Combining inflation-free and iterative ensemble Kalman filters for strongly nonlinear systems

Marc Bocquet, Pavel Sakov

Nonlin. Processes Geophys., 19, 383–399, 2012
www.nonlin-processes-geophys.net/19/383/2012/
doi:10.5194/npg-19-383-2012

Nonlinear Processes in Geophysics

Combining inflation-free and iterative ensemble Kalman filters for strongly nonlinear systems

M. Bocquet[1,2] **and P. Sakov**[3]

[1]Université Paris-Est, CEREA joint laboratory École des Ponts ParisTech and EDF R&D, France
[2]INRIA, Paris Rocquencourt research center, France
[3]Nansen Environment and Remote Sensing Center, Bergen, Norway

Correspondence to: M. Bocquet (bocquet@cerea.enpc.fr)

Received: 22 January 2012 – Revised: 30 April 2012 – Accepted: 22 May 2012 – Published: 25 June 2012

Abstract. The finite-size ensemble Kalman filter (EnKF-N) is an ensemble Kalman filter (EnKF) which, in perfect model condition, does not require inflation because it partially accounts for the ensemble sampling errors. For the Lorenz '63 and '95 toy-models, it was so far shown to perform as well or better than the EnKF with an optimally tuned inflation. The iterative ensemble Kalman filter (IEnKF) is an EnKF which was shown to perform much better than the EnKF in strongly nonlinear conditions, such as with the Lorenz '63 and '95 models, at the cost of iteratively updating the trajectories of the ensemble members. This article aims at further exploring the two filters and at combining both into an EnKF that does not require inflation in perfect model condition, and which is as efficient as the IEnKF in very nonlinear conditions.

In this study, EnKF-N is first introduced and a new implementation is developed. It decomposes EnKF-N into a cheap two-step algorithm that amounts to computing an optimal inflation factor. This offers a justification of the use of the inflation technique in the traditional EnKF and why it can often be efficient. Secondly, the IEnKF is introduced following a new implementation based on the Levenberg-Marquardt optimisation algorithm. Then, the two approaches are combined to obtain the finite-size iterative ensemble Kalman filter (IEnKF-N). Several numerical experiments are performed on IEnKF-N with the Lorenz '95 model. These experiments demonstrate its numerical efficiency as well as its performance that offer, at least, the best of both filters. We have also selected a demanding case based on the Lorenz '63 model that points to ways to improve the finite-size ensemble Kalman filters. Eventually, IEnKF-N could be seen as the first brick of an efficient ensemble Kalman smoother for strongly nonlinear systems.

1 Introduction

Let us first introduce two recently developed and complementary ensemble Kalman filters.

1.1 An ensemble Kalman filter without the intrinsic need for inflation

The finite-size ensemble Kalman filter (EnKF-N), introduced in Bocquet (2011), relies on the statistical modelling assumption that the ensemble used in the analysis is a collection of samples of the prior probability density function $p(\mathbf{x}|\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$, where $\mathbf{x}_k$ is the k-th member of the N-member forecast ensemble. The idea behind EnKF-N is that the empirical moments of the ensemble,

$$\overline{\mathbf{x}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_k\,, \qquad \mathbf{P} = \frac{1}{N-1}\sum_{k=1}^{N}(\mathbf{x}_k-\overline{\mathbf{x}})\,(\mathbf{x}_k-\overline{\mathbf{x}})^{\mathrm{T}}\,, \quad (1)$$

do not necessarily match the mean $\mathbf{x}_b$ and the error covariance matrix $\mathbf{B}$ of the prior probability density function (pdf). In the large ensemble size limit $N \to \infty$, we expect that they coincide. But for any finite N, sampling errors may induce a discrepancy.

An effective form for the prior pdf was proposed. It is the result of an integration over all possible $\mathbf{x}_b$ and $\mathbf{B}$. The effective prior pdf which is advocated to be used in the analysis step of the ensemble Kalman filter is:

$$p(\mathbf{x}|\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N) \propto \left|(\mathbf{x}-\overline{\mathbf{x}})\,(\mathbf{x}-\overline{\mathbf{x}})^{\mathrm{T}} + \varepsilon_N(N-1)\mathbf{P}\right|^{-\frac{N}{2}}\,, \quad (2)$$

rather than the Gaussian prior

$$p(\mathbf{x}|\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N) \propto \exp\left\{-\frac{1}{2}(\mathbf{x}-\overline{\mathbf{x}})^{\mathrm{T}}\mathbf{P}^{-1}(\mathbf{x}-\overline{\mathbf{x}})\right\} \quad (3)$$

which is implicitly assumed in traditional EnKF. Notation $|\mathbf{X}|$ designates the determinant of matrix $\mathbf{X}$. Note that the determinant and the inverse operators in formula Eqs. (2) and (3) are meant to operate in the vector space spanned by the anomalies $\{\mathbf{x}_k-\overline{\mathbf{x}}\}_{k=1,\ldots,N}$. Otherwise they would often be zero for the determinant and undefined for the inverse. The constant ε_N depends on the assumptions made to derive the prior. Two classes of filter that depend on these assumptions have been introduced. For the first, both $\mathbf{x}_b$ and $\mathbf{P}_b$ are uncertain, and $\varepsilon_N = 1+\frac{1}{N}$. For the second, it is assumed that the empirical mean $\overline{\mathbf{x}}$ is a fine approximation of $\mathbf{x}_b$. In this case: $\varepsilon_N = 1$.

The analysis step of EnKF-N follows from Bayes rule:

$$p(\mathbf{x}|\mathbf{y},\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N) \propto p(\mathbf{y}|\mathbf{x})\,p(\mathbf{x}|\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N), \quad (4)$$

where $\mathbf{y}\in\mathbb{R}^d$ is the observation vector at a given update step. It is assumed that the observational likelihood is Gaussian:

$$p(\mathbf{y}|\mathbf{x}) \propto \exp\left\{-\frac{1}{2}(\mathbf{y}-H(\mathbf{x}))^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{y}-H(\mathbf{x}))\right\}, \quad (5)$$

where H is the observation operator and $\mathbf{R}$ is the observation error covariance matrix.

In Bocquet (2011), the filter was seen as a deterministic filter and, in particular, the focus was on its ensemble transform variant. It is assumed that the system state $\mathbf{x}$ we would like to estimate, can be decomposed into

$$\mathbf{x} = \overline{\mathbf{x}} + \mathbf{A}\mathbf{w}, \quad (6)$$

where $\mathbf{A} = [\mathbf{x}_1-\overline{\mathbf{x}},\ldots,\mathbf{x}_N-\overline{\mathbf{x}}]$ is the matrix of the anomalies. The weights $\mathbf{w}\in\mathbb{R}^N$ are the (redundant) coordinates of $\mathbf{x}$ in ensemble space.

The filter follows the same algorithm as the ensemble transform Kalman filter of Hunt et al. (2007), except that the optimal vector of weight $\mathbf{w}_a$ at the analysis step is obtained from the minimisation of the cost function

$$\widetilde{\mathcal{J}}(\mathbf{w}) = \frac{1}{2}(\mathbf{y}-H(\overline{\mathbf{x}}+\mathbf{A}\mathbf{w}))^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{y}-H(\overline{\mathbf{x}}+\mathbf{A}\mathbf{w})) + \frac{N}{2}\ln\left(\varepsilon_N+\mathbf{w}^{\mathrm{T}}\mathbf{w}\right), \quad (7)$$

where H is the observation operator and $\mathbf{R}$ is the observation error covariance matrix. The tilde symbol indicates that the function is defined in ensemble space. This variational analysis step has similarities with that of Zupanski (2005); Harlim and Hunt (2007).

The other modification is the computation of the posterior error covariance matrix, which, instead of being based on the Hessian in ensemble space

$$\widetilde{\mathcal{H}} = (\mathbf{H}\mathbf{A})^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}\mathbf{A} + (N-1)\mathbf{I}_N, \quad (8)$$

where $\mathbf{H}$ is the Jacobian matrix of H as is done in the traditional scheme, is based on the Hessian of $\widetilde{\mathcal{J}}$ in Eq. (7)

$$\widetilde{\mathcal{H}} = (\mathbf{H}\mathbf{A})^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}\mathbf{A} + N\frac{(\varepsilon_N+\mathbf{w}^{\mathrm{T}}\mathbf{w})\,\mathbf{I}_N - 2\mathbf{w}\mathbf{w}^{\mathrm{T}}}{(\varepsilon_N+\mathbf{w}^{\mathrm{T}}\mathbf{w})^2}, \quad (9)$$

where $\mathbf{I}_N$ is the identity matrix in ensemble space. The complete algorithm is recalled in algorithm 1. In this algorithm, $\mathbf{U}$ is an arbitrary orthogonal matrix in ensemble space that preserves the ensemble mean (Sakov and Oke, 2008).

Algorithm 1 Finite-size ensemble Kalman filter.

Require: The forecast ensemble $\{\mathbf{x}_k\}_{k=1,\ldots,N}$, the observations $\mathbf{y}$ and error covariance matrix $\mathbf{R}$

1: Compute the mean $\overline{\mathbf{x}}$ and the anomalies $\mathbf{A}$ from $\{\mathbf{x}_k\}_{k=1,\ldots,N}$.
2: Compute $\mathbf{Y}=\mathbf{H}\mathbf{A}$, $\boldsymbol{\delta}=\mathbf{y}-\mathbf{H}\overline{\mathbf{x}}$
3: Find the minimum:

$$\mathbf{w}_a = \min_{\mathbf{w}}\left\{(\boldsymbol{\delta}-\mathbf{Y}\mathbf{w})^{\mathrm{T}}\mathbf{R}^{-1}(\boldsymbol{\delta}-\mathbf{Y}\mathbf{w}) + N\ln\left(\varepsilon_N+\mathbf{w}^{\mathrm{T}}\mathbf{w}\right)\right\}$$

4: Compute $\Omega_a = \left(\mathbf{Y}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{Y} + N\frac{(\varepsilon_N+\mathbf{w}_a^{\mathrm{T}}\mathbf{w}_a)\mathbf{I}_N - 2\mathbf{w}_a\mathbf{w}_a^{\mathrm{T}}}{(\varepsilon_N+\mathbf{w}_a^{\mathrm{T}}\mathbf{w}_a)^2}\right)^{-1}$
5: Compute $\mathbf{x}^a = \overline{\mathbf{x}} + \mathbf{A}\mathbf{w}_a$.
6: Compute $\mathbf{W}^a = ((N-1)\Omega_a)^{1/2}\mathbf{U}$
7: Compute $\mathbf{x}_k^a = \mathbf{x}^a + \mathbf{A}\mathbf{W}_k^a$

The filter was, in particular, tested on the Lorenz '95 toy-model (Lorenz and Emmanuel, 1998), using the root-mean-square error of the analysis as a criterion. In this context, EnKF-N was used without inflation which is usually required to stabilise or optimise the performance of such system (Anderson and Anderson, 1999). As a consequence the burden of tuning inflation was avoided. Yet, EnKF-N ($\varepsilon_N=1$ type) systematically levelled or slightly outperformed EnKF with optimally tuned inflation, for a large range of time intervals between updates. The extra numeral cost of using EnKF-N instead of EnKF was deemed marginal, especially for high-dimensional systems. This statement will be confirmed and clarified in the present article.

1.2 An ensemble Kalman filter for strongly nonlinear systems

The extended Kalman filter propagates the error covariance matrix from time t_1 to time t_2, based on a tangent linear model that has been computed at the previous analysis step around the analysis $\mathbf{x}_1^{(0)}$. However, when new observations $\mathbf{y}_2$ are assimilated at time t_2, a new estimation of the state at time t_1 conditional on the future observations $\mathbf{y}_2$ can be obtained. For strongly nonlinear systems and large update

intervals, this new estimation may significantly differ from $\mathbf{x}_1^{(0)}$, and the tangent linear model should be corrected and obtained at this new state. One would usually solve for the optimal analysed state at time t_2 by minimizing

$$\mathcal{J}_2(\mathbf{x}_2) = \frac{1}{2}\left(\mathbf{y}_2 - H(\mathbf{x}_2)\right)^{\mathrm{T}}\mathbf{R}^{-1}\left(\mathbf{y}_2 - H(\mathbf{x}_2)\right) + \frac{1}{2}\left(\mathbf{x}_2 - \mathbf{x}_2^f\right)\mathbf{P}_2^{-1}\left(\mathbf{x}_2 - \mathbf{x}_2^f\right), \quad (10)$$

where $\mathbf{P}_2$ is the background error covariance at time t_2. Instead, it is preferable, in strongly nonlinear conditions, to minimise for the (re-)analysed state $\mathbf{x}_1$ conditional on the future observations:

$$\mathcal{J}_1(\mathbf{x}_1) = \frac{1}{2}\left(\mathbf{y}_2 - H(\mathcal{M}(\mathbf{x}_1))\right)^{\mathrm{T}}\mathbf{R}^{-1}\left(\mathbf{y}_2 - H(\mathcal{M}(\mathbf{x}_1))\right) + \frac{1}{2}\left(\mathbf{x}_1 - \mathbf{x}_1^{(0)}\right)\mathbf{P}_1^{-1}\left(\mathbf{x}_1 - \mathbf{x}_1^{(0)}\right), \quad (11)$$

where $\mathcal{M}$ is the transition model from time t_1 to time t_2, and $\mathbf{P}_1$ is the background error covariance at time t_1 obtained by a forecast from the assimilation step prior to t_1 and does not depend on future observations. This is equivalent to solving a one-lag extended Kalman smoother. Even with a linear observation operator, cost function $\mathcal{J}_1$ is difficult to solve when the model is nonlinear, since as opposed to $\mathcal{J}_2$, $\mathcal{J}_1$ is non-quadratic, so that this cost function needs to be iteratively optimised. This approach has been suggested by (Wishner et al., 1969; Jazwinski, 1970; Tarantola, 2005).

Cost function $\mathcal{J}_1$ can be, for instance, optimised using a Newton approach

$$\mathbf{x}_1^{(p+1)} = \mathbf{x}_1^{(p)} - \mathcal{H}_{(p)}^{-1}\nabla\mathcal{J}_1(\mathbf{x}_1^{(p)}), \quad (12)$$

where p is the iteration index and where the gradient and the Hessian are respectively given by

$$\nabla\mathcal{J}_1^{(p)} = -\mathbf{M}_{(p)}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\left(\mathbf{y}_2 - H\mathcal{M}(\mathbf{x}_1^{(p)})\right) + \mathbf{P}_1^{-1}\left(\mathbf{x}_1^{(p)} - \mathbf{x}_1^{(0)}\right) \quad (13)$$

$$\mathcal{H}_{(p)} = \mathbf{P}_1^{-1} + \mathbf{M}_{(p)}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}\mathbf{M}_{(p)}, \quad (14)$$

where $\mathbf{M}_{(p)}$ is the tangent linear model computed at $\mathbf{x}_1^{(p)}$.

More recently it has been suggested to use a similar approach, but with the EnKF (Gu and Oliver, 2007). Kalnay and Yang (2010) suggested repeating the assimilation cycle by applying the ensemble transform calculated at t_2 to the ensemble from the previous iteration at t_1 until the best fit to observations is obtained. One advantage of the EnKF framework is that the use of the tangent linear model and its adjoint is replaced with the use of the ensemble. It has been properly formalized as the iterative ensemble Kalman filter (IEnKF) in the framework of the deterministic filters and tested on the Lorenz '63 and '95 by Sakov et al. (2012). At the cost of additional iterations and, hence, of additional propagations of the ensemble, this IEnKF was shown to significantly outperform EnKF especially for large time interval between updates.

Two versions of the filter were put forward. The first one mimics the use of the tangent linear model by the propagation of a rescaled ensemble, a *bundle* of trajectories. It is called IEKF by Sakov et al. (2012). In this study, we shall not use IEKF, but a close variant that will be called the *bundle* variant. It will turn out to offer a performance improvement over the IEKF. The second one consists in transforming the ensemble before its propagation, using the ensemble transform

$$\mathbf{T}_{(p)} = \left((N-1)\mathbf{I}_N + \mathbf{Y}_{(p)}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{Y}_{(p)}\right)^{-1/2}, \quad (15)$$

obtained at the previous iteration. The inverse transformation is applied after propagation. It performs a form of preconditioning of the optimisation problem in ensemble space. We shall call it here the *transform* variant.

The IEnKF still requires inflation. In strongly nonlinear conditions, the inflation factor may be large, although Sakov et al. (2012) remark that the sensitivity to it is weaker and that the required magnitudes are smaller than in non-iterative schemes.

Note that the minimisation scheme implicitly adopted by Sakov et al. (2012) is the iterative Newton scheme. Because the Newton method is one of many schemes to minimise cost function Eq. (11), there is a large freedom in choosing the iterative scheme. In this article, we shall choose the Levenberg-Marquardt algorithm (Levenberg, 1944; Marquardt, 1963), in order to have a better control of the optimisation and safely generalize the iterative ensemble Kalman filter of Sakov et al. (2012) to an inflation-free iterative ensemble Kalman filter, which is the final goal of this article.

1.3 Outline

In this article, the focus is on the deterministic EnKFs rather than the stochastic EnKF. The variants of the filters with localization are deliberately not studied. It is well beyond the objective of this article and, to some extent, a disconnected topic.

To start with, we shall come back to the EnKF-N and IEnKF filters. Our starting points are the results of Bocquet (2011) and Sakov et al. (2012), and we develop on these two filters. In Sect. 2 we shall first give another interpretation of EnKF-N. It makes an explicit connection with inflation and provides an efficient way to optimise cost function Eq. (7). Besides it sheds light on the use of inflation and why it can be surprisingly efficient when accounting for sampling errors. In Sect. 3, we introduce an implementation of IEnKF that is based on a Levenberg-Marquardt algorithm. As a precursor of the trust-region methods, it allows the controlling of the update which will be useful in the generalization to a finite-size version of the filter, which we introduce in Sect. 4.

Numerical experiments are carried out on the new systems in Sect. 5. Conclusions and leads for improvements are given in Sect. 6.

2 The finite-size ensemble Kalman filter

In this section, we give a new insight on the EnKF-N, whose principles have been recalled in the introduction. For the sake of simplicity, the observation operator H is assumed linear. The generalization to a nonlinear H is not difficult and will be treated in the framework of iterative EnKFs anyway (Sects. 3 and 4). Equation (7) can be written

$$\widetilde{\mathcal{J}}(\mathbf{w}) = \frac{1}{2}\left(\boldsymbol{\delta}-\mathbf{Y}\mathbf{w}\right)^{\mathrm{T}}\mathbf{R}^{-1}\left(\boldsymbol{\delta}-\mathbf{Y}\mathbf{w}\right) + \frac{N}{2}\ln\left(\varepsilon_N+\mathbf{w}^{\mathrm{T}}\mathbf{w}\right), \tag{16}$$

where $\mathbf{Y}=\mathbf{H}\mathbf{A}$ is the observation anomaly matrix and $\boldsymbol{\delta}=\mathbf{y}-H(\overline{\mathbf{x}})$ is the innovation. The minimisation of $\widetilde{\mathcal{J}}$ is performed over ensemble space, which is numerically efficient in high-dimensional applications. However, this cost function was shown to be non-convex. Besides it has one global minimum and possibly additional local minima. In practice, in Bocquet (2011), the L-BFGS-B minimizer of Byrd et al. (1995) was used. The quasi-Newton algorithm prevents from forming a Hessian with a negative eigenvalue, which might occur with the joint use of such cost function and of a Newton optimisation method.

2.1 Dual scheme

Although this path is successful and efficient, we would like to find a more explicit scheme in order to establish stronger connections with traditional EnKF with inflation. We wish to split the radial degree of freedom of $\mathbf{w}$, that is $\sqrt{\mathbf{w}^{\mathrm{T}}\mathbf{w}}$, from its angular degrees of freedom, that is $\mathbf{w}/\sqrt{\mathbf{w}^{\mathrm{T}}\mathbf{w}}$.

To alleviate the notations, we define the functions:

$$g(\mathbf{w}) = \left(\boldsymbol{\delta}-\mathbf{Y}\mathbf{w}\right)^{\mathrm{T}}\mathbf{R}^{-1}\left(\boldsymbol{\delta}-\mathbf{Y}\mathbf{w}\right), \tag{17}$$

$$f(\rho) = N\ln\left(\varepsilon_N+\rho\right). \tag{18}$$

Having in mind $\widetilde{\mathcal{J}}(\mathbf{w})$ of Eq. (16), a related Lagrangian is introduced:

$$\mathcal{L}(\mathbf{w},\rho,\zeta) = \frac{1}{2}g(\mathbf{w}) + \frac{1}{2}\zeta\left(\mathbf{w}^{\mathrm{T}}\mathbf{w}-\rho\right) + \frac{1}{2}f(\rho). \tag{19}$$

The dual cost function, that we define for $\zeta>0$, is given by:

$$\mathcal{D}(\zeta) = \inf_{\mathbf{w}}\sup_{\rho\geq 0}\mathcal{L}(\mathbf{w},\rho,\zeta). \tag{20}$$

It is easy to check that the maximum and minimum that define $\mathcal{D}(\zeta)$ exist. The dual problem consists in minimizing this dual cost function:

$$\Delta = \inf_{\zeta>0}\mathcal{D}(\zeta), \tag{21}$$

whereas the original problem

$$\Pi = \inf_{\mathbf{w}}\widetilde{\mathcal{J}}(\mathbf{w}) \tag{22}$$

is called the primal problem. In Appendix A, we demonstrate that the two global minima of $\mathcal{D}$ and $\widetilde{\mathcal{J}}$ coincide. In particular $\Pi=\Delta$. This is a remarkable non-quadratic case of so-called strong duality (Borwein and Lewis, 2000). This means that the problem of finding the global minimum of our original (primal) problem can rigorously be traded with the problem of finding the global minimum of the dual function.

To connect the corresponding optimal $\zeta^\star$, $\rho^\star$ and $\mathbf{w}_\star$, and to compute the dual cost function, we write the condition of stationarity of the Lagrangian over ρ and over $\mathbf{w}$. It yields

$$\zeta^\star = \frac{\mathrm{d}f}{\mathrm{d}\rho}(\rho^\star) = \frac{N}{\varepsilon_N+\rho^\star}, \tag{23}$$

$$\zeta^\star\mathbf{w}_\star = \nabla_{\mathbf{w}}g(\mathbf{w}_\star) = -\mathbf{Y}^{\mathrm{T}}\mathbf{R}^{-1}(\boldsymbol{\delta}-\mathbf{Y}\mathbf{w}_\star). \tag{24}$$

Since $\rho^\star\geq 0$, Eq. (23) implies that $\zeta^\star$ belongs to $[0, N/\varepsilon_N]$, the interval from 0 (excluded) to N/ε_N (included). The solutions of Eqs. (23) and (24) are

$$\rho^\star = \frac{N}{\zeta^\star}-\varepsilon_N, \tag{25}$$

$$\mathbf{w}_\star = \left(\mathbf{Y}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{Y}+\zeta^\star\mathbf{I}_N\right)^{-1}\mathbf{Y}^{\mathrm{T}}\mathbf{R}^{-1}\boldsymbol{\delta}. \tag{26}$$

By inserting these solutions in the Lagrangian, one obtains the dual cost function

$$\mathcal{D}(\zeta) = \mathcal{L}(\mathbf{w}_\star,\rho^\star,\zeta) = \frac{1}{2}\boldsymbol{\delta}^{\mathrm{T}}\left(\mathbf{R}+\mathbf{Y}\zeta^{-1}\mathbf{Y}^{\mathrm{T}}\right)^{-1}\boldsymbol{\delta} + \frac{\varepsilon_N\zeta}{2}+\frac{N}{2}\ln\frac{N}{\zeta}-\frac{N}{2}. \tag{27}$$

The dual cost function is a function of one single variable. Hence, it is easy to find its global minimum, even in the presence of several minima.

As a result, instead of minimizing $\widetilde{\mathcal{J}}(\mathbf{w})$ over $\mathbf{w}$, one can equivalently:

1. find the global minimum $\zeta^\star$ of $\mathcal{D}(\zeta)$ in $]0, N/\varepsilon_N]$,
2. compute $\mathbf{w}_\star = \left(\mathbf{Y}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{Y}+\zeta^\star\mathbf{I}_N\right)^{-1}\mathbf{Y}^{\mathrm{T}}\mathbf{R}^{-1}\boldsymbol{\delta}$.

The implementation of the corresponding EnKF-N is detailed in algorithm 2.

2.2 Assets of the dual approach

Let us discuss this alternate minimisation. It has several advantages.

Firstly, even though the primal problem seems to be efficiently solved with the help of a quasi-Newton minimizer, it is only guaranteed to find one minimum, not necessarily the global one. On the contrary, because the search of the global

Algorithm 2 Dual finite-size ensemble Kalman filter.

Require: The forecast ensemble $\{\mathbf{x}_k\}_{k=1,\ldots,N}$, the observations $\mathbf{y}$, and error covariance matrix $\mathbf{R}$

1: Compute the mean $\overline{\mathbf{x}}$ and the anomalies $\mathbf{A}$ from $\{\mathbf{x}_k\}_{k=1,\ldots,N}$.
2: Compute $\mathbf{Y} = \mathbf{HA}$, $\boldsymbol{\delta} = \mathbf{y} - \mathbf{H}\overline{\mathbf{x}}$
3: Find the minimum:

$$\zeta^a = \min_{\zeta \in]0, N/\varepsilon_N]} \left\{ \boldsymbol{\delta}^{\mathrm{T}} \left(\mathbf{R} + \mathbf{Y}\zeta^{-1}\mathbf{Y}^{\mathrm{T}} \right)^{-1} \boldsymbol{\delta} + \varepsilon_N \zeta + N \ln \frac{N}{\zeta} - N \right\} \quad (28)$$

4: Compute $\Omega_a = \left(\mathbf{Y}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{Y} + \zeta^a \mathbf{I}_N \right)^{-1}$
5: Compute $\mathbf{w}_a = \Omega_a \mathbf{Y}^{\mathrm{T}}\mathbf{R}^{-1}\boldsymbol{\delta}$.
6: Compute $\mathbf{x}^a = \overline{\mathbf{x}} + \mathbf{A}\mathbf{w}_a$.
7: Compute $\mathbf{W}^a = ((N-1)\Omega_a)^{1/2}\mathbf{U}$
8: Compute $\mathbf{x}^a_k = \mathbf{x}^a + \mathbf{A}\mathbf{W}^a_k$

minimum is reduced to a one-dimensional problem, the dual problem allows to easily find the global minimum. We have performed numerical tests about this issue on the Lorenz '63 model (Lorenz, 1963) which are discussed in Sect. 5. We found that in marginal cases, the dual approach (global optimisation) would perform better than the primal scheme (local optimisation). However, for most of the tests (Lorenz '63 and Lorenz '95) we found no significant performance differences between the dual and primal approaches. Specifically for the numerical tests we have performed with Lorenz '95, the primal and dual algorithm systematically led to the same optimum.

Secondly, the algorithm clearly exhibits the extra cost of EnKF-N against EnKF: minimizing Eq. (27) over scalar ζ. Note that the inverse matrix in the first term of Eq. (27) can be obtained from a singular value decomposition that can also be used in the gain computation Eq. (26). In practice, for the experiments of Sect. 5, we found that the extra cost is negligible.

Thirdly, this algorithm parallels the traditional EnKF scheme. Comparing Eq. (26) with Eqs. (20) and (21) of Hunt et al. (2007), it is clear that ζ replaces $N-1$ found in the traditional deterministic filters. That is why ζ can be seen as the effective size of the ensemble: the mean number of members that truly contribute in the effective prior. The Lagrange multiplier ζ is also connected to an inflation of the prior error covariance matrix. It can be absorbed into a rescaling of the ensemble anomalies by a factor $\sqrt{(N-1)/\zeta}$, so that the analysis Eq. (26) coincides with the usual formula. Therefore, if we define the inflation operation as:

$$\mathbf{x}_k \longrightarrow \overline{\mathbf{x}} + \lambda(\mathbf{x}_k - \overline{\mathbf{x}}), \quad (29)$$

EnKF-N forecasts the following *optimal* prior inflation factor

$$\lambda^\star = \sqrt{\frac{N-1}{\zeta^\star}}. \quad (30)$$

That EnKF-N can equivalently be rewritten as a a traditional EnKF with an optimal (prior) inflation factor sheds light as to why inflation can be so successful in dealing with sampling errors. At a fundamental level, the introduction of ζ was possible because of the rotational invariance of the prior in ensemble space. In short, the inflation works well to compensate sampling errors because of the exchangeability of members in the ensemble.

Finally, we found the dual approach to be more stable than the primal one. Indeed, in the primal algorithm the BFGS iterations cannot lead to a singular Hessian by construction. However, at the end of the minimisation, one has to generate the new ensemble by computing Ω_a in algorithm 1. In infinite machine precision, the Hessian is positive-definite. However, it might be singular in finite numerical conditions, in very demanding conditions, that we only found in the severe Lorenz '63 test case of Sect. 5. The dual algorithm does not meet this problem because the value of ζ_a that enters the definition of Ω_a is controlled and is guaranteed to be positive, since $\mathcal{D}(\zeta)$ goes to $+\infty$ when ζ vanishes.

3 The iterative ensemble Kalman filter

In this section, we follow the steps of Sakov et al. (2012). One minor difference is in the formulation of the iterative ensemble filter, which is written here in ensemble space. Another one is an improvement in terms of performance over the IETKF of Sakov et al. (2012), whose updated version will be called the bundle IEnKF. Another significant difference consists of noticing that one has the freedom to choose any iterative optimisation scheme. Here we shall choose the Levenberg-Marquardt scheme.

3.1 The IEnKF in ensemble space

Let us note $\overline{\mathbf{x}}$ the ensemble mean at time t_1 and $\mathbf{A}$ the anomaly ensemble matrix at time t_1. Equivalently to using cost function Eq. (11) to perform the analysis, one can use a cost function depending on the coordinates $\mathbf{w}$ of $\mathbf{x}_1 = \overline{\mathbf{x}} + \mathbf{A}\mathbf{w}$ in ensemble space:

$$\widetilde{\mathcal{J}}(\mathbf{w}) = \frac{1}{2} \left(\mathbf{y}_2 - H(\mathcal{M}(\overline{\mathbf{x}} + \mathbf{A}\mathbf{w})) \right)^{\mathrm{T}} \mathbf{R}^{-1} \times \left(\mathbf{y}_2 - H(\mathcal{M}(\overline{\mathbf{x}} + \mathbf{A}\mathbf{w})) \right) + \frac{1}{2}(N-1)\mathbf{w}^{\mathrm{T}}\mathbf{w}, \quad (31)$$

The derivation of the background term can be read in Hunt et al. (2007). The iterative minimisation of the cost function following a Newton algorithm reads:

$$\mathbf{w}^{(p+1)} = \mathbf{w}^{(p)} - \widetilde{\mathcal{H}}^{-1}_{(p)} \nabla \widetilde{\mathcal{J}}(\mathbf{w}^{(p)}), \quad (32)$$

where p is the iteration index and where the gradient and the Hessian are given by

$$\nabla\widetilde{\mathcal{J}}_{(p)} = -\mathbf{Y}_{(p)}^{\mathrm{T}}\mathbf{R}^{-1}\left(\mathbf{y}_2 - H\mathcal{M}(\overline{\mathbf{x}}+\mathbf{A}\mathbf{w}^{(p)})\right) + (N-1)\mathbf{w}^{(p)}, \quad (33)$$

$$\widetilde{\mathcal{H}}_{(p)} = (N-1)\mathbf{I}_N + \mathbf{Y}_{(p)}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{Y}_{(p)}, \quad (34)$$

where $\mathbf{Y}_{(p)} = [HM\mathbf{A}]'_{(p)}$ is the tangent linear of the operator from ensemble space to the observation space that propagates the ensemble anomalies $\mathbf{A}$ through the model $\mathcal{M}$ and the observation operator H, and computed at $\mathbf{x}_1^{(p)} = \overline{\mathbf{x}} + \mathbf{A}\mathbf{w}^{(p)}$.

3.2 Levenberg-Marquardt algorithm

The Levenberg-Marquardt algorithm (Levenberg, 1944; Marquardt, 1963) is a precursor of the trust-region method in the sense that it seeks to determine when the (superlinear) Newton method is applicable and when it is not, and should be replaced by the slower but safer gradient descent method (also known as steepest descent). The distinction between the two regimes is obtained by the ratio

$$\theta = \frac{\widetilde{\mathcal{J}}(\mathbf{w}) - \widetilde{\mathcal{J}}(\mathbf{w}')}{L(\mathbf{0}) - L(\mathbf{w}'-\mathbf{w})}, \quad (35)$$

where $\mathbf{w}'$ is the new tentative vector, and L is the quadratic local expansion of $\widetilde{\mathcal{J}}$:

$$L(\Delta\mathbf{w}) = \widetilde{\mathcal{J}}(\mathbf{w}) + (\Delta\mathbf{w})^{\mathrm{T}}\nabla\widetilde{\mathcal{J}} + \frac{1}{2}(\Delta\mathbf{w})^{\mathrm{T}}\widetilde{\mathcal{H}}\Delta\mathbf{w}. \quad (36)$$

Instead of determining a region of confidence as in modern trust-region methods, it shifts the Hessian in ensemble space used in the Newton method: $\widetilde{\mathcal{H}} \longrightarrow \widetilde{\mathcal{H}} + \mu\mathbf{I}_N$, where μ is a positive constant. If the quadratic expansion of the cost function allowed by the gradient and the Hessian matches the behaviour of the exact cost function, which corresponds to a large θ, then μ is reduced, otherwise μ is increased (small or negative θ). When μ is small the algorithm is close to a Gauss-Newton method which has a superlinear convergence. When μ is large enough the algorithm is close to a gradient descent method. A textbook and a clear synthesis on this well-established technique are given by Nocedal and Wright (2006) and by Madsen et al. (2004). In the following, the Levenberg-Marquardt algorithm described in Madsen et al. (2004) is adapted to our ensemble Kalman filter context.

In the part of the algorithm which seeks a satisfying μ, a single model propagation is necessary. At each successful Newton step, the propagation of the full ensemble is required. The resulting IEnKF scheme is detailed in algorithm 3 of Appendix B. The algorithm describes both the bundle and the transform variants. When algorithmic steps differ in the two variants, the variant is explicitly mentioned.

In the bundle variant, the ensemble is shrunk by a small ϵ factor before propagation. It is chosen to be $\epsilon = 10^{-4}$ throughout this article. It is the same as that of Sakov et al. (2012), and we found this value to be suitable for the experiments described in Sect. 5. After propagation, the ensemble is inflated by a factor $1/\epsilon$.

Note that in the transform variant, the last propagation of the ensemble is often unnecessary. But when the algorithm exits on, e.g., the maximum iteration criterion, it may be necessary to propagate the ensemble, because it may not have been updated at the latest $\mathbf{w}$ (as opposed to the Gauss-Newton implementation of Sakov et al., 2012). Therefore, it is possible to avoid this last propagation. However, for the sake of simplicity we have preferred to keep the simpler implementation.

4 Combining the finite-size and iterative ensemble Kalman filters

The EnKF-N and IEnKF are complementary since, when looking at the underlying cost function of the analysis, the EnKF-N modifies the prior likelihood, while IEnKF modifies the observational likelihood. Combining the outcome of the previous sections, the cost function used in the analysis as a function of coordinates $\mathbf{w}$ of $\mathbf{x}_1 = \overline{\mathbf{x}} + \mathbf{A}\mathbf{w}$ should be

$$\widetilde{\mathcal{J}}(\mathbf{w}) = \frac{1}{2}\left(\mathbf{y}_2 - H(\mathcal{M}(\overline{\mathbf{x}}+\mathbf{A}\mathbf{w}))\right)^{\mathrm{T}}\mathbf{R}^{-1} \times \left(\mathbf{y}_2 - H(\mathcal{M}(\overline{\mathbf{x}}+\mathbf{A}\mathbf{w}))\right) + \frac{N}{2}\ln\left(\varepsilon_N + \mathbf{w}^{\mathrm{T}}\mathbf{w}\right), \quad (37)$$

$\widetilde{\mathcal{J}}$ has a global minimum in $\mathbb{R}^N$ since it is a positive function and since it goes to infinity as $\mathbf{w}^{\mathrm{T}}\mathbf{w}$ goes to infinity. Several strategies are certainly possible to minimise this cost function.

4.1 Dual approach

The first one consists in using the dual transformation put forward for the EnKF-N in Sect. 2. The derivation holds if one replaces g with

$$g(\mathbf{w}) = \left(\mathbf{y}_2 - H\mathcal{M}(\overline{\mathbf{x}}+\mathbf{A}\mathbf{w})\right)^{\mathrm{T}}\mathbf{R}^{-1} \times \left(\mathbf{y}_2 - H\mathcal{M}(\overline{\mathbf{x}}+\mathbf{A}\mathbf{w})\right). \quad (38)$$

In particular, the strong duality holds (this can be checked going through Appendix A). As a consequence, it demonstrates that there should be an optimal inflation factor in this context. However, each one of the subproblems, indexed by ζ,

$$\inf_{\mathbf{w}}\left(g(\mathbf{w}) + \zeta\mathbf{w}^{\mathrm{T}}\mathbf{w}\right) \quad (39)$$

which were equivalent to solving the traditional EnKF analysis in ensemble space Eq. (26), is now equivalent to solving an IEnKF analysis, with a prior inflation factor $\sqrt{(N-1)/\zeta}$.

Hence, such a path may be numerically costly. Solutions such as primal-dual algorithms may be contemplated, but we

prefer to explore a more straightforward way to minimise Eq. (37). However, the existence of an optimal inflation factor is proven in this context.

4.2 Primal approach

The second and more direct way to minimise Eq. (37) is to minimise with respect to the $\mathbf{w}$ coordinates, with the sole guarantee to find a local minimum. The gradient and Hessian are

$$\nabla\widetilde{\mathcal{J}} = -\mathbf{Y}^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{y}_2 - H\mathcal{M}(\overline{\mathbf{x}}+\mathbf{A}\mathbf{w})) + N\frac{\mathbf{w}}{\varepsilon_N+\mathbf{w}^{\mathrm{T}}\mathbf{w}}, \tag{40}$$

$$\widetilde{\mathcal{H}} = N\frac{\left(\varepsilon_N+\mathbf{w}^{\mathrm{T}}\mathbf{w}\right)\mathbf{I}_N - 2\mathbf{w}\mathbf{w}^{\mathrm{T}}}{\left(\varepsilon_N+\mathbf{w}^{\mathrm{T}}\mathbf{w}\right)^2} + \mathbf{Y}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{Y}. \tag{41}$$

The main drawback is that the Hessian may have a non-positive eigenvalue. A priori, this precludes Newton approaches. One solution around it is to use a quasi-Newton minimizer such as L-BFGS-B (Byrd et al., 1995). Only the gradient is provided to the minimizer. This approach was successfully tested. However, on very rare occasions and only for very non-Gaussian systems, the filter can break because the approximate tangent linear leading to an approximate adjoint leads the minimizer into re-initializing the quasi-Hessian, which may break the filter. This is also a very economical approach, as part of the work is carried out by the minimizer, known to be very efficient.

However, for this article, we prefer to report results obtained in a more controlled environment. We can use the Levenberg-Marquardt algorithm as it uses a shifted Hessian, which can be made positive-definite, for large enough μ. Besides, we know that at the minimum of the cost function, the Hessian is positive-definite, so there is a neighborhood around the minimum where the Hessian is positive-definite. This means that in the vicinity of the minimum, the Levenberg-Marquardt can safely rely on the Hessian of the cost function.

As a diagnostics, an intermediate result of the dual approach can be used. Indeed, from Eq. (23), and using the fact that $\rho^\star = \mathbf{w}_\star^{\mathrm{T}}\mathbf{w}_\star$ at the saddle-point, an equivalent optimal prior inflation factor can be obtained from the analysis in ensemble space:

$$\lambda^\star = \sqrt{\frac{N-1}{N}\left(\varepsilon_N+\mathbf{w}_\star^{\mathrm{T}}\mathbf{w}_\star\right)}. \tag{42}$$

For instance, the statistics of $\zeta^\star$ over a long data assimilation experiment can tell us about the need to enforce adaptive inflation or not. Cases where $\zeta^\star$ is strongly fluctuating are cases where IEnKF-N may outperform IEnKF with optimal but constant inflation.

The details of this primal scheme are given in algorithm 4 of Appendix B. In severe conditions, such as the one studied at the end of Sect. 5, it might be tedious to define a program in which μ is always large enough so as to guarantee a positive-definite Hessian. One rigorous way around it is to truncate the Hessian to a positive-definite matrix in the iterative minimisation, while the gradient is left unchanged. This truncation of the Hessian does not apply when building the new ensemble. This is equivalent to a slightly sub-optimal pre-conditioning of the minimisation problem. For instance, one can choose:

$$\widetilde{\mathcal{H}} = \frac{N}{\varepsilon_N+\mathbf{w}^{\mathrm{T}}\mathbf{w}}\mathbf{I}_N + \mathbf{Y}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{Y}, \tag{43}$$

which is an always positive-definite substitute matrix for the Hessian.

5 Numerical experiments

Most numerical tests will be performed on the Lorenz '95 toy-model (Lorenz and Emmanuel, 1998). It has $M = 40$ variables and its dynamics reads for $m = 1, \ldots, M$:

$$\frac{\mathrm{d}x_m}{\mathrm{d}t} = (x_{m+1} - x_{m-2})x_{m-1} - x_m + F \tag{44}$$

where $F = 8$, and the boundary conditions are chosen cyclic. It is integrated using a fourth-order Runge-Kutta scheme with a time-step of 0.05. With this choice of F, the model is strongly chaotic with 13 positive Lyapunov exponents and a doubling time of 0.42 time unit. Hence, it is difficult to control by filtering techniques and can offer simple but severe tests for new methods. Since the model is a simplistic representation of a mid-latitude band of the global atmosphere, a time step of 0.05 in the model's time represents 6 h of physical time.

The data assimilation experiment setup we have chosen consists in computing a reference run of the model, defined as the truth (Sakov and Oke, 2008; Bocquet, 2011; Sakov et al., 2012). This truth is observed for every variable each Δt. Each observation is perturbed by an independent draw from a Gaussian variable of mean 0 and variance 1. Because we do not want to use localization that would mask some of the properties of the filters, the ensemble size is chosen to be $N = 40$ in the rank-sufficient regime. Nevertheless, one can contemplate building local versions of the filters similarly to what was done by Hunt et al. (2007); Bocquet (2011). The performance of a data assimilation run is assessed computing the average root-mean-square of the difference between the data assimilation analysis and the truth (denoted rmse in the following), for a sufficiently long run. In the following, the runs' duration, that does not include the burn-in period, is 5×10^4 days (physical time), and we have checked that the convergence of the statistical indicators, such as the rmse, is satisfying.

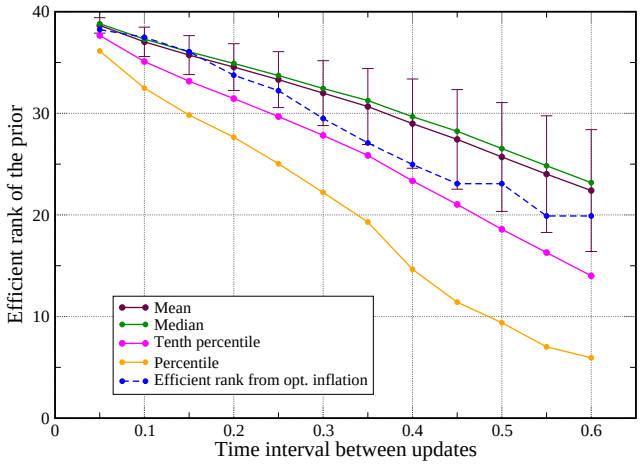

Fig. 1. Mean, standard deviation (shown by errors bars), median, tenth percentile and percentile of the efficient rank ζ^a for a long run of the EnKF-N, and an ensemble size of $N = 40$. The dashed line represents the efficient rank diagnosed from the inflation of EnKF leading to the best analysis rmse.

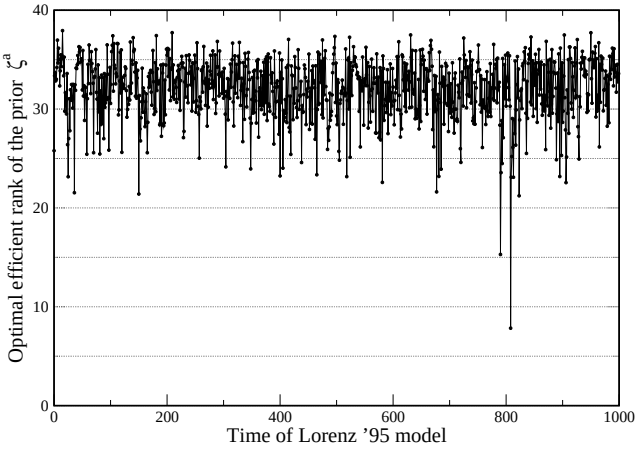

Fig. 2. Sequence of efficient ranks ζ^a in a long run of EnKF-N applied to Lorenz '95 model, with $\Delta t = 0.30$ and with an ensemble of $N = 40$ members.

5.1 Complementary experiments on EnKF-N

EnKF-N has been tested on the Lorenz '63 and Lorenz '95 models in Bocquet (2011). Here, we wish to report some complementary experiments related to the dual implementation of EnKF-N, introduced in Sect. 2, in its $\varepsilon_N = 1$ variant.

The efficient rank ζ^a of the ensemble prior, as estimated by the EnKF-N formalism, is computed using Eq. (23), and $\rho^a = \mathbf{w}_a^{\mathrm{T}}\mathbf{w}_a$ at optimality:

$$\zeta^a = \frac{N}{\varepsilon_N + \mathbf{w}_a^{\mathrm{T}}\mathbf{w}_a}, \tag{45}$$

for a long run of EnKF-N applied to Lorenz '95, with the setup described above. Note that $\zeta^a = N - 1$ would correspond to a deterministic EnKF without inflation. The time interval between updates Δt is varied: $\Delta t = 0.05, 0.10, \ldots, 0.60$, so as to probe the critical cases where inflation is strong (when ζ^a/N is small). The statistics of ζ^a are plot in Fig. 1: mean, standard deviation, tenth percentile and percentile.

This allows to study the variability of the efficient rank, or, indirectly, of the *optimal* prior inflation factor (as seen by EnKF-N). It is clear that the efficient rank decreases with Δt. More importantly, the variability increases very significantly since the tenth percentile decreases below 20 for $\Delta t \geq 0.50$. Because of this variability of the rank in time, the optimal inflation factor diagnosed by EnKF-N is not constant and varies with time. That is why we believe EnKF-N may outperform EnKF with optimised constant inflation, when the system becomes significantly nonlinear. Using Eq. (42), the efficient rank has been diagnosed from the optimal inflation of the EnKF obtained from selecting the best rmse configuration. It is also plotted in Fig. 1, and shows a similar evolution as the efficient rank of EnKF-N. Since the mean and median do not differ much, it also suggests that, most of the time, the efficient rank is rather constant, but that it occasionally suffers sudden drops. We have checked this examining sequences of ζ^a. In Fig. 2 is displayed a typical sequence in the case $\Delta t = 0.30$.

5.2 Testing IEnKF, Levenberg-Marquardt implementation

The IEnKF is first tested in its Levenberg-Marquardt implementation, for the bundle and transform variants. Since the focus is on the ability of the IEnKF to handle strong nonlinearity, the time interval between updates Δt is varied: $\Delta t = 0.05, 0.10, \ldots, 0.60$. The filters are run with optimal inflation: the inflation factors leading to the best rmses are selected. The termination criteria of the iterative minimisation are such that the maximal number of iterations is $p_{\max} = 40$, and the precision has reached $||\Delta\mathbf{w}|| = \sqrt{\mathbf{w}^{\mathrm{T}}\mathbf{w}} \simeq e_2$, with $e_2 = 10^{-3}$. Provided $p_{\max}$ is large enough (especially fort large Δt), the results are largely independent from this choice. The choice of the influential criterion e_2 is more critical. A compromise must be found between precision and limiting the number of iterations. This leads to choosing $e_2 = 10^{-3}$, found to be satisfying for all experiments reported in this article. The condition on the gradient was not implemented (see algorithm 3), which means that $e_1 = 0$. We also chose $\tau = 10^{-3}$ which has an influence on the iteration number, since it tells whether the minimisation at the beginning is closer to a Newton method (small τ) or to a steepest descent method. We found that in the weakly nonlinear regime of small Δt, a smaller τ could decrease the number

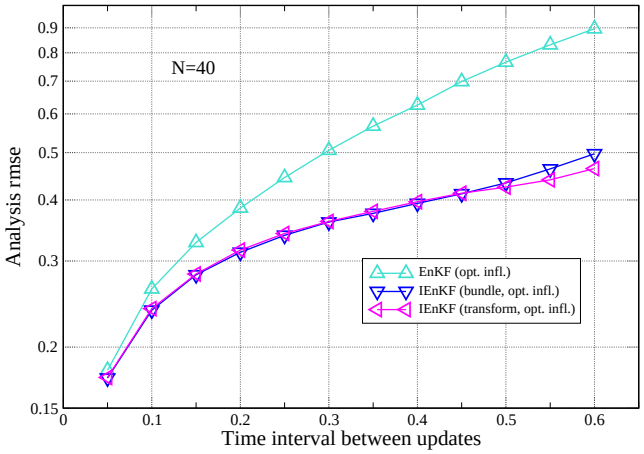

Fig. 3. Analysis root-mean-square error of the IEnKF (bundle and transform versions) for the Lorenz '95 model and for several time intervals Δt between two subsequent updates. For comparison the same results for EnKF (ETKF version) are given for optimal inflation.

of required iterations. This is consistent with the intuition that a steepest descent method is unnecessary in this regime and favouring Newton's method results in a finer convergence.

The results are reported in Fig. 3.

As a comparison, the rmse for the EnKF with optimal inflation is also reported. It clearly shows that, at the cost of cycling the ensemble propagation, the more nonlinear the propagation is the more IEnKF outperforms EnKF.

As opposed to Sakov et al. (2012), we did not find a significant difference between the transform and bundle variants for small and moderate Δt, as far as analysis rmses are concerned. The main difference is not in the Levenberg-Marquardt implementation, but in the fact that after the analysis full cycle, we propagate the ensemble from t_1 to t_2 avoiding to use the rescaled ensemble used within the cycle. This final propagation of the non-rescaled ensemble makes the bundle scheme used in this study no longer based on the extended Kalman filter, as the IEKF in Sakov et al. (2012).

However, note that for $\Delta t \geq 0.5$, the transform outperforms the bundle variant. As opposed to Sakov et al. (2012) implementation of IEnKF, we found that the rmses are rapidly increasing beyond $\Delta t \geq 0.60$. Beyond this time interval, which is more than the doubling time of the dynamical system, multiple minima are likely to form in the underlying cost function Eq. (11) as pointed out by Pires et al. (1996).

The average number of model runs used until convergence is reported in Fig. 4 for the bundle and transform variants. In addition to the $N = 40$, an ensemble of $N = 25$ members is also considered in order to quantitatively compare with the results of Sakov et al. (2012). However, we did not notice any qualitative change between the outcomes of the two ensemble configurations.

It is clear that the bundle variant requires less iterations than the transform variant, between 1 to 2 fewer iterations. This is different from the implementation of Sakov et al. (2012), who showed that the IEKF variant usually requires slightly more iterations than the transform variant, when not using the Levenberg-Marquardt framework. Note that in the transform variant algorithm, as opposed to the bundle variant, it is legitimate to skip the computation of the forecasted ensemble (lines 41–44) because it has been done earlier in line 29. This could lead to a gain of up to one iteration. However, in practice, we found it was inefficient in our test cases, since it had a negative impact on the precision which may have required additional iterations at a later cycle.

The large number of iterations needed at small Δt, between 3 to 4, is not very significant because the Levenberg-Marquardt is not designed to optimally perform a convergence of a very few iterations. In this case, the more straightforward Gauss-Newton method, such as the implementation of Sakov et al. (2012) is preferable.

5.3 Testing IEnKF-N

Here, the IEnKF-N, bundle and transform versions, are tested on the same configuration as IEnKF. The results are reported in Fig. 5.

First the two variants of IEnKF-N offer the same rmses in this configuration over the full range of Δt. Secondly, in this example, they are essentially equally performing or better than the IEnKF with optimal inflation. This shows that, as hoped for, some of the properties of the finite-size EnKF apply as well to iterative variants of the IEnKF.

The average number of model runs used until convergence is reported in Fig. 4 for the bundle and transform variants of IEnKF-N. Like for the IEnKF experiments, the same termination criterion is chosen ($p_{\max} = 40$, $e_1 = 0$, $e_2 = 10^{-3}$, and $\tau = 10^{-3}$) for all runs. The finite-size versions of the filters require a similar number of iterations (or less) as their optimised inflation counterparts. Here again, the bundle variant is doing better than the transform variant: it requires 1 to 2 less ensemble propagations on average required by the transform variants to achieve the same precision, using the same termination criterion.

Similarly to Fig. 1 in the case of EnKF-N, the statistics of ζ^a are plot in Fig. 6: mean, standard deviation, median, tenth percentile, and percentile.

The results are qualitatively similar. However, as could be expected, the efficient rank is larger in the IEnKF-N case than EnKF-N in the same configuration. Even though the percentile remains high below $\Delta t \leq 0.5$, it rapidly falls off beyond. Consistently this is where the IEnKF-N starts to slightly outperform IEnKF with optimal inflation.

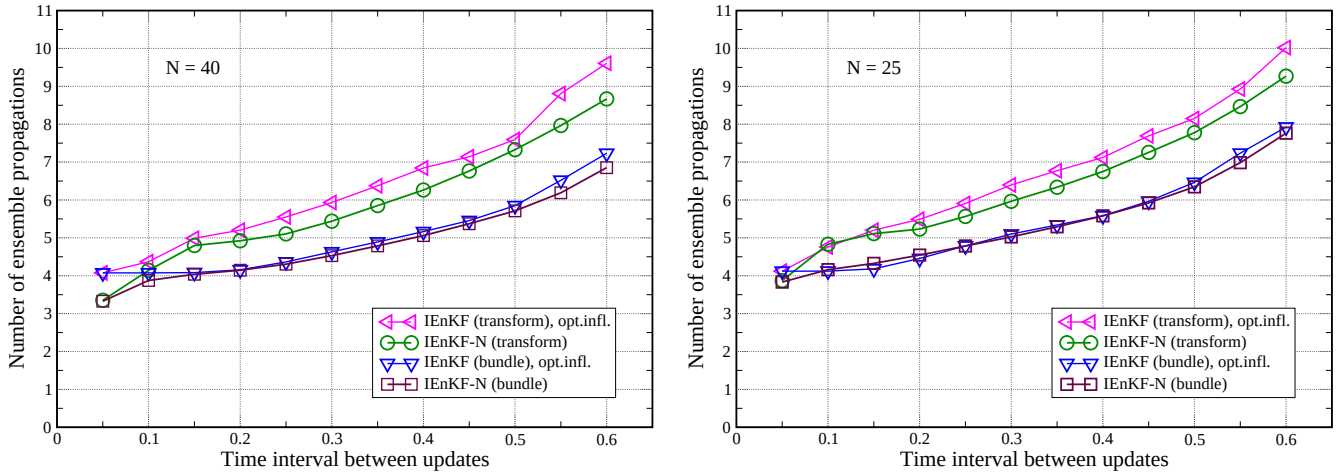

Fig. 4. Average number of model runs, divided by the size of the ensemble for the two variants of the Levenberg-Marquardt IEnKF, for the Lorenz '95 model and for several time intervals Δt between two subsequent updates. On the left, the ensemble size is $N = 40$, whereas on the right, it is $N = 25$.

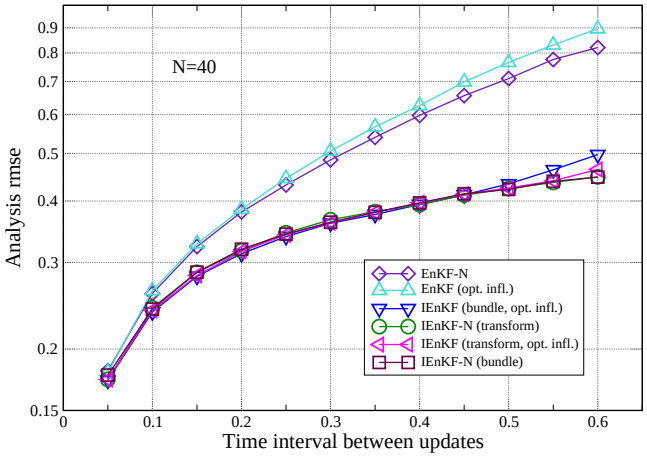

Fig. 5. Analysis root-mean-square error of the IEnKF-N (bundle and transform versions) for the Lorenz '95 model and for several time intervals Δt between two subsequent updates. For comparison the same results are given for EnKF-N, IEnKF (bundle and transform variants) with optimal inflation (repeated from Fig. 3).

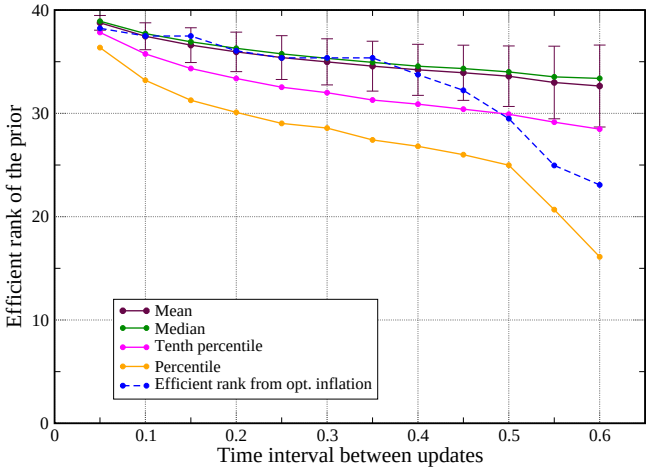

Fig. 6. Mean, standard deviation (shown by errors bars), median, tenth percentile and percentile of the efficient rank ζ^a for a long run of the IEnKF-N, bundle variant and an ensemble size of $N = 40$. The dashed line represents the efficient rank diagnosed from the inflation of EnKF leading to the best analysis rmse.

5.4 Focusing on the weakly nonlinear regime

To mitigate the optimistic results obtained on Lorenz '95, we would like, in this section, to illustrate and discuss apparent limitations of the formalism. To do so, we choose the quite demanding case of the 3-variable Lorenz '63 toy model, with an ensemble of 3 members. It might not be directly relevant to high-dimensional geophysical systems. But the goal of this section is to find out about flaws or inconsistencies of the schemes and to point to directions of possible improvement.

5.4.1 Diagnosis

The Lorenz '63 model (Lorenz, 1963) is defined by the set of three ordinary differential equations:

$$\begin{aligned} \frac{\mathrm{d}x}{\mathrm{d}t} &= \sigma(y - x)\,, \\ \frac{\mathrm{d}y}{\mathrm{d}t} &= \rho x - y - xz\,, \\ \frac{\mathrm{d}z}{\mathrm{d}t} &= xy - \beta z\,, \end{aligned} \tag{46}$$

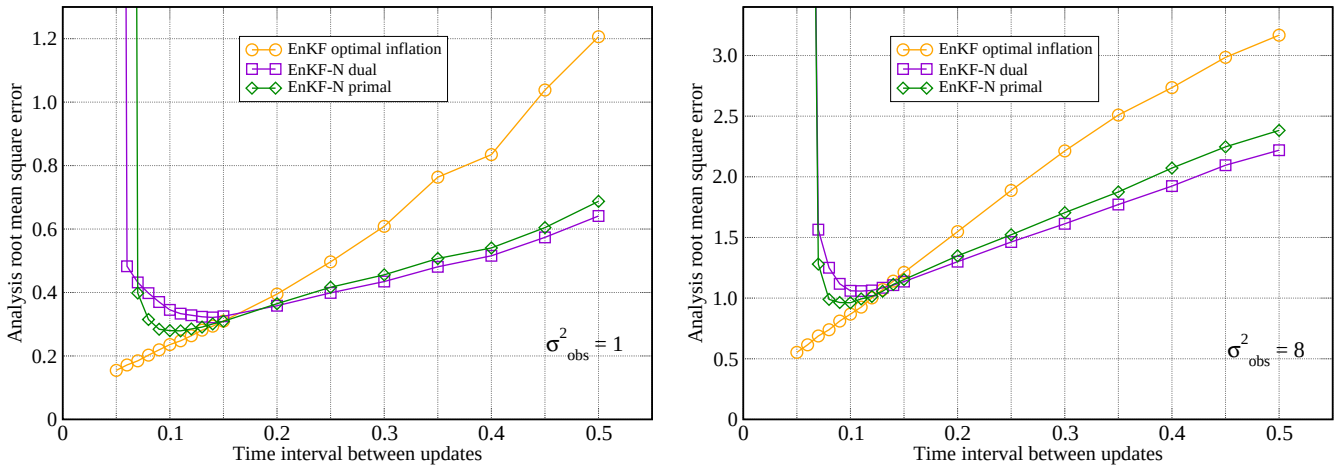

Fig. 7. Analysis root-mean-square error for three filters applied to the Lorenz '63 model ($N = 3$): the EnKF with optimally tuned inflation, the primal EnKF-N and the dual EnKF-N. Left panel illustrates the $\sigma^2_{\rm obs} = 1$ case and the right panel illustrates the $\sigma^2_{\rm obs} = 8$ case.

where $\sigma = 10$, $\rho = 28$, and $\beta = 8/3$. This model is chaotic with a doubling time of 0.78 time unit. It is integrated using a fourth-order Runge-Kutta scheme with a time-step of 0.01 time unit. In the data assimilation experiments ahead, all three variables are observed every Δt. These observations have normal uncorrelated errors of standard deviation $\sigma_{\rm obs}$. All runs are 5×10^5-cycle long with an additional spin-up period of 5×10^4 cycles. To illustrate an apparent limitation of the finite-size formalism in either the EnKF or the iterative EnKF, we vary the time interval between updates between $\Delta t = 0.05$ (nearly linear regime between update steps) and $\Delta t = 0.50$ (strongly nonlinear regime). The error covariance will either be $\sigma^2_{\rm obs} = 1$, or $\sigma^2_{\rm obs} = 8$.

The results are first obtained for three non-iterative ensemble Kalman filters. EnKF with optimally tuned inflation (the inflation factor that yields the best analysis rmse), EnKF-N using its primal implementation, and EnKF-N using its dual implementation are considered. As discussed in Bocquet (2011) about the application of the EnKF-N filters to the Lorenz '63 model, the $\varepsilon_N = 1 + \frac{1}{N}$ variant should be used for such a small ensemble where the analysed state vector is not well estimated by the ensemble mean. The fundamental difference between the primal and dual formalism is that the primal implementation picks up *one* minimum for the analysis, whereas the dual implementation determines the global minimum of the analysis cost function.

The analysis root-mean-square errors are reported on Fig. 7.

First of all, it is clear that the finite-size filters underperform below some threshold equals to $\Delta t = 0.16$ for $\sigma^2_{\rm obs} = 1$, and $\Delta t = 0.14$ for $\sigma^2_{\rm obs} = 8$. They ultimately diverge for too small Δt. This divergence is not directly due to the quasi-linearity of the model in that regime, but to the fact that even in that regime it still diverges without a properly tuned inflation. Indeed we have checked that the EnKF-N behaves to a large extent similarly to an EnKF on a linear advection model (LA model) used in Sakov and Oke (2008). It is, therefore, likely that this underperformance could be traced back to the determination of the optimal prior inflation through the minimisation of Eq. (27) that may become inaccurate in that regime. If one trusts the Bayesian formalism underlying EnKF-N, then according to the discussion in Bocquet (2011) this inaccuracy can be ascribed to either (i) an inappropriate choice of the hyperprior which is at the heart of the derivation of Eq. (2) and *in fine* specifies the particular form of Eq. (27), (ii) or the use of the prior $p(\mathbf{x}|\mathbf{x}_1, \ldots, \mathbf{x}_N)$ rather than $p(\mathbf{x}|\mathbf{y}, \mathbf{x}_1, \ldots, \mathbf{x}_N)$, (iii) or the use of a local minimum rather than the global minimum. Point (iii) can be ruled out, because the present study essentially solved this problem.

Beyond time interval $\Delta t \simeq 0.15$, the EnKF-N primal and dual implementations significantly outperform the EnKF with optimally tuned inflation. As should have been expected, the dual implementation is better than the primal implementation. Note that when $\Delta t \leq 0.15$, the primal variant can beat the dual one. However, since it is in a regime where the formalism breaks down for the likely reasons given above, this difference is essentially irrelevant.

A similar experiment was performed but for $N = 9$, closer to the asymptotic ensemble size limit. In that case, the turning point is at about $\Delta t \simeq 0.04$.

The same study was conducted for the iterative ensemble Kalman filters. The results are reported in Fig. 8. Note that we have not introduced any dual variant of the IEnKF-N. That is why only the IEnKF with optimally tuned inflation and the IEnKF-N in its primal implementation have been tested. The bundle variant was chosen. The results are qualitatively similar to the non-iterative filters. However, since the flow between updates is made more linear thanks to the

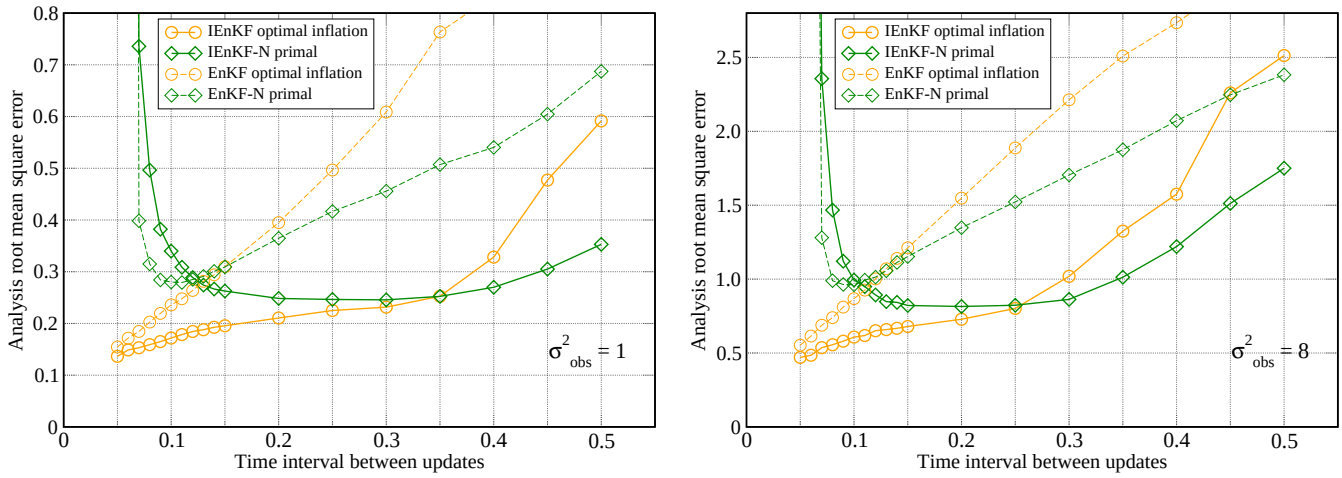

Fig. 8. Analysis root-mean-square error for two iterative filters applied to the Lorenz '63 model ($N = 3$): the IEnKF, bundle variant with optimally tuned inflation, and the primal algorithm for the IEnKF-N, bundle variant. Left panel illustrates the $\sigma^2_{\rm obs} = 1$ case and the right panel illustrates the $\sigma^2_{\rm obs} = 8$ case. The EnKF results of Fig. 7 are reported for comparison.

iterative corrections, the regime is pushed toward the qualitative behaviour of small Δt. Consistently the turning point beyond which the IEnKF-N outperforms, is pushed towards higher Δt.

5.4.2 An empirical solution

Exploring the small Δt regime and the behaviour of the dual cost function Eq. (27), we found that, in that regime, the argument $\zeta^\star$ of its minimum is mostly given by the maximum of the interval, that is N/ε_N. But if N/ε_N asymptotically behaves like $N-1$, it is bigger than $N-1$ at finite N, for $\varepsilon_N = 1$ or $\varepsilon_N = 1+1/N$. For instance, in the case $N = 3$ and $\varepsilon_N = 1+1/N$, one has $N/\varepsilon_N = 2.25$ as compared to $N-1 = 2$. Hence, in that regime, EnKF-N tends to often implicitly deflate the ensemble, which may lead to an overconfident Kalman filter and to its divergence.

From this educated guess, we propose to cap the argument $\zeta^\star$ of the minimum of Eq. (27) at $N-1$. Note that the case $\zeta^\star = N-1$ corresponds to an absence of inflation and does not a priori guarantee the filter's stability. Moreover, rather than capping $\zeta^\star$, we prefer to modify Eq. (27) in such a way that the maximum value of $\zeta^\star$ cannot exceed $N-1$. The first reason for doing so is that the duality result of Appendix A is derived on the full interval $]0, N/\varepsilon_N]$, and an abrupt truncation would invalidate the duality equivalence. The second reason is that a direct capping of $\zeta^\star$ requires an access to Eq. (27), which is not straightforward in the case of the primal schemes. For these two reasons, we propose to renormalize ε_N to $\varepsilon_N = N/(N-1)$. That way, it is easy to check through Eq. (23) that the maximum value of $\zeta^\star$ is $\zeta^\star = N-1$ at $\rho^\star = 0$. In the following *capping* will refer to this renormalization of ε_N, and not to the abrupt truncation of Eq. (27) to the interval $]0, N-1]$.

The resulting performances of the capped EnKF-N, primal and dual variants, are displayed in Fig. 9. With our setup, the dual EnKF-N outperforms or equals EnKF with optimally tuned inflation over the enlarged range $\Delta t \in [0.05, 0.50]$. The benefit of the dual filter over the primal variant is now even clearer in the almost linear regime.

In that regime, $\zeta^\star$ remains close to $N-1$, with occasional excursions to smaller values. These events are mostly provoked by transitions of the dynamics of the Lorenz '63 model between lobes. Because of the short Δt, it is only on these occasions that the ensemble departs from the control run, whereas most of the time the system is in a quasi-linear regime where almost no inflation is needed. The behaviour of $\zeta^\star$ is illustrated in Fig. 10 in the case $\Delta t = 0.05$.

The corresponding results for the bundle IEnKF-N are reported in Fig. 11 for its primal variant. Even though the capping essentially solves the divergence problem diagnosed earlier, IEnKF-N underperforms IEnKF with optimal inflation in the nearly linear regime. Guided by the capped primal and dual EnKF-N results, we conjecture that a dual variant of the capped IEnKF-N would at least partially close this gap. However, implementing the dual IEnKF-N is certainly challenging and left as an open question.

In addition to being empirical, the capping solution cannot be fully satisfying since it leads back to some tuning. However, in the context of this numerical experiment, no scalar was tuned. It was only necessary to distinguish the weakly nonlinear regime from the rest of the Δt range. This also suggests that, in the context of this experiment, a genuinely satisfactory solution that avoids tuning of inflation

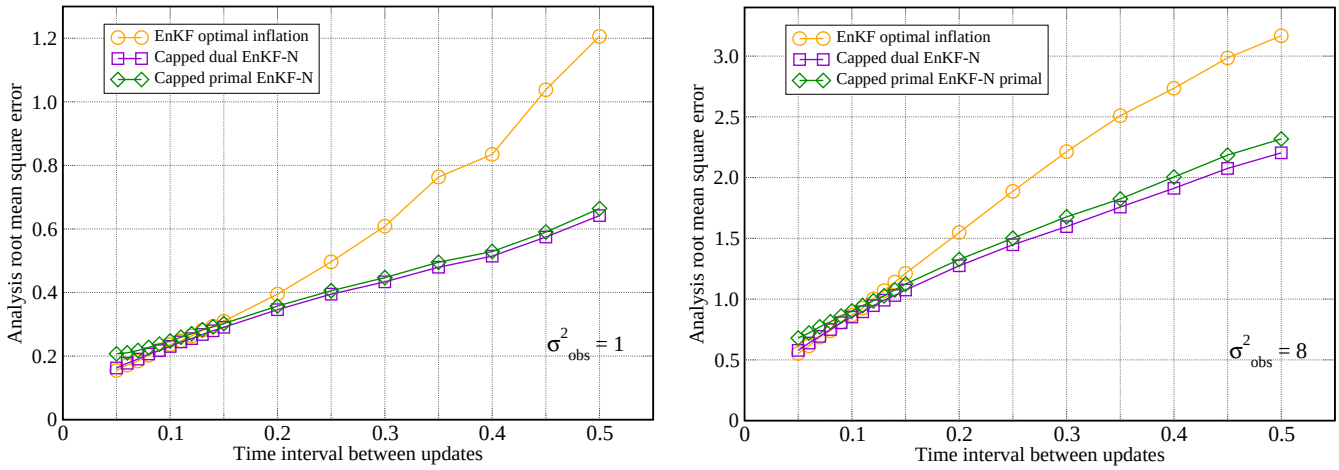

Fig. 9. Analysis root-mean-square error for three filters applied to the Lorenz '63 model ($N = 3$): the EnKF with optimally tuned inflation, the primal EnKF-N with capping and the dual EnKF-N with capping. Left panel illustrates the $\sigma^2_{\text{obs}} = 1$ case and the right panel illustrates the $\sigma^2_{\text{obs}} = 8$ case.

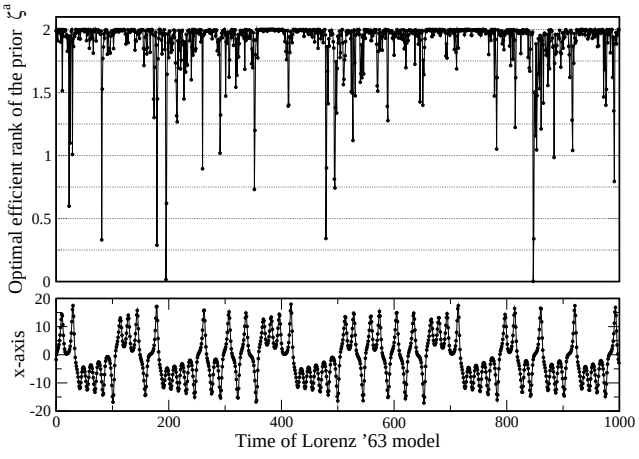

Fig. 10. Sequence of efficient ranks ζ^a in a long run of a capped dual EnKF-N applied to the Lorenz '63 model, with $\sigma^2_{\text{obs}} = 1$ and $\Delta t = 0.05$ which corresponds to a nearly linear regime, and with an ensemble of $N = 3$ members. The lower graph displays the x variable of the Lorenz '63 reference trajectory (the truth). Note the good correlation between the period of relative stability for $\zeta^\star$ and the presence of the true state in orbit within a lobe.

or distinguishing between regimes, would necessitate a fine modelling of the hyperprior, beyond Jeffrey's prior that leads to the particular form of the dual cost function Eq. (27).

6 Conclusion

In this article, we have further more explored two recently developed ensemble Kalman filters meant to operate in high dimensional geophysical systems. We have presented and justified their algorithms. They have been mostly tested on the Lorenz '95 chaotic toy model.

The first filter, the finite-size ensemble Kalman filter, is built to reduce the impact of sampling errors, and was shown, in a perfect model context, to significantly reduce the need for inflation. We have shown that the scheme can be made equivalent to a traditional deterministic EnKF, but with an inflation of the prior, whose value is determined by the minimum of a one-dimensional non-necessarily convex cost-function. Assuming that EnKF-N significantly reduces the need for inflation accounting for sampling errors, this suggests why inflation is usually so efficient in accounting for sampling errors. It tells that this efficiency mathematically comes from the invariance of the ensemble by permutation of its members. Through a dual transformation, it was shown how to find the global minimum of the EnKF-N analysis cost function. This solves one of the open questions raised by Bocquet (2011).

The second filter, the iterative ensemble Kalman filter, outperforms EnKF in strongly nonlinear regime, at the cost of iterating the ensemble propagation between updates. The implementation of Sakov et al. (2012) corresponds to a Newton method in solving the underlying analysis cost function. In this article, we have proposed to use a Levenberg-Marquardt implementation of the filter instead. It offers a better control on the convergence, the positive-definiteness of the Hessian and seems to require less iterations in strongly nonlinear conditions. However, it is less efficient in mild nonlinear conditions where the number of required iterations is small. The transform and bundle variant of IEnKF lead to very similar results and we suggest that they should be considered as variants of a same IEnKF filter.

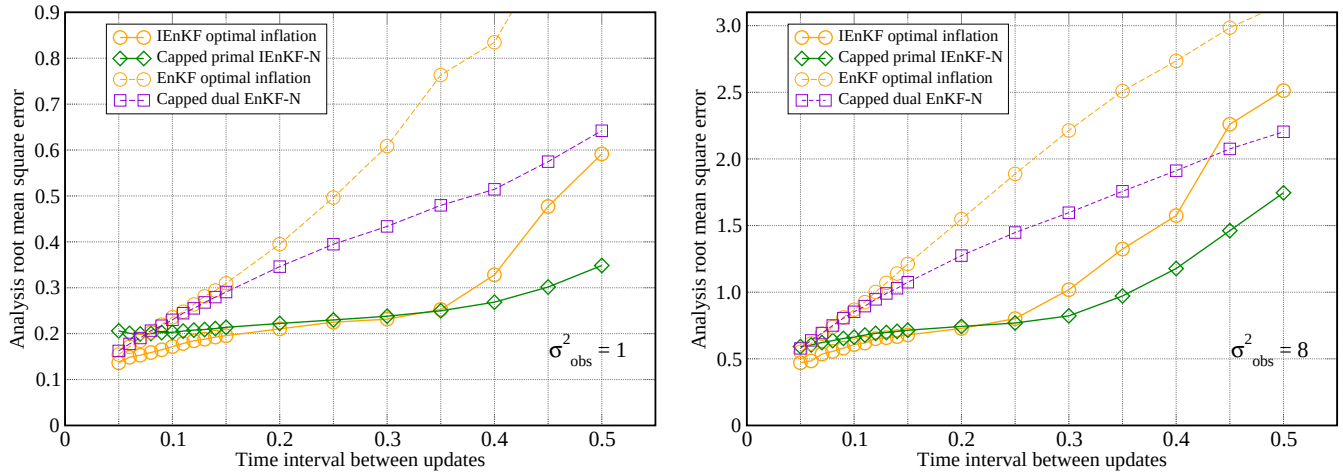

Fig. 11. Analysis root-mean-square error for two iterative filters applied to the Lorenz '63 model ($N = 3$): the IEnKF, bundle variant with optimally tuned inflation, and the primal algorithm for the IEnKF-N with capping, bundle variant. Left panel illustrates the $\sigma^2_{\rm obs} = 1$ case and the right panel illustrates the $\sigma^2_{\rm obs} = 8$ case. The EnKF with optimally tuned inflation and dual EnKF-N with capping results of Fig. 7 are reported for comparison.

Exploiting their complementarity, the two filtering approaches have been combined into an inflation-free iterative ensemble Kalman filter. The Levenberg-Marquardt scheme provides the necessary control on the minimisation of the non-convex underlying cost function. Their performances are close to that of the IEnKF with optimised inflation. The number of ensemble propagation required by IEnKF-N seems to be very similar to that of IEnKF in the same context.

In this article, we have deliberately eluded the rank problem aspect of the EnKF. To go beyond, and extend the results to low-rank ensemble, one should additionally built localization into this algorithm as was for instance done for EnKF-N in Bocquet (2011). But it should bring us far from the preliminary objectives of this article.

A possible improvement over the iterative filters in their Levenberg-Marquardt implementation, would be to re-shape the algorithm so that it avoids extra fundamentally unnecessary iterations for small time interval between updates. One has to keep in mind that the Levenberg-Marquardt method was designed to reduce the large number of iterations needed in the optimisation of a system built on a nonlinear model and, in the present form, is not optimal for system built on a mildly nonlinear model.

Another previously identified challenge is to build an optimisation algorithm for the cost function of the *dual* IEnKF-N.

In the last section of this article, we have identified a demanding regime, the almost (but not exactly) linear regime with a small ensemble, where the finite-size (iterative or not) EnKFs do not apparently succeed in determining a proper effective inflation. We have proposed a motivated but empirical solution which consists in capping the diagnosed value of degrees of freedom in the ensemble at $N-1$ through a modification of the dual cost function Eq. (27). For the cases under scrutiny, this offers a satisfying solution in the EnKF-N case and a promising one in the IEnKF-N case. However, it surely requires a robust argument such as the derivation of the modified dual cost function from a more fitted hyperprior. Additionally it was shown in this special case that knowledge of the global minimum of the analysis, which is provided by the dual algorithm, leads to a systematically better performance than the primal algorithm.

For the longer term, we believe that this finite-size iterative ensemble Kalman filter can be seen as a first elementary one-lag brick of an efficient ensemble Kalman smoother.

Appendix A

Strong duality of the EnKF-N optimisation problem

In this appendix, proof is given that the dual and primal problems lead to the same global minimum. We define the particular Legrendre-Fenchel transform of a function $\alpha > 0 \to h(\alpha)$ as the function $\beta > 0 \to h^*(\beta)$ defined by

$$h^*(\beta) = \inf_{\alpha>0} (\alpha\beta - h(\alpha)) \,. \tag{A1}$$

Applying this transform to function f defined by Eq. (18), it is easy to check that:

$$f^*(\zeta) = \begin{cases} N - \varepsilon_N \zeta - N \ln \frac{N}{\zeta} & \text{for } \zeta \in]0, N/\varepsilon_N] \\ -N \ln \varepsilon_N & \text{for } \zeta > N/\varepsilon_N \,, \end{cases} \tag{A2}$$

and, using the concavity of the ln function and, in particular, for any $\rho \geq 0$

$$N \ln(\varepsilon_N + \rho) \leq N \ln \varepsilon_N + \frac{N}{\varepsilon_N} \rho \,, \tag{A3}$$

it is found that for any $\rho > 0$

$$f^{**}(\rho) = N \ln(\varepsilon_N + \rho) \,. \tag{A4}$$

Therefore, f coincides with its bi-conjugate f^{**}.

To prove strong-duality, we consider the dual problem and we gradually transform it into the primal one:

$$\begin{aligned}
\Delta &= \inf_{\zeta>0} \mathcal{D}(\zeta) \\
&= \inf_{\zeta>0} \left(\inf_{\mathbf{w}} \sup_{\rho \geq 0} \mathcal{L}(\mathbf{w}, \rho, \zeta) \right) \\
&= \inf_{\zeta>0} \inf_{\mathbf{w}} \left(\sup_{\rho \geq 0} \mathcal{L}(\mathbf{w}, \rho, \zeta) \right) \\
&= \frac{1}{2} \inf_{\zeta>0} \inf_{\mathbf{w}} \left(g(\mathbf{w}) + \zeta \mathbf{w}^{\mathrm{T}} \mathbf{w} - \inf_{\rho \geq 0} (\zeta \rho - f(\rho)) \right) \\
&= \frac{1}{2} \inf_{\zeta>0} \inf_{\mathbf{w}} \left(g(\mathbf{w}) + \zeta \mathbf{w}^{\mathrm{T}} \mathbf{w} - f^*(\zeta) \right) \\
&= \frac{1}{2} \inf_{\mathbf{w}} \left(g(\mathbf{w}) + \inf_{\zeta>0} \left(\zeta \mathbf{w}^{\mathrm{T}} \mathbf{w} - f^*(\zeta) \right) \right) \\
&= \frac{1}{2} \inf_{\mathbf{w}} \left(g(\mathbf{w}) + f^{**}(\mathbf{w}^{\mathrm{T}} \mathbf{w}) \right) \\
&= \frac{1}{2} \inf_{\mathbf{w}} \left(g(\mathbf{w}) + f(\mathbf{w}^{\mathrm{T}} \mathbf{w}) \right) \\
&= \inf_{\mathbf{w}} \widetilde{\mathcal{J}}(\mathbf{w}) \\
&= \Pi \,.
\end{aligned} \tag{A5}$$

The key assumptions that make this derivation possible are the following. First the infima of $\mathcal{D}(\zeta)$, and $\widetilde{\mathcal{J}}(\mathbf{w})$ in Eq. (A5) are attained and make the derivation meaningful. Secondly, we used the fact that f defined by Eq. (18) coincides with its bi-conjugate: $f^{**} = f$, as proven above. It can also be checked that the infimum of $\mathcal{D}(\zeta)$ over $\zeta > 0$ is attained in $]0, N/\varepsilon_N]$.

Appendix B

Algorithms of the iterative filters

Algorithm 3 Levenberg-Marquardt IEnKF.

Require: Transition model from t_1 to t_2; $\mathcal{M}$, an observation operator H. Algorithm parameters: ϵ, τ, e_1, e_2, $p_{\max}$. $\mathbf{E}_1^f$, the forecast ensemble at t_1, $\mathbf{y}$ the observation at t_2

Compute $\overline{\mathbf{x}}$ and $\mathbf{A}$ from $\mathbf{E}_1^f$
$p = 0$, $\nu = 2$, $\mathbf{w} = \mathbf{0}$
$\mathbf{x}_1 = \overline{\mathbf{x}} + \mathbf{A}\mathbf{w}$
$\mathbf{y}_2 = H\mathcal{M}(\mathbf{x}_1)$
$\mathbf{T} = \mathbf{I}_N$ (transform)
$\mathbf{E}_1 = \mathbf{x}_1 + \epsilon \mathbf{A}$ (bundle),
$\mathbf{E}_1 = \mathbf{x}_1 + \mathbf{A}\mathbf{T}$ (transform)
$\mathbf{E}_2 = \mathcal{M}(\mathbf{E}_1)$
$\mathbf{Y}_2 = (H(\mathbf{E}_2) - \mathbf{y}_2)/\epsilon$ (bundle),
$\mathbf{Y}_2 = (H(\mathbf{E}_2) - \mathbf{y}_2)\mathbf{T}^{-1}$ (transform)
$\widetilde{\mathcal{J}} = \frac{1}{2}(\mathbf{y} - \mathbf{y}_2)^{\mathrm{T}} \mathbf{R}^{-1} (\mathbf{y} - \mathbf{y}_2) + \frac{N-1}{2} \mathbf{w}^{\mathrm{T}} \mathbf{w}$
$\nabla \widetilde{\mathcal{J}} = (N-1)\mathbf{w} - \mathbf{Y}_2^{\mathrm{T}} \mathbf{R}^{-1} (\mathbf{y} - \mathbf{y}_2)$
$\widetilde{\mathcal{H}} = (N-1)\mathbf{I}_N + \mathbf{Y}_2^{\mathrm{T}} \mathbf{R}^{-1} \mathbf{Y}_2$
flag $= \left(||\nabla \widetilde{\mathcal{J}}||_\infty > e_1 \right)$, $\mu = \tau \max\left(\widetilde{\mathcal{H}}_{kk}\right)$
while flag **and** $p < p_{\max}$ **do**
  $p := p + 1$
  Solve $\left(\widetilde{\mathcal{H}} + \mu \mathbf{I}_N\right) \Delta\mathbf{w} = -\nabla \widetilde{\mathcal{J}}$
  if $||\Delta\mathbf{w}|| \leq e_2$ **then**
    flag = FALSE
  else
    $\mathbf{w}' = \mathbf{w} + \Delta\mathbf{w}$
    $\mathbf{x}_1 = \overline{\mathbf{x}} + \mathbf{A}\mathbf{w}'$
    $\mathbf{y}_2 = H\mathcal{M}(\mathbf{x}_1)$
    $L = \frac{1}{2} \Delta\mathbf{w}^{\mathrm{T}} \left(\mu \Delta\mathbf{w} - \nabla \widetilde{\mathcal{J}}\right)$
    $\widetilde{\mathcal{J}}' = \frac{1}{2}(\mathbf{y} - \mathbf{y}_2)^{\mathrm{T}} \mathbf{R}^{-1} (\mathbf{y} - \mathbf{y}_2) + \frac{N-1}{2} \mathbf{w}'^{\mathrm{T}} \mathbf{w}'$
    $\theta = (\widetilde{\mathcal{J}} - \widetilde{\mathcal{J}}')/L$
    if $\theta > 0$ **then**
      $\widetilde{\mathcal{J}} = \widetilde{\mathcal{J}}'$
      $\mathbf{w} = \mathbf{w}'$
      $\mathbf{E}_1 = \mathbf{x}_1 + \epsilon \mathbf{A}$ (bundle),
      $\mathbf{E}_1 = \mathbf{x}_1 + \mathbf{A}\mathbf{T}$ (transform)
      $\mathbf{E}_2 = \mathcal{M}(\mathbf{E}_1)$
      $\mathbf{Y}_2 = (H(\mathbf{E}_2) - \mathbf{y}_2)/\epsilon$ (bundle),
      $\mathbf{Y}_2 = (H(\mathbf{E}_2) - \mathbf{y}_2)\mathbf{T}^{-1}$ (transform)
      $\nabla \widetilde{\mathcal{J}} = (N-1)\mathbf{w} - \mathbf{Y}_2^{\mathrm{T}} \mathbf{R}^{-1} (\mathbf{y} - \mathbf{y}_2)$
      $\widetilde{\mathcal{H}} = (N-1)\mathbf{I}_N + \mathbf{Y}_2^{\mathrm{T}} \mathbf{R}^{-1} \mathbf{Y}_2$
      $\mathbf{T} = \widetilde{\mathcal{H}}^{-\frac{1}{2}}$ (transform)
      flag $= \left(||\nabla \widetilde{\mathcal{J}}||_\infty > e_1 \right)$
      $\mu := \mu \max\left\{\frac{1}{3}, 1 - (2\theta - 1)^3\right\}$, $\nu = 2$
    else
      $\mu := \mu\,\nu$, $\nu := 2\,\nu$
    end if
  end if
end while
$\mathbf{x}_1 = \overline{x} + \mathbf{A}\mathbf{w}$
$\mathbf{T} = \widetilde{\mathcal{H}}^{-\frac{1}{2}}$ (bundle)
$\mathbf{E}_1 = \mathbf{x}_1 + \mathbf{A}\mathbf{T}$
$\mathbf{E}_2 = \mathcal{M}(\mathbf{E}_1)$
$\mathbf{E}_2 := \overline{\mathbf{x}}_2 + \lambda (\mathbf{E}_2 - \overline{\mathbf{x}}_2)$

Algorithm 4 Levenberg-Marquardt IEnKF-N.

Require: Transition model from t_1 to t_2: $\mathcal{M}$, an observation operator H. Algorithm parameters: ϵ, τ, e_1, e_2, $p_{\max}$. $\mathbf{E}_1^f$, the forecast ensemble at t_1, $\mathbf{y}$ the observation at t_2

Compute $\overline{\mathbf{x}}$ and $\mathbf{A}$ from $\mathbf{E}_1^f$
$p = 0$, $\nu = 2$, $\mathbf{w} = \mathbf{0}$
$\mathbf{x}_1 = \overline{\mathbf{x}} + \mathbf{A}\mathbf{w}$
$\mathbf{y}_2 = H\mathcal{M}(\mathbf{x}_1)$
$\mathbf{T} = \mathbf{I}_N$ (transform)
$\mathbf{E}_1 = \mathbf{x}_1 + \epsilon\mathbf{A}$ (bundle),
$\mathbf{E}_1 = \mathbf{x}_1 + \mathbf{A}\mathbf{T}$ (transform)
$\mathbf{E}_2 = \mathcal{M}(\mathbf{E}_1)$
$\mathbf{Y}_2 = (H(\mathbf{E}_2) - \mathbf{y}_2)/\epsilon$ (bundle),
$\mathbf{Y}_2 = (H(\mathbf{E}_2) - \mathbf{y}_2)\mathbf{T}^{-1}$ (transform)
$\widetilde{\mathcal{J}} = \frac{1}{2}(\mathbf{y} - \mathbf{y}_2)^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{y} - \mathbf{y}_2) + \frac{N}{2}\ln\left(\varepsilon_N + \mathbf{w}^{\mathrm{T}}\mathbf{w}\right)$
$\nabla\widetilde{\mathcal{J}} = N\frac{\mathbf{w}}{\varepsilon_N + \mathbf{w}^{\mathrm{T}}\mathbf{w}} - \mathbf{Y}_2^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{y} - \mathbf{y}_2)$
$\widetilde{\mathcal{H}} = N\frac{(\varepsilon_N + \mathbf{w}^{\mathrm{T}}\mathbf{w})\mathbf{I}_N - 2\mathbf{w}\mathbf{w}^{\mathrm{T}}}{(\varepsilon_N + \mathbf{w}^{\mathrm{T}}\mathbf{w})^2} + \mathbf{Y}_2^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{Y}_2$
flag $= \left(||\nabla\widetilde{\mathcal{J}}||_\infty > e_1\right)$, $\mu = \tau\max\left(\widetilde{\mathcal{H}}_{kk}\right)$
while flag **and** $p < p_{\max}$ **do**
  $p := p + 1$
  Solve $\left(\widetilde{\mathcal{H}} + \mu\mathbf{I}_N\right)\Delta\mathbf{w} = -\nabla\widetilde{\mathcal{J}}$
  if $||\Delta\mathbf{w}|| \leq e_2$ **then**
    flag = FALSE
  else
    $\mathbf{w}' = \mathbf{w} + \Delta\mathbf{w}$
    $\mathbf{x}_1 = \overline{\mathbf{x}} + \mathbf{A}\mathbf{w}$
    $\mathbf{y}_2 = H\mathcal{M}(\mathbf{x}_1)$
    $L = \frac{1}{2}\Delta\mathbf{w}^{\mathrm{T}}\left(\mu\Delta\mathbf{w} - \nabla\widetilde{\mathcal{J}}\right)$
    $\widetilde{\mathcal{J}}' = \frac{1}{2}(\mathbf{y} - \mathbf{y}_2)^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{y} - \mathbf{y}_2) + \frac{N}{2}\ln\left(\varepsilon_N + \mathbf{w}'^{\mathrm{T}}\mathbf{w}'\right)$
    $\theta = (\widetilde{\mathcal{J}} - \widetilde{\mathcal{J}}')/L$
    if $\theta > 0$ **then**
      $\widetilde{\mathcal{J}} = \widetilde{\mathcal{J}}'$
      $\mathbf{w} = \mathbf{w}'$
      $\mathbf{E}_1 = \mathbf{x}_1 + \epsilon\mathbf{A}$ (bundle),
      $\mathbf{E}_1 = \mathbf{x}_1 + \mathbf{A}\mathbf{T}$ (transform)
      $\mathbf{E}_2 = \mathcal{M}(\mathbf{E}_1)$
      $\mathbf{Y}_2 = (H(\mathbf{E}_2) - \mathbf{y}_2)/\epsilon$ (bundle),
      $\mathbf{Y}_2 = (H(\mathbf{E}_2) - \mathbf{y}_2)\mathbf{T}^{-1}$ (transform)
      $\nabla\widetilde{\mathcal{J}} = N\frac{\mathbf{w}}{\varepsilon_N + \mathbf{w}^{\mathrm{T}}\mathbf{w}} - \mathbf{Y}_2^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{y} - \mathbf{y}_2)$
      $\widetilde{\mathcal{H}} = N\frac{(\varepsilon_N + \mathbf{w}^{\mathrm{T}}\mathbf{w})\mathbf{I}_N - 2\mathbf{w}\mathbf{w}^{\mathrm{T}}}{(\varepsilon_N + \mathbf{w}^{\mathrm{T}}\mathbf{w})^2} + \mathbf{Y}_2^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{Y}_2$
      $\mathbf{T} = \widetilde{\mathcal{H}}^{-\frac{1}{2}}$ (transform)
      flag $= \left(||\nabla\widetilde{\mathcal{J}}||_\infty > e_1\right)$
      $\mu := \mu\max\left\{\frac{1}{3}, 1 - (2\theta - 1)^3\right\}$, $\nu = 2$
    else
      $\mu := \mu\,\nu$, $\nu := 2\,\nu$
    end if
  end if
end while
$\mathbf{x}_1 = \overline{x} + \mathbf{A}\mathbf{w}$
$\mathbf{T} = \widetilde{\mathcal{H}}^{-\frac{1}{2}}$ (bundle)
$\mathbf{E}_1 = \mathbf{x}_1 + \mathbf{A}\mathbf{T}$
$\mathbf{E}_2 = \mathcal{M}(\mathbf{E}_1)$

Acknowledgements. The authors thank two anonymous reviewers for their useful comments.

Edited by: J. Duan
Reviewed by: two anonymous referees

References

Anderson, J. L. and Anderson, S. L.: A Monte Carlo Implementation of the Nonlinear Filtering Problem to Produce Ensemble Assimilations and Forecasts, Mon. Weather Rev., 127, 2741–2758, 1999.

Bocquet, M.: Ensemble Kalman filtering without the intrinsic need for inflation, Nonlin. Processes Geophys., 18, 735–750, doi:10.5194/npg-18-735-2011, 2011.

Borwein, J. M. and Lewis, A. S.: Convex analysis and nonlinear optimization: theory and examples, Springer, 2000.

Byrd, R. H., Lu, P., and Nocedal, J.: A Limited Memory Algorithm for Bound Constrained Optimization, J. Sci. Stat. Comput., 16, 1190–1208, 1995.

Gu, Y. and Oliver, D. S.: An Iterative Ensemble Kalman Filter for Multiphase Fluid Flow Data Assimilation, Soc. Petrol. Eng. J., 12, 438–446, 2007.

Harlim, J. and Hunt, B.: A non-Gaussian ensemble filter for assimilating infrequent noisy observations, Tellus A, 59, 225–237, 2007.

Hunt, B. R., Kostelich, E. J., and Szunyogh, I.: Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter, Physica D, 230, 112–126, 2007.

Jazwinski, A. H.: Stochastic Processes and Filtering Theory, Academic Press, 1970.

Kalnay, E. and Yang, S.-C.: Accelerating the spin-up of Ensemble Kalman Filtering, Q. J. Roy. Meteor. Soc., 136, 1644–1651, 2010.

Levenberg, K.: A Method for the Solution of Certain Problems in Least Squares, Quart. Appl. Math., 2, 164–168, 1944.

Lorenz, E. N.: Deterministic nonperiodic flow, J. Atmos. Sci., 20, 130–141, 1963.

Lorenz, E. N. and Emmanuel, K. E.: Optimal sites for supplementary weather observations: simulation with a small model, J. Atmos. Sci., 55, 399–414, 1998.

Madsen, K., Nielsen, H. B., and Tingleff, O.: Methods for nonlinear least square problems, Tech. rep., Informatics and Mathematical Modelling, Technical University of Denmark, 2nd Edn., 2004.

Marquardt, D.: An Algorithm for Least-Squares Estimation of Nonlinear Parameters, SIAM J. Appl. Math., 11, 431–441, 1963.

Nocedal, J. and Wright, S. J.: Numerical Optimization, Springer Series in Operations Research, Springer, 2006.

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

Sakov, P. and Oke, P. R.: Implications of the Form of the Ensemble Transformation in the Ensemble Square Root Filters, Mon. Weather Rev., 136, 1042–1053, 2008.

Sakov, P., Oliver, D., and Bertino, L.: An iterative EnKF for strongly nonlinear systems, Mon. Weather Rev., 140, 1988–2004, doi:10.1175/MWR-D-11-00145.1, 2012.

Tarantola, A.: Inverse Problem Theory and Methods for Model Parameter Estimation, SIAM, 352 pp., 2005.

Wishner, R. P., Tabaczynski, J. A., and Athans, M.: A Comparison of Three Non-Linear Filters, Automatica, 5, 487–496, 1969.

Zupanski, M.: Maximum Likelihood Ensemble Filter: Theoretical Aspects, Mon. Weather Rev., 133, 1710–1726, 2005.