Ocean Dynamics (2004) 54: 539–560
DOI 10.1007/s10236-004-0099-2

Geir Evensen

Sampling strategies and square root analysis schemes for the EnKF

Received: 08. 12. 03 / Accepted: 19. 08. 04

Abstract The purpose of this paper is to examine how different sampling strategies and implementations of the analysis scheme influence the quality of the results in the EnKF. It is shown that by selecting the initial ensemble, the model noise and the measurement perturbations wisely, it is possible to achieve a significant improvement in the EnKF results, using the same number of members in the ensemble. The results are also compared with a square root implementation of the EnKF analysis scheme where the analyzed ensemble is computed without the perturbation of measurements. It is shown that the measurement perturbations introduce sampling errors which can be reduced using improved sampling schemes in the standard EnKF or fully eliminated when the square root analysis algorithm is used. Further, a new computationally efficient square root algorithm is proposed which allows for the use of a low-rank representation of the measurement error covariance matrix. It is shown that this algorithm in fact solves the full problem at a low cost without introducing any new approximations.

Keywords Data assimilation · Ensemble Kalman Filter

1 Introduction

The Ensemble Kalman Filter (EnKF), in its native formulation as originally introduced by Evensen (1994) and Burgers et al. (1998), used pure Monte Carlo sampling when generating the initial ensemble, the model noise and the measurement perturbations. This has been a useful approach since it has made it very easy to interpret and understand the method (see Evensen, 2003). Further, sampling errors can be reduced by an increase of the ensemble size.

Based on the works by Pham (2001) and Nerger (2004) it should be possible to introduce some improvements in the EnKF, by using a more clever sampling for the initial ensemble, the model noise and the measurement perturbations. Further, the works by Anderson (2001), Whitaker and Hamill (2002), Bishop et al. (2001), and see also the review by Tippett et al. (2003), have developed implementations of the analysis scheme where the perturbation of measurements is avoided.

This paper proposes a sampling scheme for the initial ensemble, the measurement perturbations and the model noise, which effectively produces results similar to the sampling used in the SEIK filter by Pham (2001). The scheme does not add significantly to the computational cost of the EnKF, and leads to a very significant improvement in the results. A further improvement can be obtained if the sampling errors introduced by the perturbation of measurements could be removed. Thus, a consistent analysis algorithm, which works without perturbation of measurements, is derived and examined in combination with the improved sampling schemes.

This paper has adopted the same notation as was used in Evensen (2003), and this together with the traditional EnKF analysis scheme is briefly reviewed in the following section. In Section 3 we derive a square root analysis algorithm which computes the analysis without the use of measurement perturbations. Then, in Section 4 we discuss the improved sampling scheme for the EnKF and a simple analysis of the expected impact of improved sampling is presented in Section 5. Several examples which quantify the impact of the improved sampling and the use of square root algorithms are presented in Section 6.

Then in the further discussion, the paper is concerned with the use of low-rank approximatons for the measurement error covariance matrix and efficient implementations of the square root analysis scheme. A

Responsible Editor: Jőrg-Olaf Wolff

G. Evensen
Hydro Research Centre, Bergen, Norway,
and Nansen Environmental and Remote Sensing Center,
Bergen, Norway
e-mail: Geir.Evensen@hydro.com

standard pseudo-inverse is discussed in Section 7.1, while in Section 7.2 we present an analysis of the potential loss of rank that may occur in the case when random measurement perturbations are used to represent the measurement error covariance matrix, as recently discussed by Kepert (2004). It is proved that this loss of rank is easily avoided by a proper sampling of measurement perturbations. Section 7.3 presents a stable algorithm for computing the pseudo-inverse required in the analysis scheme. This forms the basis for the derivation of a very efficient implementation of the square root algorithm in Section 7.4, where the inverse of a matrix of dimension equal to the number of measurements is eliminated. The present algorithm avoids the approximate factorization introduced in Evensen (2003) and solves the full problem. Results from the new algorithm and a variant of it is then presented in Section 8, where we examine the impact of assimilating a large number of measurements together with the use of a low-rank representation for the measurement error covariance matrix.

2 The EnKF

The EnKF is now briefly described with focus on notation and the standard analysis scheme. The notation follows that used in Evensen (2003).

2.1 Ensemble representation for P

As in Evensen (2003), we have defined the matrix holding the ensemble members $\psi_i \in \Re^n$,

$$\boldsymbol{A} = (\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \ldots, \boldsymbol{\psi}_N) \in \Re^{n \times N}, \tag{1}$$

where N is the number of ensemble members and n is the size of the model state vector.

The ensemble mean is stored in each column of $\overline{\boldsymbol{A}}$ which can be defined as

$$\overline{\boldsymbol{A}} = \boldsymbol{A}\boldsymbol{1}_N, \tag{2}$$

where $\boldsymbol{1}_N \in \Re^{N \times N}$ is the matrix where each element is equal to $1/N$. We can then define the ensemble perturbation matrix as

$$\boldsymbol{A}' = \boldsymbol{A} - \overline{\boldsymbol{A}} = \boldsymbol{A}(\boldsymbol{I} - \boldsymbol{1}_N). \tag{3}$$

The ensemble covariance matrix $\boldsymbol{P}_\mathrm{e} \in \Re^{n \times n}$ can be defined as

$$\boldsymbol{P}_\mathrm{e} = \frac{\boldsymbol{A}'(\boldsymbol{A}')^\mathrm{T}}{N-1}. \tag{4}$$

2.2 Measurement perturbations

Given a vector of measurements $\boldsymbol{d} \in \Re^m$, with m being the number of measurements, we can define the N vectors of perturbed observations as

$$\boldsymbol{d}_j = \boldsymbol{d} + \boldsymbol{\epsilon}_j, \quad j = 1, \ldots, N, \tag{5}$$

which can be stored in the columns of a matrix

$$\boldsymbol{D} = (\boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_N) \in \Re^{m \times N}, \tag{6}$$

while the ensemble of perturbations, with ensemble mean equal to zero, can be stored in the matrix

$$\boldsymbol{E} = (\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2, \ldots, \boldsymbol{\epsilon}_N) \in \Re^{m \times N}, \tag{7}$$

from which we can construct the ensemble representation of the measurement error covariance matrix

$$\boldsymbol{R}_\mathrm{e} = \frac{\boldsymbol{E}\boldsymbol{E}^\mathrm{T}}{N-1}. \tag{8}$$

2.3 Analysis equation

The analysis equation, expressed in terms of the ensemble covariance matrices, is

$$\boldsymbol{A}^\mathrm{a} = \boldsymbol{A} + \boldsymbol{P}_\mathrm{e}\boldsymbol{H}^\mathrm{T}(\boldsymbol{H}\boldsymbol{P}_\mathrm{e}\boldsymbol{H}^\mathrm{T} + \boldsymbol{R}_\mathrm{e})^{-1}(\boldsymbol{D} - \boldsymbol{H}\boldsymbol{A}). \tag{9}$$

Using the ensemble of innovation vectors defined as

$$\boldsymbol{D}' = \boldsymbol{D} - \boldsymbol{H}\boldsymbol{A} \tag{10}$$

and the definitions of the ensemble error covariance matrices in Eqs. (4) and (8) the analysis can be expressed as

$$\boldsymbol{A}^\mathrm{a} = \boldsymbol{A} + \boldsymbol{A}'\boldsymbol{A}'^\mathrm{T}\boldsymbol{H}^\mathrm{T}\left(\boldsymbol{H}\boldsymbol{A}'\boldsymbol{A}'^\mathrm{T}\boldsymbol{H}^\mathrm{T} + \boldsymbol{E}\boldsymbol{E}^\mathrm{T}\right)^{-1}\boldsymbol{D}'. \tag{11}$$

When the ensemble size, N, is increased by adding random samples, the analysis computed from this equation will converge towards the exact solution of Eq. (9) with $\boldsymbol{P}_\mathrm{e}$ and $\boldsymbol{R}_\mathrm{e}$ replaced by the exact covariance matrices $\boldsymbol{P}$ and $\boldsymbol{R}$.

We now introduce the matrix holding the measurements of the ensemble perturbations, $\boldsymbol{S} = \boldsymbol{H}\boldsymbol{A}' \in \Re^{m \times N}$, and we define the matrix $\boldsymbol{C} \in \Re^{m \times m}$ as

$$\boldsymbol{C} = \boldsymbol{S}\boldsymbol{S}^\mathrm{T} + (N-1)\boldsymbol{R}, \tag{12}$$

and the ensemble approximation, $\boldsymbol{C}_\mathrm{e}$, of $\boldsymbol{C}$ as

$$\boldsymbol{C}_\mathrm{e} = \boldsymbol{S}\boldsymbol{S}^\mathrm{T} + (N-1)\boldsymbol{R}_\mathrm{e} \tag{13}$$

$$= \boldsymbol{S}\boldsymbol{S}^\mathrm{T} + \boldsymbol{E}\boldsymbol{E}^\mathrm{T}. \tag{14}$$

Thus, we will consider the use of both a full-rank and exact measurement error covariance matrix $\boldsymbol{R}$ and the low-rank representation defined in Eq. (8).

The analysis equation (11) can then be written

$$\boldsymbol{A}^\mathrm{a} = \boldsymbol{A} + \boldsymbol{A}'\boldsymbol{S}^\mathrm{T}\boldsymbol{C}^{-1}\boldsymbol{D}' \tag{15}$$

$$= \boldsymbol{A} + \boldsymbol{A}(\boldsymbol{I} - \boldsymbol{1}_N)\boldsymbol{S}^\mathrm{T}\boldsymbol{C}^{-1}\boldsymbol{D}' \tag{16}$$

$$= \boldsymbol{A}(\boldsymbol{I} + (\boldsymbol{I} - \boldsymbol{1}_N)\boldsymbol{S}^\mathrm{T}\boldsymbol{C}^{-1}\boldsymbol{D}') \tag{17}$$

$$= \boldsymbol{A}(\boldsymbol{I} + \boldsymbol{S}^\mathrm{T}\boldsymbol{C}^{-1}\boldsymbol{D}') \tag{18}$$

$$= \boldsymbol{A}\boldsymbol{X}, \tag{19}$$

where we have used Eq. (3) and $\boldsymbol{1}_N\boldsymbol{S}^\mathrm{T} \equiv \boldsymbol{0}$. The matrix $\boldsymbol{X} \in \Re^{N \times N}$ is defined as

$$\boldsymbol{X} = \boldsymbol{I} + \boldsymbol{S}^\mathrm{T}\boldsymbol{C}^{-1}\boldsymbol{D}'. \tag{20}$$

Thus, the EnKF analysis becomes a combination of the forecast ensemble members and is sought for in the space spanned by the forecast ensemble.

3 A square root algorithm for the EnKF analysis

Several authors have pointed out that the perturbation of measurements used in the EnKF standard analysis equation may be an additional source of errors. Some methods for computing the analysis without introducing measurement noise have recently been presented, e.g. the square root schemes presented by Anderson (2001), Whitaker and Hamill (2002), Bishop et al. (2001) and in the review by Tippett et al. (2003). Based on these results, we have used a simpler and more direct variant of the square root analysis schemes. The perturbation of measurements is avoided and the scheme solves for the analysis without imposing any additional approximations, such as the assumption of uncorrelated measurement errors or knowledge of the inverse of the measurement error covariance matrix. It does require the inverse of the matrix, $\boldsymbol{C}$, but it will later be shown how this can be computed very efficiently using the low-rank $\boldsymbol{C}_{\mathrm{e}}$.

The new algorithm is used to update the ensemble perturbations and is derived starting from the traditional analysis equation for the covariance update in the Kalman Filter (the time index has been dropped for convenience),

$$\boldsymbol{P}^{\mathrm{a}} = \boldsymbol{P}^{\mathrm{f}} - \boldsymbol{P}^{\mathrm{f}}\boldsymbol{H}^{\mathrm{T}}(\boldsymbol{H}\boldsymbol{P}^{\mathrm{f}}\boldsymbol{H}^{\mathrm{T}} + \boldsymbol{R})^{-1}\boldsymbol{H}\boldsymbol{P}^{\mathrm{f}}. \tag{21}$$

When using the ensemble representation for the error covariance matrix, $\boldsymbol{P}$, as defined in Eq. (4), this equation can be written

$$\boldsymbol{A}^{\mathrm{a}\prime}\boldsymbol{A}^{\mathrm{a}\prime\mathrm{T}} = \boldsymbol{A}'(\boldsymbol{I} - \boldsymbol{S}^{\mathrm{T}}\boldsymbol{C}^{-1}\boldsymbol{S})\boldsymbol{A}'^{\mathrm{T}}, \tag{22}$$

where we have used the definitions of $\boldsymbol{S}$ and $\boldsymbol{C}$ from Section 2.

The analyzed ensemble mean is computed from the standard Kalman Filter analysis equation which may be obtained by multiplication of Eq. (15) from the right with $\boldsymbol{1}_N$, i.e. each column in the resulting equation for the mean, becomes

$$\overline{\psi}^{\mathrm{a}} = \overline{\psi}^{\mathrm{f}} + \boldsymbol{A}'\boldsymbol{S}^{\mathrm{T}}\boldsymbol{C}^{-1}(\boldsymbol{d} - \boldsymbol{H}\overline{\psi}^{\mathrm{f}}). \tag{23}$$

In the remainder of the derivation we have dropped the f superscripts.

The following derives an equation for the ensemble analysis by defining a factorization of Eq. (22) where there are no references to the measurements or measurement perturbations.

Note that in the original EnKF the analysis equation (11) was derived by inserting the standard update equation for each ensemble member on the left-hand side of Eq. (22) and showing that if measurements are perturbed this will give a result which is consistent with Eq. (22).

3.1 Derivation of algorithm

We start by forming $\boldsymbol{C}$ as defined in either of Eqs. (12–14). We will also for now assume that $\boldsymbol{C}$ is of full rank such that $\boldsymbol{C}^{-1}$ exists.

We can then compute the eigenvalue decomposition $\boldsymbol{Z}\Lambda\boldsymbol{Z}^{\mathrm{T}} = \boldsymbol{C}$, and we obtain:

$$\boldsymbol{C}^{-1} = \boldsymbol{Z}\Lambda^{-1}\boldsymbol{Z}^{\mathrm{T}}, \tag{24}$$

where all matrices are of dimension $m \times m$. The eigenvalue decomposition may be the most demanding computation required in the analysis when m is large. However, a more efficient algorithm is presented below.

We now write Eq. (22) as follows:

$$\boldsymbol{A}^{\mathrm{a}\prime}\boldsymbol{A}^{\mathrm{a}\prime\mathrm{T}} = \boldsymbol{A}'(\boldsymbol{I} - \boldsymbol{S}^{\mathrm{T}}\boldsymbol{Z}\Lambda^{-1}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{S})\boldsymbol{A}'^{\mathrm{T}} \tag{25}$$

$$= \boldsymbol{A}'\left[\boldsymbol{I} - (\Lambda^{-\frac{1}{2}}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{S})^{\mathrm{T}}(\Lambda^{-\frac{1}{2}}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{S})\right]\boldsymbol{A}'^{\mathrm{T}} \tag{26}$$

$$= \boldsymbol{A}'(\boldsymbol{I} - \boldsymbol{X}_2^{\mathrm{T}}\boldsymbol{X}_2)\boldsymbol{A}'^{\mathrm{T}}, \tag{27}$$

where we have defined $\boldsymbol{X}_2 \in \Re^{m\times N}$ as

$$\boldsymbol{X}_2 = \Lambda^{-\frac{1}{2}}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{S}, \tag{28}$$

where rank $(\boldsymbol{X}_2) = \min(m, N-1)$.

When computing the singular value decomposition[1] (SVD) of $\boldsymbol{X}_2$,

$$\boldsymbol{U}_2\boldsymbol{\Sigma}_2\boldsymbol{V}_2^{\mathrm{T}} = \boldsymbol{X}_2, \tag{29}$$

with $\boldsymbol{U}_2 \in \Re^{m\times m}, \boldsymbol{\Sigma}_2 \in \Re^{m\times N}$ and $\boldsymbol{V}_2 \in \Re^{N\times N}$, Eq. (27) can be written

$$\boldsymbol{A}^{\mathrm{a}\prime}\boldsymbol{A}^{\mathrm{a}\prime\mathrm{T}} = \boldsymbol{A}'(\boldsymbol{I} - [\boldsymbol{U}_2\boldsymbol{\Sigma}_2\boldsymbol{V}_2^{\mathrm{T}}]^{\mathrm{T}}[\boldsymbol{U}_2\boldsymbol{\Sigma}_2\boldsymbol{V}_2^{\mathrm{T}}])\boldsymbol{A}'^{\mathrm{T}} \tag{30}$$

$$= \boldsymbol{A}'(\boldsymbol{I} - \boldsymbol{V}_2\boldsymbol{\Sigma}_2^{\mathrm{T}}\boldsymbol{\Sigma}_2\boldsymbol{V}_2^{\mathrm{T}})\boldsymbol{A}'^{\mathrm{T}} \tag{31}$$

$$= \boldsymbol{A}'\boldsymbol{V}_2(\boldsymbol{I} - \boldsymbol{\Sigma}_2^{\mathrm{T}}\boldsymbol{\Sigma}_2)\boldsymbol{V}_2^{\mathrm{T}}\boldsymbol{A}'^{\mathrm{T}} \tag{32}$$

$$= \left(\boldsymbol{A}'\boldsymbol{V}_2\sqrt{\boldsymbol{I} - \boldsymbol{\Sigma}_2^{\mathrm{T}}\boldsymbol{\Sigma}_2}\right)\left(\boldsymbol{A}'\boldsymbol{V}_2\sqrt{\boldsymbol{I} - \boldsymbol{\Sigma}_2^{\mathrm{T}}\boldsymbol{\Sigma}_2}\right)^{\mathrm{T}}. \tag{33}$$

Thus, a solution for the analysis ensemble perturbations is

$$\boldsymbol{A}^{\mathrm{a}\prime} = \boldsymbol{A}'\boldsymbol{V}_2\sqrt{\boldsymbol{I} - \boldsymbol{\Sigma}_2^{\mathrm{T}}\boldsymbol{\Sigma}_2\Theta^{\mathrm{T}}}. \tag{34}$$

Note that the additional multiplication with a random orthogonal matrix Θ^{T} also results in a valid solution. Such a random redistribution of the variance reduction among the ensemble memebers, is in some cases necessary and should be used by default. The matrix Θ^{T} is easily constructed, e.g., by using the right singular vectors from a singular value decomposition of a random $N \times N$ matrix.

[1] The singular value decomposition of a rectangular matrix $\boldsymbol{A} \in \Re^{m\times n}$ is $\boldsymbol{A} = \boldsymbol{U}\Sigma\boldsymbol{V}^{\mathrm{T}}$, where $\boldsymbol{U} \in \Re^{m\times m}$ and $\boldsymbol{V} \in \Re^{n\times n}$ are orthogonal matrices and $\Sigma \in \Re^{m\times n}$ contains the $p = \min(m, n)$ singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0$ on the diagonal. Further, $\boldsymbol{U}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{V} = \Sigma$. Note that numerical algorithms for computing the SVD when $m > n$ often offer to compute only the first p singular vectors in $\boldsymbol{U}$ since the remaining singular vectors (columns in $\boldsymbol{U}$) are normally not needed. However, for the expression $\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}} = \boldsymbol{I}$ to be true the full $\boldsymbol{U}$ must be used.

3.2 Implementation of algorithm

The algorithm requires the following steps:

1. Form $\boldsymbol{C}$ and compute the eigenvalue decomposition: $\boldsymbol{Z}\Lambda\boldsymbol{Z}^{\mathrm{T}} = \boldsymbol{C}$.
2. Update the ensemble mean from the equation

$$\overline{\boldsymbol{\psi}}^{\mathrm{a}} = \overline{\boldsymbol{\psi}}^{\mathrm{f}} + A'\boldsymbol{S}^{\mathrm{T}}\boldsymbol{Z}\Lambda^{-1}\boldsymbol{Z}^{\mathrm{T}}(\boldsymbol{d} - \boldsymbol{H}\overline{\boldsymbol{\psi}}^{\mathrm{f}}), \tag{35}$$

using the following sequence of matrix-vector multiplications:

a) $\boldsymbol{y}_1 = \boldsymbol{Z}^{\mathrm{T}}(\boldsymbol{d} - \boldsymbol{H}\overline{\boldsymbol{\psi}}^{\mathrm{f}})$,
b) $\boldsymbol{y}_2 = \Lambda^{-1}\boldsymbol{y}_1$,
c) $\boldsymbol{y}_3 = \boldsymbol{Z}\boldsymbol{y}_2$,
d) $\boldsymbol{y}_4 = \boldsymbol{S}^{\mathrm{T}}\boldsymbol{y}_3$,
e) $\overline{\boldsymbol{\psi}}^{\mathrm{a}} = \overline{\boldsymbol{\psi}}^{\mathrm{f}} + A'\boldsymbol{y}_4$.

3. Evaluate the matrix: $\boldsymbol{X}_2 = \Lambda^{-\frac{1}{2}}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{S}$.
4. Compute the SVD: $\boldsymbol{U}_2\boldsymbol{\Sigma}_2\boldsymbol{V}_2^{\mathrm{T}} = \boldsymbol{X}_2$.
5. Then evaluate the analyzed ensemble perturbations from $A^{\mathrm{a}\prime} = A'\boldsymbol{V}_2\sqrt{\boldsymbol{I} - \boldsymbol{\Sigma}_2^{\mathrm{T}}\boldsymbol{\Sigma}_2\Theta^{\mathrm{T}}}$ and add the ensemble mean, $\overline{\boldsymbol{\psi}}^{\mathrm{a}}$, to the analyzed perturbations.

In Appendix B it is shown that the analysis ensemble resulting from the square root algorithm still becomes a combination of the forecast ensemble as was discussed in Evensen (2003).

4 An improved sampling scheme

We start by defining an error covariance matrix $\boldsymbol{P}$. We can assume this to be the initial error covariance matrix for the model state. Given $\boldsymbol{P}$ we can compute the eigenvalue decomposition

$$\boldsymbol{P} = \boldsymbol{Z}\Lambda\boldsymbol{Z}^{\mathrm{T}}, \tag{36}$$

where the matrices $\boldsymbol{Z}$ and Λ contain the eigenvectors and eigenvalues of $\boldsymbol{P}$. In the SEIK filter by Pham (2001) an algorithm was used where the initial ensemble was sampled from the first dominant eigenvectors of $\boldsymbol{P}$. This introduces a maximum rank and conditioning of the ensemble matrix and also ensures that the ensemble, best possibly represents the error covariance matrix for a given ensemble size. In other words, we want to generate $\boldsymbol{A}$ such that rank $(\boldsymbol{A}) = N$ and the condition number defined as the ratio between the singular values, $\kappa_2(\boldsymbol{A}) = \sigma_1(\boldsymbol{A})/\sigma_N(\boldsymbol{A})$, is minimal. If the ensemble members stored in the columns of $\boldsymbol{A}$ are nearly dependent, then $\kappa_2(\boldsymbol{A})$ is large.

Now, approximate the error covariance matrix with its ensemble representation, $\boldsymbol{P}_{\mathrm{e}} \simeq \boldsymbol{P}$. We can then write

$$\boldsymbol{P}_{\mathrm{e}} = \frac{1}{N-1}A'(A')^{\mathrm{T}} \tag{37}$$

$$= \frac{1}{N-1}\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{U}^{\mathrm{T}} \tag{38}$$

$$= \frac{1}{N-1}\boldsymbol{U}\boldsymbol{\Sigma}^2\boldsymbol{U}^{\mathrm{T}}, \tag{39}$$

where A' contains the ensemble perturbations as defined in Eq. (3), $\boldsymbol{U}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{V}^{\mathrm{T}}$ result from a singular value decomposition and contain the singular vectors and singular values of A'. In the limit when the ensemble size goes to infinity the n singular vectors in $\boldsymbol{U}$ will converge towards the n eigenvectors in $\boldsymbol{Z}$ and the square of the singular values, $\boldsymbol{\Sigma}^2$, divided by $N-1$, will converge towards the eigenvalues, Λ.

This shows that there are two strategies for defining an accurate approximation, $\boldsymbol{P}_{\mathrm{e}}$, of $\boldsymbol{P}$.

1. We can increase the ensemble size, N, by sampling additional model states and adding these to the ensemble. As long as the addition of new ensemble members increases the space spanned by the overall ensemble, this will result in an ensemble covariance, $\boldsymbol{P}_{\mathrm{e}}$, which is a more accurate representation of $\boldsymbol{P}$.
2. Alternatively we can improve the rank/conditioning of the ensemble by ensuring that the first N singular vectors in $\boldsymbol{U}$ are similar to the N first eigenvectors in $\boldsymbol{Z}$. Thus, the absolute error in the representation $\boldsymbol{P}_{\mathrm{e}}$ of $\boldsymbol{P}$ will be smaller for ensembles generated with such an improved sampling than for Monte Carlo ensembles of a given moderate ensemble size.

This first approach is the standard Monte Carlo method used in the traditional EnKF where the convergence is slow. The second approach has a flavour of quasirandom sampling, which ensures much better convergence with increasing sample size, i.e. we choose ensemble members which have less linear dependence and therefore span a larger space. These two strategies are, of course, used in combination when the initial ensemble is created.

For most applications the size of $\boldsymbol{P}$ is too large to allow for the direct computation of eigenvectors. An alternative algorithm for generating an N member ensemble with better conditioning is to first generate a start ensemble which is larger than N, and then to resample N members along the first N dominant singular vectors of this larger start ensemble.

The algorithm goes as follows. First, sample a large ensemble of model states with, e.g., β times N members, and store the ensemble perturbations in $\hat{A}' \in \Re^{n\times\beta N}$. Then perform the following steps:

1. Compute the SVD, $\hat{A}' = \hat{\boldsymbol{U}}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{V}}^{\mathrm{T}}$, where the columns of $\hat{\boldsymbol{U}}$ are the singular vectors and the diagonal of $\hat{\boldsymbol{\Sigma}}$ contains the singular values σ_i (note that with a multivariate state it may be necessary to scale the variables in $\hat{A}'$ first).
2. Retain only the first $N \times N$ quadrant of $\hat{\boldsymbol{\Sigma}}$ which is stored in $\boldsymbol{\Sigma} \in \Re^{N\times N}$, i.e. $\sigma_i = 0, \forall i > N$.
3. Scale the non-zero singular values with $\sqrt{\beta}$ (needed to retain the correct variance in the new ensemble).
4. Generate a random orthogonal matrix $\boldsymbol{V}_1^{\mathrm{T}} \in \Re^{N\times N}$, (could be the matrix $\boldsymbol{V}_1^{\mathrm{T}}$ from a SVD of a random $N \times N$ matrix, $\boldsymbol{Y} = \boldsymbol{U}_1\boldsymbol{\Sigma}_1\boldsymbol{V}_1^{\mathrm{T}}$, which is computed very quickly).
5. Generate an N-member ensemble using only the first N singular vectors in $\hat{\boldsymbol{U}}$ (stored in $\boldsymbol{U}$), the non-zero

singular values stored in Σ and the orthogonal matrix $\boldsymbol{V}_1^{\mathrm{T}}$.

Thus, we are using the formula

$$\boldsymbol{A}' = \boldsymbol{U}\frac{1}{\sqrt{\beta}}\Sigma\boldsymbol{V}_1^{\mathrm{T}}. \tag{40}$$

When the size of the start ensemble approaches infinity, the singular vectors will converge towards the eigenvectors of $\boldsymbol{P}$. It is, of course, assumed that the ensemble perturbations are sampled with the correct covariance as given by $\boldsymbol{P}$. As long as the initial ensemble is chosen large enough, this algorithm will provide an ensemble which is similar to what is used in the SEIK filter, and the SVD algorithm has a lower computational cost than the explicit eigenvalue decomposition of $\boldsymbol{P}$ when n is large.

Before the ensemble perturbation matrix, $\boldsymbol{A}'$, is used, it is important to ensure that the mean is zero and the variance is as specified. This can be done by subtracting an eventual ensemble mean and then rescaling the ensemble members to obtain the correct variance. As will be seen below, this has a positive impact on the quality of the results. Note that the removal of the mean of the ensemble leaves the maximum possible rank of $\boldsymbol{A}'$ to be $N-1$.

As an example, a 100-member ensemble has been generated using start ensembles of $100, 200 \ldots 800$ members. The size of the one-dimensional model state is 1001 and the characteristic length scale of the solution is four grid cells. The singular values (normalized to the first singular value) for the resulting ensemble is plotted in Fig. 1 for the different sizes of start ensemble. Clearly, there is a benefit in using this sampling strategy. The ratio between singular values 100 and 1 is 0.21 when standard sampling is used. With increasing size of the start ensemble the conditioning improves until it reaches 0.59 for 800 members in the start ensemble.

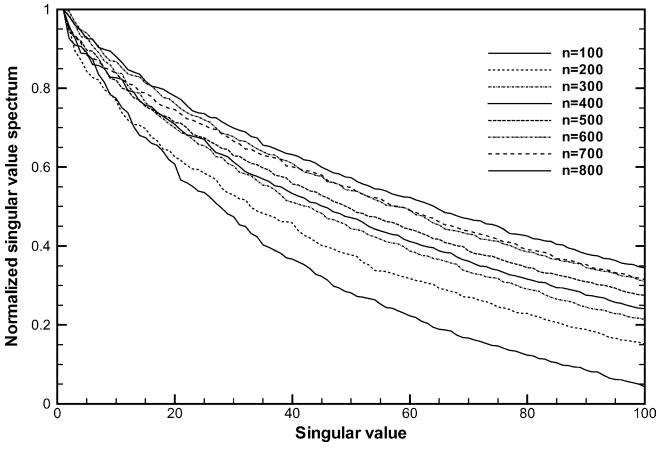

Fig. 1 The plot shows the normalized singular values of ensembles which are generated using start ensembles of different sizes. Clearly, the condition of the ensemble improves when a larger start ensemble is used

5 Impact of improved sampling

The following discussion provides some background on the relative impact and importance of the use of improved sampling for the initial ensemble, the model noise and the measurement perturbations.

5.1 Preservation of rank during model integration

Assume a linear model operator is defined by the full rank matrix $\boldsymbol{F}$. With zero model noise the ensemble at a later time, t_k, can then be written as

$$\boldsymbol{A}_k = \boldsymbol{F}^k\boldsymbol{A}_0. \tag{41}$$

Thus, the rank introduced in the initial ensemble will be preserved as long as $\boldsymbol{F}$ is full-rank, and $\boldsymbol{A}_k$ will span the same space as $\boldsymbol{A}_0$.

With system noise the time evolution of the ensemble becomes

$$\boldsymbol{A}_k = \boldsymbol{F}^k\boldsymbol{A}_0 + \sum_{i=1}^{k}\boldsymbol{F}^{k-i}\boldsymbol{Q}_i, \tag{42}$$

where $\boldsymbol{Q}_i$ denote the ensemble of model noise used at time t_i. Thus, the rank and conditioning of the ensemble will also depend on the rank and conditioning of the model noise introduced.

For a non-linear model operator, $\boldsymbol{f}(\boldsymbol{\psi}, \boldsymbol{q})$, where $\boldsymbol{q}$ is the model noise, the evolution of the ensemble can be written as

$$\boldsymbol{A}_k = \boldsymbol{f}_k(\ldots\boldsymbol{f}_2(\boldsymbol{f}_1(\boldsymbol{A}_0, \boldsymbol{Q}_1), \boldsymbol{Q}_2)\ldots\boldsymbol{Q}_k). \tag{43}$$

Using a non-linear model there is no guarantee that the non-linear transformations will preserve the rank of $\boldsymbol{A}$, and the introduction of wisely sampled model noise may be crucial to maintain an ensemble with good rank properties during the simulation. Thus, the same procedure as used when generating the initial ensemble should be used when simulating the system noise. This will ensure that a maximum rank is introduced into the ensemble, and this may also counteract any rank reduction introduced by the model operator.

5.2 The EnKF with a linear perfect model

Let us examine the EnKF with a linear model with no model errors. Given the initial ensemble stored in $\boldsymbol{A}_0$ and assume a finite number of time instances, distributed at regular time intervals, where the forecasted ensemble is stored and measurements are assimilated.

An ensemble forecast at time t_k is then expressed by Eq. (41).

If the EnKF is used to update the solution at every time t_i, where $i = 1, k$, the ensemble solution at time t_k becomes

$$\boldsymbol{A}_k = \boldsymbol{F}^k\boldsymbol{A}_0\prod_{i=1}^{k}\boldsymbol{X}_i, \tag{44}$$

where $\boldsymbol{X}_i$ is the matrix defined by Eq. (20) which, when multiplied with the ensemble forecast matrix at time t_i, produces the analysis ensemble at that time (i.e. the $\boldsymbol{X}_5$ matrix of Evensen 2003). Thus, starting with $\boldsymbol{A}_0$, the assimilation solution at time t_1 is obtained by multiplication of $\boldsymbol{F}$ with $\boldsymbol{A}_0$ to produce the forecast at time t_1 followed by the multiplication of the forecast with $\boldsymbol{X}_1$.

Note that the expression $\boldsymbol{A}_0 \prod_{i=1}^{k} \boldsymbol{X}_i$ is the smoother (EnKS) solution at time t_0. Thus, for this model, Eq. (44) can also be interpreted as a forward integration of the smoother solution from the initial time, t_0, until t_k, where $\boldsymbol{A}_k$ is produced.

This also means that for a linear model in the case without model errors, the EnKF solution at all times is a combination of the initial ensemble members, and the dimension of the affine space spanned by the initial ensemble does not change with time as long as the operators $\boldsymbol{F}$ and $\boldsymbol{X}_i$'s are full-rank. Thus, the quality of the EnKF solution is dependent on the rank and conditioning of the initial ensemble matrix.

5.3 Generation of measurement perturbations

When the EnKF analysis algorithm in Eq. (15) with measurement perturbations is used, then the improved sampling procedure should also be used when generating the perturbations. This will lead to a better conditioning of the ensemble of perturbations, which will have an ensemble representation $\boldsymbol{R}_e$, which is closer to $\boldsymbol{R}$. The impact of improved sampling of measurement perturbations is significant and will be demonstrated in the examples below.

6 Experiments

The impact of the improved sampling scheme and the use of an analysis scheme where measurements are not perturbed will now be examined in some detail.

6.1 Model description and initialization

The model used in the following experiments is a one-dimensional linear advection model on a periodic domain of length 1000. The model has a constant advection speed, $u = 1.0$, the grid spacing is $\Delta x = 1.0$ and the time step is $\Delta t = 1.0$. Given an initial condition, the solution of this model is exactly known, and this allows us to run realistic experiments with zero model errors to better examine the impact of the choice of initial ensemble in relation to the choice of analysis scheme.

The true initial state is sampled from a distribution, $\mathcal{N}$, with mean equal to 0, variance equal to 1, and a spatial decorrelation length of 20. Thus, it is a smooth periodic pseudorandom solution consisting of a superposition of waves with different wave lengths, where the shorter waves are penalized, and where each wave has a random phase.

The first guess solution is generated by drawing another sample from $\mathcal{N}$ and adding this to the true state. The initial ensemble is then generated by adding samples drawn from $\mathcal{N}$ to the first guess solution. Thus, the initial state is assumed to have error variance equal to 1.

Four measurements of the true solution, distributed regularly in the model domain, are assimilated every fifth time step. The measurements are contaminated by errors of variance equal to 0.01, and we have assumed uncorrelated measurement errors.

The length of the integration is 300, which is 50 time units longer than the time needed for the solution to advect from one measurement to the next (i.e. 250 time units).

In most of the following experiments an ensemble size of 100 members is used. A larger start ensemble is used in many of the experiments to generate ensemble members and/or measurement perturbations which provide a better representation of the error covariance matrix. Otherwise, the experiments differ in the sampling of measurement perturbations and the analysis scheme used.

In Fig. 2 an example is shown from one of the experiments. The plots illustrate the convergence of the estimated solution at various times during the experiment, and show how information from measurements is propagated with the advection speed and how the error variance is reduced.

6.2 Analysis schemes

Two versions of the analysis scheme are available:

Analysis 1: This is the standard EnKF analysis solving Eq. (15), using perturbation of measurements.

Analysis 2: This is a so-called square root implementation of the analysis scheme where the perturbation of measurements is avoided. The algorithm is described in Section 3.

Both algorithms form a full-rank matrix $\boldsymbol{C} = \boldsymbol{S}\boldsymbol{S}^{\mathrm{T}} + (N-1)\boldsymbol{R}$ and then factorize it by computing the eigenvalue decomposition. In cases with many measurements, the computational cost becomes large since Nm^2 operations are required to form the matrix and the eigenvalue decomposition requires $\mathcal{O}(m^3)$ operations. An alternative inversion algorithm which reduces the factorization of the $m \times m$ matrix to a factorization of an $N \times N$ matrix is presented in Section 7.3.

6.3 Overview of experiments

Several experiments have been carried out as listed in Table 1. For each of the experiments, 50 EnKF simulations were performed to allow for a statistical comparison. In each simulation, the only difference is the random

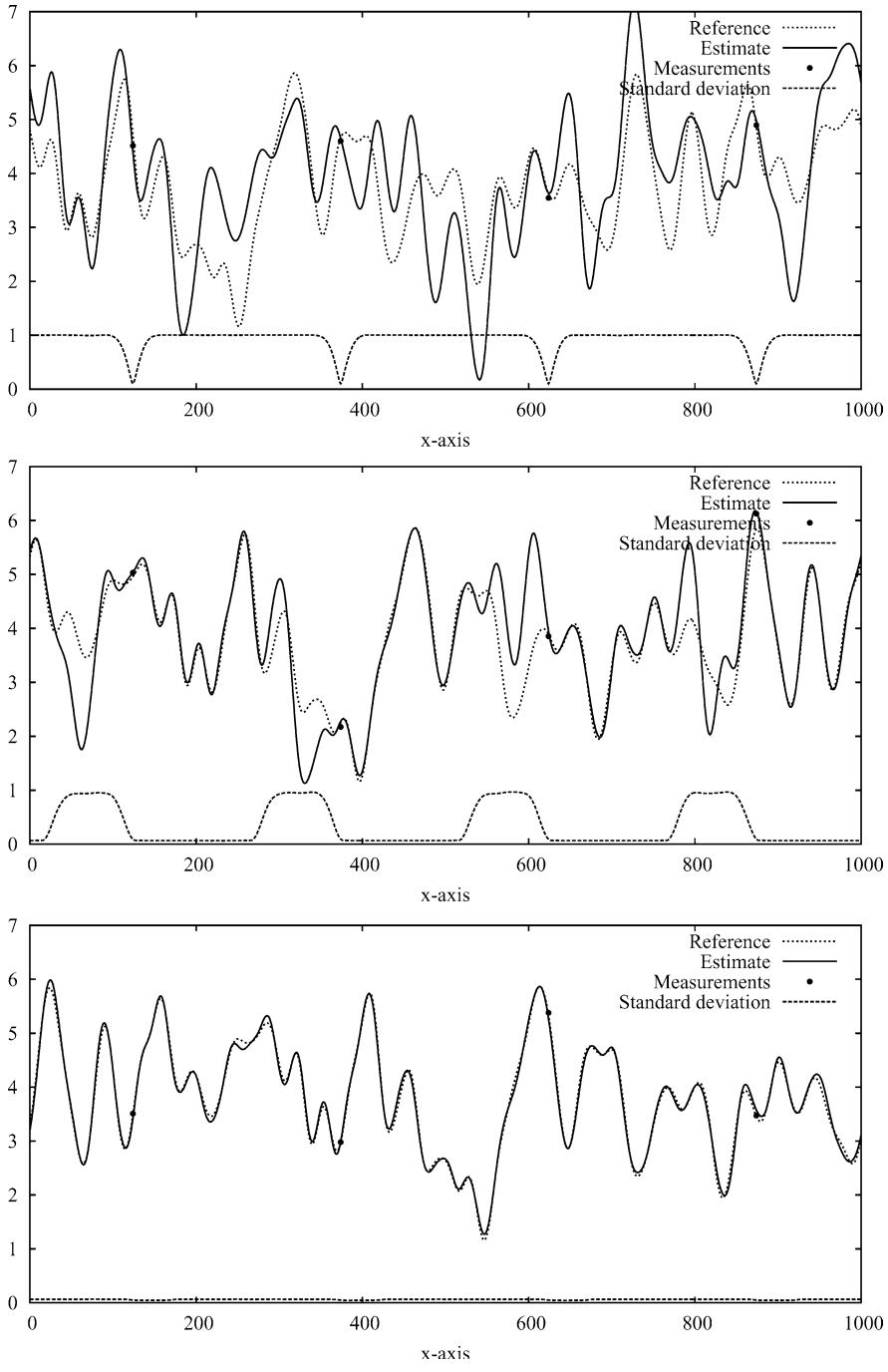

Fig. 2 Example of an EnKF experiment: reference solution, measurements, estimate and standard deviation at three different times $t = 5.0$ (*top*), $t = 150.0$ (*centre*), and $t = 300.0$ (*bottom*)

seed used. Thus, every simulation will have a different and random true state, first guess, initial ensemble, set of measurements and measurement perturbations.

In all the experiments the residuals were computed as the root mean square (RMS) errors of the difference between the estimate and the true solution taken over the complete space and time domain. For each of the experiments we plotted the mean and standard deviation of the residuals in Fig. 3.

Table 2 gives the probabilities that the average residuals from the experiments are equal, as computed from the Student's *t*-Test. Probabilities lower than, say, 0.5, indicate statistically that the distributions from two experiments are significantly different. When selecting a method or approach, one should use the one which has a distribution with the lowest average residual, and possibly also a low variance of the residuals.

It is also of interest to examine how well the predicted errors represent the actual residuals (RMS as a function of time). In Figs. 4 and 5 we have plotted the average of the predicted errors from the 50 simulations as the thick full line. The thin full lines indicate the one standard deviation spread of the predicted errors from the 50 simulations. The average of the RMS errors from the 50 simulations is plotted as the thick dotted line, with associated one standard deviation spread shown by the dotted thin lines.

Table 1 Summary of experiments. The *first column* is the experiment name, in the *second column N* is the ensemble size used, then Sample fix is true or false and indicates if the sample mean and variance is corrected, β_{ini} is a number which defines the size of the start ensemble used for generating the initial ensemble as $\beta_{\mathrm{ini}}N$, similarly β_{mes} denote the size of the start ensemble used for generating the measurement perturbations, followed by the analysis algorithm used. The last two columns contain the average RMS errors of the 50 simulations in each experiment and the standard deviation of these

Experiment	N	Sample fix	β_{ini}	β_{mes}	Analysis	Residual	Std. dev.
A	100	F	1	1	1	0.762	0.074
B	100	T	1	1	1	0.759	0.053
C	100	T	2	1	1	0.715	0.065
D	100	T	4	1	1	0.683	0.062
E	100	T	6	1	1	0.679	0.071
F	100	T	1		2	0.688	0.060
G	100	T	6		2	0.586	0.046
H	100	T	6	30	1	0.627	0.053
I	100	T	1	30	1	0.706	0.060
B150	150	T	1	1	1	0.681	0.053
B200	200	T	1	1	1	0.651	0.061
B250	250	T	1	1	1	0.626	0.058
G50	50	T	12		2	0.784	0.088
G52	52	T	12		2	0.738	0.072
G55	55	T	11		2	0.693	0.068
G60	60	T	10		2	0.643	0.069
G75	75	T	8		2	0.604	0.042

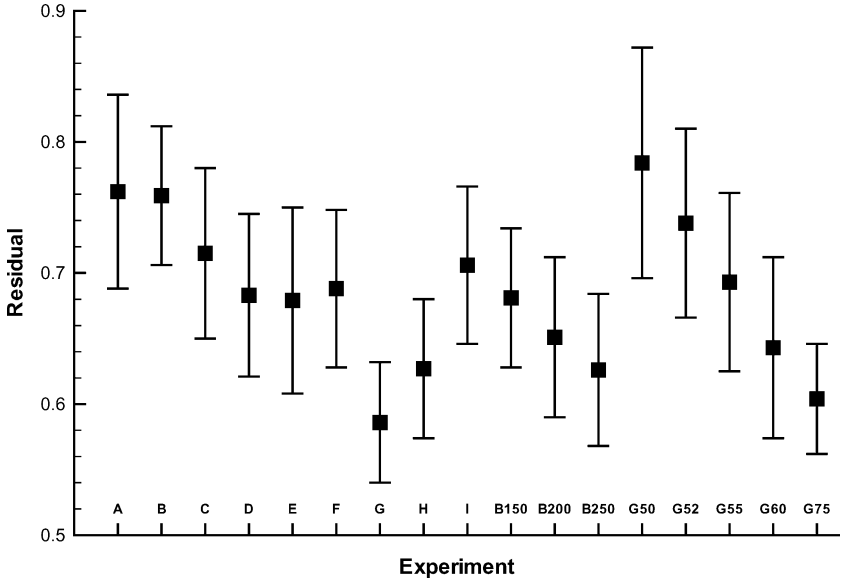

Fig. 3 Mean and Standard deviation of the residuals from each of the experiments

Table 2 Statistical probability that two experiments provide an equal mean for the residuals as computed using the Student's *t*-Test. A probability close to one indicates that it is likely that the two experiments provide distributions of residuals with similar mean. The *t*-test numbers higher than 0.5 are printed *in bold*

Exp	B	B150	B200	B250	C	D	E	F	G	G50	G52	G55	G60	G75	H	I
A	**0.86**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.17	0.11	0.00	0.00	0.00	0.00	0.00
B		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.09	0.10	0.00	0.00	0.00	0.00	0.00
B150			0.01	0.00	0.01	**0.86**	**0.86**	**0.53**	0.00	0.00	0.00	0.34	0.00	0.00	0.00	0.03
B200				0.04	0.00	0.01	0.04	0.00	0.00	0.00	0.00	0.00	**0.52**	0.00	0.04	0.00
B250					0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.19	0.03	**0.91**	0.00
C						0.01	0.01	0.03	0.00	0.00	0.10	0.10	0.00	0.00	0.00	0.48
D							**0.75**	**0.68**	0.00	0.00	0.00	0.45	0.00	0.00	0.00	0.06
E								0.48	0.00	0.00	0.00	0.32	0.01	0.00	0.00	0.04
F									0.00	0.00	0.00	**0.72**	0.00	0.00	0.00	0.13
G										0.00	0.00	0.00	0.00	0.04	0.00	0.00
G50											0.01	0.00	0.00	0.00	0.00	0.00
G52												0.00	0.00	0.00	0.00	0.02
G55													0.00	0.00	0.00	0.29
G60														0.00	0.21	0.00
G75															0.02	0.00
H																0.00

The further details of the different experiments are described below.

Exp. A is the pure Monte Carlo case using a start ensemble of 100 members where all random variables are sampled "randomly". Thus, the mean and variance of the initial ensemble and the measurement perturbations will fluctuate within the accuracy that can be expected using a 100-member sample size. The analysis is computed using the standard EnKF Analysis 1 algorithm.

Exp. B is similar to Exp. A except that the sampled ensemble perturbations are corrected to have mean zero and the correct specified variance. This is done by subtracting an eventual mean from the random sample and then dividing the members by the square root of the ensemble variance. As will be seen below, this leads to a small improvement in the assimilation results and this correction is therefore used in all the following experiments. This experiment is used as a reference case in the further discussion which illustrates the performance of the standard EnKF Analysis 1 algorithm.

Exps. C, D and E are similar to Exp. B except that the start ensembles used to generate the initial 100-member ensemble contain respectively 200, 400 and 600 members.

Exp. E is used as a reference case illustrating the impact of the improved initial sampling algorithm.

Exp. F uses the square root algorithm implemented as Analysis 2, where the perturbation of measurements is avoided and a start ensemble of 100 members is used as in Exp. B.

Exp. G uses the square root algorithm as in Exp. F except that the initial ensemble is sampled from a start ensemble of 600 members as in Exp. E. It examines the benefit of combined use of improved initial sampling and the square root algorithm.

Exp. H examines the impact of improved sampling of measurement perturbations using the standard EnKF Analysis 1 algorithm, but is otherwise similar to Exp. E with improved sampling of the initial ensemble.

Exp. I is similar to Exp. H with improved sampling of measurement perturbations and using the EnKF Analysis 1 algorithm but, as in Exp. B, it does not use the improved sampling of initial conditions.

Exps. B150, B200 and B250 are similar to Exp. B but using respectively ensemble sizes of 150, 200 and 250 members.

Exps. G50, G52, G55, G60 and G75 are similar to Exp. G but using, respectively, ensemble sizes of 50, 52, 55, 60 and 75 members.

6.4 Impact of improved sampling for the initial ensemble

Using the procedure outlined in Section 4, several experiments have been performed using start ensembles of 100–600 members to examine the impact of using an initial ensemble with better properties.

As can be seen from Fig. 3, the pure Monte Carlo Exp. A has the poorest performance among the Exps. A–E. Starting with Exp. A, we find a very small improvement when we apply the sample correction (correcting for ensemble mean and variance) in Exp. B. This correction may become more important if smaller ensembles are used.

In the Exps. C, D and E, larger start ensembles of respectively 200, 400 and 600 members are used to generate the initial 100 member ensemble. Just doubling the size of the start ensemble to 200 members (Exp. C) has a significant positive effect on the results, and using a start ensemble of 400 members (Exp. D) leads to a further improvement.

The use of an even larger start ensemble of 600 members (Exp. E) does not provide a statistically significant improvement over Exp. D in this particular application, although this may become different if a more complex model with a larger state space is used.

When comparing the time evolution of the residuals and the estimated standard deviations for the Exps. B–E in Fig. 4, we observe that the residuals show a larger spread between the simulations than the estimated standard deviations. The estimated standard deviations are internally consistent between the simulations performed in each of the experiments. The residuals are also generally larger than the ensemble standard deviations, although there is a slight improvement observed due to the improved sampling of the initial ensemble. It will become visible below that this discrepancy is to a large extent caused by the perturbation of measurements in the analysis scheme, although the initial sampling also leads to some improvement.

These experiments clearly show that the improved sampling is justified for the initial ensemble. It is computed only once and the additional computational cost is marginal.

6.5 Impact of square root analysis algorithms

It has been pointed out by several authors that the perturbation of measurements introduces an additional source of sampling errors in the results. In Section 3 we have presented the algorithm, Analysis 2, which computes the analysis without the use of perturbed measurements.

Let us use the Exps. B and E discussed above as the EnKF reference cases with and without improved sampling for the initial ensemble. Then we run two experiments, Exp. F and Exp. G, using the square root algorithm implemented in Analysis 2. Exp. F uses standard sampling of 100 members as in Exp. B while Exp. G uses a 600-member start ensemble as in Exp. E.

Referring again to the residuals plotted in Fig. 3, it is clear that Exp. F provides a significant improvement compared to Exp. B, which uses the standard EnKF analysis algorithm. The improvement is similar to what was found using the improved sampling in Exp. E. The

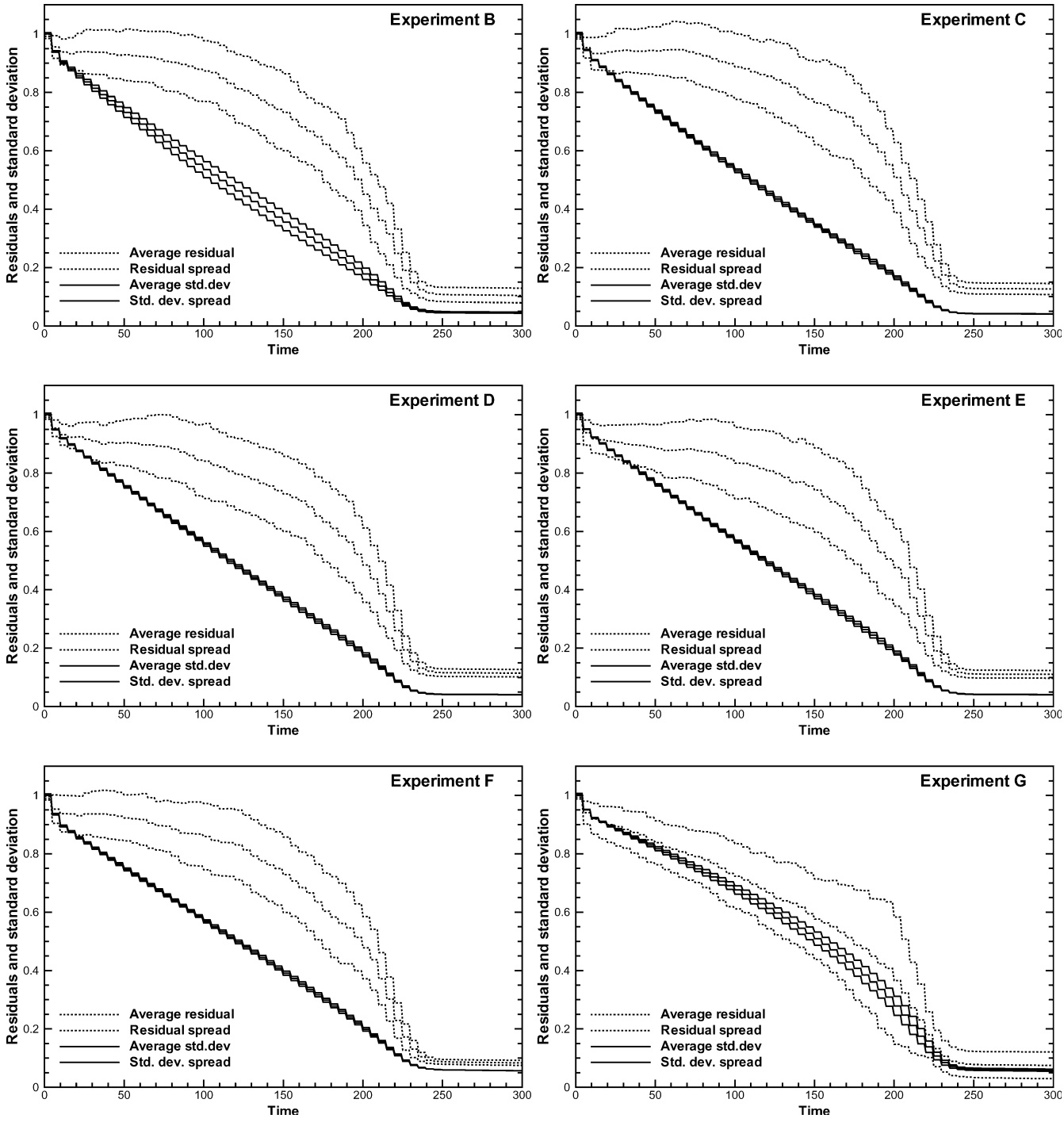

Fig. 4 Time evolution for RMS residuals (*dashed lines*) and estimated standard deviations (*full lines*). The *thick lines* show the means over the 50 simulations and the *thin lines* show the means plus/minus one standard deviation

combined use of improved sampling of the initial ensemble and the square root algorithm is illustrated in Exp. G and this leads to a significant further improvement in the results.

The time evolution of the residuals in Exp. G, plotted in Fig. 4, shows a fairly good consistency with the estimated standard deviations. This must be attributed to the elimination of the measurement perturbations and the improved sampling used for the initial ensemble. There is now only a minor underestimation of the predicted errors compared with the actual residuals. Clearly, we can only explain total error projected onto the ensemble space and a larger ensemble should lead to even better consistency.

In Exp. F, where the standard sampling was used, the residuals show similar behaviour as in Exp. E.

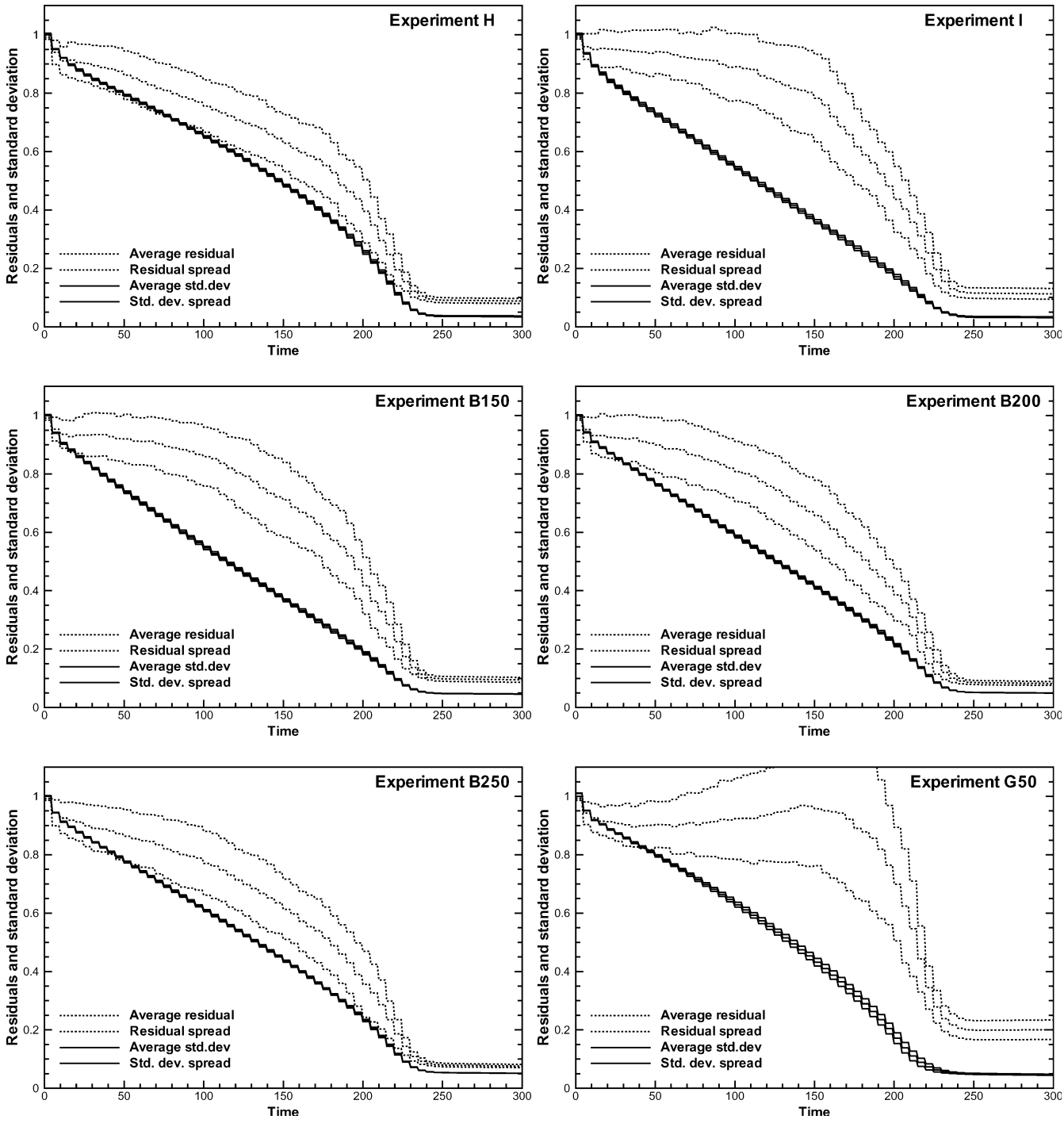

Fig. 5 Time evolution for RMS residuals (*dashed lines*) and estimated standard deviations (*full lines*). The *thick lines* show the means over the 50 simulations and the *thin lines* show the means plus/minus one standard deviation

6.6 Improved sampling of measurement perturbations

We have shown that the square root algorithm, which avoids the perturbation of measurements, provides a significant improvement in the results in the EnKF (Exps. F and G). However, it is also possible to improve the sampling of the measurement perturbations using the same algorithm as was used for the initial ensemble, and this should lead to results closer to those obtained in Exps. F and G, when using the standard EnKF analysis algorithms.

The Exps. H and I use the improved sampling of measurement perturbations with a large start ensemble of perturbations (30 times the ensemble size) and the solution is found using Analysis 1. The impact of this improved sampling is illustrated by comparing the Exp. I with Exps. B and F, and then Exp. H with Exps. E and G, in Fig. 3.

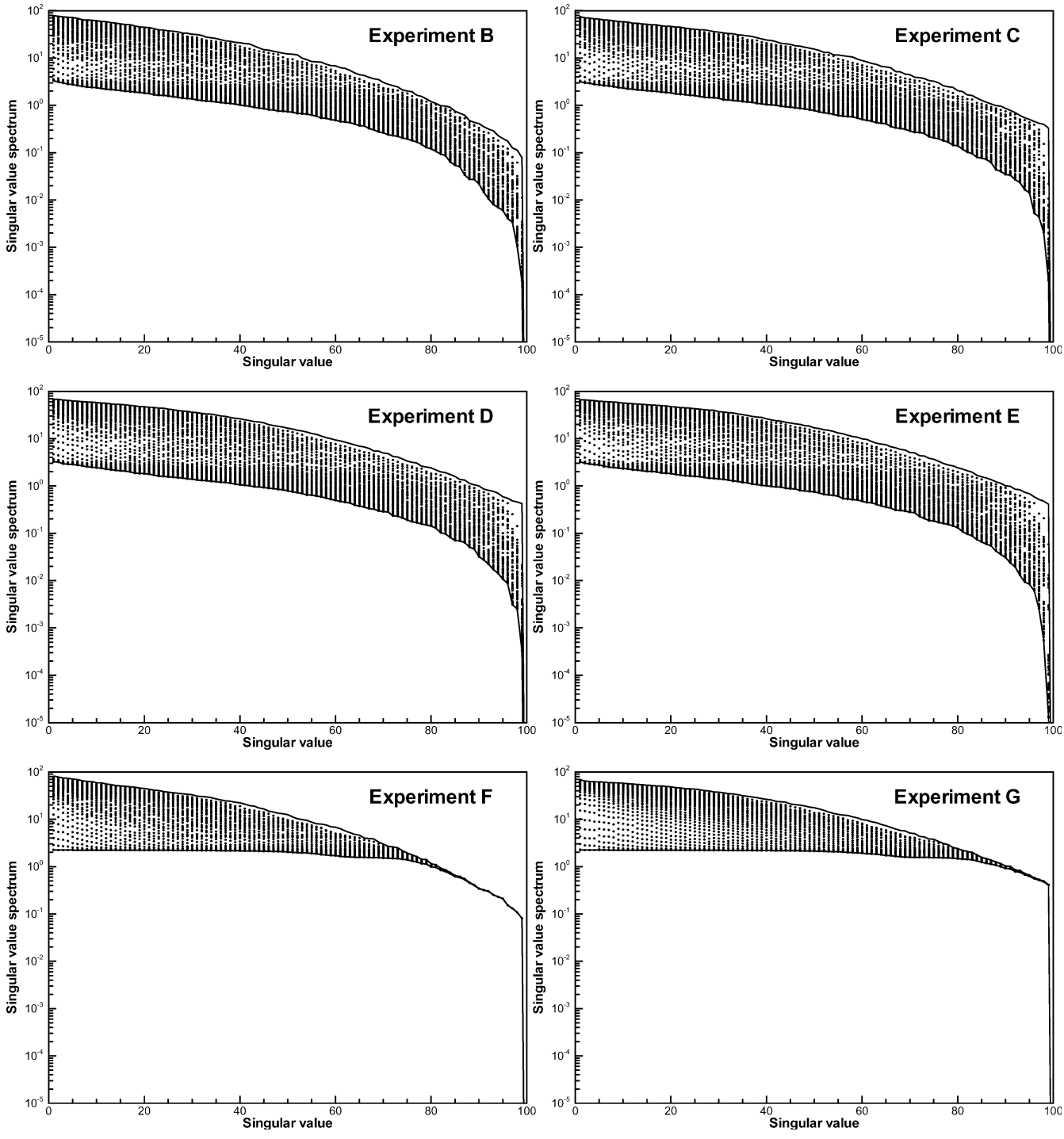

Fig. 6 Time evolution of the ensemble singular value spectra for some of the experiments

There is clearly a significant positive impact resulting from the improved sampling of measurement perturbations. The Exp. I leads to an improvement which is nearly as good as was obtained using the square root algorithm in Exp. F. The use of improved sampling for both the initial ensemble and the measurement perturbations in Exp. H results in a solution with residuals located somewhere between the Exps. E and G. Thus, a significant improvement is obtained from the improved sampling of the measurement perturbations, but the square root algorithm is still superior in the example shown here.

Comparison of the time evolution of residuals for Exps. H and I in Fig. 5 also confirms that the perturbation of measurements is a major cause for the overestimate of residuals. In Exps. H and I the improved sampling scheme is used for the measurement perturbations and this significantly improves the results compared to the Exps. B–E.

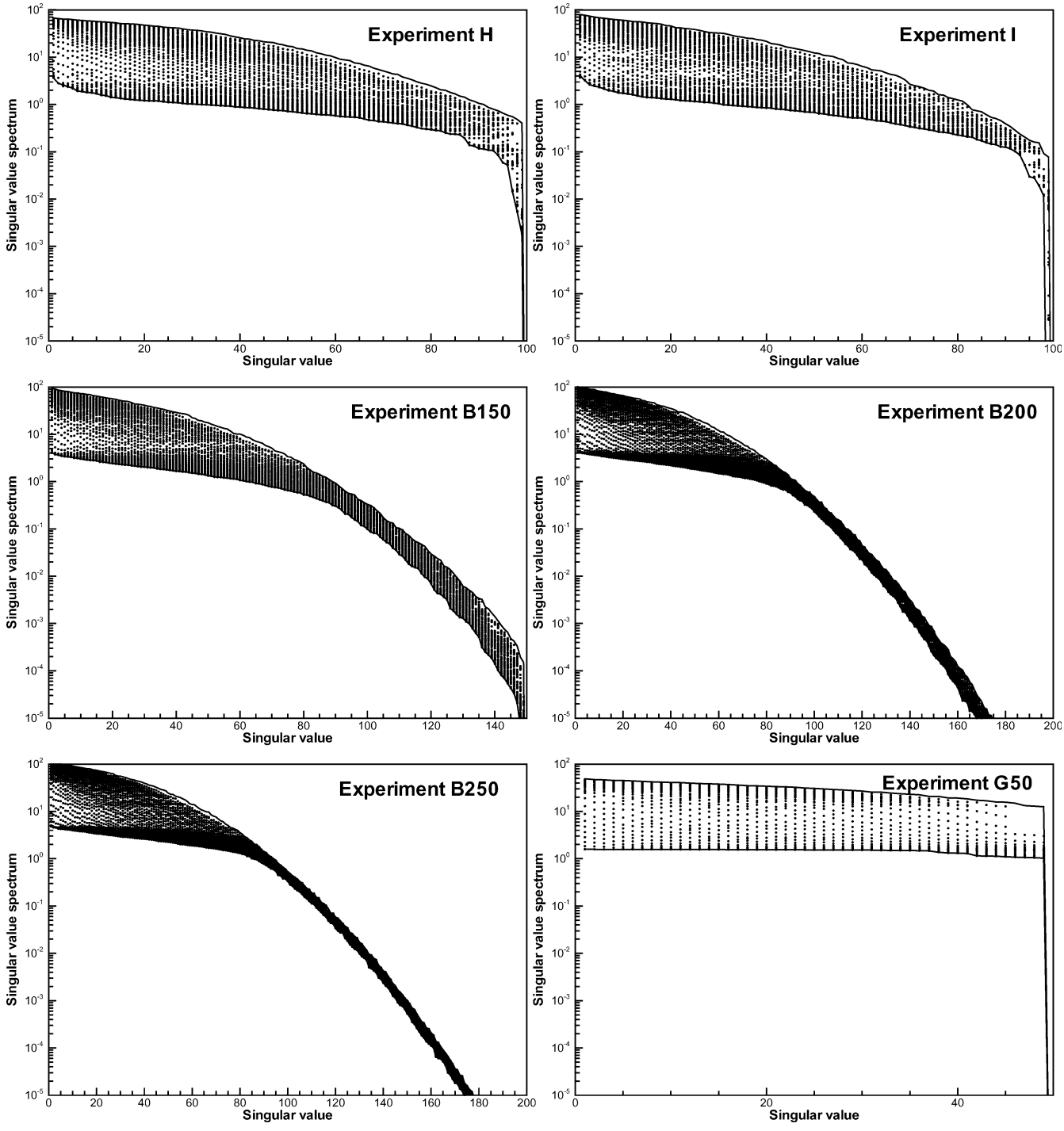

Fig. 7 Time evolution of the ensemble singular value spectra for some of the experiments

6.7 Impact from ensemble size

It is of interest to quantify the benefit of using the improved algorithms by running the traditional EnKF with increasing ensemble size. In these experiments we correct only for the mean and variance in the ensemble, something which is always done in practical applications. We have run the Exps. B150, B200 and B250 which are equivalent to Exp. B except for the ensemble size, which is respectively 150, 200 and 250.

From Fig. 3 it is seen that the traditional EnKF with between 150 and 200 members gives results which are of a quality similar to those of the EnKF, with improved initial sampling using a start ensemble of 600. When using 200 and 250 members, an additional improvement in the results is found but it is still far to go before we obtain the same result as in Exp. G. Also in Fig. 5 it is seen that an improvement is obtained when the ensemble size is increased.

It was also found when comparing Exps. B150 and B200 with Exp. F that an ensemble size between 150 and 200 is needed in the standard EnKF to obtain improvement similar to that obtained from the use of the Analysis 2 with standard sampling of a 100-member initial ensemble.

From these experiments it is clear that a rather large ensemble must be used in the standard EnKF to match the results of Exp. G. Additional experiments were run using improved initial sampling in combination with Analysis 2 (as in Exp. G), but with smaller ensembles. From these experiments we obtained results similar to Exp. B using only around 50–52 members in the ensemble, as seen in Fig. 3. The time evolution of the residuals for Exp. G50 is shown in Fig. 5 and is similar to Exp. B. There seemed to be a breakpoint between 60 and 75 members where the residuals in the G experiments became consistent with the estimated standard deviations.

Thus, it seems that the convergence properties of the EnKF using the Analysis 2 together with the improved sampling are faster with increasing ensemble size than in the standard EnKF. Using Exp. G50 as a reference, we need to double the ensemble size with the standard EnKF in Exp. B, to obtain similar results. Further, the Exp. B200 is still far from the results obtained with Exp. G using 100 members of the ensemble. This can be expected from the theory of quasi-random sampling, which converges proportional to N rather than proportional to $\sqrt{N}$ in standard Monte Carlo sampling.

The improvement obtained by the improved sampling could be utilized to apply the filter algorithm with an ensemble size smaller than used for the normal EnKF algorithm while still obtaining a comparable residual. This configuration will lead to a much shorter computing time.

6.8 Evolution of ensemble singular spectra

Finally, it is of interest to examine how the rank and conditioning of the ensemble evolves in time and is impacted by the computation of the analysis. In Figs. 6 and 7 we have plotted the singular values for the ensemble at each analysis time for the same experiments as shown in Figs. 4 and 5. The initial singular spectrum of the ensemble is plotted as the upper thick line. Then the dotted lines indicate the reduction of the ensemble variance introduced at each analysis update, until the end of the experiment, where the singular spectrum is given by the lower thick line.

It is clear from Exps. B, C, D and E that the conditioning of the initial ensemble improves when the new sampling scheme is used. It is also seen that the traditional EnKF analysis equation lead to a reduction of variance for all the singular values.

The spectra change dramatically when the square root algorithm is used in Exps. F and G. The initial spectra are similar to the ones in Exps. B and E, but now the reduction of variance is more confined to the dominant singular values and there is no reduction at all for the least significant singular values. At the final time the singular spectrum is almost flat up to a certain singular value, which is in contrast to the singular spectra obtained when measurements were perturbed. This shows that the square root scheme weights the singular vectors in an equal way. In these experiments it appears that the error space is large enough for the solution to converge. In the Exp. G50 all singular values experience a similar reduction of variance, indicating that additional ensemble members should be included to better represent the error space.

As expected, the Exp. H shows a behaviour which is somewhere in between what we found in Exps. E and G, while the Exp. I shows a behaviour which is somewhere in between what we found in Exps. B and F.

Finally, it is seen from Exps. B, B150, B200 and B250 that increasing the ensemble size does not add much to the representation of variance in the error subspace. This can be expected with the simple low-dimensional model state considered here.

7 A low-rank square root analysis scheme

In the following we will present an analysis of the case when a low-rank representation is used for the measurement error covariance matrix $\boldsymbol{R}$. The use of the pseudo-inverse is discussed, and the rank issues presented by Kepert (2004) are further analyzed. It is shown that a stable inverse can be computed in the case when a low-rank $\boldsymbol{R}_e$ is used instead of the full-rank R. This leads to the derivation of a very efficient analysis scheme which can be used both when $\boldsymbol{R}_e$ is of full rank and when a low-rank representation is used.

7.1 A pseudo-inverse for $\boldsymbol{C}$

When the dimension of $\boldsymbol{C}$ is large, it is possible that $\boldsymbol{C}$ becomes numerically singular even when using a full-rank $\boldsymbol{R}$ in the definition (12). If the low-rank approximations in Eqs. (13) and (14) are used, then $\boldsymbol{C}$ may become singular, as is further discussed in the following section.

When $\boldsymbol{C}$ is singular it is possible to compute the pseudo-inverse $\boldsymbol{C}^+$ of $\boldsymbol{C}$. It is convenient to formulate the analysis schemes in terms of the pseudo-inverse, since the pseudo-inverse $\boldsymbol{C}^+ \equiv \boldsymbol{C}^{-1}$ when $\boldsymbol{C}$ is of full rank. The algorithm will then be valid in the general case.

The pseudo-inverse of the quadratic matrix $\boldsymbol{C}$ with eigenvalue factorization

$$\boldsymbol{C} = \boldsymbol{Z}\boldsymbol{\Lambda}\boldsymbol{Z}^{\mathrm{T}}, \tag{45}$$

is defined as

$$\boldsymbol{C}^+ = \boldsymbol{Z}\boldsymbol{\Lambda}^+\boldsymbol{Z}^{\mathrm{T}}. \tag{46}$$

The matrix $\boldsymbol{\Lambda}^+$ is diagonal and with $p = \mathrm{rank}(\boldsymbol{C})$ it is defined as

$$\text{diag}\ (\boldsymbol{\Lambda}^{+}) = (\lambda_1^{-1}, \ldots, \lambda_p^{-1}, 0, \ldots, 0), \tag{47}$$

with the eigen values $\lambda_i \geq \lambda_{i+1}$.

It is useful to attempt an interpretation of the algorithm used in the square root analysis schemes. We start by storing the p non-zero elements of $\text{diag}(\boldsymbol{\Lambda}^{+})$ on the diagonal of $\boldsymbol{\Lambda}_p^{-1}$, i.e.

$$\text{diag}(\boldsymbol{\Lambda}_p^{-1}) = (\lambda_1^{-1}, \ldots, \lambda_p^{-1}). \tag{48}$$

We then define the matrix containing the first p eigenvectors in $\boldsymbol{Z}$ as $\boldsymbol{Z}_p = (\boldsymbol{z}_1 \ldots \boldsymbol{z}_p) \in \Re^{m \times p}$. It is clear that the product, $\boldsymbol{Z}_p \boldsymbol{\Lambda}_p^{-1} \boldsymbol{Z}_p^{\mathrm{T}}$, is the Moore-Penrose or pseudo-inverse of the original matrix, $\boldsymbol{C}$.

We now define the rotated measurement operator $\widetilde{\boldsymbol{H}} \in \Re^{p \times n}$ as

$$\widetilde{\boldsymbol{H}} = \boldsymbol{Z}_p^{\mathrm{T}} \boldsymbol{H}, \tag{49}$$

the p rotated measurements

$$\widetilde{\boldsymbol{d}} = \boldsymbol{Z}_p^{\mathrm{T}} \boldsymbol{d}, \tag{50}$$

and the p rotated measurements of the ensemble perturbations, $\widetilde{\boldsymbol{S}} \in \Re^{p \times N}$, as

$$\widetilde{\boldsymbol{S}} = \boldsymbol{Z}_p^{\mathrm{T}} \boldsymbol{H} \boldsymbol{A}' = \widetilde{\boldsymbol{H}} \boldsymbol{A}' = \boldsymbol{Z}_p^{\mathrm{T}} \boldsymbol{S}. \tag{51}$$

This corresponds to the use of a measurement antenna which is oriented along the p dominant principal directions of the original matrix (see Bennett 1992, Chap. 6). The analysis equation for the ensemble mean Eq. (23) can then be written

$$\overline{\boldsymbol{\psi}}^{\mathrm{a}} = \overline{\boldsymbol{\psi}}^{\mathrm{f}} + \boldsymbol{A}' \widetilde{\boldsymbol{S}}^{\mathrm{T}} \boldsymbol{\Lambda}_p^{-1} (\widetilde{\boldsymbol{d}} - \widetilde{\boldsymbol{H}} \overline{\boldsymbol{\psi}}^{\mathrm{f}}). \tag{52}$$

Thus, the analysis is just the assimilation of the p rotated measurements where the matrix in the inverse is diagonal.

The update for the measurement perturbations is defined by Eq. (22), which can be rewritten as

$$\boldsymbol{A}^{\mathrm{a}\prime} \boldsymbol{A}^{\mathrm{a}\prime \mathrm{T}} = \boldsymbol{A}' \left(\boldsymbol{I} - \widetilde{\boldsymbol{S}}^{\mathrm{T}} \boldsymbol{\Lambda}_p^{-1} \widetilde{\boldsymbol{S}} \right) \boldsymbol{A}'^{\mathrm{T}}. \tag{53}$$

This pseudo-inverse will in some cases lead to a loss of rank in the analyzed ensemble, as is discussed in the following section.

7.2 Rank issues

It has recently been shown by Kepert (2004) that the use of an ensemble representation, $\boldsymbol{R}_{\mathrm{e}}$, for $\boldsymbol{R}$ in some cases leads to a loss of rank in the ensemble when $m > N$. However, it is not obvious that the case with $m > N$ and the use of a low-rank representation $\boldsymbol{R}_{\mathrm{e}}$ of $\boldsymbol{R}$, should pose a problem. After all, the final coefficient matrix which is multiplied with the ensemble forecast to produce the analysis is an $N \times N$ matrix. The rank problem may occur using both the EnKF analysis scheme with perturbation of measurements and the square root algorithm presented in Section 3.

The following will revisit the proof by Kepert (2004) and extend it to a more general situation. Further, it will be shown that the rank problem can be avoided when the measurement perturbations, used to represent the low-rank measurement error covariance matrix, are sampled under specific constraints.

The EnKF analysis equation (19) can be rewritten as

$$\begin{aligned} \boldsymbol{A} = \overline{\boldsymbol{A}} &+ \boldsymbol{A}' \boldsymbol{S}^{\mathrm{T}} (\boldsymbol{S}\boldsymbol{S}^{\mathrm{T}} + \boldsymbol{E}\boldsymbol{E}^{\mathrm{T}})^{+} (\overline{\boldsymbol{D}} - \boldsymbol{H}\overline{\boldsymbol{A}}) \\ &+ \boldsymbol{A}' + \boldsymbol{A}' \boldsymbol{S}^{\mathrm{T}} (\boldsymbol{S}\boldsymbol{S}^{\mathrm{T}} + \boldsymbol{E}\boldsymbol{E}^{\mathrm{T}})^{+} (\boldsymbol{E} - \boldsymbol{S}), \end{aligned} \tag{54}$$

where the first line is the update of the mean and the second line is the update of the ensemble perturbations. Thus, for the standard EnKF it suffices to show that rank $(\boldsymbol{W}) = N - 1$ is sufficient to conserve the full rank of the state ensemble, with $\boldsymbol{W}$ defined as

$$\boldsymbol{W} = \boldsymbol{I} - \boldsymbol{S}^{\mathrm{T}} (\boldsymbol{S}\boldsymbol{S}^{\mathrm{T}} + \boldsymbol{E}\boldsymbol{E}^{\mathrm{T}})^{+} (\boldsymbol{S} - \boldsymbol{E}). \tag{55}$$

Similary, for the square root algorithm $\boldsymbol{W}$ is redefined from Eq. (22) as

$$\boldsymbol{W} = \boldsymbol{I} - \boldsymbol{S}^{\mathrm{T}} (\boldsymbol{S}\boldsymbol{S}^{\mathrm{T}} + \boldsymbol{E}\boldsymbol{E}^{\mathrm{T}})^{+} \boldsymbol{S}. \tag{56}$$

We consider the case when $m > N - 1$, which was shown to cause problems in Kepert (2004). Define $\boldsymbol{S} \in \Re^{m \times N}$ with rank $(\boldsymbol{S}) = N - 1$, where the columns of $\boldsymbol{S}$ span a subspace $\mathscr{S}$ of dimension $N - 1$. Further, we define $\boldsymbol{E} \in \Re^{m \times q}$ with $\text{rank}(\boldsymbol{E}) = \min(m, q - 1)$, where $\boldsymbol{E}$ contains an arbitrary number, q, of measurement perturbations.

As in Kepert (2004), one can define the matrix $\boldsymbol{Y} \in \Re^{m \times (N+q)}$ as

$$\boldsymbol{Y} = (\boldsymbol{S}, \boldsymbol{E}), \tag{57}$$

and the matrix $\boldsymbol{C}$ becomes

$$\boldsymbol{C} = \boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}}, \tag{58}$$

with rank

$$p = \text{rank}(\boldsymbol{Y}) = \text{rank}(\boldsymbol{C}). \tag{59}$$

Dependent on the definition for $\boldsymbol{E}$, we have $\min(m, N - 1) \leq p \leq \min(m, N + q - 2)$. One extreme is the case where $q \leq N$ and $\boldsymbol{E}$ is fully contained in $\mathscr{S}$, in which case we have $p = N - 1$. The case considered in Kepert (2004) is another extreme. It had $q = N$, and $p = \min(m, 2N - 2)$, which was also implicitely assumed when setting $\boldsymbol{S}\boldsymbol{E}^{\mathrm{T}} = 0$ in the approximate factorization introduced in Evensen (2003). This also corresponds to a situation which is likely to occur when $\boldsymbol{E}$ is sampled randomly with components along the $N - 1$ directions in $\mathscr{S}^{\perp}$.

We define the SVD of $\boldsymbol{Y}$ as

$$\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}} = \boldsymbol{Y}, \tag{60}$$

with $\boldsymbol{U} \in \Re^{m \times m}$, $\boldsymbol{\Sigma} \in \Re^{m \times (N+q)}$ and $\boldsymbol{V} \in \Re^{(N+q) \times (N+q)}$. The pseudo-inverse of $\boldsymbol{Y}$ is defined as

$$\boldsymbol{Y}^{+} = \boldsymbol{V}\boldsymbol{\Sigma}^{+}\boldsymbol{U}^{\mathrm{T}}, \tag{61}$$

where $\boldsymbol{\Sigma}^{+} \in \Re^{(N+q) \times m}$ is diagonal and defined as $\text{diag}(\boldsymbol{\Sigma}^{+}) = (\sigma_1^{-1}, \sigma_2^{-1}, \ldots, \sigma_p^{-1}, 0, \ldots, 0)$.

Both the equations for $\boldsymbol{W}$ in Eqs. (56) and (57) can be rewritten in a form similar to that used by Kepert (2004). Introducing the expressions (60) and (61) in Eq. (56),

and defining $\boldsymbol{I}_N$ to be the N-dimensional identity matrix, we obtain

$$\boldsymbol{W} = \boldsymbol{I}_N - (\boldsymbol{I}_N, \boldsymbol{0})\boldsymbol{Y}^{\mathrm{T}}(\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}})^{+}\boldsymbol{Y}(\boldsymbol{I}_N, \boldsymbol{0})^{\mathrm{T}} \tag{62}$$

$$= \boldsymbol{I}_N - (\boldsymbol{I}_N, \boldsymbol{0})\boldsymbol{V}\boldsymbol{\Sigma}^{\mathrm{T}}\boldsymbol{\Sigma}^{+\mathrm{T}}\boldsymbol{\Sigma}^{+}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}(\boldsymbol{I}_N, \boldsymbol{0})^{\mathrm{T}} \tag{63}$$

$$= (\boldsymbol{I}_N, \boldsymbol{0})\boldsymbol{V}\left\{\boldsymbol{I}_{N+q} - \begin{pmatrix} \boldsymbol{I}_p & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix}_{N+q}\right\} \times \boldsymbol{V}^{\mathrm{T}}(\boldsymbol{I}_N, \boldsymbol{0})^{\mathrm{T}} \tag{64}$$

$$= (\boldsymbol{I}_N, \boldsymbol{0})\boldsymbol{V}\begin{pmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{N+q-p} \end{pmatrix}_{N+q}\boldsymbol{V}^{\mathrm{T}}(\boldsymbol{I}_N, \boldsymbol{0})^{\mathrm{T}}. \tag{65}$$

The similar expression for $\boldsymbol{W}$ in Eq. (55) is obtained by replacing the rightmost matrix, $(\boldsymbol{I}_N, \boldsymbol{0}) \in \Re^{N\times(N+q)}$, with $(\boldsymbol{I}_N, -\boldsymbol{I}_N, \boldsymbol{0}) \in \Re^{N\times(N+q)}$.

We need the $N+q$ matrix in Eq. (65) to have rank at least $N-1$ to maintain the rank of the updated ensemble perturbations. Thus, we require that $N+q-p \geq N-1$ and obtain the general condition

$$p \leq q+1. \tag{66}$$

With $q = N$ this condition requires $p \leq N+1$. This is only possible when all singular vectors of $\boldsymbol{E}$, except two, are contained in $\mathscr{S}$. Thus, it is clear that a low-rank representation of $\boldsymbol{R}$ using N measurement perturbations, $\boldsymbol{E}$, can be used as long as the selected perturbations do not increase the rank of $\boldsymbol{Y}$ to more than $N+1$.

It is also clear that if this constrained low-rank representation, $\boldsymbol{E} \in \Re^{m\times N}$, is unable to properly represent the real measurement error covariance, it is possible to increase the number of perturbations to an arbitrary number $q > N$ as long as the rank, p, satisfies the condition (66).

In Kepert (2004) it was assumed that the rank $p = 2N-2$, i.e. $\boldsymbol{E}$ has components in $N-1$ directions of $\mathscr{S}^{\perp}$. Then, clearly, the condition (66) is violated and this results in a loss of rank. It was showed that this can be resolved using a full-rank measurement error covariance matrix (corresponding to the limiting case when $q \geq m+1$). Then, $p = \text{ rank }(\boldsymbol{Y}) = \text{ rank }(\boldsymbol{R}_{\mathrm{e}}) = m$ and the condition (66) is always satisfied.

Assume now that we have removed r columns from $\boldsymbol{E} \in \Re^{m\times(q=m+1)}$. We then obtain the reduced $\tilde{\boldsymbol{E}} \in \Re^{m\times(q=m+1-r)}$ of rank equal to $m-r$. In this situation we can consider two cases. First, if the removed perturbations are also fully contained in $\mathscr{S}$, then this does not lead to a reduction of p, which still equals m. In this case we can write the condition (66), for $r \leq N-1$, as

$$p = m \leq m+2-r, \tag{67}$$

which is violated for $r > 2$. Secondly, assume that the removed perturbations are fully contained in $\mathscr{S}^{\perp}$. Then the rank p will be reduced with r and we write the condition (66) as

$$p = m-r \leq m+2-r. \tag{68}$$

We can continue to remove columns of $\boldsymbol{E}$ contained in $\mathscr{S}^{\perp}$, without violating the condition (66), until there are only $N-1$ columns left in $\tilde{\boldsymbol{E}}$, all contained in $\mathscr{S}$.

From this discussion, it is clear that we need the measurement error perturbations to explain variance within the ensemble space $\mathscr{S}$. This will now be used in the following discussion to reformulate the square root analysis scheme as an efficient and stable algorithm which exploits that the analysis is computed in the ensemble subspace.

7.3 A stable pseudo-inverse of $\boldsymbol{C}$

From the previous discussion it is clear that we need to use measurement perturbations which are contained in $\mathscr{S}$ to avoid the loss of rank in the analyzed ensemble. This now forms the foundation for the derivation of an algorithm where the inverse is computed in the N-dimensional ensemble space rather than the m-dimensional measurement space.

The key to this algorithm is a new approach for computing the inverse of $\boldsymbol{C}$ in the case when $m > N-1$. The case when $m \leq N-1$ is trivial since then $\boldsymbol{C}$ will have full rank.

We assume again that $\boldsymbol{S}$ has rank equal to $N-1$, which will be the case if the ensemble is chosen properly and the measurement operator has full rank. The SVD of $\boldsymbol{S}$ is

$$\boldsymbol{U}_0\boldsymbol{\Sigma}_0\boldsymbol{V}_0^{\mathrm{T}} = \boldsymbol{S}, \tag{69}$$

with $\boldsymbol{U}_0 \in \Re^{m\times m}$, $\boldsymbol{\Sigma}_0 \in \Re^{m\times N}$ and $\boldsymbol{V}_0 \in \Re^{N\times N}$. The subspace $\mathscr{S}$ is more precisely defined by the first $N-1$ singular vectors of $\boldsymbol{S}$ as contained in $\boldsymbol{U}_0$.

The pseudo-inverse of $\boldsymbol{S}$ is defined as

$$\boldsymbol{S}^{+} = \boldsymbol{V}_0\boldsymbol{\Sigma}_0^{+}\boldsymbol{U}_0^{\mathrm{T}}, \tag{70}$$

where $\boldsymbol{\Sigma}_0^{+} \in \Re^{N\times m}$ is diagonal and defined as $\mathrm{diag}(\boldsymbol{\Sigma}_0^{+}) = (\sigma_1^{-1}, \sigma_2^{-1}, \ldots, \sigma_{N-1}^{-1}, 0)$.

The matrix product $\boldsymbol{\Sigma}_0\boldsymbol{\Sigma}_0^{+} = \tilde{\boldsymbol{I}}_{N-1} \in \Re^{m\times m}$ where $\tilde{\boldsymbol{I}}_{N-1}$ has the first $N-1$ diagonal elements equal to one and the rest of the elements in the matrix are zero.

We now use this in the expression for $\boldsymbol{C}$, as defined in Eq. (12), to obtain

$$\boldsymbol{C} = (\boldsymbol{U}_0\boldsymbol{\Sigma}_0\boldsymbol{\Sigma}_0^{\mathrm{T}}\boldsymbol{U}_0^{\mathrm{T}} + (N-1)\boldsymbol{R}) \tag{71}$$

$$= \boldsymbol{U}_0(\boldsymbol{\Sigma}_0\boldsymbol{\Sigma}_0^{\mathrm{T}} + (N-1)\boldsymbol{U}_0^{\mathrm{T}}\boldsymbol{R}\boldsymbol{U}_0)\boldsymbol{U}_0^{\mathrm{T}} \tag{72}$$

$$\approx \boldsymbol{U}_0\boldsymbol{\Sigma}_0(\boldsymbol{I} + (N-1)\boldsymbol{\Sigma}_0^{+}\boldsymbol{U}_0^{\mathrm{T}}\boldsymbol{R}\boldsymbol{U}_0\boldsymbol{\Sigma}_0^{+\mathrm{T}})\boldsymbol{\Sigma}_0^{\mathrm{T}}\boldsymbol{U}_0^{\mathrm{T}} \tag{73}$$

$$= \boldsymbol{S}\boldsymbol{S}^{\mathrm{T}} + (N-1)(\boldsymbol{S}\boldsymbol{S}^{+})\boldsymbol{R}(\boldsymbol{S}\boldsymbol{S}^{+})^{\mathrm{T}}. \tag{74}$$

In Eq. (72), the matrix $\boldsymbol{U}_0^{\mathrm{T}}\boldsymbol{R}\boldsymbol{U}_0$ is the projection of the measurement error covariance matrix, $\boldsymbol{R}$, onto the space spanned by the m singular vectors of $\boldsymbol{S}$, contained in the columns of $\boldsymbol{U}_0$.

Then in Eq. (73) we introduce an approximation by effectively multiplying $\boldsymbol{U}_0^{\mathrm{T}}\boldsymbol{R}\boldsymbol{U}_0$ from left and right by the matrix $\boldsymbol{\Sigma}_0\boldsymbol{\Sigma}_0^{+} = \tilde{\boldsymbol{I}}_{N-1} \in \Re^{m\times m}$, Thus, we extract the part of $\boldsymbol{R}$ contained in the subspace consisting of the $N-1$ dominant directions in $\boldsymbol{U}_0$, i.e. the subspace $\mathscr{S}$.

The matrix $\boldsymbol{S}\boldsymbol{S}^{+} = \boldsymbol{U}_0\tilde{\boldsymbol{I}}_{N-1}\boldsymbol{U}_0^{\mathrm{T}}$ in Eq. (74) is a Hermitian and normal matrix. It is also an orthogonal projection onto $\mathscr{S}$. Thus we essentially adopt a low-rank representation for $\boldsymbol{R}$ which is contained in the same subspace as the ensemble perturbations in $\boldsymbol{S}$.

It is now interesting to observe that if we replace $\boldsymbol{R}$ with a low-rank version $(N-1)\boldsymbol{R}_{\mathrm{e}} = \boldsymbol{E}\boldsymbol{E}^{\mathrm{T}}$, then

$$\boldsymbol{C} = \boldsymbol{S}\boldsymbol{S}^{\mathrm{T}} + \boldsymbol{E}\boldsymbol{E}^{\mathrm{T}} \tag{75}$$

$$\approx \boldsymbol{S}\boldsymbol{S}^{\mathrm{T}} + (\boldsymbol{S}\boldsymbol{S}^{+})\boldsymbol{E}\boldsymbol{E}^{\mathrm{T}}(\boldsymbol{S}\boldsymbol{S}^{+}) \tag{76}$$

$$= \boldsymbol{S}\boldsymbol{S}^{\mathrm{T}} + \widehat{\boldsymbol{E}}\widehat{\boldsymbol{E}}^{\mathrm{T}}, \tag{77}$$

where $\widehat{\boldsymbol{E}} = (\boldsymbol{S}\boldsymbol{S}^{+})\boldsymbol{E}$ is the projection of $\boldsymbol{E}$ onto the first $N-1$ singular vectors in $\boldsymbol{U}_0$. Thus, we can only account for the measurement variance contained in the subspace $\mathscr{S}$. When we project $\boldsymbol{E}$ onto $\mathscr{S}$ we reject all possible contributions in $\mathscr{S}^{\perp}$. It is this rejection which ensures that we avoid the loss of rank reported by Kepert (2004).

There are now two cases to consider. The first is when a full measurement error covariance matrix is specified. It may be of full rank or not. The second case considers the use of a low-rank representation using ensemble perturbations, $\boldsymbol{E}$.

If we use Eqs. (69) and (24) in Eq. (22), we obtain

$$\boldsymbol{I} - \boldsymbol{S}^{\mathrm{T}}\boldsymbol{C}^{-1}\boldsymbol{S} = \boldsymbol{I} - \boldsymbol{V}_0\boldsymbol{\Sigma}_0^{\mathrm{T}}\boldsymbol{U}_0^{\mathrm{T}}\boldsymbol{Z}\boldsymbol{\Lambda}^{-1}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{U}_0\boldsymbol{\Sigma}_0\boldsymbol{V}_0^{\mathrm{T}}. \tag{78}$$

Thus, it is seen that only the part of $\boldsymbol{C}$ contained in $\mathscr{S}$ is accounted for since $\boldsymbol{Z}$ is projected onto the $N-1$ significant singular vectors in $\boldsymbol{U}_0$. Thus, there is really no benefit from using a full-rank $\boldsymbol{R}$ except that one then knows for sure that it will have non-zero contributions in all of $\mathscr{S}$, which is necessary to avoid loss of rank in the updated ensemble perturbations. When using the low-rank representation it is important to ensure that the ensemble stored in $\boldsymbol{E}$ spans the space $\mathscr{S}$. Thus, the projected $\boldsymbol{E}$ is orthogonal to $\mathscr{S}^{\perp}$.

Thus, we now proceed with a low-rank representation for $\boldsymbol{R}$ as the general case. Replacing $(N-1)\boldsymbol{R}$ with $\boldsymbol{E}\boldsymbol{E}^{\mathrm{T}}$ in Eq. (73) we obtain:

$$\boldsymbol{C} \approx \boldsymbol{U}_0\boldsymbol{\Sigma}_0(\boldsymbol{I} + \boldsymbol{\Sigma}_0^{+}\boldsymbol{U}_0^{\mathrm{T}}\boldsymbol{E}\boldsymbol{E}^{\mathrm{T}}\boldsymbol{U}_0\boldsymbol{\Sigma}_0^{+\mathrm{T}})\boldsymbol{\Sigma}_0^{\mathrm{T}}\boldsymbol{U}_0^{\mathrm{T}} \tag{79}$$

$$= \boldsymbol{U}_0\boldsymbol{\Sigma}_0(\boldsymbol{I} + \boldsymbol{X}_0\boldsymbol{X}_0^{\mathrm{T}})\boldsymbol{\Sigma}_0^{\mathrm{T}}\boldsymbol{U}_0^{\mathrm{T}}, \tag{80}$$

where we have defined

$$\boldsymbol{X}_0 = \boldsymbol{\Sigma}_0^{+}\boldsymbol{U}_0^{\mathrm{T}}\boldsymbol{E}, \tag{81}$$

which is an $N \times N$ matrix of with rank equal to $N-1$ and it requires $mN^2 + N^2$ floating point operations to form it. The approximative equality sign introduced in Eq. (79) just denotes that all components in $\boldsymbol{E}$ contained in $\mathscr{S}^{\perp}$ have now been removed. This does not impact the analysis result and if $\boldsymbol{E}$ was sampled from $\mathscr{S}$ to exactly represent $\boldsymbol{U}_0^{\mathrm{T}}\boldsymbol{R}\boldsymbol{U}_0$, the exact equality would be satisfied, thus we do not continue to use the approximative equality sign below.

We then proceed with a singular value decomposition

$$\boldsymbol{U}_1\boldsymbol{\Sigma}_1\boldsymbol{V}_1^{\mathrm{T}} = \boldsymbol{X}_0, \tag{82}$$

where all matrices are $N \times N$, and insert this in Eq. (80) to obtain:

$$\boldsymbol{C} = \boldsymbol{U}_0\boldsymbol{\Sigma}_0(\boldsymbol{I} + \boldsymbol{U}_1\boldsymbol{\Sigma}_1^2\boldsymbol{U}_1^{\mathrm{T}})\boldsymbol{\Sigma}_0^{\mathrm{T}}\boldsymbol{U}_0^{\mathrm{T}} \tag{83}$$

$$= \boldsymbol{U}_0\boldsymbol{\Sigma}_0\boldsymbol{U}_1(\boldsymbol{I} + \boldsymbol{\Sigma}_1^2)\boldsymbol{U}_1^{\mathrm{T}}\boldsymbol{\Sigma}_0^{\mathrm{T}}\boldsymbol{U}_0^{\mathrm{T}}. \tag{84}$$

Now the pseudo-inverse of $\boldsymbol{C}$ becomes

$$\boldsymbol{C}^{+} = (\boldsymbol{U}_0\boldsymbol{\Sigma}_0^{+\mathrm{T}}\boldsymbol{U}_1)(\boldsymbol{I} + \boldsymbol{\Sigma}_1^2)^{-1}(\boldsymbol{U}_0\boldsymbol{\Sigma}_0^{+\mathrm{T}}\boldsymbol{U}_1)^{\mathrm{T}} \tag{85}$$

$$= \boldsymbol{X}_1(\boldsymbol{I} + \boldsymbol{\Sigma}_1^2)^{-1}\boldsymbol{X}_1^{\mathrm{T}}, \tag{86}$$

where we have defined $\boldsymbol{X}_1 \in \Re^{m\times N}$ of rank $N-1$ as

$$\boldsymbol{X}_1 = \boldsymbol{U}_0\boldsymbol{\Sigma}_0^{+\mathrm{T}}\boldsymbol{U}_1. \tag{87}$$

7.4 Efficient square root algorithm with low rank $\boldsymbol{R}$

A slight modification is now introduced to the square root scheme derived in Section 3 which exploits the use of a low-rank representation for $\boldsymbol{R}$.

7.4.1 Derivation of algorithm

We start by defining the same SVD of $\boldsymbol{S}$ as in Eq. (69), then use the definition (14) for $\boldsymbol{C}$.

Using the expression (86) for the inverse in Eq. (22) we obtain the following derivation of the analysis scheme

$$\boldsymbol{A}^{\mathrm{a}\prime}\boldsymbol{A}^{\mathrm{a}\prime\mathrm{T}} = \boldsymbol{A}'\left(\boldsymbol{I} - \boldsymbol{S}^{\mathrm{T}}\boldsymbol{C}^{+}\boldsymbol{S}\right)\boldsymbol{A}'^{\mathrm{T}} \tag{88}$$

$$= \boldsymbol{A}'\left(\boldsymbol{I} - \boldsymbol{S}^{\mathrm{T}}\boldsymbol{X}_1(\boldsymbol{I} + \boldsymbol{\Sigma}_1^2)^{-1}\boldsymbol{X}_1^{\mathrm{T}}\boldsymbol{S}\right)\boldsymbol{A}'^{\mathrm{T}} \tag{89}$$

$$= \boldsymbol{A}'\left(\boldsymbol{I} - [(\boldsymbol{I} + \boldsymbol{\Sigma}_1^2)^{-\frac{1}{2}}\boldsymbol{X}_1^{\mathrm{T}}\boldsymbol{S}]^{\mathrm{T}}[(\boldsymbol{I} + \boldsymbol{\Sigma}_1^2)^{-\frac{1}{2}}\boldsymbol{X}_1^{\mathrm{T}}\boldsymbol{S}]\right)\boldsymbol{A}'^{\mathrm{T}} \tag{90}$$

$$= \boldsymbol{A}'\left(\boldsymbol{I} - \boldsymbol{X}_2^{\mathrm{T}}\boldsymbol{X}_2\right)\boldsymbol{A}'^{\mathrm{T}}, \tag{91}$$

where we have defined $\boldsymbol{X}_2$ as

$$\boldsymbol{X}_2 = (\boldsymbol{I} + \boldsymbol{\Sigma}_1^2)^{-\frac{1}{2}}\boldsymbol{X}_1^{\mathrm{T}}\boldsymbol{S} = (\boldsymbol{I} + \boldsymbol{\Sigma}_1^2)^{-\frac{1}{2}}\boldsymbol{U}_1^{\mathrm{T}}\tilde{\boldsymbol{I}}_{N-1}\boldsymbol{V}_0^{\mathrm{T}}, \tag{92}$$

We then end up with the same final update Eq. (34) by following the derivation defined in Eqs. (29–33).

Thus, we have replaced the explicit factorization of $\boldsymbol{C} \in \Re^{m\times m}$, with a SVD of $\boldsymbol{S} \in \Re^{m\times N}$, and this is a significant saving when $m \gg N$. Further, by using a low rank version for $\boldsymbol{R}$ we replace the matrix multiplication $\boldsymbol{\Sigma}_0^{+}\boldsymbol{U}_0^{\mathrm{T}}\boldsymbol{R}$ in Eq. (73) with the less expensive $\boldsymbol{\Sigma}_0^{+}\boldsymbol{U}_0^{\mathrm{T}}\boldsymbol{E}$. The sampling of $\boldsymbol{E}$ can be done efficiently, e.g. using the algorithm described in Evensen (2003). Thus, there are no matrix operations which requires $\mathscr{O}(m^2)$ floating point operations in the new algorithm.

In Appendix A we have presented the alternative derivation of this algorithm where a full $\boldsymbol{R}$ (possibly of low rank) is specified.

7.4.2 Implementation of algorithm

The following steps are performed to compute the analysis:

1. Compute the SVD from Eq. (69): $\boldsymbol{U}_0\boldsymbol{\Sigma}_0\boldsymbol{V}_0^{\mathrm{T}} = \boldsymbol{S}$.
2. Form the matrix product in Eq. (81): $\boldsymbol{X}_0 = \boldsymbol{\Sigma}_0^{+}\boldsymbol{U}_0^{\mathrm{T}}\boldsymbol{E}$.
3. Compute the singular value decomposition of $\boldsymbol{X}_0$ in Eq. (82): $\boldsymbol{U}_1\boldsymbol{\Sigma}_1\boldsymbol{V}_1^{\mathrm{T}} = \boldsymbol{X}_0$.

4. Then form the matrix product as defined in Eq. (87): $\boldsymbol{X}_1 = \boldsymbol{U}_0\boldsymbol{\Sigma}_0^{+\mathrm{T}}\boldsymbol{U}_1$.
5. Update the ensemble mean from the equation

$$\bar{\boldsymbol{\psi}}^{\mathrm{a}} = \bar{\boldsymbol{\psi}}^{\mathrm{f}} + \boldsymbol{A}'\boldsymbol{S}^{\mathrm{T}}\boldsymbol{X}_1(\boldsymbol{I} + \boldsymbol{\Sigma}_1^2)^{-1}\boldsymbol{X}_1^{\mathrm{T}}(\boldsymbol{d} - \boldsymbol{H}\bar{\boldsymbol{\psi}}^{\mathrm{f}}), \tag{93}$$

using the following sequence of matrix-vector multiplications:
   a) $\boldsymbol{y}_0 = \boldsymbol{X}_1^{\mathrm{T}}(\boldsymbol{d} - \boldsymbol{H}\bar{\boldsymbol{\psi}}^{\mathrm{f}})$,
   b) $\boldsymbol{y}_2 = (\boldsymbol{I} + \boldsymbol{\Sigma}_1^2)^{-1}\boldsymbol{y}_0$,
   c) $\boldsymbol{y}_3 = \boldsymbol{X}_1\boldsymbol{y}_2$,
   d) $\boldsymbol{y}_4 = \boldsymbol{S}^{\mathrm{T}}\boldsymbol{y}_3$,
   e) $\bar{\boldsymbol{\psi}}^{\mathrm{a}} = \bar{\boldsymbol{\psi}}^{\mathrm{f}} + \boldsymbol{A}'\boldsymbol{y}_4$.
6. Form the matrix product defined by Eq. (92): $\boldsymbol{X}_2 = (\boldsymbol{I} + \boldsymbol{\Sigma}_1^2)^{-\frac{1}{2}}\boldsymbol{X}_1^{\mathrm{T}}\boldsymbol{S}$.
7. Compute the SVD of $\boldsymbol{X}_2$: $\boldsymbol{U}_2\boldsymbol{\Sigma}_2\boldsymbol{V}_2^{\mathrm{T}} = \boldsymbol{X}_2$.
8. Then evaluate the analyzed ensemble perturbations from $\boldsymbol{A}^{\mathrm{a}\prime} = \boldsymbol{A}'\boldsymbol{V}_2\sqrt{\boldsymbol{I} - \boldsymbol{\Sigma}_2^{\mathrm{T}}\boldsymbol{\Sigma}_2\boldsymbol{\Theta}^{\mathrm{T}}}$ and add the mean to arrive at the final analyzed ensemble.

8 Experiments with $m \gg N$

The following experiments were performed to evaluate the properties of the analysis schemes in the case where $m \gg N$. An experimental setup, similar to the one used in the previous section, is adapted. The differences are the following: now 500 measurements are assimilated in each step, the error variance of the measurements is set to 0.5, the number of assimilation steps is five and improved sampling was used for the initial ensemble. The following six experiments are compared:

Exp. 2: The standard EnKF scheme (Analysis 1) with perturbation of measurements and a full rank prescribed $\boldsymbol{R}$.
Exp. 4: The square root algorithm (Analysis 2 from Sec. 3) with a full rank prescribed $\boldsymbol{R}$.
Exp. 5a: The square root algorithm from Appendix A with a full rank prescribed $\boldsymbol{R}$.
Exp. 5b: Same as Exp. 5a but with $\boldsymbol{R} = \boldsymbol{R}_{\mathrm{e}}$ of rank $N - 1$ as defined in Eq. (8). Here, $\boldsymbol{R}_{\mathrm{e}}$ is generated by first sampling $\boldsymbol{E}$ randomly and then computing $\boldsymbol{R}_{\mathrm{e}}$ using Eq. (8).
Exp. 5c: Same as Exp. 5b but with $\boldsymbol{R}_{\mathrm{e}}$ constructed from an $\boldsymbol{E}$ generated using improved sampling.
Exp. 6: The new square root algorithm from Section 7.4, where $\boldsymbol{E}$ is used directly without $\boldsymbol{R}$ being formed.

In these experiments we assume that we know the statistics of the measurement errors, i.e. they are uncorrelated with zero mean and the variance set to 0.5. Thus, the exact diagonal $\boldsymbol{R}$ with the observation error variance on the diagonal, is easily constructed in the Exps. 2, 4 and 5a. For Exps. 5b, 5c and 6 it is straightforward to sample normal independent perturbations for each element of $\boldsymbol{E}$ with the correct statistics. Note that $\boldsymbol{E}$ is sampled with rank equal to $N - 1$. When projected onto $\boldsymbol{U}_0$ it is not guaranteed that the rank of $\boldsymbol{U}_0^{\mathrm{T}}\boldsymbol{R}_{\mathrm{e}}\boldsymbol{U}_0$ or $\boldsymbol{U}_0\boldsymbol{E}$ is equal to $N - 1$. If $\boldsymbol{E}$ has columns which are orthogonal to $\boldsymbol{U}_0$; these do not contribute when projected onto $\boldsymbol{U}_0$. This corresponds to the assimilation of perfect measurements and will lead to a corresponding loss of rank in the updated ensemble. We did not experience this to be a problem in the present experiments, as seen in Fig. 9, but it may be wise to monitor the rank of $\boldsymbol{X}_0$ in Eq. (81) when computing the singular value decomposition.

If the measurement errors are correlated there are different sampling algorithms which can be used, including the one described in Evensent (2003).

Both the standard EnKF analysis algorithm and the square root algorithm from Section 3 worked well in the previous experiments discussed in Section 6 where $\boldsymbol{C}$ was of full rank. However, from the previous discussion we do not expect them to be useful when an $\boldsymbol{R}$ of low rank is randomly sampled and a large number of measurements are used. This, in fact, leads to a loss of rank in the analyzed ensemble as well some singular values in $\boldsymbol{\Sigma}_2$ which were slightly larger than 1, and the square root in Eq. (34) did not always exist. Thus, these algorithms were used only with the full-rank and exactly specified $\boldsymbol{R}$. This problem is eliminated when the more sophisticated inverse algorithm from Section 7.3 is used.

The time evolution of the residuals and singular spectra are presented in Figs. 8 and 9. Comparing Exps. 2 and 4 it is clear that we obtain results which are similar to those from the previous experiments where only four measurements were assimilated. Clearly both of these algorithms are valid for the case when $m > N$ but, as before the square root algorithm produces more accurate results than the traditional EnKF scheme.

Exp. 5a produces results which are nearly identical to those found in Exp. 4, which is expected since the same equation is solved and only the algorithms differ. This illustrates that the new inversion algorithm is consistent and that the rejection of the part of the exactly specified $\boldsymbol{R}$ which is not contained in $\mathscr{S}$ does not change the results.

In *Exp. 5b* we have used $\boldsymbol{R} = \boldsymbol{R}_{\mathrm{e}}$ of rank $N - 1$ as input. It is clear that there is no loss of rank in the analyzed ensemble, although the residuals increase and are no longer consistent with the predicted standard deviations. Further, there is no positive impact from using improved sampling for the perturbations in $\boldsymbol{E}$ as is seen from *Exp. 5c*, except for lower spread of the predicted standard deviations.

The use of a low rank representation for $\boldsymbol{R}$ is valid, and the results will be the same if $\boldsymbol{U}_0^{\mathrm{T}}\boldsymbol{R}_{\mathrm{e}}\boldsymbol{U}_0 = \boldsymbol{U}_0^{\mathrm{T}}\boldsymbol{R}\boldsymbol{U}_0$. This is not the case here since random sampling was used for $\boldsymbol{E}$, and in particular a diagonal matrix can be difficult to represent properly by a low rank random sample. In other words, if one samples $\boldsymbol{E}$ completely within the space spanned by $\boldsymbol{U}_0$, the low rank schemes will give the same result as when a full rank $\boldsymbol{R}$ is used in *Exps. 4* and *5a*.

In the final *Exp. 6* we use the algorithm as defined in Section 7.4 where we avoid the formation of the full measurement error covariance matrix. In this case we obtain results which are almost indentical to the results from *Exps. 5b* and *5c* where a low rank measurement error covariance matrix, $\boldsymbol{R}_{\mathrm{e}}$, is used.

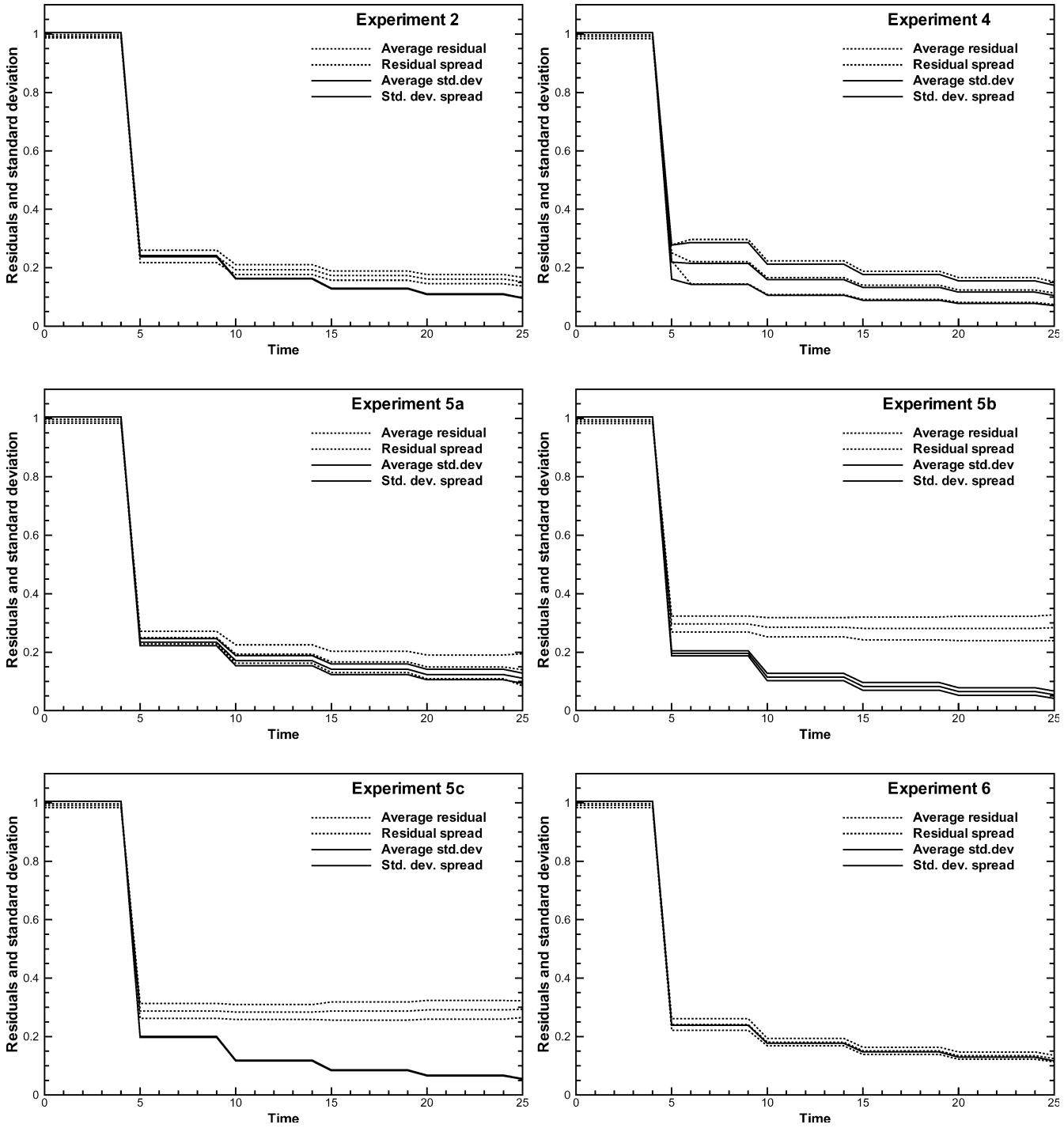

Fig. 8 Time evolution for RMS residuals (*dotted lines*) and estimated standard deviations (*full lines*) for all 50 simulations in the respective experiments

Further examination of these schemes in more realistic settings are clearly required before they are adapted in operational systems. From the previous theoretical analysis, the new low-rank square root scheme derived in Section 7.4, does not introduce any additional approximations compared to the traditional EnKF algorithm. It only introduces measures to stabilise the computation of the analysis and also makes it computationally much more efficient. However, when a low rank $\boldsymbol{R}_e$ is used, a scheme is required for the proper sampling of measurement perturbations in $\mathscr{S}$.

9. Discussion

This paper has quantified the impact of using some improved sampling schemes as well as different analysis

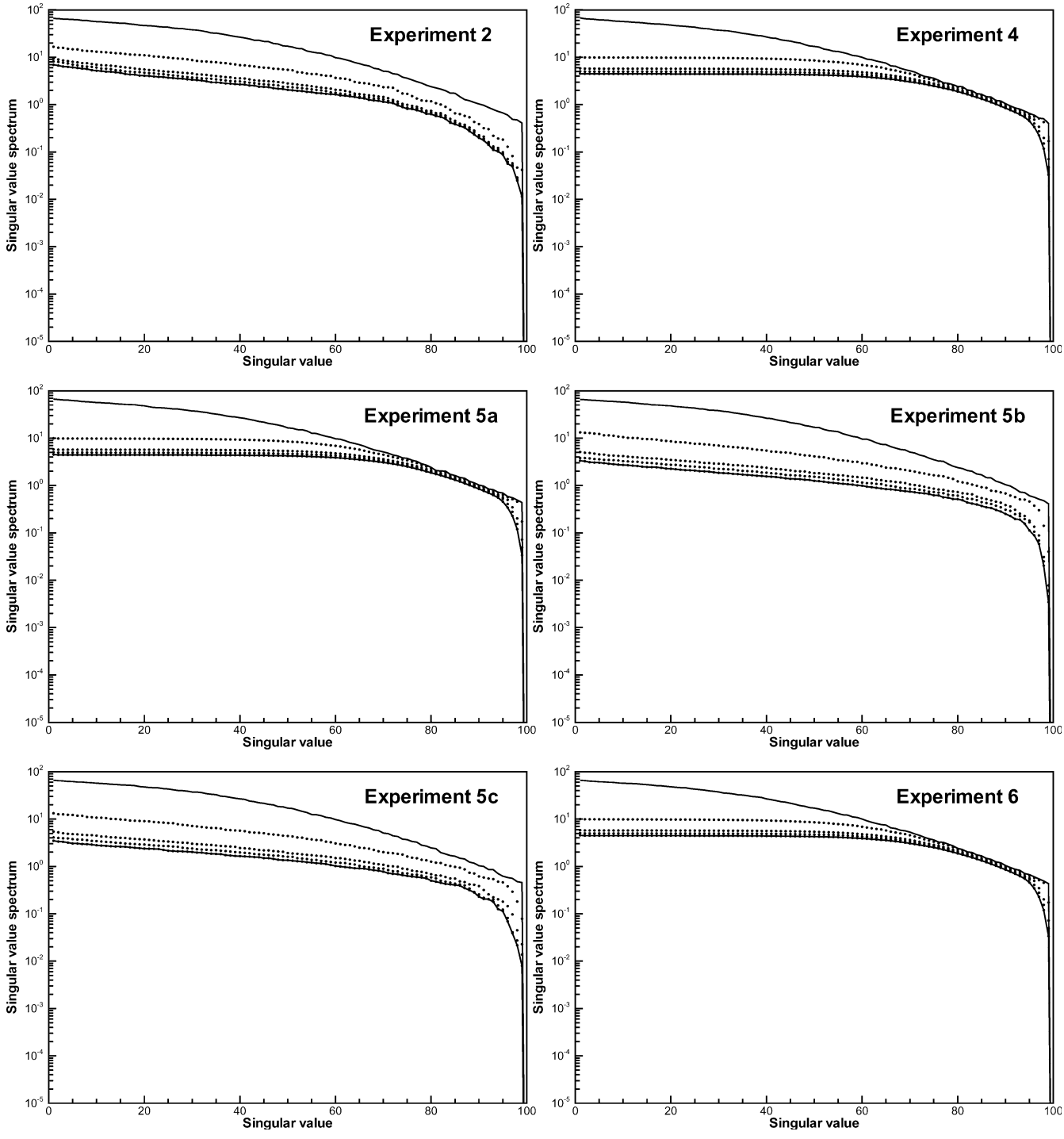

Fig. 9 Time evolution of the ensemble singular value spectra for some of the experiments

algorithms in the EnKF. The improved sampling attempts to generate ensembles with full rank and a conditioning which is better than can be obtained using random sampling. The improved sampling has been used for the generation of the initial ensemble as well as for the sampling of measurement noise.

A new analysis algorithm, which is similar to recently developed square root algorithms and which completely avoids the perturbation of measurements, has been derived and examined. It was found to provide superior results due to the removal of sampling errors which are introduced by the measurement perturbations in the standard algorithm.

In the experiments discussed here it was possible to obtain a significant improvement in the results from the standard EnKF analysis scheme if improved sampling is used both for the initial ensemble and the measurement perturbations. These results were nearly as good as those obtained using the square root algo-

rithm together with improved sampling of the initial ensemble. It is expected that similar improvements can be obtained in general since the improved sampling provides a better representation of the ensemble error covariances and of the space where the solution is sought.

A comprehensive analysis was devoted to the use of low-rank representations of the measurement error covariance matrix and it was possible to derive computationally efficient variants of the square root algorithm which exploits that the solution is searched for in a space spanned by the forecast ensemble members. This resulted in a scheme presented in Section 7.4, where a low-rank representation can be used for the measurement error covariance matrix. This algorithm solves the full problem, and turns out to be extremely efficient, and computationally more stable, compared to the traditional EnKF algorithm.

Additional experiments were performed to examine the new analysis schemes when used with a large number of measurements and low-rank representations for the measurement error covariance matrix. It was shown both theoretically and from experiments that this is does not introduce additional errors or approximations in the analysis.

It is important to point out that these results may not be directly transferable to other more complex dynamical models. In the cases discussed here the dimension of the state vector (1001 grid cells) is small compared to typical applications with ocean and atmospheric models. Thus, although we expect that the use of improved sampling schemes and/or an analysis scheme without perturbation of measurements will always lead to an improvement in the results, it is not possible to quantify this improvement in general.

We have not at all examined the potential impact a non-linear model will have on the ensemble evolution. The use of non-linear models will change the basis from that of the initial ensemble, and may even reduce the rank of the ensemble. This suggests that the improved sampling should be used for the model noise as well, to help maintain the conditioning of the ensemble during the forward integration.

The following recommendations can be given:

1. The use of high-order sampling should always be used both for the initial ensemble and the sampling of model errors.
2. The square root algorithm was superior in all experiments and a version of it should be used for the computation of the analysis.
3. The low-rank square root algorithm from Section 7.4 is computationally stable and efficient and should be used as the default algorithm. It will work properly also in the case when $m \geq N$.

It has been shown that the use of a full-rank $\boldsymbol{R}$ does not lead to any improvement in the results compared to what is obtained using low-rank approximations $\boldsymbol{R}_e = \boldsymbol{U}_0^T \boldsymbol{R} \boldsymbol{U}_0$. When an ensemble representation is used to represent $\boldsymbol{R}_e$, the ensemble of measurement perturbations needs to span the full space $\mathscr{S}$ and needs to represent $\boldsymbol{U}_0^T \boldsymbol{R} \boldsymbol{U}_0$ exactly to get the same result.

In summary, this paper has evaluated the various sampling strategies in combination with the different analysis schemes, using a very simple linear advection model. The experiments have shown that there is a potential for either a significant reduction of the computing time or an improvement of the EnKF results, using the improved sampling schemes together with the square root analysis algorithm. Further, a theoretical analysis has shown that the analysis can also be computed very efficiently using a low rank representation of the measurement error covariance matrix. Thus the new algorithm is an improvement over the algorithm from Evensen (2003) which shows the rank-loss problem found by Kepert (2004).

Appendix A Inversion and square root analysis with full-rank $\boldsymbol{R}$

This case is presented since it allows for implementation of the square root analysis scheme in existing assimilation systems where $\boldsymbol{R}$ is already given.

A.1 Pseudo-inverse of $\boldsymbol{C}$

We use the expression for $\boldsymbol{C}$ as given in Eq. (73), i.e.

$$\boldsymbol{C} \approx \boldsymbol{U}_0 \boldsymbol{\Sigma}_0 (\boldsymbol{I} + (N-1) \boldsymbol{\Sigma}_0^+ \boldsymbol{U}_0^T \boldsymbol{R} \boldsymbol{U}_0 \boldsymbol{\Sigma}_0^{+T}) \boldsymbol{\Sigma}_0^T \boldsymbol{U}_0^T \tag{94}$$

$$= \boldsymbol{U}_0 \boldsymbol{\Sigma}_0 (\boldsymbol{I} + \boldsymbol{X}_0) \boldsymbol{\Sigma}_0^T \boldsymbol{U}_0^T, \tag{95}$$

where we have defined

$$\boldsymbol{X}_0 = (N-1) \boldsymbol{\Sigma}_0^+ \boldsymbol{U}_0^T \boldsymbol{R} \boldsymbol{U}_0 \boldsymbol{\Sigma}_0^{+T} \tag{96}$$

which is an $N \times N$ matrix of with rank equal to $N-1$ and it requires $m^2 N + m N^2 + mN$ floating point operations to form it.

We then proceed with an eigenvalue decomposition

$$\boldsymbol{Z} \Lambda \boldsymbol{Z}^T = \boldsymbol{X}_0, \tag{97}$$

where all matrices are $N \times N$, and insert this in Eq. (95) to obtain

$$\boldsymbol{C} = \boldsymbol{U}_0 \boldsymbol{\Sigma}_0 (\boldsymbol{I} + \boldsymbol{Z} \Lambda \boldsymbol{Z}^T) \boldsymbol{\Sigma}_0^T \boldsymbol{U}_0^T \tag{98}$$

$$= \boldsymbol{U}_0 \boldsymbol{\Sigma}_0 \boldsymbol{Z} (\boldsymbol{I} + \Lambda) \boldsymbol{Z}^T \boldsymbol{\Sigma}_0^T \boldsymbol{U}_0^T. \tag{99}$$

Now the pseudo-inverse of $\boldsymbol{C}$ becomes

$$\boldsymbol{C}^+ = (\boldsymbol{U}_0 \boldsymbol{\Sigma}_0^{+T} \boldsymbol{Z}) (\boldsymbol{I} + \Lambda)^{-1} (\boldsymbol{U}_0 \boldsymbol{\Sigma}_0^{+T} \boldsymbol{Z})^T \tag{100}$$

$$= \boldsymbol{X}_1 (\boldsymbol{I} + \Lambda)^{-1} \boldsymbol{X}_1^T, \tag{101}$$

where we have defined $\boldsymbol{X}_1 \in \Re^{m \times N}$ of rank $N-1$ as

$$\boldsymbol{X}_1 = \boldsymbol{U}_0 \boldsymbol{\Sigma}_0^{+T} \boldsymbol{Z}. \tag{102}$$

A.2 Square root analysis algorithm

The following algorithm is an extended version of the previous square root algorithm which exploits that the solution is searched for in the space with dimension equal to the number of ensemble members. This suggests that it should not be necessary to invert an $m \times m$ matrix, to compute the analysis, when $m > N$.

Using the expression (101) for the inverse we obtain the following derivation of the analysis scheme

$$\boldsymbol{A}^{\mathrm{a}\prime}\boldsymbol{A}^{\mathrm{a}\prime\mathrm{T}} = \boldsymbol{A}'(\boldsymbol{I} - \boldsymbol{S}^{\mathrm{T}}\boldsymbol{C}^{+}\boldsymbol{S})\boldsymbol{A}'^{\mathrm{T}} \tag{103}$$

$$= \boldsymbol{A}'(\boldsymbol{I} - \boldsymbol{S}^{\mathrm{T}}\boldsymbol{X}_1(\boldsymbol{I}+\Lambda)^{-1}\boldsymbol{X}_1^{\mathrm{T}}\boldsymbol{S})\boldsymbol{A}'^{\mathrm{T}} \tag{104}$$

$$= \boldsymbol{A}'(\boldsymbol{I} - [(\boldsymbol{I}+\Lambda)^{-\frac{1}{2}}\boldsymbol{X}_1^{\mathrm{T}}\boldsymbol{S}]^{\mathrm{T}}[(\boldsymbol{I}+\Lambda)^{-\frac{1}{2}}\boldsymbol{X}_1^{\mathrm{T}}\boldsymbol{S}])\boldsymbol{A}'^{\mathrm{T}} \tag{105}$$

$$= \boldsymbol{A}'(\boldsymbol{I} - \boldsymbol{X}_2^{\mathrm{T}}\boldsymbol{X}_2)\boldsymbol{A}'^{\mathrm{T}}, \tag{106}$$

where we have defined $\boldsymbol{X}_2$ as

$$\boldsymbol{X}_2 = (\boldsymbol{I}+\Lambda)^{-\frac{1}{2}}\boldsymbol{X}_1^{\mathrm{T}}\boldsymbol{S} = (\boldsymbol{I}+\Lambda)^{-\frac{1}{2}}\boldsymbol{Z}^{\mathrm{T}}\tilde{\boldsymbol{I}}_{N-1}\boldsymbol{V}_0^{\mathrm{T}}, \tag{107}$$

which also has rank equal to $N-1$. We then end up with the same final update equation (34) by following the derivation defined in Eqs. (29–33). Thus, here we compute the factorization of an $N \times N$ matrix, but we still need to evaluate one expensive matrix multiplication involving the $m \times m$ matrix $\boldsymbol{R}$ in Eq. (96).

Appendix B Final update equation in the square root algorithms

In Evensen (2003) it was shown that the EnKF analysis update can be written as

$$\boldsymbol{A}^{\mathrm{a}} = \boldsymbol{A}^{\mathrm{f}}\boldsymbol{X}, \tag{108}$$

where $\boldsymbol{X}$ is an $N \times N$ matrix of coefficients. The square root schemes presented in this paper can also be written in the same simple form.

This is illustrated using the algorithm from Section 3, where the analysis is written as the updated ensemble mean plus the updated ensemble perturbations,

$$\boldsymbol{A}^{\mathrm{a}} = \bar{\boldsymbol{A}}^{\mathrm{a}} + \boldsymbol{A}^{\mathrm{a}\prime}. \tag{109}$$

The updated mean can, using Eq. (23), be written as

$$\overline{\boldsymbol{A}}^{\mathrm{a}} = \overline{\boldsymbol{A}}^{\mathrm{f}} + \boldsymbol{A}^{\mathrm{f}\prime}\boldsymbol{S}^{\mathrm{T}}\boldsymbol{C}^{-1}(\overline{\boldsymbol{D}} - \boldsymbol{H}\overline{\boldsymbol{A}}^{\mathrm{f}}) \tag{110}$$

$$= \boldsymbol{A}^{\mathrm{f}}\boldsymbol{1}_N + \boldsymbol{A}^{\mathrm{f}}(\boldsymbol{I} - \boldsymbol{1}_N)\boldsymbol{S}^{\mathrm{T}}\boldsymbol{C}^{-1}(\boldsymbol{D} - \boldsymbol{H}\boldsymbol{A}^{\mathrm{f}})\boldsymbol{1}_N, \tag{111}$$

and the updated perturbations are, from Eq. (34),

$$\boldsymbol{A}^{\mathrm{a}\prime} = \boldsymbol{A}^{\mathrm{f}\prime}\boldsymbol{V}_2\sqrt{\boldsymbol{I} - \boldsymbol{\Sigma}_2^{\mathrm{T}}\boldsymbol{\Sigma}_2} \tag{112}$$

$$= \boldsymbol{A}^{\mathrm{f}}(\boldsymbol{I} - \boldsymbol{1}_N)\boldsymbol{V}_2\sqrt{\boldsymbol{I} - \boldsymbol{\Sigma}_2^{\mathrm{T}}\boldsymbol{\Sigma}_2\Theta^{\mathrm{T}}}. \tag{113}$$

Combining the previous equations we get (108) with $\boldsymbol{X}$ defined as

$$\boldsymbol{X} = \boldsymbol{1}_N + (\boldsymbol{I} - \boldsymbol{1}_N)\boldsymbol{S}^{\mathrm{T}}\boldsymbol{C}^{-1}(\boldsymbol{D} - \boldsymbol{H}\boldsymbol{A}^{\mathrm{f}})\boldsymbol{1}_N + (\boldsymbol{I} - \boldsymbol{1}_N)\boldsymbol{V}_2\sqrt{\boldsymbol{I} - \boldsymbol{\Sigma}_2^{\mathrm{T}}\boldsymbol{\Sigma}_2\Theta^{\mathrm{T}}}. \tag{114}$$

Thus, we still search for the solution as a combination of ensemble members as was discussed in Evensen (2003). It also turns out that forming $\boldsymbol{X}$ and then computing the matrix multiplication in Eq. (108) is the most efficient algorithm for computing the analysis when many measurements are used.

Acknowlegements The author is grateful for the detailed comments provided by an anonymous reviewer, which contributed significantly to improving the readability of the manuscript.

References

Anderson JL (2001) An ensemble adjustment Kalman filter for data assimilation. Mon Weather Rev 129: 2884–2903

Bennett, AF (1992) Inverse methods in physical oceanography. Cambridge University Press

Bishop CH, Etherton BJ, Majumdar SJ (2001) Adaptive sampling with the ensemble transform Kalman filter, part I: theoretical aspects. Mon Weather Rev 129: 420–436

Burgers G, van Leeuwen PJ, Evensen G (1998) Analysis scheme in the ensemble Kalman filter. Mon Weather Rev 126: 1719–1724

Evensen G (1994) Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. J Geophys Res 99: 10,143–10,162

Evensen, G (2003) The Ensemble Kalman Filter: theoretical formulation and practical implementation. Ocean Dynamics 53: 343–367

Kepert JD (2004) On ensemble representation of the observation – error convariance in the Ensemble Kalman Filter. Ocean Dynamics 6: 561–569

Nerger L (2004) Parallel filter algorithms for data assimilation in oceanography, Reports on Polar and Marine Research 487, Alfred Wegener Institute for Polar and Marine Research, Bremerhaven, Germany, Postfach 12 0161, D-27515 Bremerhaven, PhD thesis, University of Bremen

Pham DT (2001) Stochastic methods for sequential data assimilation in strongly nonlinear systems. Mon Weather Rev 129: 1194–1207

Tippett MK, Anderson JL, Bishop CH, Hamill TM, Whitaker JS (2003) Ensemble square-root filters Mon Weather Rev 131: 1485–1490

Whitaker JS, Hamill TM (2002) Ensemble data assimilation without perturbed observations. Mon Weather Rev 130: 1913–1924