{w}^{(j)} - \widetilde{\mathcal{H}}_{(j)}^{-1}\nabla\widetilde{\mathcal{J}}_{(j)}(\mathbf{w}^{(j)}), \tag{11}$$

where j is the iteration index and where the gradient and an approximate Hessian are given by

$$\mathbf{x}_0^{(j)} = \mathbf{x}_0^{(0)} + \mathbf{A}_0\mathbf{w}^{(j)}, \tag{12}$$

$$\begin{aligned} \nabla\widetilde{\mathcal{J}}_{(j)} =& -\mathbf{Y}_{(j)}^{\mathrm{T}}\mathbf{R}_1^{-1}\left[\mathbf{y}_1 - H_1 \circ \mathcal{M}_{1\leftarrow 0}(\mathbf{x}_0^{(j)})\right] \\ &+ (N-1)\mathbf{w}^{(j)}, \end{aligned} \tag{13}$$

$$\widetilde{\mathcal{H}}_{(j)} = (N-1)\mathbf{I}_N + \mathbf{Y}_{(j)}^{\mathrm{T}}\mathbf{R}_1^{-1}\mathbf{Y}_{(j)}, \tag{14}$$

where $\mathbf{I}_N$ is the identity matrix in ensemble space. $\mathbf{Y}_{(j)} = [H_1 \circ \mathcal{M}_{1\leftarrow 0}]'_{|\mathbf{x}_0^{(j)}}\mathbf{A}_0$ is the tangent linear of the operator from ensemble space to the observation space. The estimation of this sensitivity is a cornerstone of the method. Similarly to the work of Gu and Oliver (2007) and Liu *et al.* (2008), it does not require the use of the adjoint of the tangent linear of $H_1 \circ \mathcal{M}_{1\leftarrow 0}$. Two

different implementations of estimating $\mathbf{Y}_{(j)}$ were advocated by Sakov *et al.* (2012) and Bocquet and Sakov (2012). Here, we use the finite-difference estimate of the tangent linear operator:

$$\mathbf{Y}_{(j)} \approx \frac{1}{\varepsilon} H_1 \circ \mathcal{M}_{1\leftarrow 0}\left(\mathbf{x}_0^{(j)}\mathbf{1}^{\mathrm{T}} + \varepsilon \mathbf{A}_0\right)\left(\mathbf{I}_N - \frac{\mathbf{1}\mathbf{1}^{\mathrm{T}}}{N}\right), \tag{15}$$

where $\varepsilon \ll 1$ is a small scaling factor. We call it the *bundle* variant (Bocquet and Sakov, 2012).

Several stopping criteria of the minimization are possible. Here, the minimization is stopped when the root mean square of the increment $\mathbf{w}^{(j+1)} - \mathbf{w}^{(j)}$ becomes smaller than a predetermined $e > 0$.

2.4. Analysis and ensemble update

Another key point is the update of the analysis ensemble at t_0 and the forecast/analysis ensemble at the present time t_1.

In our scheme, a Gaussian approximation of $p(\mathbf{x}_0|\mathbf{y}_1)$ is computed using the argument $\mathbf{w}^\star$ of the minimum of $\widetilde{\mathcal{J}}(\mathbf{w})$ and the approximate Hessian of $\widetilde{\mathcal{J}}(\mathbf{w})$ at $\mathbf{w}^\star$,

$$\widetilde{\mathcal{H}}_\star = (N-1)\mathbf{I}_N + \mathbf{Y}_1^{\star\mathrm{T}}\mathbf{R}_1^{-1}\mathbf{Y}_1^\star, \tag{16}$$

as the approximate inverse error covariance matrix in ensemble space. The retrospective ensemble analysis at t_0 is then

$$\mathbf{E}_0^\star = \mathbf{x}_0^\star\mathbf{1}^{\mathrm{T}} + \sqrt{N-1}\mathbf{A}_0\widetilde{\mathcal{H}}_\star^{-1/2}\mathbf{U}, \tag{17}$$

where $\mathbf{x}_0^\star = \overline{\mathbf{x}}_0 + \mathbf{A}_0\mathbf{w}^\star$ and $\mathbf{U}$ is an orthogonal matrix satisfying $\mathbf{U}\mathbf{1} = \mathbf{1}$. We choose $\mathbf{U} = \mathbf{I}_N$ in this article. At the next cycle of the IEnKF, a Gaussian approximation of the pdf of $\mathbf{x}_1$ is needed to serve as a prior of the future analysis. From Eq. (7), the ensemble at t_1, which will serve as a prior for the next cycle, will be approximated as the forecast ensemble: $\mathbf{E}_1^\star = \mathcal{M}_{1\leftarrow 0}(\mathbf{E}_0^\star)$. Indeed, if

$$p(\mathbf{x}_0|\mathbf{y}_1) \simeq \frac{1}{N}\sum_{n=1}^{N}\delta\left(\mathbf{x}_0 - \mathbf{x}_{[n],1}^\star\right), \tag{18}$$

one has approximately

$$p(\mathbf{x}_1|\mathbf{y}_1) \simeq \frac{1}{N}\sum_{n=1}^{N}\delta\left[\mathbf{x}_1 - \mathcal{M}_{1\leftarrow 0}\left(\mathbf{x}_{[n],1}^\star\right)\right]. \tag{19}$$

From the present time t_1 it is now possible to produce any ensemble forecast using the updated ensemble.

2.5. Numerical implementation

A complete cycle of the bundle IEnKF is detailed in the pseudo-code of Algorithm 1 (see Table 1) using a Gauss–Newton scheme. The Levenberg–Marquardt counterparts can be found in Bocquet and Sakov (2012).

The final propagation of the ensemble in Algorithm 1 is essential for the fine performance of the method. It explains why the bundle IEnKF outperforms the IEKF of Sakov *et al.* (2012).

Note that there are a number of different inflation techniques. These, along with the multiplicative inflation (Anderson and Anderson, 1999) applied at line 19 of Algorithm 1, include additive inflation (Houtekamer and Mitchell, 2005) and relaxation-to-prior (Zhang *et al.*, 2004). Inflation can also be applied either prior to assimilation or posterior to assimilation.

2.6. Assets of the approach

Let us emphasize the numerical assets of this data assimilation scheme. Firstly, as an ensemble method, the minimization operates in a reduced space and Newton methods (and even more advanced methods) can be used because the ensemble size is not targeted to exceed $\mathcal{O}(10^2)$, whereas 4D-Var on realistic systems would rely on a minimization by Hessian-free methods: quasi-Newton, conjugate gradient. Secondly and very importantly, the method is free of any tangent linear model or any adjoint model.

Table 1. Algorithm 1: A cycle of the bundle/Gauss–Newton IEnKF.

Require: Transition model from t_0 to t_1; $\mathcal{M}_{1\leftarrow 0}$, observation operator H_1 at t_1. Algorithm parameters: ϵ, e, $j_{\max}$. $\mathbf{E}_0$, the ensemble at t_0, $\mathbf{y}_1$ the observation at t_1. λ is the inflation factor. $\mathbf{U}$ is an orthogonal matrix of size $N \times N$ satisfying $\mathbf{U}\mathbf{1} = \mathbf{1}$.

1: $j = 0$, $\mathbf{w} = \mathbf{0}$
2: $\mathbf{x}_0^{(0)} = \mathbf{E}_0\mathbf{1}/N$
3: $\mathbf{A}_0 = \mathbf{E}_0 - \mathbf{x}_0^{(0)}\mathbf{1}^{\mathrm{T}}$
4: **repeat**
5: $\quad\mathbf{x}_0 = \mathbf{x}_0^{(0)} + \mathbf{A}_0\mathbf{w}$
6: $\quad\mathbf{E}_0 = \mathbf{x}_0\mathbf{1}^{\mathrm{T}} + \epsilon\mathbf{A}_0$
7: $\quad\mathbf{E}_1 = \mathcal{M}_{1\leftarrow 0}(\mathbf{E}_0)$
8: $\quad\overline{\mathbf{y}}_1 = H_1(\mathbf{E}_1)\mathbf{1}/N$
9: $\quad\mathbf{Y}_1 = (H_1(\mathbf{E}_1) - \overline{\mathbf{y}}_1)/\epsilon$
10: $\quad\nabla\widetilde{\mathcal{J}} = (N-1)\mathbf{w} - \mathbf{Y}_1^{\mathrm{T}}\mathbf{R}_1^{-1}(\mathbf{y}_1 - \overline{\mathbf{y}}_1)$
11: $\quad\widetilde{\mathcal{H}} = (N-1)\mathbf{I}_N + \mathbf{Y}_1^{\mathrm{T}}\mathbf{R}_1^{-1}\mathbf{Y}_1$
12: $\quad$Solve $\widetilde{\mathcal{H}}\Delta\mathbf{w} = \nabla\widetilde{\mathcal{J}}$
13: $\quad\mathbf{w} := \mathbf{w} - \Delta\mathbf{w}$
14: $\quad j := j + 1$
15: **until** $||\Delta\mathbf{w}|| \leq e$ **or** $j \geq j_{\max}$
16: $\mathbf{E}_0 = \mathbf{x}_0\mathbf{1}^{\mathrm{T}} + \sqrt{N-1}\mathbf{A}_0\widetilde{\mathcal{H}}^{-\frac{1}{2}}\mathbf{U}$
17: $\mathbf{E}_1 = \mathcal{M}_{1\leftarrow 0}(\mathbf{E}_0)$
18: $\overline{\mathbf{x}}_1 = \mathbf{E}_1\mathbf{1}/N$
19: $\mathbf{E}_1 := \overline{\mathbf{x}}_1\mathbf{1}^{\mathrm{T}} + \lambda\left(\mathbf{E}_1 - \overline{\mathbf{x}}_1\mathbf{1}^{\mathrm{T}}\right)$

3. The iterative ensemble Kalman smoother (IEnKS)

The iterative ensemble Kalman smoother (IEnKS) extends the iterative ensemble Kalman filter, using an analysis over a window of L cycles (time length $L\Delta t$) instead of just $L = 1$ for the IEnKF. If, by definition, the present time is t_L and $\mathbf{A}_0$ is the ensemble anomaly matrix at t_0, i.e. L cycles earlier, then the goal is to estimate an $\mathbf{x}_0$ of the form $\mathbf{x}_0 = \overline{\mathbf{x}}_0 + \mathbf{A}_0\mathbf{w}$. With these notations, time indices are relative to the current DAW, in contrast to the double-index absolute-time notations of e.g. Cohn *et al.* (1994), which are suited for smoothers. Our choice is consistent with the notations often used in 4D-Var (e.g. Fisher *et al.*, 2005).

As time unfolds, the DAW of fixed length will slide. It is assumed that the DAW is shifted by $S\Delta t$, when moving to the next data assimilation cycle. Figure 1 is a schematic of how a DAW slides.

If $1 \leq S < L$, the DAWs overlap. If $S = L$, the DAWs do not overlap. Note that the DAWs of IEnKF, i.e. $S = L = 1$, do not overlap. Fast schemes would imply non-overlapping DAWs: the shift corresponds to the full length of the DAW, $S = L$. Focused on performance rather than computational efficiency, we choose in this article to restrict ourselves to a scheme where the shift is equal to Δt, i.e. $S = 1$. The study of other schemes with larger shifts is left for future work. As a consequence, the schemes considered in this article can be viewed as quasi-static, since, when the DAW is shifted, only one batch of observations comes in and only one batch leaves. In the process, the underlying analysis cost function is often only slightly deformed and its global minimum can hopefully be tracked from one cycle to another. As a consequence of $S = 1$, the DAWs overlap in this study (as seen in Figure 1), except with the IEnKF ($L = 1$).

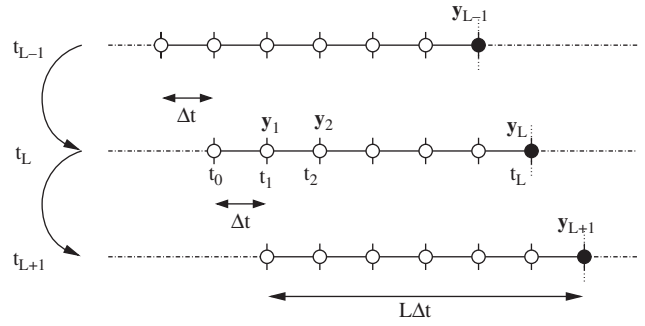

Figure 1. Chaining of the IEnKS cycles. This illustrates the case $L = 6$ and $S = 1$. $S = 1$ corresponds to the quasi-static case studied in this article. Here, the DAWs overlap. Additionally, in the SDA IEnKS, for a given analysis, only the latest observation vector of the DAW is assimilated (full black dot). The other five observation vectors have already been assimilated in previous cycles. In the MDA IEnKS, for a given analysis, all observations vectors of the DAW could be assimilated (each with a certain weight).

Whether a given observation vector is assimilated just once through a whole run (single data assimilation, denoted SDA) or several times (multiple data assimilation, denoted MDA) is a distinct choice. Because the SDA implementation of IEnKS turns out to be much simpler and parallels the IEnKF, it is introduced first.

3.1. Single data assimilation scheme

With the SDA scheme and specifically $S = 1$, only the latest observation vector is assimilated in the analysis over a DAW of length L. The other $L - 1$ observation vectors within the DAW have already been assimilated in the previous cycles.

3.1.1. Derivation of the scheme

At time t_0, L cycles in the past, the pdf of state $\mathbf{x}_0$, $p(\mathbf{x}_0|\mathbf{y}_{:L})$, is assumed to follow a Gaussian pdf with mean state $\mathbf{x}_0^{(0)}$ and covariance matrix $\mathbf{P}_0$. This is the outcome of the previous analysis at t_{-1} forecast at t_0. These moments now serve as the first guess and forecast-error covariance matrix of the subsequent smoother analysis. As opposed to the presentation of the IEnKF, the conditional dependence on past observations is now explicit: the notation $\mathbf{y}_{:L}$ denotes all past observation vectors up to and including $\mathbf{y}_L$.

One assumes that the observation likelihood at t_k is

$$p(\mathbf{y}_k|\mathbf{x}_k) = n\left(\mathbf{y}_k - H_k(\mathbf{x}_k)|\mathbf{0}, \mathbf{R}_k\right). \tag{20}$$

As for the IEnKF, nonlinear observation operators can be handled by the method as well. We have assumed that the observation errors are white in time. This restriction might be lifted for the MDA variants to be defined later, but cannot be so in this SDA approach.

Within the DAW, the system evolves between t_k and t_{k+1} according to a model $\mathcal{M}_{k+1\leftarrow k}$:

$$p(\mathbf{x}_k|\mathbf{x}_{k-1}) \propto \delta\left[\mathbf{x}_k - \mathcal{M}_{k\leftarrow k-1}(\mathbf{x}_{k-1})\right]. \tag{21}$$

The smoothing pdf of state $\mathbf{x}_0$ conditional on all observations up to $\mathbf{y}_L$ is obtained using Bayes' rule and the perfect model assumption:

$$\begin{aligned} p(\mathbf{x}_0|\mathbf{y}_{:L}) &\propto p(\mathbf{x}_0|\mathbf{y}_{:L-1})p(\mathbf{y}_L|\mathbf{x}_0, \mathbf{y}_{:L-1}) \\ &\propto p(\mathbf{x}_0|\mathbf{y}_{:L-1})p(\mathbf{y}_L|\mathbf{x}_L = \mathcal{M}_{L\leftarrow 0}(\mathbf{x}_0)). \end{aligned} \tag{22}$$

Then, the pdfs within the DAW are obtained via

$$\begin{aligned} p(\mathbf{x}_k|\mathbf{y}_{:L}) &= \int \mathrm{d}\mathbf{x}_0\, p(\mathbf{x}_k|\mathbf{x}_0, \mathbf{y}_{:L})p(\mathbf{x}_0|\mathbf{y}_{:L}) \\ &= \int \mathrm{d}\mathbf{x}_0\, \delta\left[\mathbf{x}_k - \mathcal{M}_{k\leftarrow 0}(\mathbf{x}_0)\right] p(\mathbf{x}_0|\mathbf{y}_{:L}), \end{aligned} \tag{23}$$

for $k = 1, \ldots, L$. This includes the filtering pdf $p(\mathbf{x}_L|\mathbf{y}_{:L})$, but also $p(\mathbf{x}_1|\mathbf{y}_{:L})$ to be used as the prior pdf of the next data assimilation cycle.

3.1.2. The cost function in ensemble space

Looking for a solution in ensemble space, $\mathbf{x}_0 = \overline{\mathbf{x}}_0 + \mathbf{A}_0\mathbf{w}$, and following the derivation of the IEnKF in section 2, one obtains from Eq. (22) the ensemble space cost function, setting $\mathbf{x}_0^{(0)} = \overline{\mathbf{x}}_0$:

$$\begin{aligned} \widetilde{\mathcal{J}}(\mathbf{w}) = &\frac{1}{2}(N-1)\mathbf{w}^{\mathrm{T}}\mathbf{w} \\ &+ \frac{1}{2}\left[\mathbf{y}_L - H_L \circ \mathcal{M}_{L\leftarrow 0}\left(\mathbf{x}_0^{(0)} + \mathbf{A}_0\mathbf{w}\right)\right]^{\mathrm{T}} \\ &\times \mathbf{R}_L^{-1}\left[\mathbf{y}_L - H_L \circ \mathcal{M}_{L\leftarrow 0}\left(\mathbf{x}_0^{(0)} + \mathbf{A}_0\mathbf{w}\right)\right]. \end{aligned} \tag{24}$$

It can be minimized without the use of the tangent linear and adjoint models, using the same schemes as the IEnKF. All the considerations of section 2.3 apply, albeit replacing indices referring to t_1 by indices referring to t_L.

3.1.3. Analysis and ensemble update

In the IEnKS, the analysis and the new ensemble at t_1 are obtained similarly to the IEnKF. The only difference is that the ensemble is extracted from the Gaussian approximation of $p(\mathbf{x}_0|\mathbf{y}_{:L})$, which is computed using the argument $\mathbf{w}^\star$ of the minimum of $\widetilde{\mathcal{J}}(\mathbf{w})$ and using its approximate Hessian at $\mathbf{w}^\star$,

$$\widetilde{\mathcal{H}}_\star = (N-1)\mathbf{I}_N + \mathbf{Y}_L^{\star\mathrm{T}}\mathbf{R}_L^{-1}\mathbf{Y}_L^\star, \tag{25}$$

as the approximate inverse error covariance matrix in ensemble space. Following Eq. (23), the new ensemble at t_1 is obtained by a forecast of the ensemble at t_0.

Optionally, an analysis and a posterior ensemble may also be computed for all the states up to time t_L (and possibly beyond), even though the cycling of the smoother does not require it. To do so, the analysis at t_k can be obtained from $\mathbf{x}_k^\star = \mathcal{M}_{k\leftarrow 0}(\mathbf{x}_0^\star)$, or from the ensemble forecast $\mathbf{E}_k^\star = \mathcal{M}_{k\leftarrow 0}(\mathbf{E}_0^\star)$.

3.1.4. Numerical implementation

A complete cycle of the IEnKS scheme is described by the pseudo-code of Algorithm 2 (see Table 2). The forecast part from t_1 to t_L is not written explicitly in the algorithm, as it is only an optional component of the cycle. As mentioned earlier, the numerical tests to follow do not require localization for a satisfying performance. However, the need for inflation has to be addressed.

In Algorithm 2, tuned inflation (λ) is used. An alternative is the inflation-free scheme of the finite-size filters (Bocquet, 2011). It has been extended to the finite-size IEnKF in Bocquet and Sakov (2012), which is used here to adapt Algorithm 2 to the finite-size variant IEnKS-N, detailed in Algorithm 3 (see Table 3). We essentially extend the so-called primal IEnKF-N version, with a minor modification compared with Bocquet and Sakov (2012): within the minimization of the cost function (not the posterior ensemble update), the Hessian of Algorithm 2 is used. This way, the spectrum of the Hessian has a lower positive bound. It deviates from the Gauss–Newton solution by making the method closer to a steepest descent algorithm when the inflation diagnosed by the finite-size scheme is expected to be much larger than 1. Hence, it borrows from the Levenberg–Marquardt approach (see the discussion in Bocquet and Sakov, 2012).

The IEnKS-N is designed mainly for mildly and strongly nonlinear conditions. However, with Jeffreys non-informative distribution as a default hyperprior, it can be suboptimal in an effectively almost linear regime (Bocquet and Sakov, 2012). In this case, one can revert to a standard filter/smoother with small inflation or use a fix of the finite-size schemes instead.

Table 2. Algorithm 2: A cycle of the lag-L/SDA/bundle/Gauss–Newton IEnKS.

Require: t_L is present time. Transition model $\mathcal{M}_{k+1\leftarrow k}$, observation operators H_k at t_k. Algorithm parameters: ϵ, e, $j_{\max}$. $\mathbf{E}_0$, the ensemble at t_0, $\mathbf{y}_k$ the observation at t_k. λ is the inflation factor. $\mathbf{U}$ is an orthogonal matrix of size $N \times N$ satisfying $\mathbf{U1} = \mathbf{1}$.

$j = 0, \mathbf{w} = \mathbf{0}$
$\mathbf{x}_0^{(0)} = \mathbf{E}_0\mathbf{1}/N$
$\mathbf{A}_0 = \mathbf{E}_0 - \mathbf{x}_0^{(0)}\mathbf{1}^{\mathrm{T}}$
repeat
  $\mathbf{x}_0 = \mathbf{x}_0^{(0)} + \mathbf{A}_0\mathbf{w}$
  $\mathbf{E}_0 = \mathbf{x}_0\mathbf{1}^{\mathrm{T}} + \epsilon\mathbf{A}_0$
  $\mathbf{E}_L = \mathcal{M}_{L\leftarrow 0}(\mathbf{E}_0)$
  $\overline{\mathbf{y}}_L = H_L(\mathbf{E}_L)\mathbf{1}/N$
  $\mathbf{Y}_L = (H_L(\mathbf{E}_L) - \overline{\mathbf{y}}_L)/\epsilon$
  $\nabla\widetilde{\mathcal{J}} = (N-1)\mathbf{w} - \mathbf{Y}_L^{\mathrm{T}}\mathbf{R}_L^{-1}(\mathbf{y}_L - \overline{\mathbf{y}}_L)$
  $\widetilde{\mathcal{H}} = (N-1)\mathbf{I}_N + \mathbf{Y}_L^{\mathrm{T}}\mathbf{R}_L^{-1}\mathbf{Y}_L$
  Solve $\widetilde{\mathcal{H}}\Delta\mathbf{w} = \nabla\widetilde{\mathcal{J}}$
  $\mathbf{w} := \mathbf{w} - \Delta\mathbf{w}$
  $j := j + 1$
until $||\Delta\mathbf{w}|| \le e$ **or** $j \ge j_{\max}$
$\mathbf{E}_0 = \mathbf{x}_0\mathbf{1}^{\mathrm{T}} + \sqrt{N-1}\mathbf{A}_0\widetilde{\mathcal{H}}^{-\frac{1}{2}}\mathbf{U}$
$\mathbf{E}_1 = \mathcal{M}_{1\leftarrow 0}(\mathbf{E}_0)$
$\overline{\mathbf{x}}_1 = \mathbf{E}_1\mathbf{1}/N$
$\mathbf{E}_1 := \overline{\mathbf{x}}_1\mathbf{1}^{\mathrm{T}} + \lambda\left(\mathbf{E}_1 - \overline{\mathbf{x}}_1\mathbf{1}^{\mathrm{T}}\right)$

Table 3. Algorithm 3: A cycle of the lag-L/SDA/bundle/Gauss–Newton IEnKS-N. Its pseudo-code is identical to Algorithm 2, with the exception of the following lines.

$\nabla\widetilde{\mathcal{J}} = N\frac{\mathbf{w}}{\varepsilon_N + \mathbf{w}^{\mathrm{T}}\mathbf{w}} - \mathbf{Y}_L^{\mathrm{T}}\mathbf{R}_L^{-1}(\mathbf{y}_L - \overline{\mathbf{y}}_L)$
$\widetilde{\mathcal{H}} = N\frac{(\varepsilon_N + \mathbf{w}_0^{\mathrm{T}}\mathbf{w}_0)\mathbf{I}_N - 2\mathbf{w}_0\mathbf{w}_0^{\mathrm{T}}}{(\varepsilon_N + \mathbf{w}_0^{\mathrm{T}}\mathbf{w}_0)^2} + \mathbf{Y}_L^{\mathrm{T}}\mathbf{R}_L^{-1}\mathbf{Y}_L$
$\mathbf{E}_0 = \mathbf{x}_0\mathbf{1}^{\mathrm{T}} + \sqrt{N-1}\mathbf{A}_0\widetilde{\mathcal{H}}^{-\frac{1}{2}}\mathbf{U}$
$\mathbf{E}_1 = \mathcal{M}_{1\leftarrow 0}(\mathbf{E}_0)$

3.2. Multiple data assimilation scheme

In this section, multiple data assimilation (Yang *et al.*, 2012; Emerick and Reynolds, 2013) within the IEnKS is studied. Note that the SDA IEnKS will appear as a particular, simpler subcase of the MDA IEnKS.

The observation likelihoods are assumed Gaussian-distributed, following Eq. (20). For the sake of simplicity, they are assumed white in time. With MDA, an observation is consistently assimilated several times with adequately inflated variances. *Consistently* means that, with a perfect linear model and Gaussian statistics, the MDA scheme theoretically yields the same outcome as an SDA scheme. Additionally, the observation prior errors used in the multiple assimilation of an observation vector are *chosen* to be independent. An observation vector $\mathbf{y}$ is said to be assimilated with weight β ($0 \le \beta \le 1$), if the following Gaussian observation likelihood is used in the analysis:

$$p(\mathbf{y}^{\beta}|\mathbf{x}) \equiv n\left(\mathbf{y} - H(\mathbf{x})|\mathbf{0}, \beta^{-1}\mathbf{R}\right). \tag{26}$$

The upper index of $\mathbf{y}^{\beta}$ refers to its partial assimilation with weight β.

Because of the assumption $S = 1$, an observation vector will be assimilated for the first time at t_L with weight β_L, one cycle later at position t_{L-1} in the DAW with weight β_{L-1}, and so on until, $L-1$ cycles later, at position t_1 in the DAW with weight β_1, and will finally leave the DAW. Hence, the sequence $\beta_1, \beta_2, \ldots, \beta_L$ characterizes the distribution of the weights within the DAW, but also the successive weights ascribed to an observation vector. For statistical consistency, the sum of the weights must satisfy $\sum_{k=1}^{L}\beta_k = 1$, so that the sum of the precision matrices for an observation vector $\mathbf{y}$, $\beta_1\mathbf{R}^{-1}, \beta_2\mathbf{R}^{-1}, \ldots, \beta_L\mathbf{R}^{-1}$, will be $\mathbf{R}^{-1}$.

3.2.1. Heuristic derivation of the scheme

We wish to perform an analysis using the weighted set of observation vectors $\mathbf{y}_{:L}^{\boldsymbol{\beta}_{:L}} = \left\{\mathbf{y}_L^{\beta_L}, \mathbf{y}_{L-1}^{\beta_{L-1}}, \ldots\right\}$. Because of the fixed DAW length, $\beta_k = 0$ is enforced for $k \le 0$ so that $\mathbf{y}_{:L}^{\boldsymbol{\beta}_{:L}} = \left\{\mathbf{y}_L^{\beta_L}, \mathbf{y}_{L-1}^{\beta_{L-1}}, \ldots, \mathbf{y}_1^{\beta_1}\right\}$. We would like to estimate the smoothing pdf of state $\mathbf{x}_0$, L cycles in the past, conditional on all past observations prior to t_1 and a fraction of the observations in the DAW: $p(\mathbf{x}_0|\mathbf{y}_{:L}^{\boldsymbol{\alpha}_{:L}}) \equiv p(\mathbf{x}_0|\mathbf{y}_L^{\alpha_L}, \mathbf{y}_{L-1}^{\alpha_{L-1}}, \ldots)$. The sequence $\boldsymbol{\alpha}_{:L}$ defined up to t_L stipulates that only a fraction α_k of observation vector $\mathbf{y}_k$ has already been assimilated and accounted for. Because of the fixed DAW length, $\alpha_k = 0$ is enforced for $k \le 0$: once an observation vector has left the DAW, it has been entirely assimilated. The α_k, with $1 \le k \le L$, will be shown shortly to be given by

$$\alpha_k = \sum_{l=k}^{L}\beta_l, \tag{27}$$

which merely sums the fractions of an observation vector that have already been assimilated.

The targeted pdf $p(\mathbf{x}_0|\mathbf{y}_{:L}^{\boldsymbol{\alpha}_{:L}})$ can be obtained using a heuristic derivation relying on Bayes' rule and the perfect model assumption:

$$\begin{aligned} p(\mathbf{x}_0|\mathbf{y}_{:L}^{\boldsymbol{\alpha}_{:L}}) &\propto p(\mathbf{x}_0|\mathbf{y}_{:L}^{\boldsymbol{\alpha}_{:L}-\boldsymbol{\beta}_{:L}})p(\mathbf{y}_{:L}^{\boldsymbol{\beta}_{:L}}|\mathbf{x}_0, \mathbf{y}_{:L}^{\boldsymbol{\alpha}_{:L}-\boldsymbol{\beta}_{:L}}) \\ &\propto p(\mathbf{x}_0|\mathbf{y}_{:L}^{\boldsymbol{\alpha}_{:L}-\boldsymbol{\beta}_{:L}})\prod_{k=1}^{L}p(\mathbf{y}_k^{\boldsymbol{\beta}_k}|\mathbf{x}_k = \mathcal{M}_{k\leftarrow 0}(\mathbf{x}_0)). \end{aligned} \tag{28}$$

To be able to repeat the data assimilation cycle conveniently, the posterior should have the same functional form as the prior except shifted by Δt. Hence, we would like $p(\mathbf{x}_0|\mathbf{y}_{:L}^{\boldsymbol{\alpha}_{:L}})$ to have the same conditional dependence as $p(\mathbf{x}_0|\mathbf{y}_{:L}^{\boldsymbol{\alpha}_{:L}-\boldsymbol{\beta}_{:L}})$, albeit with a shift of one Δt. Consequently, one first requires

$$\alpha_L = \beta_L, \tag{29}$$

but also, for any $k \le L$,

$$\alpha_{k-1} = \alpha_k + \beta_{k-1}. \tag{30}$$

Therefore a given sequence $\boldsymbol{\beta}_{1:L}$ fully determines the sequence $\boldsymbol{\alpha}_{:L}$. Moreover, one obtains Eq. (27). As a consequence of $\sum_{k=1}^{L}\beta_k = 1$, one has $0 \le \alpha_k \le 1$.

One simple choice for the β_k consists of the sequence $\beta_L = 1$ and $\beta_k = 0$ for any $k \le L-1$. This implies that $\alpha_k = 1$ for any $k \le L$. This is merely the SDA scheme.

Other schemes with MDA are possible. We may want to use them for the sake of numerical stability because, for lengthy DAWs, $\mathbf{x}_0$ and $\mathbf{y}_L$ may only be connected by a very long run of the model (more on that in the final discussion of section 5). We have tested several weight distributions with similar performance. For instance, we could choose $\beta_k = L^{-1}$ for $k = 1, \ldots, L$. Consequently, $\alpha_k = 1 - (k-1)/L$ for $k = 1, \ldots, L$. This *uniform* scheme assigns equal weights to all observations.

Equation (28) stipulates how to estimate $p(\mathbf{x}_0|\mathbf{y}_{:L}^{\boldsymbol{\alpha}_{:L}})$ recursively. However, the pdf of interest is instead $p(\mathbf{x}_0|\mathbf{y}_{:L})$, which can

nevertheless be obtained by

$$\begin{aligned} p(\mathbf{x}_0|\mathbf{y}_{:L}) &\propto p(\mathbf{x}_0|\mathbf{y}_{:L}^{\boldsymbol{\alpha}_{:L}}) p(\mathbf{y}_{:L}^{1-\boldsymbol{\alpha}_{:L}}|\mathbf{x}_0, \mathbf{y}_{:L}^{\boldsymbol{\alpha}_{:L}}) \\ &\propto p(\mathbf{x}_0|\mathbf{y}_{:L}^{\boldsymbol{\alpha}_{:L}}) p(\mathbf{y}_{:L}^{1-\boldsymbol{\alpha}_{:L}}|\mathbf{x}_0) \\ &\propto p(\mathbf{x}_0|\mathbf{y}_{:L}^{\boldsymbol{\alpha}_{:L}}) \prod_{k=1}^{L} p(\mathbf{y}_k^{1-\alpha_k}|\mathbf{x}_k = \mathcal{M}_{k\leftarrow 0}(\mathbf{x}_0)). \end{aligned} \tag{31}$$

This shall be called the *balancing* step, as it effectively redistributes the weights within the DAW. Obviously, the balancing step is unnecessary in the SDA scheme, since, in this case, the two targeted pdfs, $p(\mathbf{x}_0|\mathbf{y}_{:L})$ and $p(\mathbf{x}_0|\mathbf{y}_{:L}^{\boldsymbol{\alpha}_{:L}})$, coincide.

Like the SDA scheme, the filtering distribution $p(\mathbf{x}_L|\mathbf{y}_{:L})$ is obtained via Eq. (23). Differently, the prior for the next cycle is obtained by

$$\begin{aligned} p(\mathbf{x}_1|\mathbf{y}_{:L}^{\boldsymbol{\alpha}_{:L}}) &= \int \mathrm{d}\mathbf{x}_0 \, p(\mathbf{x}_1|\mathbf{x}_0, \mathbf{y}_{:L}^{\boldsymbol{\alpha}_{:L}}) p(\mathbf{x}_0|\mathbf{y}_{:L}^{\boldsymbol{\alpha}_{:L}}) \\ &= \int \mathrm{d}\mathbf{x}_0 \, \delta\left[\mathbf{x}_1 - \mathcal{M}_{k\leftarrow 0}(\mathbf{x}_0)\right] p(\mathbf{x}_0|\mathbf{y}_{:L}^{\boldsymbol{\alpha}_{:L}}) \,. \end{aligned} \tag{32}$$

Like the IEnKF and IEnKS, $p(\mathbf{x}_0|\mathbf{y}_{:L}^{\boldsymbol{\alpha}_{:L}})$ will subsequently be assumed to follow approximately a Gaussian distribution with mean state $\mathbf{x}_0^{(0)}$ and covariance matrix $\mathbf{P}_0$.

3.2.2. *The cost function in ensemble space*

Looking for a solution in ensemble space, $\mathbf{x}_0 = \overline{\mathbf{x}}_0 + \mathbf{A}_0\mathbf{w}$, and following the derivation of the SDA IEnKS, one obtains the ensemble space cost function from Eq. (28):

$$\begin{aligned} \widetilde{\mathcal{J}}(\mathbf{w}) = &\frac{1}{2}(N-1)\mathbf{w}^{\mathrm{T}}\mathbf{w} \\ &+ \frac{1}{2}\sum_{k=1}^{L}\left[\mathbf{y}_k - H_k \circ \mathcal{M}_{k\leftarrow 0}\left(\mathbf{x}_0^{(0)} + \mathbf{A}_0\mathbf{w}\right)\right]^{\mathrm{T}} \\ &\times \beta_k \mathbf{R}_k^{-1}\left[\mathbf{y}_k - H_k \circ \mathcal{M}_{k\leftarrow 0}\left(\mathbf{x}_0^{(0)} + \mathbf{A}_0\mathbf{w}\right)\right]. \end{aligned} \tag{33}$$

All the considerations of section 2.3 apply, with straightforward adaptations.

3.2.3. *Analysis and ensemble update*

In the IEnKS, the analysis and the new ensemble at t_1 are obtained similarly to the IEnKF. The only difference is that the Gaussian approximation of $p(\mathbf{x}_0|\mathbf{y}_{:L}^{\boldsymbol{\alpha}_{:L}})$ is computed using the argument $\mathbf{w}^\star$ of the minimum of $\mathcal{J}(\mathbf{w})$ and using its approximate Hessian:

$$\widetilde{\mathcal{H}}_\star = (N-1)\mathbf{I}_N + \sum_{k=1}^{L} \mathbf{Y}_k^{\star\mathrm{T}} \beta_k \mathbf{R}_k^{-1} \mathbf{Y}_k^\star \tag{34}$$

as the approximate inverse error covariance matrix in ensemble space. Again, the new and future prior ensemble at t_1 is obtained by a forecast of the ensemble at t_0.

Optionally, an analysis and a posterior ensemble may also be computed for all the states up to time t_L, even though the cycling of the smoother does not require it. To do so, $\mathbf{x}_k^\star = \mathcal{M}_{k\leftarrow 0}(\mathbf{x}_0^\star)$ provides an estimate at t_k, while $\mathbf{E}_k^\star = \mathcal{M}_{k\leftarrow 0}(\mathbf{E}_0^\star)$ alternatively provides an ensemble estimate. However, for MDA it is only an approximation since it is meant to sample $p(\mathbf{x}_k|\mathbf{y}_{:L}^{\boldsymbol{\alpha}_{:L}})$ rather than $p(\mathbf{x}_k|\mathbf{y}_{:L})$.

A more accurate solution that accounts for the balancing step and targets $p(\mathbf{x}_k|\mathbf{y}_{:L})$ consists of exploiting Eq. (31). The outcome of the previous analysis $p(\mathbf{x}_0|\mathbf{y}_{:L}^{\boldsymbol{\alpha}_{:L}})$ is a Gaussian pdf with mean $\mathbf{x}_0^\star$ and with covariance matrix $\mathbf{P}_0^\star$. A solution in ensemble space is sought: $\mathbf{x}_0 = \overline{\mathbf{x}}_0^\star + \mathbf{A}_0^\star\mathbf{w}$, which would minimize

$$\begin{aligned} \widetilde{\mathcal{J}}^\star(\mathbf{w}) = &\frac{1}{2}(N-1)\mathbf{w}^{\mathrm{T}}\mathbf{w} \\ &+ \frac{1}{2}\sum_{k=1}^{L}(1-\alpha_k)\left[\mathbf{y}_k - H_k \circ \mathcal{M}_{k\leftarrow 0}\left(\overline{\mathbf{x}}_0^\star + \mathbf{A}_0^\star\mathbf{w}\right)\right]^{\mathrm{T}} \\ &\times \mathbf{R}_k^{-1}\left[\mathbf{y}_k - H_k \circ \mathcal{M}_{k\leftarrow 0}\left(\overline{\mathbf{x}}_0^\star + \mathbf{A}_0^\star\mathbf{w}\right)\right]. \end{aligned} \tag{35}$$

The minimum and Hessian of $\widetilde{\mathcal{J}}^\star$ lead to a Gaussian posterior pdf of moments $\mathbf{x}_0^\odot$ and $\mathbf{P}_0^\odot$, from which a forecast or an ensemble forecast can be performed up to t_L and beyond using Eq. (23).

For the test cases ahead, we did not find obvious differences in the performance of $\mathbf{x}_0^\star$ or $\mathbf{x}_0^\odot$ used as a smoothing estimation of $\mathbf{x}_0$. Therefore, if retrospective analysis is the objective, the extra computation of $\mathbf{x}_0^\odot$ may not be needed. However, in the examples ahead, we found significant differences in the forecast of $\mathbf{x}_L$ from $\mathbf{x}_0^\star$ or $\mathbf{x}_0^\odot$, as anticipated.

An alternative to computing the Gaussian approximation of $p(\mathbf{x}_0|\mathbf{y}_{:L})$ is to minimize

$$\begin{aligned} \widetilde{\mathcal{J}}^\star(\mathbf{w}) = &\frac{1}{2}(N-1)\mathbf{w}^{\mathrm{T}}\mathbf{w} \\ &+ \frac{1}{2}\sum_{k=1}^{L}(1-\alpha_{k-1})\left[\mathbf{y}_k - H_k \circ \mathcal{M}_{k\leftarrow 0}\left(\overline{\mathbf{x}}_0 + \mathbf{A}_0\mathbf{w}\right)\right]^{\mathrm{T}} \\ &\times \mathbf{R}_k^{-1}\left[\mathbf{y}_k - H_k \circ \mathcal{M}_{k\leftarrow 0}\left(\overline{\mathbf{x}}_0 + \mathbf{A}_0\mathbf{w}\right)\right], \end{aligned} \tag{36}$$

using the prior anomalies $\mathbf{A}_0$ rather than the newly generated $\mathbf{A}_0^\star$. Both choices were implemented, but we did not find any significant difference in the outcome of both approaches in the experiments of section 4.

3.2.4. *Numerical implementation*

A complete cycle of the MDA IEnKS scheme is described by Algorithm 4 (see Table 4). The forecast part from t_0 to t_L and, if required, the balancing step are not written explicitly in the algorithm, since they are optional.

3.2.5. *Adaptive inflation schemes and MDA*

In Algorithm 4, tuned inflation is used. For some of the numerical experiments ahead, which require hundreds of smoother runs in various conditions, tuning of inflation is very costly. As an alternative, it is straightforward to develop a finite-size MDA IEnKS from Algorithms 3 and 4. However, we found that, in practice, this leads to the divergence of the filter (except in the SDA case). The finite-size approach can be seen as an adaptive inflation scheme (Bocquet and Sakov, 2012). Like other adaptive inflation schemes (Anderson, 2007; Miyoshi, 2011; Liang *et al.*, 2012), the inflation of the prior is implicitly tuned so as to optimize a criterion, related to the likelihood of the innovations given their prior statistics. For instance, according to Bocquet and Sakov (2012), the *optimal* inflation would be directly inferred in this context from the minimum $\zeta^\star$ of

$$\mathcal{D}(\zeta) = \inf_{\mathbf{w}}\left[\widetilde{\mathcal{J}}_o(\mathbf{w}) + \frac{\zeta}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}\right] + \frac{\varepsilon_N\zeta}{2} + \frac{N}{2}\ln\frac{N}{\zeta} - \frac{N}{2}, \tag{37}$$

where $\widetilde{\mathcal{J}}_o(\mathbf{w})$ is the observation term of $\widetilde{\mathcal{J}}(\mathbf{w})$ in Eq. (33), which depends on the innovations accounted for in the DAW and on their prior statistics. (The *optimal* inflation of the prior covariance matrix is given by $(N-1)/\zeta^\star$.) However, in the MDA case, the true errors in the observation vector $\mathbf{y}_k$ with weight $0 < \beta_k < 1$, and hence the related innovations, do not match the prior variances in $\beta_k\mathbf{R}_k^{-1}$ and hence the same related prior

Table 4. Algorithm 4: A cycle of the lag-L/MDA/bundle/Gauss–Newton IEnKS.

Require: t_L is present time. Transition model $\mathcal{M}_{k+1\leftarrow k}$, observation operators H_k at t_k. Algorithm parameters: ϵ, e, $j_{\max}$. $\mathbf{E}_0$, the ensemble at t_0, $\mathbf{y}_k$ the observation at t_k. λ is the inflation factor. $\mathbf{U}$ is an orthogonal matrix of size $N \times N$ satisfying $\mathbf{U1} = \mathbf{1}$. β_k, $1 \leq k \leq L$, are the observation weights within the DAW.

$j = 0$, $\mathbf{w} = \mathbf{0}$
$\mathbf{x}_0^{(0)} = \mathbf{E}_0\mathbf{1}/N$
$\mathbf{A}_0 = \mathbf{E}_0 - \mathbf{x}_0^{(0)}\mathbf{1}^{\mathrm{T}}$
repeat
  $\mathbf{x}_0 = \mathbf{x}_0^{(0)} + \mathbf{A}_0\mathbf{w}$
  $\mathbf{E}_0 = \mathbf{x}_0\mathbf{1}^{\mathrm{T}} + \epsilon\mathbf{A}_0$
  for $k = 1, \ldots, L$ **do**
    $\mathbf{E}_k = \mathcal{M}_{k\leftarrow k-1}(\mathbf{E}_{k-1})$
    $\overline{\mathbf{y}}_k = H_k(\mathbf{E}_k)\mathbf{1}/N$
    $\mathbf{Y}_k = (H_k(\mathbf{E}_k) - \overline{\mathbf{y}}_k)/\epsilon$
  end for
  $\nabla\widetilde{\mathcal{J}} = (N-1)\mathbf{w} - \sum_{k=1}^{L}\mathbf{Y}_k^{\mathrm{T}}\beta_k\mathbf{R}_k^{-1}(\mathbf{y}_k - \overline{\mathbf{y}}_k)$
  $\widetilde{\mathcal{H}} = (N-1)\mathbf{I}_N + \sum_{k=1}^{L}\mathbf{Y}_k^{\mathrm{T}}\beta_k\mathbf{R}_k^{-1}\mathbf{Y}_k$
  Solve $\widetilde{\mathcal{H}}\Delta\mathbf{w} = \nabla\widetilde{\mathcal{J}}$
  $\mathbf{w} := \mathbf{w} - \Delta\mathbf{w}$
  $j := j + 1$
until $||\Delta\mathbf{w}|| \leq e$ **or** $j \geq j_{\max}$
$\mathbf{E}_0 = \mathbf{x}_0\mathbf{1}^{\mathrm{T}} + \sqrt{N-1}\mathbf{A}_0\widetilde{\mathcal{H}}^{-\frac{1}{2}}\mathbf{U}$
$\mathbf{E}_1 = \mathcal{M}_{1\leftarrow 0}(\mathbf{E}_0)$
$\overline{\mathbf{x}}_1 = \mathbf{E}_1\mathbf{1}/N$
$\mathbf{E}_1 := \overline{\mathbf{x}}_1\mathbf{1}^{\mathrm{T}} + \lambda\left(\mathbf{E}_1 - \overline{\mathbf{x}}_1\mathbf{1}^{\mathrm{T}}\right)$

innovation statistics. The tuning of ζ (and hence of inflation) cannot account for this misfit, which leads to a failure of the adaptive scheme.

A solution to this would be to relax the constraint $\sum_{k=1}^{L}\beta_k = 1$ and, for instance, to choose $\beta_k \equiv 1$. This is inconsistent with a view to estimating the pdf $p(\mathbf{x}_0|\mathbf{y}_{:L})$. However, the choice $\beta_k \equiv 1$ is consistent with the estimation of the Lth power of the same pdf. To give a heuristic proof, Eq. (22) is considered, together with the uniform scheme for the weights:

$$\begin{aligned} p(\mathbf{x}_0|\mathbf{y}_L^{1/L}, \mathbf{y}_{L-1}^{2/L}, \ldots, \mathbf{y}_1, \mathbf{y}_0, \ldots) \\ \propto p(\mathbf{x}_0|\mathbf{y}_{L-1}^{1/L}, \mathbf{y}_{L-2}^{2/L}, \ldots, \mathbf{y}_1^{(L-1)/L}, \mathbf{y}_0, \ldots) \\ \times \prod_{k=1}^{L} p(\mathbf{y}_k^{1/L}|\mathbf{x}_k = \mathcal{M}_{k\leftarrow 0}(\mathbf{x}_0)). \end{aligned} \tag{38}$$

Elevating both members of the equation to the power L, one finds

$$\begin{aligned} p^L(\mathbf{x}_0|\mathbf{y}_L^{1/L}, \mathbf{y}_{L-1}^{2/L}, \ldots, \mathbf{y}_1, \mathbf{y}_0, \ldots) \\ \propto p^L(\mathbf{x}_0|\mathbf{y}_{L-1}^{1/L}, \mathbf{y}_{L-2}^{2/L}, \ldots, \mathbf{y}_1^{(L-1)/L}, \mathbf{y}_0, \ldots) \\ \times \prod_{k=1}^{L} p(\mathbf{y}_k|\mathbf{x}_k = \mathcal{M}_{k\leftarrow 0}(\mathbf{x}_0)). \end{aligned} \tag{39}$$

As for the forecast, one has

$$p(\mathbf{x}_1|\mathbf{y}_{:L}^{\boldsymbol{\alpha}_{:L}}) = \int \mathrm{d}\mathbf{x}_0\, p(\mathbf{x}_1|\mathbf{x}_0, \mathbf{y}_{:L}^{\boldsymbol{\alpha}_{:L}}) p(\mathbf{x}_0|\mathbf{y}_{:L}^{\boldsymbol{\alpha}_{:L}}). \tag{40}$$

Under perfect model assumptions and neglecting the impact of the Jacobian of the tangent linear of $\mathcal{M}_{1\leftarrow 0}$, which is assumed constant near $\mathbf{x}_0^\star$, one obtains

$$p^L(\mathbf{x}_1|\mathbf{y}_{:L}^{\boldsymbol{\alpha}_{:L}}) \simeq \int \mathrm{d}\mathbf{x}_0\, \delta\left[\mathbf{x}_1 - \mathcal{M}_{1\leftarrow 0}(\mathbf{x}_0)\right] p^L(\mathbf{x}_0|\mathbf{y}_{:L}^{\boldsymbol{\alpha}_{:L}}). \tag{41}$$

For the balancing step, one similarly obtains

$$\begin{aligned} p^L(\mathbf{x}_0|\mathbf{y}_{:L}) \propto p^L(\mathbf{x}_0|\mathbf{y}_L^{1/L}, \mathbf{y}_{L-1}^{2/L}, \ldots, \mathbf{y}_1, \mathbf{y}_0, \ldots) \\ \times \prod_{k=1}^{L} p(\mathbf{y}_k^{k-1}|\mathbf{x}_k = \mathcal{M}_{k\leftarrow 0}(\mathbf{x}_0)). \end{aligned} \tag{42}$$

This scheme with weights 1 is very similar to the uniform case with weights $1/L$. Note that, as specified by Eq. (42), in the cost function of the balancing step Eq. (35), $1 - \alpha_k$ should be replaced by $k - 1$. However, the main difference lies in the fact that the estimated pdf is the power L of the targeted pdf, rather than the pdf itself. This has no impact on the chaining of the cycles. It does not impact the forecast of the system state if it is chosen to be the maximum *a posteriori* of the system state pdf. However, to perform an ensemble forecast from the original posterior pdf, one needs first to rescale the posterior error covariance matrix by a factor L.

3.3. Linearized variant (Lin-IEnKS)

Here, we introduce a linearized variant of the IEnKS, denoted Lin-IEnKS. The cost function is quadratically expanded as

$$\begin{aligned} \widetilde{\mathcal{J}}(\mathbf{w}) \simeq \frac{1}{2}\sum_{k=1}^{L}(\boldsymbol{\delta}_k - \mathbf{Y}_k\mathbf{w})^{\mathrm{T}}\,\beta_k\mathbf{R}_k^{-1}\,(\boldsymbol{\delta}_k - \mathbf{Y}_k\mathbf{w}) \\ + \frac{1}{2}(N-1)\mathbf{w}^{\mathrm{T}}\mathbf{w}, \end{aligned} \tag{43}$$

where $\boldsymbol{\delta}_k = \mathbf{y}_k - H_k \circ \mathcal{M}_{k\leftarrow 0}(\mathbf{x}_0^{(0)})$ and $\mathbf{Y}_k = [H_k \circ \mathcal{M}_{k\leftarrow 0}]'_{|\mathbf{x}_0^{(0)}}\mathbf{A}_0$ is the tangent linear of the operator from ensemble space to observation space, computed at $\mathbf{x}_0^{(0)}$ using the propagation of an ensemble, as for the IEnKF and IEnKS. Obviously, the solution to the minimization is

$$\begin{aligned} \mathbf{w}^\star = \left[(N-1)\mathbf{I}_N + \sum_{k=1}^{L}\mathbf{Y}_k^{\mathrm{T}}\beta_k\mathbf{R}_k^{-1}\mathbf{Y}_k\right]^{-1} \\ \times \sum_{k=1}^{L}\mathbf{Y}_k^{\mathrm{T}}\beta_k\mathbf{R}_k^{-1}\boldsymbol{\delta}_k. \end{aligned} \tag{44}$$

This analysis part of the linearized variant, which could be seen as an inner loop of the IEnKS, resembles the EnVar analysis step of Liu *et al.* (2008). It merely amounts to performing the first iteration of the minimization of the IEnKS cost function.

3.4. General remarks with regard to other methods

The IEnKS performs the task of a 4D-Var method by minimizing a cost function, with the help of an ensemble used to simulate the adjoint of the tangent linear models of the dynamical and observation models. By construction, the IEnKS also belongs to the class of reduced-rank variational methods (Robert *et al.*, 2005; Hoteit and Köhl, 2006).

Further, the IEnKS (specifically in the case $S = 1$) is a sequential quasi-static method (Pires *et al.*, 1996). Indeed, the analysis focuses on the beginning of the DAW by optimizing the initial condition. At the start of the algorithm, the DAW is progressively extended until it reaches the predetermined fixed DAW length. Once the fixed DAW length is reached, the IEnKS becomes a fixed-lag smoother, acquiring a new batch of observations and discarding the oldest one. Moreover, Gaussian statistics of the errors are passed on from one cycle to the next. Hence, the initial condition $\mathbf{w}$ at the beginning of the window is progressively updated and should remain close to a nearly global minimum of the cost function.

Chen and Oliver (2012) have developed an iterative ensemble smoother (IEnS) targeted for oil-reservoir modelling. It is based on the ensemble smoother by van Leeuwen and Evensen (1996) and uses a global DAW. It does not suit the chaotic systems encountered in meteorology and oceanography, because in chaotic systems observations that are distant in time become uncorrelated with the current model state. Similarly to the EnRML (Gu and Oliver, 2007), the IEnS uses a stochastic framework and an explicit estimation of sensitivities (correlations) between model state and forecast observations from the ensemble.

3.5. The Gaussian smoother (EnKS)

Besides the new methods that we have introduced (SDA and MDA IEnKS, Lin-IEnKS), an already existing, non-iterative smoother based on assumptions of linearity and Gaussianity will also be tested and compared in section 4. In the following, we briefly recall the principles.

We will denote by EnKS the ensemble square-root variant of the non-iterative smoother discussed in Evensen and van Leeuwen (2000), Evensen (2003, 2009) and Cosme *et al.* (2012). It has been tested on the Lorenz '95 model by Khare *et al.* (2008), and on an ocean model by Cosme *et al.* (2010). In terms of ensemble runs, it is remarkably cheap as it does not require any more ensemble propagation than the underlying EnKF. As a trade-off, the retrospective analysis relies on assumptions of linearity and Gaussian statistics. Therefore, a comparison of the EnKS with the IEnKS is of great interest here, especially for long DAWs. We emphasize that the nowcasting and forecasting abilities of this smoother are those of the EnKF, since, by construction, the EnKS only updates the past analyses.

Our implementation of the EnKS is a straightforward adaptation of the algorithm described in Cosme *et al.* (2010). Building a finite-size EnKS that accounts for sampling errors is a simple task: to obtain an EnKS-N, the smoothing part of the EnKS is simply appended to the EnKF-N. Hence, the filtering performance of the EnKS-N is that of the EnKF-N. It is important to mention that, in the smoothing part of the algorithm, inflation or alternatively the finite-size scheme are not used. Indeed, the priors used in the smoothing part already result from a past analysis and therefore do not require additional inflation.

3.6. Numerical cost

In this article, we do not focus on optimizing the numerical cost of the smoothers. But counting the number of ensemble propagation required by each of the smoothers is an easy, albeit gross, estimation of the algorithm complexity. Remember that the length of the DAW is $L\Delta t$. We shall assume here that the number of iterations required to reach convergence in the optimization of the cost function is R. Let us count Ω, the number of ensemble propagations through each update time interval Δt required by a full cycle of the smoother. Note that the forecast through the whole DAW needed to compute the analysis at $t_2, \ldots, t_L$ is not part of the smoother cycle. This cost has to be added and depends on the kind of analysis/forecast that is needed and may include the balancing step.

The EnKS, just like the EnKF, requires $\Omega_{\text{EnKS}} = 1$. The Lin-IEnKS requires a full ensemble propagation through the window in order to compute the sensitivity. After obtaining the optimal $\mathbf{w}^\star$, it additionally requires the propagation of the final ensemble from t_0 to t_1. As a consequence, one has $\Omega_{\text{Lin-EnKS}} = L + 1$. The IEnKS requires R ensemble propagations through the window in order to compute the sensitivities and optimize the cost function. After obtaining the optimal $\mathbf{w}^\star$, it additionally requires the propagation of the final ensemble from t_0 to t_1. Therefore, one has $\Omega_{\text{IEnKS}} = R \times L + 1$.

To obtain an analysis at t_L (present time), one would additionally require the forecast of the analysis at t_0, which comes with an extra $\Omega = (L-1)/N$. In the MDA case, when the filtering performance matters, the balancing step might be much more costly, since a new minimization is required. To obtain an ensemble analysis at t_L (present time), one would additionally require the forecast of the analysis ensemble at t_0, which comes with an extra $\Omega = L - 1$.

4. Numerical tests and comparisons

In the following, we perform twin experiments with two low-order models, using the assumption of a perfect model (identical twin). The observations, extracted from the truth run, are perturbed using errors from a normal distribution. The analysis RMSE is the root mean square of the analysis minus the truth. The algorithmic parameters of the iterative smoothers are $\varepsilon = 10^{-4}$, $e = 10^{-3}$ and $j_{\max} = 50$. In the following experiments, the sensitivity of the performance of the smoothers to a moderate change in these parameters is very weak. In particular, the results barely depend on the value of ε when chosen in the range 10^{-6}–10^{-1}. We believe this is due to an ensemble spread σ that is already small enough that the growth of the rescaled ensemble anomalies of spread $\varepsilon\sigma$, even for an ε that is not too small, remains driven by the tangent linear model over the DAW.

4.1. The Lorenz '95 model

The Lorenz '95 model (Lorenz and Emmanuel, 1998) represents a mid-latitude zonal circle of the global atmosphere. It has $M = 40$ variables $\{x_m\}_{m=1,\ldots,M}$. Its dynamics is given by a set of ordinary differential equations:

$$\frac{\mathrm{d}x_m}{\mathrm{d}t} = (x_{m+1} - x_{m-2})x_{m-1} - x_m + F, \tag{45}$$

for $m = 1, \ldots, M$, where $F = 8$ and the domain is periodic. The dynamics is chaotic, with a doubling time of about 0.42 time units. A time step of $\Delta t = 0.05$ is meant to represent a time interval of 6 h in the real atmosphere. This low-order model is commonly used as a benchmark for new data assimilation methods. This model is integrated using a fourth-order Runge–Kutta scheme, with a time-step of 0.05.

The truth run is assumed to be fully observed ($d = 40$), so that $\mathbf{H} = \mathbf{I}_d$, with an observation-error covariance matrix $\mathbf{R} = \mathbf{I}_d$. The goal is to track the nature run. Unless otherwise stated, the size of the ensemble will be $N = 20$, which is greater than the size of the unstable subspace and, in the case of this model, makes localization unnecessary. Unless otherwise stated, the time interval between observational updates will be $\Delta t = 0.05$, meant to be representative of a data assimilation cycle of global meteorological models.

All of the data assimilation runs extend over 10^5 cycles, after a burn-in period of 5×10^3 cycles, which guarantees satisfying convergence of the statistics.

For the Lorenz '95 case, we have chosen to solve the inflation issue by using the finite-size counterparts of the filters/smoothers. For this model, and except in quasi-linear conditions ($\Delta t \sim 0.01$), it leads to performances quantitatively very close to the same filter/smoother with optimally tuned uniform inflation (Bocquet, 2011). Therefore, in the following, IEnKS/EnKS and other variants will interchangeably mean IEnKS-N/EnKS-N or IEnKS/EnKS with optimally tuned inflation.

4.2. Smoother performance as a function of lag

We first study the performance of the (finite-size variants of) EnKF, SDA IEnKS, SDA Lin-IEnKS and EnKS as a function of the lag for $\Delta t = 0.05$, corresponding to a mild nonlinearity. Here, the lag is defined as the time length in Δt that separates an estimate within the DAW from the end of the DAW. A DAW of $L = 20$ cycles is first considered. The RMSEs of the present (lag 0) and retrospective (lag > 0) analyses are reported in Figure 2 as a function of the lag.

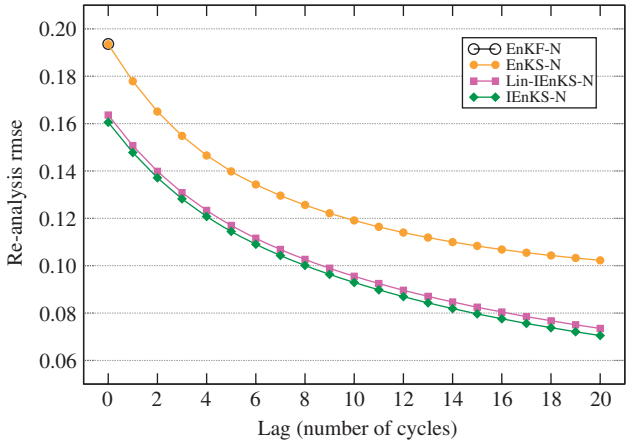

Figure 2. Comparison of the EnKF-N, SDA IEnKS-N, SDA Lin-IEnKS-N and EnKS-N in a mildly non-Gaussian system with a DAW of $L = 20$ cycles and $\Delta t = 0.05$, as a function of the distance from the end of the DAW.

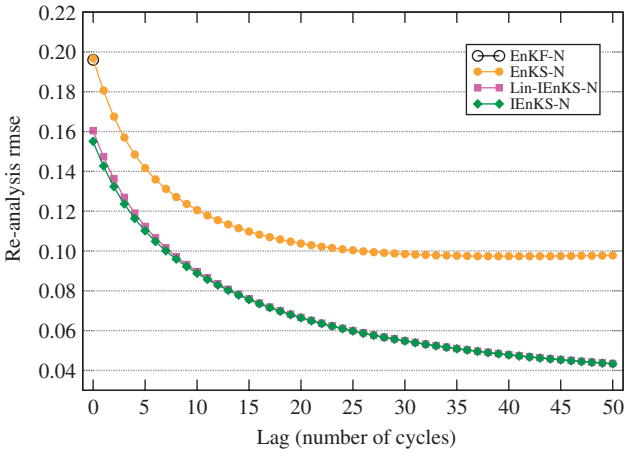

Figure 3. Comparison of the EnKF-N, MDA IEnKS-N, MDA Lin-IEnKS-N and EnKS-N in a mildly non-Gaussian system with a DAW of $L = 50$ cycles and $\Delta t = 0.05$, as a function of the distance from the end of the DAW.

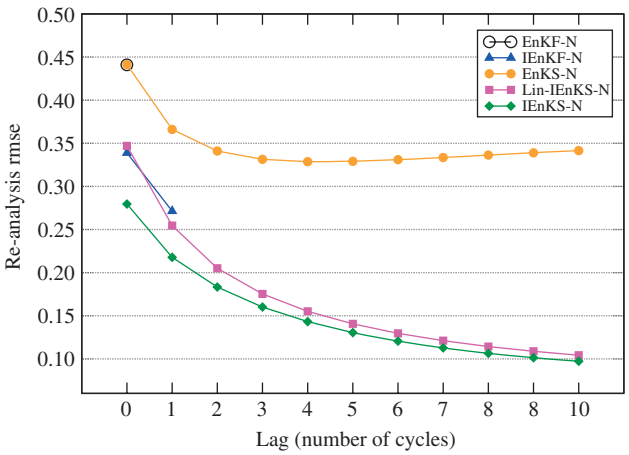

Figure 4. Comparison of the EnKF-N, MDA IEnKS-N, MDA Lin-IEnKS-N and EnKS-N in a significantly non-Gaussian system with a DAW of $L = 10$ cycles and $\Delta t = 0.20$, as a function of the distance from the end of the DAW.

At first, we have checked that the EnKF gives the same present-time analysis as the EnKS, which is merely a check of consistency. The filtering performance of the IEnKS is clearly better than that of the EnKF and the EnKS, in accordance with the marginal improvement diagnosed in Bocquet and Sakov (2012) for the bundle IEnKF but improving further. Logically, all smoothers achieve a better reanalysis along with the increase of lag.

As the lag increases, the advantage of the IEnKS over the EnKS becomes even larger. This is due to the fact that nonlinearity is handled by the EnKS within each interval Δt, whereas the IEnKS handles it within the full DAW of length $L\Delta t$. More precisely, the small nonlinearity in between two updates builds up over the full DAW of length $L\Delta t$, possibly leading to a significantly nonlinear forecast from t_0 to t_L. In contrast, by truncating the statistics to second order at each Δt, the EnKS partially misses this build-up. The linearized IEnKS, with only one iteration through the DAW, performs almost as well as the IEnKS. For $e = 10^{-3}$, the IEnKS requires here an average of $\Omega = 49$ ensemble propagations (in units of Δt), whereas the Lin-IEnKS requires by construction $\Omega = 21$. This means that, here, the threshold e could have been increased.

We have also performed tests with DAWs as long as $L = 50$ cycles. Although it is still stable, the performance of the SDA IEnKS degrades. It may be due to the very large time difference between the control $\mathbf{x}_0$ and the observation $\mathbf{y}_L$. That is why we introduced the MDA schemes for the IEnKS. Most of the MDA schemes we tested were stable up to $L = 80$. Here, we shall use the uniform scheme with equal weights for observation vectors within the DAW. Many other non-uniform schemes lead to very similar performances. In order to avoid tuning inflation, we choose the scheme with $\beta_k \equiv 1$, which is compatible with adaptive inflation schemes. However, it was checked that the scheme $\beta_k \equiv 1/L$ with a tuning of optimal inflation leads to similar RMSEs. The scores are reported in Figure 3.

Because we opted for an MDA scheme, we also systematically performed additional analyses (useless in the SDA case) using the balancing scheme described in section 3.2.3. With the MDA IEnKS, the retrospective analysis keeps improving down to values of the RMSE about 0.043. Instead of using a finite-size scheme that avoids the need to tune inflation, a careful tuning of the inflation at 1.01 could help achieve a slightly better RMSE of 0.039. Remarkably, the filtering RMSE keeps improving down to values of about 0.15, in spite of several numerical obstacles such as the precision of the gradient simulated by the finite-size ensemble and the unstable propagation of the model through several doubling times.

As opposed to the EnKS, the IEnKS does not reach clear saturation in its smoothing performance when L is increased and before numerical degradation in the filtering performance appears. Therefore, it is not clear to us whether the IEnKS smoothing performance should reach a positive theoretical lower bound with a chaotic dynamical model. Indeed, on the one hand, a longer DAW theoretically allows us to get closer to the truth. On the other hand, however, the ensemble needed to compute the approximate gradient needs to be propagated through the entire window with fewer and fewer stable integrations (due to positive Lyapunov spectrum). This could result in a compromise and a lower bound as L is increased. For a stable, possibly nonlinear model, we do not expect such a lower bound.

The linearized MDA IEnKS performs almost as well as the MDA IEnKS, even for a DAW length of $L = 50$. This is consistent with the average number of iterations required by the IEnKS (very close to 1). The fact that, most of the time, only one iteration is required by the IEnKS is consistent with the quasi-static view of the algorithm: the new observations at the present time lead only to a minor adjustment of the ensemble trajectory.

A quite different picture emerges when the observation update interval Δt is increased, so that the model's nonlinearities can develop in between, leading to increasingly non-Gaussian error statistics. Figure 4 is an illustration of this when $\Delta t = 0.20$ and $L = 10$, so that the DAW length is $L\Delta t = 2$, about five times the doubling time. In this case, the Lin-IEnKS becomes less efficient. It turns out that it diverges beyond $\Delta t = 0.30$. The EnKS is much less efficient since it is driven by the EnKF. The IEnKS still leads to an improvement over the IEnKF for the filtering analysis RMSE and remains very well-performing as a filter or smoother.

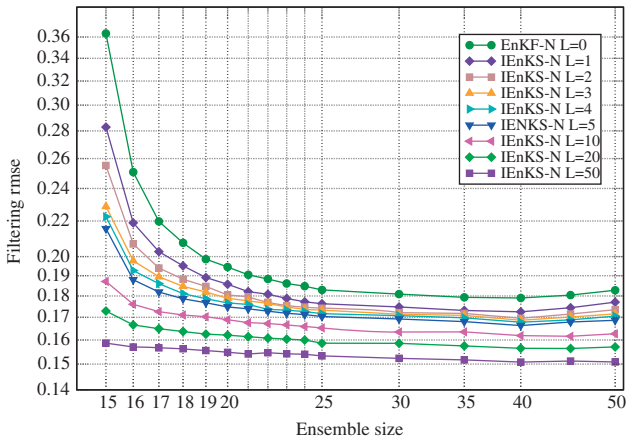

Figure 5. Filtering performance of the EnKF-N and MDA IEnKS-N when the ensemble size is varied, in a mildly non-Gaussian system with $\Delta t = 0.05$.

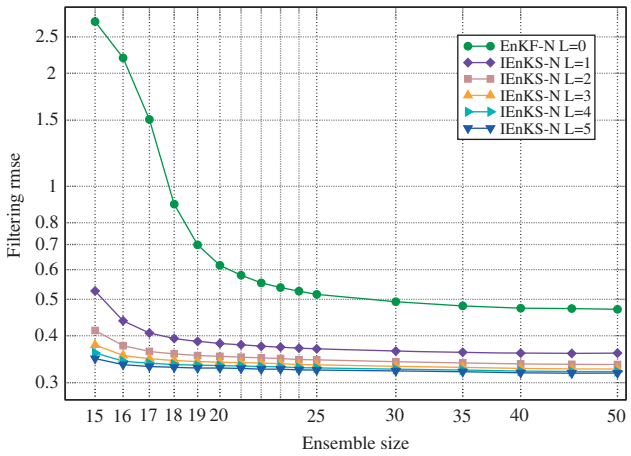

Figure 6. Filtering performance of the EnKF-N and MDA IEnKS-N when the ensemble size is varied, in a significantly non-Gaussian system with $\Delta t = 0.30$.

4.3. *Smoother performance as a function of ensemble size*

The ensemble size is varied from $N = 15$ to $N = 50$, using the (finite-size) EnKF and the MDA IEnKS with DAWs of length $L\Delta t$ with $L = 1, 2, 3, 4, 5, 10, 20, 50$ for $\Delta t = 0.05$ and with DAWs of length $L = 1, 2, 3, 4, 5$ and $\Delta t = 0.30$. The filtering analysis RMSEs are reported in Figure 5 for $\Delta t = 0.05$ and in Figure 6 for $\Delta t = 0.30$. In both cases, the gain of the smoother at small ensemble size, even slightly above the unstable subspace dimension, is very significant. It also allows us to see much more clearly the impact of multiple iterations on the filtering performance. As expected, the improvement is more striking when the nonlinearity is stronger ($\Delta t = 0.30$). At fixed ensemble size, the gain in RMSE is greater for small L, especially in nonlinear conditions. This gain is smaller with larger L. This type of result was first reported by Cohn *et al.* (1994) for a Kalman smoother (see their figure 7).

4.4. *Global view of the filtering and smoothing performance*

We have made extensive runs to compute the analysis RMSE at the end of the DAW (filtering performance) and the retrospective analysis RMSE at the beginning of the DAW (smoothing performance) for the SDA IEnKS, SDA Lin-IEnKS and EnKS. The length of the DAW in units of Δt is varied from $L = 1$ to $L = 50$ and Δt is also varied from $\Delta t = 0.05$ to $\Delta t = 0.60$ in steps of 0.05. The non-Gaussianity of the data assimilation system increases with Δt, since the system remains unobserved for longer and longer times, so that nonlinearities impact the system more and more. The use of finite-size versions of the smoothers allowed us to avoid tuning of the inflation and significantly reduced the computation load. The results are reported in Figure 7.

It is clear from the figure and expected from the past IEnKF results that the IEnKS can accommodate very large Δt and can very significantly outrank the EnKS. In the absence of iterations (outer loop in the language of 4D-Var), the Lin-IEnKS and EnKS cannot compete with the IEnKS when Δt becomes large.

For large $L\Delta t$, the performance of both IEnKS and Lin-IEnKS is impeded by numerical instabilities. In contrast, the EnKS can accommodate much larger DAW, even though there is no gain (and possibly a degradation) in the smoothing performance.

Since the IEnKS focuses on optimization of the initial conditions at t_0, it is unsurprising that the best smoothing performance is obtained for the largest L before a strong degradation in the performance is observed. Theoretically, the optimal filtering performance should also be obtained when L is the largest numerically stable value. This is indeed what is observed.

To assert the stability gain granted by MDA, the same study was also carried out with the MDA IEnKS and MDA Lin-IEnKS. Again the finite-size versions of the smoothers were employed. We used the scheme $\beta_k \equiv 1$ in order to use the adaptive inflation scheme. Additionally, the balancing procedure described in section 3.2.3 is implemented to compute the analysis within the DAW. The RMSEs are reported in Figure 8. Firstly, it appears that the algorithm accommodates longer DAWs. Secondly, because the stability extends over larger DAWs, the smoothing and even filtering performance are often better when compared with the SDA IEnKS.

4.5. *Quality of the ensemble forecast*

The quality of the ensemble forecast through the whole DAW is not the focus of this article. However, since nonlinearity is well accounted for within the DAW, we wanted to check the potential of such a forecast ensemble.

We have considered the finite-size SDA IEnKS with $L = 20$ and the finite-size MDA IEnKS with $L = 50$. For each cycle, the posterior ensemble at t_0 has been forecast to t_L. The rank histogram of the ensemble (Anderson, 1996; Talagrand *et al.*, 1997; Hamill, 2000) has been computed in the two cases and reported in Figure 9. In the MDA case, following section 3.2.5, the posterior ensemble has been inflated by a factor $\sqrt{L}$ before forecasting. In both IEnKS cases the histograms are quite flat, with mild deformations at the edges. The approximate Gaussian posterior from which the ensemble is extracted at t_0 is an excellent *local* approximation of the pdf. By construction, it is meant to remain so throughout the whole DAW. We have also performed the same experiments but with the EnKS. In both cases, the ensemble was severely underdispersed (U-shape). This was to be expected, since the true ensemble forecast within the DAW as seen by the EnKS consists of successive ensemble forecasts of Δt followed by Gaussian resamplings, which adds to the uncertainty. This is not accounted for in the spread of the forecasted ensemble. This explains why the forecast ensembles are underdispersed. This has been checked by comparing the RMSEs and the ensemble spreads.

4.6. *Two-dimensional turbulence*

The smoothers have also been tested in the framework of a forced two-dimensional turbulence model. The selected model was very similar to the barotropic model used by van Leeuwen (2013). In particular, we choose $\Delta t = 2$, which leads to a significant temporal decorrelation of 0.82 and significant nonlinearity. The vorticity field was discretized on a grid of 64×64. Within a twin experiment, the vorticity field was fully observed. The ensemble size was chosen to be 40 in order to avoid rank deficiency even without localization. We computed RMSEs as a function of the lag, with qualitative results very similar to those of the Lorenz '95

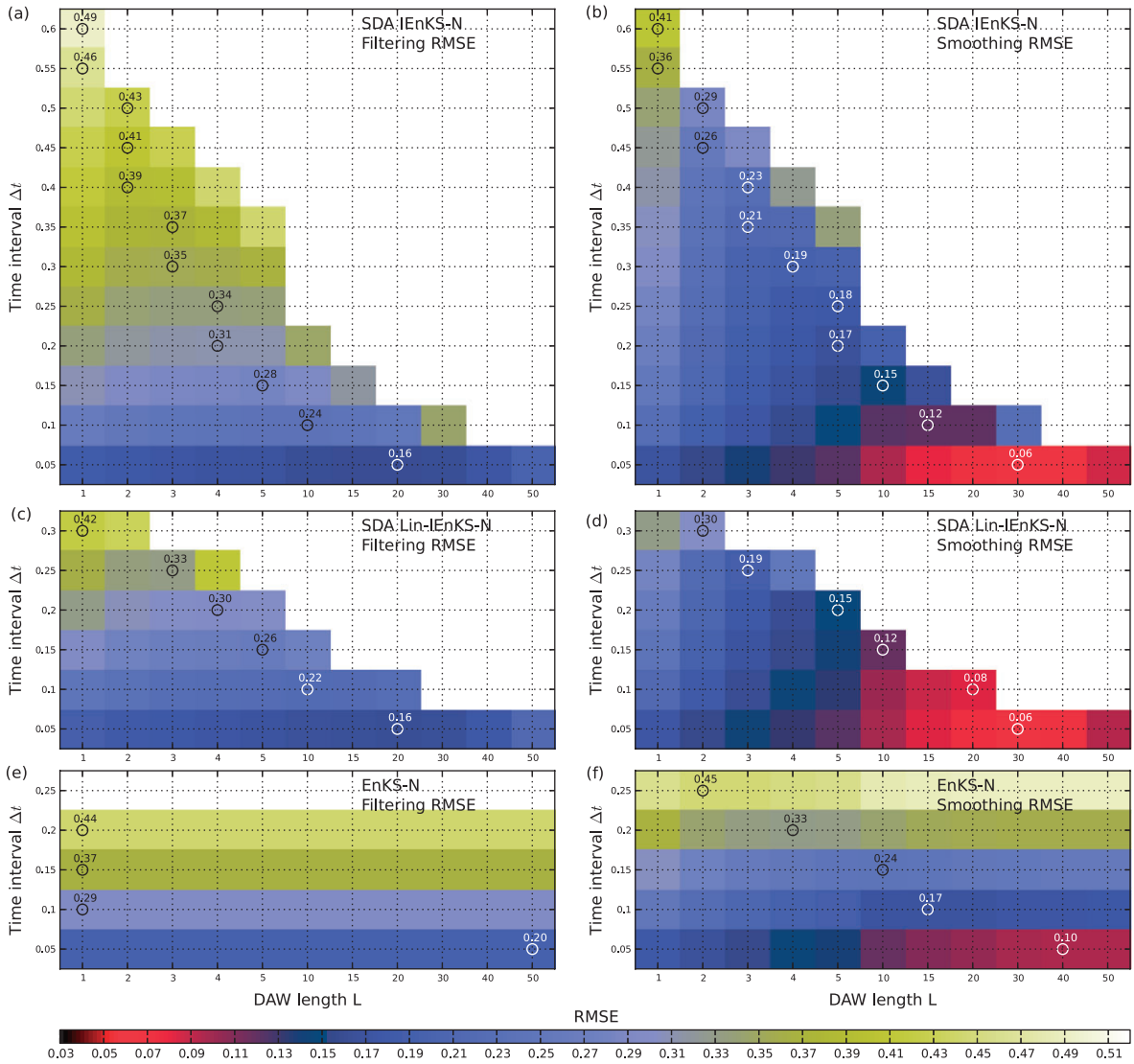

Figure 7. Comparison of filtering and smoothing performance of (a, b) the SDA IEnKS-N, (c, d) the SDA Lin-IEnKS-N and (e, f) the EnKS-N, when the time interval between observations Δt and the DAW length L are varied. The circle indicates the best analysis or reanalysis RMSE for each row (at fixed Δt). The value of minimum RMSE is displayed near each circle. White pixels correspond to RMSEs greater than 0.52.

model experiments, for DAWs as long as $L = 50$. We found the following:

- a saturation of the EnKS performance for lags longer than 15;
- a clear superiority of the IEnKS with an RMSE three times smaller for retrospective analysis (states with a lag of $50\Delta t$);
- a linearized IEnKS performing almost as well as the IEnKS.

Two differences were the need to use the MDA IEnKS as soon as the DAW reaches $L = 10$ and the numerical difficulties encountered in performing the balancing step for the longest tested DAWs.

5. Summary and further discussion

In this article, we have extended the iterative ensemble Kalman filter (IEnKF) to the iterative ensemble Kalman smoother (IEnKS). If the IEnKF is understood as a lag-one smoother then the IEnKS represents a fixed-lag smoother of length L. It is a four-dimensional ensemble variational method. Even though it seeks to minimize a cost function iteratively over the fixed-lag temporal window, it does not require tangent-linear models nor adjoint models of the evolution and observation, because the ensemble is used to simulate them instead. Seeking performance, we have chosen to study a variant where the DAW is shifted by only one observation time interval per cycle. In this case, it also qualifies as a quasi-static method. As a reduced-rank variational method, its cost function can be minimized using many different optimization schemes. In this article, we chose the Gauss–Newton approach because of its simplicity and because the test cases chosen here seldom required more advanced minimization schemes.

With such minimal shifts between cycles, the system must deal with overlapping data assimilation windows. In the simplest implementation (single data assimilation: SDA), the latest observation set entering the window is assimilated once and for all. In the multiple data assimilation (MDA) implementation, the observation sets are assimilated several times according to a flexible schedule.

The smoother has been tested on the Lorenz '95 model and on a two-dimensional forced turbulence model. We have studied the filtering performance of the method, i.e. the estimation of the present state (analysis), as well as its smoothing performance, i.e. the estimation of past states (retrospective analysis).

In a mildly nonlinear regime, for instance when the interval between two observations batches is sufficiently short (typically $\Delta t = 0.05$ in the Lorenz '95 model), the filtering performance of the IEnKS is marginally but systematically better than that of the EnKF, at a higher computational cost. The improvement is larger with larger L. It can be very significant for small ensemble size but there seems to be an asymptotic limit to the improvement, which may be due to the fundamental Gaussian hypothesis used to generate the ensembles and explains why we kept 'Kalman' in the smoother's name.

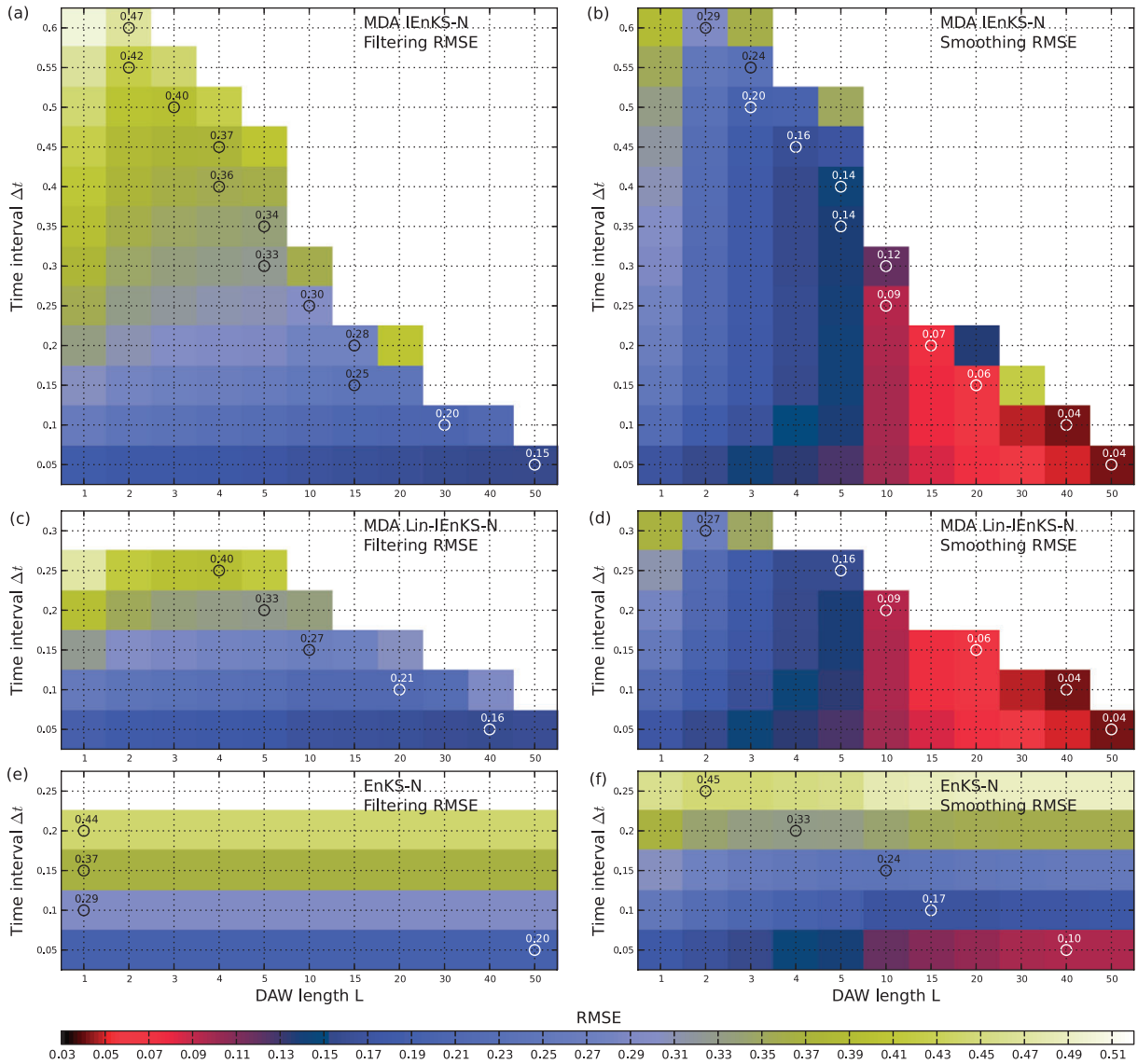

Figure 8. Comparison of filtering and smoothing performance of (a,b) a typical MDA IEnKS-N, (c,d) the MDA Lin-IEnKS-N and (e,f) the EnKS-N, when the time interval between observations Δt and the DAW length L are varied. The circle indicates the best analysis or reanalysis RMSE for each row (at fixed Δt). The value of minimum RMSE is displayed near each circle. White pixels correspond to RMSEs greater than 0.52.

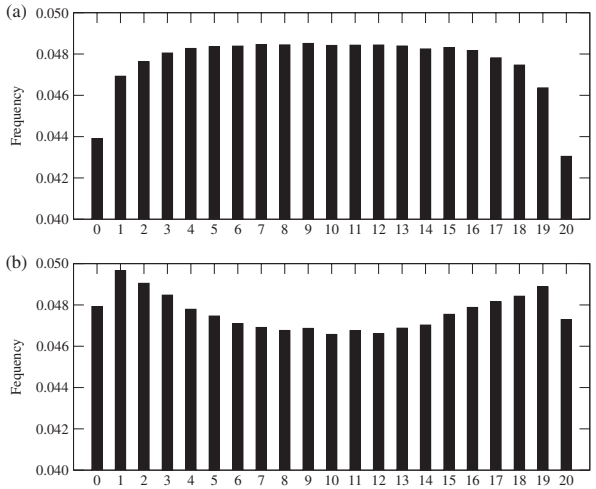

Figure 9. Average rank histograms for the 20 member ensemble at t_L forecast from the ensemble generated at t_0, for (a) $L = 20$/SDA/bundle IEnKS-N and (b) $L = 50$/MDA/bundle IEnKS-N.

In the same regime, the smoothing performance of the IEnKS is compared with a standard and numerically efficient non-iterative ensemble Kalman smoother (EnKS). When the lag is more than a few times the doubling time, the IEnKS very significantly outranks the EnKS. It is difficult from our results to decide whether there is a lower limit to the smoothing root-mean-square error (RMSE) of the IEnKS or not. This contrasts with the EnKS, which shows a lower bound for the smoothing RMSE.

In a strongly nonlinear regime, for instance when the time interval between two observation batches is sufficiently long (typically $\Delta t \gg 0.05$ in the Lorenz '95 model), the filtering performance of the the IEnKS is much better than that of the EnKF.

The linearized version of the MDA IEnKS, which has similarities with the EnVar approaches currently discussed in (operational) meteorology, corresponds to the IEnKS with a single iteration or its inner loop. In mildly nonlinear conditions (typical of synoptic meteorology) its performance is almost as good as that of the full IEnKS. This shows that nonlinearities do not have a strong impact on the data assimilation system and concurs with the empirical equivalence of the EnKF and 4D-Var in a synoptic meteorological context (Kalnay *et al.*, 2007). However, in strongly nonlinear conditions the Lin-IEnKS is inoperative, as more iterations are definitely needed to track the truth. It is very likely that the use of multiple iterations will be needed in challenging meteorological data assimilation systems operating at convective scales or in atmospheric chemistry data assimilation systems, as well as geophysical data assimilation systems relying on parameter estimation with a nonlinear relationship to the observation.

We assumed no temporal correlations of observation-error vectors. In principle, they can be accounted for in the MDA schemes. However, this seems at variance with the choice of independence made when splitting observations. That is why such generalization would require further investigation of the MDA schemes.

Localization of the iterative schemes introduced in this article is the next step of their development, so that they can be applied to realistic high-dimensional systems. It deserves special attention. Localization of an EnKF with a local observation operator H is known to be efficient (Hamill *et al.*, 2001). However, in the context of the iterative schemes, $H \circ \mathcal{M}$ plays the role of H, as can be seen for instance in Eq. (24). It is a non-local map. Unfortunately, the standard covariance localization and the model flow do not commute. Consequently, the less local this map is, the less efficient standard covariance localization techniques are expected to be. This also pertains to the delicate use of localization in the context of non-local H (such as column sounding).

This challenge has a counterpart when using local analysis, also known as domain localization (Sakov and Bertino, 2011; Nerger *et al.*, 2011). While local analysis is characterized in the literature as scheme-independent (Sakov and Bertino, 2011), this is less obvious in the context of iterative schemes. Specifically, iterative schemes involve multiple propagation of the model during the minimization. Unlike the update, the propagation has to remain global in most applications. This may be expected to result in the numerical cost of a scheme being defined by the most slowly converging local region (which may be quite low: for example, just one iteration is required for the linearized version of the IEnKS).

The algorithms developed in this study are the basis for an extension to local algorithms. The results obtained in this article in the rank-sufficient regime for Lorenz '95 and two-dimensional turbulence, i.e. without localization but with a sufficiently large ensemble, will stand as a valuable baseline (for different values of Δt and L) by which to estimate the performance of a small-ensemble local variant of the IEnKS for these models.

One intriguing result of this study is that the MDA algorithm proved more robust than the SDA algorithm with very large DAWs (up to 80×0.05 with the Lorenz '95 model, i.e. an equivalent of 20 days), without resorting to a weak-constraint variational formalism. So far, we have not managed to justify this behaviour mathematically. We believe that the difference between the two schemes does not stem from numerical imprecision. Indeed, in the Lorenz '95 case, a DAW of 50×0.05 corresponds to a time length of about six times the doubling time and a mild error growth of $2^6 = 64$. In addition, we checked that perturbations of the gradients computed from the ensemble do not significantly degrade the overall performance. However, we risk offering another explanation. As put forward in the study, we believe that the excellent smoothing performance of the IEnKS is partly due to its quasi-static nature. Both SDA and MDA schemes are quasi-static, in the sense that they are based on sliding the DAW incrementally. However, in the MDA scheme one vector of observations truly enters the cost function and one leaves it, so that the cost function is smoothly deformed and the global minimum can be tracked. With the SDA scheme, most of the observations retained in the MDA cost function are not explicitly accounted for but are taken into account in the Gaussian prior. As a result, the cost function varies substantially and it becomes difficult to track the global minimum, especially when L becomes large and the cost function becomes rougher due to nonlinearity of the model over the whole DAW. In this case, we have observed an oscillating behaviour of the RMSE.

Finally, since computational efficiency was not our main concern, we did not explore the IEnKS where the shift of the data assimilation window is of several time intervals between observation updates ($2 \leq S \leq L$), leading to more computationally efficient, but perhaps more poorly performing, less non-quasi-static schemes. This should be studied.

Acknowledgement

The authors are grateful to two anonymous reviewers and Martin Leutbecher, acting as Associate Editor, for their useful suggestions that helped to improve the manuscript.

References

Anderson JL. 1996. A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *J. Climate* **9**: 1518–1530.

Anderson JL. 2007. An adaptive covariance inflation error correction algorithm for ensemble filters. *Tellus A* **59**: 210–224.

Anderson JL, Anderson SL. 1999. A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. *Mon. Weather Rev.* **127**: 2741–2758.

Andersson E, Fisher M, Hólm E, Isaksen L, Radnóti G, Trémolet Y. 2005.. '*Will the 4D-Var approach be defeated by nonlinearity?*' Technical Report 479. ECMWF: Reading, UK.

Bocquet M. 2011. Ensemble Kalman filtering without the intrinsic need for inflation. *Nonlin. Processes Geophys.* **18**: 735–750.

Bocquet M, Sakov P. 2012. Combining inflation-free and iterative ensemble Kalman filters for strongly nonlinear systems. *Nonlin. Processes Geophys.* **19**: 383–399. DOI: 10.5194/npg-19-383-2012.

Bocquet M, Pires CA, Wu L. 2010. Beyond Gaussian statistical modeling in geophysical data assimilation. *Mon. Weather Rev.* **138**: 2997–3023.

Chen Y, Oliver DS. 2012. Ensemble randomized maximum likelihood method as an iterative ensemble smoother. *Math. Geosci.* **44**: 1–26.

Cohn SE, Sivakumaran NS, Todling R. 1994. A fixed-lag Kalman smoother for retrospective data assimilation. *Mon. Weather Rev.* **122**: 2838–2867.

Cosme E, Brankart JM, Verron J, Brasseur P, Krysta M. 2010. Implementation of a reduced-rank, square-root smoother for ocean data assimilation. *Ocean Modelling* **33**: 87–100.

Cosme E, Verron J, Brasseur P, Blum J, Auroux D. 2012. Smoothing problems in a Bayesian framework and their linear Gaussian solutions. *Mon. Weather Rev.* **140**: 683–695.

Emerick AA, Reynolds AC. 2013. Ensemble smoother with multiple data assimilation. *Comput. Geosci.* **55**: 3–15.

Evensen G. 2003. The ensemble Kalman filter: theoretical formulation and practical implementation. *Ocean Dyn.* **53**: 343–367.

Evensen G. 2009. *Data Assimilation: The Ensemble Kalman Filter* (2nd edn). Springer-Verlag: Berlin, Heildelberg.

Evensen G, van Leeuwen PJ. 2000. An ensemble Kalman smoother for nonlinear dynamics. *Mon. Weather Rev.* **128**: 1852–1867.

Fisher M, Leutbecher M, Kelly GA. 2005. On the equivalence between Kalman smoothing and weak-constraint four-dimensional variational data assimilation. *Q. J. R. Meteorol. Soc.* **131**: 3235–3246.

Gu Y, Oliver DS. 2007. An iterative ensemble Kalman filter for multiphase fluid flow data assimilation. *SPE J.* **12**: 438–446.

Hamill TM. 2000. Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Weather Rev.* **129**: 550–560.

Hamill TM, Whitaker JS, Snyder C. 2001. Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter. *Mon. Weather Rev.* **129**: 2776–2790.

Hoteit I, Köhl A. 2006. Efficiency of reduced-order, time-dependent adjoint data assimilation approaches. *J. Oceanogr.* **62**: 539–550.

Houtekamer PL, Mitchell HL. 2005. Ensemble Kalman filtering. *Q. J. R. Meteorol. Soc.* **131**: 3269–3289.

Hunt BR, Kostelich EJ, Szunyogh I. 2007. Efficient data assimilation for spatiotemporal chaos: a local ensemble transform Kalman filter. *Physica D* **230**: 112–126.

Jazwinski AH. 1970. *Stochastic Processes and Filtering Theory*. Academic Press: New York, NY.

Kalnay E, Li H, Miyoshi T, Yang SC, Ballabrera-Poy J. 2007. 4D-Var or ensemble Kalman filter? *Tellus A* **59A**: 758–773.

Khare SP, Anderson JL, Hoar TJ, Nychka D. 2008. An investigation into the application of an ensemble Kalman smoother to high-dimensional geophysical systems. *Tellus A* **60**: 97–112.

Levenberg K. 1944. A method for the solution of certain problems in least squares. *Q. Appl. Math.* **2**: 164–168.

Liang X, Zheng X, Zhang S, Wu G, Dai Y, Li Y. 2012. Maximum likelihood estimation of inflation factors on error covariance matrices for ensemble Kalman filter assimilation. *Q. J. R. Meteorol. Soc.* **138**: 263–273.

Liu C, Xiao Q, Wang B. 2008. An ensemble-based four-dimensional variational data assimilation scheme. Part I: technical formulation and preliminary test. *Mon. Weather Rev.* **132**: 3363–3373.

Lorenz EN, Emmanuel KE. 1998. Optimal sites for supplementary weather observations: simulation with a small model. *J. Atmos. Sci.* **55**: 399–414.

Marquardt D. 1963. An algorithm for least-squares estimation of nonlinear parameters. *SIAM J. Appl. Math.* **11**: 431–441.

Miyoshi T. 2011. The Gaussian approach to adaptive covariance inflation and its implementation with the local ensemble transform Kalman filter. *Mon. Weather Rev.* **139**: 1519–1535.

Nerger L, Janjić T, Schröter J, Hiller W. 2011. A regulated localization scheme for ensemble-based Kalman filters. *Q. J. R. Meteorol. Soc.* **138**: 802–812.

Pires C, Vautard R, Talagrand O. 1996. On extending the limits of variational assimilation in nonlinear chaotic systems. *Tellus A* **48**: 96–121.

Robert C, Durbiano S, Blayo E, Verron J, Blum J, Le Dimet FX. 2005. A reduced-order strategy for 4D-Var data assimilation. *J. Marine Syst.* **57**: 70–82.

Sakov P, Bertino L. 2011. Relation between two common localization methods for the EnKF. *Comput. Geosci.* **15**: 225–237.

Sakov P, Oliver D, Bertino L. 2012. An iterative EnKF for strongly nonlinear systems. *Mon. Weather Rev.* **140**: 1988–2004.

Talagrand O, Vautard R, Strauss B. 1997. 'Evaluation of probabilistic prediction systems'. In *ECMWF Workshop on Predictability* 1–25. ECMWF: Reading, UK.

van Leeuwen PJ. 2009. Particle filtering in geophysical systems. *Mon. Weather Rev.* **137**: 4089–4114.

van Leeuwen PJ. 2013. Efficient fully nonlinear data assimilation for geophysical fluid dynamics. *Comput. Geosci.* **55**: 16–27.

van Leeuwen PJ, Evensen G. 1996. Data assimilation and inverse methods in terms of a probabilistic formulation. *Mon. Weather Rev.* **124**: 2898–2913.

Yang SC, Kalnay E, Hunt B. 2012. Handling nonlinearity in an ensemble Kalman filter: experiments with the three-variable Lorenz model. *Mon. Weather Rev.* **140**: 2628–2646.

Zhang F, Snyder C, Sun J. 2004. Impacts of initial estimate and observation availability on convective-scale data assimilation with an ensemble Kalman filter. *Mon. Weather Rev.* **134**: 1238–1253.

