# Advances in data assimilation techniques and their relevance to satellite data assimilation

**Andrew C. Lorenc**

*Met Office, Exeter,*
*EX1 3PB, United Kingdom*
*andrew.lorenc@metoffice.gov.uk*

## 1 Introduction

I have been given a very broad topic – I have to be selective. I have chosen to consider only one type of advance in data assimilation (DA) methods for NWP: the use of forecast errors of the day (EOTD). At the heart of the THORPEX programme of the last decade was the ability, thanks to better science and bigger computers, to predict the evolution and growth of forecast errors. It has been used within the short time-window of 4DVar. Ensemble methods can predict over longer periods to estimate the EOTD; section 2 describes DA methods making use of this: hybid-4DVar, 4DEnVar and the EnKF (focussing on the LETKF as a popular "flavour").

Developments are driven by the evolving capabilities of, and requirements for, NWP:

**Computing:** processors have stopped getting faster, so the hoped-for continuation of the steady increase in power has to come from massively parallel machines. While we expect to soon be able to run global convective scale models, this may be achieved by major revisions of both the scientific model design and the software, e.g. GungHo, MPAS and NICAM models (Ford *et al.*, 2013; Skamarock *et al.*, 2010; Satoh *et al.*, 2008).

**Nonlinearity:** customer requirements require higher resolution and more use of observations affected by cloud and precipitation, at scales where error evolution is more nonlinear (Hohenegger and Schär, 2007).

**Ensembles:** there is demand for probabilistic forecasts and warnings of extreme events, best met using ensemble forecasts (Golding, 2009).

Then in section 3 I change perspective, and look at some potentially difficult problems relevant to satellite DA and how well the different methods handle them.

Finally in section 4 I abandon my attempts at balance, to give some personal conclusions.

## 2 DA methods for NWP

The current generation of DA methods for NWP typically assimilate many millions of observations to produce the initial conditions for forecast models with $O\left(10^9\right)$ state variables, often in less than an hour. All methods must incorporate significant approximations to make such calculations possible on available supercomputers. The main approximation is that errors are Gaussian, giving linear DA equations – some non-Gaussian distributions can be handled by iteration of the linear equations, but we cannot afford fully nonlinear DA.

## 2.1 4-Dimensional DA Methods

All the methods in this paper perform a 4D best-fit to observations in a window, assuming Gaussian background and observation errors. I use an underline to extend the standard notation of Ide *et al.* (1997) to four dimensions, for instance $\underline{\mathbf{x}}^b$ is the background trajectory. The expected error covariance of $\underline{\mathbf{x}}^b$ is $\underline{\mathbf{P}}$. This defines a Gaussian pdf for the 4D increment $\delta\underline{\mathbf{x}}$:

$$\delta\underline{\mathbf{x}} \sim N\left(0, \underline{\mathbf{P}}\right), \tag{1}$$

which gives the probability density that $\underline{\mathbf{x}}^b + \delta\underline{\mathbf{x}}$ is the true trajectory. All the methods variationally determine the $\delta\underline{\mathbf{x}}$ which maximizes the posterior Bayesian likelihood by minimizing a penalty function measuring the distances from the background and the observations:

$$J\left(\delta\underline{\mathbf{x}}\right) = \frac{1}{2}\delta\underline{\mathbf{x}}^T\underline{\mathbf{P}}^{-1}\delta\underline{\mathbf{x}} + \frac{1}{2}\left(\underline{\mathbf{y}} - \underline{\mathbf{y}}^o\right)^T\underline{\mathbf{R}}^{-1}\left(\underline{\mathbf{y}} - \underline{\mathbf{y}}^o\right), \tag{2}$$

where the second term is the observational penalty, measuring the difference of the observations in the time-window ($\underline{\mathbf{y}}^o$) from their model estimates ($\underline{\mathbf{y}}$). The latter are calculated as accurately as possible, using the nonlinear observation operator. For the purposes of this paper we can simplify this to

$$\underline{\mathbf{y}} = \underline{H}\left(\underline{\mathbf{x}}^b + \delta\underline{\mathbf{x}}\right). \tag{3}$$

$\underline{\mathbf{P}}$ is **BIG!** We cannot even estimate it fully (Dee, 1991), let alone compute $\frac{1}{2}\delta\underline{\mathbf{x}}^T\underline{\mathbf{P}}^{-1}\delta\underline{\mathbf{x}}$. The solution is to model $\underline{\mathbf{P}}$ using a sequence of operations we can compute, then use these to transform $\delta\underline{\mathbf{x}}$ so that $\frac{1}{2}\delta\underline{\mathbf{x}}^T\underline{\mathbf{P}}^{-1}\delta\underline{\mathbf{x}}$ simplifies.

## 2.2 Hybrid-4DVar

### 2.2.1 Using climatological covariance B

The traditional 4DVar models the 3D covariance using transforms

$$\mathbf{B} = \mathbf{U}\mathbf{U}^T \tag{4}$$

and use these to construct the 3D analysis increment

$$\delta\mathbf{x}_0 = \mathbf{U}\mathbf{v}^c \tag{5}$$

which is made 4D using linear forecast model $\underline{\mathbf{M}}$

$$\delta\underline{\mathbf{x}} = \underline{\mathbf{M}}\delta\mathbf{x}_0 \tag{6}$$

This gives an implicit 4D prior covariance

$$\underline{\mathbf{P}} = \underline{\mathbf{M}}\mathbf{B}\underline{\mathbf{M}}^T \tag{7}$$

and a transformed penalty function

$$J\left(\mathbf{v}^c\right) = \frac{1}{2}\mathbf{v}^{cT}\mathbf{v}^c + \frac{1}{2}\left(\underline{\mathbf{y}} - \underline{\mathbf{y}}^o\right)^T\underline{\mathbf{R}}^{-1}\left(\underline{\mathbf{y}} - \underline{\mathbf{y}}^o\right) \tag{8}$$

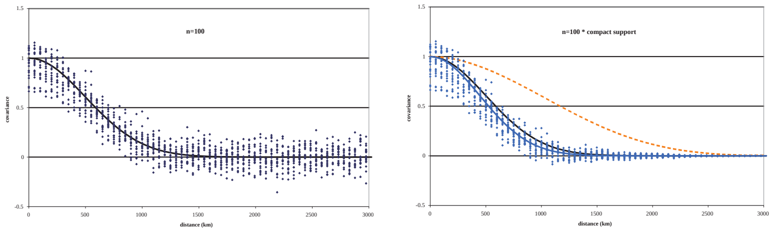

*Figure 1: Left: covariances from 100 samples from the solid curve. Right: Same covariance samples, localised by a Schur product with the red localisation function. (Lorenc, 2003)*

### 2.2.2 Weaknesses of traditional 4DVar

1. Climatological covariances **B**
2. Lack of parallelisation
3. Lack of an analysis ensemble

Weakness 1 can be addressed by introducing EOTD from an ensemble; 2 can be addressed by 4DEnVar, as discussed below. I do not have time to discuss 3 further. It is currently addressed either by using an external EnKF or by running ensembles of 4DVar (Bonavita *et al.*, 2012). The latter is very expensive for a large ensemble. Research is underway into less expensive methods, e.g. Auligné (2012).

### 2.2.3 Ensemble covariance filtering

**B** is **BIG!** To get anything like a reasonable estimate from an EOTD ensemble we need a large ensemble PLUS clever covariance filtering, based on two ideas:

- We make assumptions about local homogeneities, and smooth accordingly to reduce sampling error. This can be done by horizontal, rotational, and time averaging.
- We make assumptions that certain correlations are near zero, and "localise" them towards zero. This can be done in the horizontal and vertical, in spectral space, and between transformed variables.

Even with these, we get better results using a hybrid of the EOTD estimate with a climatological **B**. There are two methods of making a hybrid in 4DVar, both can use some of the above ideas:

1. Train [part of] a covariance model, like that used for **B**, using current (or recent) ensembles. This is the method used at ECMWF and Meteo-France.
2. Augment **B** by using localised ensemble perturbations (Clayton *et al.*, 2013).

I will show how to implement method 2 later. First I discuss some examples of covariance filtering in the two methods:

- I will start with basic horizontal covariance localisation (Hamill *et al.*, 2001; Houtekamer and Mitchell, 2001), since this is usually the only filtering idea in EnKF and it was the first tried in

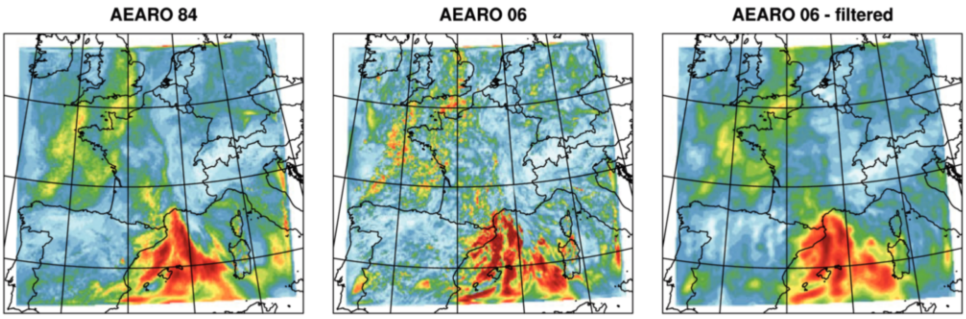

*Figure 2: AROME ensemble s.d. of humidity at ∼945 hPa: Left using 84 members; Centre using 6 members; Right using 6 members and horizontal filter (Ménétrier et al., 2014).*

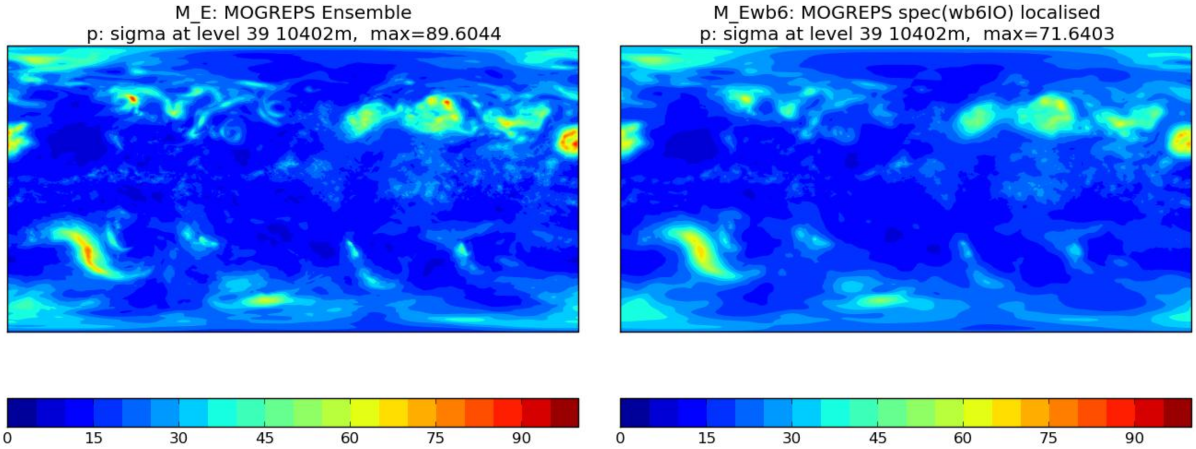
M_E: MOGREPS Ensemble
p: sigma at level 39 10402m, max=89.6044
M_Ewb6: MOGREPS spec(wb6IO) localised
p: sigma at level 39 10402m, max=71.6403
0
15
30
45
60
75
90
0
15
30
45
60
75
90

*Figure 3: MOGREPS ensemble s.d. of pressure at ∼10 km: Left raw ensemble (22 members); Right after spectral localisation in 6 wavebands.*

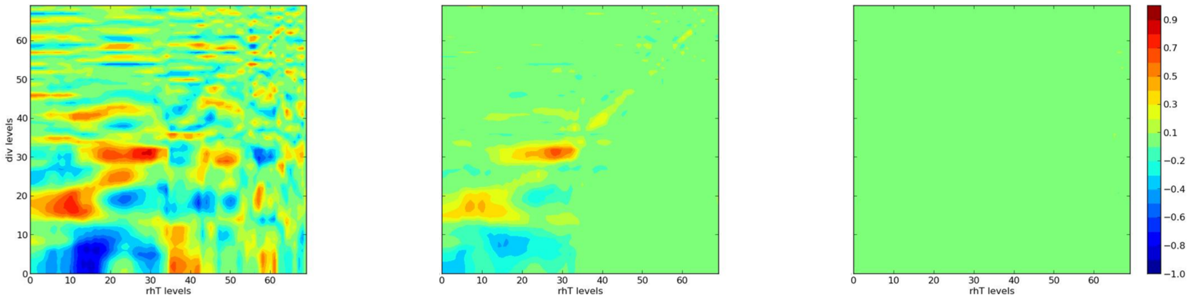

*Figure 4: Correlations between levels for divergence (vertical axis) and rh (horizontal axis) for a precipitating column from the ensemble in fig. 3. Left: raw ensemble; Centre: horizontally, vertically and waveband localised ensemble; Right: inter-variable localised ensemble.*

En-Var methods. Figure 1(left) shown covariances from 100 samples drawn from a covariance function which drops to zero at long distance. It seems clear that the noise is relatively more important where the correlation is small (Lorenc (2003) showed one way of quantifying this) so the idea is to reduce such samples towards zero. We have to retain a valid covariance function – this is ensured by using a Schur product with a valid correlation function, in this case the red curve in Figure 1(right). The Schur product, denoted $\mathbf{C} \circ \mathbf{P}^e$, is an element wise product of the localisation matrix $\mathbf{C}$ and the sampled covariance $\mathbf{P}^e$. Localisation is designed to mitigate the small ensemble size used in EnKF and En-Var methods. It should be less needed if training a covariance model with a larger sample. A covariance model can impose similar effects by assumptions about the shape of the covariance function being fitted to the training ensemble.

- Many covariance models used in method 1 separate into variance fields and a correlation model, in which case it is straightforward to provide the variance fields from an ensemble. Raynaud *et al.* (2009) suggested filtering them, and this is applied by Bonavita *et al.* (2012). Figure 2 comes from a study to apply the same method at the convective scale. Using method 2 it would be harder to apply explicit filtering to the variances, but figure 3 shows that similar smoothing can be achieved by spectral localisation (Buehner and Charron, 2007; Buehner, 2012); the amount of filtering is controlled by the spectral width of each waveband. (The ECMWF wavelet covariance model (Fisher, 2003) uses similar wavebands.)

- Time-averaging is natural in method 1 – simply use the ensembles from a long enough period to derive coefficients in **B**. It is even possible to use different periods for different parts, e.g. early attempts update only the variances using EOTD. Some time-averaging can be added to method 2 by using lagged ensembles (i.e. longer forecasts from an earlier cycle). Another approach is to time shift the current ensemble trajectories, e.g. by +-1 hour, to increase the effective ensemble size.

- The ECMWF system, by training covariances grouped by global wavenumber, performs a directional averaging in method 1. This is not essential in covariance models, e.g. Purser *et al.* (2003a,b). Examples such as figure 5 suggest this is often not appropriate – which is lucky because there is no easy way to do it in method 2.

- The covariance models used in method 1 perform a variable transform (Derber and Bouttier, 1999; Lorenc *et al.*, 2000), then assume there is no correlation between the different variables. While correct on average, this does not correctly model covariances involving divergence, vertical motion, and moisture in precipitating situations (Montmerle and Berre, 2010). The Met Office system using method 2 (Clayton *et al.*, 2013) transforms the ensemble in the same way (to avoid imbalance from horizontal localisation), so it is easy to provide an optional "localisation" between the variables – its effect is seen in figure 4. The raw ensemble correlations are noisy; the localised ensemble has less noise while retaining plausible features like a positive correlation of rh with convergence below and divergence above; the inter variable localisation, as expected, removes all correlations between rh and wind.

The two approaches to hybrid covariances start from different ends: method 1 starts from a climatological covariance model, then adds ensemble-derived coefficients; method 2 starts from a raw ensemble then filters the covariances. Eventually they might approach each other in the middle. As we shall see later, there is less scope for these methods in the EnKF, other than simple spatial localisation.

### 2.2.4 En-4DVar: using an ensemble of 3D states which samples background errors

It is helpful to define an ensemble perturbation matrix

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}'_1 & \cdots & \mathbf{x}'_N \end{bmatrix} \tag{9}$$

where

$$\mathbf{x}'_k = \frac{1}{\sqrt{N-1}} \left(\mathbf{x}_k - \bar{\mathbf{x}}\right) \tag{10}$$

We model the 3D background error covariance as localised ensemble covariance

$$\mathbf{P} = \mathbf{C} \circ \mathbf{X}\mathbf{X}^T \tag{11}$$

then model $\mathbf{C}$ using transforms

$$\mathbf{C} = \mathbf{U}^\alpha \mathbf{U}^{\alpha T} \tag{12}$$

The initial increment $\delta\mathbf{x}_0$ is a linear combination of ensemble perturbations $\mathbf{x}'_k$, localise by weights $\alpha_k$ which are constrained to be smooth, consistent with (12)

$$\alpha_k = \mathbf{U}^\alpha \mathbf{v}_k^\alpha \tag{13}$$

$$\delta\mathbf{x}_0 = \textstyle\sum_{k=1}^N \alpha_k \circ \mathbf{x}'_k \tag{14}$$

The 4D increment uses the linear forecast model $\underline{\mathbf{M}}$ as in normal 4DVar

$$\delta\underline{\mathbf{x}} = \underline{\mathbf{M}} \textstyle\sum_{k=1}^N \alpha_k \circ \mathbf{x}'_k \tag{15}$$

Lorenc (2003) showed that localising $\mathbf{x}'_k$ with $\alpha_k$ sampled such that $\left\langle \alpha_k \alpha_k^T \right\rangle = \mathbf{C}$ is equivalent to using a localized covariance $\mathbf{C} \circ \mathbf{X}\mathbf{X}^T$. So the 4D covariance is

$$\underline{\mathbf{P}} = \underline{\mathbf{M}} \left(\mathbf{C} \circ \mathbf{X}\mathbf{X}^T\right) \underline{\mathbf{M}}^T \tag{16}$$

To use the same software, with the background penalty transformed into a dot-product, we concatenated control vectors $\mathbf{v}^T = \left[\mathbf{v}_1^{\alpha T} \cdots \mathbf{v}_N^{\alpha T}\right]$, giving the transformed penalty function

$$J(\mathbf{v}) = \frac{1}{2}\mathbf{v}^T\mathbf{v} + \frac{1}{2}\left(\underline{\mathbf{y}} - \underline{\mathbf{y}}^o\right)^T \underline{\mathbf{R}}^{-1} \left(\underline{\mathbf{y}} - \underline{\mathbf{y}}^o\right) \tag{17}$$

### 2.2.5 hybrid-4DVar

This is a simple combination of the climatological and ensemble methods. We use a 4D analysis increment

$$\delta\underline{\mathbf{x}} = \underline{\mathbf{M}} \left(\beta_c \mathbf{U}\mathbf{v}^c + \beta_e \sum_{k=1}^N \mathbf{U}^\alpha \mathbf{v}_k^\alpha \circ \mathbf{x}'_k\right) \tag{18}$$

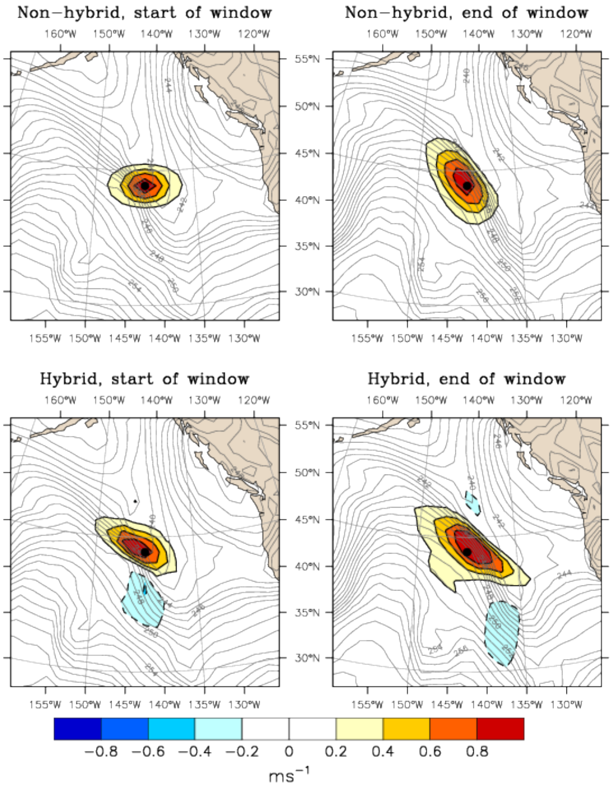

*Figure 5: Zonal wind responses (filled thick contours, with negative contours dashed) to a single zonal wind observation at the start (left-hand plots) and end (right-hand plots) of the 6-hour 4D-Var window. The plots are for the same time and model level ($\approx$ 500 hPa) as the observation. Upper plots are for the non-hybrid configuration; lower plots for the hybrid. The observation location is marked with a black dot at the centre of each plot. The unfilled contours show the background temperature field. (Clayton* et al.*, 2013)*

and concatenated the control vectors $\mathbf{v}^T = \left[\mathbf{v}^{cT}, \mathbf{v}_1^{\alpha T} \cdots \mathbf{v}_N^{\alpha T}\right]$, giving a localized 4D covariance

$$\underline{\mathbf{P}} = \underline{\mathbf{M}}\left(\beta_c^2\mathbf{B} + \beta_e^2\mathbf{C} \circ \mathbf{X}\mathbf{X}^T\right)\underline{\mathbf{M}}^T \tag{19}$$

Hybrid-4DVar was made operational at the Met Office on 20 July 2011, giving about a 1% reduction in RMS forecast errors (Clayton *et al.*, 2013). Figure 5 illustrates the effect of the hybrid, particularly on covariances at the beginning of the 4DVar window which previously had no flow dependence. At the end of the window the effective covariances are $\mathbf{MBM}^T$ and even for a 6 hour window show stretching along the jet. The hybrid covariances show a similar stretching from the start.

### 2.2.6 Parallelisation

4DVar has several potential problems looming in the next decade; they will affect different centres at different times depending on their computers and models:

1. To run fast enough on massively-parallel (MPP) computers we will need to use millions of parallel threads. Horizontal domain-decomposition is not enough, especially in the sequential runs of lower-resolution linear (PF) and Adjoint models inside a 4DVar minimisation.

2. Projects are underway promising significant redesign of forecast models, to address the MPP issue for the forecast model. This means we may have to re-write the PF and Adjoint models.

While some mitigation is possible (e.g. Fisher and Auvinen (2011)), the simplest solution to both problems is to use the ensemble trajectories, which can be pre-calculated in parallel, instead of the models inside 4DVar.

4DEnVar, described in the next section, does this while retain the rest of the 4DVar infrastructure. Note that this still needs global operations such as smoothing (currently done using spectral transforms) and Poisson solvers. For some models even these may not be available; if so the LETKF is an easier approach (Kondo and Tanaka, 2009).

### 2.3 4DEnVar: using an ensemble of 4D trajectories which samples background errors

The is simply derived by adding a time-dimension to all the variables in (9), (10), (11), (12), (13) and (14):

$$\underline{\mathbf{X}} = \left[ \begin{array}{ccc} \underline{\mathbf{x}}'_1 & \cdots & \underline{\mathbf{x}}'_N \end{array} \right] \tag{20}$$

$$\underline{\mathbf{x}}'_k = \frac{1}{\sqrt{N-1}} \left( \underline{\mathbf{x}}_k - \underline{\bar{\mathbf{x}}} \right) \tag{21}$$

$$\underline{\mathbf{P}} = \underline{\mathbf{C}} \circ \underline{\mathbf{X}}\underline{\mathbf{X}}^T \tag{22}$$

$$\underline{\mathbf{C}} = \underline{\mathbf{U}}^{\alpha} \underline{\mathbf{U}}^{\alpha T}. \tag{23}$$

In general $\underline{\mathbf{U}}^{\alpha}$ needs a time component, which should follow the expected propagation of information (Bishop and Hodyss, 2009a,b; Ota *et al.*, 2013)

$$\underline{\alpha}_k = \underline{\mathbf{U}}^{\alpha} \mathbf{v}_k^{\alpha} \tag{24}$$

$$\delta \underline{\mathbf{x}} = \sum_{k=1}^{N} \underline{\alpha}_k \circ \underline{\mathbf{x}}'_k, \tag{25}$$

but current large NWP systems simply build a 4D $\underline{\mathbf{C}}$ which does not vary in time, using a persistence forecast $\underline{\mathbf{I}}$ and the same 3D $\mathbf{C} = \mathbf{U}^{\alpha}\mathbf{U}^{\alpha T}$. In this case the 4D $\delta\underline{\mathbf{x}}$ can be built one time-level at a time, from 3D localised perturbations and constant $\alpha_k = \mathbf{U}^{\alpha}\mathbf{v}_k^{\alpha}$

$$\delta \mathbf{x}(t) = \textstyle\sum_{k=1}^{N} \alpha_k \circ \mathbf{x}'_k(t) \tag{26}$$

To make the link with the LETKF later it is helpful to rewrite this using a matrix whose columns are the $N$ ensemble $\alpha$s, and use $\mathbf{1}_N$ to denote a column vector of N 1s

$$\mathbf{A} = \left[ \begin{array}{ccc} \alpha_1 & \cdots & \alpha_N \end{array} \right] \tag{27}$$

$$\delta \mathbf{x}(t) = \left( \mathbf{A} \circ \mathbf{X}(t) \right) \mathbf{1}_N \tag{28}$$

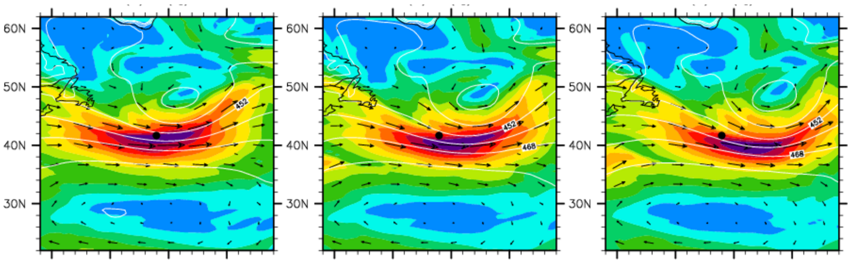

*Figure 6: The background field at 3-hourly intervals. The test observation was placed at the black dot at the first time (Lorenc et al., 2014).*

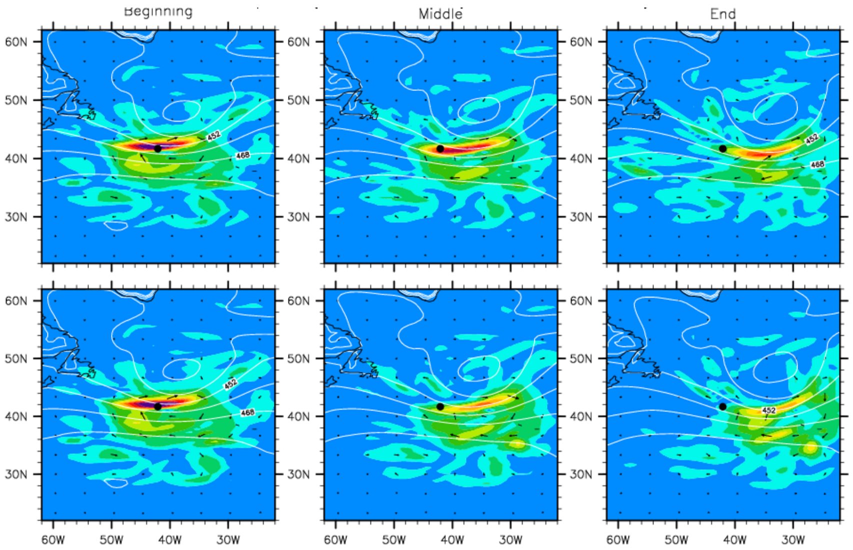

*Figure 7: Top: propagation of the increment $\delta\underline{\mathbf{x}}$ in 4DEnVar with localisation scale 1200 km. Bottom: propagation of the increment $\delta\underline{\mathbf{x}}$ in En-4DVar (Lorenc et al., 2014).*

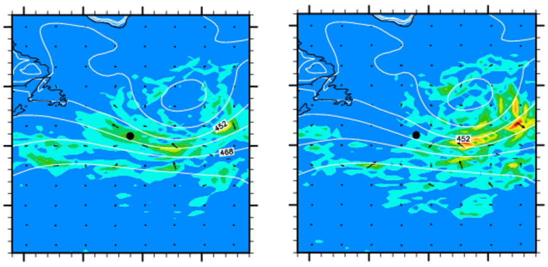

*Figure 8: $E_{SC}$ field calculated using (29) for the final time in 4DEnVar (left) and En-4DVar (right) (Lorenc et al., 2014).*

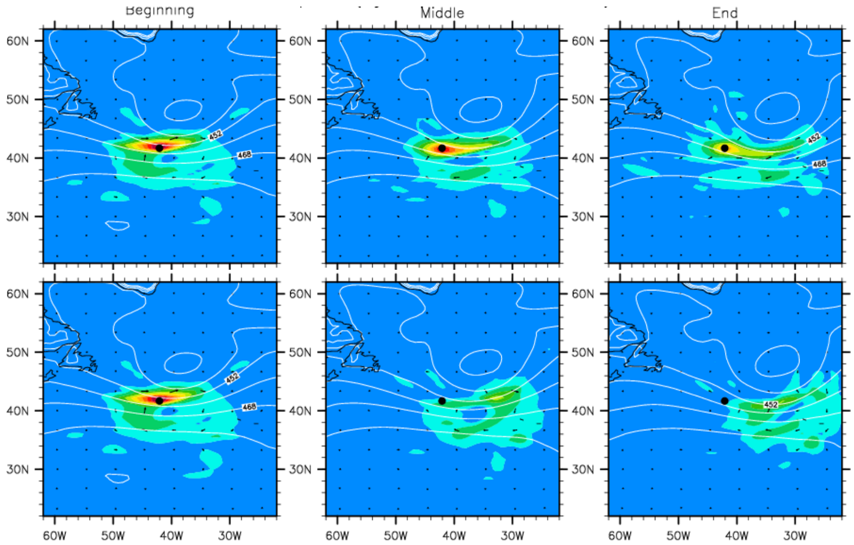

*Figure 9: As figure 7 but for 50% hybrids: Top: propagation of the increment $\delta\underline{\mathbf{x}}$ in hybrid-4DEnVar. Bottom: propagation of the increment $\delta\underline{\mathbf{x}}$ in hybrid-4DVar (Lorenc* et al.*, 2014).*

The concatenated control vectors and transformed penalty function (17) are unchanged.

Comparing (15) with (25), we see we have a different type of linear model – 4DVar uses $\underline{\mathbf{M}}$ to calculate $\delta\underline{\mathbf{x}}$, whereas 4DEnVar directly calculates $\delta\underline{\mathbf{x}}$ from a linear combination of ensemble trajectories. The behaviour of these models is best seen in a single observation experiment in a jet stream (figure 6). Figure 7 shows the propagation of the increments in 4DEnVar (top) and En-4DVar (bottom). At the initial time they are virtually identical, as expected because the observation is at the start, making them both 3D. The subsequent propagations are very similar - it is hard to judge between them. An objective measure is given using the difference from a nonlinear model trajectory[1] (29). Figure 8 shows that 4DEnVar increments are slightly more consistent with the nonlinear model. On a hurricane case (not shown here) 4DEnVar was noticeably more consistent.

$$E_{SC} = \left(M_{0\rightarrow 6}\left(\mathbf{x}_b + \delta\mathbf{x}_0^a\right)\right) - \left(M_{0\rightarrow 6}\left(\mathbf{x}_b\right) + \delta\mathbf{x}_6^a\right) \tag{29}$$

Hybrid-4DEnVar is constructed like hybrid-4DVar with an important different – because we want to avoid using any model inside 4DEnVar the climatological part is persisted, as in 3DVar.

$$\delta\underline{\mathbf{x}} = \beta_c\underline{\mathbf{I}}\delta\mathbf{x}_0 + \beta_e\sum_{k=1}^{N}\underline{\alpha}_k \circ \underline{\mathbf{x}}'_k \tag{30}$$

$$\underline{\mathbf{P}} = \beta_c^2\underline{\mathbf{IBI}}^T + \beta_e^2\underline{\mathbf{C}} \circ \underline{\mathbf{XX}}^T \tag{31}$$

The first Met Office trial of 4DEnVar (Lorenc *et al.*, 2014) copied its settings from the hybrid-4DVar (Clayton *et al.*, 2013), in particular the localisation matrix $\mathbf{C}$ with scale 1200 km and the hybrid weights $\beta_c^2 = 0.8$, $\beta_e^2 = 0.5$. Results were disappointing: while hybrid-3DEnVar performed equally to hybrid-3DVar, as they should, but while hybrid 4DVar beat hybrid-3DVar by 3.6% [2], hybrid-4DEnVar only beat hybrid-3DEnVar by 0.5% (measured on a basket of RMS verification scores). The reason was the large weight given to the climatological covariance, which is treated the same in 3DEnVar and 4DEnVar. This effect is illustrated in figure 9, where the hybrid-4DEnVar propagation is split - half of the increment does not propagate at all.

[1] Neither method here makes any allowance for model error, although they can.

[2] Rather more than expected – more than on 4DVar's implementation Rawlins *et al.* (2007).

Lorenc *et al.* (2014) concluded that we need to reduce the weight on climatological **B** relative to the ensemble covariance. But these weights are usually determined by experiment; both components provide some benefit (Etherton and Bishop, 2004; Clayton *et al.*, 2013). Increasing the ensemble weight requires us to first improve the covariances derived from the ensemble by:

- a bigger ensemble;
- better ensemble generation;
- better covariance filtering.

As pointed out above, the parallelisation issues will hit different centres at different time. Canada are already replacing 4DVar by 4DEnVar. Their 4DVar is not hybrid, and very inefficient. Their global model is being replaced by one with a Ying-Yan grid, which does not have an adjoint. In trials global 4DEnVar analysis (~10 min) is ~6 times faster than 4DVar (~1 hr) on half as many cpus (320 vs 640), even though much higher resolution increments (50km vs 100km). Scores were at least as good (Mark Buehner, personal communication).

## 2.4 EnKF – common properties

EnKFs produce an analysis ensemble, so we need to extend the notation to distinguish background values ($\underline{\mathbf{X}}^b$, previously $\underline{\mathbf{X}}$) and analysis values ($\underline{\mathbf{X}}^a$). The computations use the matrix of ensemble model-ob perturbations. For linear $H$ this would be given by $\underline{\mathbf{Y}}^b = \mathbf{H}\underline{\mathbf{X}}^b$, but it is calculated using nonlinear $H$:

$$\underline{\mathbf{y}}'_k = \frac{1}{\sqrt{N-1}}\left(\underline{H}\left(\underline{\mathbf{x}}_k^b\right) - \overline{\underline{H}\left(\underline{\mathbf{x}}^b\right)}\right) \tag{32}$$

$$\underline{\mathbf{Y}}^b = \left[\begin{array}{ccc} \underline{\mathbf{y}}'_1 & \cdots & \underline{\mathbf{y}}'_N \end{array}\right] \tag{33}$$

Most EnKF use the localised ob-gridpoint covariance

$$\underline{\mathbf{C}} \circ \underline{\mathbf{Y}}^b \underline{\mathbf{X}}^{bT}. \tag{34}$$

There are two methods for calculation the analysis ensemble:

- Stochastic filters such as the operational EnKF in Environment Canada (Houtekamer and Mitchell, 2001; Houtekamer *et al.*, 2014) use the same analysis equation for each member and perturb observations (as in ensembles of 4DVar).
- SQRT filters (Tippett *et al.*, 2003) analyses the ensemble mean, then calculate perturbations such that $\underline{\mathbf{X}}^a\underline{\mathbf{X}}^{aT} = \underline{\mathbf{P}}^a$.

Stochastic filters are more robust to wrong assumptions, while SQRT filters have fewer sampling errors and hence are usually more accurate (at least for toy problems).

### 2.4.1 LETKF

(Note below that I have put the factor $1/\sqrt{N-1}$ in (32), to match that in (21). The equations of Hunt *et al.* (2007); Harlim and Hunt (2007) apply the factor $1/\sqrt{N-1}$ to **w** and $\alpha$ rather than $\mathbf{X}^b$.)

The ETKF equation for the mean analysis is a simple linear combination of the ensemble perturbations. The weight use the Kalman gain expressed in ensemble space – the required inverse is solved directly:

$$\delta\underline{\mathbf{x}} = \underline{\mathbf{X}}^b\mathbf{w} \tag{35}$$

$$\mathbf{w} = \tilde{\mathbf{P}}^a\left(\underline{\mathbf{Y}}^b\right)^T\mathbf{R}^{-1}\left(\underline{\mathbf{y}}^o - \underline{H}\left(\underline{\mathbf{x}}^b\right)\right) \tag{36}$$

$$\tilde{\mathbf{P}}^a = \left[\mathbf{I} + \left(\underline{\mathbf{Y}}^b\right)^T\mathbf{R}^{-1}\underline{\mathbf{Y}}^b\right]^{-1} \tag{37}$$

The ETKF and LETKF are SQRT-filters; the analysis perturbations are calculated directly:

$$\underline{\mathbf{X}}^a = \left(\tilde{\mathbf{P}}^a\right)^{1/2}\underline{\mathbf{X}}^b$$

The LETKF solves these equations separately for each grid-point, with local observations. This observation selection is instead of (34), which cannot be applied in the ETKF. Each **w** is one row of matrix **A** defined in (27), then, as in 4DEnVar, the analysis is given by 28. Like 4DEnVar, the LETKF is a 4D method, using ensemble background values which span the window. Again like current 4DEnVar implementations, there is no suggestion that the weights could vary in time.

# 3 Some Difficult Issues for DA Methods

In this section I go through a few issues relevant to satellite DA, discussing how the methods might cope. The issues are chosen to be difficult for at least one of the methods, sometimes all of them. Many are simplified or theoretical studies – whether they are relevant depends on the application and NWP system. For instance several issues become less visible with realistic (larger) observation errors.

## 3.1 Dense but incomplete observations, tracers

Remote sensing usually gives observations that are dense in space and time, but incomplete in that they do not observe all the variables necessary to define a state, for instance satellite soundings of temperature or radar observations of radial wind. Diagnostic relationships such as geostrophy are not accurate enough to help in modern NWP systems. We need to use the prognostic equations, which link space and time gradients of different variables. All the methods discussed are four-dimensional and can do this within the window considered. In principle a Kalman filter can do it between windows because the forecast covariance reflects the positions of previous observations. A localised ensemble cannot however reflect all the detail, so information depending on accurate observed tendencies will not be fully utilised.

An example of this process is the extraction of wind information from a sequence of tracer fields. Lorenc (1988) shown that nonlinear 4DVar could use a sequence of tracer observation in one time-window in a toy model. Daley (1996) showed that a sequential Extended Kalman Filter could recover wind fields from tracers, as long as there were sufficient observations so that the filter-estimated background field stayed close enough to the truth for gradients to be accurate. If the displacements are seen with reference to the background, then linear incremental 4DVar, 4DEnVar or the EnKF can work. Andersson *et al.* (1994) saw impacts on winds in an early version of 4DVar (but did to verify them); Peubey and McNally (2009) measured an impact. This will be discussed more in Mary Forsythe's talk at this Seminar.

Simple spatial localisation causes a problem for 4DEnVar or the EnKF. The severe localisation needed to assimilate dense observations will limit the ability to extract a large-scale wind field – more generally

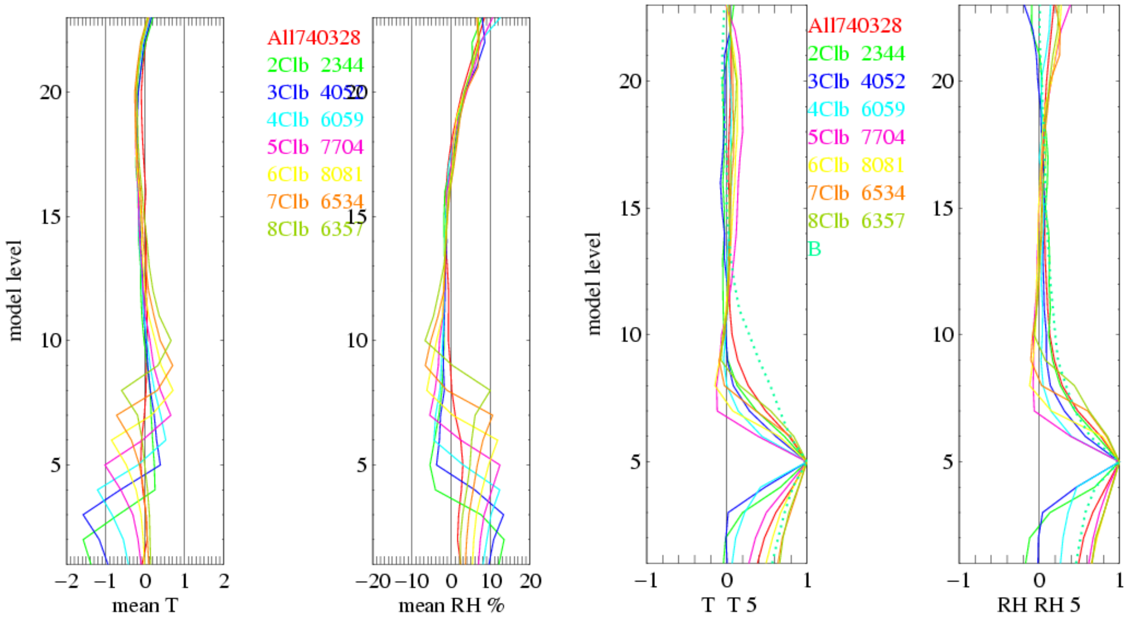

*Figure 10: Left: Mean errors in model T and RH, when composited by the cloudy inversion level in the model. Each curve if for a different inversion level. Right: Correlations of T and RH with level 5, for the same composites (Lorenc, 2007).*

the fine details in the imagery are useful in determining large-scale wind, so it is best to treat all scales together, as variational methods do. 4DEnVar does give the option of scale-dependent localisation (Buehner, 2012). This is research in progress.

### 3.2 Synergistic observations

The advantage of using synergistic observations together was a motivation for ECMWF's first DA scheme, which used a then unprecedented 191 observations at once! Lorenc (1981) table 1 showed the benefit of using surface pressure, thickness and wind together. This is because the model footprint of their observation operators overlapped. Much more relevant today are the overlapping weighting functions of a radiometer, or the use of a vertically integrated observation with an in-situ surface observation. An extreme example is variational bias correction, where the co-location of one biased satellite observation with an accurate in-situ observation allows others some way away to be used more accurately. The benefit of synergistic observations is easiest to see in idealised examples with accurate observations, but likely has a small effect with normal (relatively large) observational errors.

Observation-space localisation effectively alters the $H$ operators and can damage the synergy. Campbell *et al.* (2010) showed that observation space localisation degraded a 1D ensemble DA of radiances. Model-space localisation used in hybrid-4DVar and 4DEnVar is better in this respect.

### 3.3 Cloudy inversions

Lorenc (2007) performed a study of 6 hour forecast profiles from the Met Office global model, compared to collocated radiosondes mapped to model levels. A simple algorithm to detect cloudy inversions was applied, and the results composited for cloud layers. Conclusions were:

1. It is common for the model to have a plausible cloudy inversion structure in the wrong place.

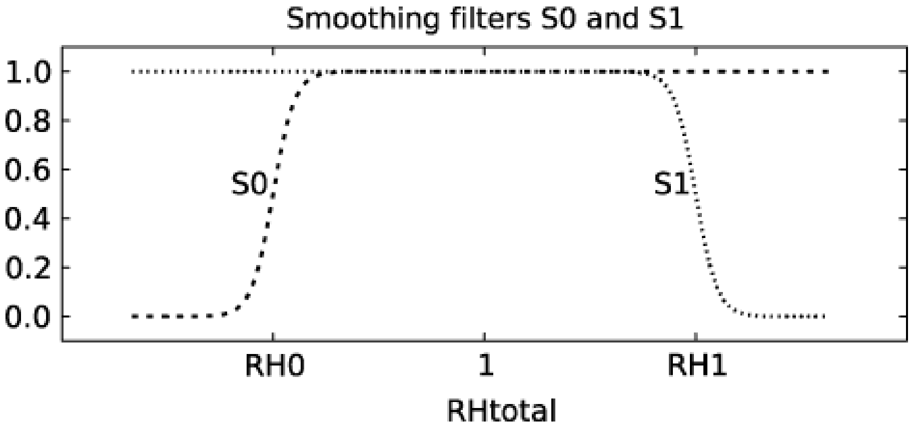

*Figure 11: Extra weights making the penalty function non-quadratic, when assimilation humidity derived from cloud observations (Renshaw and Francis, 2011).*

This leads to error distributions which are non-Gaussian, with the mean of the PDF giving a profile which is unrealistically smooth and apparently biased. This is a fundamental problem of minimum-variance best-estimate methods; it cannot be cured by better covariances within such methods.

2. The climatological error variances are about right for average conditions. But they should be about doubled for temperature and humidity near cloud-topped inversions.

3. The climatological vertical correlations are about right for average conditions, but correlations are too large across cloud-topped inversions, where they should be near zero.

Conclusion 1 applies to all the methods discussed here - it just shows the difficulty of our subject and the importance of the forecast model as discussed in 3.5. Assimilating IR imagery information on cloud tops is likely to remain difficult for all the methods described in this paper.

Conclusions 2 and 3 show the importance of situation-dependent covariances. These could be from ensembles as in the methods discussed here, or cleverer covariance models such as that of Piccolo and Cullen (2011).

### 3.4 Non-Gaussian observed variables

Some observations are "simple" nonlinear functions of the model (e.g. wind speed). Some observations only give limit information on the model state (e.g. the presence or absence of cloud; figure 11). They can be handled by a nonlinear observation operator $H$ in (3) or a variable $\mathbf{R}$ in (2). Variational method need to linearise this, but this can be about a current best estimate, which should be more accurate than the background. In most variational schemes, following Courtier *et al.* (1994), (3) is approximated using the innovations $\underline{\mathbf{d}}$ from the background or guess, as in

$$\underline{\mathbf{y}}^o - \underline{H}\left(\underline{\mathbf{x}}^b\right) = \underline{\mathbf{d}} \tag{38}$$

$$\underline{\mathbf{y}}^o - \underline{\mathbf{y}} = \underline{\mathbf{d}} - \underline{\mathbf{H}}\delta\underline{\mathbf{x}}. \tag{39}$$

This gives a quadratic penalty function, but means $\mathbf{H}$ can only be re-linearised in a full outer-loop. The Met Office variational DA system retains the nonlinear (3), so it can re-linearise $\mathbf{H}$ without rerunning the (expensive) full model.

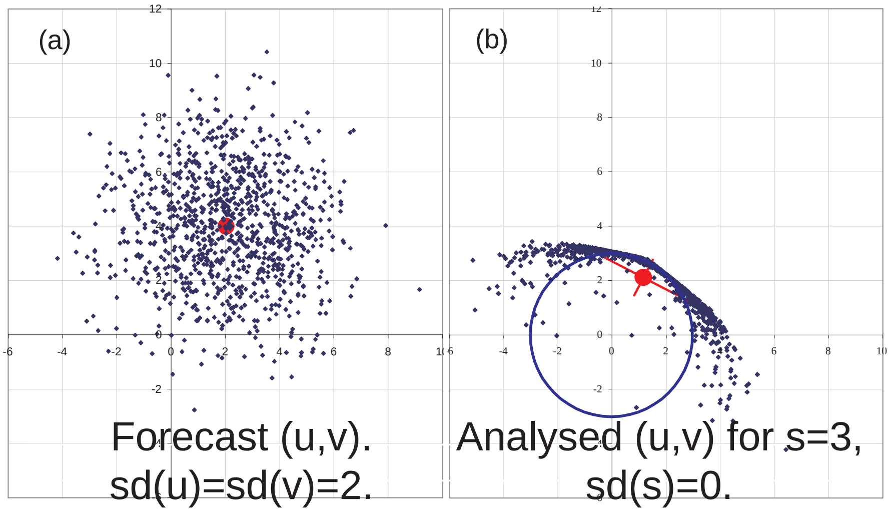

*Figure 12: Idealised EnKF of a perfect wind speed observation. Left: background ensemble. Right: analysis ensemble, showing also ensemble mean and s.d. in red. To fit the observation, the ensemble members should lie on the circle (Lorenc, 2003).*

EnKF method on the other hand effectively use the average linearisation over the ensemble members. This is on average less accurate, as illustrated in figure 12, but makes them more robust in on-off situations – as long as one member has the observed values the EnKF will draw partly towards it whereas variational methods can get stuck because the gradient of zero is zero.

For observations of variables predicted by complex physical paprametrisations in the model (e.g. radar reflectivity) the robustness and simplicity of ensemble approaches of 4DEnVar and the EnKF are an advantage. They simply need the ensemble predictions of the observed parameter (although they work better if it is transformed to have a near-Gaussian distribution). 4DVar requires the model parametrisation to be simplified, regularised (i.e. made more smooth) and linearised – this is much more work.

### 3.5 Initialisation – Spin-up – Staying near the attractor

Although the methods considered here are basically linear, we have to be thankful the world in nonlinear. Otherwise, some perturbations would grow and grow. Instead Abarbanel *et al.* (2010) suggested we only need as many observations as the number of positive conditional Lyapunov exponents of the nonlinear dynamics of the model – a much smaller number than the degrees of freedom of a high-resolution model. As discussed in 3.3 for cloudy inversions, it is impossible for a Gaussian DA scheme to put a structure maintained by nonlinear processes in place just from observations that it is there; we would need detailed coverage. Since model resolution is increasing faster than the number of observations, this problem is increasing – convective scales are more nonlinear (Hohenegger and Schär, 2007). There has been little fundamental research in this area relevant to NWP, yet most DA developers know the technical steps which help keep the assimilated state near the attractor (i.e. in the set of plausible states): only alter the model near observations which disagree with it, initialise the increments to be close to some sort of balance, and allow/encourage the model to adapt towards its attractor.

Of the methods considered in this paper, perhaps 4DVar with a long window has the best chance (although the weak-constraint method of Fisher and Auvinen (2011) loses this by not doing long model runs). More pragmatically, the 4DIAU is addressing the issue for 4DEnVar (Lorenc *et al.*, 2014).

## 4 My Personal Conclusions

Nearly 6 years ago in Buenos Aires at the WWRP/THORPEX Workshop on 4D-Var and Ensemble Kalman Filter Intercomparisons, I said that global 4DVar was good for perhaps a decade. While that was probably right, I would not repeat the prediction now. I think the scientific advantages of 4DVar are decreasing and will not in the long-term outweigh the increasing technical difficulties of parallelisation and software maintenance. As radically new models aimed at efficient parallelisation are introduced, centres will not judge it worthwhile to develop 4DVar. This has already happened in Canada (Buehner *et al.*, 2013) – at NCEP they never implemented 4DVar and are happy to by-pass it (Kleist and Ide, 2014a,b).

At the convective-scale which will be the research focus for the next decade, ensemble systems are essential because of the forecast uncertainty. The difficulty of developing accurate models, accessing and processing enough observations and the sheer computational cost will dominate. Perhaps this will militate in favour of the simple EnKF, nevertheless my personal favourite is 4DEnVar because of its ability to handle large-scale errors by also using a global ensemble, and for consistency with the global system.

At longer time-scales who knows? Computers may continue to increase at a rate that enables convective-scale global NWP (although I believe global NWP may fall further behind the leaders, so resources for global NWP do not increase as fast as in the past). Unless we have a corresponding increase in observations, I predict that convective-scale global NWP models with be practicable long before we have DA methods to initialise all the scales they resolve.

## References

Abarbanel HDI, Kostuk M, Whartenby W. 2010. Data assimilation with regularized nonlinear instabilities. *Q. J. R. Meteorol. Soc.* **136**(648): 769–783, doi:http://dx.doi.org/10.1002/qj.600.

Andersson E, Pailleux J, Thépaut JN, Eyre JR, McNally AP, Kelly GA, Courtier P. 1994. Use of cloud-cleared radiances in three/four-dimensional variational data assimilation. *Q. J. R. Meteorol. Soc.* **120**: 627–653.

Aulignè T. 2012. An integrated ensemble/variational hybrid data assimilation. Presented at International Conference on Ensemble Methods in Geophysical Sciences Toulouse, France. 12-16 November 2012.

Bishop CH, Hodyss D. 2009a. Ensemble covariances adaptively localized with ECO-RAP. part 1: tests on simple error models. *Tellus A* **61**: 84–96, doi:http://dx.doi.org/10.1111/j.1600-0870.2008.00371.x.

Bishop CH, Hodyss D. 2009b. Ensemble covariances adaptively localized with ECO-RAP. part 2: a strategy for the atmosphere. *Tellus A* **61**: 97–111, doi:http://dx.doi.org/10.1111/j.1600-0870.2008.00372.x.

Bonavita M, Isaksen L, Holm E. 2012. On the use of EDA background error variances in the ECMWF 4D-Var. *Q. J. R. Meteorol. Soc.* **138**: 1540–1559, doi:http://dx.doi.org/10.1002/qj.1899.

Buehner M. 2012. Evaluation of a spatial/spectral covariance localization approach for atmospheric data assimilation. *Mon. Weather Rev.* **140**: 617–636, doi:http://dx.doi.org/10.1175/MWR-D-10-05052.1.

Buehner M, Charron M. 2007. Spectral and spatial localization of background-error correlations for data assimilation. *Q. J. R. Meteorol. Soc.* **133**: 615–630, doi:http://dx.doi.org/10.1002/qj.50.

Buehner M, Morneau J, Charette C. 2013. Four-dimensional ensemble-variational data assimilation for global deterministic weather prediction. *Nonlinear Processes in Geophysics* **20**: 669–682, doi:10.5194/npg-20-669-2013, URL `http://www.nonlin-processes-geophys.net/20/669/2013/`.

Campbell WF, Bishop CH, Hodyss D. 2010. Vertical Covariance Localization for Satellite Radiances in Ensemble Kalman Filters. *Mon. Weather Rev.* **138**(1): 282–290.

Clayton AM, Lorenc AC, Barker DM. 2013. Operational implementation of a hybrid ensemble/4D-Var global data assimilation system at the Met Office. *Q. J. R. Meteorol. Soc.* **139**: 1445–1461, doi: http://dx.doi.org/10.1002/qj.2054.

Courtier P, Thépaut JN, Hollingsworth A. 1994. A strategy for operational implementation of 4D-Var, using an incremental approach. *Q. J. R. Meteorol. Soc.* **120**: 1367–1387, doi:http://dx.doi.org/10.1002/qj.49712051912.

Daley R. 1996. Recovery of the one and two dimensional windfields from chemical constituent observations using the constituent transport equation and an extended Kalman filter. *Met. and Atmos. Phys.* **60**: 119–136.

Dee DP. 1991. Simplification of the Kalman filter for meteorological data assimilation. *Q. J. R. Meteorol. Soc.* **117**: 365–384.

Derber J, Bouttier F. 1999. A reformulation of the background error covariance in the ECMWF global data assimilation system. *Tellus A* **51A**: 195–221.

Etherton BJ, Bishop CH. 2004. Resilience of hybrid ensemble/3DVAR analysis schemes to model error and ensemble covariance error. *Mon. Weather Rev.* **132**: 1065–1080, doi:http://dx.doi.org/10.1175/1520-0493(2004)132⟨1065:ROHDAS⟩2.0.CO;2.

Fisher M. 2003. Background error covariance modelling. In: *Seminar on Recent developments in data assimilation for atmosphere and ocean, 8-12 September 2003*. ECMWF: Shinfield Park, Reading, pp. 45–64.

Fisher M, Auvinen H. 2011. Long window weak-constraint 4D-Var. In: *ECMWF Seminar on Data assimilation for atmosphere and ocean, 6 - 9 September 2011*. ECMWF: Shinfield Park, Reading, URL http://old.ecmwf.int/publications/library/ecpublications/_pdf/seminar/2011/Fisher.pdf.

Ford R, Glover M, Ham D, Maynard C, Pickles S, Riley G. 2013. GungHo phase 1 - computational science recommendations. Technical Report 587, Met Office, URL http://www.metoffice.gov.uk/media/pdf/8/o/FRTR587Tagged.pdf. http://www.metoffice.gov.uk/media/pdf/8/o/FRTR587Tagged.pdf.

Golding BW. 2009. Long lead time flood warnings: reality or fantasy? *Meteorological Applications* **16**(1): 3–12, doi:10.1002/met.123, URL http://dx.doi.org/10.1002/met.123.

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.

Harlim J, Hunt BR. 2007. Four-dimensional local ensemble transform Kalman filter: numerical experiments with a global circulation model. *Tellus A* **59**(5): 731–748, doi:http://dx.doi.org/10.1111/j.1600-0870.2007.00255.x.

Hohenegger C, Schär C. 2007. Atmospheric predictability at synoptic versus cloud-resolving scales. *Bull. Am. Meteorol. Soc.* **88**(11): 1783–1793.

Houtekamer P, Deng X, Mitchell HL, Baek SJ, Gagnon N. 2014. Higher resolution in an operational ensemble Kalman filter. *Mon. Weather Rev.* **142**(3): 1143–1162.

Houtekamer PL, Mitchell HL. 2001. A sequential ensemble Kalman filter for atmospheric data assimilation. *Mon. Weather Rev.* **129**: 123–137, doi:http://dx.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-NONLINEAR PHENOMENA* **230**: 112–126, doi:10.1016/j.physd.2006.11.008.

Ide K, Courtier P, Ghil M, Lorenc A. 1997. Unified notation for data assimilation: Operational, sequential and variational. *J. Met. Soc. of Japan* **75**: 181–189.

Kleist DT, Ide K. 2014a. An OSSE-based evaluation of hybrid variational-ensemble data assimilation for the NCEP GFS, part i: System description and 3d-hybrid results. *Mon. Weather Rev.* **submitted**.

Kleist DT, Ide K. 2014b. An OSSE-based evaluation of hybrid variational-ensemble data assimilation for the NCEP GFS, part ii: 4d envar and hybrid variants. *Mon. Weather Rev.* **submitted**.

Kondo K, Tanaka HL. 2009. Applying the local ensemble transform kalman filter to the nonhydrostatic icosahedral atmospheric model (NICAM). *SOLA* : 121–124doi:10.2151/sola.2009-031.

Lorenc A. 1988. Optimal nonlinear objective analysis. *Q. J. R. Meteorol. Soc.* **114**: 205–240, doi:http://dx.doi.org/10.1002/qj.49711447911.

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

Lorenc AC. 2003. The potential of the ensemble Kalman filter for NWP - a comparison with 4D-Var. *Q. J. R. Meteorol. Soc.* **129**: 3183–3203, doi:http://dx.doi.org/10.1256/qj.02.132.

Lorenc AC. 2007. A study of o-b monitoring statistics from radiosondes, composited for low-level cloud layers. Met Office Forecasting Research Tech. Rept. 504, Met Office, URL http://www.metoffice.gov.uk/archive/forecasting-research-technical-report-504.

Lorenc AC, Ballard SP, Bell RS, Ingleby NB, Andrews PLF, Barker DM, Bray JR, Clayton AM, Dalby T, Li D, Payne TJ, Saunders FW. 2000. The Met. Office global three-dimensional variational data assimilation scheme. *Q. J. R. Meteorol. Soc.* **126**: 2991–3012, doi:http://dx.doi.org/10.1002/qj.49712657002.

Lorenc AC, Bowler NE, Clayton AM, Fairbairn D, Pring SR. 2014. Comparison of hybrid-4DEnVar and hybrid-4DVar data assimilation methods for global NWP. *Mon. Weather Rev.* **accepted**: 19pp.

Ménétrier B, Montmerle T, Berre L, Michel Y. 2014. Estimation and diagnosis of heterogeneous flow-dependent background-error covariances at the convective scale using either large or small ensembles. *Q. J. R. Meteorol. Soc.* doi:10.1002/qj.2267.

Montmerle T, Berre L. 2010. Diagnosis and formulation of heterogeneous background-error covariances at the mesoscale. *Q. J. R. Meteorol. Soc.* **136**: 1408–1420, doi:http://dx.doi.org/10.1002/qj.655.

Ota Y, Derber JC, Kalnay E, Miyoshi T. 2013. Ensemble-based observation impact estimates using the NCEP GFS. *Tellus A* **65**, doi:http://dx.doi.org/10.3402/tellusa.v65i0.20038.

Peubey C, McNally A. 2009. Characterization of the impact of geostationary clear-sky radiances on wind analyses in a 4d-var context. *Q. J. R. Meteorol. Soc.* **135**(644): 1863–1876.

Piccolo C, Cullen M. 2011. Adaptive mesh method in the met office variational data assimilation system. *Q. J. R. Meteorol. Soc.* **137**(656): 631–640, doi:http://dx.doi.org/10.1002/qj.801.

Purser RJ, Wu WS, Parrish DF, Roberts NM. 2003a. Numerical aspects of the application of recursive filters to variational statistical analysis. part i: Spatially homogeneous and isotropic gaussian covariances. *Mon. Weather Rev.* **131**: 1524–1535, doi:http://dx.doi.org/10.1175//1520-0493(2003)131⟨1524:NAOTAO⟩2.0.CO;2.

Purser RJ, Wu WS, Parrish DF, Roberts NM. 2003b. Numerical aspects of the application of recursive filters to variational statistical analysis. part ii: Spatially inhomogeneous and anisotropic general covariances. *Mon. Weather Rev.* **131**: 1536–1548, doi:http://dx.doi.org/10.1175//2543.1.

Rawlins F, Ballard SP, Bovis KJ, Clayton AM, Li D, Inverarity GW, Lorenc AC, Payne TJ. 2007. The Met Office global 4-dimensional variational data assimilation system. *Q. J. R. Meteorol. Soc.* **133**: 347–362, doi:http://dx.doi.org/10.1002/qj.32.

Raynaud L, Berre L, Desroziers G. 2009. Objective filtering of ensemble-based background-error variances. *Q. J. R. Meteorol. Soc.* **135**(642): 1177–1199, doi:http://dx.doi.org/10.1002/qj.438.

Renshaw R, Francis PN. 2011. Variational assimilation of cloud fraction in the operational met office unified model. *Q. J. R. Meteorol. Soc.* **137**(661): 1963–1974, doi:http://dx.doi.org/10.1002/qj.980.

Satoh M, Matsuno T, Tomita H, Miura H, Nasuno T, Iga Si. 2008. Nonhydrostatic icosahedral atmospheric model (nicam) for global cloud resolving simulations. *Journal of Computational Physics* **227**(7): 3486–3514.

Skamarock W, Klemp J, Duda M, Fowler L, Park S. 2010. Global nonhydrostatic modeling using voronoi meshes: the mpas model. In: *Proceedings of the ECMWF Workshop on Non-hydrostatic Modelling, Reading*. pp. 8–10.

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