Covariance inflation in the ensemble Kalman filter: a residual nudging perspective and some implications

Xiaodong Luo *

International Research Institute Of Stavanger (IRIS), 5008 Bergen, Norway

Ibrahim Hoteit

King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia

* *Corresponding author address:* International Research Institute Of Stavanger (IRIS), Thormøhlens Gate 55, 5008 Bergen, Norway

E-mail: xiaodong.luo@iris.no

ABSTRACT

This note examines the influence of covariance inflation on the distance between the measured observation and the simulated (or predicted) observation with respect to the state estimate. In order for the aforementioned distance to be bounded in a certain interval, some sufficient conditions are derived, indicating that the covariance inflation factor should be bounded in a certain interval, and that the inflation bounds are related to the maximum and minimum eigenvalues of certain matrices. Implications of these analytic results are discussed, and a numerical experiment is presented to verify the validity of our analysis.

1. Data assimilation with residual nudging

A finite, often small, ensemble size has some well known effects that may substantially influence the behaviour of an ensemble Kalman filter (EnKF). These effects include, for instance, rank deficient sample error covariance matrices, systematically underestimated error variances, and in contrast, exceedingly large error cross-covariances of the model state variables (Whitaker and Hamill 2002). In the literature, the latter two issues are often tackled through covariance localization (Hamill et al. 2001), while the first issue, under-estimation of sample variances, is often handled by covariance inflation (Anderson and Anderson 1999), in which one artificially increases the sample variances, either multiplicatively (see, for example, Anderson and Anderson 1999; Anderson 2007, 2009; Bocquet and Sakov 2012; Miyoshi 2011), or additively (see, for example, Hamill and Whitaker 2011), or in a hybrid way by combining both multiplicative and additive inflation methods (see, for example, Whitaker and Hamill 2012), or through other ways such as relaxation to the prior

(Zhang et al. 2004), multi-scheme ensembles (Meng and Zhang 2007), modification of the eigenvalues of sample error covariance matrices (Altaf et al. 2013; Luo and Hoteit 2011; Ott et al. 2004; Triantafyllou et al. 2013), back projection of the residuals to construct new ensemble members Song et al. (2010) to name but a few. In general, covariance inflation tends to increase the robustness of the EnKF against uncertainties in data assimilation (Luo and Hoteit 2011), and often also improves the filter performance in terms of estimation accuracy.

The focus of this note is to study the effect of covariance inflation from the point of view of residual nudging (Luo and Hoteit 2012). Here, the "residual" with respect to an m-dimensional system state $\mathbf{x}$ is a vector in the observation space, defined as $\mathbf{H}\mathbf{x} - \mathbf{y}$ [1], where $\mathbf{H} : \mathbb{R}^m \rightarrow \mathbb{R}^p$ is a linear observation operator, and $\mathbf{y}$ the corresponding p-dimensional observation vector. Throughout this note, our discussion is confined to the filtering (or analysis) step of the EnKF, so that the time index in the EnKF is dropped. The linearity assumption in the observation operator $\mathbf{H}$ is taken in order to simplify our discussion. The result to be presented later, though, might also provide insights into more complex situations.

Before introducing the concept of residual nudging, let us define some additional notations. We assume that the observation system is given by

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{v} \, , \tag{1}$$

where $\mathbf{v}$ is the vector of observation error, with zero mean and a non-singular covariance matrix $\mathbf{R}$. We further decompose $\mathbf{R}$ as $\mathbf{R} = \mathbf{R}^{1/2} \, \mathbf{R}^{T/2}$, where $\mathbf{R}^{1/2}$ is a non-singular square root of $\mathbf{R}$ and $\mathbf{R}^{T/2}$ denotes the transpose of $\mathbf{R}^{1/2}$.

[1] In the literature, the vector with the opposite sign, $\mathbf{y} - \mathbf{H}\mathbf{x}$, is often called "innovation".

To measure the length of a vector $\mathbf{z}$ in the observation space, we adopt the following weighted Euclidean norm

$$\|\mathbf{z}\|_{\mathbf{R}} \equiv \sqrt{\mathbf{z}^T \, \mathbf{R}^{-1} \, \mathbf{z}} \, . \tag{2}$$

One may convert the weighted Euclidean norm to the standard Euclidean norm by noticing that $\|\mathbf{z}\|_{\mathbf{R}} = \|\mathbf{R}^{-1/2} \, \mathbf{z}\|_2$, where $\| \bullet \|_2$ denotes the standard Euclidean norm. As a result, many topological properties with respect to the standard Euclidean norm, e.g., the triangle inequality (see (3) below), still hold with respect to the weighted Euclidean norm.

The idea of data assimilation with residual nudging (DARN) is the following. Let $\mathbf{x}^{tr}$ be the true system state (truth), $\mathbf{y}^o = \mathbf{H}\mathbf{x}^{tr} + \mathbf{v}^o$ the recorded observation for a specific realization $\mathbf{v}^o$ of the observation error, and $\hat{\mathbf{x}}$ the state estimate (e.g., either the prior or posterior estimate) obtained from a data assimilation (DA) algorithm. Then the residual $\hat{\mathbf{r}} = \mathbf{H}\hat{\mathbf{x}} - \mathbf{y}^o = \mathbf{H}\hat{\mathbf{x}} - \mathbf{H}\mathbf{x}^{tr} - \mathbf{v}^o$. By the triangle inequality, the weighted Euclidean norm of the residual (residual norm hereafter) satisfies

$$\|\hat{\mathbf{r}}\|_{\mathbf{R}} \leq \|\mathbf{H}\hat{\mathbf{x}} - \mathbf{H}\mathbf{x}^{tr}\|_{\mathbf{R}} + \|\mathbf{v}^o\|_{\mathbf{R}} \, . \tag{3}$$

If the DA algorithm performs reasonably well, one may expect that the magnitude of $\|\mathbf{H}\hat{\mathbf{x}} - \mathbf{H}\mathbf{x}^{tr}\|_{\mathbf{R}}$ not be significantly larger than $\|\mathbf{v}^o\|_{\mathbf{R}}$. As a result, one may obtain an upper bound of $\|\hat{\mathbf{r}}\|_{\mathbf{R}}$ in terms of $\|\mathbf{v}^o\|_{\mathbf{R}}$, e.g, in the form of $\beta\|\mathbf{v}^o\|_{\mathbf{R}}$, where β is a non-negative scalar coefficient. In practice, though, $\|\mathbf{v}^o\|_{\mathbf{R}}$ is often unknown. As a remedy, we replace $\|\mathbf{v}^o\|_{\mathbf{R}}$ by an upper bound of the expectation $\mathbb{E}(\|\mathbf{v}\|_{\mathbf{R}})$ of the weighted Euclidean norm of the observation error $\mathbf{v}$, where $\mathbb{E}$ denotes the expectation operator. One such upper bound can be obtained by noticing that

$$\left(\mathbb{E}(\|\mathbf{v}\|_{\mathbf{R}})\right)^2 \leq \mathbb{E}(\|\mathbf{v}\|_{\mathbf{R}}^2) = \mathrm{trace}\left(\mathbf{R}^{-1}\mathbb{E}(\mathbf{v}\mathbf{v}^T)\right) = \mathrm{trace}(\mathbf{I}_p) = p \, , \tag{4}$$

where the operator "trace" evaluates the trace of a matrix, and $\mathbf{I}_p$ the p-dimensional identity matrix. From (4), we have the upper bound $\mathbb{E}(\|\mathbf{v}\|_{\mathbf{R}}) \leq \sqrt{p}$. Consequently, we want to find a state estimate $\hat{\mathbf{x}}$ whose residual norm $\|\hat{\mathbf{r}}\|_{\mathbf{R}}$ satisfies

$$\|\hat{\mathbf{r}}\|_{\mathbf{R}} \leq \beta\sqrt{p} \tag{5}$$

for a pre-chosen β. It is worthy of mentioning that in general it may be difficult to identity which β gives the best state estimation accuracy with respect to the truth $\mathbf{x}^{tr}$. Therefore, in Luo and Hoteit (2012) we mainly used DARN as a safeguard strategy, that is, if a state estimate $\hat{\mathbf{x}}$ is found to have a too large residual norm, then we try to introduce some correction to the state estimate in order to reduce its residual norm, which in turn might also improve the estimation accuracy.

In Luo and Hoteit (2012) we introduced DARN to the analysis $\hat{\mathbf{x}}^a$ in the ensemble adjustment Kalman filter (EAKF, see Anderson 2001). In the EAKF with residual nudging (EAKF-RN), if the residual norm of $\hat{\mathbf{x}}^a$ is less than $\beta\sqrt{p}$, then we accept $\hat{\mathbf{x}}^a$ as a reasonable estimate and no change is made. Otherwise, a correction is introduced to $\hat{\mathbf{x}}^a$ in a way such that the residual norm of the modified state estimate $\tilde{\mathbf{x}}^a$ is exactly $\beta\sqrt{p}$, and that among all possible state estimates whose residual norms are equal to $\beta\sqrt{p}$, the simulated (or predicted) observation $\mathbf{H}\tilde{\mathbf{x}}^a$ of the modified state estimate $\tilde{\mathbf{x}}^a$ has the shortest distance to the one $\mathbf{H}\hat{\mathbf{x}}^a$ of the original state estimate $\hat{\mathbf{x}}^a$. Numerical results in Luo and Hoteit (2012) show that the EAKF-RN exhibits (sometimes substantially) improved filter performance, in terms of estimation accuracy and/or stability against filter divergence, compared to the EAKF. Extension of DARN to other types of filters is also possible, for example, see Luo and Hoteit (2013).

2. Covariance inflation from the point of view of residual nudging

Here we examine the effect of covariance inflation on the analysis residual norm. To this end, we first recall that the mean update formula in the EnKF (without perturbing the observation) is given by

$$\begin{aligned} \hat{\mathbf{x}}^a &= \hat{\mathbf{x}}^b + \mathbf{K}\left(\mathbf{y}^o - \mathbf{H}\hat{\mathbf{x}}^b\right) \,, \\ \mathbf{K} &= \hat{\mathbf{C}}^b \mathbf{H}^T \left(\mathbf{H}\hat{\mathbf{C}}^b\mathbf{H}^T + \mathbf{R}\right)^{-1} \,, \end{aligned} \tag{6}$$

where $\hat{\mathbf{x}}^b$ and $\hat{\mathbf{x}}^a$ are the sample means of the background and analysis ensembles, respectively; $\mathbf{K}$ is the Kalman gain; and $\hat{\mathbf{C}}^b$ is a certain symmetric, positive semi-definite matrix in accordance to the chosen inflation scheme. In general $\hat{\mathbf{C}}^b$ may be related, but not necessarily proportional, to the sample error covariance matrix $\hat{\mathbf{P}}^b$ of the background ensemble. For instance, in the hybrid EnKF $\hat{\mathbf{C}}^b$ can be a mixture of $\hat{\mathbf{P}}^b$ and a "background covariance" $\mathbf{B}$ (Hamill and Snyder 2000), or partially time-varying as in Hoteit et al. (2002).

Our objective is to examine under which conditions the residual norm $\|\hat{\mathbf{r}}^a\|_{\mathbf{R}}$ of the analysis $\hat{\mathbf{x}}^a$ satisfies $\beta_l \sqrt{p} \le \|\hat{\mathbf{r}}^a\|_{\mathbf{R}} \le \beta_u \sqrt{p}$, where β_l and β_u ($0 \le \beta_l \le \beta_u$) represents the lower and upper values of β that one wants to set for the analysis residual norm in DARN. Different from the previous works (Luo and Hoteit 2012, 2013), the lower bound $\beta_l \sqrt{p}$ is introduced here in order to make our discussion below slightly more general. In practice it may also be used to prevent too small residual norms in certain circumstances in order to avoid, for instance, a state estimate that over-fits the observation, a phenomenon that may be caused by "over-inflation", as will be shown later.

Inserting Eq. (6) into $\hat{\mathbf{r}}^a = \mathbf{H}\hat{\mathbf{x}}^a - \mathbf{y}^o$, one has

$$\hat{\mathbf{r}}^a = \mathbf{R}\left(\mathbf{H}\hat{\mathbf{C}}^b\mathbf{H}^T + \mathbf{R}\right)^{-1}\hat{\mathbf{r}}^b\,, \tag{7}$$

where $\hat{\mathbf{r}}^b = \mathbf{H}\hat{\mathbf{x}}^b - \mathbf{y}^o$. Multiplying both sides of Eq. (7) by $\mathbf{R}^{-1/2}$, one obtains

$$(\mathbf{R}^{-1/2}\hat{\mathbf{r}}^a) = \left(\mathbf{R}^{-1/2}\mathbf{H}\hat{\mathbf{C}}^b\mathbf{H}^T\mathbf{R}^{-T/2} + \mathbf{I}_p\right)^{-1}(\mathbf{R}^{-1/2}\hat{\mathbf{r}}^b)\,. \tag{8}$$

To derive the bounded residual norm, we first consider under which conditions the upper bound $\|\hat{\mathbf{r}}^a\|_{\mathbf{R}} \le \beta_u\sqrt{p}$ is guaranteed to hold. Given that (cf (19) later)

$$\|\hat{\mathbf{r}}^a\|_{\mathbf{R}} = \|\mathbf{R}^{-1/2}\hat{\mathbf{r}}^a\|_2 \le \|(\mathbf{R}^{-1/2}\mathbf{H}\hat{\mathbf{C}}^b\mathbf{H}^T\mathbf{R}^{-T/2} + \mathbf{I}_p)^{-1}\|_2\;\|\hat{\mathbf{r}}^b\|_{\mathbf{R}}\,, \tag{9}$$

a sufficient condition is thus

$$\|(\mathbf{R}^{-1/2}\mathbf{H}\hat{\mathbf{C}}^b\mathbf{H}^T\mathbf{R}^{-T/2} + \mathbf{I}_p)^{-1}\|_2 \le \frac{\beta_u\sqrt{p}}{\|\hat{\mathbf{r}}^b\|_{\mathbf{R}}}\,. \tag{10}$$

Let

$$\mathbf{A} = \mathbf{R}^{-1/2}\mathbf{H}\hat{\mathbf{C}}^b\mathbf{H}^T\mathbf{R}^{-T/2}\,, \tag{11}$$

and λ_{max} and λ_{min} be the maximum and minimum eigenvalues of $\mathbf{A}$, respectively. Recalling that the induced 2-norm of a symmetric positive semi-definite matrix is exactly the maximum eigenvalue of that matrix (Horn and Johnson 1990, §5.6.6), we have

$$\|(\mathbf{A} + \mathbf{I}_p)^{-1}\|_2 = (\lambda_{min} + 1)^{-1}\,. \tag{12}$$

Therefore (10) leads to

$$\lambda_{min} + 1 \ge \frac{\|\hat{\mathbf{r}}^b\|_{\mathbf{R}}}{\beta_u\sqrt{p}}\,. \tag{13}$$

If $\|\hat{\mathbf{r}}^b\|_{\mathbf{R}}$ is relatively small such that $\|\hat{\mathbf{r}}^b\|_{\mathbf{R}} \le \beta_u\sqrt{p}$, then (13) automatically holds. However, if $\|\hat{\mathbf{r}}^b\|_{\mathbf{R}} > \beta_u\sqrt{p}$, and that λ_{min} is very small, then there is no guarantee that (13) will hold. A small λ_{min} may appear, for instance, when the ensemble size n is smaller than the dimension p of the observation space. In such circumstances, the matrix $\mathbf{A}$ may be singular with $\lambda_{min} = 0$, and the singularity may not be avoided only through the multiplicative covariance inflation. If one cannot afford to increase the ensemble size n, then a few alternative strategies may be adopted to address (or at least mitigate) the problem of singularity. These include, for instance, (a) introducing covariance localization (Hamill et al. 2001) to $\hat{\mathbf{P}}^b$ in order to increase its rank (Hamill et al. 2009); (b) replacing the sample error covariance $\hat{\mathbf{P}}^b$ by a hybrid of $\hat{\mathbf{P}}^b$ and some full-rank matrix, similar to that in Hamill and Snyder (2000); and (c) reducing the dimension p of the observation in the update formula, for instance, by assimilating the observation in a serial way (see, for example, Whitaker and Hamill 2002), or by assimilating the observation in the framework of local EnKF (see, for example, Bocquet 2011; Ott et al. 2004). Once the problem of singularity is solved so that the smallest eigenvalue of $\mathbf{A}$ becomes positive, a (large enough) multiplicative inflation factor can be introduced to make sure that (13) holds.

Inequality (13) provides insights of what the constraints there may be in choosing the inflation factor. In what follows, we study the problem in a slightly more general setting. Concretely, we consider a family of mean update formulae in the form of

$$\hat{\mathbf{x}}^a = \hat{\mathbf{x}}^b + \mathbf{G}\left(\mathbf{y}^o - \mathbf{H}\hat{\mathbf{x}}^b\right) , \tag{14a}$$

$$\mathbf{G} = \alpha\, \hat{\mathbf{C}}^b\mathbf{H}^T \left(\delta\, \mathbf{H}\hat{\mathbf{C}}^b\, \mathbf{H}^T + \gamma\, \mathbf{R}\right)^{-1} , \tag{14b}$$

where α, δ and γ are some positive coefficients, and $\mathbf{G}$ is the gain matrix which in general differs from the Kalman gain $\mathbf{K}$ in Eq. (6) with the presence of these three extra coefficients. Without loss of generality, though, one may let $\alpha = 1$ (e.g., by moving α inside the parentheses) so that the gain matrix is simplified to

$$\mathbf{G} = \hat{\mathbf{C}}^b \mathbf{H}^T \left(\delta \, \mathbf{H} \hat{\mathbf{C}}^b \mathbf{H}^T + \gamma \, \mathbf{R} \right)^{-1} \, , \text{ with } \delta > 0 \text{ and } \gamma > 0. \tag{15}$$

If $\delta = 1$, then $\mathbf{G}$ resembles the Kalman gain in the EnKF, with $1/\gamma$ being analogous to the multiplicative covariance inflation factor as used in Anderson and Anderson (1999). In our discussion below, we first derive some inflation constraints in the general case with $\delta > 0$, and then examine the more specific situation with $\delta = 1$. It is expected that one can also obtain constraints for other types of inflations in a similar way, but the results themselves may be case-dependent.

Using Eqs. (14a) and (15) as the update formulae and with some algebra, the weighted residual is given by

$$(\mathbf{R}^{-1/2} \hat{\mathbf{r}}^a) = \left[\mathbf{I}_p - \mathbf{A} \left(\delta \, \mathbf{A} + \gamma \mathbf{I}_p \right)^{-1} \right] (\mathbf{R}^{-1/2} \hat{\mathbf{r}}^b) \, , \tag{16}$$

where $\hat{\mathbf{r}}^a$, $\hat{\mathbf{r}}^b$ and $\mathbf{A}$ are defined as previously. Let

$$\begin{aligned} \Phi &\equiv \mathbf{I}_p - \mathbf{A} \left(\delta \, \mathbf{A} + \gamma \mathbf{I}_p \right)^{-1} \\ &= \frac{\delta - 1}{\delta} \mathbf{I}_p + \frac{\gamma}{\delta} \left(\delta \, \mathbf{A} + \gamma \mathbf{I}_p \right)^{-1} \; , \end{aligned} \tag{17}$$

then one has

$$\| \hat{\mathbf{r}}^a \|_{\mathbf{R}} = \| \mathbf{R}^{-1/2} \hat{\mathbf{r}}^a \|_2 = \| \Phi \, (\mathbf{R}^{-1/2} \hat{\mathbf{r}}^b) \|_2 \, . \tag{18}$$

For our purpose, the following two matrix inequalities are useful. Firstly, given a matrix $\mathbf{M}$ and a vector $\mathbf{z}$ with suitable dimensions, one has

$$\| \mathbf{M} \, \mathbf{z} \|_2 \leq \| \mathbf{M} \|_2 \, \| \mathbf{z} \|_2 \, , \tag{19}$$

where $\|\mathbf{M}\|_2$, the induced 2-norm of $\mathbf{M}$, is the maximum of the absolute singular values of $\mathbf{M}$, or equivalently, $\|\mathbf{M}\|_2$ is equal to the square root of the largest eigenvalue of $\mathbf{M}\,\mathbf{M}^T$

(Horn and Johnson 1990, ch. 5). Secondly, if in addition $\mathbf{M}$ is non-singular, then (see, e.g., Grcar 2010 and the references therein)

$$\|\mathbf{M}^{-1}\|_2^{-1}\,\|\mathbf{z}\|_2 \le \|\mathbf{M}\mathbf{z}\|_2\,. \tag{20}$$

The first inequality, (19), can be applied to obtain the sufficient conditions under which the inequality $\|\hat{\mathbf{r}}^a\|_{\mathbf{R}} \le \beta_u\sqrt{p}$ is achieved. Let the maximum and minimum eigenvalues of Φ be μ_{max} and μ_{min}, respectively. Then by Eq. (17)

$$\mu_{max} = \frac{\delta-1}{\delta} + \frac{\gamma}{\delta}\,(\delta\,\lambda_{min}+\gamma)^{-1}\,, \tag{21a}$$

$$\mu_{min} = \frac{\delta-1}{\delta} + \frac{\gamma}{\delta}\,(\delta\,\lambda_{max}+\gamma)^{-1}\,. \tag{21b}$$

We remark that both μ_{max} and μ_{min} can be negative (e.g., when $\delta < 1$ and $\gamma \to 0$), therefore $\|\Phi\|_2 = \max(|\mu_{max}|, |\mu_{min}|)$. By (18) and (19), a sufficient condition for $\|\hat{\mathbf{r}}^a\|_{\mathbf{R}} \le \beta_u\sqrt{p}$ is $\max(|\mu_{max}|, |\mu_{min}|) \le \beta_u\sqrt{p}/\|\hat{\mathbf{r}}^b\|_{\mathbf{R}}$. For notational convenience, we define $\xi_u \equiv \beta_u\sqrt{p}/\|\hat{\mathbf{r}}^b\|_{\mathbf{R}}$ and $\xi_l \equiv \beta_l\sqrt{p}/\|\hat{\mathbf{r}}^b\|_{\mathbf{R}}$.

Depending on the signs and magnitudes of μ_{max} and μ_{min}, there are in general four possible scenarios: (a) $\mu_{max} \ge 0$ and $\mu_{min} \ge 0$, so that $\|\Phi\|_2 = \mu_{max}$; (b) $\mu_{max} \le 0$ and $\mu_{min} \le 0$, so that $\|\Phi\|_2 = -\mu_{min}$; (c) $\mu_{max} \ge 0$, $\mu_{min} \le 0$ and $\mu_{max} + \mu_{min} \ge 0$, so that $\|\Phi\|_2 = \mu_{max}$; and (d) $\mu_{max} \ge 0$, $\mu_{min} \le 0$ and $\mu_{max} + \mu_{min} \le 0$, so that $\|\Phi\|_2 = -\mu_{min}$. Inserting Eq. (21) into the above conditions one obtains some inequalities with respect to the variables δ and γ (subject to $\delta > 0$ and $\gamma > 0$), which are omitted in this note for brevity.

Similarly, the second inequality, (20), can be used to find the sufficient conditions for $\beta_l\sqrt{p} \le \|\hat{\mathbf{r}}^a\|_{\mathbf{R}}$. By (18) and (20), one such sufficient condition can be $\|\Phi^{-1}\|_2 \le \|\hat{\mathbf{r}}^b\|_{\mathbf{R}}/(\beta_l\sqrt{p}) = 1/\xi_l$. By Eq. (17) it can be shown that

$$\Phi^{-1} = \mathbf{I}_p + \left((\delta-1)\,\mathbf{I}_p + \gamma\,\mathbf{A}^{-1}\right)^{-1}\,. \tag{22}$$

Let the maximum and minimum eigenvalues of Φ^{-1} be ν_{max} and ν_{min}, respectively, then

$$\nu_{max} = 1 + \lambda_{max} \left((\delta - 1) \lambda_{max} + \gamma \right)^{-1} , \tag{23a}$$

$$\nu_{min} = 1 + \lambda_{min} \left((\delta - 1) \lambda_{min} + \gamma \right)^{-1} . \tag{23b}$$

Similar to the previous discussion, we require that $\|\Phi^{-1}\|_2 = \max(|\nu_{max}|, |\nu_{min}|) \leq 1/\xi_l$, which also leads to four possible scenarios: (a) $\nu_{max} \geq 0$ and $\nu_{min} \geq 0$, so that $\|\Phi^{-1}\|_2 = \nu_{max}$; (b) $\nu_{max} \leq 0$ and $\nu_{min} \leq 0$, so that $\|\Phi^{-1}\|_2 = -\nu_{min}$; (c) $\nu_{max} \geq 0$, $\nu_{min} \leq 0$ and $\nu_{max} + \nu_{min} \geq 0$, so that $\|\Phi^{-1}\|_2 = \nu_{max}$; and (d) $\nu_{max} \geq 0$, $\nu_{min} \leq 0$ and $\nu_{max} + \nu_{min} \leq 0$, so that $\|\Phi^{-1}\|_2 = -\nu_{min}$. Again, inserting Eq. (23) into the above conditions one obtains some inequalities with respect to the variables δ and γ.

Despite the complexity in the general situation, the analysis in the case of $\delta = 1$ (corresponding to the update formula in the EnKF) is significantly simplified. Indeed, when $\delta = 1$, the maximum and minimum eigenvalues in Eqs. (21) and (23) are all positive. Therefore the following conditions

$$\mu_{max} = \gamma \left(\lambda_{min} + \gamma \right)^{-1} \leq \xi_u , \tag{24a}$$

$$\nu_{max} = 1 + \lambda_{max}/\gamma \leq 1/\xi_l . \tag{24b}$$

are sufficient for the objective $\beta_l \sqrt{p} \leq \|\hat{\mathbf{r}}^a\|_{\mathbf{R}} \leq \beta_u \sqrt{p}$. Note that if $\xi_u \geq 1$, i.e., $\|\hat{\mathbf{r}}^b\|_{\mathbf{R}} \leq \beta_u \sqrt{p}$, then any $\gamma > 0$ would guarantee that $\|\hat{\mathbf{r}}^a\|_{\mathbf{R}} \leq \beta_u \sqrt{p}$ (indeed by Eqs. (16) and (19) the analysis residual norm $\|\hat{\mathbf{r}}^a\|_{\mathbf{R}}$ is guaranteed to be no larger than $\|\hat{\mathbf{r}}^b\|_{\mathbf{R}}$ since $\|\Phi\|_2 \leq 1$ with $\delta = 1$), and that inequality (24a) holds. On the other hand, if $\xi_l \geq 1$ such that $\|\hat{\mathbf{r}}^b\|_{\mathbf{R}} \leq \beta_l \sqrt{p}$, then in most cases[2] it is impossible for the EnKF to have $\|\hat{\mathbf{r}}^a\|_{\mathbf{R}}$ no less

[2] An exception is in the case that $\gamma = +\infty$ and $\xi_l = 1$. This implies that $\|\hat{\mathbf{r}}^a\|_{\mathbf{R}} = \|\hat{\mathbf{r}}^b\|_{\mathbf{R}} = \beta_l \sqrt{p}$, and that no mean update is conducted (i.e., $\hat{\mathbf{x}}^a = \hat{\mathbf{x}}^b$).

than $\|\hat{\mathbf{r}}^b\|_{\mathbf{R}}$ (hence $\beta_l\sqrt{p}$), for the same aforementioned reason. Therefore the inequality (24b) becomes infeasible. With these said, in what follows we focus on the cases in which $\xi_u, \xi_l \in [0, 1)$. With some algebra, it can be shown that γ should be bounded by

$$\frac{\xi_l}{1-\xi_l}\,\lambda_{max} \leq \gamma \leq \frac{\xi_u}{1-\xi_u}\,\lambda_{min}\,. \tag{25}$$

Let $\kappa = \lambda_{max}/\lambda_{min}$ be the condition number of the (normalized) matrix $\mathbf{A} = \mathbf{R}^{-1/2}\mathbf{H}\hat{\mathbf{C}}^b\mathbf{H}^T\mathbf{R}^{-T/2}$. From (25) we have $\dfrac{\xi_l}{1-\xi_l}\,\lambda_{max} \leq \dfrac{\xi_u}{1-\xi_u}\,\lambda_{min}$, which leads to a constraint in choosing β_l and β_u, in terms of

$$\beta_l \leq \frac{\beta_u}{\kappa + (1-\kappa)\,\xi_u}\,. \tag{26}$$

Inequality (25) suggests that the upper and lower bounds of γ are related to the minimum and maximum eigenvalues of $\mathbf{A}$, respectively. In particular, to avoid a too small residual norm, i.e., observation over-fitting, γ should be lower bounded, hence its inverse $1/\gamma$, resembling the multiplicative inflation factor, should be upper bounded, as mentioned previously.

In practice, if the dimension p of the observation space is large, then it may be expensive to evaluate λ_{max} and λ_{min}. In certain circumstances, though, there may be cheaper ways to compute an interval for γ. For instance, if $\hat{\mathbf{C}}^b$ in the mean update formula is in the form of $c_1\,\hat{\mathbf{P}}^b + c_2\,\mathbf{B}$ with c_1 and c_2 being some positive scalars and $\mathbf{B}$ a constant, symmetric and positive-definite matrix, then

$$\mathbf{A} = c_1\,\mathbf{R}^{-1/2}\mathbf{H}\hat{\mathbf{P}}^b\mathbf{H}^T\mathbf{R}^{-T/2} + c_2\,\mathbf{R}^{-1/2}\mathbf{H}\mathbf{B}\mathbf{H}^T\mathbf{R}^{-T/2}\,.$$

The additive Weyl inequality (Horn and Johnson 1991, ch. 3) suggests that the following bounds hold for λ_{max} and λ_{min}.

$$\lambda_{max} \leq c_1 \, \tau_{max} + c_2 \, \rho_{max} \, , \qquad \lambda_{min} \geq c_1 \, \tau_{min} + c_2 \, \rho_{min} \geq c_2 \, \rho_{min} \, , \tag{27}$$

where τ and ρ are the eigenvalues of $\mathbf{R}^{-1/2}\mathbf{H}\hat{\mathbf{P}}^b\mathbf{H}^T\mathbf{R}^{-T/2}$ and $\mathbf{R}^{-1/2}\mathbf{H}\mathbf{B}\mathbf{H}^T\mathbf{R}^{-T/2}$, respectively. In many situations, $\hat{\mathbf{P}}^b$ may be rank deficient, therefore a singular value decomposition (SVD) analysis shows that τ_{max} is equal to the largest eigenvalue of $(\mathbf{H}\hat{\mathbf{S}}^b)^T\mathbf{R}^{-1}(\mathbf{H}\hat{\mathbf{S}}^b)$, where $\hat{\mathbf{S}}^b$ is a square root of $\hat{\mathbf{P}}^b$ that can be directly constructed based on the background ensemble (Bishop et al. 2001; Luo and Moroz 2009; Wang et al. 2004). Note that $(\mathbf{H}\hat{\mathbf{S}}^b)^T\mathbf{R}^{-1}(\mathbf{H}\hat{\mathbf{S}}^b)$ is a matrix with its dimension determined by the ensemble size n, and is in fact the same as the one used in the ensemble transform Kalman filter (ETKF) (Bishop et al. 2001; Wang et al. 2004) in order to obtain the transform matrix. Therefore τ_{max} can be taken as a by-product in the framework of ETKF. On the other hand, if both $\mathbf{H}$ and $\mathbf{R}$ are time-invariant, then the eigenvalues ρ_{max} and ρ_{min} of $\mathbf{R}^{-1/2}\mathbf{H}\mathbf{B}\mathbf{H}^T\mathbf{R}^{-T/2}$ can be calculated off-line once and for all. Taking these considerations into account, (25) can be modified as follows

$$\frac{\xi_l}{1-\xi_l} \left(c_1 \, \tau_{max} + c_2 \, \rho_{max}\right) \leq \gamma \leq \frac{\xi_u}{1-\xi_u} \left(c_2 \, \rho_{min}\right) . \tag{28}$$

Accordingly, (26) is changed to

$$\beta_l \leq \frac{\beta_u}{\tilde{\kappa} + (1-\tilde{\kappa}) \, \xi_u} \, , \tag{29}$$

with $\tilde{\kappa} = (c_1 \, \tau_{max} + c_2 \, \rho_{max})/(c_2 \, \rho_{min})$ being a modified "condition number".

Remark: Inequalities (25) and (26), or alternatively, (28) and (29), are sufficient, but not necessary, conditions. Therefore, even though γ does not lie in the interval in (25) or (28), it may be still possible for the analysis residual norm to satisfy $\beta_l \sqrt{p} \leq \|\hat{\mathbf{r}}^a\|_{\mathbf{R}} \leq \beta_u \sqrt{p}$.

3. Numerical verification

Here we focus on using the 40-dimensional Lorenz 96 (L96) model (Lorenz and Emanuel 1998) to verify the above analytic results, while more intensive filter (with residual nudging) performance investigations are reported in Luo and Hoteit (2012). The experiment settings are the following. A reference trajectory (truth) is generated by numerically integrating the L96 model (with the driving force term $F = 8$) forward through the fourth-order Runge-Kutta method, with the integration step being 0.05 and the total number of integration steps being 1500. The first 500 steps are discarded to avoid the transition effect, and the rest 1000 steps are used for data assimilation. To obtain a long-term "background covariance" $\mathbf{B}^{lt}$ ("background mean" $\mathbf{x}^B$, respectively), we also conduct a separate long model run with $100,000$ integration steps, and take $\mathbf{B}^{lt}$ ($\mathbf{x}^B$) as the temporal covariance (mean) of the generated model trajectory. The synthetic observations are generated by adding the Gaussian white noise $N(0,1)$ to each odd number elements $(x_1, x_3, \cdots, x_{39})$ of the state vector $\mathbf{x} = [x_1, x_2, \cdots, x_{40}]^T$ every 4 integration steps. This corresponds to the 1/2 observation scenario used in Luo and Hoteit (2012). An initial ensemble with 20 ensemble members is generated by drawing samples from the Gaussian distribution $N(\mathbf{x}^B, \mathbf{B}^{lt})$, and the ETKF is adopted for data assimilation.

For distinction later, we call the ETKF without residual nudging the normal ETKF, and the ETKF with residual nudging the ETKF-RN. In the normal ETKF, Eq. (6) is used for mean update with $\hat{\mathbf{C}}^b$ equal to the sample error covariance $\hat{\mathbf{P}}^b$ of the background ensemble[3]. Neither covariance inflation nor covariance localization is introduced to the normal

[3]One may also let $\hat{\mathbf{C}}^b$ be the hybrid of $\hat{\mathbf{P}}^b$ and $\mathbf{B}^{lt}$. In this case, both residual norms and root mean square errors (RMSEs) of the normal ETKF may become smaller (results not shown), while the validity of

ETKF, since for our purpose we wish to use this plain filter setting as the baseline for comparison. One may adopt various inflation and localization techniques to enhance the filter performance, but such an investigation is beyond the scope of this note.

In the ETKF-RN, we adopt the hybrid scheme $\hat{\mathbf{C}}^b = 0.5\hat{\mathbf{P}}^b + 0.5\mathbf{B}^{lt}$ to address the issue of possible singularity in the matrix $\mathbf{A}$ (cf. Eq. 11). Eq. (14) is adopted for mean update, with $\alpha = \delta = 1$, and γ constrained by (28) and (29). For convenience, we denote the lower and upper bounds of γ in (28) by γ_{min} and γ_{max}, respectively, and re-write γ in terms of $\gamma = \gamma_{min} + c\,(\gamma_{max} - \gamma_{min})$ with c being a corresponding scalar coefficient that is involved in our discussion later. Note that in general the background residual norm $\|\hat{\mathbf{r}}^b\|_{\mathbf{R}}$ changes with time, so are the values of ξ_u and ξ_l in Eq. (25). This implies that in general γ_{min} and γ_{max} (hence γ) also change with time, therefore they need to be calculated at each data assimilation cycle.

An additional remark is that the normal ETKF and the ETKF-RN share the same square root update formula as in Wang et al. (2004), where it is the sample error covariance $\hat{\mathbf{P}}^b$, rather than its hybrid with $\mathbf{B}^{lt}$, which is used to generate the background square root. Such a choice is based on the following considerations. On the one hand, if one uses the hybrid covariance for square root update, then it would require a matrix factorization (e.g., singular value decomposition) in order to compute a square root of the hybrid covariance at each data assimilation cycle, which can be very expensive in large-scale applications. On the other hand, for the L96 model used here, numerical investigations show that using the hybrid covariance for square root update does not necessarily improve the filter performance (results not shown).

---

the analytic results in the previous section is not affected.

The procedures in the ETKF-RN are summarized as follows. Because the matrix $\mathbf{R}^{-1/2}\mathbf{H}\mathbf{B}\mathbf{H}^T\mathbf{R}^{-T/2}$ is time invariant, its maximum and minimum eigenvalues, ρ_{max} and ρ_{min} (cf. (28)), respectively, are calculated and saved for later use. Then, with the background ensemble at each data assimilation cycle, calculate the sample mean $\hat{\mathbf{x}}^b$, the corresponding background residual norm $\|\hat{\mathbf{r}}^b\|_{\mathbf{R}}$, and a square root $\hat{\mathbf{S}}^b$ of the sample error covariance $\hat{\mathbf{P}}^b$ following Bishop et al. (2001); Luo and Moroz (2009); Wang et al. (2004). Update $\hat{\mathbf{S}}^b$ to its analysis counterpart $\hat{\mathbf{S}}^a \equiv \hat{\mathbf{S}}^b\mathbf{T}\mathbf{U}$ by calculating a transform matrix $\mathbf{T}$, together with a "centering" matrix $\mathbf{U}$ following Wang et al. (2004). During the square root update process, the maximum eigenvalue τ_{max} of $\mathbf{R}^{-1/2}\mathbf{H}\hat{\mathbf{P}}^b\mathbf{H}^T\mathbf{R}^{-T/2}$ is obtained as a by-product following our discussion in the previous section. With these information, one is ready to calculate the interval bounds γ_{min} and γ_{max} in (28), hence obtain $\gamma = \gamma_{min} + c\,(\gamma_{max} - \gamma_{min})$ for a given value of c (c can be constant or variable during the whole data assimilation time window). This γ value is then inserted into Eq. (14) (with $\alpha = \delta = 1$ there) to obtain the analysis mean $\hat{\mathbf{x}}^a$. With $\hat{\mathbf{x}}^a$ and $\hat{\mathbf{S}}^a$, an analysis ensemble can be generated in the same way as in Bishop et al. (2001); Wang et al. (2004). Propagating this ensemble forward in time, one starts a new data assimilation cycle, and so on. Comparing the above procedures to those in Luo and Hoteit (2012), the observation inversion used in Luo and Hoteit (2012) is avoided.

The experiment below aims to show that, at each data assimilation cycle, if a γ value lies in the interval $\mathbb{C}_\gamma = [\gamma_{min}, \gamma_{max}]$ given by (28), then the corresponding analysis residual norm $\|\hat{\mathbf{r}}^a\|_{\mathbf{R}}$ is bounded by the interval $\mathbb{C}_{rn} = [\beta_l\sqrt{p}, \beta_u\sqrt{p}]$, with β_l and β_u satisfying the constraint (29). In the experiment we fix $\beta_u = 2$, and let $\beta_l = 0.1 \times (\beta_u/(\tilde{\kappa} + (1-\tilde{\kappa})\,\xi_u))$, where the small fraction 0.1 is introduced for convenience of visualization[4].

[4]In some cases $\beta_u/(\tilde{\kappa} + (1-\tilde{\kappa})\,\xi_u)$ in (29) may be very close to β_u. Therefore if β_l is close to this value,

Fig. 1 shows the time series of the background (dash-dotted) and analysis (thick solid) residual norms in different filter settings (for convenience of visualization, the residual norm values are plotted in the logarithmic scale). For reference we also plot the targeted lower and upper bounds (dash and thin solid lines, respectively), $\beta_l \sqrt{p}$ and $\beta_u \sqrt{p}$ ($p = 20$), respectively. In the normal ETKF (Fig. 1(a)), in most of the time the analysis residual norms are larger than the targeted upper bound (no targeted lower bound is calculated and plotted in this case). With residual nudging, the analysis residual norms of the ETKF-RN migrate into the targeted interval, as long as the coefficient c lies in $[0, 1]$ (Figs. 1(b) – 1(d). Also see the caption of Fig. 1 to find out how the corresponding c values are chosen). When c is outside the interval $[0, 1]$, the corresponding γ is not bounded by $[\gamma_{min}, \gamma_{max}]$, hence there is no guarantee that the corresponding analysis residual norms are bounded by $[\beta_l \sqrt{p}, \beta_u \sqrt{p}]$. Two such examples are presented in Fig. 1(e) and 1(f), with c being 2.5 and -0.005, respectively (e.g., for $c = -0.005$ in Fig. 1(f), breakthroughs of the lower bound are found around time step 220 and a few other places). As side results, we also report in Table 1 the time mean root mean square errors (RMSEs) (see Eq. (13) of Luo and Hoteit 2012) that correspond to different filter settings in Fig. 1. In these tested cases, the filter performance of the ETKF-RN appears improved, in terms of the time mean RMSE, when compared to that of the normal ETKF.

the difference $(\beta_u - \beta_l)$, hence the interval $\mathbb{C}_{rn}$, may be very small.

4. Discussion and conclusion

We derived some sufficient inflation constraints in order for the analysis residual norm to be bounded in a certain interval. The analytic results showed that these constraints are related to the maximum and minimum eigenvalues of certain matrices (cf. Eq. (11)). In certain circumstances, the constraint with respect to the minimum eigenvalue (e.g., Eq. (13)) may impose a non-singularity requirement on relevant matrices. A few strategies in the literature that can be adopted to address or mitigate this issue are highlighted.

Some remaining issues are manifest in our deduction. These include, for instance, the nonlinearity in the observation operator and the choice of β_u and β_l. For the former problem, under a suitable smoothness assumption on the observation operator, one may also obtain inflation constraints similar to those in Section 2. On the other hand, though, more investigations may be needed to make the results more practical in terms of computational complexity. For the latter problem, numerical results in Luo and Hoteit (2012) show that the β values influence the overall performance of the EnKF in terms of filter stability and accuracy. Intuitively, smaller (larger) β values tend to make residual nudging happen more (less) often. Therefore, if the normal EnKF performs well (poorly), then a larger (smaller) β value may be suitable. In this aspect, it is expected that an objective criterion is needed. This will be investigated in the future.

Acknowledgement

We would like to thank two anonymous reviewers for their constructive comments and suggestions. The first author would also like to thank the IRIS/CIPR cooperative research project "Integrated Workflow and Realistic Geology" which is funded by industry partners ConocoPhillips, Eni, Petrobras, Statoil, and Total, as well as the Research Council of Norway (PETROMAKS) for financial support.

REFERENCES

Altaf, U. M., T. Butler, X. Luo, C. Dawson, T. Mayo, and H. Hoteit, 2013: Improving short range ensemble Kalman storm surge forecasting using robust adaptive inflation. *Mon. Wea. Rev.*, accepted.

Anderson, J. L., 2001: An ensemble adjustment Kalman filter for data assimilation. *Mon. Wea. Rev.*, **129**, 2884–2903.

Anderson, J. L., 2007: An adaptive covariance inflation error correction algorithm for ensemble filters. *Tellus*, **59A (2)**, 210–224.

Anderson, J. L., 2009: Spatially and temporally varying adaptive covariance inflation for ensemble filters. *Tellus*, **61A**, 72–83.

Anderson, J. L. and S. L. Anderson, 1999: A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. *Mon. Wea. Rev.*, **127**, 2741–2758.

Bishop, C. H., B. J. Etherton, and S. J. Majumdar, 2001: Adaptive sampling with ensemble transform Kalman filter. Part I: theoretical aspects. *Mon. Wea. Rev.*, **129**, 420–436.

Bocquet, M., 2011: Ensemble Kalman filtering without the intrinsic need for inflation. *Nonlinear Processes in Geophysics*, **18 (5)**, 735–750.

Bocquet, M. and P. Sakov, 2012: Combining inflation-free and iterative ensemble Kalman filters for strongly nonlinear systems. *Nonlinear Processes in Geophysics*, **19 (3)**, 383–399.

Grcar, J. F., 2010: A matrix lower bound. *Linear Algebra and its Applications*, **433**, 203–220.

Hamill, T. M. and C. Snyder, 2000: A hybrid ensemble Kalman filter-3d variational analysis scheme. *Mon. Wea. Rev.*, **128**, 2905–2919.

Hamill, T. M. and J. S. Whitaker, 2011: What constrains spread growth in forecasts initialized from ensemble Kalman filters? *Mon. Wea. Rev.*, **139**, 117–131.

Hamill, T. M., J. S. Whitaker, J. L. Anderson, and C. Snyder, 2009: Comments on "Sigma-point Kalman filter data assimilation methods for strongly nonlinear systems". *J. Atmos. Sci.*, **66**, 3498–3500.

Hamill, T. M., J. S. Whitaker, and C. Snyder, 2001: Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter. *Mon. Wea. Rev.*, **129**, 2776–2790.

Horn, R. and C. Johnson, 1990: *Matrix analysis*. Cambridge University Press.

Horn, R. and C. Johnson, 1991: *Topics in matrix analysis*. Cambridge University Press.

Hoteit, I., D. T. Pham, and J. Blum, 2002: A simplified reduced order Kalman filtering and application to altimetric data assimilation in Tropical Pacific. *Journal of Marine Systems*, **36**, 101–127.

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

Luo, X. and I. Hoteit, 2011: Robust ensemble filtering and its relation to covariance inflation in the ensemble Kalman filter. *Mon. Wea. Rev.*, **139**, 3938–3953.

Luo, X. and I. Hoteit, 2012: Ensemble Kalman filtering with residual nudging. *Tellus A*, **64**, 17 130.

Luo, X. and I. Hoteit, 2013: Efficient particle filtering through residual nudging. *Quart. J. Roy. Meteor. Soc.*, in press.

Luo, X. and I. M. Moroz, 2009: Ensemble Kalman filter with the unscented transform. *Physica D*, **238**, 549–562.

Meng, Z. and F. Zhang, 2007: Tests of an ensemble Kalman filter for mesoscale and regional-scale data assimilation. part II: Imperfect model experiments. *Mon. Wea. Rev*, **135 (4)**, 1403–1423.

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

Ott, E., et al., 2004: A local ensemble Kalman filter for atmospheric data assimilation. *Tellus*, **56A**, 415–428.

Song, H., I. Hoteit, B. Cornuelle, and A. Subramanian, 2010: An adaptive approach to mitigate background covariance limitations in the ensemble Kalman filter. *Mon. Wea. Rev.*, **138 (7)**, 2825–2845.

Triantafyllou, G., I. Hoteit, X. Luo, K. Tsiaras, and G. Petihakis, 2013: Assessing a robust ensemble-based Kalman filter for efficient ecosystem data assimilation of the Cretan sea. *Journal of Marine Systems*, in press.

Wang, X., C. H. Bishop, and S. J. Julier, 2004: Which is better, an ensemble of positive-negative pairs or a centered simplex ensemble. *Mon. Wea. Rev.*, **132**, 1590–1605.

Whitaker, J. S. and T. M. Hamill, 2002: Ensemble data assimilation without perturbed observations. *Mon. Wea. Rev.*, **130**, 1913–1924.

Whitaker, J. S. and T. M. Hamill, 2012: Evaluating methods to account for system errors in ensemble data assimilation. *Monthly Weather Review*, **140**, 3078–3089.

Zhang, F., C. Snyder, and J. Sun, 2004: Impacts of initial estimate and observation availability on convective-scale data assimilation with an ensemble Kalman filter. *Mon. Wea. Rev*, **132 (5)**, 1238–1253.

List of Tables

Table 1. Time mean RMSEs in the normal ETKF and the ETKF-RN with the same c values as in Fig. 1.

	Normal ETKF	ETKF-RN with				
		$c = 0$	$c = 1$	$c \in [0, 1]$	$c = 2.5$	$c = -0.005$
Background RMSE	4.3148	1.8252	2.4095	2.2182	2.6857	2.0394
Analysis RMSE	4.2645	1.6953	2.2764	2.0894	2.5679	1.9054

List of Figures

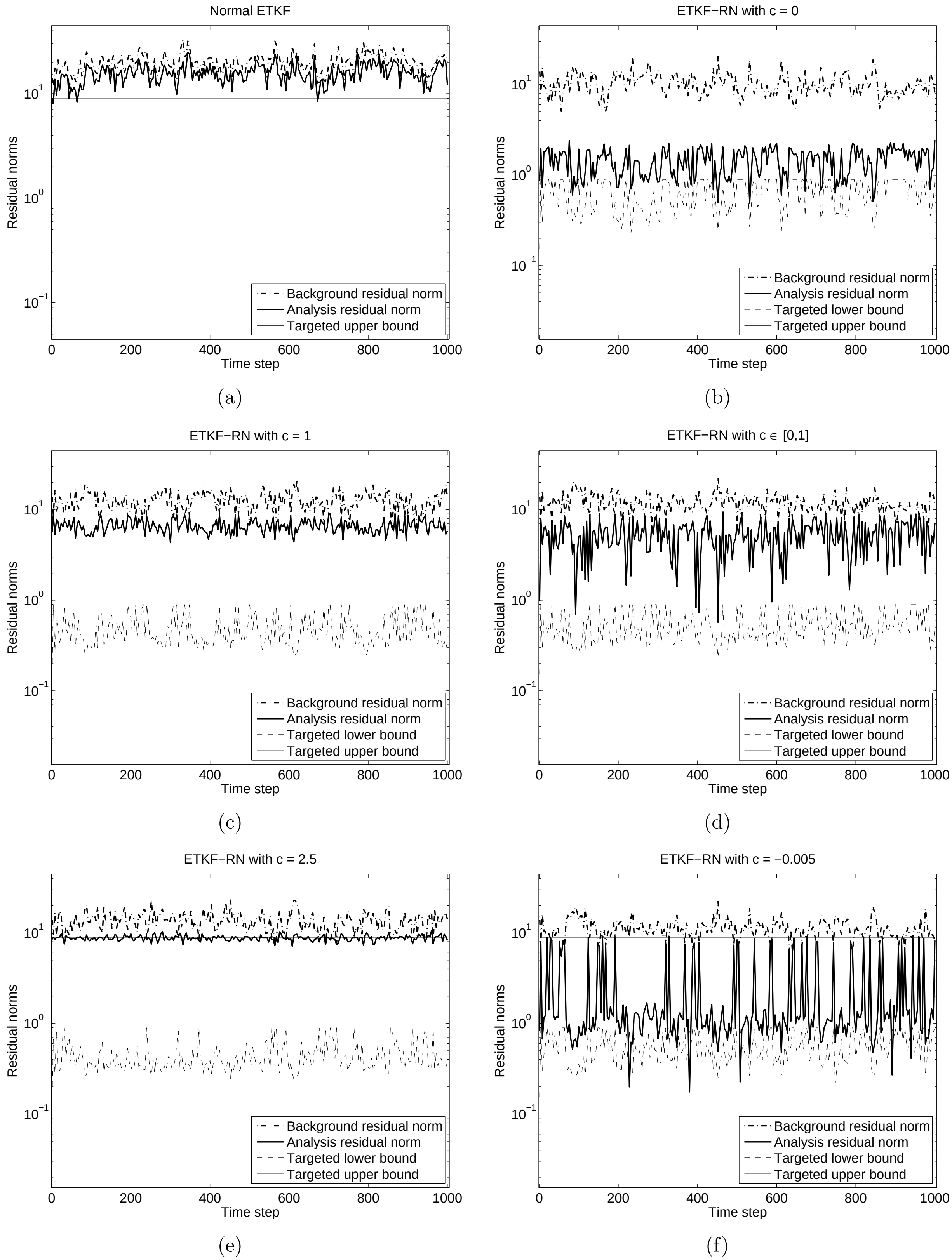

FIG. 1. Time series of the analysis residual norms in: (a): the normal ETKF without residual nudging; (b) – (f) the ETKF-RN with different c values. For the normal ETKF there are no targeted lower and upper residual norm bounds. For reference, though, we still plot the targeted upper bound ($= 2\sqrt{20}$) in (a). We also note that the c value in Fig. 1(d) is randomly drawn from the uniform distribution on the interval $[0, 1]$ at each data assimilation cycle, while in the rest of the sub-figures the c values are constant during the assimilation time window.