RMetS
Royal Meteorological Society

A comparison of 4DVar with ensemble data assimilation methods

D. Fairbairn,[a,b*] S. R. Pring,[b†] A. C. Lorenc[b†] and I. Roulstone[c]

[a]*Centre for Environmental Strategy, University of Surrey, UK*
[b]*Met Office, Exeter, UK*
[c]*Mathematics Department, University of Surrey, UK*

*Correspondence to: D. Fairbairn, Met Office, FitzRoy Road, Exeter EX1 3PB, UK.
E-mail: David.Fairbairn@metoffice.gov.uk
†The contributions of these authors were written in the course of their employment at the Met Office, UK and are published with the permission of the Controller of HMSO and the Queen's Printer for Scotland.

Three data assimilation methods are compared for their ability to produce the best analysis: (i) 4DVar, four-dimensional variational data assimilation using linear and adjoint models with either a (perfect) 3D climatological background-error covariance or a 3D ensemble background-error covariance; (ii) EDA, an ensemble of 4DEnVars, which is a variational method using a 4D ensemble covariance; and (iii) the deterministic ensemble Kalman filter (DEnKF, also using a 4D ensemble covariance).

The accuracy of the deterministic analysis from each method was measured for both perfect and imperfect toy model experiments. With a perfect model, 4DVar with the climatological covariance is easily beaten by the ensemble methods, due to the importance of flow-dependent background-error covariances. When model error is present, 4DVar is more competitive and its relative performance is improved by increasing the observation density. This is related to the model error representation in the background-error covariance.

The dynamical time-consistency of the 4D ensemble background-error covariance is degraded by the localization, since the localization function and the nonlinear model do not commute. As a result, 4DVar with the ensemble covariance performs significantly better than the other ensemble methods when severe localization is required, i.e. for a small ensemble.

Key Words: model error; localization; 4DEnVar; observation density; covariance; additive inflation

Received 11 September 2012; Revised 11 December 2012; Accepted 5 February 2013; Published online in Wiley Online Library 2 May 2013

Citation: Fairbairn D, Pring SR, Lorenc AC, Roulstone I. 2014. A comparison of 4DVar with ensemble data assimilation methods. *Q. J. R. Meteorol. Soc.* **140**: 281–294. DOI:10.1002/qj.2135

1. Introduction

In recent years, there has been a growing use of ensemble forecast data to perform data assimilation (DA) in numerical weather prediction (NWP). We look at methods that are relevant to current operational practice and compare them for their ability to produce the best 'deterministic' analysis:

4DVar/4DVar-Ben. Four-dimensional variational DA, using linear and adjoint models. 4DVar uses a (perfect) 3D climatological background-error covariance matrix and 4DVar-Ben uses a 3D localized flow-dependent background-error covariance matrix coming from the EDA.

EDA-D/EDA-S. Ensemble of DAs that generates its own localized 4D ensemble covariance. This covariance is used and localized in a 4DEnVar variational analysis of each member, without the

need for linear or adjoint models. The spread is either maintained using a deterministic (EDA-D) or stochastic (EDA-S) formulation.

DEnKF. Ensemble Kalman filter using a localized 4D ensemble covariance and a deterministic analysis update step.

While some of the methods necessarily generate ensembles, the focus of the article is on the best estimate 'deterministic' analysis from the ensemble mean, since that best matches current NWP practice. The experiments will be performed on an idealized 1D toy model, but we continue to use the 3D and 4D terminology used for the methods in NWP. The results of this article will focus on judging the analyses produced by the DA methods, since we are able to verify against the 'truth'.

The 4DVar method has been the preferred method at most leading operational NWP centres since the 2000s (Rabier, 2005). More recently, there have been advocates for replacing 4DVar by an Ensemble Kalman Filter (EnKF) (Kalnay *et al.*, 2007). The main reasons for this transition from 4DVar to ensemble DA include the difficulty of improving the 4DVar background-error covariance (in particular to take into account the 'errors of the day'), the development and maintenance costs of the linear and adjoint models, the increasingly widespread use of ensemble forecasting systems, and the increasing parallelisation of supercomputers (which lends itself better to ensemble methods than 4DVar). Both approaches have advantages and disadvantages (Lorenc, 2003b), so it is attractive to seek a method combining advantages of each – the EDA and 4DVar-Ben methods both use variational algorithms but make use of an ensemble background-error covariance matrix.

One motivation for the experiments comes from Buehner *et al.* (2010a,b), who gave an intercomparison of the Environment Canada variational and ensemble DA systems on the global forecast model. Their methods were similar to those listed above. Their 4DVar with the ensemble background-error covariance (4DVar-Ben) performed better overall than with the climatological background-error covariance (4DVar). The difference in performance was more significant in the Southern Extratropics, which has sparser conventional observations than the Northern Extratropics. Their 4DEnVar method (similar to our EDA) gave similar performance to an EnKF. Buehner *et al.* (2010a,b) discussed a range of effects that could have explained their results, some of which were unique to their operational system. We use the Lorenz (2005) model to gain understanding of three of these effects:

1. The background-error covariance matrix construction (i.e. climatological or flow-dependent);
2. The different ways the ensemble methods perform localization over the assimilation window;
3. Differences in the observation densities between the Southern and Northern Extratropics.

Using a simple toy model, we can investigate a wider range of parameters than in Buehner *et al.* (2010b), such as various temporal and spatial observation densities, ensemble sizes, tuned covariance localization and inflation.

The layout of the article is as follows. Section 2 describes the DA methods used in this paper and section 3 describes the experimental set-up of the Lorenz (2005) model and the localization and inflation techniques used. Section 4 presents results for the different DA methods for the perfect model case (no model error) and section 5 presents the results for the case of an imperfect model (with model error present). Both sections contain a discussion of the results and section 6 concludes the article and provides a brief summary of the results in relation to NWP.

2. The DA methods

We next describe the following DA methods: 4DVar, 4DEnVar (a single deterministic implementation of four-dimensional ensemble variational DA), EDA (an ensemble of 4DEnVars), 4DVar-Ben, and finally the DEnKF. We note that no experiments are carried out with a single deterministic implementation of 4DEnVar, but a discussion of it is given to enable a description of the EDA. We also mention two different ways the EDA can maintain the ensemble spread: using a deterministic analysis update step (EDA-D) or perturbed observations (EDA-S). A summary of the different methods is given in Table 1.

2.1. 4DVar

Four-dimensional variational DA (4DVar; Le Dimet and Talagrand, 1986) provides a least-squares fit of a model trajectory to the background and the observations in an assimilation window. The propagation of the background $\mathbf{x}^{\mathrm{b}}$ (of dimension n) from the beginning of the window t_0 to the end of the window t_N using the forecast model is represented as:

$$\underline{\mathbf{x}}^{\mathrm{b}} = \underline{M}\{\mathbf{x}^{\mathrm{b}}(t_0)\}, \tag{1}$$

where $\underline{\mathbf{x}}^{\mathrm{b}}$ is the four-dimensional representation over the assimilation window (the underscore representing a time-dimensional vector). In this article we use incremental 4DVar (Courtier *et al.*, 1994), which finds the increment

$$\delta\mathbf{x}(t_0) = \mathbf{x}(t_0) - \mathbf{x}^{\mathrm{b}}(t_0), \tag{2}$$

which minimizes a cost function J. In order to improve the conditioning of the cost function, a control variable transform is applied (Lorenc, 2003a) so that the increment is defined as a function of $\mathbf{v}$:

$$\delta\mathbf{x}(t_0) = \mathbf{U}\mathbf{v}, \tag{3}$$

where $\mathbf{U}$ is designed such that $\mathbf{U}\mathbf{U}^{\mathrm{T}} = \mathbf{B}$ gives the climatological background-error covariance matrix. In this article $\mathbf{B}$ is homogeneous and $\mathbf{U}$ consists of a scaling by the square root of the estimated power spectrum and a spectral transform (as in Rawlins *et al.*, 2007). The increment is propagated using the tangent linear model $\underline{\mathbf{M}}$:

$$\delta\underline{\mathbf{x}} = \underline{\mathbf{M}}\delta\mathbf{x}(t_0). \tag{4}$$

The cost function obtained is

$$J[\mathbf{v}] = \frac{1}{2}\mathbf{v}^{\mathrm{T}}\mathbf{v} + [\underline{\mathbf{H}}\delta\underline{\mathbf{x}} - \underline{\mathbf{d}}]^{\mathrm{T}}\underline{\mathbf{R}}^{-1}[\underline{\mathbf{H}}\delta\underline{\mathbf{x}} - \underline{\mathbf{d}}], \tag{5}$$

where $\mathbf{R}$ is the observation-error covariance matrix and the innovation vector $\mathbf{d}$ is defined as:

$$\underline{\mathbf{d}} = \underline{\mathbf{y}}^{\mathrm{o}} - \underline{H}(\underline{\mathbf{x}}^{\mathrm{b}}). \tag{6}$$

Table 1. The data assimilation methods, for which results are given in sections 4 and 5.

Method	Type	Background cov. matrix	Background cov. source	Cov. propagation	Maintaining ensemble spread
4DVar	Variational	Climatological	Perfect cov. matrix (from bootstrapping method)	Tangent linear/ adjoint model	N/A
4DVar-Ben	Variational	3D-localized, flow-dependent	EDA-D ensemble	Tangent linear/ adjoint model	N/A
EDA-S	Ensemble of 4DEnVars	4D-localized, flow-dependent	Ensemble	Ensemble forecasts	Perturbed observations
EDA-D	Ensemble of 4DEnVars	4D-localized, flow-dependent	Ensemble	Ensemble forecasts	Deterministic (see Eq. (18), equivalent to Eq. (23))
DEnKF	Sequential	4D-localized, flow-dependent	Ensemble	Ensemble forecasts	Deterministic (see Eq. (23))

The linearisation of the observation operator H, mapping from model space to observation space, is denoted by $\mathbf{H}$. The gradient of the cost function is

$$\left[\frac{\partial J}{\partial \mathbf{v}}\right] = \mathbf{v} + \mathbf{U}^{\mathrm{T}}\underline{\mathbf{M}}^{\mathrm{T}}\underline{\mathbf{H}}^{\mathrm{T}}\underline{\mathbf{R}}^{-1}[\underline{\mathbf{H}}\delta\underline{\mathbf{x}} - \underline{\mathbf{d}}], \tag{7}$$

where $\mathbf{M}^{\mathrm{T}}$ is the adjoint model.

2.2. 4DEnVar

Four-dimensional ensemble variational DA (4DEnVar; Liu *et al.*, 2008, 2009; Buehner *et al.*, 2010a), also referred to as En4DVar in the literature, weights an ensemble of model trajectories according to how well the ensemble fits the observations and the background state in an assimilation window. The ensemble of model trajectories comes from a separate DA method and is only used to calculate perturbations and hence an implicit background-error covariance. The actual background, used in (6) and to which the analysed increments are added, comes from a high-resolution forecast from the previous analysis.

The design adopts as much as possible from 4DVar. The fundamental difference is that 4DVar uses a propagated climatological covariance which is replaced in 4DEnVar by a localized 4D ensemble covariance. From a background ensemble of model trajectories:

$$\underline{\mathbf{x}}^{\mathrm{b}}_j = \underline{M}\{\mathbf{x}^{\mathrm{b}}_j(t_0)\}, \qquad \text{for } j = 1, \ldots, m, \tag{8}$$

we calculate the mean trajectory $\overline{\underline{\mathbf{x}}}^{\mathrm{b}}$ and hence the perturbations from the mean ($\delta\underline{\mathbf{x}}^{\mathrm{b}}_j$). Equation (4), which uses the linear model, is replaced by a locally weighted linear combination of these perturbation trajectories:

$$\delta\underline{\mathbf{x}} = \sum_{j=1}^{m} \frac{1}{\sqrt{m-1}}\delta\underline{\mathbf{x}}^{\mathrm{b}}_j \circ \underline{\boldsymbol{\alpha}}_j, \tag{9}$$

where $\underline{\boldsymbol{\alpha}}_j$ is the smooth 4D field of weights given to the jth perturbation trajectory and $\circ$ denotes an element-by-element product between two matrices known as a Schur product. In the limit of constant $\underline{\boldsymbol{\alpha}}_j$ this is simply making the analysis increment equal to a linear combination of the perturbations – there are only $m-1$ degrees of freedom to fit the observations (one is lost because the perturbations sum to zero by construction). The smoothly varying $\underline{\boldsymbol{\alpha}}_j$ means this is still the case in local regions, but new values of $\underline{\boldsymbol{\alpha}}_j$ with new degrees of freedom can be selected some distance away. The smoothness of $\underline{\boldsymbol{\alpha}}_j$ is controlled by adding a penalty term equivalent to the background term in 4DVar:

$$J_\alpha\left(\underline{\boldsymbol{\alpha}}_1, \ldots, \underline{\boldsymbol{\alpha}}_m\right) = \sum_{j=1}^{m} \frac{1}{2}\underline{\boldsymbol{\alpha}}_j^{\mathrm{T}}\underline{\mathbf{C}}^{-1}\underline{\boldsymbol{\alpha}}_j, \tag{10}$$

where $\mathbf{C}$ is the localization matrix. As in (3), we use a control variable transform to condition J_α:

$$\underline{\boldsymbol{\alpha}}_j = \underline{\mathbf{U}}^{\alpha}\mathbf{v}^{\alpha}_j \quad \text{for } j = 1, \ldots, m, \tag{11}$$

where

$$\underline{\mathbf{U}}^{\alpha}(\underline{\mathbf{U}}^{\alpha})^{\mathrm{T}} = \underline{\mathbf{C}}. \tag{12}$$

The sequence of control vectors $\mathbf{v}^{\alpha}_j$ is concatenated to make the full variational control vector $\mathbf{v}^{\alpha s}$. The rest of the algorithm is the same as 4DVar, using the same cost function (5). We can define a new operator $\underline{\mathbf{U}}^{\alpha s}$ to represent (9) and (11):

$$\delta\underline{\mathbf{x}} = \underline{\mathbf{U}}^{\alpha s}\mathbf{v}^{\alpha s}. \tag{13}$$

The adjoint transform appears in the equation for the gradient:

$$\left[\frac{\partial J}{\partial \mathbf{v}^{\alpha s}}\right] = \mathbf{v}^{\alpha s} + \underline{\mathbf{U}}^{\alpha s\mathrm{T}}\underline{\mathbf{H}}^{\mathrm{T}}\underline{\mathbf{R}}^{-1}[\underline{\mathbf{H}}\delta\underline{\mathbf{x}} - \underline{\mathbf{d}}]. \tag{14}$$

In our experiments, as in most other implementations of 4DEnVar, we have enforced total smoothness of $\underline{\boldsymbol{\alpha}}_j$ in time, which means there is no explicit localization of the correlations in time. This removes any time dimension from $\mathbf{v}^{\alpha}_j$; the transform $\underline{\mathbf{U}}^{\alpha}$ replicates a single 3D field $\boldsymbol{\alpha}_j$ to give the 4D $\underline{\boldsymbol{\alpha}}_j$.

The localization technique applied in 4DEnVar (the 'alpha control variable') at each timestep is equivalent to the localized ensemble covariance of an EnKF (Lorenc, 2003b; Wang *et al.*, 2007). The EnKF notation is simplified by defining a rectangular $n \times m$ matrix $\mathbf{X}^{\mathrm{b}}$ from m column vectors $\delta\mathbf{x}^{\mathrm{b}}_j$:

$$\mathbf{X}^{\mathrm{b}} = \frac{1}{\sqrt{m-1}}\left[\delta\mathbf{x}^{\mathrm{b}}_1 \quad \ldots \quad \delta\mathbf{x}^{\mathrm{b}}_m\right]. \tag{15}$$

If the ensemble members are randomly sampled from a distribution with covariance $\mathbf{P}^{\mathrm{b}}$, then a noisy unbiased estimate of the covariance is

$$\mathbf{P}^{\mathrm{b}}_{\mathrm{ens}} = \mathbf{X}^{\mathrm{b}}(\mathbf{X}^{\mathrm{b}})^{\mathrm{T}}. \tag{16}$$

Localization is applied to the EnKF as a Schur product with $\mathbf{P}^{\mathrm{b}}_{\mathrm{ens}}$ (Hamill *et al.*, 2001):

$$\mathbf{P}^{\mathrm{b}} = \{\mathbf{X}^{\mathrm{b}}(\mathbf{X}^{\mathrm{b}})^{\mathrm{T}}\} \circ \mathbf{C}. \tag{17}$$

Lorenc (2003b) showed that this is equivalent to using (9) and then constraining the $\boldsymbol{\alpha}$s to be smooth using the same $\mathbf{C}$ in (10). In 4DEnVar we use the ensemble trajectories over the time window in 4D versions of (15), (16) and (17). This gives a 4D background-error covariance $\{\underline{\mathbf{X}}^{\mathrm{b}}(\underline{\mathbf{X}}^{\mathrm{b}})^{\mathrm{T}} \circ \underline{\mathbf{C}}\}$ which replaces that used in 4DVar ($\underline{\mathbf{M}}\mathbf{B}\underline{\mathbf{M}}^{\mathrm{T}}$).

2.3. EDA

The EDA method presented in this article is an ensemble of 4DEnVars. This set-up makes it yet another flavour of the ensemble Kalman filter. The analysis which we evaluate comes from the ensemble mean. The same ensemble perturbations are used to calculate the analysis of each ensemble member i.e. the same $\delta\mathbf{x}^{\mathrm{b}}_j$ are used in (9) for each cost function. But each analysis ensemble member has its own innovations, from (6) applied to that member's forecast.

The spread of the EDA will collapse without modification. The same problem was discovered for the EnKF (Burgers *et al.*, 1998), but it can be alleviated by randomly perturbing the observations. The pseudo-random observation perturbations are constructed such that they sum to zero over the ensemble. Sakov and Oke (2008) and Bowler *et al.* (2013) demonstrated an alternative to this stochastic formulation, which involves a deterministic analysis update step. This deterministic analysis update step is also used by the DEnKF algorithm (section 2.5). The analysis perturbations are halved:

$$\begin{aligned}\mathbf{x}^{\mathrm{a}}_j(t_0) \leftarrow \overline{\mathbf{x}}^{\mathrm{a}}(t_0) \quad &+ \quad \frac{1}{2}\{\mathbf{x}^{\mathrm{b}}_j(t_0) - \overline{\mathbf{x}}^{\mathrm{b}}(t_0)\} \\ &+ \quad \frac{1}{2}\{\mathbf{x}^{\mathrm{a}}_j(t_0) - \overline{\mathbf{x}}^{\mathrm{a}}(t_0)\}.\end{aligned} \tag{18}$$

Both the perturbed observations and the deterministic versions of the EDA will be compared. In this article, they will be referred to as the EDA-S and EDA-D respectively.

2.4. 4DVar-Ben

Our 4DVar-Ben method is identical to 4DVar, except it uses $\mathbf{P}^{\mathrm{b}}_{\mathrm{ens}}(t_0)$ from EDA-D instead of using the climatological $\mathbf{B}$. The 4DVar increment at the beginning of the window has the same equation as the 4DEnVar increment (9):

$$\delta\mathbf{x}(t_0) = \sum_{j=1}^{m} \frac{1}{\sqrt{m-1}} \delta\mathbf{x}^{\mathrm{b}}_j(t_0) \circ \boldsymbol{\alpha}_j. \tag{19}$$

However, localization is only applied at t_0 and the increment in (19) is propagated forwards in time using the tangent linear model (4). The purpose of 4DVar-Ben is to take advantage of the tangent linear model in 4DVar, which calculates an implicit 4D background-error covariance $\underline{\mathbf{M}}\mathbf{P}^{\mathrm{b}}_{\mathrm{ens}}\underline{\mathbf{M}}^{\mathrm{T}}$ with time correlations which are accurate to within the linear approximation of $\mathbf{M}$. 4DVar-Ben allows for 'errors of the day' by using $\mathbf{P}^{\mathrm{b}}_{\mathrm{ens}}(t_0)$ instead of the climatological $\mathbf{B}$ of 4DVar.

2.5. DEnKF

The deterministic Ensemble Kalman Filter (DEnKF) method was introduced by Sakov and Oke (2008). The DEnKF approximates a proper Ensemble Square Root Filter (EnSRF; Whitaker and Hamill, 2002) when the analysis correction for each window is small ($\mathbf{H}\mathbf{P}^{\mathrm{b}}\mathbf{H}^{\mathrm{T}} << \mathbf{R}$). The analysis ensemble mean is given by

$$\overline{\mathbf{x}}^{\mathrm{a}} = \overline{\mathbf{x}}^{\mathrm{b}} + \mathbf{K}\{\mathbf{y}^{\mathrm{o}} - \overline{H(\mathbf{x}^{\mathrm{b}})}\}. \tag{20}$$

The Kalman gain is calculated by

$$\mathbf{K} = (\mathbf{P}^{\mathrm{b}})^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}(\mathbf{H}\mathbf{P}^{\mathrm{b}}\mathbf{H}^{\mathrm{T}} + \mathbf{R})^{-1}. \tag{21}$$

We localize the ensemble background-error covariance matrix $\mathbf{P}^{\mathrm{b}}_{\mathrm{ens}}$ in observation space to get $\mathbf{P}^{\mathrm{b}}$. This is different to 4DEnVar, where the 'alpha control variable' implies that $\mathbf{P}^{\mathrm{b}}_{\mathrm{ens}}$ is localized in model space (17). In general, localizing in observation space is not the same as localizing in model space, since the observation operator does not commute:

$$\mathbf{H}\mathbf{C}\mathbf{H}^{\mathrm{T}} \circ \mathbf{H}\mathbf{P}^{\mathrm{b}}_{\mathrm{ens}}\mathbf{H}^{\mathrm{T}} \neq \mathbf{H}(\mathbf{C} \circ \mathbf{P}^{\mathrm{b}}_{\mathrm{ens}})\mathbf{H}^{\mathrm{T}}. \tag{22}$$

However, in our experiments the observations fall on model gridpoints, which means the left- and right-hand sides of (22) are identical. Therefore our localization in observation space is equivalent to using (17) in (21). We note that there is no theoretical justification for localization in observation space – EnKF methods do it for technical reasons. 4DEnVar is theoretically correct to localize in model space.

The DEnKF maintains the ensemble spread by halving the Kalman gain matrix in the analysis perturbation update step:

$$\delta\mathbf{x}^{\mathrm{a}}_j = \delta\mathbf{x}^{\mathrm{b}}_j - \frac{1}{2}\mathbf{K}H(\delta\mathbf{x}^{\mathrm{b}}_j). \tag{23}$$

This is equivalent to halving the influence of the analysis perturbations (as used in (18) by EDA-D). The ensemble mean remains equivalent to (20). We formulate the DEnKF to be 4D within the assimilation window (Hunt *et al.*, 2004, give details), which makes it theoretically equivalent to EDA-D.

3. Methodology

3.1. The toy model

The Lorenz-05 model II (Lorenz, 2005) is a 1D cyclic system and represents the propagation of waves, of some unspecified atmospheric quantity, around a latitude circle (see Appendix). It is an extension of the Lorenz-95 model to incorporate smooth spatial correlations. The parameter K in (A1) controls the wave length and the amount of correlation between grid points. For the perfect model experiments:

- The number of gridpoints of the forecast model and the truth model are equal to 180 ($n_{\mathrm{truth}} = n_{\mathrm{forecast}} = 180$);
- The parameter K is set equal to 6 ($K_{\mathrm{truth}} = K_{\mathrm{forecast}} = 6$).

Model error is introduced through the truth model being at a higher resolution than the DA model. For the imperfect model experiments:

- The number of gridpoints $n_{\text{truth}} = 240$ and $n_{\text{forecast}} = 180$;
- The parameter K has values $K_{\text{truth}} = 8$ and $K_{\text{forecast}} = 6$.

For both the perfect and imperfect forecast model experiments, the ratio $n_{\text{truth}}/K_{\text{truth}} = n_{\text{forecast}}/K_{\text{forecast}}$, so the wavelengths of the truth and DA models stay approximately equivalent (Lorenz, 2005, gives details). The forcing F in (A1) is set equal to 15 for both models, which corresponds to an error doubling time of about 2 days. In all the experiments, the models are integrated using a fourth-order Runge–Kutta method.

3.2. *Experimental parameters*

All the DA methods use an integration timestep length of 0.1 time units and the assimilation window consists of five integration steps, giving a length of 0.5 time units, which is equivalent to about 6 h in the real atmosphere. We denote the initial value as timestep 1 and the last value as timestep 6. Most operational centres currently use 6 h long assimilation windows for 4DVar, which is a good length when model error is unaccounted for. In all our experiments, the observations come from the truth model, but the observations are randomly selected to be at the forecast model gridpoints. This is straightforward for the perfect model since $n_{\text{truth}} = n_{\text{forecast}}$. For the imperfect model experiments, where $n_{\text{truth}} > n_{\text{forecast}}$ the observation values are interpolated using cubic interpolation. The same pseudo-random errors are generated for the observations for each method. The observation error is uncorrelated and drawn from a Gaussian distribution with standard deviation 0.1 ($\mathbf{R} = 0.01\mathbf{I}$). The magnitude of the observation error was calculated to be much greater than the representivity error from the cubic interpolation. Hence the representivity error is neglected in $\mathbf{R}$.

We test the five DA methods by varying the spatial and temporal density of the observations through the assimilation window. The total number of observations in the assimilation window is defined as n_{obs}, with n_{obs} a value within the set: $n_{\text{obs}} \in \{25, 50, 75, 100, 125, 150\}$. The point of changing the observation densities is to simulate differences in conventional observation densities that exist between different regions of the globe. We then test the temporal distribution of these observations using two scenarios:

1. Observations only at the beginning of the window (timestep 1);
2. Observations evenly distributed over timesteps 2, 3, 4, 5, 6 in the assimilation window.

The first scenario means the 4D background-error covariance is irrelevant; all that matters is the 3D covariance at the time of the observations. The second scenario is a test of the full 4D covariances implicit in each method.

The EDA-D, EDA-S, 4DVar-Ben and DEnKF are all tested for ensemble sizes $n_{\text{ens}} \in \{3, 4, 5, 9, 13, 17, 25, 50, 75, 100, 125, 150\}$. The smaller ensemble sizes suffer significantly from sampling error, since the ensemble size is far less than the forecast model dimension. Localization and fixed multiplicative covariance inflation are applied in order to reduce the effects of sampling error; they are tuned to produce the smallest analysis errors. The localization is described in section 3.4. The multiplicative inflation is described in section 3.5. Additive inflation is also used for the model error experiments and is sampled from the true model error (section 3.6 gives details). The ensemble methods were tuned for each ensemble size whenever the spatial or temporal distribution of the observations was changed. The tuned parameters for the DEnKF were used in EDA-D, since the methods are theoretically equivalent.

The analysis errors are measured against the truth after the truth has been spectrally interpolated to the DA resolution. In both scenarios the analysis error is taken at the mean time of the observations; timestep 1 for scenario 1 and timestep 4 for scenario 2. The errors are averaged over the last 3000 windows of a 4000 window run. All the methods use the analysed increments at the same timestep as the analysis errors were taken.

3.3. *Tuning the climatological* **B**

Since we had available the 'truth', the 4DVar climatological background-error covariance matrix, **B**, could be calculated directly from the background forecasts. However, since the accuracy of these depended on the **B** used, the calculation had to be bootstrapped. We started by using a **B** calculated from the time-averaged EDA-D forecast errors and we enforced homogeneity by averaging the covariances. Using this **B**, we ran 4DVar for 4000 cycles and calculated a new **B** from the last 3000 cycles. The process was repeated until the analysis errors converged to three significant figures. Using this method, a different **B** was calculated whenever the spatial or temporal distribution of the observations was changed.

3.4. *Localization*

We followed other EnKF methods (Hamill *et al.*, 2001) and localized the ensemble covariances using a near-Gaussian function which sets them to zero beyond a certain distance (Gaspari and Cohn, 1994, Eq. (4.10)) and increases the rank of the background-error covariance matrix. The Gaspari–Cohn (GC) half-width is normally quoted as the length-scale and is defined as half this distance.

The localization matrix **C** is required to be positive semi-definite in order that 4DEnVar can use the control variable transform (11). We found that the GC function using large gridpoint distances between points may not be positive semi-definite. We follow the method of Bowler *et al.* (2013) by placing the Lorenz model on a circle and calculate the GC function using the Euclidean distance d between gridpoints:

$$d = \frac{n}{\pi} \sin\left\{\frac{\pi}{n}(i-j)\right\}, \tag{24}$$

where i and j are the gridpoints and n is the dimension. This ensures that **C** remains positive semi-definite. The GC half-width was tuned using a range of values: 0.5, 1.0,. . . , 10 (in steps of 0.5), 10, 15,. . . , 120 (in steps of 5). For the largest ensemble size and the perfect model, the ensemble methods did not require localization. The smallest GC half-width used in the experiments is 6, while the largest value used is 115.

We demonstrate here the effect of these localization length-scales by localizing the climatological background-error covariance matrix (**B**) of 4DVar with a perfect model and 150 observations at the start of the window. We localize

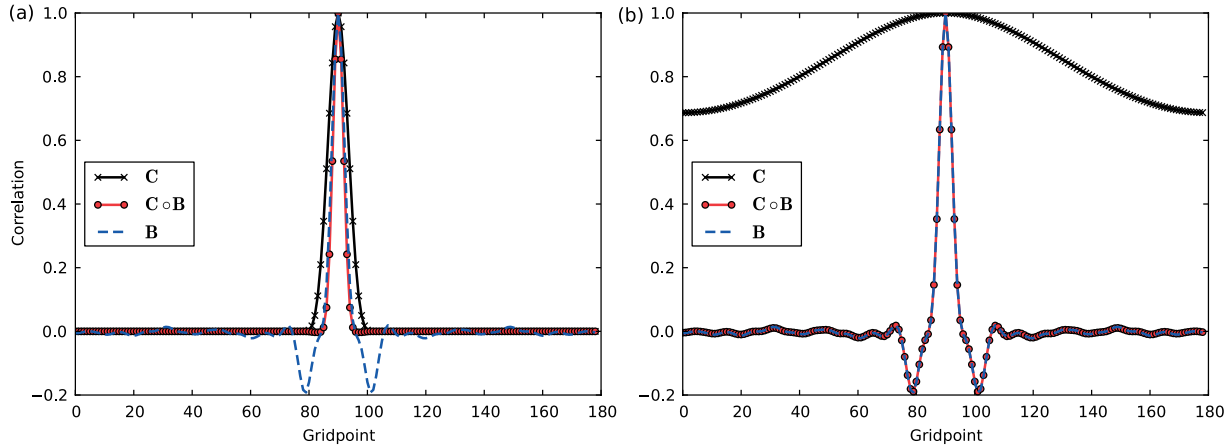

Figure 1. 90th gridpoint correlations of **C**, **C** ∘ **B** and **B** for (a) GC half-width of 6; here **C** ∘ **B** is similar to **B** only near the origin, and elsewhere is it similar to **C**; and (b) GC halfwidth of 115; here **C** ∘ **B** is similar to **B** everywhere. A range of localizations between these values are used in the experiments. This figure is available in colour online at wileyonlinelibrary.com/journal/qj

B using the Schur product **B** ∘ **C**. Figure 1(a) shows the correlations at the 90th gridpoint of **C**, **B** and **B** ∘ **C** for a GC half-width of 6 and Figure 1(b) for a GC half-width of 115.

With a GC half-width of 6, it is only near the origin that the correlations are similar to **B**; elsewhere they are like the GC correlation function. On the other extreme, with a GC half-width of 115, the **C** given by (24) is everywhere near 1, so the correlations are approximately the same as the background-error correlations. Most of the time, the ensemble methods required localization somewhere between these values.

3.5. *Multiplicative inflation*

Having insufficient ensemble members often causes the ensemble spread to be underestimated. In operational practice, the analysis ensemble perturbations are inflated in order to achieve a better spread. A fixed multiplicative covariance inflation can be applied by linearly inflating the analysis ensemble perturbations by a factor γ (Anderson and Anderson, 1999):

$$\mathbf{X}^{\mathrm{a}}_j = \gamma(\mathbf{x}^{\mathrm{a}}_j - \overline{\mathbf{x}}^{\mathrm{a}}). \tag{25}$$

Fixed covariance inflation was applied to all the ensemble methods. The factor γ was selected from the range: 1.00, 1.01,. . . , 1.20 (in steps of 0.01), so as to minimize the analysis error.

3.6. *Additive inflation*

When model error is present, then multiplicative inflation alone may be inadequate. Although multiplicative inflation increases the ensemble spread, it does nothing to capture the structure of the model-error covariance matrix **Q**. We follow Mitchell *et al.* (2002) and use additive inflation to add new directions to the ensemble perturbations. This consists of adding noise $\boldsymbol{\eta}_j$ to the analysis ensemble member perturbations $\delta\mathbf{x}^{\mathrm{a}}_j$ at the end of the window, before the update of the next cycle:

$$\delta\mathbf{x}^{\mathrm{a}}_j \leftarrow \delta\mathbf{x}^{\mathrm{a}}_j + \boldsymbol{\eta}_j. \tag{26}$$

The noise $\boldsymbol{\eta}_j$ is randomly sampled from the true **Q** ($\boldsymbol{\eta} \in \mathbf{N}(0, \mathbf{Q})$), which avoids the need to inflate it. Note that using additive inflation also causes sampling error, which is alleviated by a large ensemble.

Q comes from the climatological model errors, which are a result of the forecast model being at a smaller dimension (180 gridpoints) than the truth model (240 gridpoints). The model error is computed from the difference between the truth trajectory, spectrally interpolated to forecast model resolution, and 6 h forecasts using the forecast model. **Q** is generated from the time-averaged covariance of differences over the 3000 window run and homogeneity is enforced.

Table 2. The tuned localizations and fixed inflations for EDA-D and various ensemble sizes. The total number of observations used is 25 with five observations (at random locations) at the five timesteps 2, 3,. . . , 6. Tuning is performed to produce the lowest analysis error and the ∞ symbol for the GC half-width denotes no localization.

Ensemble size	GC half-width	Covariance inflation	RMS error	RMS spread
3	7	1.08	0.0813	0.0730
4	20	1.10	0.0522	0.0524
5	30	1.06	0.0418	0.0371
9	75	1.08	0.0281	0.0354
13	105	1.02	0.0213	0.0248
17	115	1.01	0.0197	0.0224
25	∞	1.01	0.0179	0.0222
150	∞	1.00	0.0181	0.0194

4. Perfect model experiments

4.1. *The effect of ensemble size and observation density*

Table 2 shows the tuned localization and inflation coefficients, and the RMS analysis error for different ensemble sizes. Table 3 shows the same quantities for different observation densities. As expected, a larger ensemble and more observations lead to lower errors.

It is worth noting that in Table 2 there is no advantage in going above a 25 ensemble member. We also note that, by comparing Tables 2 and 3, the (optimally tuned) localization length-scale and covariance inflation factor are strongly dependent on the ensemble size and not on the number of observations. Finally we note that the analysis spread is typically not equal to the analysis error – they should be if we

Table 3. The tuned localizations and fixed inflations for a range of observation densities. An observation density of 150 represents 30 observations taken at random spatial locations at five different timesteps. The EDA-D method is used with a relatively small ensemble size of five.

Obs. density	GC half-width	Covariance inflation	RMS error	RMS spread
25	30	1.06	0.0418	0.0371
50	30	1.07	0.0294	0.0264
75	30	1.07	0.0272	0.0213
100	30	1.07	0.0222	0.0183
125	30	1.06	0.0174	0.0154
150	30	1.08	0.0166	0.0156

were using a linear model and a perfect ensemble Kalman filter. In certain cases the ensemble is slightly underspread compared to the analysis error and in other cases it is overspread.

4.2. *Comparing EDA-D with 4DVar*

Figure 2 plots the RMS analysis errors of EDA-D against 4DVar for different observation densities. Figure 2(a) is the case when all observations are at the start of the assimilation window and Figure 2(b) is the case when the observations are distributed over five different timesteps.

Since EDA-D is using an ensemble-based covariance matrix, then for a large enough ensemble we would expect it to beat 4DVar. This occurs in Figure 2(a) for a relatively small ensemble of four. The situation does change however when the observations are distributed over time, as is the case in Figure 2(b). In this situation the 4D covariance is important; that implicit in 4DVar (propagated using the linear model and its adjoint) is better than the localized 4D covariance of EDA-D. The localization of the EDA-D background-error covariance introduces errors because the Schur product and the nonlinear model do not commute. So a bigger ensemble (five in this case), with less localization, is needed before EDA-D beats 4DVar. In the next section, where we consider model error, the 4DVar method becomes significantly more competitive.

4.3. *Comparing EDA-D with 4DVar-Ben*

The only theoretical difference between EDA-D and 4DVar-Ben is the way in which the 4D background-error covariance matrix is propagated through time. We demonstrate this theoretical difference in Figure 3 for the case again when observations are either distributed all at the start or evenly in time during the assimilation window.

When all the observations are at the start of the window, Figures 3(a, c) show exact agreement between the two methods of EDA-D and 4DVar-Ben. Note that both of these methods are using the same ensemble-based background-error covariance matrix at the start of the window. When the observations are distributed in time then, as in Figure 2, 4DVar-Ben makes better use of the time-distributed observations. Again, the results can be explained by the fact that 4DVar-Ben does not suffer from the same localization issues as EDA-D. 4DVar-Ben localizes the initial background-error covariance matrix $\mathbf{P}_e^{\mathrm{b}}(t_0)$ and then propagates this forwards in time using the tangent linear model (4). Provided the tangent linear model is a good representation of the nonlinear model within the assimilation window, the 4D structure of the background-error covariance matrix will not be degraded. We show 4DVar in Figure 3 since it also benefits from the use of the tangent linear model when the observations are distributed in time.

In order to be certain that the differences between EDA-D and 4DVar-Ben are indeed due to localization and nothing else, we compare in Table 4 EDA-D with 4DVar-Ben for a large ensemble of 50 members. In this table, severe localization (with a GC half-width of 5) is switched on and off and evidently the difference between the methods exists when the severe localization is there, but the methods are almost identical when localization is not present. We note in Table 4 that the EDA-D method and DEnKF method give similar RMS errors and this is discussed next.

4.4. *Comparing the deterministic and stochastic ensemble methods*

Finally, 4DVar, EDA-D and EDA-S are compared with the DEnKF. Note that the DEnKF and EDA-D are theoretically

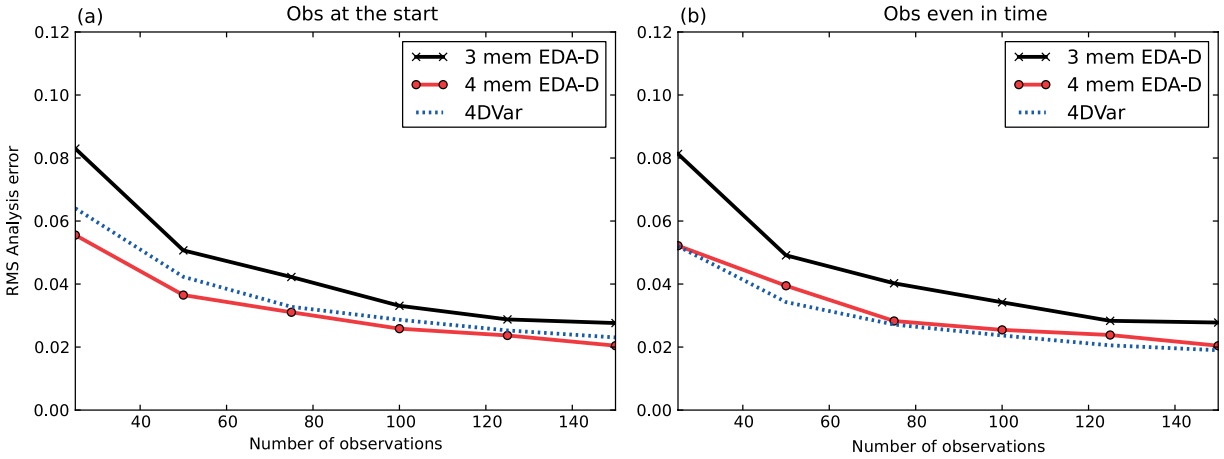

Figure 2. (a) has all observations at the start of the window whilst (b) has observations at the five timesteps $2, 3, \ldots, 6$. The perfect model is used and it can be seen that 4DVar makes a large improvement in the analysis errors in (b) over (a). On the other hand, EDA-D makes very little use of the time-distributed observations, due to the severe localization being used in the EDA-D method at all timesteps during the assimilation window. This figure is available in colour online at wileyonlinelibrary.com/journal/qj

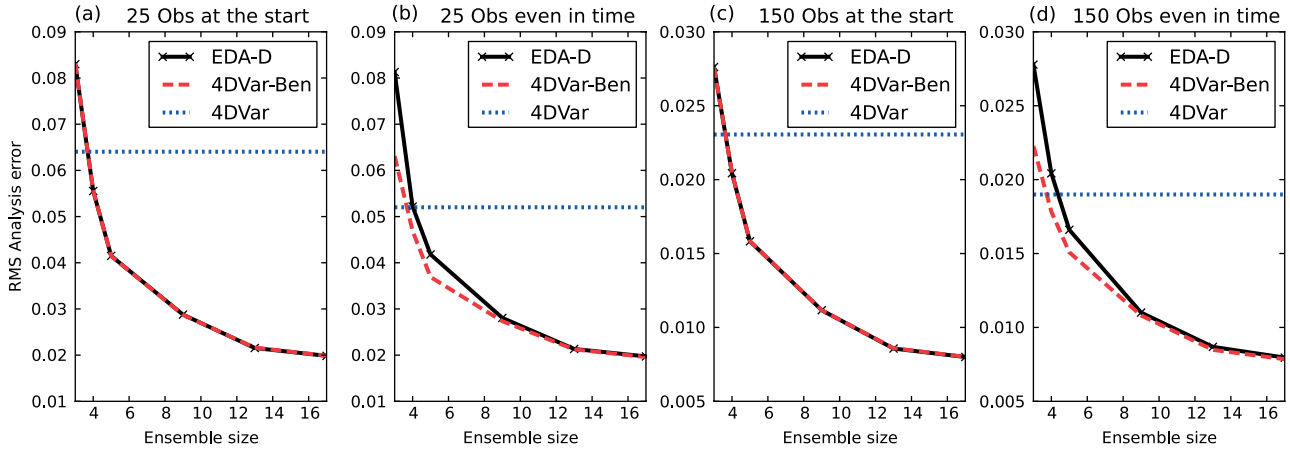

Figure 3. RMS analysis errors for 25/150 observations at the start of the window and distributed evenly over the five timesteps 2, 3,. . . , 6, using the perfect model. EDA-D and 4DVar-Ben are equivalent when observations are all at the start, since they both use the same 3D ensemble covariance matrix. 4DVar-Ben and 4DVar improve significantly when observations are changed from being all at the start to being distributed in time. On the other hand, EDA-D makes very little use of the time-distributed observations when severe localization is applied (i.e. for a small ensemble). This figure is available in colour online at wileyonlinelibrary.com/journal/qj

Table 4. RMS analysis errors for 25 and 150 observations distributed evenly over the 5 timesteps 2,3,. . . ,6 for the perfect model. Severe localization is applied by using a GC halfwidth of 5. This is compared with the unlocalized case. It is evident that EDA-D and 4DVar-Ben are similar when no localization is applied, but that 4DVar-Ben performs best when localization is applied.

Obs. density	GC half-width	RMS error		
		EDA-D	4DVar-Ben	DEnKF
25	5	0.0770	0.0657	0.0776
25	∞	0.0179	0.0180	0.0179
150	5	0.0315	0.0242	0.0315
150	∞	0.00682	0.00683	0.00682

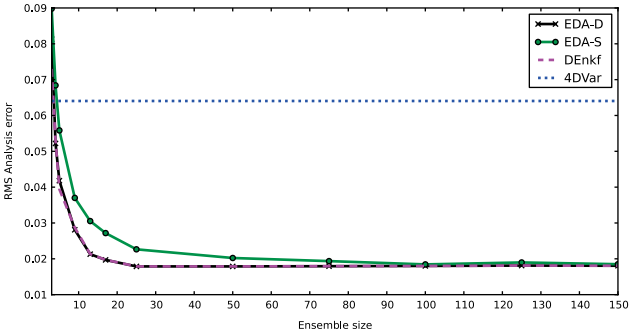

Figure 4. Comparisons beteen EDA-D, EDA-S, the DEnKF and 4DVar for five observations at the five timesteps 2, 3,. . . , 6 using the perfect model. As expected, EDA-D and the DEnKF are nearly always equivalent. These deterministic methods perform better than EDA-S due to the sampling error introduced from perturbed observations. This figure is available in colour online at wileyonlinelibrary.com/journal/qj

equivalent. The only difference between EDA-S and EDA-D is that EDA-S perturbs the observations to maintain the ensemble spread, while EDA-D uses a deterministic analysis perturbation update step. Figure 4 shows RMS analysis error against ensemble size for these methods with 25 observations evenly distributed in time.

As expected, the analysis errors for EDA-D and the DEnKF are statistically indistinguishable most of the time*

For a small ensemble, both the DEnKF and EDA-D perform significantly better than EDA-S, since EDA-S is affected by sampling error from the perturbed observations. EDA-S also requires more severe localization and larger fixed covariance inflation than EDA-D. These results agree with experiments by Whitaker and Hamill (2002) and Sakov and Oke (2008), who also demonstrated that sampling error was introduced from perturbed observations. The DEnKF approximates a proper EnSRF when the analysis correction is small ($\mathbf{HP}^{\mathrm{b}}\mathbf{H}^{\mathrm{T}} << \mathbf{R}$) (Sakov and Oke, 2008). It was found that generally $\mathbf{HP}^{\mathrm{b}}\mathbf{H}^{\mathrm{T}} << \mathbf{R}$ for the perfect model, especially when the ensemble size was large. Hence the DEnKF and EDA-D would be expected to behave like an EnSRF.

*The difference between the DEnKF and EDA-D for the small ensembles was found to be due to the 'chaotic attractor effect'. Small differences in the computational round-off errors (of order 10^{-8}) cause the DEnKF and EDA-D to diverge to slightly different solutions over the length of the run (4000 windows).

All the ensemble methods perform significantly better than 4DVar for a large enough ensemble, due to the benefit of a flow-dependent background-error covariance matrix.

5. Imperfect model

5.1. *The effect of ensemble size and observation density*

All the experiments in section 4 are now repeated for the imperfect model, where the model error results from the forecast model being at a smaller dimension (180 gridpoints) than the truth model (240 gridpoints). Since model error is present, we now give ensemble results where additive inflation is either used or not used. The form of additive inflation which we use is described in section 3.6. Tables 5 and 6 give results without and with additive inflation respectively. These Tables can be compared against Table 2 for the perfect model and the first point to notice is the increase in analysis errors. We also note that increasing the ensemble size in the case of no additive inflation has a much less significant effect on the analysis errors than when additive inflation is used. The effectiveness of the additive inflation appears to increase with ensemble size and this is

Table 5. The tuned localizations and fixed inflations for EDA-D and various ensemble sizes for the imperfect model. The total number of observations used is 25 with five observations at the five timesteps $2, 3, \ldots, 6$. Analysis errors are larger and localization more severe for the imperfect model.

Ensemble size	GC half-width	Covariance inflation	RMS error	RMS spread
5	9	1.08	0.0898	0.0800
9	15	1.08	0.0810	0.0801
13	15	1.08	0.0786	0.0832
17	20	1.08	0.0762	0.0783
25	35	1.08	0.0752	0.0674
150	35	1.07	0.0732	0.0778

Table 6. As Table 5, but the EDA-D has additive inflation, which is less effective for a small ensemble, due to sampling error.

Ensemble size	GC half-width	Covariance inflation	RMS error	RMS spread
5	9	1.08	0.0883	0.0801
9	15	1.08	0.0771	0.0739
13	20	1.07	0.0720	0.0670
17	20	1.07	0.0710	0.0709
25	35	1.07	0.0693	0.0775
150	35	1.05	0.0649	0.0642

Table 7. Tuned localizations and fixed inflations for a range of observation densities for the imperfect model. An observation density of 150 represents 30 observations taken at random spatial locations at five different timesteps. The EDA-D method is used with a large ensemble size of 150. Note that larger observation densities requires more severe localization, unlike the perfect model results, Table 3.

Obs. density	GC half-width	Covariance inflation	RMS error	RMS spread
25	35	1.07	0.0732	0.0778
50	20	1.09	0.0605	0.0728
75	15	1.11	0.0548	0.0721
100	15	1.11	0.0504	0.0591
125	15	1.12	0.0474	0.0559
150	10	1.12	0.0451	0.0518

Table 8. As Table 7, but with additive inflation, which significantly reduces the analysis errors.

Obs. density	GC half-width	Covariance inflation	RMS error	RMS spread
25	35	1.05	0.0649	0.0642
50	25	1.06	0.0532	0.0538
75	25	1.06	0.0472	0.0449
100	20	1.06	0.0439	0.0407
125	20	1.07	0.0416	0.0401
150	20	1.07	0.0396	0.0372

likely due to sampling error issues (of the additive inflation) for small ensemble sizes.

We now look at Tables 7 and 8 where the observation density is changed; Table 7 (8) gives results of the case without (with) additive inflation. A 150-member ensemble is chosen in both Tables to alleviate any sampling error issues of the additive inflation. An effect which we see in Table 7 is that when additive inflation is not used, the (optimally tuned) localization and covariance inflation are strongly dependent on the observation density, more dense observations benefiting from more severe localization and larger covariance inflation. This effect is much smaller with additive inflation (Table 8) and was not seen at all for the perfect model (Table 3). We will link the effect of model error and the structure of the ensemble background-error covariances in the next section 5.2 when we compare EDA-D with 4DVar.

5.2. *Comparing EDA-D with 4DVar*

We now compare EDA-D with 4DVar for various observation densities. Figure 5 is the imperfect model equivalent of Figure 2 but showing only the 9 ensemble member EDA-D with and without additive inflation. Figure 6 is with a larger ensemble of 150, where sampling error is less significant. 4DVar improves significantly when the observations are distributed over time compared to when they are all at the start of the window, while EDA-D shows no improvement. As was explained for the perfect model,

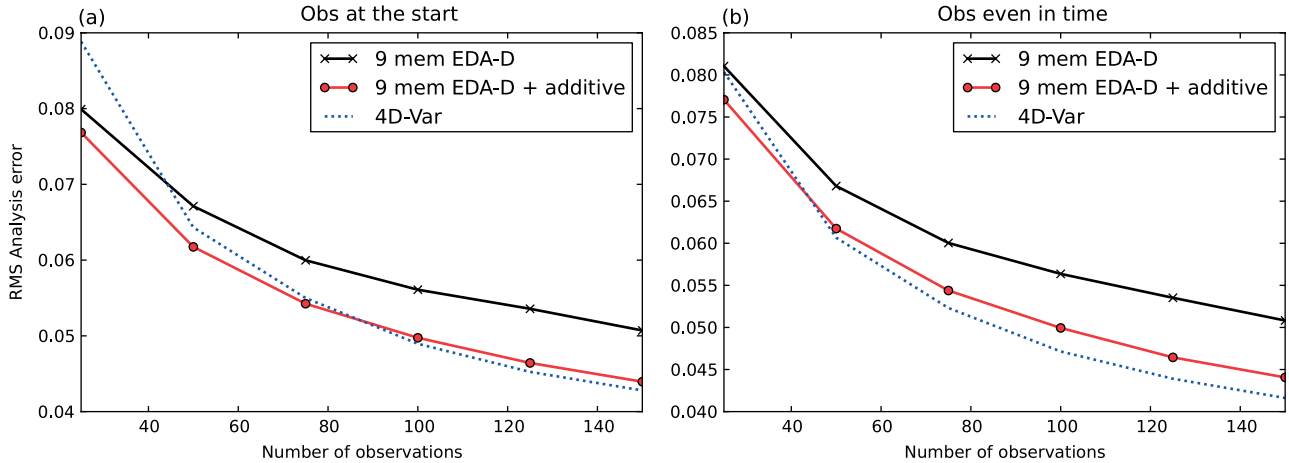

Figure 5. The imperfect model results from (a) all observations at the start of the window and (b) observations at the five timesteps $2, 3, \ldots, 6$. Comparison with the perfect model results (Figure 2) shows two important effects that the model error has on the results: 4DVar handles the model error better than EDA-D, and 4DVar improves more rapidly than EDA-D for denser observations. EDA-D with additive inflation reduces both these effects because it can represent the climatological model-error covariance matrix **Q**. This figure is available in colour online at wileyonlinelibrary.com/journal/qj

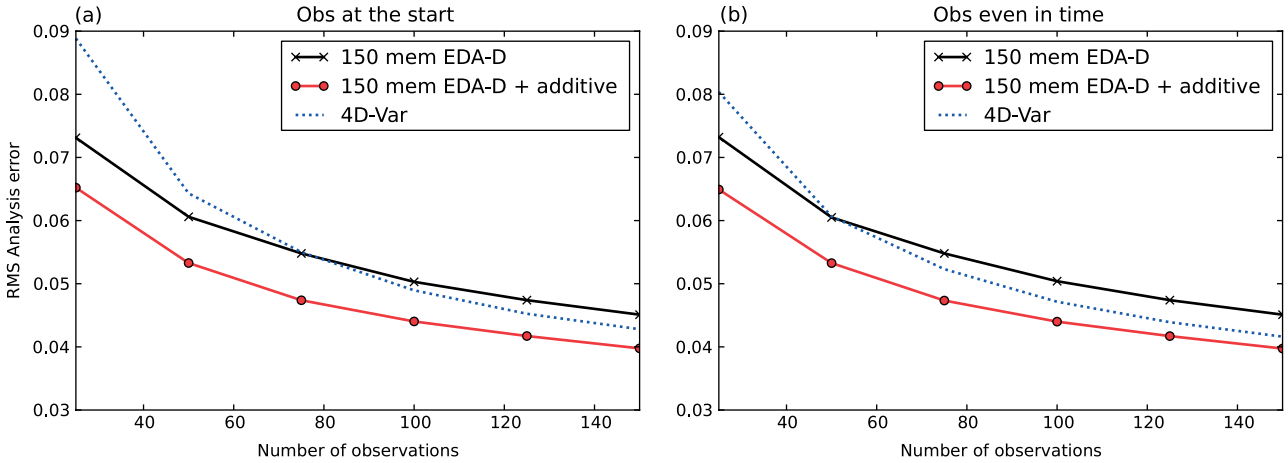

Figure 6. As Figure 5, but for a larger ensemble of 150, which means that sampling error is less significant. This figure is available in colour online at wileyonlinelibrary.com/journal/qj

this is due to the localization degrading the 4D structure of the EDA-D background-error covariance matrix.

Two important differences between the perfect and the imperfect model results are firstly, 4DVar is more competitive with model error and secondly, 4DVar improves more rapidly than EDA-D as the observation density is increased. These effects are reduced significantly by additive inflation. We believe three interacting effects can explain these results:

1. Kalman filter theory says that the true background-error covariance is $\mathbf{M}\mathbf{P}^{a}\mathbf{M}^{T} + \mathbf{Q}$, where $\mathbf{P}^{a}$ measures the previous analysis-error covariance. A positive feedback exists, where increases in $\mathbf{Q}$ cause even larger increases in analysis and background errors;
2. In our experiments, 4DVar has a 'perfect' climatological covariance $\mathbf{B}$, which we construct to represent $\mathbf{M}\mathbf{P}^{a}\mathbf{M}^{T} + \mathbf{Q}$. On the other hand the EDA-D (without additive inflation) uses ensemble perturbations to construct an inflated first term $\mathbf{M}\mathbf{P}^{a}\mathbf{M}^{T}$. The EDA-D with additive inflation approximates both terms $\mathbf{M}\mathbf{P}^{a}\mathbf{M}^{T} + \mathbf{Q}$. The second term is sampled from the climatological model-error covariance;
3. A common result is that, if an assimilation scheme is using an incorrect background-error covariance then it is usually improved by increased localization, especially when there are sufficient observations. This increased localization degrades the 4DEnVar used in EDA-D when observations are distributed in time.

Evidence of the above three points can be found in the covariance matrices themselves. Firstly we look at point 1 in the above list. Figure 7 shows the 90th row of $\mathbf{Q}$ for gridpoints 50 to 130. The size of the $\mathbf{Q}$ variance is 0.00013. Adding $\mathbf{Q}$ to $\mathbf{M}\mathbf{P}^{a}\mathbf{M}^{T}$ will increase the background errors, which will then increase the analysis errors in the next cycle. This positive feedback means that introducing model error will increase the background errors more than the sum of the individual contributions of $\mathbf{Q}$ and the perfect model $\mathbf{M}\mathbf{P}^{a}\mathbf{M}^{T}$. For example, the 4DVar analysis error with 25 observations even in time and with a perfect model is 0.064 and with an imperfect model is 0.080. The difference between the squares (i.e. the variances) is 0.0023, which is more than 10 times the $\mathbf{Q}$ variance.

In order to demonstrate points 2 and 3, we investigate how the climatological $\mathbf{B}$ and ensemble $\mathbf{P}^{b}$ background-error

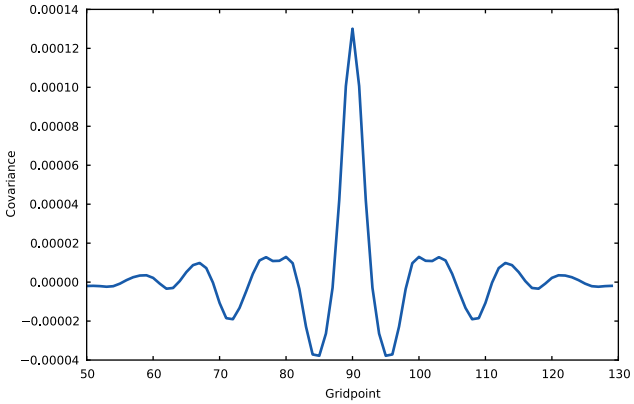

Figure 7. Plot of the model-error covariance matrix $\mathbf{Q}$. This figure is available in colour online at wileyonlinelibrary.com/journal/qj

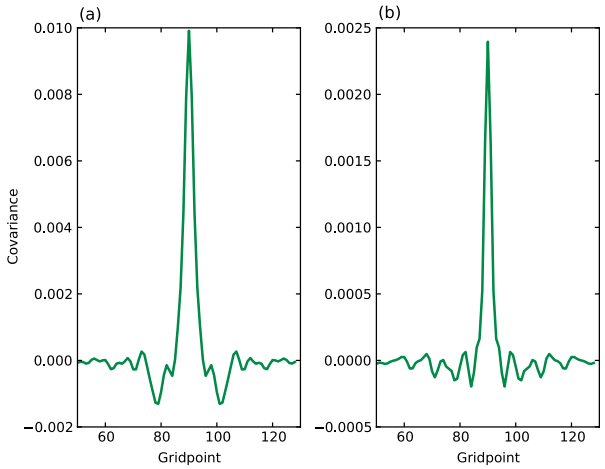

Figure 8. $\mathbf{B}$ for (a) 25 observations, and (b) 150 observations. All the observations are at the start of the window. As the observation density increases, $\mathbf{B}$ becomes more similar to $\mathbf{Q}$ (Figure 7). This figure is available in colour online at wileyonlinelibrary.com/journal/qj

covariance matrices are affected by model error and by the observation density.

Figures 8(a, b) show the 4DVar $\mathbf{B}$ for 25 and 150 observations respectively at the beginning of the window. As the observation density is increased, $\mathbf{B}$ tends to $\mathbf{Q}$ (Figure 7).

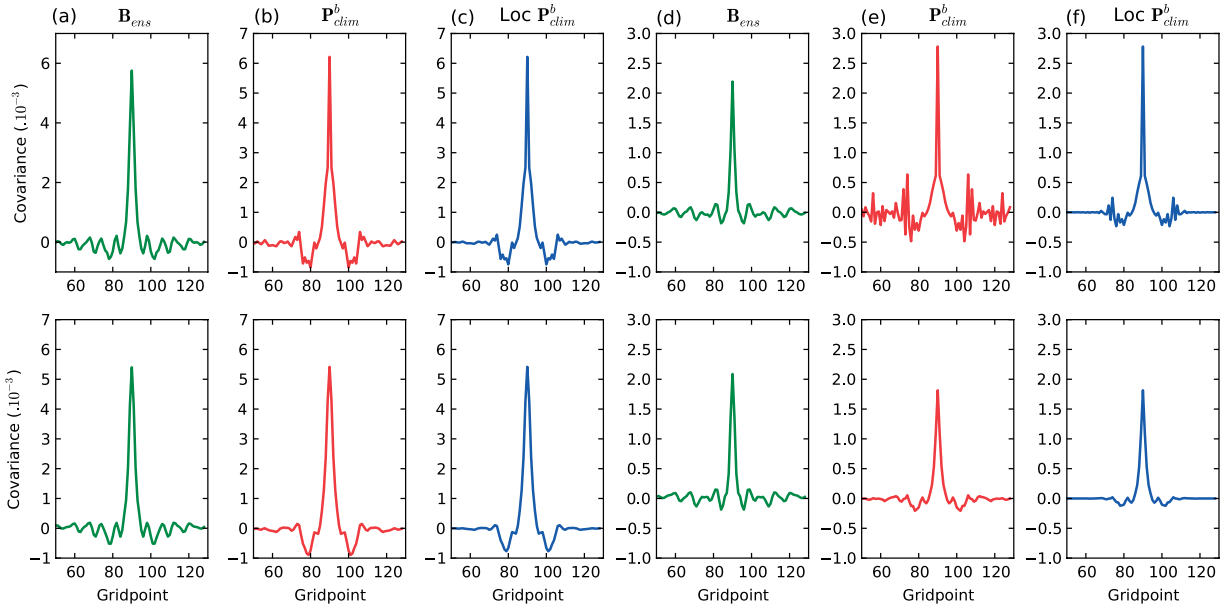

Figure 9. Top: without additive inflation, (a) $\mathbf{B}_{\text{ens}}$, (b) $\mathbf{P}^{\text{b}}_{\text{clim}}$, and (c) localized $\mathbf{P}^{\text{b}}_{\text{clim}}$. These have 25 observations at the start of the window. (d–f) are as (a–c), but have 150 observations. Bottom: as top but with additive inflation. If $\mathbf{P}^{\text{b}}$ is accurate then $\mathbf{P}^{\text{b}}_{\text{clim}}$ should look like $\mathbf{B}_{\text{ens}}$. Additive inflation significantly improves the $\mathbf{P}^{\text{b}}_{\text{clim}}$ structure with dense observations, although some localization is still beneficial. This figure is available in colour online at wileyonlinelibrary.com/journal/qj

Over the range of observation densities we are looking at, $\mathbf{B}$ never actually looks like $\mathbf{Q}$, since other background errors are affecting the covariances. However, the shape of $\mathbf{B}$ changes significantly between 25 and 150 observations, to be more like $\mathbf{Q}$.

We look at the EDA-D $\mathbf{P}^{\text{b}}$ with and without additive inflation and compare it with the true background-error covariance. We generate the climatological ensemble covariance matrix $\mathbf{P}^{\text{b}}_{\text{clim}}$ by averaging the 150 member EDA-D $\mathbf{P}^{\text{b}}$ (before inflation and localization) over 3000 windows and then averaging the covariances. The true climatological ensemble background-error covariance $\mathbf{B}_{\text{ens}}$ comes from the difference between the ensemble mean background and the truth, and we average this matrix over the same 3000 windows.

Figures 9(a, b, c) show $\mathbf{B}_{\text{ens}}$, $\mathbf{P}^{\text{b}}_{\text{clim}}$ and the localized $\mathbf{P}^{\text{b}}_{\text{clim}}$ respectively, with 25 observations at the start of the window. We start with the top half, where additive inflation has not been used. It is evident that $\mathbf{P}^{\text{b}}_{\text{clim}}$ is quite different to $\mathbf{B}_{\text{ens}}$. Some localization is therefore beneficial. Increasing the observation density to 150 (Figures 9(d, e, f)) makes the $\mathbf{P}^{\text{b}}_{\text{clim}}$ correlations diverge further from the true correlations, and hence shows why severe localization is beneficial. The EDA-D $\mathbf{P}^{\text{b}}$ can only measure an inflated $\mathbf{M}\mathbf{P}^{\text{a}}\mathbf{M}^{\text{T}}$. As the number of observations increases, $\mathbf{M}\mathbf{P}^{\text{a}}\mathbf{M}^{\text{T}}$ decreases but $\mathbf{Q}$ stays the same, so the $\mathbf{P}^{\text{b}}$ structure becomes less accurate.

The lower half of Figure 9 is with additive inflation. The additive inflation significantly improves the $\mathbf{P}^{\text{b}}_{\text{clim}}$ structure, especially with dense observations. It is not clear why the $\mathbf{P}^{\text{b}}_{\text{clim}}$ benefits from localization and multiplicative inflation. There are two possible explanations: the additive inflation suffers from sampling error, and the additive inflation does not represent flow-dependent model errors. These two factors also suggest the relative performance of 4DVar should be improved by increasing the observation density, which agrees with our results.

5.3. *Comparing EDA-D with 4DVar-Ben*

Figure 10 compares EDA-D (with additive inflation), 4DVar-Ben and 4DVar for the imperfect model and this shows that 4DVar-Ben beats EDA-D when severe localization is applied and when the observations are evenly distributed in time. By doing the same experiments as for the perfect model (Table 4), we linked this to localization degrading the 4D structure of the EDA-D background-error covariance matrix. For a given ensemble size, the difference between EDA-D and 4DVar-Ben is larger for the imperfect model than for the perfect model, since more severe localization is required by the methods when model error is present. Figure 10 also shows that 4DVar benefits from the use of the tangent linear model when observations are distributed in time. With model error present, sparser observations favour the ensemble methods, while denser observations favour 4DVar (explained in the previous section).

5.4. *Comparing the deterministic and stochastic EDA methods*

Finally, 4DVar is compared with the EDA methods (with and without additive inflation) for the imperfect model in Figure 11. We choose not to show the DEnKF since we have already demonstrated that it is approximately equivalent to EDA-D for the perfect model. The EDA-S analysis errors are significantly inferior to EDA-D for a small ensemble size but EDA-S performs slightly better than EDA-D beyond an ensemble size of about 50. When additive inflation is introduced, both EDA-D and EDA-S improve significantly. The improvement is greater in EDA-D and it is not clear why this is. It may be that the EDA-D square root filter approximation improves as the background errors get smaller, which occurs with additive inflation. Additive inflation is most effective for a large ensemble because the sampling error is less.

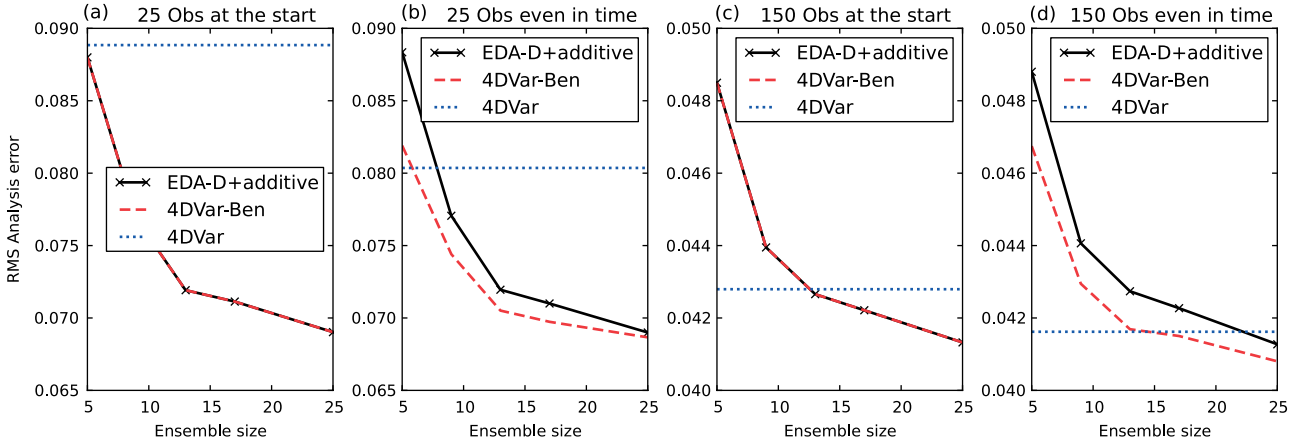

Figure 10. Similar to Figure 3, but the imperfect model is used. As for the perfect model, 4DVar-Ben and 4DVar benefit more than EDA-D when observations are changed to being distributed in time, due to the localization. In the presence of model error, 4DVar is significantly more competitive for dense observations than for sparse observations. This figure is available in colour online at wileyonlinelibrary.com/journal/qj

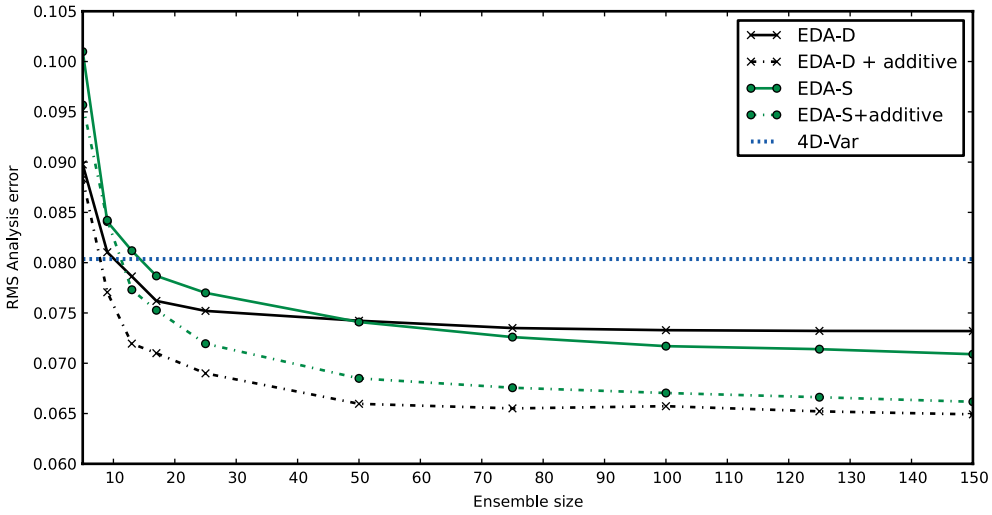

Figure 11. These imperfect model results show comparisons between EDA-D, EDA-S and 4DVar for 25 observations at the 5 timesteps 2,3,. . . ,6. Additive inflation significantly reduces the analysis errors of both the EDA methods. This figure is available in colour online at wileyonlinelibrary.com/journal/qj

4DVar is significantly more competitive for the imperfect model than the perfect model (Figure 4).

6. Summary and discussion

In this article, 4DVar has been compared with a variety of flow-dependent DA methods. One motivation for the toy model experiments came from the near-operational experiments by Buehner *et al.* (2010b), where similar methods were compared. It was hoped that some of their results could be explained using a toy model test bed.

It was shown that, even with a perfect climatological background-error covariance matrix **B**, 4DVar is beaten by the DA methods that use a flow-dependent background-error covariance matrix. The 4DVar-Ben method is a way of tackling the weakness of a climatological **B**. The flow-dependent background-error covariance matrix from EDA-D is used in the 4DVar cost function. The 4DVar analysis errors are then similar to the other flow-dependent methods. Similar idealized model results have been shown by Zhang *et al.* (2009), who used an EnKF background-error covariance matrix in 4DVar.

It is not essential to use an ensemble to provide flow-dependent covariances; an alternative approach is long-window 4DVar (Fisher *et al.*, 2005). Weak-constraint 4DVar, allowing for model errors, must be used. The problems with estimating model errors are discussed below, and implementation is likely to be expensive (Temolet, 2007).

Bannister (2009) describes three practical difficulties that exist in estimating **B** in NWP, which were not replicated in the toy model experiments. Firstly, the climatological **B** can only be approximated in NWP since the true background errors are unknown. Secondly, the **B** matrix is too big to model, which requires further approximations, such as homogeneity and isotropy. Thirdly, there are not enough samples of errors (or surrogate states) to calculate **B** accurately. As a result of these practical difficulties, the climatological **B** is incorrect in NWP. These effects mean that our 4DVar with a perfect climatological **B** has an advantage. On the other hand, the NWP **B** has some advantages which we do not properly simulate in our toy system. For example, **B** does not suffer from sampling error. Sampling error and localization of $\mathbf{P}^{b}$ mean that balance constraints are more difficult to implement with ensemble methods (Lorenc, 2003b, gives details).

Operational NWP methods which allow for model error are still very much a subject of research. Our toy model

experiments had accurate knowledge of the 'truth', enabling us to measure model errors – operational systems cannot do this. Some early methods for estimating **B** used a surrogate truth, either a different length forecast in the NMC method (Parrish and Derber, 1992) or observations (Hollingsworth and Lönnberg, 1986). The resulting estimate therefore included the contribution of model error. But such methods have several difficulties, so most recent developments have estimated **B** from EDA schemes (Fisher, 2003), as we did to get $\mathbf{P}^{\mathrm{b}}_{\mathrm{clim}}$ in Figure 9. They therefore rely on the EDA scheme to allow for the contribution of model errors. This can either be by additive inflation (Mitchell *et al.*, 2002), as we did in this article, or by a more physically based model of error sources such as imperfect resolution and parametrisation schemes (Tennant *et al.*, 2011); such methods can give some flow dependence, but it is hard to validate them. Model errors are also likely to have significant time correlations, something ignored in our study. Because of the uncertainty about the size and structure of NWP model errors, we must be cautious about our toy model results. It seems likely that these conclusions about model error are relevant:

- Additive inflation improves ensemble covariance estimates. However, even in our idealized experiments, it did not fully correct for the effect of model error; localization and multiplicative inflation were still needed;
- Model error is more important in data-rich areas. This seems to improve the relative performance of 4DVar.

An ensemble of 4DEnVars provides a 4D fit between the observations and an ensemble of model trajectories in an assimilation window, without the need for tangent linear or adjoint models. It was shown that 4DVar-Ben and EDA-D performed about the same without localization but that 4DVar-Ben performed best with localization. The 4D localization of the 4DEnVar background-error covariance degrades the time correlations because the nonlinear model and the localization function do not commute. 4DVar-Ben avoids this problem by propagating the initial localized background-error covariance using the tangent linear model. For our set-up, the tangent linear model was an accurate representation of the full nonlinear model. The results agree with the quasi-operational experiments by Buehner *et al.* (2010b) and Zhang and Zhang (2012). However, the following reasons mean the potential operational benefit of 4DEnVar is likely to be much more significant than the toy model results suggest:

- 4DVar-Ben is more expensive both to maintain and to run, being less suitable for ensembles and parallel computers. It may be possible to implement higher-resolution and larger ensembles by moving to 4DEnVar methods, compensating for the degrading effect of localization compared to 4DVar-Ben;
- There is ongoing research to reduce the errors from 4D localization (e.g. Bishop and Hodyss, 2007; Buehner and Charron, 2007).

We demonstrated the equivalence between an ensemble of 4DEnVars (EDA-D) and an ensemble Kalman filter (DEnKF). We did not explore the advantages of 4DEnVar over the EnKF, which are important for operational systems:

- 4DEnVar can use a hybrid background-error covariance matrix;
- 4DEnVar should make better use of nonlinear observation operators and indirect observations (such as satellite radiances) because it localizes in model space;
- 4DEnVar can use many of the same functions as 4DVar, giving less overheads for an operational centre with a 4DVar infrastructure.

Although an ensemble of 4DEnVars is likely to be more expensive than an EnKF (since the latter can re-use the Kalman gain matrix), the difference is not very significant since the dominant cost is that of the forecasts.

Lastly, we compared a stochastic ensemble of 4DEnVars (EDA-S) with a deterministic version (EDA-D). It was found that the deterministic ensemble generally performed better than the stochastic ensemble, although the difference was small with a large ensemble of 150. This matches the results of Whitaker and Hamill (2002) and Sakov and Oke (2008). EDA-D approximates a proper square root filter when the analysis correction is small. Because of this approximation, it is unlikely that operational systems would implement EDA-D, even though there is currently no way of implementing an ensemble of 4DEnVars as an exact square root filter. However, operational ensembles are of the order $10–10^2$ and are expected to grow, so the sampling error introduced from perturbed observations should become less significant.

Appendix

The Lorenz (2005) model

The equations for model II are given by

$$\frac{\mathrm{d}X_n}{\mathrm{d}t} = [X,X]_{K,n} - X_n + F, \tag{A1}$$

where

$$[X,X]_{K,n} = \sum_{j=-J}^{J}{}' \sum_{i=-J}^{J}{}' \big(- X_{n-2K-i}X_{n-K-j} + X_{n-K+j-i}X_{n+K+j} \big) \Big/ K^2. \tag{A2}$$

The parameter $J = K/2$ if K is even and $J = (K-1)/2$ if K is odd. $\sum'$ implies K is even, which is replaced by $\sum$ if K is odd. The variable X represents the cyclic grid points on a latitude circle, with dimension n. The quadratic and linear terms in (A1) represent advection and dissipation respectively and F is the forcing.

Acknowledgements

We would like to thank Neill Bowler, Adam Clayton and others at the Met Office for their help. Craig Bishop also gave some good suggestions when he visited the Met Office. The comments from the two reviewers were constructive, especially in improving the model error section of the article. Finally, the work would not have been completed without funding from EPSRC and the Met Office.

References

Anderson JL, Anderson SL. 1999. A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. *Mon. Weather Rev.* **127**: 2741–2758.

Bannister RN. 2009. A review of forecast-error covariance statistics in atmospheric variational data assimilation. I: Characteristics and measurements of forecast-error covariances. *Tellus A* **61**: 97–111.

Bishop CH, Hodyss D. 2007. Flow-adaptive moderation of spurious ensemble correlations and its use in ensemble-based data assimilation. *Q. J. R. Meteorol. Soc.* **133**: 2029–2044.

Bowler NE, Flowerdew J, Pring SR. 2013. Tests of different flavours of EnKF on a simple model. *Q. J. R. Meteorol. Soc.* in press. DOI: 10.1002/qj.2055.

Buehner M, Charron M. 2007. Spectral and spatial localization of background-error correlations for data assimilation. *Q. J. R. Meteorol. Soc.* **133**: 615–630.

Buehner M, Houtekamer PL, Charette C, Mitchell HL, He B. 2010a. Intercomparison of variational data assimilation and the ensemble Kalman filter for global deterministic NWP. Part I: Description and single-observation experiments. *Mon. Weather Rev.* **138**: 1550–1566.

Buehner M, Houtekamer PL, Charette C, Mitchell HL, He B. 2010b. Intercomparison of variational data assimilation and the ensemble Kalman filter for global deterministic NWP. Part II: One-month experiments with real observations. *Mon. Weather Rev.* **138**: 1567–1586.

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

Courtier P, Thepaut J-N, Hollingsworth A. 1994. A strategy for operational implementation of 4D-Var, using an incremental approach. *Q. J. R. Meteorol. Soc.* **120**: 1389–1408.

Fisher M. 2003. 'Background-error covariance modelling'. In *Proceedings of Seminar on recent developments in data assimilation for atmosphere and ocean*, 8–12 September 2003. 45–64 ECMWF: Reading, UK.

Fisher M, Leutbecher M, Kelly GA. 2005. On the equivalence between Kalman smoothing and weak-constraint four-dimensional variational data assimilation. *Q. J. R. Meteorol. Soc.* **131**: 3235–3246.

Gaspari G, Cohn SE. 1994. Construction of correlation functions in two and three dimensions. *Q. J. R. Meteorol. Soc.* **125**: 723–757.

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.

Hollingsworth A, Lönnberg P. 1986. The statistical structure of short-range forecast errors as determined from radiosonde data. Part 1: The wind field. *Tellus A* **38A**: 111–136.

Hunt BR, Kalnay E, Kostelich EJ, Ott E, Patil DJ, Sauer T, Szunyogh I, Yorke JA, Zimin AV. 2004. Four-dimensional ensemble Kalman filtering. *Tellus A* **56**: 273–277.

Kalnay E, Li H, Miyoshi T, Yang S-C, Ballabrera-Poy J. 2007. 4D-Var or ensemble Kalman filter? *Tellus A* **59**: 758–773.

Le Dimet FX, Talagrand O. 1986. Variational algorithms for analysis and assimilation of meteorological observations: Theoretical aspects. *Tellus A* **38**: 97–110.

Liu C, Xiao Q, Wang B. 2008. An ensemble-based four-dimensional variational data assimilation scheme. Part I: Technical formulation and preliminary test. *Mon. Weather Rev.* **136**: 3363–3373.

Liu C, Xiao Q, Wang B. 2009.. An ensemble-based four-dimensional variational data assimilation scheme. Part II: Observing System Simulation Experiments with Advanced Research WRF (ARW). *Mon. Weather Rev.* **137**: 1687–1704.

Lorenc AC. 2003a. Modelling of error covariances by four-dimensional variational data assimilation. *Q. J. R. Meteorol. Soc.* **129**: 3167–3182.

Lorenc AC. 2003b. The potential of the Ensemble Kalman filter for NWP – a comparison with 4D-Var. *Q. J. R. Meteorol. Soc.* **129**: 3183–3203.

Lorenz EN. 2005. Designing chaotic models. *J. Atmos. Sci.* **62**: 1574–1587.

Mitchell HL, Houtekamer PL, Pellerin G. 2002. Ensemble size, balance and model-error representation in an ensemble Kalman filter. *Mon. Weather Rev.* **130**: 2791–2808.

Parrish DF, Derber JC. 1992. The National Meteorological Centre's spectral statistical interpolation analysis scheme. *Mon. Weather Rev.* **120**: 1747–1763.

Rabier F. 2005. Overview of global data assimilation developments in numerical weather prediction centres. *Q. J. R. Meteorol. Soc.* **131**: 3215–3233.

Rawlins F, Ballard SP, Bovis KJ, Clayton AM, Li D, Inverarity GW, Lorenc AC, Payne TJ. 2007. The Met Office four-dimensional variational data assimilation scheme. *Q. J. R. Meteorol. Soc.* **133**: 347–362.

Sakov P, Oke PR. 2008. A deterministic formulation of the ensemble Kalman filter: an alternative to ensemble square root filters. *Tellus A* **60**: 3035–3049.

Tennant WJ, Shutts GJ, Arribas A, Thompson SA. 2011. Using a stochastic kinetic energy backscatter scheme to improve MOGREPS probabilistic forecast skill. *Mon. Weather Rev.* **139**: 1190–1206.

Tremolet Y. 2007. Model-error estimation in 4D-Var. *Q. J. R. Meteorol. Soc.* **133**: 1267–1280.

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

Wang X, Snyder C, Hamill TM. 2007. On the theoretical equivalence of differently proposed ensemble–3D-Var hybrid analysis schemes. *Mon. Weather Rev.* **135**: 222–227.

Zhang M, Zhang F. 2012. E4DVar: Coupling an Ensemble Kalman Filter with four-dimensional variational data assimilation in a limited-area weather prediction model. *Mon. Weather Rev.* **140**: 587–600.

Zhang F, Zhang M, Hansen JA. 2009. Coupling ensemble Kalman filter with four-dimensional variational data assimilation. *Adv. Atmos. Sci.* **26**: 1–8. DOI: 10.1007/s00376-009-0001-8.