Information-based data selection for ensemble data assimilation

RMetS
Royal Meteorological Society

Information-based data selection for ensemble data assimilation

S. Migliorini*
Department of Meteorology, University of Reading, UK

*Correspondence to: S. Migliorini, Department of Meteorology, University of Reading, Earley Gate, PO Box 243, Reading RG6 6BB, UK. E-mail: s.migliorini@reading.ac.uk

Ensemble-based data assimilation is rapidly proving itself as a computationally efficient and skilful assimilation method for numerical weather prediction, which can provide a viable alternative to more established variational assimilation techniques. However, a fundamental shortcoming of ensemble techniques is that the resulting analysis increments can only span a limited subspace of the state space, whose dimension is less than the ensemble size. This limits the amount of observational information that can effectively constrain the analysis. In this paper, a data selection strategy that aims to assimilate only the observational components that matter most and that can be used with both stochastic and deterministic ensemble filters is presented. This avoids unnecessary computations, reduces round-off errors and minimizes the risk of importing observation bias in the analysis. When an ensemble-based assimilation technique is used to assimilate high-density observations, the data selection procedure allows the use of larger localization domains that may lead to a more balanced analysis. Results from the use of this data selection technique with a two-dimensional linear and a nonlinear advection model using both *in situ* and remote sounding observations are discussed.

Key Words: data assimilation; ensemble filtering; information content

Received 11 July 2012; Revised 5 December 2012; Accepted 13 December 2012; Published online in Wiley Online Library

Citation: S. Migliorini. 2013. Information-based data selection for ensemble data assimilation. *Q. J. R. Meteorol. Soc.* DOI:10.1002/qj.2104

1. Introduction

A necessary condition for an ideal assimilation is that the true state should be statistically indistinguishable from any analysis ensemble members that are randomly sampled from the posterior (or analysis) probability density function (pdf) resulting from data assimilation (see, for example, Anderson and Anderson 1999, their section 4). A key shortcoming of ensemble filtering, due to its representation of the posterior pdf with a limited number of analysis ensemble members, is that the mean analysis error covariance $\tilde{\mathbf{P}}^a$ estimated by means of a limited-size ensemble is negatively biased (Sacher and Bartello, 2008). In the scalar case this is equivalent to stating that the mean analysis error variance estimated by means of a limited-size ensemble underestimates the optimal analysis error variance estimated using an infinite number of ensemble members. Another implication of the use of a limited ensemble size is that the sample covariance of forecast error $\tilde{\mathbf{P}}^f$ is rank deficient when $K < n + 1$ (see, for example, section 3), where K is the number of ensemble members and n is the dimension of the state space. This implies that the analysis increments can only belong to the range of $\tilde{\mathbf{P}}^f$, i.e. the subspace of the state space defined by the columns of $\tilde{\mathbf{P}}^f$ (e.g. Bannister, 2008, his section 3.3), with potentially adverse effects on the reliability of the analysis ensemble. It follows that ensemble filtering can lead to filter divergence (e.g. Houtekamer and Mitchell, 1998; van Leeuwen, 1999), where the magnitude of the true analysis error becomes much larger than its estimate, as a result of the fact that observations are progressively ignored by the filter. Sampling error may also lead to a misrepresentation of forecast error covariance values between two different locations, and this can be particularly detrimental when long-range spatial correlations are overestimated, leading to spurious analysis increments (e.g. Hamill *et al.*, 2001).

To minimize these shortcomings, ensemble filtering usually makes use of procedures such as covariance inflation

and covariance localization. The purpose of inflation methods (Anderson and Anderson, 1999) is to enlarge the spread of the forecast ensemble either by multiplying the ensemble member perturbations from the mean by a factor greater than one or by adding random perturbations to the ensemble members (e.g. Whitaker *et al.*, 2008; Houtekamer *et al.*, 2009). More recently, adaptive multiplicative inflation schemes have also been introduced (Anderson, 2007a, 2009; Li *et al.*, 2009; Miyoshi, 2011). Localization techniques aim to eliminate long-range correlations either via an element-wise multiplication (or Schur product) of $\tilde{\mathbf{P}}^f$ with a correlation matrix with compactly supported correlation functions (see Houtekamer and Mitchell, 1998, 2001; Janjić *et al.*, 2011) or by estimating the analysis on a local domain (e.g. Ott *et al.*, 2004; Hunt *et al.*, 2007) using only observations within a given radius of influence from each grid point. Adaptive localization techniques have also been proposed (Anderson, 2007b; Bishop and Hodyss, 2009a, 2009b).

As noted by Lorenc (2003, his section 3b), another consequence of using a rank-deficient estimate of the forecast error covariance matrix is that at most $K-1$ degrees of freedom are available to ensemble-based data assimilation schemes in order to fit the observations. This means that observations that are sensitive to components of the state vector that do not belong to the range of $\tilde{\mathbf{P}}^f$ do not improve the analysis estimate. Both distance-dependent or Schur product localization procedures ease the rank deficiency problem as the localized $\tilde{\mathbf{P}}^f$ is only supposed to represent the covariance of the local forecast error. However, the radius of influence should be large enough not to disturb the balances that act at given spatial scales and that are well represented by the ensemble error covariance (e.g. Lorenc, 2003, his section 3c). The radius of influence should also be large enough to include sufficient observations to constrain the analysis effectively. At the same time, a radius of influence that is too large may not substantially reduce the number of assimilated observations, particularly over data-dense areas.

In this paper, a data selection strategy based on the information content of the measurements is proposed, which ensures that only the observational components that are able to constrain the analysis are assimilated using ensemble filtering techniques. The paper is organized as follows. In section 2 the relationship between the measurements and the true state of the system is described. Section 3 provides a detailed derivation of a square-root filter and of its expression when the true forecast uncertainty is approximated by using a given number of forecast ensemble members. Section 4 discusses measures of observational information content that can be used in the context of ensemble-based data assimilation, while two versions of the proposed data selection procedure are detailed in section 5. Section 6 provides details of the numerical model used in this work for the assimilation of both *in situ* and remote sounding observations. Results of a number of ensemble data assimilation experiments to test the proposed data selection methodology are provided in section 7. Finally, some conclusions are drawn in section 8.

2. Characterization of measurements for assimilation

The relationship between a measurement vector $\mathbf{y}^o \in \mathbb{R}^q$ and the true state $\mathbf{x}^t \in \mathbb{R}^n$ of a system (e.g. the atmosphere) can be expressed as

$$\mathbf{y}^o = H(\mathbf{x}^t) + \boldsymbol{\epsilon}^o, \tag{1}$$

where $H(\mathbf{x}^t)$ is the observation operator calculated in $\mathbf{x}^t$ and where $\boldsymbol{\epsilon}^o \in \mathbb{R}^q$ is the measurement error, assumed Gaussian, unbiased and with covariance $\mathbf{R} \in \mathbb{R}^{q\times q}$. When it is a linear function of the state, the observation operator is represented as the matrix $\mathbf{H} \in \mathbb{R}^{q\times n}$ and Eq. (1) becomes

$$\mathbf{y}^o = \mathbf{H}\mathbf{x}^t + \boldsymbol{\epsilon}^o. \tag{2}$$

When instead it is a nonlinear function of the state, the observation operator can be linearized about a given $\mathbf{x}_i$. In this case, from Eq. (1) we can write

$$\mathbf{y}^o \simeq H(\mathbf{x}_i) + \mathbf{H}^{(i)}(\mathbf{x}^t - \mathbf{x}_i) + \boldsymbol{\epsilon}^o, \tag{3}$$

where $\mathbf{H}^{(i)} \equiv (\partial H/\partial \mathbf{x})_{\mathbf{x}=\mathbf{x}_i} \in \mathbb{R}^{q\times n}$ is the Jacobian matrix of $H(\mathbf{x})$ calculated in $\mathbf{x} = \mathbf{x}_i$. We can also define $\mathbf{y}^{(i)}$ as (e.g. Migliorini, 2012)

$$\mathbf{y}^{o(i)} \equiv \mathbf{y}^o - H(\mathbf{x}_i) + \mathbf{H}^{(i)}\mathbf{x}_i \simeq \mathbf{H}^{(i)}\mathbf{x}^t + \boldsymbol{\epsilon}^o. \tag{4}$$

A succession of linearized measurements $\mathbf{y}^{o(i)}$ can be used within the analysis update step of a locally iterated extended Kalman filter (e.g. Cohn, 1997, his section 5.2) as well as within an iterated ensemble Kalman smoother. To simplify the notation, hereafter we will not provide explicit indication of the iteration index when considering either a linear measurement or a measurement linearized about a given state.

The matrix $\mathbf{R}$ can be expressed in terms of its eigenvector decomposition as $\mathbf{R} = \mathbf{L}\boldsymbol{\Lambda}\mathbf{L}^T$, where $\mathbf{L}$ is a matrix whose columns are its eigenvectors and $\boldsymbol{\Lambda}$ a diagonal matrix whose diagonal elements are the corresponding eigenvalues. When the number m of non-zero (or not too small) eigenvalues is less than q, it is possible to define a truncated eigenvector decomposition $\mathbf{R} \simeq \mathbf{L}_m\boldsymbol{\Lambda}_m\mathbf{L}_m^T$ where the columns of $\mathbf{L}_m$ are the eigenvectors corresponding to the m largest eigenvalues of $\mathbf{R}$ and where the elements of the diagonal of $\boldsymbol{\Lambda}_m$ are the m largest eigenvalues of $\mathbf{R}$. For $m \leq q$ it is now possible to define $\mathbf{y}^{o\prime} \in \mathbb{R}^m$ as $\mathbf{y}^{o\prime} \equiv \boldsymbol{\Lambda}_m^{-1/2}\mathbf{L}_m^T\mathbf{y}^o$ so that from Eq. (2) or (4) we can write

$$\mathbf{y}^{o\prime} = \boldsymbol{\Lambda}_m^{-1/2}\mathbf{L}_m^T\mathbf{H}\mathbf{x}^t + \boldsymbol{\Lambda}_m^{-1/2}\mathbf{L}_m^T\boldsymbol{\epsilon}^o = \mathbf{H}'\mathbf{x}^t + \boldsymbol{\epsilon}^{o\prime}, \tag{5}$$

where $\mathbf{H}' \in \mathbb{R}^{m\times n}$ is defined as $\mathbf{H}' \equiv \boldsymbol{\Lambda}_m^{-1/2}\mathbf{L}_m^T\mathbf{H}$ and where the covariance of $\boldsymbol{\epsilon}^{o\prime} \equiv \boldsymbol{\Lambda}_m^{-1/2}\mathbf{L}_m^T\boldsymbol{\epsilon}^o$ is the unit matrix of rank m, i.e. $\boldsymbol{\epsilon}^{o\prime} \in \mathbb{R}^m$ is the result of the application of a whitening filter to $\boldsymbol{\epsilon}^o$. Finally, an alternative definition of $\mathbf{y}^{o\prime}$ that preserves the nonlinear relationship with $\mathbf{x}^t$ (when applicable) is given by

$$\mathbf{y}^{o\prime} = \boldsymbol{\Lambda}_m^{-1/2}\mathbf{L}_m^T H(\mathbf{x}^t) + \boldsymbol{\Lambda}_m^{-1/2}\mathbf{L}_m^T\boldsymbol{\epsilon}^o = H'(\mathbf{x}^t) + \boldsymbol{\epsilon}^{o\prime}, \tag{6}$$

where $H'(\mathbf{x}^t) \equiv \boldsymbol{\Lambda}_m^{-1/2}\mathbf{L}_m^T H(\mathbf{x}^t) \in \mathbb{R}^m$.

3. Square root filters and their approximations

In this paper we will initially concentrate on the square root formulation of the Kalman filter algorithm, as it guarantees that error covariances remain positive definite (Maybeck, 1982, chapter 7), and on the ensemble square root filter,

which avoids the need for perturbing the observations (e.g. Whitaker and Hamill, 2002). In this way, it is possible to provide the most accurate description of the observational information content, as reviewed below and in the following section.

The analysis error covariance $\mathbf{P}^a$ is related to the forecast error covariance $\mathbf{P}^f$ according to the Kalman filter solution of the cycling problem for a linear stochastic-dynamic system and given by (e.g. Cohn, 1997, his section 4)

$$\mathbf{P}^a = \mathbf{P}^f - \mathbf{P}^f\mathbf{H}'^T(\mathbf{H}'\mathbf{P}^f\mathbf{H}'^T + \mathbf{I}_m)^{-1}\mathbf{H}'\mathbf{P}^f. \tag{7}$$

If we now express $\mathbf{P}^f$ as $\mathbf{P}^f \equiv \mathbf{X}^f\mathbf{X}^{fT}$ and $\mathbf{S} \equiv \mathbf{H}'\mathbf{X}^f \in \mathbb{R}^{m\times n}$, it follows that Eq. (7) can be written as

$$\mathbf{P}^a = \mathbf{X}^f(\mathbf{I}_n - \mathbf{S}^T(\mathbf{S}\mathbf{S}^T + \mathbf{I}_m)^{-1}\mathbf{S})\mathbf{X}^{fT}. \tag{8}$$

We can express $\mathbf{S}$ in terms of its singular value decomposition as $\mathbf{S} = \mathbf{E}\boldsymbol{\Gamma}\mathbf{V}^T$, where $\mathbf{E} \in \mathbb{R}^{m\times m}$ and $\mathbf{V} \in \mathbb{R}^{n\times n}$ are orthogonal matrices whose columns are the left and right singular vectors of $\mathbf{S}$, respectively, and where $\boldsymbol{\Gamma} \in \mathbb{R}^{m\times n}$. Note that the only non-zero elements of $\boldsymbol{\Gamma}$ have the same row and column indexes and are equal to the $p \leq \min(m, n)$ positive non-dimensional singular values γ_i of $\mathbf{S}$. Let us now express $\boldsymbol{\Gamma}$ as

$$\boldsymbol{\Gamma} = \begin{pmatrix} \boldsymbol{\Gamma}_p & \mathbf{0}_{p\times(n-p)} \\ \mathbf{0}_{(m-p)\times p} & \mathbf{0}_{(m-p)\times(n-p)} \end{pmatrix} \tag{9}$$

and

$$\mathbf{Y}_{m,p} \equiv \boldsymbol{\Gamma}\boldsymbol{\Gamma}^T = \begin{pmatrix} \boldsymbol{\Gamma}_p^2 & \mathbf{0}_{p\times(m-p)} \\ \mathbf{0}_{(m-p)\times p} & \mathbf{0}_{m-p} \end{pmatrix}, \tag{10}$$

where $\boldsymbol{\Gamma}_p$ and $\mathbf{Y}_{m,p}$ are $p \times p$ and $m \times m$ diagonal matrices, respectively, with p non-zero eigenvalues. In this way, by substituting the quantities defined above, it is now possible to write

$$\begin{aligned}\mathbf{S}^T(\mathbf{S}\mathbf{S}^T + \mathbf{I}_m)^{-1}\mathbf{S} &= \mathbf{V}\boldsymbol{\Gamma}^T(\mathbf{Y}_{m,p} + \mathbf{I}_m)^{-1}\boldsymbol{\Gamma}\mathbf{V}^T \\ &= \mathbf{V}\begin{pmatrix} \boldsymbol{\Gamma}_p(\boldsymbol{\Gamma}_p^2 + \mathbf{I}_p)^{-1}\boldsymbol{\Gamma}_p & \mathbf{0}_{p\times(n-p)} \\ \mathbf{0}_{(n-p)\times p} & \mathbf{0}_{n-p} \end{pmatrix}\mathbf{V}^T \\ &= \mathbf{V}\begin{pmatrix} \mathbf{I}_p - (\boldsymbol{\Gamma}_p^2 + \mathbf{I}_p)^{-1} & \mathbf{0}_{p\times(n-p)} \\ \mathbf{0}_{(n-p)\times p} & \mathbf{0}_{n-p} \end{pmatrix}\mathbf{V}^T, \end{aligned} \tag{11}$$

given the commutative property of diagonal matrices of the same dimensions. From Eqs (10) and (11) we can then write

$$\begin{aligned}\mathbf{I}_n - \mathbf{S}^T(\mathbf{S}\mathbf{S}^T + \mathbf{I}_m)^{-1}\mathbf{S} &= \mathbf{V}\begin{pmatrix} (\boldsymbol{\Gamma}_p^2 + \mathbf{I}_p)^{-1} & \mathbf{0}_{p\times(n-p)} \\ \mathbf{0}_{(n-p)\times p} & \mathbf{I}_{n-p} \end{pmatrix}\mathbf{V}^T \\ &= \mathbf{V}(\mathbf{Y}_{n,p} + \mathbf{I}_n)^{-1}\mathbf{V}^T. \end{aligned} \tag{12}$$

It follows that Eq. (8) can then be written as

$$\mathbf{P}^a = \mathbf{X}^f\mathbf{T}\mathbf{T}^T\mathbf{X}^{fT}, \tag{13}$$

where

$$\mathbf{T} = \mathbf{V}(\mathbf{Y}_{n,p} + \mathbf{I}_n)^{-1/2}\mathbf{C} \in \mathbb{R}^{n\times n}, \tag{14}$$

with $\mathbf{C}\mathbf{C}^T = \mathbf{I}_n$. By defining $\mathbf{P}^a = \mathbf{X}^a\mathbf{X}^{aT}$, from Eq. (13) we can write

$$\mathbf{X}^a = \mathbf{X}^f\mathbf{T} \in \mathbb{R}^{n\times n}. \tag{15}$$

Note that when $\mathbf{C} = \mathbf{V}^T$ the transform matrix $\mathbf{T}$ is symmetric.

The ensemble transform Kalman filter (ETKF, Bishop *et al.*, 2001), which is the ensemble square root filter we will concentrate on, provides an approximation of $\mathbf{X}^a$ by means of the analysis perturbations matrix $\mathbf{X}'^a$, calculated as

$$\mathbf{X}'^a = \mathbf{X}'^f\tilde{\mathbf{T}} \in \mathbb{R}^{n\times K}, \tag{16}$$

where

$$\begin{aligned}\mathbf{X}'^f = \frac{1}{\sqrt{K-1}}(&\mathbf{x}_1^f - \overline{\mathbf{x}^f}, \mathbf{x}_2^f - \overline{\mathbf{x}^f}, \ldots, \\ &\mathbf{x}_i^f - \overline{\mathbf{x}^f}, \ldots, \mathbf{x}_K^f - \overline{\mathbf{x}^f}) \in \mathbb{R}^{n\times K},\end{aligned} \tag{17}$$

with K being the number of ensemble forecast members $\mathbf{x}_i^f$ with mean $\overline{\mathbf{x}^f}$, and where $\tilde{\mathbf{T}} \in \mathbb{R}^{K\times K}$ is a suitable approximation of $\mathbf{T}$. If we now define $\tilde{\mathbf{P}}^f \equiv \mathbf{X}'^f\mathbf{X}'^{fT}$ and $\tilde{\mathbf{P}}^a \equiv \mathbf{X}'^a\mathbf{X}'^{aT}$, from Eq. (16) we can write

$$\tilde{\mathbf{P}}^a = \mathbf{X}'^f\tilde{\mathbf{T}}\tilde{\mathbf{T}}^T\mathbf{X}'^{fT}. \tag{18}$$

Note that $\mathrm{rank}(\tilde{\mathbf{P}}^f) = \mathrm{rank}(\mathbf{X}'^f) \leq \min(K-1, n)$. The reason for having $K-1$ rather than K in the expression constraining the rank of $\tilde{\mathbf{P}}^f$ is that the sum of the columns of $\mathbf{X}'^f$ is, by definition, equal to zero. We now want to determine an expression for $\tilde{\mathbf{T}}$ in a way that is fully consistent with the derivation in Bishop *et al.* (2001) and instrumental to the implementation of the data selection strategy discussed in section 5. To this end, $\mathbf{S}$ is approximated by $\tilde{\mathbf{S}} \equiv \mathbf{H}'\mathbf{X}'^f \in \mathbb{R}^{m\times K}$ so that Eq. (8) can be written as

$$\tilde{\mathbf{P}}^a = \mathbf{X}'^f(\mathbf{I}_K - \tilde{\mathbf{S}}^T(\tilde{\mathbf{S}}\tilde{\mathbf{S}}^T + \mathbf{I}_m)^{-1}\tilde{\mathbf{S}})\mathbf{X}'^{fT}. \tag{19}$$

It is worth noting that it is possible to avoid linearizing the observation operator as in Eq. (4) if we define $\mathbf{y}^{o\prime}$ and $H'(\mathbf{x}^t)$ as in Eq. (6). In this case $\tilde{\mathbf{S}}$ can be defined as

$$\begin{aligned}\tilde{\mathbf{S}} = \frac{1}{\sqrt{K-1}}(&H'(\mathbf{x}_1^f) - \overline{H'(\mathbf{x}^f)}, \ldots, \\ &H'(\mathbf{x}_j^f) - \overline{H'(\mathbf{x}^f)}, \ldots, H'(\mathbf{x}_K^f) - \overline{H'(\mathbf{x}^f)}),\end{aligned} \tag{20}$$

where

$$\overline{H'(\mathbf{x}^f)} \equiv \frac{1}{K}\sum_{j=1}^{K} H'(\mathbf{x}_j^f). \tag{21}$$

Similarly to $\mathbf{S}$, it is possible to express $\tilde{\mathbf{S}}$ in terms of its singular value decomposition as $\tilde{\mathbf{S}} = \tilde{\mathbf{E}}\tilde{\boldsymbol{\Gamma}}\tilde{\mathbf{V}}^T$, where $\tilde{\mathbf{E}} \in \mathbb{R}^{m\times m}$ and $\tilde{\mathbf{V}} \in \mathbb{R}^{K\times K}$ are orthogonal matrices whose columns are the left and right singular vectors of $\tilde{\mathbf{S}}$, respectively, and where the only non-zero elements of

$\tilde{\mathbf{\Gamma}} \in \mathbb{R}^{m\times K}$ have the same row and column indexes and are equal to the $\tilde{p}$ positive non-dimensional singular values $\tilde{\gamma}_i$ of $\tilde{\mathbf{S}}$. In this way, Eq. (19) can be expressed as

$$\tilde{\mathbf{P}}^a = \mathbf{X}'^f \tilde{\mathbf{V}}(\tilde{\mathbf{Y}}_{K,\tilde{p}} + \mathbf{I}_K)^{-1}\tilde{\mathbf{V}}^T \mathbf{X}'^{fT}, \tag{22}$$

where

$$\tilde{\mathbf{\Gamma}} = \begin{pmatrix} \tilde{\mathbf{\Gamma}}_{\tilde{p}} & \mathbf{0}_{\tilde{p}\times(K-\tilde{p})} \\ \mathbf{0}_{(m-\tilde{p})\times\tilde{p}} & \mathbf{0}_{(m-\tilde{p})\times(K-\tilde{p})} \end{pmatrix} \tag{23}$$

and

$$\tilde{\mathbf{Y}}_{K,\tilde{p}} \equiv \begin{pmatrix} \tilde{\mathbf{\Gamma}}^2_{\tilde{p}} & \mathbf{0}_{\tilde{p}\times(K-\tilde{p})} \\ \mathbf{0}_{(K-\tilde{p})\times\tilde{p}} & \mathbf{0}_{K-\tilde{p}} \end{pmatrix}, \tag{24}$$

with $\tilde{p} = \mathrm{rank}(\tilde{\mathbf{S}}) \leq \min(m, K-1)$. From Eqs (18) and (22) it follows that $\tilde{\mathbf{T}}$ can be written as

$$\tilde{\mathbf{T}} = \tilde{\mathbf{V}}(\tilde{\mathbf{Y}}_{K,\tilde{p}} + \mathbf{I}_K)^{-1/2}\tilde{\mathbf{V}}^T \in \mathbb{R}^{K\times K}, \tag{25}$$

where we have chosen a symmetric form of the ensemble transform matrix $\tilde{\mathbf{T}}$ so as to ensure that $\mathbf{X}'^a$ is unbiased (e.g. Hunt *et al.*, 2007; Livings *et al.*, 2008; Sakov and Oke, 2008). Note that the expression of $\tilde{\mathbf{T}}$ in Eq. (25) is equivalent to that given in Bishop *et al.*(2001, their equation 18b).

4. Information considerations

In the previous section, the expression for the analysis error covariance as given by the Kalman filter when the observation error covariance is given by the unit matrix was compared to the estimate of the analysis error covariance given by the ETKF when the observation error covariance is also given by the unit matrix, for either a linearized or nonlinear observation operator. It is now useful to introduce in this present context some indicators that can be used to quantify the extent of uncertainty reduction when the analysis estimate rather than a forecast from a previous analysis is considered. In particular, it is here useful to focus on three indicators: the signal-to-noise ratio, the number of degrees of freedom for signal and the information content of the measurements, which are discussed below.

By definition, $\mathbf{S}$ is the product of the inverse of the square root of the measurement error covariance and the square root of the forecast error covariance multiplied on the left by the observation operator. This means that when a single direct observation of a given model variable at model grid point i is considered, the ith element of $\mathbf{s}^T$ – in this case the matrix $\mathbf{S}$ becomes a row vector denoted as $\mathbf{s}^T$ – is equal to the ratio between the forecast error and the measurement error standard deviation, which defines the signal-to-noise ratio of the observation. It follows that the singular values γ_i of $\mathbf{S}$ represent the signal-to-noise ratio values of the components of the measurements along the corresponding left singular vectors of $\mathbf{S}$ or columns of $\mathbf{E}$. This means that there are only $p \leq \min(m, n)$ measurement components with positive signal-to-noise ratio, along the left singular vectors of $\mathbf{S}$ that correspond to the positive singular values of $\mathbf{S}$. Also, the $r \leq p$ left singular vectors of $\mathbf{S}$ corresponding to singular values of $\mathbf{S}$ that are greater than about one define the directions where measurements can detect variations in the state that are larger than measurement noise. It is also possible to show that the number of degrees of freedom for signal d_s is given by (e.g. Rodgers, 2000, his section 2.4.2)

$$d_s = \mathrm{tr}(\mathbf{S}^T(\mathbf{S}\mathbf{S}^T + \mathbf{I}_m)^{-1}\mathbf{S}) = \sum_{i=1}^{p} \frac{\gamma_i^2}{1+\gamma_i^2}. \tag{26}$$

This means that only p out m measurement components contribute – each for an amount given by $\gamma_i^2/(1+\gamma_i^2)$ – to the total number of degrees of freedom for signal. Finally, the information content h of the measurements in the case of Gaussian errors can be calculated as (e.g. Rodgers, 2000, his section 2.5)

$$h = \frac{1}{2}\log_2|\mathbf{S}\mathbf{S}^T + \mathbf{I}_m| = \frac{1}{2}\sum_{i=1}^{p}\log_2(1+\gamma_i^2), \tag{27}$$

where $|\mathbf{S}\mathbf{S}^T + \mathbf{I}_m|$ denotes the determinant of $\mathbf{S}\mathbf{S}^T + \mathbf{I}_m$. Again, only p out m measurement components contribute – each for an amount given by $1+\gamma_i^2$ – to the total information content of the measurements.

As discussed in section 3, when the square root of the forecast error covariance is approximated by means of an ensemble of forecasts, the matrix $\mathbf{S}$ is approximated by $\tilde{\mathbf{S}}$. In this case there are only $\tilde{p} \leq \min(m, K-1)$ measurements that provide information, i.e. with $\tilde{\gamma}_i > 0$, so that the effective number of degrees of freedom for signal $\tilde{d}_s$ resulting from the use of a reduced-rank forecast error covariance can be written as (Rodgers, 2000; Zupanski *et al.*, 2007b)

$$\tilde{d}_s = \mathrm{tr}(\tilde{\mathbf{S}}^T(\tilde{\mathbf{S}}\tilde{\mathbf{S}}^T + \mathbf{I}_m)^{-1}\tilde{\mathbf{S}}) = \sum_{i=1}^{\tilde{p}} \frac{\tilde{\gamma}_i^2}{1+\tilde{\gamma}_i^2}. \tag{28}$$

It follows that *when the true forecast error covariance matrix is approximated as the sample covariance matrix calculated using a forecast ensemble of size K, there are at most* $K-1$ *components of the measurement vector* $\mathbf{y}^{o\prime}$ *that can provide information*. Note that the above result is consistent with the discussion provided in Lorenc (2003, his section 3b and Appendix A), where the special case of a perfect observation is considered. The importance of this consideration is that it is now possible to define an appropriate threshold and decide which observational components are worth assimilating. For example, we may want to assimilate a given component only if it provides: (a) a signal-to-noise ratio greater than about 1, (b) an information content $h_i = \frac{1}{2}\log_2(1+\tilde{\gamma}_i^2)$ greater than about 0.5 or (c) more than about half a degree of freedom for signal. Conditions (a), (b) and (c) are all equivalent. It is also possible, however, to choose a more or less restrictive threshold according to the circumstances. It follows that when $m \gg K$, only the $r < K$ leading singular values and vectors of $\tilde{\mathbf{S}}$ need to be determined, e.g. by using the Lanczos method (e.g. Golub and van Loan, 1996, section 9.3.3).

5. Data selection strategy

We now illustrate a practical strategy to reduce the number of observational components without significant information loss. To this end, let us define $\mathbf{y}^{o\prime\prime} \in \mathbb{R}^r$ as $\mathbf{y}^{o\prime\prime} \equiv \tilde{\mathbf{E}}_r^T\mathbf{y}^{o\prime}$, where $\tilde{\mathbf{E}}_r \in \mathbb{R}^{m\times r}$ is the matrix whose columns are the r left singular vectors corresponding to the r positive singular values of $\tilde{\mathbf{S}}$

that are greater than a given threshold, with $r \leq \tilde{p}$. From Eq. (5) we can write

$$\mathbf{y}^{o\prime\prime} = \tilde{\mathbf{E}}_r^T \mathbf{H}' \mathbf{x}^t + \tilde{\mathbf{E}}_r^T \boldsymbol{\epsilon}^{o\prime} = \mathbf{H}'' \mathbf{x}^t + \boldsymbol{\epsilon}^{o\prime\prime}, \tag{29}$$

while from Eq. (6) we can write

$$\mathbf{y}^{o\prime\prime} = \mathbf{E}_r^T H'(\mathbf{x}^t) + \mathbf{E}_r^T \boldsymbol{\epsilon}^{o\prime} = H''(\mathbf{x}^t) + \boldsymbol{\epsilon}^{o\prime\prime}, \tag{30}$$

where $\mathbf{H}'' \in \mathbb{R}^{r\times n}$ and $H''(\mathbf{x}^t) \in \mathbb{R}^r$ are defined as $\mathbf{H}'' \equiv \tilde{\mathbf{E}}_r^T \mathbf{H}'$ and $H''(\mathbf{x}^t) \equiv \mathbf{E}_r^T H'(\mathbf{x}^t)$, respectively. Note that the covariance of $\boldsymbol{\epsilon}^{o\prime\prime} \equiv \mathbf{E}_r^T \boldsymbol{\epsilon}^{o\prime} \in \mathbb{R}^r$ is $\mathbf{I}_r$, the unit matrix of rank r. We will now derive an expression for the analysis ensemble that considers only informative measurements, both in the ensemble square root and the ensemble Kalman filter (EnKF) case.

5.1. Square root filter algorithm

From Eq. (19), the analysis error covariance can now be written as

$$\tilde{\mathbf{P}}^a = \mathbf{X}'^f (\mathbf{I}_K - \tilde{\mathbf{S}}'^T (\tilde{\mathbf{S}}' \tilde{\mathbf{S}}'^T + \mathbf{I}_r)^{-1} \tilde{\mathbf{S}}') \mathbf{X}'^{fT}, \tag{31}$$

where $\tilde{\mathbf{S}}' \in \mathbb{R}^{r\times K}$ is defined as $\tilde{\mathbf{S}}' \equiv \tilde{\mathbf{E}}_r^T \tilde{\mathbf{S}}$. Note that from our definition it follows that $\tilde{\mathbf{S}}'$ can be written either as $\tilde{\mathbf{S}}' = \mathbf{H}'' \mathbf{X}'^f$ or as in Eqs (20) and (21) when all primed quantities are replaced with double-primed ones, depending on the relationship between $\mathbf{y}^{o\prime\prime}$ and $\mathbf{x}^t$. From the definition of $\tilde{\mathbf{S}}$ it follows that $\tilde{\mathbf{S}}' = \tilde{\boldsymbol{\Gamma}}_r \tilde{\mathbf{V}}_r^T$ so that Eq. (31) becomes

$$\begin{aligned}
\tilde{\mathbf{P}}^a &= \mathbf{X}'^f (\mathbf{I}_K - \tilde{\mathbf{V}}_r \tilde{\boldsymbol{\Gamma}}_r (\tilde{\boldsymbol{\Gamma}}_r^2 + \mathbf{I}_r)^{-1} \tilde{\boldsymbol{\Gamma}}_r \tilde{\mathbf{V}}_r^T) \mathbf{X}'^{fT} \\
&= \mathbf{X}'^f (\mathbf{I}_K - \tilde{\mathbf{V}}_r \tilde{\mathbf{G}}_r \tilde{\mathbf{V}}_r^T) \mathbf{X}'^{fT} \\
&= \mathbf{X}'^f \tilde{\mathbf{V}} (\mathbf{I}_K - \tilde{\mathbf{G}}) \tilde{\mathbf{V}}^T \mathbf{X}'^{fT} \\
&= \mathbf{X}'^f \tilde{\mathbf{V}} (\tilde{\mathbf{Y}}_{K,r} + \mathbf{I}_K)^{-1} \tilde{\mathbf{V}}^T \mathbf{X}'^{fT},
\end{aligned} \tag{32}$$

with

$$\tilde{\mathbf{G}} \in \mathbb{R}^{K\times K} = \begin{pmatrix} \tilde{\mathbf{G}}_r & \mathbf{0}_{r\times(K-r)} \\ \mathbf{0}_{(K-r)\times r} & \mathbf{0}_{K-r} \end{pmatrix}, \tag{33}$$

where $\tilde{\mathbf{G}}_r \in \mathbb{R}^{r\times r} \equiv \tilde{\boldsymbol{\Gamma}}_r^2 (\tilde{\boldsymbol{\Gamma}}_r^2 + \mathbf{I}_r)^{-1}$ is a diagonal matrix with diagonal elements equal to the degrees of freedom for signal provided by each of the r components of $\mathbf{y}^{o\prime\prime}$ with $\tilde{\gamma}_i$ greater than a given threshold. Note also that $(\tilde{\mathbf{Y}}_{K,r} + \mathbf{I}_K)^{-1}$ is a diagonal matrix with diagonal elements equal to the degrees of freedom for noise provided by each of the r components of $\mathbf{y}^{o\prime\prime}$ (e.g. Rodgers 2000, his section 2.4.2). This can be interpreted as follows: the more a measurement component or degree of freedom is related to the signal, the more it reduces the corresponding variance of the analysis error in observation space (see also Wang and Bishop, 2003, Appendix A) and the more it is worth assimilating.

From Eqs (16) and (32) it follows that the analysis perturbation matrix can be written as

$$\mathbf{X}'^a = \mathbf{X}'^f \tilde{\mathbf{V}} (\tilde{\mathbf{Y}}_{K,r} + \mathbf{I}_K)^{-1/2} \tilde{\mathbf{V}}^T, \tag{34}$$

while the analysis ensemble mean can be calculated as (see Evensen, 2004, his section 3)

$$\begin{aligned}
\overline{\mathbf{x}^a} &= \overline{\mathbf{x}^f} + \mathbf{X}'^f \tilde{\mathbf{S}}'^T (\tilde{\mathbf{S}}' \tilde{\mathbf{S}}'^T + \mathbf{I}_r)^{-1} \mathbf{d}'' \\
&= \overline{\mathbf{x}^f} + \mathbf{X}'^f \tilde{\mathbf{V}}_r \tilde{\boldsymbol{\Gamma}}_r (\tilde{\boldsymbol{\Gamma}}_r^2 + \mathbf{I}_r)^{-1} \mathbf{d}'',
\end{aligned} \tag{35}$$

where $\mathbf{d}'' \in \mathbb{R}^r$ is the innovation vector, defined either as $\mathbf{d}'' \equiv \mathbf{y}^{o\prime\prime} - \mathbf{H}'' \overline{\mathbf{x}^f}$ or as $\mathbf{d}'' \equiv \mathbf{y}^{o\prime\prime} - \overline{H''(\mathbf{x}^f)}$ (see Eq. (21) with double-primed quantities) depending on the relationship between $\mathbf{y}^{o\prime}$ and $\mathbf{x}^t$. From Eq. (35) we can then write

$$\begin{aligned}
\overline{\mathbf{x}^a} &= \overline{\mathbf{x}^f} + \mathbf{X}'^f \tilde{\mathbf{V}}_r \boldsymbol{\delta} \\
&= \overline{\mathbf{x}^f} + \sum_{i=1}^{r} \delta_i \mathbf{X}'^f \tilde{\mathbf{v}}_i,
\end{aligned} \tag{36}$$

where $\boldsymbol{\delta} \equiv \tilde{\boldsymbol{\Gamma}}_r (\tilde{\boldsymbol{\Gamma}}_r^2 + \mathbf{I}_r)^{-1} \mathbf{d}'' \in \mathbb{R}^r$, $\delta_i \equiv \gamma_i (1 + \gamma_i^2)^{-1} d_i''$ is the ith component of $\boldsymbol{\delta}$ and $\tilde{\mathbf{v}}_i \in \mathbb{R}^K$ is the ith right singular vector of $\tilde{\mathbf{S}}$.

5.2. Ensemble Kalman filter algorithm

From Eq. (35), the analysis update equation for the jth member of the forecast ensemble can be written as (Evensen, 2003, his equation 20)

$$\mathbf{x}_j^a = \mathbf{x}_j^f + \mathbf{X}'^f \tilde{\mathbf{S}}'^T (\tilde{\mathbf{S}}' \tilde{\mathbf{S}}'^T + \mathbf{I}_r)^{-1} (\mathbf{y}_j^{o\prime\prime} - \mathbf{H}'' \mathbf{x}_j^f), \tag{37}$$

or, when $\mathbf{y}^{o\prime\prime}$ is nonlinearly related to $\mathbf{x}^t$, as

$$\mathbf{x}_j^a = \mathbf{x}_j^f + \mathbf{X}'^f \tilde{\mathbf{S}}'^T (\tilde{\mathbf{S}}' \tilde{\mathbf{S}}'^T + \mathbf{I}_r)^{-1} (\mathbf{y}_j^{o\prime\prime} - H''(\mathbf{x}_j^f)), \tag{38}$$

where $\mathbf{y}_j^{o\prime\prime} \in \mathbb{R}^r$ is the jth member of an observation ensemble with ensemble mean $\mathbf{y}^{o\prime\prime}$ and where the ensemble error variance for each component of $\mathbf{y}_j^{o\prime\prime}$ is equal to 1. Note that in Eqs (37) and (38) we have assumed use of the full-rank expression for the observation error covariance matrix rather than its ensemble representation, given that it is simply given by $\mathbf{I}_r$. This also ensures that $\tilde{\mathbf{S}}' \tilde{\mathbf{S}}'^T + \mathbf{I}_r$ is non-singular. Let us now define $\mathbf{X}^a \equiv (\mathbf{x}_1^a, \mathbf{x}_2^a, \ldots, \mathbf{x}_K^a)$, $\mathbf{X}^f \equiv (\mathbf{x}_1^f, \mathbf{x}_2^f, \ldots, \mathbf{x}_K^f)$ and $\mathbf{D} \equiv (\mathbf{d}_1'', \mathbf{d}_2'', \ldots, \mathbf{d}_K'') \in \mathbb{R}^{r\times K}$, where the jth innovation ensemble member $\mathbf{d}_j''$ is defined either as $\mathbf{d}_j'' \equiv \mathbf{y}_j^{o\prime\prime} - \mathbf{H}'' \mathbf{x}_j^f$ or as $\mathbf{d}_j'' \equiv \mathbf{y}_j^{o\prime\prime} - H''(\mathbf{x}_j^f)$ according to whether Eq. (37) or Eq. (38) is used, respectively. From Eqs (37) and (38) we can write

$$\mathbf{X}^a = \mathbf{X}^f + \mathbf{X}'^f \tilde{\mathbf{S}}'^T (\tilde{\mathbf{S}}' \tilde{\mathbf{S}}'^T + \mathbf{I}_r)^{-1} \mathbf{D}, \tag{39}$$

which is equivalent to (see Evensen, 2004, his section 2.3)

$$\mathbf{X}^a = \mathbf{X}^f (\mathbf{I}_K + \tilde{\mathbf{S}}'^T (\tilde{\mathbf{S}}' \tilde{\mathbf{S}}'^T + \mathbf{I}_r)^{-1} \mathbf{D}). \tag{40}$$

From Eq. (35) we can write

$$\begin{aligned}
\mathbf{X}^a &= \mathbf{X}^f (\mathbf{I}_K + \tilde{\mathbf{V}}_r \tilde{\boldsymbol{\Gamma}}_r (\tilde{\boldsymbol{\Gamma}}_r^2 + \mathbf{I}_r)^{-1} \mathbf{D}) \\
&= \mathbf{X}^f (\mathbf{I}_K + \tilde{\mathbf{V}}_r \Delta),
\end{aligned} \tag{41}$$

where $\Delta \equiv (\boldsymbol{\delta}_1, \boldsymbol{\delta}_2, \ldots, \boldsymbol{\delta}_K) \equiv \tilde{\boldsymbol{\Gamma}}_r (\tilde{\boldsymbol{\Gamma}}_r^2 + \mathbf{I}_r)^{-1} \mathbf{D} \in \mathbb{R}^{r\times K}$ and where the jth scaled-innovation ensemble member $\boldsymbol{\delta}_j \in \mathbb{R}^r$ is defined as $\boldsymbol{\delta}_j \equiv \tilde{\boldsymbol{\Gamma}}_r (\tilde{\boldsymbol{\Gamma}}_r^2 + \mathbf{I}_r)^{-1} \mathbf{d}_j''$. From Eq. (41) it follows that the jth analysis ensemble member $\mathbf{x}_j^a$ can be written as

$$\mathbf{x}_j^a = \mathbf{x}_j^f + \sum_{i=1}^{r} \delta_{i,j} \mathbf{X}^f \tilde{\mathbf{v}}_i, \tag{42}$$

where $\delta_{i,j}$ is the i-th component of $\boldsymbol{\delta}_j$.

5.3. Localization considerations

It is important to note that the data selection strategy presented above is compatible with existing localization procedure that may be used for ensemble data assimilation, as localization is enforced prior to the use of the data selection algorithms described in sections 5.1 and 5.2. For example, it is possible first to select only observations within a given distance from an analysis grid point or to multiply the elements of the inverse of the measurement error covariance matrix by a correlation function decreasing with distance from a given analysis grid point (Hunt *et al.*, 2007). The algorithms in sections 5.1 and 5.2 can then be applied without modification. When localization is used, the data selection procedure will result in a further data reduction over the local domain. In this way it is possible to use localization procedures with a larger radius of influence or correlation functions whose support spans a larger part of their domain, while only assimilating a number of measurement components – in general, not greater than the rank of the forecast error covariance matrix used for assimilation – whose signal-to-noise ratio, information content or contribution to the number of degrees of freedom for signal is below a chosen threshold. The problem of choosing the appropriate dimension of the local domain for the system under consideration then merely becomes that of finding a trade-off between the need, on one side, of reducing the rank deficiency of the forecast error covariance matrix for a given ensemble size as well as the risk of the occurrence of spurious covariances between distant parts of the domain and, on the other side, of avoiding the risk of shortening the natural correlation length scales of model fields that may lead to unbalanced initial conditions (Cohn *et al.*, 1998; Mitchell *et al.*, 2002; Lorenc, 2003; Kepert, 2009; Greybush *et al.*, 2011) as well as the need for constraining the analysis with an adequate number of observations.

Finally, note that Zupanski *et al.* (2007a) discusses a localization procedure also based on the information content of the measurements, where the region of influence of the observations is defined as the area of the domain where the ratio between the forecast error and the analysis error standard deviation is greater than an empirically chosen cut-off value. Also, note that the data selection procedure discussed in this paper is closely related to the efficient subspace pseudo inversion method described in Evensen (2004, his section 7.3) and Evensen (2009, his section 14.2), but it avoids introducing any approximations in the case of a non-diagonal observation error covariance matrix and it determines the number of independent observational components that is useful to assimilate based on their information content.

6. Experimental set-up

In this section, the prognostic model and the assimilation strategy used for a number of data assimilation experiments, whose aim is to test the data selection strategy described in this paper, are presented.

6.1. Description of the model

The model used in this study is an extension of that presented in Evensen (2004) and given by the one-dimensional temperature advection equation

$$\frac{\partial T(\mathbf{x},t)}{\partial t} + u(\mathbf{x},t)\frac{\partial T(\mathbf{x},t)}{\partial x} = 0, \tag{43}$$

with random initial condition $T(\mathbf{x},0) = T_0(\mathbf{x})$, where $\mathbf{x} = (x,y,p)^T$ with x and y being the zonal and meridional coordinates of a grid point at a given pressure p. The origin of the y coordinate is chosen to correspond to a latitude of 45°. The advection speed $u(\mathbf{x},t)$ is first assumed constant (linear advection equation case, $u(\mathbf{x},t) = u_0$) and then to be dependent on $T(\mathbf{x},t)$ (nonlinear advection case). In the latter case, we assume that u and T are related according to the thermal wind equation, given by (e.g. Gill, 1982, p. 217)

$$f\frac{\partial u}{\partial p} = -\rho^{-2}\frac{\partial \rho}{\partial y}, \tag{44}$$

where ρ is the atmospheric density and $f = 10^{-4}$ s^{-1} is the Coriolis parameter appropriate to 45° latitude. For a perfect

Table 1. Model's vertical levels and reference temperature. Model height values correspond to those for an atmosphere in hydrostatic balance with a 1013.25 hPa surface pressure and a 7.5 km scale height.

Level	Pressure (hPa)	Height (km)	Temperature (K)
1	0.100	69.176	219.10
2	0.290	61.191	249.82
3	0.690	54.690	255.99
4	1.420	49.277	257.54
5	2.611	44.709	251.72
6	4.407	40.783	243.48
7	6.950	37.366	235.02
8	10.370	34.365	229.33
9	14.810	31.692	226.62
10	20.400	29.290	223.05
11	27.260	27.116	218.62
12	35.510	25.133	215.08
13	45.290	23.309	212.13
14	56.730	21.620	207.95
15	69.970	20.046	202.51
16	85.180	18.571	194.34
17	102.050	17.216	193.15
18	122.040	15.874	198.05
19	143.840	14.642	205.49
20	167.950	13.479	212.48
21	194.360	12.384	219.84
22	222.940	11.355	226.74
23	253.710	10.385	233.39
24	286.600	9.471	240.02
25	321.500	8.609	246.66
26	358.280	7.797	253.05
27	396.810	7.031	258.76
28	436.950	6.308	263.40
29	478.540	5.626	267.21
30	521.460	4.982	271.02
31	565.540	4.374	275.13
32	610.600	3.799	278.31
33	656.430	3.256	281.06
34	702.730	2.745	283.33
35	749.120	2.265	285.48
36	795.090	1.818	287.82
37	839.950	1.407	290.44
38	882.800	1.034	292.76
39	922.460	0.704	294.82
40	957.440	0.425	296.76
41	985.880	0.205	299.19
42	1005.430	0.058	300.97
43	1013.250	0.000	301.64

Table 2. IASI channel selection for temperature retrieval (adapted from Collard, 2007, his Appendix A).

Channel	Wave number (cm^{-1})	Noise SD (K)
72	662.75	0.37
87	666.50	0.37
89	667.00	0.33
92	667.75	0.31
95	668.50	0.32
97	669.00	0.33
99	669.50	0.33
125	676.00	0.33
135	678.50	0.34
138	679.25	0.33
141	680.00	0.34
148	681.75	0.34
154	683.25	0.34
167	686.50	0.34
199	694.50	0.33
205	696.00	0.32
243	705.50	0.29
249	707.00	0.29
252	707.75	0.29
254	708.25	0.29
260	709.75	0.29
262	710.25	0.29
265	711.00	0.28
267	711.50	0.28
269	712.00	0.28
275	713.50	0.29
282	715.25	0.27
294	718.25	0.29
296	718.75	0.30
303	720.50	0.29
306	721.25	0.28
323	725.50	0.26
327	726.50	0.26
329	727.00	0.27
335	728.50	0.27
345	731.00	0.26
347	731.50	0.28
350	732.25	0.25
354	733.25	0.26
356	733.75	0.25
360	734.75	0.26
366	736.25	0.27
371	737.50	0.27
373	738.00	0.25
375	738.50	0.25
377	739.00	0.27
379	739.50	0.26
381	740.00	0.26
383	740.50	0.28
386	741.25	0.28
389	742.00	0.27
398	744.25	0.25
401	745.00	0.26
404	745.75	0.24
407	746.50	0.26
410	747.25	0.24
414	748.25	0.26
416	748.75	0.24
426	751.25	0.25
428	751.75	0.24
432	752.75	0.25
434	753.25	0.24
439	754.50	0.25
445	756.00	0.24
457	759.00	0.24
2239	1204.50	0.22

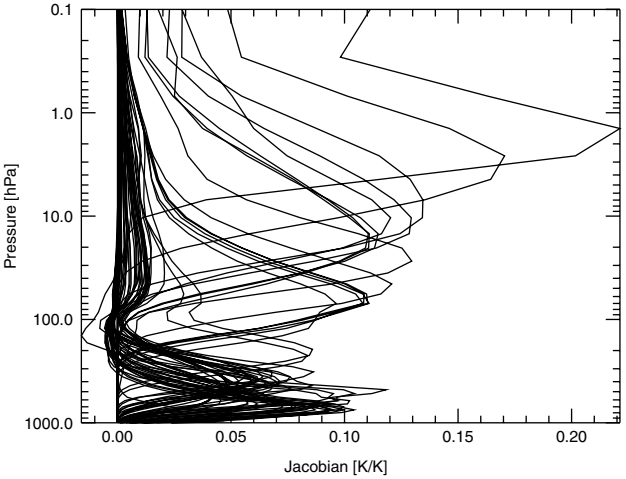

Figure 1. Temperature Jacobians for the 66 IASI channels shown in Table 2.

gas, Eq. (44) can also be written as

$$f\frac{\partial u}{\partial p} = \frac{R}{p}\left(\frac{\partial T}{\partial y}\right)_p, \tag{45}$$

where $R = 287.05\ \mathrm{J\,kg^{-1}\,K^{-1}}$ is the specific gas constant for dry air.

Let us now express $T(\mathbf{x}, t)$ as $T(\mathbf{x}, t) = T(x, y = 0, p, t)\Theta(\phi(y), p)/\Theta(45^\circ, p)$, where ϕ is the latitude and $\Theta(\phi(y), p)$ is the meridional temperature variation given by (Stull and Ahrens, 2000, their chapter 11)

$$\Theta(\phi(y), p) = c_1 + c_2(p)\left[\frac{3}{2}\left(\frac{2}{3} + \sin^2\phi\right)\cos^3\phi\right], \tag{46}$$

where $c_1 = 261$ K and

$$c_2(p) = c_3\frac{\mathrm{H}}{z_T}\ln(p/p_T), \tag{47}$$

where $c_3 = 40$ K, $\mathrm{H} = 7.5$ km is the atmospheric scale height, $z_T = 11$ km is the average depth of the troposphere and $p_T = 233.75$ hPa is the average tropospheric pressure. In this case, Eq. (45) can be written as

$$\frac{\partial u}{\partial p} = -\frac{15Rc_2(p)}{2afp\Theta(45^\circ, p)}T(x, y = 0, p, t)\sin^3\phi\cos^2\phi, \tag{48}$$

where $a = 6371 \times 10^3$ m is the radius of the Earth. Hereafter we will only consider the evolution of temperature on the $y = 0$ domain, representing a section of the atmosphere through a circle of latitude at 45°, which is periodic (with period L) in the zonal direction represented by the x-axis. This means that from now on we will redefine T as $T \equiv T(x, y = 0, p, t) = T(\mathbf{x}, t)$, with $\mathbf{x} = (x, p)^T$, and from Eq. (48) we can write

$$\frac{\partial u}{\partial p} = -\frac{\mathrm{H}u_0\ln(p/p_T)}{z_T\Theta(45^\circ, p)}\frac{T}{p}, \tag{49}$$

where $u_0 \equiv 15\sqrt{2}Rc_3/(16af) \simeq 23.9\ \mathrm{ms^{-1}}$. Note that Eq. (49) is consistent with a meridional temperature gradient that changes sign in the stratosphere, which accounts for the fact that in the stratosphere the Tropics are colder than the polar regions.

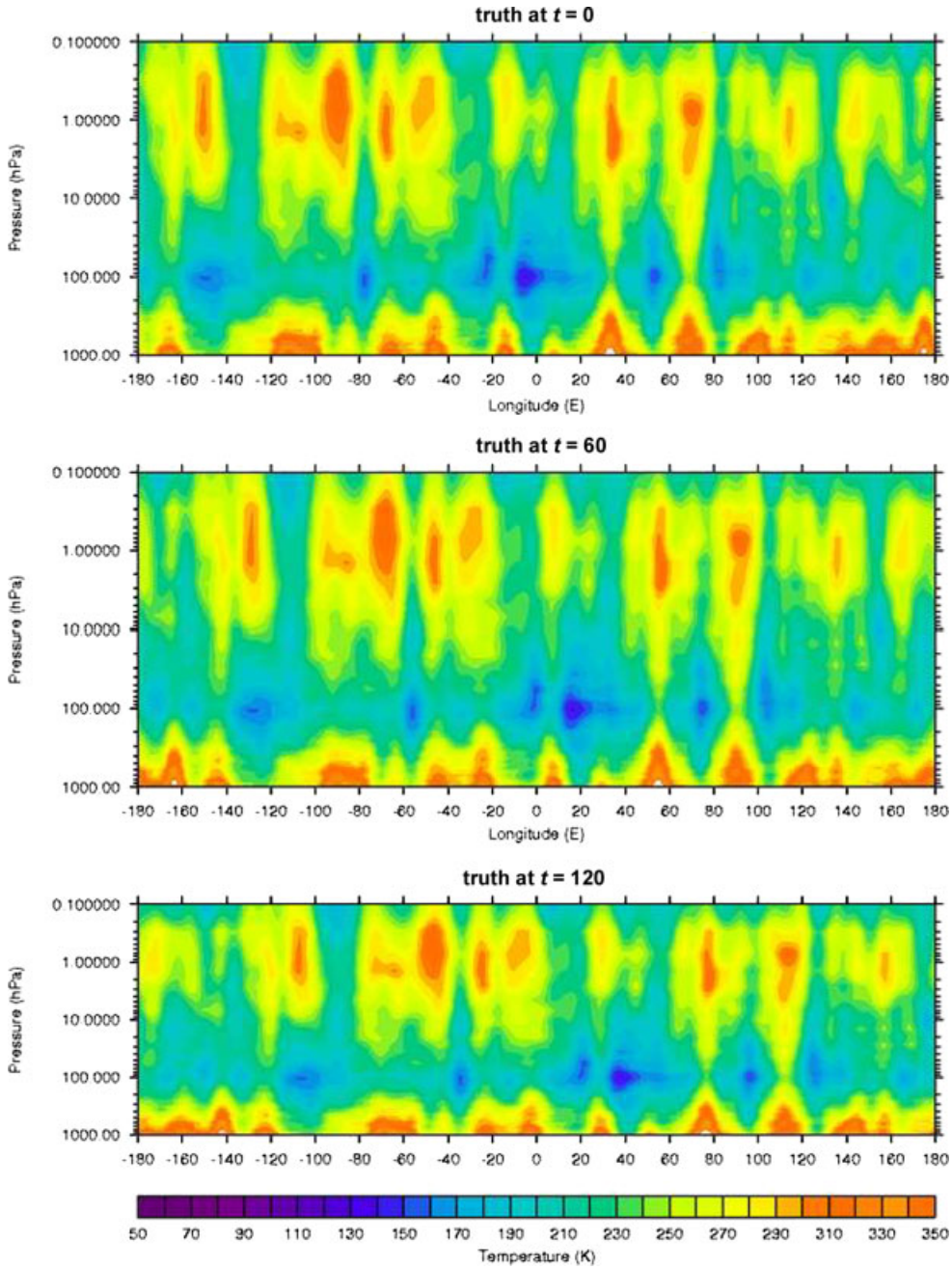

Figure 2. True temperature field at $t = 0$ (top), $t = 60\Delta t$ (middle) and $t = 120\Delta t$ (bottom), with no assimilation. This figure is available in colour online at wileyonlinelibrary.com/journal/qj

The zonal length of the domain considered for our experiments is $L = 1000\Delta x$, while there are 43 pressure levels in the vertical between 0.1 hPa and 1013.25 hPa (see Table 1). The advection equation (43) is discretized using a forward-upstream finite difference scheme considering different zonal grid lengths and time steps in the two advection cases. In the linear advection case, $\Delta x = 1$ (in arbitrary length units), a time step $\Delta t = 1$ (in arbitrary time units) and a prescribed constant zonal-only advection speed $u = 1$ (in the chosen length divided by time units) are assumed. In this way, the distance travelled in one time step is equal to the grid length and the Courant number C is equal to one so that no damping or erroneous dispersion (e.g. Pielke, 2002, chapter 10) are introduced. In the nonlinear advection case the wind speed is calculated from Eq. (49) using the Euler method with a constant value $u = 0\ \mathrm{m\,s^{-1}}$ at lowest model level (no-slip condition). The zonal grid length is in this case given by $\Delta x = 28.3$ km, appropriate for a 45° circle of latitude subdivided in 1000 grid cells. Also, the time step Δt is chosen to be equal to 60 s so as to ensure that the Courant number is never greater than one for the nonlinear advection experiments performed in this work, so that the advection scheme is always linearly stable.

In the linear advection case the reference trajectory describes the evolution of the true atmosphere over a period of time slightly greater than that required for the temperature field to be advected from one observation location to the next. Longer time periods are considered in the nonlinear advection case. The initial conditions for the reference trajectory are represented by a realization of a random field that is constructed as the following. First, a one-dimensional Gaussian correlation function with standard

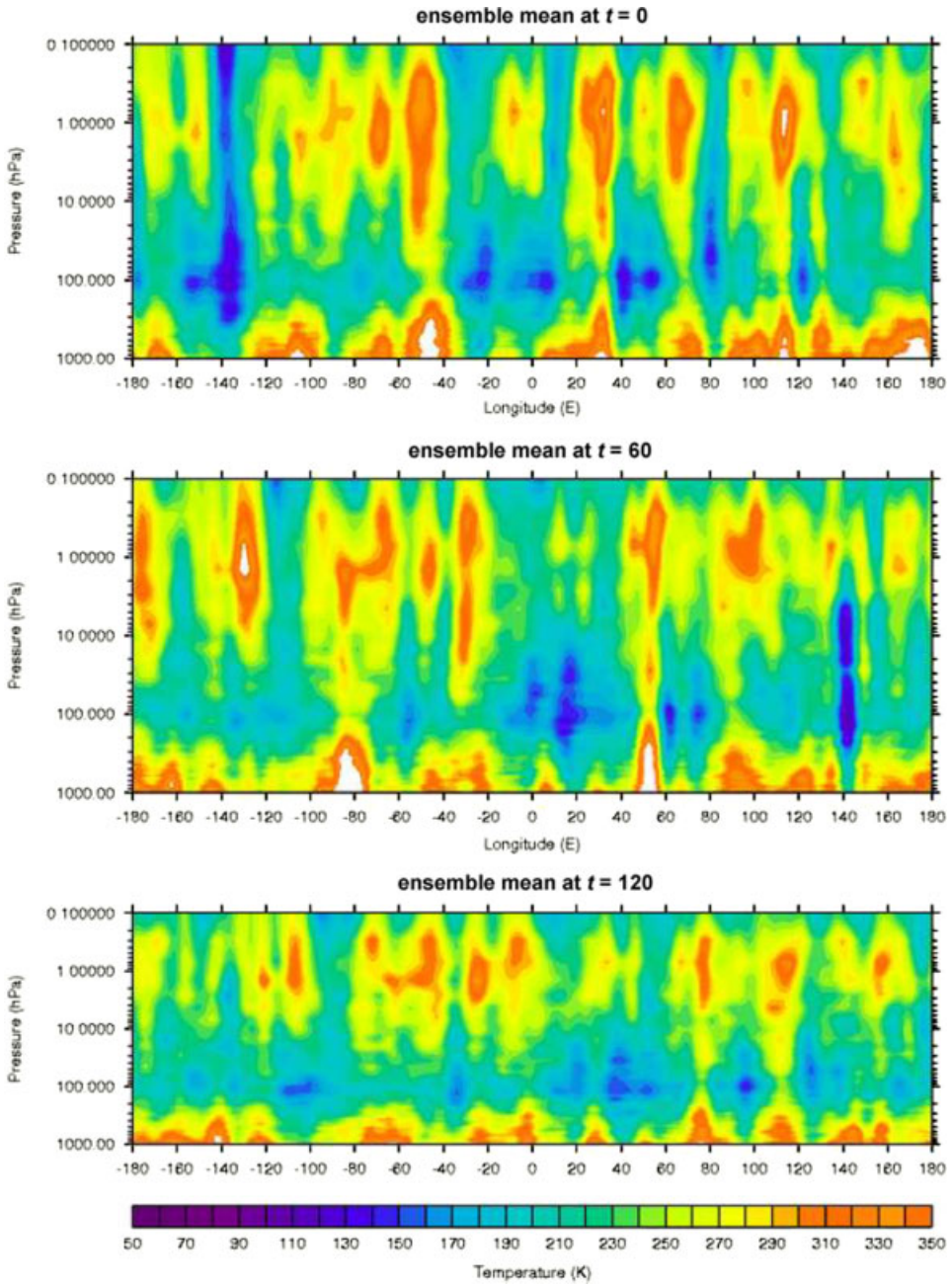

Figure 3. As Figure 2 but for the case when all $43 \times 8 = 344$ *in situ* observations are assimilated every $5\Delta t$ using a standard ensemble square root filter. This figure is available in colour online at wileyonlinelibrary.com/journal/qj

deviation chosen equal to $10\sqrt{2}$ is Fourier transformed and the Fourier coefficients are square-rooted and multiplied by $\exp(2\pi i\phi)$, where i is the imaginary unit and ϕ is a realization of a random number that is uniformly distributed between 0 and 1 (see Evensen, 2009, section 11.2). The resulting random vector is then inverse-Fourier transformed after imposing the conditions for the resulting field to have no imaginary part, to determine $w_1(x)$, where x is the zonal coordinate and subscript denotes a given model vertical level, starting from the top. Then, we set $\psi_1(x) = w_1(x)$ and for $j = 2, \ldots, 43$, we define

$$\psi_j(x) = \rho_{j-1}\psi_{j-1}(x) + \sqrt{1-\rho_{j-1}^2}\, w_j(x), \qquad (50)$$

where $\rho_{j-1} = \exp[-\Delta z_{j-1}/H_z]$, $\Delta z_{j-1} \equiv z_{j-1} - z_j$ is the thickness in kilometres of the jth model layer, given by the hypsometric equation with a 7.5 km scale height, and $H_z = 50$ km is the vertical de-correlation length. From the first-order autoregressive model in Eq. (50) it follows that, at a given x, $\psi_j(x)$ and $\psi_i(x)$ are vertically correlated, with correlation $\exp[-|z_j - z_i|/H_z]$. The true temperature field $T_j(x)$ at initial time at level j is determined from $\psi_j(x)$ as

$$T_j(x) = T_j^{\mathrm{ref}} + \sigma_{T_j}\psi_j(x), \qquad (51)$$

where T_j^{ref} is the reference temperature at level j as given in Table 1 and σ_{T_j} is chosen as 10% of T_j^{ref}. The initial conditions for the 'background' trajectory are defined as

$$T_j^a(x) = T_j(x) + \sigma_{T_j}\psi_j'(x), \qquad (52)$$

where $\psi_j'(x)$ is another realization of the same random field used to generate $\psi_j(x)$. Finally, an ensemble of K initial conditions is determined in a similar manner, and is constrained to have $T_j^a(\boldsymbol{x})$ as its ensemble mean.

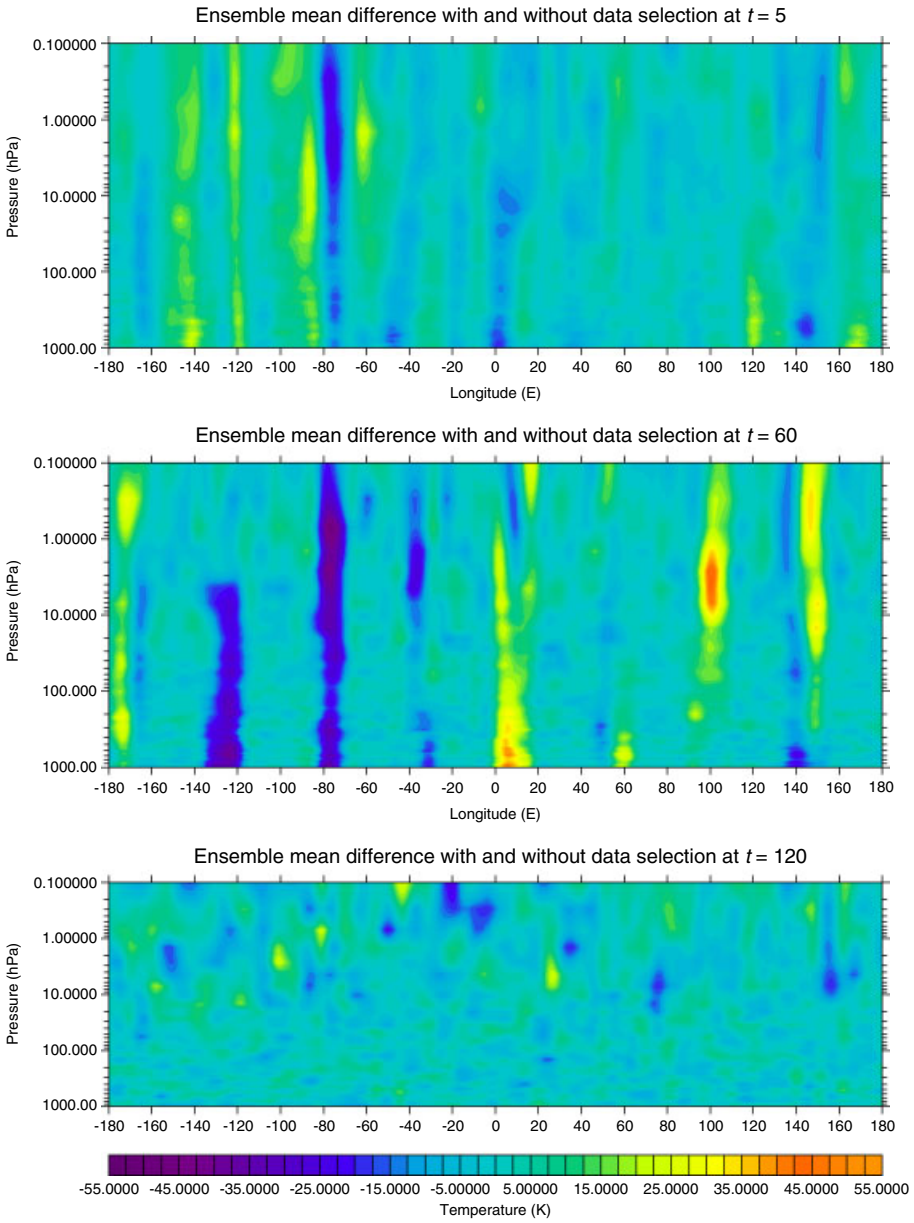

Figure 4. Difference between the ensemble mean temperature field at $t = 5$ (top), $t = 60\Delta t$ (middle) and $t = 120\Delta t$ (bottom) when only observations with signal-to-noise ratio $\tilde{\gamma}_i > 1$ are assimilated every $5\Delta t$ using the data-selective square root filter and the ensemble mean temperature field shown in Figure 3. The experiment made use of a 300-member forecast ensemble with no localization. Note that for $0 \leq t < 5$ the temperature difference is identically zero by construction. This figure is available in colour online at wileyonlinelibrary.com/journal/qj

6.2. *Assimilation strategy*

Each initial condition is propagated forward in time until observation time, when an analysis scheme based either on a standard mean-preserving ensemble square-root method (see Evensen, 2009, section 13.1) or on one of the two data-selective schemes described in section 5 generates a new set of initial conditions. The relative performance of the data-selective square root scheme, described in section 5.1, and of data-selective EnKF, described in section 5.2, is also investigated. The experiments described in this paper make use of a $5\Delta t$ observation frequency, and at each observation time the observation vector is composed of 8 or 16 regularly spaced vertical temperature profiles with 43 elements each or of a set of satellite radiances emerging from eight regularly spaced zonal locations. All observations are simulated from the truth by using an appropriate observation operator, which in the case of *in situ* observations amount to a trivial vertical interpolation operator (as observation levels are assumed to coincide with model levels), and additional zero-mean random noise with standard deviation $\sigma^o_{T_j}$ chosen as 0.1% of T^{ref}_j.

Satellite observations are assumed to represent radiances measured by the Infrared Atmospheric Sounding Interferometer (IASI, Siméoni *et al.*, 1997) over 66 spectral channels that are suitable for atmospheric temperature profile retrieval (Collard, 2007), which are specified in Table 2. These are responsible for 62% of the total number of d_s for temperature calculated using all 8461 IASI channels, with the exception of those excluded through pre-screening. Note that a subset of 30 channels accounts for 55% of the total number of d_s for temperature. The observation operator for IASI radiances sensitive to temperature was determined using RTTOV version 8.7 (Saunders and Brunel, 2005), a radiative transfer model that is suitable for use within an

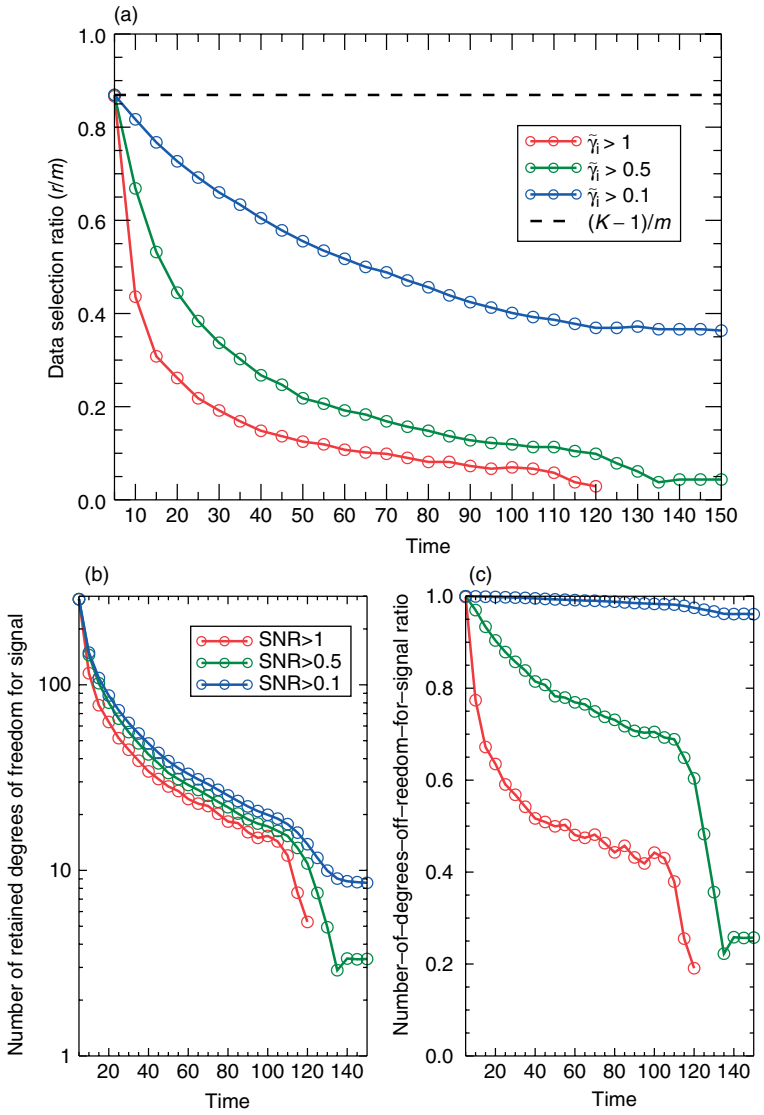

Figure 5. (a) Number of components r – relative to the total number $m = 344$ *in situ* observations – of $\mathbf{y}^{o\prime\prime}$ that have signal-to-noise ratio $\tilde{\gamma}_i > 1$ (red solid line), $\tilde{\gamma}_i > 0.5$ (green solid line) and $\tilde{\gamma}_i > 0.1$ (blue solid line) for a 300-member ensemble size with no localization. For reference, the highest achievable data selection ratio $\tilde{\gamma}_i = (K-1)/m$ (black dashed line) is also shown; (b) number of degrees of freedom for signal $\tilde{d}_s$ considering only observation components with $\tilde{\gamma}_i > 1$ (red solid line), $\tilde{\gamma}_i > 0.5$ (green solid line) and $\tilde{\gamma}_i > 0.1$ (blue solid line); (c) as in (b), with quantities scaled by the number of degrees of freedom for signal when all observation components are considered. This figure is available in colour online at wileyonlinelibrary.com/journal/qj

operational data assimilation system. For the purpose of this work, the observation operator for IASI temperature channels is assumed to be nearly linear about $\mathbf{T}^{\mathrm{ref}}$, so that Eq. (4) needs to be calculated only once. The temperature Jacobians, which are the rows of $\mathbf{H}$ corresponding to the 66 IASI channels described in Table 2, are shown in Figure 1.

A localization procedure has also been implemented, which selects for each grid point a subset of observations within a cylinder of a given radius l (Hunt *et al.*, 2007). The analysis at each grid point is then computed by considering only the innovations, the measurement error covariance matrix blocks and the rows of $\tilde{\mathbf{S}}$ corresponding to observations whose distance from the given grid point is not greater then l. For ease of use with satellite observations, the localization procedure has not been applied in the vertical, i.e. the height of the cylinder is taken to be equal to the depth of the atmosphere. Each analysis value at the centre of the local domain has also been averaged with its neighbouring values resulting from $2l' + 1$ local analysis (Ott *et al.*, 2004), with $l' < l$. In our experiments we found that this averaging contributes to a substantial reduction of high-spatial-frequency noise in the analysis. Finally, note that our assimilation experiments do not include the use of any inflation algorithms. This is to facilitate the direct comparison of the results of the experiments with and without the use of the data selection procedure and to prevent inconsistencies due to the possible occurrence of different inflation factors in the two sets of experiments. The absence of inflation also avoids increasing artificially the information content of the observations and this gives us the opportunity to assess the data selection procedure when it can discard a significant amount of observational components. Also note that our experiments, discussed in section 7, do not show indications that the magnitude of the ensemble-mean error gets progressively larger (see, for example, Figure 6) or other signs of the detrimental filter divergence effects arising from lack of inflation. This is arguably due to the constraints provided by the observation

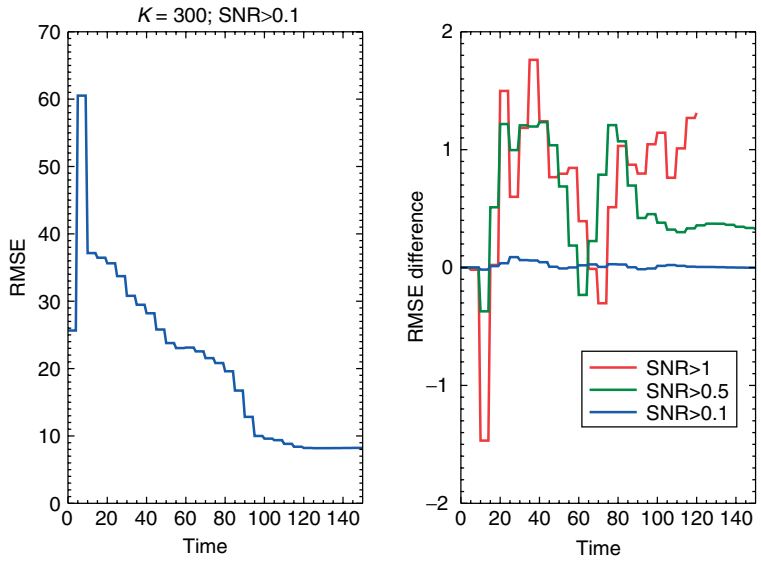

Figure 6. Root mean square error (RMSE) calculated as the difference between the ensemble mean and the true temperature field for a 300-member ensemble size with no localization, in the case when only observation components with $\tilde{\gamma}_i > 0.1$ are assimilated every $5\Delta t$ (left); evolution of the difference between the RMSE values obtained when considering only selected observations – with $\tilde{\gamma}_i > 1$ (red solid line), $\tilde{\gamma}_i > 0.5$ (green solid line) and $\tilde{\gamma}_i > 0.1$ (blue solid line) – and the RMSE values obtained when all observations are assimilated (right). This figure is available in colour online at wileyonlinelibrary.com/journal/qj

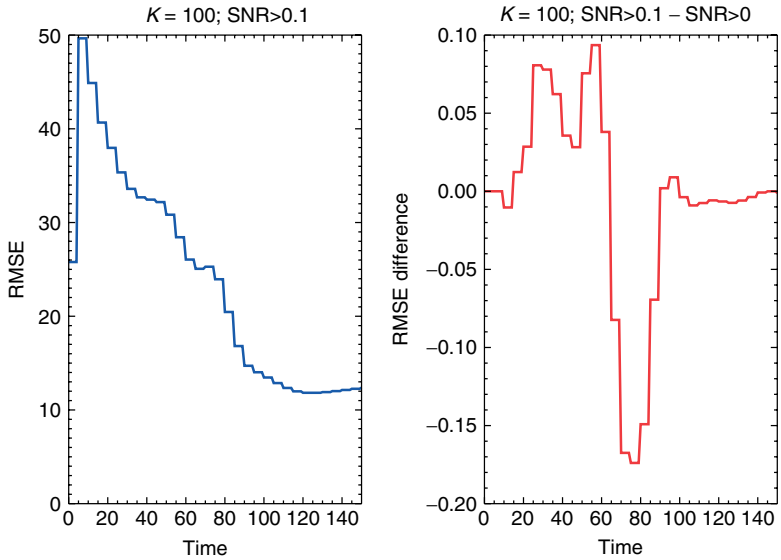

Figure 7. Root mean square error (RMSE) calculated as the difference between the ensemble mean and the true temperature field for a 100-member ensemble size and localization with a $200\Delta x$ radius of influence, in the case when only observations with $\tilde{\gamma}_i > 0.1$ are considered (left); evolution of the difference between the RMSE values obtained when considering only observations with $\tilde{\gamma}_i > 0.1$ and the RMSE values obtained when all observations are considered (right). This figure is available in colour online at wileyonlinelibrary.com/journal/qj

components assimilated in our experiments that have signal-to-noise ratio greater than a given threshold (see, for example, Figure 5).

7. Experiment results

In this section we discuss the results of a number of data assimilation experiments, described in section 6.2, which are carried out to assess whether we can safely discard observational components with information content below a given threshold and achieve comparable results as in the case when all observational components are retained. The effects of the use of localization procedures, remotely sensed observations and stochastic filters (i.e. data-selective EnKF) are also discussed. Unless stated otherwise, the experiments discussed below consider only *in situ* observations for assimilation.

Let us first consider the linear advection case. Figure 2 shows the true temperature field at times $t = 0$, $t = 60\Delta t$ and $t = 120\Delta t$, while Figure 3 shows the ensemble mean temperature estimated at the same times using a standard ensemble square root filter with a 300-member ensemble size, no localization and where all $43 \times 8 = 344$ available observations are assimilated every $5\Delta t$. The difference between the ensemble mean temperature estimated using a data-selective square root filter with a 300-member ensemble size and no localization, when only observations with signal-to-noise ratio $\tilde{\gamma}_i > 1$ are assimilated every $5\Delta t$, and the ensemble mean temperature estimated at the same times using a standard ensemble square root filter is shown in

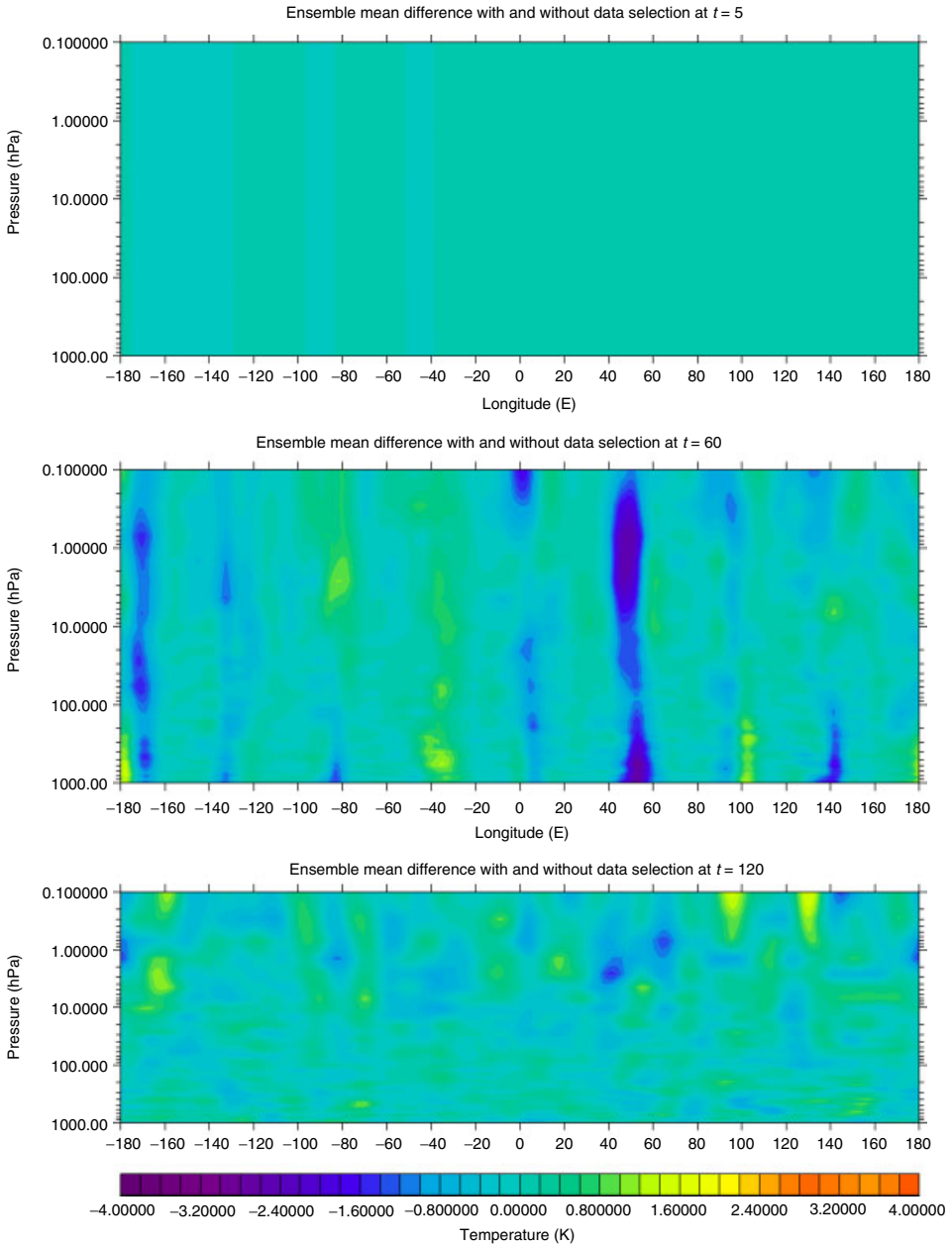

Figure 8. Difference between the ensemble mean temperature field at $t = 5$ (top), $t = 60\Delta t$ (middle) and $t = 120\Delta t$ (bottom) when only observations with signal-to-noise ratio $\tilde{\gamma}_i > 0.1$ are assimilated every $5\Delta t$ using the data-selective square root filter and the ensemble mean temperature field when all 344 *in situ* observations are assimilated every $5\Delta t$. The data-selective experiment made use of a 100-member forecast ensemble and localization with a 200-grid-point radius of influence. Note that for $0 \leq t < 5$ the temperature difference is identically zero by construction. This figure is available in colour online at wileyonlinelibrary.com/journal/qj

Figure 4. The last considered time is $120\Delta t$, as this is the last assimilation time in the data-selective experiment that measurements components with $\tilde{\gamma}_i > 1$ are available for assimilation. From Figure 4 it is possible to see that at the end of the two experiments most of the ensemble-mean temperature differences – namely about 72% – have a magnitude of less than 5 K, which is smaller than the minimum root mean square error (RMSE) value of the ensemble mean temperature field that is achieved when all observations are assimilated (see Figure 6).

Experiments with different data selection thresholds were also carried out. Figure 5(a) shows the evolution of the ratio between the number r of components of $\mathbf{y}^{o\prime\prime}$ that have signal-to-noise ratio $\tilde{\gamma}_i$ greater than a given threshold and the total number m of available measurements with independent errors. Also shown is the maximum value of the ratio defined above, given by $(K-1)/m$ in the considered case where $m > K$, as discussed in section 4. From Figure 5(a) it is evident that the largest number of informative measurement components, which is about 13% less than the total number of available measurements, is achieved at the beginning of the experiment, as expected, where forecast error uncertainty is largest. Also, the time when $r/m \simeq 0.5$ is reached relatively quickly when the threshold is 1 and 0.5 ($t \approx 10\Delta t$ and $t \approx 15\Delta t$, respectively), while the same value is reached much later in the 0.1-threshold experiment ($t \approx 65\Delta t$). At the end of the experiments, when forecast uncertainty has significantly reduced, only a minority of measurement components have signal-to-noise ratio above the considered threshold. But while for thresholds 1 and 0.5 there are no or only about 4% of measurements with sufficient signal-to-noise ratio, for threshold 0.1 there are still about 36%

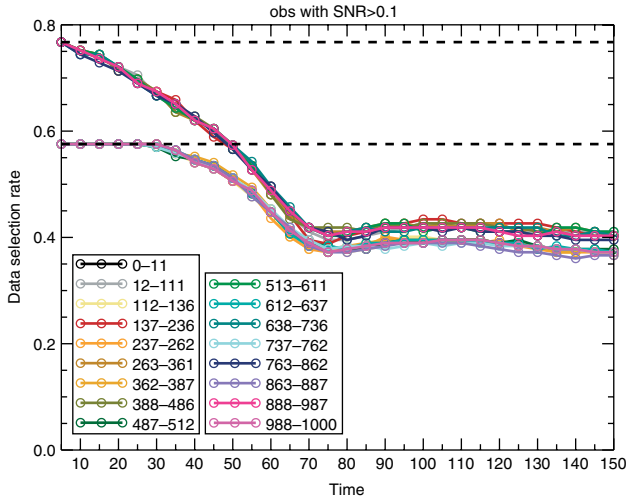

Figure 9. Number of components – relative to the total number of *in situ* observations available at each assimilation time, for localization domains (whose extents in grid-length units are provided in the key) that include 17 different observation configurations at each assimilation time – of $\mathbf{y}^{o\prime\prime}$ that have signal-to-noise ratio greater than 0.1 for a 100-member ensemble size and localization with a $200\Delta x$ radius of influence, for a $125\Delta x$ observation separation distance. The total number of observations at a given assimilation time for each localization domain is either $m_1 = 129$ (43×3) or $m_2 = 172$ (43×4). The dashed lines denote the theoretical maximum values $(K-1)/m_1$ and $(K-1)/m_2$, with $K = 100$, for the data selection rate when considering either m_1 or m_2 observations. This figure is available in colour online at wileyonlinelibrary.com/journal/qj

of measurements that can be assimilated using the data-selective filter. Also, Figure 5(b) shows the number of degrees of freedom for signal $\tilde{d}_s(r_{\tilde{\gamma}_t})$ (see Eq. (28)) as a function of time for the measurement components with signal-to-noise ratio above a given threshold $\tilde{\gamma}_t$, with $\tilde{\gamma}_t = 0.1, 0.5, 1$ while Figure 5(c) shows the ratio $\tilde{d}_s(r_{\tilde{\gamma}_t})/\tilde{d}_s$ as a function of time. From Figure 5 it follows that when a $\gamma_t = 0.1$ signal-to-noise threshold is chosen, at the end of the data assimilation experiment about 36% of the measurement components that are above the threshold are responsible for about 97% of the number of degrees of freedom for signal achieved when all observation components are assimilated. In the case, instead, when a $\gamma_t = 0.5$ signal-to-noise threshold is considered, at the end of the data assimilation experiment only about 4% of the measurement components are above the threshold and represent only 26% of the total number of degrees of freedom for signal. These results show that $\gamma_t = 0.1$ is a convenient signal-to-noise-ratio threshold to choose in this case and is the one adopted for the remaining data-selective assimilation experiments described in this paper.

In Figure 6 the evolution of the root mean square difference between the ensemble mean and the true temperature field in the case when only observations with $\tilde{\gamma}_i > 0.1$ are considered (left panel) and the evolution of the difference between the root mean square values obtained when only observations with $\tilde{\gamma}_i > 0.1, 0.5, 1$ and the root mean square values obtained when all observations are considered (right panel) are shown. The results obtained when the data selection procedure is used show that it is possible to avoid assimilating a considerable number of observational components (up to about 64% towards the end of the $\tilde{\gamma}_i > 0.1$ experiment, as discussed above) without significantly affecting the accuracy of the assimilation results.

Let us now consider the case when only 100 ensemble members are considered and a localization procedure is used that at each grid point only includes for assimilation the observations that are within a cylinder of $200\Delta x$ radius of influence. Note that each temperature analysis value at a given grid point is averaged with those within a $5\Delta x$ radius, so as to damp higher horizontal frequencies. As discussed in Ott *et al.* (2004, their section 6.4), this is found to be beneficial for relatively small ensemble sizes. Figure 7 shows the RMSE results when a data-selective square root filter is used with the localization procedure discussed above and when only observation components with $\tilde{\gamma}_i > 0.1$ are assimilated using a 100-member ensemble size. By comparing Figure 7 (left panel) with Figure 6 (left panel) it is evident that at the end of the experiments a lower accuracy is achieved when using localization and a smaller ensemble size, i.e. when less observational information is available for assimilation. However, from Figures 7 (right panel) and 8 it follows that the differences between the case where data selection is either used or not used are relatively very small, also in the case when localization and a 100-member ensemble size are used. This is despite the considerable amount of observational components discarded for each local domain, as can be seen from Figure 9, which shows that at the end of the experiment only about 40% of the observation components provide sufficient information for the chosen threshold, a data selection rate consistent with that achieved at the end of the 300-member ensemble experiment with no localization for the same data selection threshold. These findings confirm that the data selection procedure is advantageous also when used in conjunction with localization procedures.

It is also interesting to investigate the performance of the data selection procedure with smaller ensemble sizes. Three experiments using 50, 25 and 15 forecast ensemble members with the same true temperature field trajectory as that shown in Figure 2 were carried out with a 0.1 signal-to-noise data selection threshold. Given the relatively small ensemble sizes, for these experiments a $50\Delta x$ localization radius was used and 16 evenly spaced observation profiles every $5\Delta t$ were considered for assimilation. By comparing Figure 10(a) with Figures 6 and 7 it follows that the RMSE values achieved just after the time needed for the temperature field at a given observation location to be advected to the next observation location (i.e. for $t > (L/16)/u = 62.5\Delta t$) when using 50, 25 or 15 ensemble members are between the RMSE values achieved at $t > (L/8)/u = 125\Delta t$ when using a 300-member ensemble with no localization and those when a 100-member ensemble with a $200\Delta x$ localization radius is considered, in the case when only eight observations are considered for assimilation. Also, by comparing Figure 10(a) with Figure 10(b) it follows that the larger the ensemble size the more observation components with sufficient signal-to-noise ratio and the lower the RMSE values at the end of the experiment, as expected. This means that reasonably accurate assimilation results can be attained even when using a relatively small ensemble size and with correspondingly high observation rejection rates, provided that the size of the local domain is not significantly larger than the ensemble size and that a sufficient number of observations with enough signal-to-noise ratio are available in each local domain. These results show that the use of the data selection procedure can be beneficial also when limitations in computational resources only allow the use of a relatively small ensemble size.

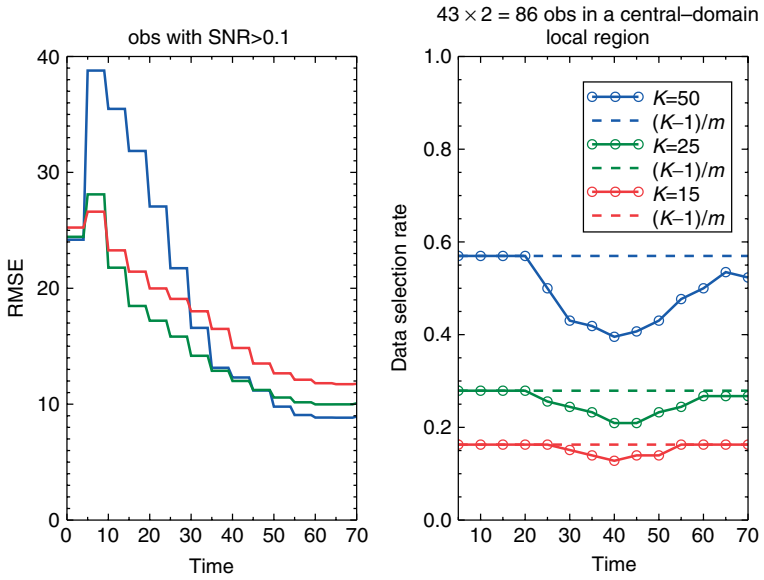

Figure 10. Root mean square error (RMSE) calculated as the difference between the ensemble mean, in the case when only observations with $\tilde{\gamma}_i > 0.1$ are considered, and the true temperature field for a 50-member (blue solid line), 25-member (green solid line) and 15-member (red solid line) ensemble size (left); number of components r – relative to the total number $m = 86$ *in situ* observations – of $\mathbf{y}^{o\prime\prime}$ that have signal-to-noise ratio $\tilde{\gamma}_i > 0.1$ in the case when $K = 50$ (blue solid line), $K = 25$ (green solid line) and $K = 15$ (red solid line) ensemble members are considered. For clarity, only data selection rate results for localization regions that include two observation profiles located in the centre of the model's domain are shown. For reference, the highest achievable data selection ratio $\tilde{\gamma}_i = (K-1)/m$ using a 50-member (blue dashed line), 25-member (green dashed line) and 15-member (red dashed line) ensemble size are also shown (right). Note that the results in these experiments were obtained by considering 16 evenly spaced observation profiles for assimilation every $5\Delta t$ using a $50\Delta x$ localization radius of influence. This figure is available in colour online at wileyonlinelibrary.com/journal/qj

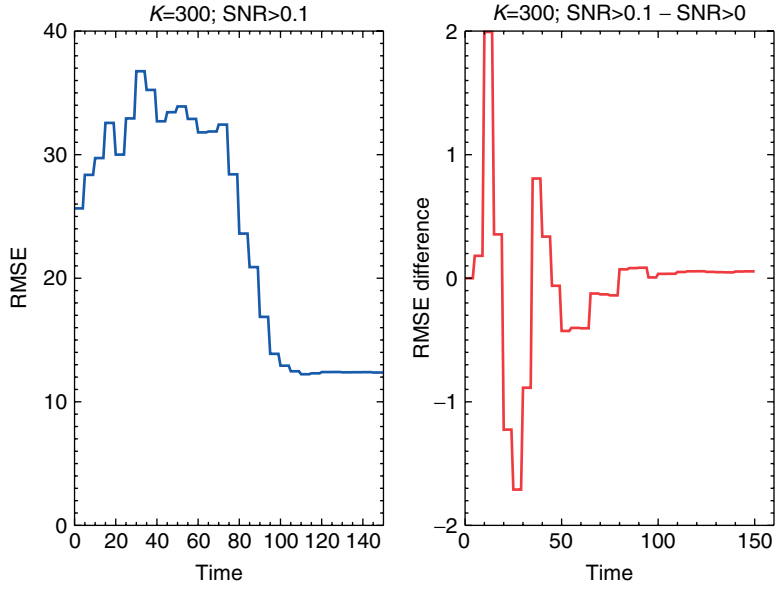

Figure 11. Root mean square error (RMSE) calculated as the difference between the ensemble mean and the true temperature field for a 300-member ensemble size and no localization, in the case when only remotely sensed observations with $\tilde{\gamma}_i > 0.1$ are considered (left); evolution of the difference between the RMSE values obtained when considering only observations with $\tilde{\gamma}_i > 0.1$ and the RMSE values obtained when all observations are considered (right). This figure is available in colour online at wileyonlinelibrary.com/journal/qj

We now want to compare the results obtained from data assimilation experiments performed using either remote sounding or *in situ* observations. Figure 11(left panel) shows the RMSE values obtained when radiances from 66 IASI channels, measured over eight uniformly spaced longitudes and assimilated every $5\Delta t$, for a 300-member ensemble size and when only observations with $\tilde{\gamma}_i > 0.1$ are considered. Note that in this case the RMSE values at the end of the experiment are larger than those obtained when *in situ* data are assimilated (see Figure 6, left panel), despite the number of satellite observations considered for assimilation being larger than that of *in situ* data. From Figure 12(c) it is evident that this can be explained by the number of degrees of freedom for signal of remote sounding data being significantly smaller than that of *in situ* data during the first part of the assimilation experiment. As shown in Figure 11(right panel), however, from our results it follows that, even with remote sounding data the data selection procedure does not affect significantly the accuracy of the assimilation. This is despite the considerable amount of data that has been discarded during assimilation (see Figure 12(a) and (b)), corresponding to a satellite data rejection rate of

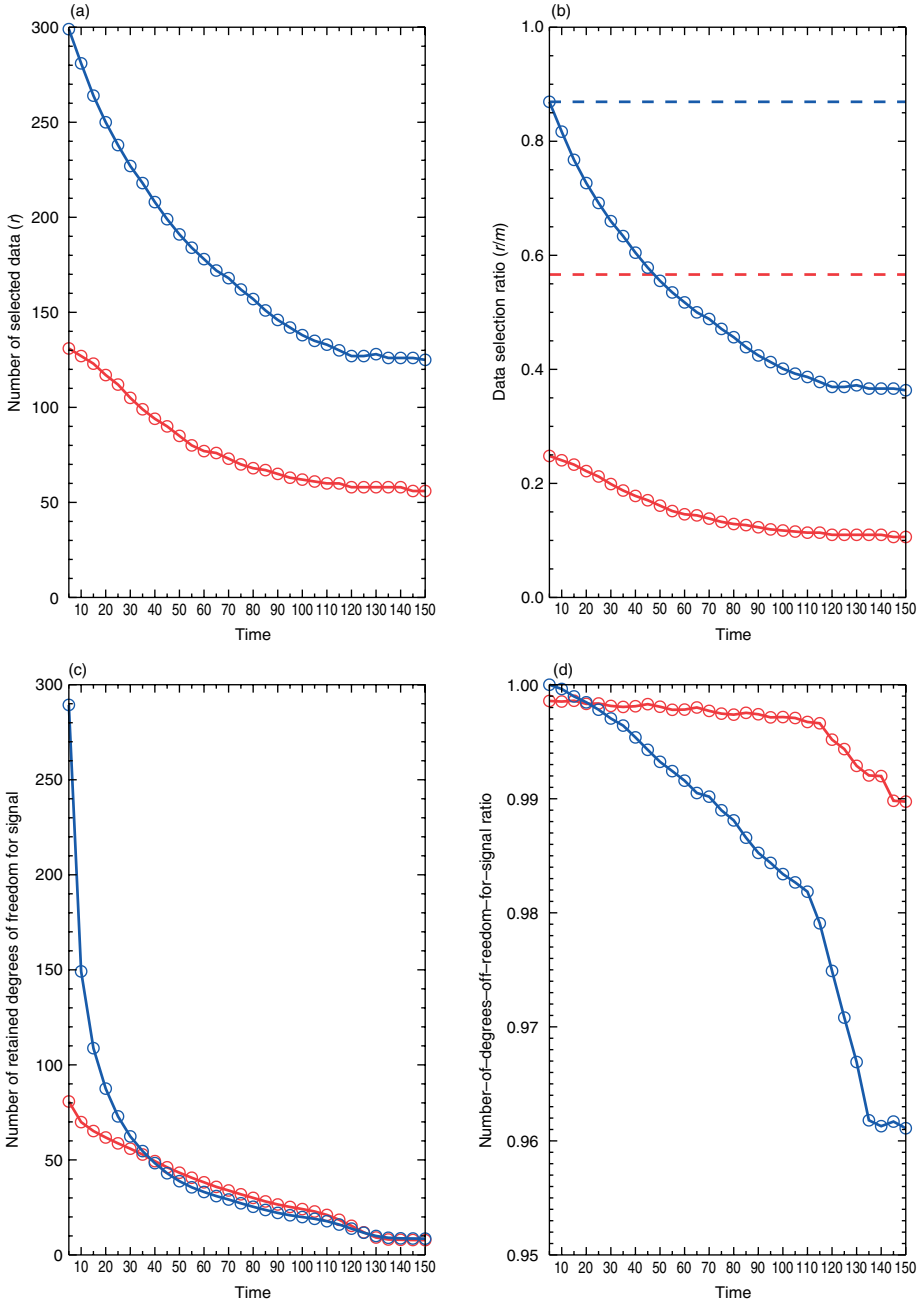

Figure 12. (a) Number of components r of $\mathbf{y}^{o\prime\prime}$ that have signal-to-noise ratio $\tilde{\gamma}_i > 0.1$ for *in situ* (blue solid line) and for remotely sensed data (red solid line) for a 300-member ensemble size with no localization; (b) as in (a), with quantities scaled either by the total number $m = 344$ (43×8) of *in situ* observations (blue solid line) or by the total number $m = 528$ (66×8) of remotely sensed observations at each assimilation time (red solid line); (c) number of degrees of freedom for signal $\tilde{d}_s$ considering either *in situ* observation components (blue solid line) or remotely sensed observations (red solid line) with $\tilde{\gamma}_i > 0.1$; (d) as in (c), with quantities scaled by the number of degrees of freedom for signal when all observation components are considered. This figure is available in colour online at wileyonlinelibrary.com/journal/qj

about 88% at the end of the assimilation experiment. Also note that Figure 12(d) shows that the number of degrees of freedom for signal that are discarded by the data selection procedure is less than about 1% of the total number of available satellite data, implying that 0.1 is here a convenient threshold also when remote sounding data are considered. Similar results are obtained when a 100-member ensemble size with a $200\Delta x$ radius of influence is considered, both in terms of RMSE values and differences (not shown) and in terms of data selection rates for localization domains including different observation locations, as can be seen by comparing Figure 13(a) to Figure 12(b) (bottom solid line).

Let us now investigate the performance of the data selection procedure when using the EnKF, discussed in section 5.2, rather than the square root filter introduced in section 5.1 and used for all experiments discussed so far. Figure 14 shows the RMSE results (left panel) obtained when a 300-member ensemble data assimilation experiment is performed and only observation components with signal-to-noise ratio greater than 0.1 are assimilated using the data-selective EnKF. The right panel of Figure 14 shows instead the RMS differences between the data-selective EnKF experiment with a 0.1 signal-to-noise threshold and the EnKF experiment using all observation components. By comparing Figures 6 and 14 it follows that the magnitude

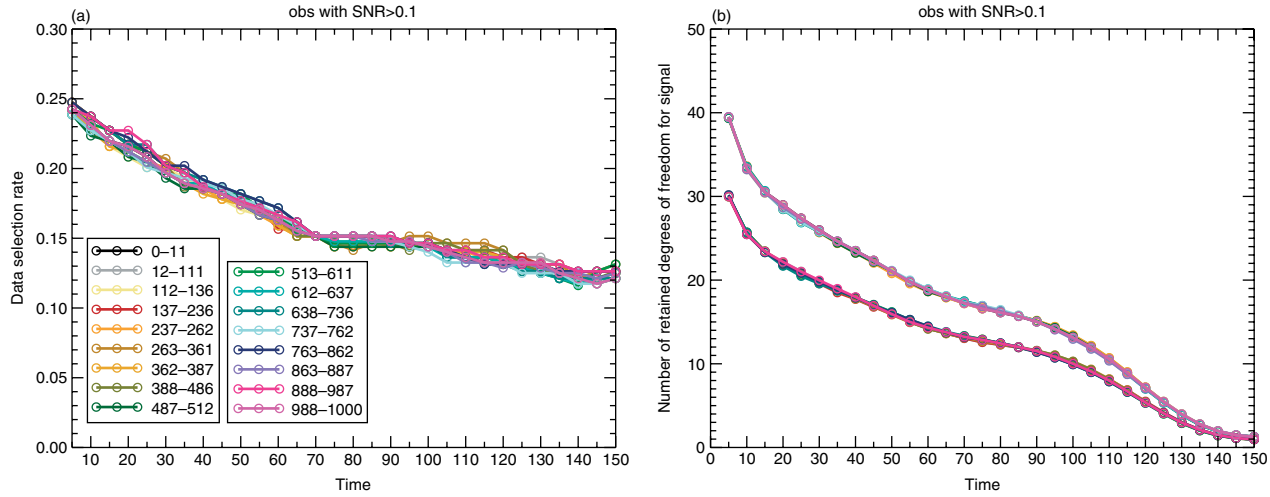

Figure 13. (a) Number of components – relative to the total number of remote sounding measurements available at each assimilation time, for localization domains (whose extents in grid-length units are provided in the key) that include 17 different observation configurations at each assimilation time – of $\mathbf{y}^{o\prime\prime}$ that have signal-to-noise ratio greater than 0.1 for a 100-member ensemble size and localization with a $200\Delta x$ radius of influence, for a $125\Delta x$ observation separation distance. The total number of observations at a given assimilation time for each localization domain is either $m_1 = 198$ (3×66) or $m_2 = 264$ (4×66). Note that the values of the theoretical maximum values $(K-1)/m_1$ and $(K-1)/m_2$, with $K = 100$, for the data selection rate when considering either m_1 or m_2 observations are 0.5 and 0.375, respectively. (b) Number of degrees of freedom for signal when only remotely sensed observations with $\tilde{\gamma}_i > 0.1$ are considered over localization domains with different observation configurations. This figure is available in colour online at wileyonlinelibrary.com/journal/qj

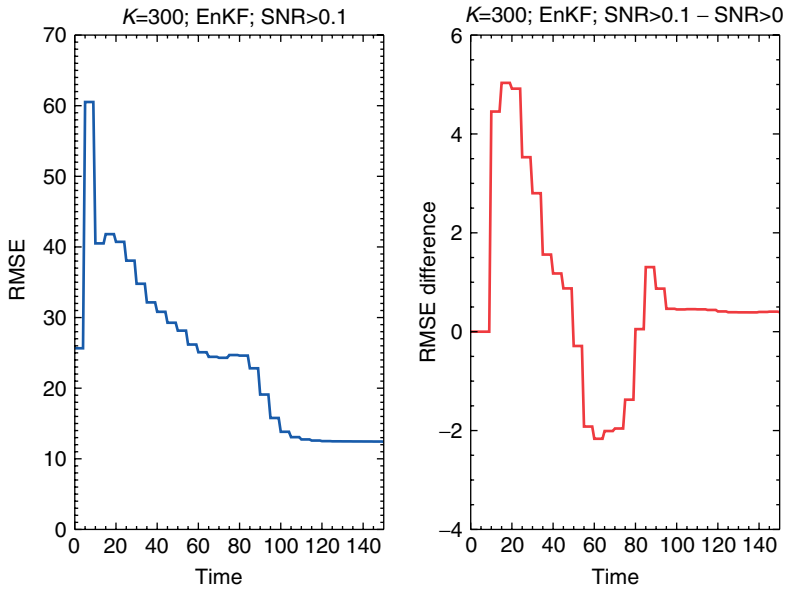

Figure 14. Root mean square error (RMSE) calculated as the difference between the ensemble mean estimated using a data-selective EnKF and the true temperature field for a 300-member ensemble size and no localization, in the case when only observations with $\tilde{\gamma}_i > 0.1$ are assimilated every $5\Delta t$ (left); evolution of the difference between the RMSE values obtained when considering only observations with $\tilde{\gamma}_i > 0.1$ using the data-selective EnKF and the RMSE values obtained when all observations are assimilated with the EnKF (right). This figure is available in colour online at wileyonlinelibrary.com/journal/qj

of the RMS differences achieved when the data selection procedure is used with a stochastic filter with the same initial ensemble, reference trajectory and observations as those used in the deterministic (e.g. a square root filter) experiment is still a small proportion (about 3% towards the end of the experiment) of the respective RMSE value. Note, however, that the data selection procedure used with the square root filter achieves a substantially lower RMSE than when it is used with the EnKF. This is consistent with the higher data rejection ratio experienced in the EnKF experiment for a 0.1 signal-to-noise threshold, as it can be seen from Figures 5(a) (top solid line) and 15. This result seems to indicate that the same set of observations provides more information when assimilated with a deterministic rather than a stochastic scheme. The robustness of this result, however, will need to be investigated further.

Finally, let us consider the case when the evolution of the state is determined by the nonlinear advection equation (43), with advection speed given by integrating the thermal wind equation (49) while assuming $u = 0$ at the lowermost model level. Differently from the linear advection case, the advection speed now depends on the state and benefits are expected from longer assimilation experiments. Figure 16 (left panel) shows the RMSE results up to $t = 1000\Delta t$ for a 300-member ensemble experiment when only observation components with signal-to-noise ratio greater than 0.1 in the

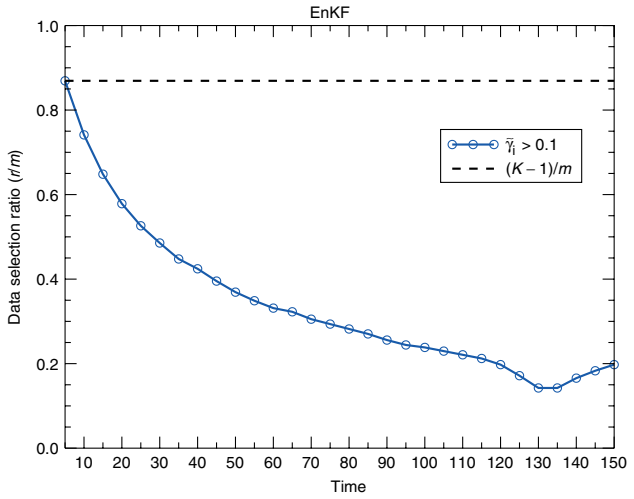

Figure 15. Number of components r – relative to the total number $m = 344$ *in situ* observations – of $\mathbf{y}^{o\prime\prime}$ that have signal-to-noise ratio $\tilde{\gamma}_i > 0.1$ (blue solid line) for a 300-member ensemble size with no localization and when the data-selective EnKF is used. For reference, the highest achievable data selection ratio $\tilde{\gamma}_i = (K-1)/m$ (black dashed line) is also shown. This figure is available in colour online at wileyonlinelibrary.com/journal/qj

usual configuration are assimilated using the data-selective square-root filter. As expected, the RMSE continues to decrease for much longer than in the linear advection case, as can be seen by comparing the results in Figures 6 and 16. The temperature values at the beginning, in the middle and at the end of the experiment for the true field are shown in Figure 17, which clearly shows the presence of zonal wind vertical gradient and of the midlatitude tropospheric jet stream resulting from thermal wind balance. Figures 18 and 16 (right panel) show, for the nonlinear advection case, the difference and the RMSE difference, respectively, between the ensemble mean temperature when using the data-selective square root filter with a 0.1 signal-to-noise threshold and the ensemble mean temperature resulting when using a square root filter with all observation components. Again, the accuracy of the data-selective square root filter is comparable to that achieved when all observations are assimilated, even in the nonlinear advection case.

Results with the data-selective EnKF for the nonlinear advection case are shown in Figure 19. Figures 16 and 19 indicate that the RMSE results of the data selective method with the nonlinear advection model are consistent with those obtained with the linear advection model, when using both deterministic and stochastic filters. Note in particular that, in agreement with the linear advection case, the lower accuracy in the nonlinear advection case of the data-selective EnKF compared to that of the data-selective square-root filter can be explained by the higher data rejection rate of the stochastic scheme, as can be seen from Figure 20.

8. Conclusions

In this paper, a procedure for assimilating only the components of the observation vector that are able to reduce the estimation uncertainty resulting from the use of an ensemble filtering technique is discussed. Our results show that most observation components end up being discarded after an initial transient (or 'spin-up') period without affecting significantly the assimilation results. This data selection procedure can be used with remote sounding measurements, achieving a rejection rate after the spin-up period of about 88% in one of the experiments performed in this work. Note, however, that the number of observation components that are discarded depends on the chosen rejection threshold and it is generally higher when inflation procedures are not used, as in the case of the experiments described in this paper. It is also important

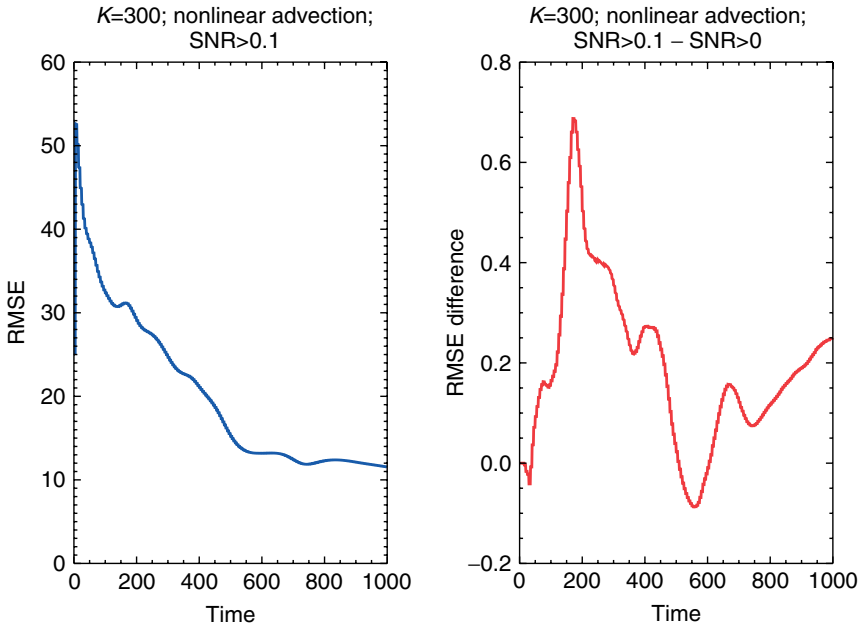

Figure 16. Root mean square error (RMSE) up to $1000\Delta t$, calculated as the difference between the ensemble mean and the true temperature field, for the nonlinear advection case and a 300-member ensemble size with no localization, when all observations are assimilated with a square root filter (left, blue solid line); evolution of the difference between the RMSE values obtained when assimilating only observations with $\tilde{\gamma}_i > 0.1$ using a data-selective square root filter and the RMSE values obtained when all observations are assimilated (right, red solid line). This figure is available in colour online at wileyonlinelibrary.com/journal/qj

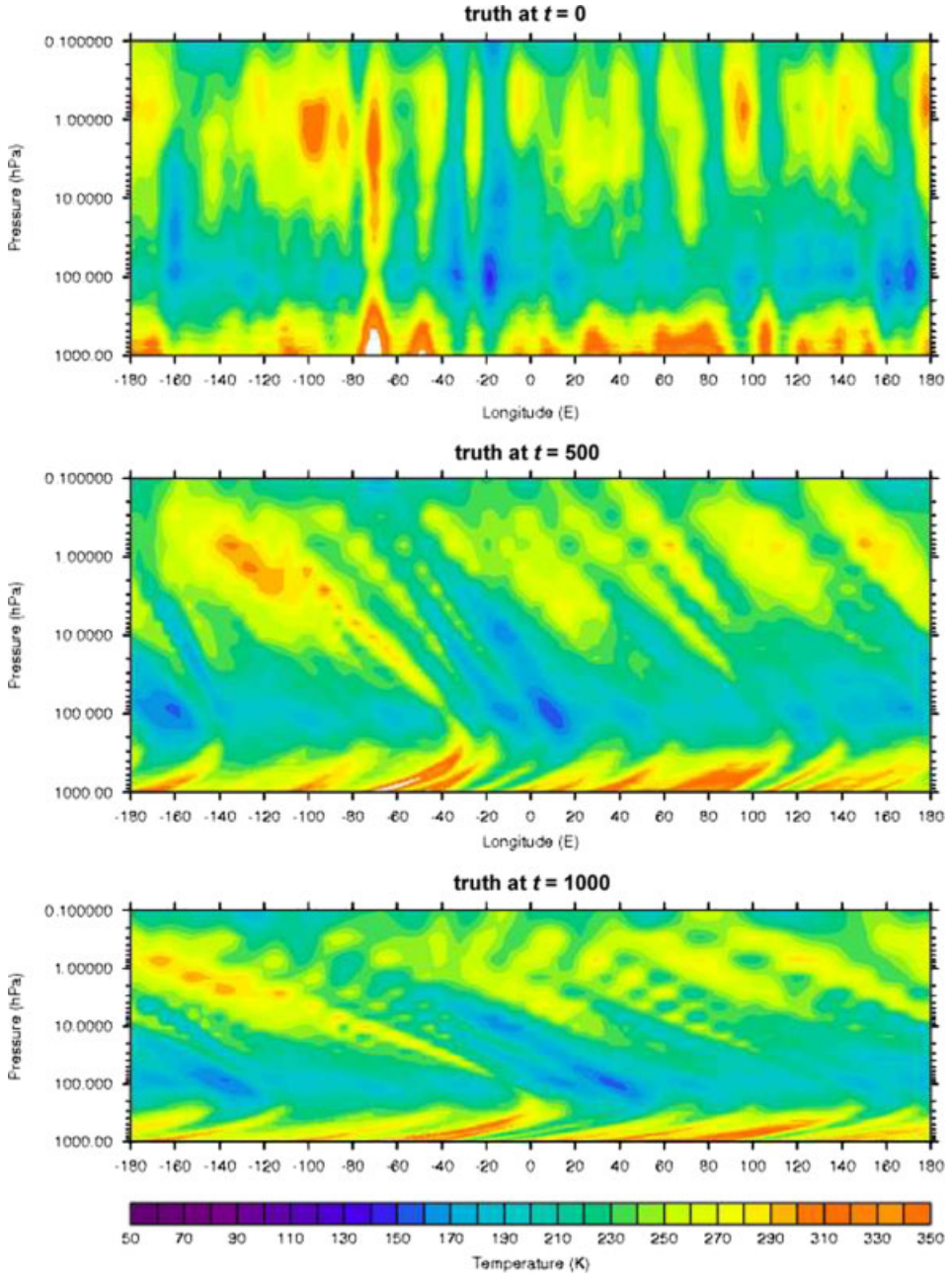

Figure 17. True temperature field at $t = 0$ (top), $t = 500\Delta t$ (middle) and $t = 1000\Delta t$ (bottom) for the nonlinear advection case. This figure is available in colour online at wileyonlinelibrary.com/journal/qj

to note that the data-selection method is compatible with existing localization and inflation procedures commonly used to ameliorate one of the fundamental and well-known shortcomings of ensemble data assimilation – which constrains its resulting analysis increments to span only a limited portion of the state space – in the case of practical importance when the size of the forecast ensemble is much smaller than the dimension of the state space. In fact, this method allows the use of a larger localization domain, which may lead to a more balanced analysis (e.g. Greybush *et al.*, 2011) even in the presence of observations with high spatial density. Note, however, that the data selection method described in this paper does not replace but should be used in addition to data-thinning and quality control procedures that are commonly used to reduce the amount of observations that are assimilated in NWP models (e.g. Bauer *et al.*, 2011). In fact, spatial data thinning avoids the need to account for error correlations between neighbouring observations (e.g. Dando *et al.*, 2007), while data quality control (e.g., Andersson and Järvinen, 1999) makes sure that only those observations that are predicted by the model with sufficient accuracy are assimilated.

Another benefit of the data selection strategy presented in this paper is that it may considerably reduce the amount of numerical computations that are required to assimilate observations in an Earth system model. This is particularly the case when a significant number of observations have a relatively low information content so that the required transformations to the observation vector and observation operator, discussed in section 5, represent a worthwhile investment. In this case, only $r \ll m$ observational components need to be considered when computing the analysis ensemble mean using Eq. (36) or the analysis ensemble using Eq. (42). Also, in this way, round-off

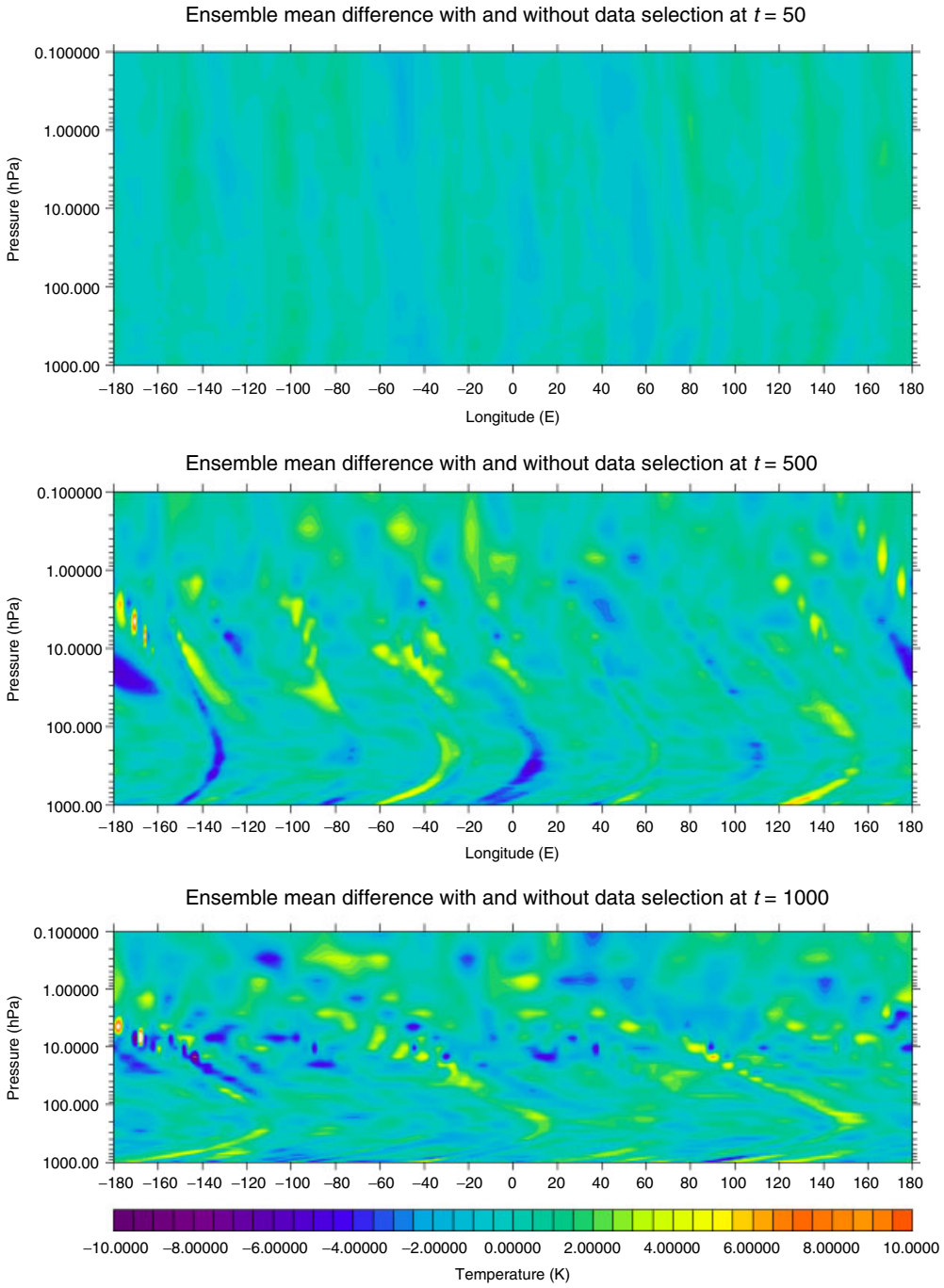

Figure 18. Nonlinear advection case: difference between the ensemble mean temperature field at $t = 50$ (top), $t = 500\Delta t$ (middle) and $t = 1000\Delta t$ (bottom) when only observations with signal-to-noise ratio $\tilde{\gamma}_i > 0.1$ are assimilated every $5\Delta t$ using the data-selective square root filter and the ensemble mean temperature field when all 344 in situ observations are assimilated every $5\Delta t$. The experiment made use of a 300-member forecast ensemble with no localization. This figure is available in colour online at wileyonlinelibrary.com/journal/qj

errors arising from the representation of a given real number as a floating point number (e.g. Golub and van Loan 1996, section 2.4), are minimized.

Finally, use of the data selection technique reduces the risk of contaminating the analysis estimate with observational bias, as discussed in the following. Consider the case of an observation that is deemed of sufficient quality to pass existing screening procedures but that is actually affected by some unknown bias. It is possible to show (e.g. Dee, 2005, his section 2) that the analysis error also becomes biased as a result of the assimilation of a biased observation. Let us now assume that this observation provides negligible information but that the observation bias is sufficiently large in magnitude to generate a significant bias in the analysis. In this case, the assimilation of this observation produces a detrimental effect on the analysis. However, the use of the data selection procedure discussed in this paper provides an effective way to avoid this shortcoming. The use of this procedure is likely to be most advantageous when satellite observations are considered for assimilation, as they currently make up 95% of all observations that are assimilated at operational meteorological centres (Bauer *et al.*, 2011) and are a major source of observational bias (e.g. Auligné *et al.*, 2007; Dee and Uppala, 2009), while having, according to our findings, the most important data reduction potential.

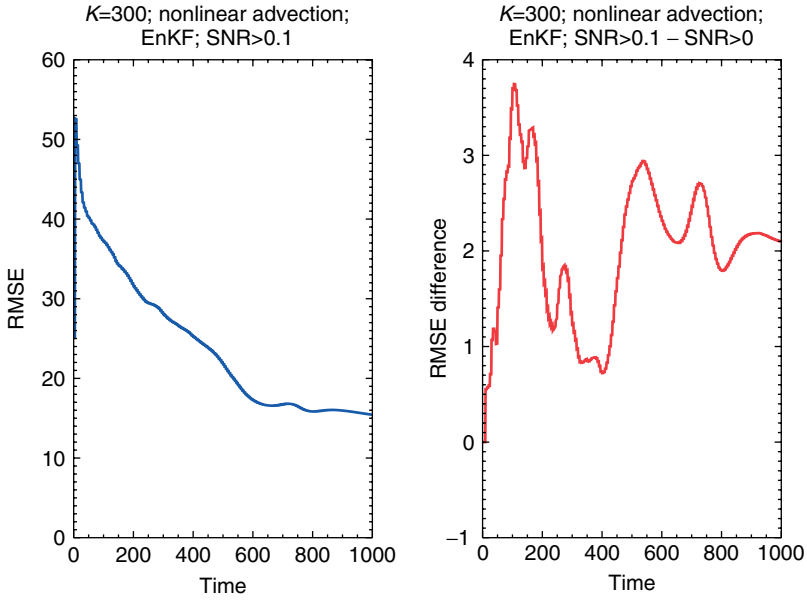

Figure 19. As Figure 16 with observations assimilated using a data-selective EnKF. This figure is available in colour online at wileyonlinelibrary.com/journal/qj

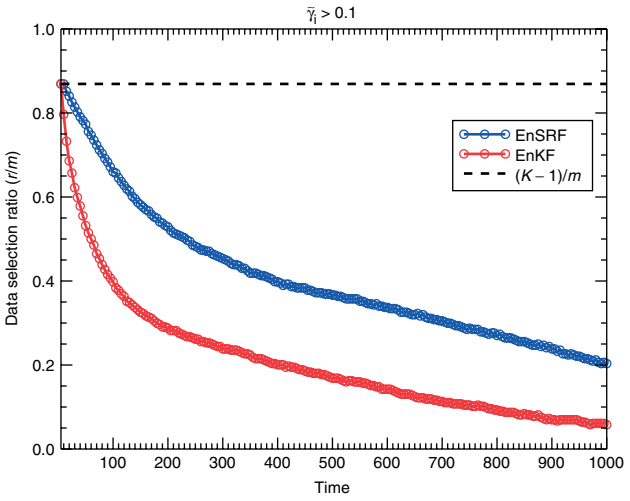

Figure 20. Number of components r – relative to the total number $m = 344$ of *in situ* observations available at each assimilation time – of $\mathbf{y}^{o\prime\prime}$ that have signal-to-noise ratio $\tilde{\gamma}_i > 0.1$ when either a data-selective square root filter (blue solid line) or a data-selective ensemble Kalman filter (red solid line) for a 300-member ensemble size with no localization in the nonlinear advection case. For reference, the highest achievable data selection ratio $\tilde{\gamma}_i = (K-1)/m$ (black dashed line) is also shown. This figure is available in colour online at wileyonlinelibrary.com/journal/qj

Acknowledgements

I would like to thank Ross Bannister and the anonymous reviewers for their useful comments that helped to improve the paper. The author is supported by the NERC National Centre for Earth Observation.

References

Anderson J. 2007a. An adaptive covariance inflation error correction algorithm for ensemble filters. *Tellus A* **59**: 210–224.
Anderson J. 2007b. Exploring the need for localization in ensemble data assimilation using a hierarchical ensemble filter. *Physica D* **230**: 99–111.
Anderson J. 2009. Spatially and temporally varying adaptive covariance inflation for ensemble filters. *Tellus A* **61**: 72–83.
Anderson J, Anderson S. 1999. A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. *Mon. Weather Rev.* **127**: 2741–2758.
Andersson E, Järvinen H. 1999. Variational quality control. *Q. J. R. Meteorol. Soc.* **125**: 697–722.
Auligné T, McNally A, Dee D. 2007. Adaptive bias correction for satellite data in a numerical weather prediction system. *Q. J. R. Meteorol. Soc.* **133**: 631–642.
Bannister R. 2008. A review of forecast error covariance statistics in atmospheric variational data assimilation. 1. Characteristics and measurements of forecast error covariances. *Q. J. R. Meteorol. Soc.* **134**: 1951–1970.
Bauer P, Buizza R, Cardinali C, Noël Thepaut J. 2011. Impact of singular-vector-based satellite data thinning on NWP. *Q. J. R. Meteorol. Soc.* **137**: 286–302.
Bishop C, Hodyss D. 2009a. Ensemble covariances adaptively localized with eco-rap. Part 1: tests on simple error models. *Tellus A* **61**: 84–96.
Bishop C, Hodyss D. 2009b. Ensemble covariances adaptively localized with eco-rap. Part 2: a strategy for the atmosphere. *Tellus A* **61**: 97–111.
Bishop C, Etherton B, Majumdar S. 2001. Adaptive sampling with the ensemble transform kalman filter. Part I: Theoretical aspects. *Mon. Weather Rev.* **129**: 420–436.
Cohn S. 1997. An introduction to estimation theory. *J. Meteorol. Soc. Jpn.* **75**: 257–288.
Cohn S, da Silva A, Guo J, Sienkiewicz M, Lamich D. 1998. Assessing the effects of data selection with the DAO physical-space statistical analysis system. *Mon. Weather Rev.* **126**: 2913–2926.
Collard A. 2007. Selection of IASI channels for use in numerical weather prediction. *Q. J. R. Meteorol. Soc.* **133**: 1977–1991.
Dando M, Thorpe A, Eyre JR. 2007. The optimal density of atmospheric sounder observations in the met office NWP system. *Q. J. R. Meteorol. Soc.* **133**: 1933–1943.
Dee D. 2005. Bias and data assimilation. *Q. J. R. Meteorol. Soc.* **131**: 3323–3343.
Dee D, Uppala S. 2009. Variational bias correction of satellite radiance data in the era-interim reanalysis. *Q. J. R. Meteorol. Soc.* **135**: 1830–1841.
Evensen G. 2003. The ensemble kalman filter: theoretical formulation and practical implementation. *Ocean Dynam.* **53**: 343–367.
Evensen G. 2004. Sampling strategies and square root analysis schemes for the enkf. *Ocean Dynam.* **54**: 539–560.
Evensen G. 2009. *Data Assimilation: The Ensemble Kalman Filter* (2nd edn). Springer: Berlin.
Gill A. 1982. *Atmosphere– Ocean Dynamics*, Vol. 30. Academic Press: New York.
Golub G, van Loan C. 1996. *Matrix Computations* (3rd edn). Johns Hopkins University Press: Baltimore, MD.

Greybush S, Kalnay E, Miyoshi T, Ide K, Hunt B. 2011. Balance and ensemble Kalman filter localization techniques. *Mon. Weather Rev.* **139**: 511–522.
Hamill T, Whitaker J, Snyder C. 2001. Distance-dependent filtering of background error covariance estimates in an ensemble kalman filter. *Mon. Weather Rev.* **129**: 2776–2790.
Houtekamer P, Mitchell H. 1998. Data assimilation using an ensemble kalman filter technique. *Mon. Weather Rev.* **126**: 796–811.
Houtekamer P, Mitchell H. 2001. A sequential ensemble Kalman filter for atmospheric data assimilation. *Mon. Weather Rev.* **129**: 123–137.
Houtekamer P, Mitchell H, Deng X. 2009. Model error representation in an operational ensemble kalman filter. *Mon. Weather Rev.* **137**: 2126–2143.
Hunt B, Kostelich E, Szunyogh I. 2007. Efficient data assimilation for spatiotemporal chaos: a local ensemble transform kalman filter. *Physica D* **230**: 112–126.
Janjić T, Nerger L, Albertella A, Schröter J, Skachko S. 2011. On domain localization in ensemble-based kalman filter algorithms. *Mon. Weather Rev.* **139**: 2046–2060.
Kepert J. 2009. Covariance localisation and balance in an ensemble kalman filter. *Q. J. R. Meteorol. Soc.* **135**: 1157–1176.
Li H, Kalnay E, Miyoshi T. 2009. Simultaneous estimation of covariance inflation and observation errors within an ensemble Kalman filter. *Q. J. R. Meteorol. Soc.* **135**: 523–533.
Livings D, Dance S, Nichols N. 2008. Unbiased ensemble square root filters. *Physica D* **237**: 1021–1028.
Lorenc A. 2003. The potential of the ensemble Kalman filter for NWPA comparison with 4D-Var. *Q. J. R. Meteorol. Soc.* **129**: 3183–3203.
Maybeck P. 1982. *Stochastic Models, Estimation and Control*, Vol. 1. Academic press: New York.
Migliorini S. 2012. On the equivalence between radiance and retrieval assimilation. *Mon. Weather Rev.* **140**: 258–265.
Mitchell H, Houtekamer P, Pellerin G. 2002. Ensemble size, balance, and model-error representation in an ensemble Kalman filter. *Mon. Weather Rev.* **130**: 2791–2811.
Miyoshi T. 2011. The Gaussian approach to adaptive covariance inflation and its implementation with the local ensemble transform Kalman filter. *Mon. Weather Rev.* **139**: 1519–1535.
Ott E, Hunt B, Szunyogh I, Zimin A, Kostelich E, Corazza M, Kalnay E, Patil D, Yorke J. 2004. A local ensemble Kalman filter for atmospheric data assimilation. *Tellus A* **56**: 415–428.
Pielke R. 2002. *Mesoscale Meteorological Modeling* (2nd edn). Academic Press: New York.
Rodgers C. 2000. *Inverse Methods for Atmospheric Sounding: Theory and Practice*. World Scientific: Singapore.
Sacher W, Bartello P. 2008. Sampling errors in ensemble Kalman filtering. Part I: theory. *Mon. Weather Rev.* **136**: 3035–3049.
Sakov P, Oke P. 2008. Implications of the form of the ensemble transformation in the ensemble square root filters. *Mon. Weather Rev.* **136**: 1042–1053.
Saunders R, Brunel P. 2005. Rttov v8.7 users guide. *NWP SAF report*: 45.
Siméoni D, Singer C, Chalon G. 1997. Infrared atmospheric sounding interferometer. *Acta Astronaut.* **40**: 113–118.
Stull R, Ahrens C. 2000. *Meteorology for Scientists and Engineers*. Brooks/Cole: Pacific Grove, CA.
van Leeuwen P. 1999. Comment on 'Data assimilation using an ensemble Kalman filter technique'. *Mon. Weather Rev.* **127**: 1374–1377.
Wang X, Bishop C. 2003. A comparison of breeding and ensemble transform Kalman filter ensemble forecast schemes. *J. Atmos. Sci.* **60**: 1140–1158.
Whitaker J, Hamill T. 2002. Ensemble data assimilation without perturbed observations. *Mon. Weather Rev.* **130**: 1913–1925.
Whitaker J, Hamill T, Wei X, Song Y, Toth Z. 2008. Ensemble data assimilation with the NCEP global forecast system. *Mon. Weather Rev.* **136**: 463–482.
Zupanski D, Denning A, Uliasz M, Zupanski M, Schuh A, Rayner P, Peters W, Corbin K. 2007a. Carbon flux bias estimation employing maximum likelihood ensemble filter (MLEF). *J. Geophys. Res.* **112**: D17107, DOI: 10.1029/2006JD008371.
Zupanski D, Hou A, Zhang S, Zupanski M, Kummerow C, Cheung S. 2007b. Applications of information theory in ensemble data assimilation. *Q. J. R. Meteorol. Soc.* **133**: 1533–1545.

