On diagnosing observation error statistics with local ensemble data assimilation

Quarterly Journal of the Royal Meteorological Society Q. J. R. Meteorol. Soc. 143: 2677–2686, October 2017 A DOI:10.1002/qj.3117

RMetS
Royal Meteorological Society

On diagnosing observation-error statistics with local ensemble data assimilation

J. A. Waller,[a*] S. L. Dance[a,b] and N. K. Nichols[a,b]
[a] *Department of Meteorology, University of Reading, UK*
[b] *Department of Mathematics and Statistics, University of Reading, UK*

*Correspondence to: J. A. Waller, Department of Meteorology, University of Reading, Earley Gate, PO Box 243, Reading, RG6 6BB, UK. E-mail: j.a.waller@reading.ac.uk

Recent research has shown that the use of correlated observation errors in data assimilation can lead to improvements in analysis accuracy and forecast skill. As a result, there is increased interest in characterizing, understanding and making better use of correlated observation errors. A simple diagnostic for estimating observation-error statistics makes use of statistical averages of observation-minus-background and observation-minus-analysis residuals. This diagnostic is derived assuming that the analysis is calculated using a best linear unbiased estimator. In this work, we consider whether the diagnostic is still applicable when the analysis is calculated using ensemble assimilation schemes with domain localization. We show that the diagnostic equations no longer hold: the statistical averages of observation-minus-background and observation-minus-analysis residuals no longer result in an estimate of the observation-error covariance matrix. Nevertheless, we are able to show that, under certain circumstances, some elements of the observation-error covariance matrix can be recovered. Furthermore, we provide a method to determine which elements of the observation-error covariance matrix can be estimated correctly. In particular, the correct estimation of correlations is dependent on both the localization radius and the observation operator. We provide numerical examples that illustrate these mathematical results.

Key Words: data assimilation diagnostic; correlated observation errors; local ensemble assimilation; domain localization; residuals; covariance matrix

Received 28 November 2016; Revised 23 June 2017; Accepted 04 July 2017; Published online in Wiley Online Library 20 September 2017

1. Introduction

A key component in numerical weather prediction (NWP) is the use of data assimilation systems. Data assimilation techniques combine model states, known as forecasts or backgrounds, with observations, weighted by their respective error statistics, to provide a best estimate of the state, known as the analysis. To obtain an accurate analysis, it is essential to have an accurate representation of the background- and observation-error statistics.

Much attention has been devoted to the estimation and representation of the background-error covariance matrix in variational assimilation systems (e.g. Bannister, 2008). In addition to this, an entire class of ensemble data assimilation schemes has been developed with the specific aim that the assimilation system itself should provide an estimate of the flow-dependent background-error statistics. First introduced by Evensen (1994), the ensemble Kalman filter estimates the background-error statistics considering a statistical sample, or ensemble, of background states during the assimilation-forecast cycle. Many forms of ensemble Kalman filter have been developed: for example Burgers *et al.* (1998), Houtekamer and Mitchell (1998), Anderson (2001), Evensen (2003) and Tippett *et al.* (2003). These methods can be split into two categories; deterministic filters and stochastic filters. Stochastic filters make use of a set of perturbed observations which are required to maintain the correct statistics of the filter (Burgers *et al.*, 1998; Lewis *et al.*, 2006). Deterministic filters do not require these perturbed observations and therefore no additional errors are introduced in the observations. It is these deterministic filters that we will focus on in this article. A key limitation of ensemble filters is the prohibitively large number of ensemble members required to obtain an accurate representation of the background-error statistics (Whitaker and Hamill, 2002). Therefore, additional constraints are required for ensemble methods to prove effective when used to provide operational weather forecasts.

One approach is to 'localize' the problem by considering only a part of the state or observation space, therefore reducing the necessary ensemble size. The two most common localization methods are state-space (covariance) localization

(Hamill *et al.*, 2001; Houtekamer and Mitchell, 2001; Petrie and Dance, 2010) and domain localization (local analysis: Houtekamer and Mitchell, 1998; Ott *et al.*, 2004; Janjić *et al.*, 2011; Nerger *et al.*, 2012), which is used in conjunction with observation localization (Sakov and Bertino, 2011). In covariance localization, the estimated background-error covariance matrix is Schur-multiplied by a localization matrix to suppress spurious background correlations that appear at long range due to sampling error. In domain localization, each model grid point is updated individually using a subset of observations within a given distance. In this article, we will consider the impact of domain localization on the estimation of observation-error statistics.

The quantification of observation errors has been a recent area of research. Typically, observation errors have been assumed uncorrelated and, in an attempt to satisfy this assumption, the data is often thinned or 'superobbed' (Lorenc, 1981). However, it is known that a number of different sources contribute to the observation error, some of which may be correlated, state-dependent and dependent on the model resolution (Lorenc, 1986; Janjić and Cohn, 2006; Waller, 2013; Waller *et al.*, 2014a, 2014b; Hodyss and Nichols, 2015). Research has shown that observation errors can indeed exhibit significant correlations. Furthermore, the inclusion of correlated interchannel errors for satellite observations in data assimilation systems has been shown to lead to a more accurate analysis, the inclusion of more observation information content and improvements in the forecast skill score (Stewart *et al.*, 2008; Stewart, 2010; Weston *et al.*, 2014; Bormann *et al.*, 2016).

One difficulty in quantifying observation-error correlations is that they can only be estimated in a statistical sense, not calculated directly. The method proposed by Desroziers *et al.* (2005) has become popular for estimating observation-error statistics due to its simplicity (a detailed discussion of this diagnostic is given in section 3). The diagnostic provides an estimate of the observation-error covariance matrix using the statistical average of observation-minus-background and observation-minus-analysis residuals, assuming that the analysis is calculated using least-variance linear statistical estimation. It has also been shown to be applicable to scenarios where the analysis is calculated using both 3D and 4D variational assimilation methods (Desroziers *et al.*, 2005; Stewart, 2010). However, the diagnostic only provides a correct estimate of the observation-error covariance matrix if the assumed background- and observation-error statistics used in the assimilation are correct. As well as the impact of the assumed error statistics, the diagnostic has further limitations, such as the error introduced when using nonlinear observation operators (Terasaki and Miyoshi, 2014) and the fact that an ergodic assumption is often made in order to obtain sufficient sample residuals (Todling, 2015). However, Desroziers *et al.* (2005) also show that the result may be improved if successive iterations of the diagnostic are applied. Furthermore, with careful interpretation of the results, the diagnostic can still provide useful information about the true observation-error statistics when the assumed statistics, used in the assimilation, are not exact (Ménard, 2016; Waller *et al.*, 2016b). Despite these limitations, the diagnostic has been used successfully in some studies to estimate observation-error variances and correlations. It has been used in simple model experiments (Li *et al.*, 2009; Stewart, 2010; Miyoshi *et al.*, 2013) and to estimate time-varying observation errors (Waller *et al.*, 2014a). The diagnostic has also been applied to operational NWP observations to calculate interchannel error covariances (Bormann and Bauer, 2010; Bormann *et al.*, 2010, 2016; Stewart *et al.*, 2014; Weston *et al.*, 2014; Waller *et al.*, 2016a) and spatial error covariances (Cordoba *et al.*, 2016; Waller *et al.*, 2016a, 2016c) in variational assimilation systems.

Another application of the diagnostics is their use in learning about the assimilation system: for example to test self-consistency in the system (Desroziers *et al.*, 2005) and to determine sources of errors (Waller *et al.*, 2016a, 2016c).

One issue that appears to have been overlooked is that the diagnostics are derived assuming that the analysis is calculated using a best linear unbiased estimator. In recent work, the diagnostic has been applied to calculate observation errors where the analysis has been calculated using an ensemble assimilation scheme employing domain and observation localization techniques (Lange and Janjić, 2016; Schraff *et al.*, 2016). In this article, we consider whether the diagnostics of Desroziers *et al.* (2005) are still appropriate for calculating observation-error statistics for observations used in a local assimilation scheme. We provide a new derivation of the diagnostics using the analysis calculated by a local ensemble assimilation. From this derivation, we show that the diagnostic equations no longer hold and that the statistical averages of observation-minus-background and observation-minus-analysis residuals no longer result in an estimate of the observation-error covariance matrix, in general. However, further analysis of our derived diagnostics shows that, under certain circumstances, some elements of the observation-error covariance matrix can, in principle, be recovered exactly. Those elements that cannot, in principle, be derived exactly we describe as *'incorrectly estimated'*. Furthermore, we provide a method to determine which elements of the observation-error covariance matrix can be estimated correctly. In particular, the correct estimation of correlations is dependent on both the localization radius and the observation operator. We provide some special cases that show, dependent on specific background- and observation-error statistics and observation operators, that one may be lucky and may, in theory, be able to recover all elements of the observation-error covariance matrix, or unlucky and able to recover none. We also use examples to show that it is possible that, theoretically, some elements will be estimated incorrectly by the diagnostic, but, due to the choice of specific background- and observation-error statistics and observation operators, the estimated values may be close to the true values. However, some prior knowledge of the true statistics is required to be able to validate the quality of the incorrect estimates. Therefore, if the estimated error statistics are to be utilized further, it is necessary to find another method to assign values to those elements of the covariance matrix that cannot be estimated correctly by the diagnostic. This may be achieved by, for example, applying techniques such as those in Higham (2002) to provide a nearest approximate correlation matrix.

Despite these problems, the estimated observation-error covariance matrices can still be useful. It is known that even the use of approximate observation-error covariance matrices in data assimilation can improve analysis accuracy (Healy and White, 2005; Stewart, 2010; Stewart *et al.*, 2013).

This article is organized as follows. We begin in section 2 by describing deterministic ensemble Kalman filters and their local implementation. We review the standard diagnostics of Desroziers *et al.* (2005) in section 3. In section 4, we derive the diagnostics where the analysis is calculated using a localized assimilation scheme and determine when the diagnostic can be used to estimate error correlations. In section 5, we demonstrate our theoretical results using numerical examples. Finally, we present our conclusions in section 6.

2. Data assimilation

2.1. Notation

Data assimilation techniques combine observations, $\mathbf{y} \in \mathbb{R}^p$, available at time t with a model prediction of the state, the background $\mathbf{x}^\mathrm{b} \in \mathbb{R}^n$, which is often determined by a previous forecast. Here, p and n denote the dimensions of the observation and model state vectors respectively. In the assimilation, the observations and background are weighted by their respective error statistics, using the background- and observation-error covariance matrices, $\mathbf{B} \in \mathbb{R}^{n\times n}$ and $\mathbf{R} \in \mathbb{R}^{p\times p}$, to provide a best estimate of the state, $\mathbf{x}^{\mathrm{a}} \in \mathbb{R}^n$, known as the analysis. After an assimilation step, the analysis is then evolved forward in time using a (possibly nonlinear) model to provide a background at the next assimilation time.

One of the simplest forms of data assimilation scheme is the best linear unbiased estimator. Using this scheme, the analysis is obtained using

$$\begin{aligned} \mathbf{x}^{\mathrm{a}} &= \mathbf{x}^{\mathrm{b}} + \widetilde{\mathbf{B}}\mathbf{H}^{\mathrm{T}}\left(\mathbf{H}\widetilde{\mathbf{B}}\mathbf{H}^{\mathrm{T}} + \widetilde{\mathbf{R}}\right)^{-1}\left(\mathbf{y} - \mathcal{H}\left(\mathbf{x}^{\mathrm{b}}\right)\right) \\ &= \mathbf{x}^{\mathrm{b}} + \widetilde{\mathbf{K}}\mathbf{d}^{\mathrm{ob}}, \end{aligned} \tag{1}$$

where $\mathcal{H} : \mathbb{R}^n \to \mathbb{R}^p$ is the (possibly nonlinear) observation operator and $\mathbf{H}$ is the observation operator linearized about the current state. Here, we restrict ourselves to the use of a linear observation operator. The innovation, or observation-minus-background residual, is denoted by

$$\mathbf{d}^{\mathrm{ob}} = \mathbf{y} - \mathbf{H}\mathbf{x}^{\mathrm{b}}. \tag{2}$$

The assumed observation- and background-error covariance matrices $\widetilde{\mathbf{R}}$ and $\widetilde{\mathbf{B}}$ are used to weight the observations and background in the assimilation. The matrix

$$\widetilde{\mathbf{K}} = \widetilde{\mathbf{B}}\mathbf{H}^{\mathrm{T}}\left(\mathbf{H}\widetilde{\mathbf{B}}\mathbf{H}^{\mathrm{T}} + \widetilde{\mathbf{R}}\right)^{-1} \tag{3}$$

is the Kalman gain ($\widetilde{\mathbf{K}} \in \mathbb{R}^{n\times p}$) used in the assimilation. We make the distinction between the assumed error statistics, denoted by $\widetilde{\cdot}$, and the exact error statistics, as this is important for the derivation of the diagnostic in section 3.

2.2. *Deterministic ensemble filters*

We now describe the general form of deterministic square-root ensemble filter. At each assimilation time we have an ensemble, a statistical sample of m state estimates $\left\{\mathbf{x}^{(k)}\right\}$ for $k = 1, \ldots, m$. From this ensemble it is possible to calculate the ensemble mean,

$$\bar{\mathbf{x}} = \frac{1}{m}\sum_{k=1}^{m} \mathbf{x}^{(k)}, \tag{4}$$

and ensemble perturbation matrix, $\mathbf{X} \in \mathbb{R}^{n\times m}$:

$$\mathbf{X} = \frac{1}{\sqrt{m-1}}\left(\begin{array}{cccc} \mathbf{x}^{(1)} - \bar{\mathbf{x}} & \mathbf{x}^{(2)} - \bar{\mathbf{x}} & \ldots & \mathbf{x}^{(m)} - \bar{\mathbf{x}} \end{array}\right). \tag{5}$$

Using the background ensemble members to calculate the background ensemble mean $\bar{\mathbf{x}}^{\mathrm{b}}$ and background ensemble perturbation matrix $\mathbf{X}^{\mathrm{b}}$ allows us to write the background ensemble covariance matrix as

$$\mathbf{P}^{\mathrm{b}} = \mathbf{X}^{\mathrm{b}}\mathbf{X}^{\mathrm{b}\,\mathrm{T}} = \widetilde{\mathbf{B}}. \tag{6}$$

We note that the background ensemble covariance matrix will not be exact and therefore can be thought of as the assumed background-error covariance matrix. The background ensemble mean is updated to provide the analysis ensemble mean, $\bar{\mathbf{x}}^{\mathrm{a}}$, as

$$\bar{\mathbf{x}}^{\mathrm{a}} = \bar{\mathbf{x}}^{\mathrm{b}} + \widetilde{\mathbf{K}}\left(\mathbf{y} - \mathbf{H}\bar{\mathbf{x}}^{\mathrm{b}}\right), \tag{7}$$

where the Kalman gain is constructed as in Eq. (3) using the background ensemble covariance matrix given in Eq. (6). (If the observation operator is nonlinear, the matrix $\widetilde{\mathbf{K}}$ is defined differently, as in e.g. Hunt *et al.* (2007). Alternatively, the nonlinearity may be dealt with using an augmented state (Evensen, 2003).)

The ensemble perturbation matrix update then gives information on the analysis-error covariance matrix. We do not describe the update of the ensemble perturbations in detail, but instead note that a number of different approaches are available (e.g. Tippett *et al.*, 2003), though one must be careful to ensure that the chosen form is unbiased (Livings *et al.*, 2008). We do not include the equations here, as we only need to understand the updated analysis mean for use with the Desroziers *et al.* (2005) diagnostic, which is introduced in section 3.

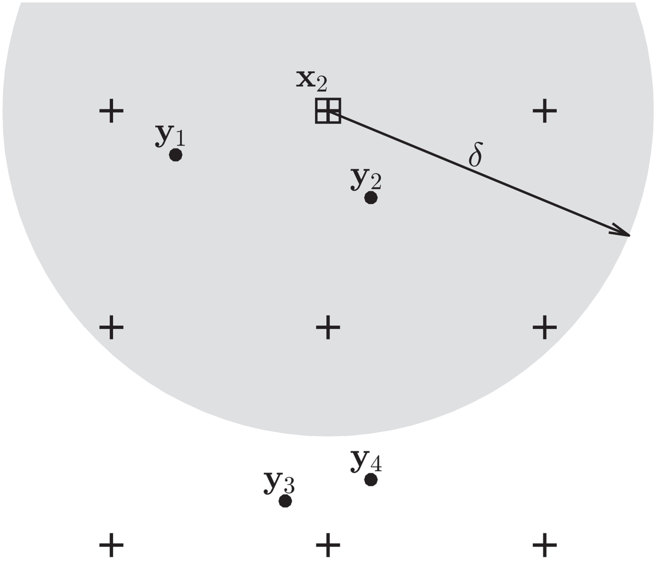

Figure 1. Schematic of domain localization with grid points (pluses), observations (dots) and local domain (shaded grey circle) with localization radius δ. Note that we order the grid points from left to right and top to bottom with $\mathbf{x}_1$ (top left) to $\mathbf{x}_9$ (bottom right). For clarity, we only label selected grid points in the figure. When domain localization is applied, the component of the state vector located at the highlighted grid point $\mathbf{x}_2$ would be updated using observations $\mathbf{y}_1$ and $\mathbf{y}_2$ that fall within the shaded area.

2.3. *Local assimilation*

In this section, we consider how domain localization is applied to ensemble assimilation schemes. In domain localization, each component of the background state vector is updated individually using a subset of observations located within a given distance, δ.

In general, the ith component of the mean background vector, $\bar{\mathbf{x}}_i$, is updated independently using a local set of $\check{p}\{i\}$ observations, $\check{\mathbf{y}}\{i\} \in \mathbb{R}^{\check{p}\{i\}}$. The local observations can be defined as a sub-vector of the full observation vector,

$$\check{\mathbf{y}}\{i\} = \boldsymbol{\Phi}\{i\}\mathbf{y}, \tag{8}$$

where $\boldsymbol{\Phi}\{i\} \in \mathbb{R}^{\check{p}\{i\}\times p}$ is a selection matrix, with elements 1 and 0, that selects only observations within a given distance, δ, from the corresponding location of $\bar{\mathbf{x}}_i$. (We remark that it is possible to update a subvector of states simultaneously if each component in the sub-vector is updated using the same local set of observations.)

Example. We present an example of the update of a single component in Figure 1. In this schematic, the component $\mathbf{x}_2$ would be updated using only those observations (black dots) that lie within a given radius (shaded grey circle). For this example, we are able to determine that the selection matrix for the highlighted grid point would be

$$\boldsymbol{\Phi}\{2\} = \left[\begin{array}{cccc} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{array}\right],$$

which selects observations $\mathbf{y}_1$ and $\mathbf{y}_2$. (For this example, it would be possible to update $\mathbf{x}_1$ and $\mathbf{x}_2$ simultaneously, as both of these points would be updated using only observations $\mathbf{y}_1$ and $\mathbf{y}_2$.)

As only a sub-vector of observations is used, it is also necessary to consider only a sub-vector of the modelled observations,

$$\check{\mathbf{y}}^{\mathrm{b}}\{i\} = \boldsymbol{\Phi}\{i\}\mathbf{H}\bar{\mathbf{x}}^{\mathrm{b}}, \tag{9}$$

and a local Kalman gain, $\breve{\widetilde{\mathbf{K}}}\{i\} \in \mathbb{R}^{n\times\check{p}\{i\}}$, defined as

$$\begin{aligned} &\breve{\widetilde{\mathbf{K}}}\{i\} = \\ &\quad \mathbf{X}^{\mathrm{b}}\mathbf{X}^{\mathrm{b}\,\mathrm{T}}\mathbf{H}^{\mathrm{T}}\boldsymbol{\Phi}\{i\}^{\mathrm{T}}\left(\boldsymbol{\Phi}\{i\}\left(\mathbf{H}\mathbf{X}^{\mathrm{b}}\mathbf{X}^{\mathrm{b}\,\mathrm{T}}\mathbf{H}^{\mathrm{T}} + \widetilde{\mathbf{R}}\right)\boldsymbol{\Phi}\{i\}^{\mathrm{T}}\right)^{-1}. \end{aligned} \tag{10}$$

The local update to the ensemble mean is calculated using

$$\bar{\mathbf{x}}_i^{\mathrm{a}} = \bar{\mathbf{x}}_i^{\mathrm{b}} + \sum_{j=1}^{\breve{p}\{i\}} \tilde{\breve{\mathbf{K}}}\{i\}_{i,j} \left(\boldsymbol{\Phi}\{i\} \left(\mathbf{y} - \mathbf{H}\bar{\mathbf{x}}^{\mathrm{b}} \right) \right)_j . \tag{11}$$

We note that only the *i*th row of the *i*th local Kalman gain is required for the update.

Following the update of the ensemble mean, it is necessary to update the ensemble perturbations. Again, we do not describe the update of the ensemble perturbations in detail; rather, we note that the ensemble perturbation updates are also performed locally using the local forms of the required matrices (Sakov and Bertino, 2011).

3. The standard observation-space diagnostic

Desroziers *et al.* (2005) present a set of diagnostics that provides estimates of the observation- and background-error covariance statistics based on combinations of observation-minus-background (OmB, also known as the innovation, given in Eq. (2)), observation-minus-analysis (OmA) and analysis-minus-background (AmB) residuals. In this article, we focus primarily on use of the diagnostics to estimate observation-error statistics. (Use of the diagnostics to estimate background-error statistics is discussed in Appendix A.) In the derivation of the diagnostic in Desroziers *et al.* (2005), it is assumed that the analysis is determined using the best linear unbiased estimator described in Eq. (1). Calculating the analysis in this way allows the analysis residual (OmA) to be defined as

$$\begin{aligned} \mathbf{d}^{\mathrm{oa}} &= \mathbf{y} - \mathbf{H}\mathbf{x}^{\mathrm{a}} \\ &= \mathbf{y} - \mathbf{H}\mathbf{x}^{\mathrm{b}} - \mathbf{H}\widetilde{\mathbf{K}}\mathbf{d}^{\mathrm{ob}}. \end{aligned} \tag{12}$$

Under the assumption that the forecast and observation errors are uncorrelated, Desroziers *et al.* (2005) show that an estimate of the observation-error correlation matrix can be obtained by taking the statistical expectation of the outer product of the analysis and background residuals:

$$\begin{aligned} E\left[\mathbf{d}^{\mathrm{oa}}\mathbf{d}^{\mathrm{ob}^{\mathrm{T}}}\right] &= \widetilde{\mathbf{R}}\left(\mathbf{H}\widetilde{\mathbf{B}}\mathbf{H}^{\mathrm{T}} + \widetilde{\mathbf{R}}\right)^{-1}\left(\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}} + \mathbf{R}\right) \\ &= \widetilde{\mathbf{R}}\widetilde{\mathbf{S}}^{-1}\mathbf{S} \\ &= \mathbf{R}^{\mathrm{e}}, \end{aligned} \tag{13}$$

where $\mathbf{R}^{\mathrm{e}}$ is the estimated observation-error covariance matrix and $\mathbf{B}$ and $\mathbf{R}$ are the exact background and observation covariance matrices. Here, we define $\mathbf{S} = \mathbf{R} + \mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}}$ and $\widetilde{\mathbf{S}} = \widetilde{\mathbf{R}} + \mathbf{H}\widetilde{\mathbf{B}}\mathbf{H}^{\mathrm{T}}$. If the observation and forecast errors used in the assimilation are exact, $\widetilde{\mathbf{R}} = \mathbf{R}$ and $\widetilde{\mathbf{B}} = \mathbf{B}$, then

$$E\left[\mathbf{d}^{\mathrm{oa}}\mathbf{d}^{\mathrm{ob}^{\mathrm{T}}}\right] = \mathbf{R}. \tag{14}$$

However, in practice the estimate will be subject to sampling error. Here, we will refer to Eq. (14) as the 'standard' form of the observation diagnostic.

The calculation of these statistics for non-homogeneous, irregular datasets, where different observations are used in each assimilation cycle, cannot be performed using the matrix multiplication above. Instead, this calculation can be achieved by pairing components of the background and analysis residuals and binning the $\mathbf{d}^{\mathrm{oa}}$ and $\mathbf{d}^{\mathrm{ob}}$ pairs. The pairing and binning of observations will depend on the type of correlation being estimated. For example, if we wish to calculate spatial correlations, we may consider pairing only residuals that occur simultaneously in time; the binning of the $\mathbf{d}^{\mathrm{oa}}$ and $\mathbf{d}^{\mathrm{ob}}$ pairs will depend on the distance between the spatial locations of $\mathbf{d}^{\mathrm{oa}}$ and $\mathbf{d}^{\mathrm{ob}}$. If we were to estimate time correlations, then we would consider pairs of residuals at the same location; the binning of the residual pairs would then be dependent on the time difference between $\mathbf{d}^{\mathrm{oa}}$ and $\mathbf{d}^{\mathrm{ob}}$. The covariance, cov(β), is then computed individually for each bin, β, using

$$\mathrm{cov}(\beta) = \frac{1}{N^{\beta}} \sum_{k=1}^{N^{\beta}} \left(\mathbf{d}_i^{\mathrm{oa}}\mathbf{d}_j^{\mathrm{ob}}\right)_k - \frac{1}{N^{\beta}} \sum_{k=1}^{N^{\beta}} \left(\mathbf{d}_i^{\mathrm{oa}}\right)_k \frac{1}{N^{\beta}} \sum_{k=1}^{N^{\beta}} \left(\mathbf{d}_j^{\mathrm{ob}}\right)_k , \tag{15}$$

where $\left(\mathbf{d}_i^{\mathrm{oa}}\mathbf{d}_j^{\mathrm{ob}}\right)_k$ is the *k*th pair of elements of $\mathbf{d}^{\mathrm{oa}}$ and $\mathbf{d}^{\mathrm{ob}}$ in bin β and N^{β} is the number of residual pairs in bin β. We note that the second term ensures that the calculation is not affected by bias (Waller *et al.*, 2016a).

4. Observation-space diagnostics using localized analyses

In this section, we revisit the derivation of the diagnostics, but in this case we assume that the analysis is calculated using the local ensemble assimilation described in section 2. For the purposes of this derivation we make several assumptions.

(1) We consider a scalar case where each individual state variable is updated using a local set of observations. We note that this can be extended to the case where a local analysis update is applied to a vector of state variables (e.g. all variables in a given column at a given latitude and longitude) that share the same set of local observations.
(2) We note that in this section we are concerned only with the background and analysis ensemble means and not the individual ensemble members, therefore to simplify the notation we drop the overbar.
(3) In section 3 we demonstrated that the standard diagnostic in Eq. (14) is only correct when $\widetilde{\mathbf{R}}$ and $\widetilde{\mathbf{B}}$ used in the assimilation are correctly specified and therefore, for the derivation in this section, we make the assumption that the error covariance statistics used in the assimilation are exact, that is $\widetilde{\mathbf{R}} = \mathbf{R}$ and $\widetilde{\mathbf{B}} = \mathbf{X}^{\mathrm{b}}\mathbf{X}^{\mathrm{b}^{\mathrm{T}}} = \mathbf{B}$.

4.1. *Derivation of diagnostic using local analyses*

If we then assume that the analysis is calculated using a local assimilation scheme, as in Eq. (11), the analysis residual is given by

$$\begin{aligned} \mathbf{d}_i^{\mathrm{oa}} &= \mathbf{y}_i - \left(\mathbf{H}\mathbf{x}^{\mathrm{a}}\right)_i \\ &= \mathbf{d}_i^{\mathrm{ob}} - \sum_{k=1}^{n} \sum_{l=1}^{\breve{p}\{k\}} \mathbf{H}_{i,k}\breve{\mathbf{K}}\{k\}_{k,l} \left(\boldsymbol{\Phi}\{i\}\mathbf{d}^{\mathrm{ob}}\right)_l . \end{aligned} \tag{16}$$

Thus, the diagnostics can be written in component form as

$$\begin{aligned} \mathbf{R}^{\mathrm{e}}_{i,j} &= E\left[\mathbf{d}_i^{\mathrm{oa}}\mathbf{d}_j^{\mathrm{ob}}\right], \\ &= E\left[\left(\mathbf{d}_i^{\mathrm{ob}} - \sum_{k=1}^{n} \sum_{l=1}^{\breve{p}\{k\}} \mathbf{H}_{i,k}\breve{\mathbf{K}}\{k\}_{k,l} \left(\boldsymbol{\Phi}\{k\}\mathbf{d}^{\mathrm{ob}}\right)_l\right) \mathbf{d}_j^{\mathrm{ob}}\right], \\ &= \mathbf{R}_{i,j} + \left(\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}}\right)_{i,j} \\ &\quad - \sum_{k=1}^{n} \mathbf{H}_{i,k} \left(\mathbf{B}\mathbf{H}^{\mathrm{T}}\boldsymbol{\Phi}\{k\}^{\mathrm{T}} \left(\boldsymbol{\Phi}\{k\}\mathbf{S}\boldsymbol{\Phi}\{k\}^{\mathrm{T}}\right)^{-1} \boldsymbol{\Phi}\{k\}\mathbf{S}\right)_{k,j}, \\ &= \mathbf{R}_{i,j} + \left(\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}}\right)_{i,j} - (\mathbf{H}\mathbf{F})_{i,j} . \end{aligned} \tag{17}$$

In Eq. (17), we define $\mathbf{F} \in \mathbb{R}^{n\times p}$ as

$$\mathbf{F}_{k,j} = \mathbf{G}\{k\}_{k,j}, \tag{18}$$

where

$$\mathbf{G}\{k\} = \mathbf{B}\mathbf{H}^{\mathrm{T}}\boldsymbol{\Phi}\{k\}^{\mathrm{T}} \left(\boldsymbol{\Phi}\{k\}\mathbf{S}\boldsymbol{\Phi}\{k\}^{\mathrm{T}}\right)^{-1} \boldsymbol{\Phi}\{k\}\mathbf{S}. \tag{19}$$

We note that $\mathbf{F} \neq \mathbf{G}\{k\}$, but that the kth row of $\mathbf{F}$ is equal to the kth row of $\mathbf{G}\{k\}$. Furthermore, the rows of the matrix $\mathbf{F}$ have a relation to the analysis grid points, with the kth row related to $\mathbf{x}_k$, and the columns of the matrix $\mathbf{F}$ have a relation to the observations, with the jth column related to the observation $\mathbf{y}_j$.

Equation (17) relies on the use of the correct error covariances in the assimilation and is based on updating the state using an assimilation scheme employing localization. The standard diagnostic gives a correct estimate of the observation-error covariance matrix. However, when local analyses are used it is not clear if the diagnostic gives a correct estimate of the matrix $\mathbf{R}$.

From Eq. (17), we see that the diagnostic will only result in $\mathbf{R}^{\mathrm{e}}_{i,j} = \mathbf{R}_{i,j}$ if $\left(\mathbf{HBH}^{\mathrm{T}}\right)_{i,j} - (\mathbf{HF})_{i,j} = 0$. To determine if this holds, we first consider which elements of the matrix $\mathbf{F}$ are equal to corresponding elements of the matrix $\mathbf{BH}^{\mathrm{T}}$. To understand which elements of $\mathbf{F}$ have the correct value, we must consider the elements of the kth row of each of the matrices $\mathbf{G}\{k\}$. We note that $\mathbf{G}\{k\}$ is dependent on $\boldsymbol{\Phi}\{k\}$, which is a selection matrix containing zeros and ones. By reordering the vector of observations, the matrix $\boldsymbol{\Phi}\{k\}$ can always be arranged into the block form $\widehat{\boldsymbol{\Phi}}\{k\} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \end{bmatrix}$, where $\mathbf{I} \in \mathbb{R}^{\check{p}\{k\}\times\check{p}\{k\}}$ is the identity matrix. Applying this rearrangement (denoted by $\widehat{\cdot}$) to all the required matrices, $\widehat{\mathbf{G}\{k\}}$ can be calculated as follows:

$$
\begin{aligned}
\widehat{\mathbf{G}\{k\}} &= \widehat{\mathbf{BH}^{\mathrm{T}}}\widehat{\boldsymbol{\Phi}}\{k\}^{\mathrm{T}}\left(\widehat{\boldsymbol{\Phi}}\{k\}\widehat{\mathbf{S}}\widehat{\boldsymbol{\Phi}}\{k\}^{\mathrm{T}}\right)^{-1}\widehat{\boldsymbol{\Phi}}\{k\}\widehat{\mathbf{S}} \\
&= \begin{bmatrix} \widehat{\mathbf{BH}^{\mathrm{T}}}_{[1,1]} & \widehat{\mathbf{BH}^{\mathrm{T}}}_{[1,2]} \end{bmatrix}\begin{bmatrix} \mathbf{I} \\ \mathbf{0} \end{bmatrix} \\
&\quad\times\left(\begin{bmatrix} \mathbf{I} & \mathbf{0} \end{bmatrix}\begin{bmatrix} \hat{\mathbf{S}}_{[1,1]} & \hat{\mathbf{S}}_{[1,2]} \\ \hat{\mathbf{S}}_{[2,1]} & \hat{\mathbf{S}}_{[2,2]} \end{bmatrix}\begin{bmatrix} \mathbf{I} \\ \mathbf{0} \end{bmatrix}\right)^{-1} \\
&\quad\times\begin{bmatrix} \mathbf{I} & \mathbf{0} \end{bmatrix}\begin{bmatrix} \hat{\mathbf{S}}_{[1,1]} & \hat{\mathbf{S}}_{[1,2]} \\ \hat{\mathbf{S}}_{[2,1]} & \hat{\mathbf{S}}_{[2,2]} \end{bmatrix} \\
&= \begin{bmatrix} \widehat{\mathbf{BH}^{\mathrm{T}}}_{[1,1]} & \widehat{\mathbf{BH}^{\mathrm{T}}}_{[1,1]}\hat{\mathbf{S}}^{-1}_{[1,1]}\hat{\mathbf{S}}_{[1,2]} \end{bmatrix}. \qquad (20)
\end{aligned}
$$

We see that the first set of columns of the matrix, i.e. the columns related to the observations selected by $\boldsymbol{\Phi}\{k\}$ and used in the local update, are equal to the corresponding elements of $\widehat{\mathbf{BH}^{\mathrm{T}}}$. In contrast, the second set of columns are not equal to $\widehat{\mathbf{BH}^{\mathrm{T}}}_{[1,2]}$. They are, in fact, related to the first set of columns, $\widehat{\mathbf{BH}^{\mathrm{T}}}_{[1,1]}$ multiplied by $\hat{\mathbf{S}}^{-1}_{[1,1]}\hat{\mathbf{S}}_{[1,2]}$ (note that in general $\hat{\mathbf{S}}^{-1}_{[1,1]}\hat{\mathbf{S}}_{[1,2]} \neq \mathbf{I}$). In Appendix B, we provide a simple example that demonstrates why the calculation of $\mathbf{F}$ fails to produce the desired elements of $\mathbf{HBH}^{\mathrm{T}}$ and how, in turn, this impacts on the estimate of the observation-error covariance matrix.

Rearranging the columns back to the original observation vector ordering results in $\mathbf{G}\{k\}$ having its jth column equal to the jth column of $\mathbf{BH}^{\mathrm{T}}$ only if the jth observation was used in the local update of $\mathbf{x}_k$. Using this information, we are able to determine that

$$
\begin{array}{ll}
\text{if } \mathbf{y}_j \in \left\{\left(\check{\mathbf{y}}\{k\}\right)_l, l = 1,\ldots,\check{p}\{k\}\right\} \text{ then} & \mathbf{F}_{k,j} = \left(\mathbf{BH}^{\mathrm{T}}\right)_{k,j}; \\
\text{otherwise} & \mathbf{F}_{k,j} \neq \left(\mathbf{BH}^{\mathrm{T}}\right)_{k,j}.
\end{array} \qquad (21)
$$

In other words, the (k,j)th element of $\mathbf{F}$ is equal to the (k,j)th element of $\mathbf{BH}^{\mathrm{T}}$ only if the analysis for state $\mathbf{x}_k$ was calculated using the observation $\mathbf{y}_j$. We remark that no calculation is required to determine which elements of $\mathbf{F}$ have the correct value. It is sufficient to know only which observations are used to update which analysis states. We are able to determine that the kth row of $\mathbf{F}$ will only have correct elements where the observation $\mathbf{y}_j$ has been used to update analysis state $\mathbf{x}_k$, i.e. is within the localization distance. As well as considering a localization distance around a grid point $\mathbf{x}_k$, it is possible to consider a 'region of influence' of an observation.

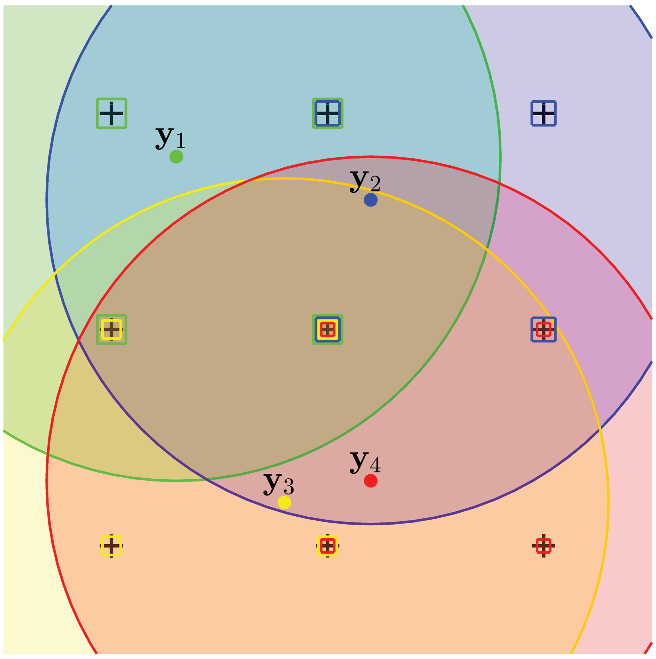

Figure 2. Schematic of regions of observation influence when domain localization is applied to ensemble assimilation schemes. Grid points (pluses) and observations (dots), with observations coloured with corresponding regions of observation influence (shaded coloured circles). Assuming that the model equivalent observations are calculated using the four nearest model states, the coloured squares around grid points select the points that would be utilized by the observation operator for the observation of the corresponding colour. Note that we order the grid points from left to right and top to bottom with $\mathbf{x}_1$ (top left) to $\mathbf{x}_9$ (bottom right). For clarity, we only label selected grid points in the figure.

> *The **region of influence** of an observation is the set of analysis states that are updated in the assimilation using the observation* $\mathbf{y}_i$.

Using this definition, we are able to determine that the jth column of $\mathbf{F}$ will only have correct elements where the analysis state $\mathbf{x}_k$ has been updated using the jth observation.

Example. In Figure 1 in section 2, we introduced a simple example where there are four observations available to update nine states. Using Eq. (21) and the diagram in Figure 1, we can determine which elements $\mathbf{F}_{i,j}$ are equal to $(\mathbf{BH}^{\mathrm{T}})_{i,j}$. If we denote elements where $\mathbf{F}_{i,j} = (\mathbf{BH}^{\mathrm{T}})_{i,j}$ by ✓and elements where $\mathbf{F}_{i,j} \neq (\mathbf{BH}^{\mathrm{T}})_{i,j}$ by ✗, we have

$$
\mathbf{F} = \begin{bmatrix}
\checkmark & \checkmark & \text{✗} & \text{✗} \\
\checkmark & \checkmark & \text{✗} & \text{✗} \\
\text{✗} & \checkmark & \text{✗} & \text{✗} \\
\checkmark & \checkmark & \checkmark & \checkmark \\
\checkmark & \checkmark & \checkmark & \checkmark \\
\text{✗} & \checkmark & \checkmark & \checkmark \\
\text{✗} & \text{✗} & \checkmark & \checkmark \\
\text{✗} & \text{✗} & \checkmark & \checkmark \\
\text{✗} & \text{✗} & \checkmark & \checkmark
\end{bmatrix}.
$$

Using Figure 1, we see that that $\mathbf{x}_2$ would be updated using only observations $\mathbf{y}_1$ and $\mathbf{y}_2$ and hence in the second row of $\mathbf{F}$ only the first two columns have correct entries.

In Figure 2, we see the same example as in Figure 1, but now the coloured shaded areas show the region of influence for each corresponding observation. Figure 2 shows that observation $\mathbf{y}_2$ will only be used to update $\mathbf{x}_1,\ldots,\mathbf{x}_6$; as a result, the second column of $\mathbf{F}$ only has the first six elements estimated correctly.

4.2. *Determining correctly estimated elements of the observation-error covariance matrix for regular datasets*

To determine which elements of the observation-error covariance matrix can be estimated correctly, we require only two things:

- to know which elements of the observation operator are zero and
- to know which elements of **F** are incorrect.

It is useful to store this information in two matrices. Let $\mathbf{C} \in \mathbb{R}^{p\times n}$ contain information on the non-zero elements of **H**, where

$$\mathbf{C}_{i,j} \overset{\text{def}}{=} \begin{cases} 0 & \text{if } \mathbf{H}_{i,j} = 0, \\ 1 & \text{if } \mathbf{H}_{i,j} \neq 0, \end{cases} \tag{22}$$

and let $\mathbf{D} \in \mathbb{R}^{n\times p}$ contain information on the incorrect elements of **F**, where

$$\mathbf{D}_{i,j} \overset{\text{def}}{=} \begin{cases} 0 & \text{if } \mathbf{F}_{i,j} = \left(\mathbf{B}\mathbf{H}^{\mathrm{T}}\right)_{i,j}, \\ 1 & \text{if } \mathbf{F}_{i,j} \neq \left(\mathbf{B}\mathbf{H}^{\mathrm{T}}\right)_{i,j}. \end{cases} \tag{23}$$

We note that, to compute **D**, it is sufficient only to know the location of the observations in relation to the location of the analysis states. The matrix **C** provides information about the 'domain of dependence' of an observation.

> *The* ***domain of dependence*** *of an observation* $\mathbf{y}_i$ *is the set of elements of the model state that are used to calculate the model equivalent of* $\mathbf{y}_i$ *(i.e. the set of states* $\{\mathbf{x}_k : \mathbf{H}_{i,k} \neq 0\}$*).*

The form of **C** and **D** ensures that product of these two matrices $\mathbf{L} = \mathbf{CD}$ can be used to determine which elements of $\mathbf{R}^{\mathrm{e}}$ can be estimated correctly using the diagnostic. In other words,

$$\begin{array}{ll} \text{if } \mathbf{L}_{i,j} = 0 \text{ then} & \mathbf{R}^{\mathrm{e}}_{i,j} = \mathbf{R}_{i,j}, \\ \text{otherwise} & \mathbf{R}^{\mathrm{e}}_{i,j} \neq \mathbf{R}_{i,j}. \end{array} \tag{24}$$

We note that, for the elements where $\mathbf{L}_{i,j} \neq 0$, the estimated value of $\mathbf{R}^{\mathrm{e}}_{i,j}$ is given by Eqs (17), (18) and (19). In section 5, we provide examples that show that the values estimated incorrectly by the diagnostic may be close to the true values or can be far from the truth; hence, without some prior knowledge of the true statistics, one cannot be certain of the quality of the incorrect estimates.

Using the definitions of the *domain of dependence* and *region of influence*, we can interpret Eq. (24) further as follows:

> *The correlation between the errors of observations* $\mathbf{y}_i$ *and* $\mathbf{y}_j$ *can be estimated using the diagnostic only if the domain of dependence for observation* $\mathbf{y}_i$ *lies within the region of influence of observation* $\mathbf{y}_j$*.*

From this statement, we can see that using the diagnostic will not necessarily result in a symmetric matrix; it is possible that the domain of dependence for observation $\mathbf{y}_i$ lies within the region of influence of observation $\mathbf{y}_j$, but that the domain of dependence for observation $\mathbf{y}_j$ does not lie within the region of influence of observation $\mathbf{y}_i$. Hence, in this case, the element $\mathbf{R}_{i,j}$ can be estimated whereas the element $\mathbf{R}_{j,i}$ cannot.

Although developed in the framework of a linear observation operator, our conclusion should hold when a nonlinear observation operator is applied, as it depends only on which states the observation operator acts on and not how the observation operator acts on the state.

Example. We must now assume a form for our observation operator in our example. We assume that, to calculate a model observation equivalent, the observation operator acts on the four closest state points to the observation itself. This is shown in Figure 2, where for each coloured observation the state points required to make the model observation are selected with a correspondingly coloured square. Using an observation operator of this form and using the previous step of our example, we are able to determine that

$$\mathbf{C} = \begin{bmatrix} 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 \end{bmatrix},$$

$$\mathbf{D} = \begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix}.$$

By comparing **D** with the **F** previously calculated in our example, we see that **D** represents in binary the correct and incorrect elements of **F**. Finally, using **C** and **D** we are able to determine that

$$\mathbf{L} = \begin{bmatrix} 0 & 0 & 2 & 2 \\ 2 & 0 & 2 & 2 \\ 2 & 2 & 0 & 0 \\ 3 & 2 & 0 & 0 \end{bmatrix}.$$

The matrix **L** tells us which elements of the observation-error covariance matrix can be estimated accurately. Using **L** along with Figure 2, we see the following.

- All the observation-error variances (diagonal elements) are estimated correctly.
- We are not able to estimate the elements $\mathbf{R}_{1,3}, \mathbf{R}_{3,1}, \mathbf{R}_{1,4}$ and $\mathbf{R}_{4,1}$. This is intuitive, since we see from Figure 2 that observations $\mathbf{y}_3$ and $\mathbf{y}_4$ do not fall within the region of influence of $\mathbf{y}_1$.
- We are able to estimate $\mathbf{R}_{3,4}$ and $\mathbf{R}_{4,3}$, since the domain of dependence for observation $\mathbf{y}_3$ lies within the region of influence of $\mathbf{y}_4$ and the domain of dependence for observation $\mathbf{y}_4$ lies within the region of influence of $\mathbf{y}_3$.
- We are able to estimate $\mathbf{R}_{1,2}$ but cannot estimate $\mathbf{R}_{2,1}$, despite the fact that $\mathbf{y}_1$ and $\mathbf{y}_2$ lie within each other's regions of influence. This is because the domain of dependence for observation $\mathbf{y}_1$ lies within the region of influence of $\mathbf{y}_2$, but the domain of dependence for observation $\mathbf{y}_2$ does not lie within the region of influence of $\mathbf{y}_1$.
- The remaining entries of **R** cannot be estimated correctly.

In our example, we are able to estimate correctly 7 of the 16 elements of **R**. In general, the number of elements that can be estimated correctly will be dependent on the localization radius and observation operator. We note two particular special cases of **H** that show that one may be lucky and recover all elements of the observation-error covariance matrix, or be unlucky and recover none.

- If **H** contains no zero entries (this scenario is likely to be unphysical), then the diagnostics may provide no correct information when applied to analyses calculated using the localization.
- If $\mathbf{H} = \mathbf{I}$ and $n = p$, i.e. the system is fully observed, then the diagnostic will have the same correct entries as matrix **F**. If, in addition, **B** and **R** are diagonal, then a local assimilation scheme does produce the correct solution (this follows, since for all $\widehat{\mathbf{G}\{k\}}$ the components of $\widehat{\mathbf{S}}_{[1,2]}$ are zero).

This methodology for determining correct elements is applicable to regular datasets where the same observation locations are used in each assimilation cycle. In this scenario, since the same set of local observations will be used for each local analysis update in each assimilation cycle, it will only ever be possible to recover certain elements of the observation-error covariance matrix.

To use the estimated statistics in an assimilation scheme, it will be necessary to 'fill in' those elements of the covariance matrix that cannot be estimated correctly by the diagnostic. It may be possible to 'fill in' some of these elements using the calculated information: e.g., if we have been able to estimate $\mathbf{R}^{\mathrm{e}}_{i,j}$ correctly, but not $\mathbf{R}^{\mathrm{e}}_{j,i}$, then, using the symmetric nature of a correlation matrix, we can replace $\mathbf{R}^{\mathrm{e}}_{j,i}$ with $\mathbf{R}^{\mathrm{e}}_{i,j}$. For the elements of the observation-error covariance matrix for which we have no reliable information, it may be necessary to assume no correlation or to estimate the elements by employing more sophisticated techniques such as those in Higham (2002) to provide a nearest approximate correlation matrix.

4.3. *Application to irregular datasets*

If the dataset is irregular, then we must modify the above approach slightly. As discussed in section 3, if we wish to calculate observation errors for irregular observations then we must use the summation form of the diagnostic and bin samples, rather than a simple multiplication of samples $\mathbf{d}^{\mathrm{oa}}$ and $\mathbf{d}^{\mathrm{ob}}$. Instead of calculating $\mathbf{R}^{\mathrm{e}}$ and determining which elements have been estimated correctly, we must be more careful and only consider pairs of $\mathbf{d}^{\mathrm{oa}}$ and $\mathbf{d}^{\mathrm{ob}}$ that give the correct result. This can be achieved by calculating **C**, **D** and **L** for each assimilation cycle. We can then use the values in the matrix **L** to determine which $\mathbf{d}^{\mathrm{oa}}$ $\mathbf{d}^{\mathrm{ob}}$ pairs can be used in the observation-error statistics calculation:

$$\begin{aligned} &\text{if } \mathbf{L}_{i,j} = 0 \quad \text{use } \mathbf{d}^{\mathrm{oa}}_i \mathbf{d}^{\mathrm{ob}}_j \text{ in the calculation,} \\ &\text{if } \mathbf{L}_{i,j} \neq 0 \quad \text{do not use } \mathbf{d}^{\mathrm{oa}}_i \mathbf{d}^{\mathrm{ob}}_j \text{ in the calculation.} \end{aligned} \tag{25}$$

Similarly, then, it is appropriate to use $\mathbf{d}^{\mathrm{oa}}_i \mathbf{d}^{\mathrm{ob}}_j$ in the calculation of observation-error correlation only if the observation operator applied to calculate observation $\mathbf{y}_i$ acts only on states that have been updated using the observation $\mathbf{y}_j$. In other words,

> *When using the diagnostic to estimate observation-error statistics, it is appropriate to use the pair of residuals,* $\mathbf{d}^{\mathrm{oa}}_i \mathbf{d}^{\mathrm{ob}}_j$, *only if the domain of dependence for observation* $\mathbf{y}_i$ *falls within the region of influence of* $\mathbf{y}_j$.

When calculating spatial correlations, one often assumes that the observation-error correlations will be isotropic. Because of this assumption, it is possible that we may be more successful in recovering the observation-error statistics than in the regular dataset case. Even so, it is possible that we will not be able to recover information on all the correlations we require. In this case, it may be necessary to fit a correlation function to the estimated data.

5. Illustration of theoretical results

5.1. *Experiment design*

In this section, we describe ensemble filter and local ensemble filter twin experiments based on the example presented in Figures 1 and 2 in sections 2 and 4. We set an explicit grid spacing of one unit and localization distance of $\delta = 1.5$. We use the analyses from these experiments to estimate the observation-error covariance matrix and use the results to illustrate the theoretical results presented in section 4.

We define the observation operator to be bilinear interpolation with equal weighting, i.e.

$$\mathbf{H} = 0.25\mathbf{C},$$

where **C** is defined in the example in section 4.2. We determine the true observation- and background-error covariance matrix elements using

$$\mathbf{R}_{i,j} = \sigma_r^2 \mathrm{e}^{(-\Delta y_{i,j}/L_r)}, \tag{26}$$

where $\Delta y_{i,j}$ is the distance between observations y_i and y_j, and

$$\mathbf{B}_{i,j} = \sigma_b^2 \mathrm{e}^{(-\Delta x_{i,j}/L_b)}, \tag{27}$$

where $\Delta x_{i,j}$ is the distance between states x_i and x_j, respectively. Both **B** and **R** are taken from Markov distributions (Wilks, 1995). For the assimilation experiments we have a single true solution; from this truth, pseudo-observations are created by adding errors drawn from $\mathcal{N}(0, \mathbf{R})$ and the background is determined by adding errors drawn from $\mathcal{N}(0, \mathbf{B})$. The background is then perturbed to create the ensemble members. We perform a single assimilation step using both a standard ensemble Kalman filter (Eq. (7)) with 1000 members and a local ensemble filter (Eq. 11)) with either 1000 or 100 members. To ensure accuracy of the background ensemble covariance matrix, $\mathbf{P}^f$, and avoid the ergodic assumption of the diagnostic, we do not implement a dynamical model. Instead, the single assimilation cycle experiment is repeated 5000 times, with new background, ensemble members and pseudo-observations created each time. The resulting OmA and OmB residuals are used to provide an estimate of the observation-error covariance matrix. (We note that we choose a large number of residual samples, since we wish to reduce the influence on the diagnostic of sampling error.) We analyse the quality of the estimated observation-error correlation matrix by calculating, for each element, the percentage error compared to the true error covariance:

$$100 \times \frac{\left|\mathbf{R}_{i,j} - \mathbf{R}^{\mathrm{e}}_{i,j}\right|}{\left|\mathbf{R}_{i,j}\right|}. \tag{28}$$

5.2. *Experimental results*

We now consider how well the diagnostic estimates $\mathbf{R}^{\mathrm{e}}$ for two different choices of **B**. We note that the example in section 4.2 shows which elements we expect to be estimated correctly and incorrectly. We will first consider the case where $\sigma_b = \sigma_r = 1$ and $L_b = L_r = 1$. We plot the percentage errors in Figure 3(a) for $\mathbf{R}^{\mathrm{e}}$ estimated using 5000 residual samples from a 1000 member ensemble filter, Figure 3(b) for $\mathbf{R}^{\mathrm{e}}$ estimated using 5000 residual samples from a 1000 member local ensemble filter and Figure 3(c) for $\mathbf{R}^{\mathrm{e}}$ estimated using 5000 residual samples from a 100 member local ensemble filter. We see that, for this choice of **B** and **R**, the errors in the 'incorrectly estimated' elements are small, with the largest error in Figure 3(b) being similar to the percentage error of the reference $\mathbf{R}^{\mathrm{e}}$ calculated using analyses from a standard ensemble filter. We also note that reducing the number of ensemble members appears to reduce the percentage error values slightly.

We next consider the case where $\sigma_b = 2$, $\sigma_r = 1$, $L_b = 0.5$ and $L_r = 1$. We plot the percentage errors in Figure 4(a) for $\mathbf{R}^{\mathrm{e}}$ estimated using 5000 residual samples from a 1000 member ensemble filter, Figure 4(b) for $\mathbf{R}^{\mathrm{e}}$ estimated using 5000 residual samples from a 1000 member local ensemble filter and Figure 4(c) for $\mathbf{R}^{\mathrm{e}}$ estimated using 5000 residual samples from a 100 member local ensemble filter. We note that the colour scale for Figure 4 is a factor of four larger than that in Figure 3. We see that, for this choice of **B**, the errors in the 'incorrectly estimated' elements are large, with some of the errors as large as the actual elements themselves. In this example, it is clear that the correctly estimated elements are a reasonable estimate of the true error statistics, whereas the incorrectly estimated elements provide poor information.

These two examples highlight that it is not possible, in general, to determine whether the incorrect elements will give estimates close to the true values. This suggests that one can never be certain about the quality of the incorrect estimates.

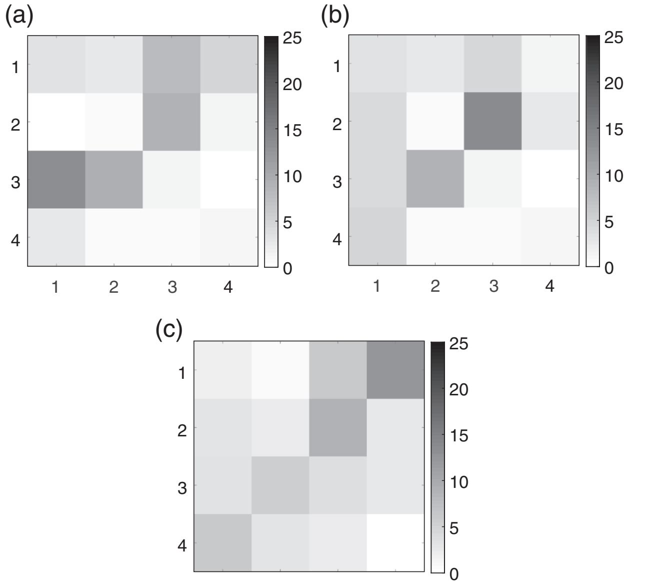

Figure 3. Percentage error in estimated observation-error covariance matrices when $\sigma_r = 1$, $\sigma_b = 1$, $L_r = 1$ and $L_b = 1$. (a) Percentage error in elements of $\mathbf{R}^e$ estimated using analysis residuals from a 1000 member ensemble Kalman filter; largest error 13.4%. (b) Percentage error in elements of $\mathbf{R}^e$ estimated using analysis residuals from a 1000 member local ensemble Kalman filter; largest error 14.0%. (c) Percentage error in elements of $\mathbf{R}^e$ estimated using analysis residuals from a 100 member local ensemble Kalman filter; largest error 12.2%. Note that the colour scale is a factor of four smaller than that in Figure 4.

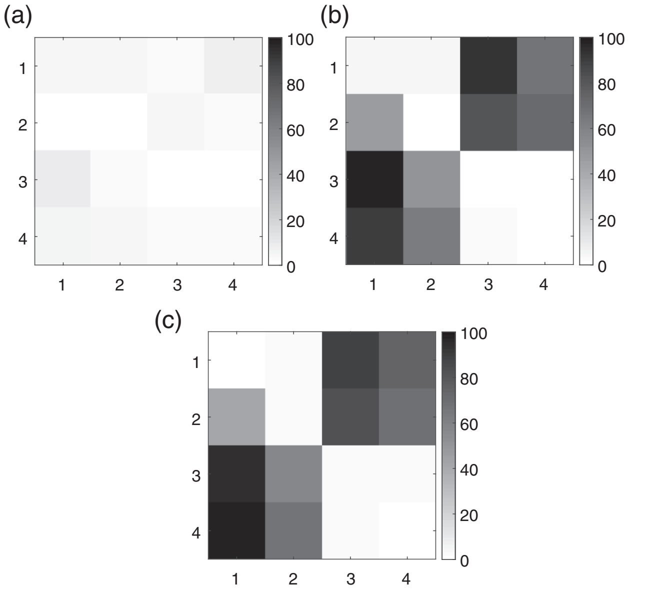

Figure 4. Percentage error in estimated observation-error covariance matrices when $\sigma_r = 1$, $\sigma_b = 2$, $L_r = 1$ and $L_b = 2$. (a) Percentage error in elements of $\mathbf{R}^e$ estimated using analysis residuals from a 1000 member ensemble Kalman filter; largest error 8.5%. (b) Percentage error in elements of $\mathbf{R}^e$ estimated using analysis residuals from a 1000 member local ensemble Kalman filter; largest error 97.7%. (c) Percentage error in elements of $\mathbf{R}^e$ estimated using analysis residuals from a 100 member local ensemble Kalman filter; largest error 97.7%. Note that the colour scale is a factor of four larger than that in Figure 3.

6. Conclusions

To obtain an accurate analysis, it is important that the observation- and background-error statistics are specified accurately in the data assimilation system. A number of operational NWP centres are now using ensemble assimilation schemes that provide flow-dependent background-error statistics. However, for the computational cost of these ensemble methods to be affordable it is necessary to employ additional constraints such as variance inflation and localization. Research involving the improved treatment of observation-error statistics has been a more recent area of research. The positive impact of including correlated observation-error statistics in operational data assimilation has highlighted the need to improve their specification and inclusion in assimilation systems. One method available to estimate observation-error statistics is the diagnostic of Desroziers *et al.* (2005). The diagnostic makes use of statistical averages of observation-minus-background and observation-minus-analysis residuals to provide an estimate of the observation-error covariance matrix. However, the diagnostic only gives a correct estimate of the observation-error statistics if the background- and observation-error statistics used in the analysis update are correctly specified. Despite the limitations, the diagnostic has been used successfully to estimate observation-error statistics for numerical weather prediction (Bormann and Bauer, 2010; Bormann *et al.*, 2010, 2016; Stewart *et al.*, 2014; Weston *et al.*, 2014; Cordoba *et al.*, 2016; Waller *et al.*, 2016a, 2016c). One important fact that appears to have been overlooked is that the diagnostic is derived assuming that the analysis is calculated using a best linear unbiased estimator.

Here, we consider whether domain localization impacts on the results of the diagnostic of Desroziers *et al.* (2005). We provide a new derivation of the diagnostic, where we assume that the analysis has been calculated using a local ensemble filter. This derivation shows that, in general, even when the assumed background- and observation-error statistics are exact, the diagnostic no longer provides a correct estimate of the observation-error covariance matrix.

Although we show that the diagnostics no longer provide a correct estimation of the observation-error covariance matrix, we are able, under the assumption that the assumed background- and observation-error statistics are correctly specified, to determine which elements of the observation-error covariance matrix are recoverable. It is possible to estimate the error correlations between two observations only if the observation operator applied to calculate the model equivalent observation $\mathbf{y}_i$ acts only on states that have been updated using the observation $\mathbf{y}_j$. In other words, one can determine the following.

> *The correlation between the errors of observations $\mathbf{y}_i$ and $\mathbf{y}_j$ can be estimated using the diagnostic only if the domain of dependence for observation $\mathbf{y}_i$ lies within the region of influence of observation $\mathbf{y}_j$.*

Using the same logic, it is possible to determine which pairs of analysis and background residuals can be used to provide an accurate estimate of the observation-error statistics. For special (unlikely) cases, we can show that one may be lucky and recover all elements of the observation-error covariance matrix or unlucky and recover none. Using examples, we show that the values estimated incorrectly by the diagnostic may be close to the true values or may be far from the truth; hence, without some prior knowledge of the true statistics, one cannot be certain of the quality of the incorrect estimates.

To use the estimated statistics in an assimilation scheme, it will be necessary to ‘fill in’ those elements of the covariance matrix that cannot be estimated correctly by the diagnostic. Some elements may be determined by using the symmetric nature of a correlation matrix. For those correlations for which we have no reliable information, it may be necessary to assume a particular correlation structure or estimate the elements by employing more sophisticated techniques, such as those in Higham (2002), to provide a nearest approximate correlation matrix.

If the assumed background- and observation-error statistics are not exact, then those elements that can be estimated will be affected. In this case, the theoretical work of Waller *et al.* (2016b) may be used to provide knowledge on how the assumed statistics may affect the estimated observation-error statistics.

We note that we have not considered the impact of variance inflation here, but instead suggest that if the inflation factor is not correct then the variance of the background-error statistics will be either under- or overestimated. In the case of under- or overestimated variances, the theoretical work of Waller *et al.* (2016b) can be used to provide an understanding of the impact on the estimated observation-error statistics. We do not discuss other forms of localization here, as a different derivation of the diagnostic would be required.

Our new derivation in this article may, at a glance, suggest that it is no longer possible to use the diagnostic to estimate observation-error statistics using analyses that have been calculated with an assimilation system employing localization. However, with careful consideration the diagnostic may be used to calculate some elements of the observation-error covariance matrix correctly.

Acknowledgements

This work is funded in part by the NERC grants NE/K008900/1 'Forecasting Rainfall exploiting new data Assimilation techniques and Novel observations of Convection' (FRANC) and NE/N006682/1 'Improving high impact weather forecasts via an international comparison of ObServation error Correlations in data Assimilation' (OSCA) and in part by the NERC National Centre for Earth Observation.

Appendix A: Diagnosing background-error statistics

In this section, we present the diagnostic described in Desroziers *et al.* (2005) used to estimate background-error statistics. We show that this diagnostic is also affected when analyses from local assimilation schemes are used.

Using the standard diagnostic to estimate the background-error statistics mapped into observation space, we make use of the innovation and analysis-minus-background (AmB) residuals:

$$\begin{aligned}\mathbf{d}^{\mathrm{ab}} &= \mathbf{H}\mathbf{x}^{\mathrm{a}} - \mathbf{H}\mathbf{x}^{\mathrm{b}}\\ &\approx \mathbf{H}\widetilde{\mathbf{K}}\mathbf{d}^{\mathrm{ob}}.\end{aligned} \tag{A1}$$

The background-error statistics in observation space, $\mathbf{H}\mathbf{B}^{\mathrm{e}}\mathbf{H}^{\mathrm{T}}$, can then be estimated using

$$\begin{aligned}&E\left[\mathbf{d}^{\mathrm{ab}}{\mathbf{d}^{\mathrm{ob}}}^{\mathrm{T}}\right]\\ &= \mathbf{H}\widetilde{\mathbf{B}}\mathbf{H}^{\mathrm{T}}\left(\mathbf{H}\widetilde{\mathbf{B}}\mathbf{H}^{\mathrm{T}} + \widetilde{\mathbf{R}}\right)^{-1}\left(\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}} + \mathbf{R}\right)\\ &= \mathbf{H}\mathbf{B}^{\mathrm{e}}\mathbf{H}^{\mathrm{T}}.\end{aligned} \tag{A2}$$

Again, $\mathbf{H}\mathbf{B}^{\mathrm{e}}\mathbf{H}^{\mathrm{T}} = \mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}}$ when the matrices $\widetilde{\mathbf{B}}$ and $\widetilde{\mathbf{R}}$ used in assimilation are exact. Practical calculation of this diagnostic for irregular datasets can be achieved in a similar way to the observation-error computation, Eq. (15).

We now consider the impact of local assimilation analyses on the diagnostics using the same assumptions as in section 4, i.e.

(1) we consider a scalar case where each individual state variable is updated using a local set of observations;
(2) all state variables refer to the ensemble mean; however, to simplify the notation we drop the overbar; and
(3) we make the assumption that the error covariance statistics used in the assimilation are exact, i.e. $\widetilde{\mathbf{R}} = \mathbf{R}$ and $\widetilde{\mathbf{B}} = \mathbf{X}^{\mathrm{b}}{\mathbf{X}^{\mathrm{b}}}^{\mathrm{T}} = \mathbf{B}$.

Now, assuming that the analysis is calculated using a local assimilation scheme, then in component form the diagnostic is given by

$$\begin{aligned}&\left(\mathbf{H}\mathbf{B}^{\mathrm{e}}\mathbf{H}^{\mathrm{T}}\right)_{i,j}\\ &= E\left[\mathbf{d}^{\mathrm{ab}}_i\mathbf{d}^{\mathrm{ob}}_j\right]\\ &= E\left[\sum_{k=1}^{n}\sum_{l=1}^{\breve{p}\{k\}}\mathbf{H}_{i,k}\breve{\mathbf{K}}\{k\}_{k,l}\left(\mathbf{\Phi}\{k\}\mathbf{d}^{\mathrm{ob}}\right)_l\mathbf{d}^{\mathrm{ob}}_j\right]\\ &= (\mathbf{H}\mathbf{F})_{i,j},\end{aligned} \tag{A3}$$

where $\mathbf{F}$ is defined as in Eq. (18). We note that the third term of Eq. (17) is equal to Eq. (A3) and hence it is possible to use the same methodology developed in section 4.2 to determine which elements of $\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}}$ may be estimated correctly using the diagnostic.

Appendix B: Simple 2 × 2 example of diagnostic failure

We now provide a simple example to show how the diagnostic fails when domain localization is employed. Suppose we have two state values, x_1 and x_2, and two direct ($\mathbf{H} = \mathbf{I}_{2\times 2}$) observations, y_1 and y_2. We assume that the localization acts such that x_1 is updated using y_1 and x_2 is updated using y_2 and therefore $\mathbf{\Phi}\{1\} = [1\ \ 0]$ and $\mathbf{\Phi}\{2\} = [0\ \ 1]$. We assume the error correlation matrices are

$$\mathbf{R} = \begin{bmatrix} a & b\\ b & c\end{bmatrix}$$

and

$$\mathbf{B} = \begin{bmatrix} d & e\\ e & f\end{bmatrix}.$$

The estimated covariance matrix is given by

$$\mathbf{R}^{\mathrm{e}} = \mathbf{R} + \mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}} - \mathbf{H}\mathbf{F}. \tag{B1}$$

We must first determine the elements of the matrix $\mathbf{F}$, which can be calculated using Eq. (19):

$$\mathbf{F} = \begin{bmatrix} d & d\left(\dfrac{b+e}{a+d}\right)\\ f\left(\dfrac{b+e}{c+f}\right) & f\end{bmatrix}$$

and

$$\mathbf{R}^{\mathrm{e}} = \begin{bmatrix} a & a\left(\dfrac{b+e}{a+d}\right)\\ c\left(\dfrac{b+e}{c+f}\right) & c\end{bmatrix}. \tag{B2}$$

We note that the variances are estimated correctly, but the covariances are not. Furthermore, the estimated matrix is not even symmetric. We now consider three cases to illustrate how the 'incorrect' elements can be very poor, or may indeed give reasonable estimates, although for the wrong reasons.

B1. Case 1

If we let $\mathbf{R} = \mathbf{B}$, then $a\left(\dfrac{b+e}{a+d}\right) = c\left(\dfrac{b+e}{c+f}\right) = b$ and all the elements of the observation-error covariance matrix are correct.

B2. Case 2

If we let $\mathbf{R}$ and $\mathbf{B}$ be correlation matrices ($a = c = d = f = 1$), with $\mathbf{B}$ uncorrelated ($e = 0$), then $a\left(\dfrac{b+e}{a+d}\right) = c\left(\dfrac{b+e}{c+f}\right) = b/2$. The estimated correlation is half the value it should be.

B3. Case 3

If we let $\mathbf{R}$ and $\mathbf{B}$ be correlation matrices ($a = c = d = f = 1$), with $-b = e$, then $a\left(\dfrac{b+e}{a+d}\right) = c\left(\dfrac{b+e}{c+f}\right) = 0$. The estimate provided by the diagnostic suggests no correlation, when in fact there should be correlations, possibly very strong.

References

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

Bannister RN. 2008. A review of forecast error covariance statistics in atmospheric variational data assimilation. I: Characteristics and measurements of forecast error covariances. *Q. J. R. Meteorol. Soc.* **134**: 1951–1970.

Bormann N, Bauer P. 2010. Estimates of spatial and interchannel observation-error characteristics for current sounder radiances for numerical weather prediction. I: Methods and application to ATOVS data. *Q. J. R. Meteorol. Soc.* **136**: 1036–1050.

Bormann N, Bonavita M, Dragani R, Eresmaa R, Matricardi M, McNally A. 2016. Enhancing the impact of IASI observations through an updated observation-error covariance matrix. *Q. J. R. Meteorol. Soc.* **142**: 1767–1780. https://dx.doi.org/10.1002/qj.2774.

Bormann N, Collard A, Bauer P. 2010. Estimates of spatial and interchannel observation-error characteristics for current sounder radiances for numerical weather prediction. II: Application to AIRS and IASI data. *Q. J. R. Meteorol. Soc.* **136**: 1051–1063.

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

Cordoba M, Dance S, Kelly G, Nichols N, Waller J. 2016. Diagnosing atmospheric motion vector observation errors for an operational high resolution data assimilation system. *Q. J. R. Meteorol. Soc.* **143**: 333–341. https://dx.doi.org/10.1002/qj.2925.

Desroziers G, Berre L, Chapnik B, Poli P. 2005. Diagnosis of observation, background and analysis-error statistics in observation space. *Q. J. R. Meteorol. Soc.* **131**: 3385–3396.

Evensen G. 1994. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res.* **99**: 10143–10162. https://dx.doi.org/10.1029/94JC00572.

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

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

Healy SB, White AA. 2005. Use of discrete Fourier transforms in the 1D-Var retrieval problem. *Q. J. R. Meteorol. Soc.* **131**: 63–72.

Higham N. 2002. Computing the nearest correlation matrix -- a problem from finance. *IMA J. Numer. Anal.* **22**: 329–343.

Hodyss D, Nichols NK. 2015. The error of representation: Basic understanding. *Tellus A* **67**: 1–17. https://dx.doi.org/10.3402/tellusa.v67.24822.

Houtekamer PL, Mitchell HL. 1998. Data assimilation using an ensemble Kalman filter technique. *Mon. Weather Rev.* **126**: 796–811.

Houtekamer PL, Mitchell HL. 2001. A sequential ensemble Kalman filter for atmospheric data assimilation. *Mon. Weather Rev.* **129**: 123–137. https://doi.org/10.1175/1520-0493(2001)129⟨0123:ASEKFF⟩2.0.CO;2.

Hunt BR, Kostelich EJ, Szunyogh I. 2007. Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter. *Physica D* **230**: 112–126. https://dx.doi.org/10.1016/j.physd.2006.11.008.

Janjić T, Cohn SE. 2006. Treatment of observation error due to unresolved scales in atmospheric data assimilation. *Mon. Weather Rev.* **134**: 2900–2915.

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. https://dx.doi.org/10.1175/2011MWR3552.1.

Lange H, Janjić T. 2016. Assimilation of Mode-S EHS aircraft observations in COSMO-KENDA. *Mon. Weather Rev.* **144**: 1697–1711. https://dx.doi.org/10.1175/MWR-D-15-0112.1.

Lewis JM, Lakshmivarahan S, Dhall SK. 2006. *Dynamic Data Assimilation*. Cambridge University Press: Cambridge, UK.

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.* **128**: 1367–1386.

Livings DM, Dance SL, Nichols NK. 2008. Unbiased ensemble square root filters. *Physica D* **237**: 1021–1028.

Lorenc AC. 1981. A global three-dimensional multivariate statistical interpolation scheme. *Mon. Weather Rev.* **109**: 701–721.

Lorenc AC. 1986. Analysis methods for numerical weather prediction. *Q. J. R. Meteorol. Soc.* **112**: 1177–1194.

Ménard R. 2016. Error covariance estimation methods based on analysis residuals: Theoretical foundation and convergence properties derived from simplified observation networks. *Q. J. R. Meteorol. Soc.* **142**: 257–273. https://doi.org/10.1002/qj.2650.

Miyoshi T, Kalnay E, Li H. 2013. Estimating and including observation-error correlations in data assimilation. *Inverse Prob. Sci. Eng.* **21**: 387–398.

Nerger L, Janj T, Schröter J, Hiller W. 2012. A regulated localization scheme for ensemble-based Kalman filters. *Q. J. R. Meteorol. Soc.* **138**: 802–812. https://dx.doi.org/10.1002/qj.945.

Ott E, Hunt BR, Szunyogh I, Zimin AV, Kostelich EJ, Corazza M, Kalnay E, Patil DJ, Yorke JA. 2004. A local ensemble Kalman filter for atmospheric data assimilation. *Tellus A* **56**: 415–428. https://dx.doi.org/10.1111/j.1600-0870.2004.00076.x.

Petrie RE, Dance SL. 2010. Ensemble-based data assimilation and the localisation problem. *Weather* **65**: 65–69. https://dx.doi.org/10.1002/wea.505.

Sakov P, Bertino L. 2011. Relation between two common localisation methods for the ENKF. *Comput. Geosci.* **15**: 225–237. https://dx.doi.org/10.1007/s10596-010-9202-6.

Schraff C, Reich H, Rhodin A, Schomburg A, Stephan K, Periáñez A, Potthast R. 2016. Kilometre-scale ensemble data assimilation for the COSMO model (KENDA). *Q. J. R. Meteorol. Soc.* **142**: 1453–1472. https://dx.doi.org/10.1002/qj.2748.

Stewart LM. 2010. 'Correlated observation errors in data assimilation', PhD thesis. University of Reading: Reading, UK. https://www.reading.ac.uk/maths-and-stats/research/theses/maths-phdtheses.aspx (accessed 28 July 2017).

Stewart LM, Dance SL, Nichols NK. 2008. Correlated observation errors in data assimilation. *Int. J. Numer. Methods Fluids* **56**: 1521–1527.

Stewart LM, Dance SL, Nichols NK. 2013. Data assimilation with correlated observation errors: Experiments with a 1-D shallow water model. *Tellus A* **65**: 1–14. https://dx.doi.org/10.3402/tellusa.v65i0.19546.

Stewart LM, Dance SL, Nichols NK, Eyre JR, Cameron J. 2014. Estimating interchannel observation-error correlations for IASI radiance data in the Met Office system. *Q. J. R. Meteorol. Soc.* **140**: 1236–1244. https://doi.org/10.1002/qj.2211.

Terasaki K, Miyoshi T. 2014. Data assimilation with error-correlated and non-orthogonal observations: Experiments with the Lorenz-96 model. *SOLA* **10**: 210–213. https://doi.org/10.2151/sola.2014-044.

Tippett MK, Anderson JL, Bishop CH, Hamil TM, Whitaker JS. 2003. Ensemble square root filters. *Mon. Weather Rev.* **131**: 1485–1490.

Todling R. 2015. A complementary note to 'A lag-1 smoother approach to system-error estimation': The intrinsic limitations of residual diagnostics. *Q. J. R. Meteorol. Soc.* **141**: 2917–2933. https://doi.org/10.1002/qj.2546.

Waller JA. 2013. 'Using observations at different spatial scales in data assimilation for environmental prediction', PhD thesis. Department of Mathematics and Statistics, University of Reading: Reading, UK. https://www.reading.ac.uk/maths-and-stats/research/theses/maths-phdtheses.aspx.

Waller JA, Dance SL, Lawless AS, Nichols NK. 2014a. Estimating correlated observation-error statistics using an ensemble transform Kalman filter. *Tellus A* **66**: 1–15. http://dx.doi.org/10.3402/tellusa.v66.23294.

Waller JA, Dance SL, Lawless AS, Nichols NK, Eyre JR. 2014b. Representativity error for temperature and humidity using the Met Office high-resolution model. *Q. J. R. Meteorol. Soc.* **140**: 1189–1197. https://doi.org/10.1002/qj.2207.

Waller JA, Ballard SP, Dance SL, Kelly G, Nichols NK, Simonin D. 2016a. Diagnosing horizontal and interchannel observation-error correlations for SEVIRI observations using observation-minus-background and observation-minus-analysis statistics. *Remote Sens.* **8**: 581.

Waller JA, Dance SL, Nichols NK. 2016b. Theoretical insight into diagnosing observation-error correlations using observation-minus-background and observation-minus-analysis statistics. *Q. J. R. Meteorol. Soc.* **142**: 418–431. https://doi.org/10.1002/qj.2661.

Waller JA, Simonin D, Dance SL, Nichols NK, Ballard SP. 2016c. Diagnosing observation-error correlations for Doppler radar radial winds in the Met Office UKV model using observation-minus-background and observation-minus-analysis statistics. *Mon. Weather Rev.* **144**: 3533–3551. https://doi.org/10.1175/MWR-D-15-0340.1.

Weston PP, Bell W, Eyre JR. 2014. Accounting for correlated error in the assimilation of high-resolution sounder data. *Q. J. R. Meteorol. Soc.* **140**: 2420–2429. https://doi.org/10.1002/qj.2306.

Whitaker JS, Hamill TM. 2002. Ensemble data assimilation without perturbed observations. *Mon. Weather Rev.* **130**: 1913–1924. https://doi.org/10.1175/1520-0493(2002)130<1913:EDAWPO>2.0.CO;2.

Wilks DS. 1995. *Statistical Methods in the Atmospheric Sciences*. Academic Press: San Diego.