Relative merits of 4D-Var and Ensemble Kalman Filter

Andrew C Lorenc

Met Office, FitzRoy Road, Exeter, EX1 3PB, United Kingdom
Andrew.Lorenc@metoffice.com

Abstract

Incremental 4D-Var and the ensemble Kalman filter have different approaches to approximating the time-evolution of the covariances which approximately characterise the probability distribution of possible states. The two methods are compared for application to NWP, both for their expected properties, and for ease of application.

1. Introduction

No practical data assimilation scheme can be fully optimal; approximations are essential. The most appropriate will depend on the application. In this paper I limit myself to comparing schemes for application to future numerical weather prediction (NWP) configurations, as envisaged for a national meteorological service such as the Met Office. In the literature one can find several "flavours" of four-dimensional variational assimilation (4D-Var), and many of ensemble Kalman filters (EnKF); I limit the main part of this comparison to incremental 4D-Var like that being developed in the Met Office (described further in section 2), and EnKF methods with a moderate ensemble size and covariance localisation (described further in section 3).

An ideal assimilation method would consider an arbitrary probability distribution function (pdf) describing our knowledge about the current state (figure 1). However for NWP it would be impossible even to define such a pdf, let alone calculate its time-evolution. Incremental 4D-Var and the EnKF both make (in different ways) the assumption that the pdf which is time-evolved can be approximated by a Gaussian, and hence represented by its mean and covariance. So the key to this comparison is their different methods of representing, and calculating the time-evolution, of covariances. Section 4 compares the expected properties of 4D-Var and EnKF is this light, while section 5 discusses more practical details of their implementation. Section 6 discusses the future.

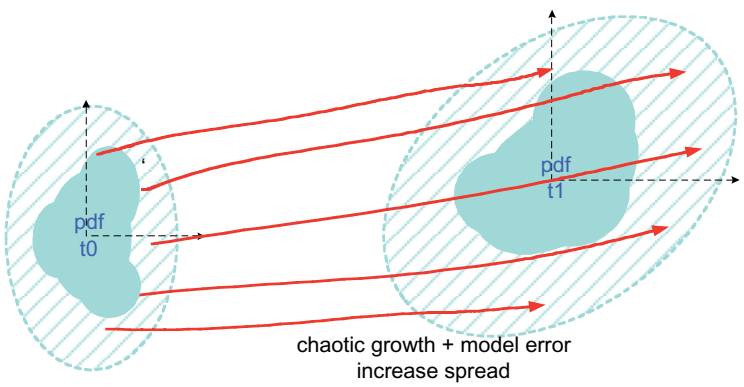

Figure 1 Schematic showing the ideal time-evolution of a general pdf

2. Incremental 4D-Var

Incremental 4D-Var uses a simplified linear model to calculate the evolution of perturbations to a (more detailed) estimate of the atmospheric state, and searches for the perturbation which gives the best fit to observations and prior knowledge. It was introduced by Courtier et al. (1994), largely justified as a cost-saving approach for making 4D-Var practicable. Lorenc (2003a) derives it differently; the simplified perturbation forecast (PF) model is introduced as part of a transformation describing the four-dimensional structure of error covariances, an extension of the transformed control variable approach used in 3D-Var.

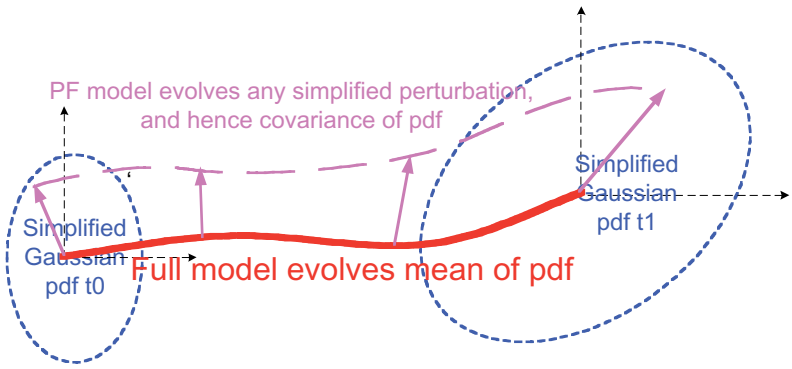

Figure 2 Schematic showing the modelled 4D pdf in 4D Var.

Incremental three-dimensional variational assimilation (3D-Var) finds the increment $\delta\mathbf{x}$ to a guess field $\mathbf{x}^g$ which minimises

$$J(\delta\mathbf{x}) = \tfrac{1}{2}(\delta\mathbf{x} - \delta\mathbf{x}^b)^T \mathbf{B}_{(\mathbf{x})}^{-1}(\delta\mathbf{x} - \delta\mathbf{x}^b) + \tfrac{1}{2}(\mathbf{y} - \mathbf{y}^o)^T \mathbf{R}^{-1}(\mathbf{y} - \mathbf{y}^o), \tag{1}$$

where $\delta\mathbf{x}^b$ is the difference between the background forecast $\mathbf{x}^f$ and the guess $\mathbf{x}^g$, and

$$\mathbf{x} = \mathbf{x}^g + \delta\mathbf{x}\,. \tag{2}$$

The model estimate of the observed values is given by

$$\mathbf{y} = H(\mathbf{x}). \tag{3}$$

Instead of direct minimisation of (1), it is normal to transform control variables, finding $\mathbf{v}$ which minimises

$$J(\mathbf{v}) = \tfrac{1}{2}\mathbf{v}^T\mathbf{v} + \tfrac{1}{2}(\mathbf{y} - \mathbf{y}^o)^T \mathbf{R}^{-1}(\mathbf{y} - \mathbf{y}^o), \tag{4}$$

$$\mathbf{x} = \mathbf{x}^g + \mathbf{U}\mathbf{v}\,. \tag{5}$$

If we choose $\mathbf{U}$ such that

$$\mathbf{B}_{(\mathbf{x})} = \mathbf{U}\mathbf{U}^T \tag{6}$$

then minimising (4) is equivalent to minimising (1). But in practice we do not have an independent means of representing $\mathbf{B}$, and (6) is better thought of as a way of modelling the three-dimensional multivariate structure of $\mathbf{B}$, via a further factorisation of $\mathbf{U}$ (e.g. Lorenc et al. 2000). The same approach can be

extended to the time dimension. Consider a finite time interval covering n assimilation cycles, and define a set of four-dimensional vectors denoted by underlining, e.g.

$$\underline{\mathbf{x}} = \begin{pmatrix} \mathbf{x}(t_0) \\ \mathbf{x}(t_1) \\ \vdots \\ \mathbf{x}(t_{n-1}) \end{pmatrix}. \tag{7}$$

Then, with appropriate extensions to the interpolations in H and $\mathbf{H}$, the four-dimensional versions of (1) and (3) have identical form:

$$J(\delta\underline{\mathbf{x}}) = \tfrac{1}{2}\left(\delta\underline{\mathbf{x}} - \delta\underline{\mathbf{x}}^b\right)^{\mathrm{T}} \mathbf{B}_{(\underline{\mathbf{x}})}^{-1}\left(\delta\underline{\mathbf{x}} - \delta\underline{\mathbf{x}}^b\right) + \tfrac{1}{2}\left(\underline{\mathbf{y}} - \underline{\mathbf{y}}^o\right)^{\mathrm{T}} \mathbf{R}^{-1}\left(\underline{\mathbf{y}} - \underline{\mathbf{y}}^o\right), \tag{8}$$

$$\underline{\mathbf{x}} = \underline{\mathbf{x}}^g + \delta\underline{\mathbf{x}}\,. \tag{9}$$

The quadratic form of (8), using the four-dimensional covariance $\mathbf{B}_{(\underline{\mathbf{x}})}$, is only appropriate if the pdf is approximately Gaussian, and the time-evolution of perturbations approximately linear. Modern NWP models are capable of representing small scales which can grow rapidly (compared to the time-window of 4D-Var), before the growth is limited as they use up their forcing. 4D-Var will not work with such a model. It is necessary to filter out scales with nonlinear, potentially chaotic, behaviour. We do this with a simplification operator S. For an NWP model the most obvious simplification is a reduction in resolution (this was the explanation for S given by Ide *et al*. 1997), but it is also useful to include simplifications to the representation of physical processes such as cloud. We introduce a simplified model state $\mathbf{w}$, and a simplification operator S that is usually linear, but may have nonlinear terms, for instance in the treatment of moisture variables. We are assuming that the desired analysis can be expressed in term of increments to this state:

$$\begin{aligned} \mathbf{w}^g &= S(\mathbf{x}^g) \\ \mathbf{w} &= S(\mathbf{x}) \\ \delta\mathbf{w} &= S(\mathbf{x}) - S(\mathbf{x}^g) \end{aligned}\,. \tag{10}$$

To use the simplified increment we need to be able to add it to $\mathbf{x}^g$:

$$\mathbf{x} = \mathbf{x}^g + S^{-I}\delta\mathbf{w}\,. \tag{11}$$

The incrementing operator S^{-I} is the generalised inverse of S; it needs extra assumptions about what to do with the filtered scales and variables. The simplification is chosen such that the time-evolution of the filtered perturbation is approximately linear; we can use a perturbation forecast (PF) model $\underline{\mathbf{M}}$ to calculate a four-dimensional $\delta\underline{\mathbf{w}}$. We use S^{-I} (applied at each time) and $\underline{\mathbf{M}}$, along with the 3D-Var control variable transform $\mathbf{U}$, to define a four-dimensional control variable transform:

$$\underline{\mathbf{x}} = \underline{\mathbf{x}}^g + S^{-I}\,\underline{\mathbf{M}}\mathbf{U}\mathbf{v}\,. \tag{12}$$

So simplified incremental 4D-Var replaces the minimisation of (8) by minimising:

$$J(\mathbf{v}) = \tfrac{1}{2}\mathbf{v}^T\mathbf{v} + \tfrac{1}{2}\left(\underline{\mathbf{y}} - \underline{\mathbf{y}}^o\right)^T \mathbf{R}^{-1}\left(\underline{\mathbf{y}} - \underline{\mathbf{y}}^o\right). \tag{13}$$

Instead of the (impracticable) full nonlinear evolution of the non-Gaussian pdf, incremental 4D-Var chooses a simplified sub-set of variables whose four-dimensional pdf can be approximated by a Gaussian, and evolves them over a finite time-window using a linear PF model, to give an implicit four-dimensional covariance model. From the simplified variables at any time it can diagnose increments to the full model using S^{-I}, and calculate model predictions of the observations using H; these latter operations may be nonlinear (as in 3D-Var), as long as the minimisation of (13) is possible.

3. Ensemble Kalman Filter

The basic idea of the EnKF (Evensen 1994) is to construct an ensemble of forecast states $\{\mathbf{x}_i^f\}(i = 1,\ldots,N)$ such that the mean of the ensemble is the best estimate of the population mean, and the sample covariance is a good estimate of the forecast error covariance of this best estimate.

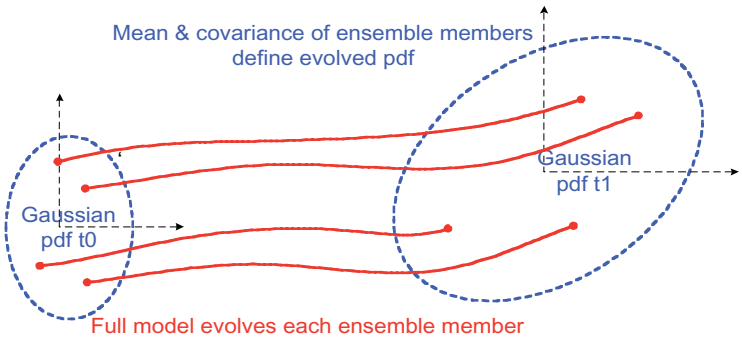

Figure 3 Schematic showing the time-evolution of the pdf in the EnKF

It is convenient to introduce a matrix notation for the ensemble, with $\mathbf{X}$ having as columns the normalised deviations from the ensemble mean:

$$\mathbf{X} = \frac{1}{\sqrt{N-1}}\left(\mathbf{x}_1 - \overline{\mathbf{x}} \quad \mathbf{x}_2 - \overline{\mathbf{x}} \quad \cdots \quad \mathbf{x}_N - \overline{\mathbf{x}}\right). \tag{14}$$

These perturbation matrices are estimates of a square-root of the covariances:

$$\mathbf{P}_e^f = \mathbf{X}^f \mathbf{X}^{fT}\,. \tag{15}$$

The time evolution of the covariances is done by evolving all the individual ensemble members using the full NWP model (figure 3). For this to be affordable for the NWP models and computers envisaged for the next decade, we are probably limited to ensemble sizes O(100). Lorenc (2003b) showed that this can give significant sampling errors in the estimated covariances; covariances near zero are regularly estimated as large as 0.2 (figure 4). If (as in NWP systems) the modelling domain is large compared to the typical correlation scale, then there will be many covariances near to zero, and the sampling errors will degrade the analysis.

To overcome this we use covariance localisation. A correlation function $\mathbf{C}$ is constructed which is near one where the ensemble covariances are significant, but which is zero where they are expected to be mainly noise. The Kalman filter equations are then implemented using a covariance formed as an element by element Schur (aka Hadamard) product between these matrices (figure 5):

$$\mathbf{P}^f \approx \mathbf{C} \circ \mathbf{P}_e^f \ . \tag{16}$$

The Kalman filter equations contain terms $\mathbf{H}^T\mathbf{P}^f\mathbf{H}$ and $\mathbf{H}^T\mathbf{P}^f$, which can be thought of as error covariances involving the predictions of the observed variables. They can be obtained by augmenting the model variables with $\mathbf{y} = H(\mathbf{x})$, and generalising (14) (15) and (16). That is, we apply the (possibly nonlinear) observation operators to each ensemble member, and calculate covariances involving $\mathbf{y}$. Note that this is treating the pdf of $\mathbf{y}$ as Gaussian, whereas 4D-Var allows for non-Gaussian pdf caused by nonlinear H.

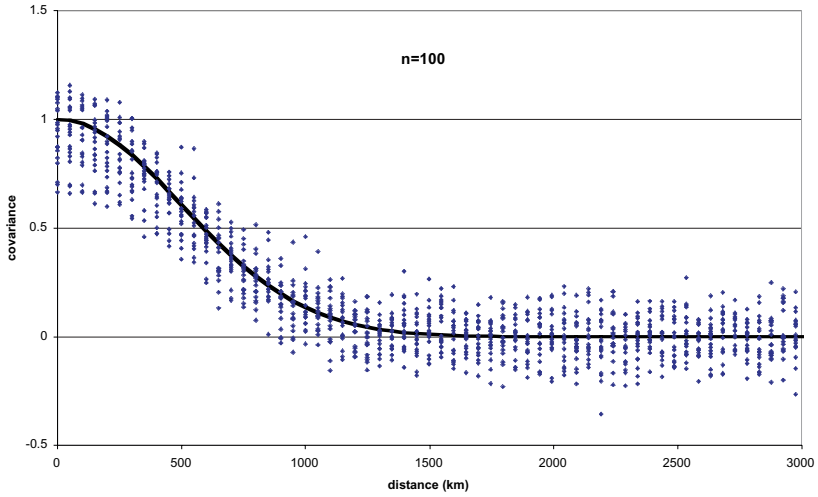

Figure 4 Ensemble estimate covariances with a point at the origin, for ensembles of 100 samples taken from a population whose covariance, shown in the solid curve, is $\exp\left(-r^2/2a^2\right)$, *with a = 500 km. (Adapted from Lorenc 2003b).*

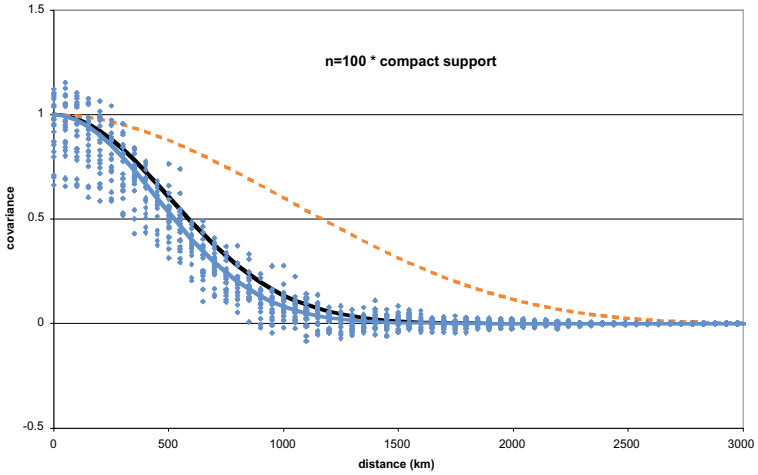

Figure 5 As figure 4, with the ensemble covariances modified with the Schur product correlation C (shown dashed). The Schur product of C with the population covariance is shown in the lowest line. (Adapted from Lorenc 2003b).

As well as using these covariances to estimate the best analysis, the ensemble Kalman filter also produces an analysis ensemble, approximating the analysis error covariance. (This is an advantage over 4D-Var, which never explicitly represents the covariances, and does not readily generate an ensemble.) There are two approaches used:

- In the original EnKF, an ensemble of analyses are produced each using a different set of pseudo-random perturbations on the observed values.
- In the ensemble square-root filter methods (EnSRF: Tippett et al. 2003), the forecast ensemble perturbations $\mathbf{X}^f$ are directly transformed into an analysis ensemble $\mathbf{X}^f$. This reduces errors for small ensembles due to noisy estimates of covariances (Whitaker and Hamill 2002). If we are to use covariance localisation, the transform method is only practicable if observations are processed sequentially.

4. Assimilation characteristics

From their different modelling of the covariances, as described above, we can compare the expected characteristics of 4D-Var and EnKF assimilations. We do this with the help of Table 1, adapted from Lorenc (2003b).

Table 1 Summary of the assimilation characteristics of the ensemble Kalman filter compared with those of 4D-Var.

	Incremental 4D-Var	EnKF
Forecast covariances	Modelled at t_0 (usually isotropic), time evolution for a finite time-window represented by linear and adjoint models.	Sampled by ensemble (flow-dependent). Noisy: must be modified to have compact support using Schur product.
Ability to fit detailed observations.	Limited by resolution of simplified model. Tendencies fitted within time-window.	Limited to fewer data (in a region) than ensemble members. Tendency information only extracted if obs properly fitted.
Balance constraints	Can be imposed through a dynamical design to the variable transform, or a separate balance penalty.	Only imposed if each forecast in the ensemble is balanced. Lost slightly in Schur product.
Nonlinear observation operators	Allowed if differentiable. (Results uncertain if pdf is bimodal in range of interest.)	Allowed, but resulting pdf modelled by Gaussian.
Non-Gaussian observational errors	Allowed if differentiable. (Results uncertain if pdf is bimodal in range of interest.)	Not allowed. Prior QC step is needed.

4D-Var only represents the time-evolution of covariances over a short time-window, limited so that the linear PF model remains a good approximation. The EnKF, like the Kalman filter, is designed to continue indefinitely.

The EnKF covariances are noisy. Unless a very large ensemble is used, or the analysis domain is limited, then steps must be taken to avoid using noisy estimates of the many covariances which are actually near zero. Introducing a Schur product covariance $\mathbf{C}$ is a convenient way of doing this. But if, because of a small ensemble, it is relied on significantly to modify many covariances, $\mathbf{C}$ becomes an arbitrary tuning

factor. Note that $\mathbf{C}$ must also define inter-variable correlations, including the augmented model variables such as radiances, so its definition is not trivial.

4D-Var can only fit information resolved by the PF model. The basic EnKF can only fit as many independent pieces of information as there are ensemble members. Covariance localisation relaxes this constraint, but the way the algorithm behaves when trying to fit dense detailed observations is determined by the (arbitrary) covariance $\mathbf{C}$.

Variational methods can easily incorporate (linear) balance constraints, either directly in the covariance model (by giving unbalanced modes small or zero variances), or in an additional so-called J_c term which can be thought of as modifying the covariance model. If the ensemble members are balanced, the ensemble covariance will similarly have small or zero variance in unbalanced modes. But, because of the small ensemble size, it also has small or zero variance in some important balanced modes, and we have to apply covariance localisation to allow these to be analysed. The localisation also removes the balance constraint.

Variational methods can cope with well-behaved nonlinearities in H, and with non-Gaussian observational error pdfs. By simultaneously handling these effects for nearby observations, variational methods are capable of allowing for the interaction of their non-Gaussian pdfs. For instance scatterometer winds can be de-aliased consistently, and quality control decisions take account of nearby observations which support each other. The EnKF can use a nonlinear H, but it treats the resulting pdf as a Gaussian. (Figure 6 gives an example.) So it cannot properly allow for the mutual support of non-Gaussian information from adjacent observations.

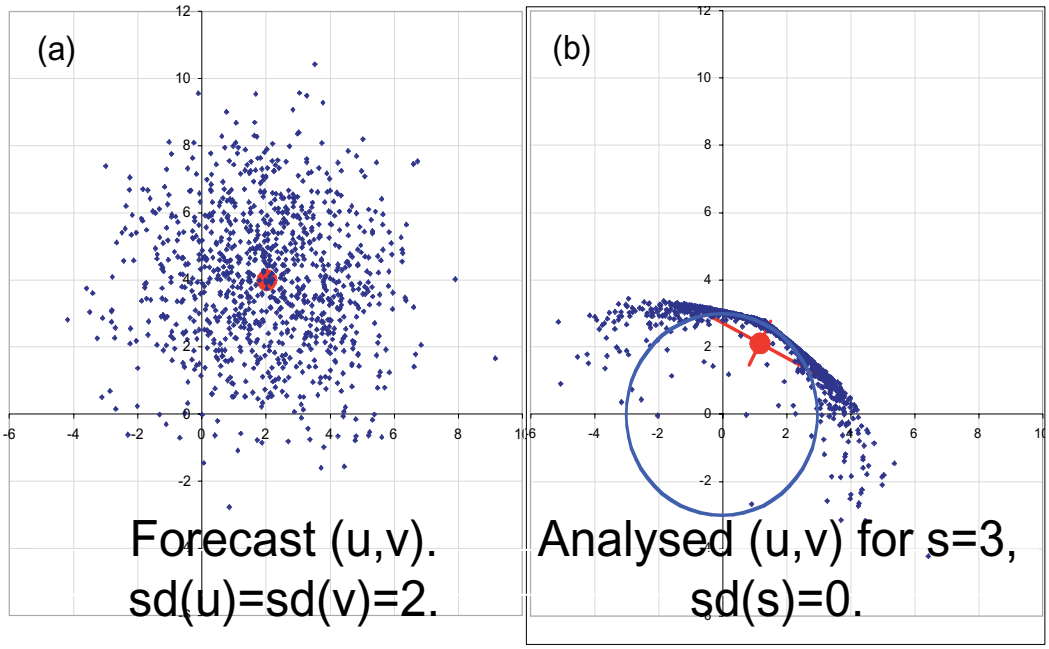

Figure 6(a) 1000 member ensemble of forecast wind vectors, with mean (2,4) and standard deviation (2,2). (b) Ensemble of wind vectors resulting from an EnKF analysis of an error-free observation of wind speed; the circle shows the observed value, 3. The large dots show the mean of each distribution, and the straight lines in (b) show its eigenvectors. If, with this observation, 4D Var had other incomplete information such as a perfect observation of wind direction, it could combine them optimally, whereas the EnKF could not since the pdfs are non-Gaussian. (Adapted from Lorenc 2003b).

4D-Var represents the evolving pdf by its mean (evolved by the main NWP model), and a covariance (evolved by the PF model). This approach gets into difficulties if we have an NWP model which can represent significant features about which we are uncertain. For instance if there are scattered showers, but

we have no radar and do not know precisely where they are, then the mean state has a smooth distribution of rainfall, unlike any actual realisation. It would be difficult to design a model which could represent and evolve such a mean state, and also represent and evolve the actual showers when we have detailed observations. If we have significant uncertainty in the model state, then ensemble methods are a natural approach. The forecast model need only evolve actual states; the smooth mean state in uncertain situations comes from the ensemble mean.

5. Ease of implementation

We do not currently have models which evolve the mean of the pdf, as described above. Incremental 4D-Var makes do with an operational NWP model, and a simplified PF model. Often the PF model is based on the tangent-linear version of the full model, with simplification coming from a reduced resolution and omission of some processes. We also require the adjoint of the PF model. Developing a model for the EnKF is much simpler; the model can be physically based to evolve actual realisations of resolved features.

Assuming we need O(100) iterations in 4D-Var, and O(100) members in the EnKF, the costs of running the model should be similar. But the 4D-Var runs are done sequentially, so each PF and adjoint model run must be parallelised, whereas the EnKF can be implemented with each model running independently. Such an EnKF needs to redistribute the model states over the processors every time covariances are needed from (15); 4D-Var needs a similar redistribution in its covariance model each iteration.

The processing of observations can proceed independently, in parallel in 4D-Var. In the EnSRF they must be processed sequentially - possibly a problem for dense observations. If the perturbed observation EnKF is used, then batches of observations can be processed in parallel (Keppenne et al. 2002), but the attractive simplicity of the EnSRF algorithm is lost.

In principle the EnKF can be run indefinitely. In practice some tuning is needed to make sure that the covariances stay realistic; this may take the form of a simple inflation, and/or some other method of allowing for model errors. 4D-Var needs an independent method of calculating parameters in its error covariance model (this may even be an EnKF).

In mesoscale NWP there are usually significant uncertainties in the larger scales. Equivalently, a limited-area model (LAM) forecast can have significant errors propagating across its boundaries, eventually affecting the errors for the whole area. 4D-Var only runs for a finite time-window, and the area can be chosen to be sufficiently large that this propagation does not occur. Simple methods of characterising boundary errors then suffice. But if one attempted to run a LAM EnKF indefinitely, then the specification of the boundary uncertainties would eventually dominate the interior covariances. Probably a LAM EnKF would need to be nested in a global ensemble; little has been published on this.

6. The way forward

There is a tendency for NWP systems to expand to fulfil many simultaneous roles. We can afford to run high-resolution forecast models to several days, and would like such a forecast to give a good "nowcast" of short-period details, as well as a good forecast of synoptic scales. We want to perform detailed quality control of individual observations, at the same time as a global estimation of satellite biases. We want to fit detailed moisture fields from remotely sensed data, and deduce the larger-scale advection fields from their observed motion. Variational methods have facilitated this tendency; much of the recent improvement in global NWP is due to variational processing of satellite data (Simmons and Hollingsworth, 2002). A single EnKF could never have such a scope. The way forward for such a comprehensive NWP system is to

enhance the 4D-Var approach with some of the attributes of the EnKF, so that some aspects of the covariances can be evolved indefinitely. This is an active area of research.

But such a system gets ever more complex, and has to make compromises. For particular tasks with a more restricted scope the EnKF is simpler, and may behave better. For instance it has been proved to work well for re-analysis of sparse observations (Whitaker et al. 2003), and in convective scale assimilation when the boundaries and larger-scales are unimportant (Snyder and Zhang, 2003). Other tasks such as quality control and bias correction can also be done by discrete systems. NWP would require a range of systems, many using ensembles; the complexity problem then being how to link them together consistently.

References

Courtier, P., J-N. Thepaut, and A. Hollingsworth, 1994: "A strategy for operational implementation of 4D-Var, using an incremental approach". *Quart. J. Roy. Met. Soc.*, **120**, 1367-1387.

Keppenne, Christian L., Rienecker, Michele M. 2002: "Initial Testing of a Massively Parallel Ensemble Kalman Filter with the Poseidon Isopycnal Ocean General Circulation Model" *Mon. Wea. Rev.*, **130**, 2951-2965

Lorenc, A. C., S. P. Ballard, R. S. Bell, N. B. Ingleby, P. L. F. Andrews, D. M. Barker, J. R. Bray, A. M. Clayton, T. Dalby, D. Li, T. J. Payne and F. W. Saunders. 2000: The Met. Office Global 3-Dimensional Variational Data Assimilation Scheme. *Quart. J. Roy. Met. Soc.*, **126**, 2991-3012.

Ide, K., Courtier, P., Ghil, M., and Lorenc, A.C. 1997: "Unified notation for data assimilation: Operational, Sequential and Variational" *J. Met. Soc. Japan*, Special issue "Data Assimilation in Meteorology and Oceanography: Theory and Practice." **75**, No. 1B, 181--189

Lorenc, A. C., 2003(a): Modelling of error covariances by four-dimensional variational data assimilation. Accepted by *Quart. J. Roy. Met. Soc.*.

Lorenc, A. C., 2003(b): The potential of the Ensemble Kalman filter for NWP - a comparison with 4D-Var. Accepted by *Quart. J. Roy. Met. Soc.*.

Simmons, A. J. and A. Hollingsworth, 2002: Some aspects of the improvement in skill of numerical weather prediction. *Quart. J. Roy. Met. Soc.*, **128**, 647-677

Snyder, C., and F. Zhang, 2003: Assimilation of simulated radar observations with an ensemble Kalman filter. *Mon. Wea. Rev.*, **131**, 1663-1677.

Tippett, M. K., J. L. Anderson, C. H. Bishop, T. M. Hamill and J. S. Whitaker, 2003: "Ensemble square-root filters." *Mon. Wea. Rev.*, **131**, in press.

Whitaker, Jeffrey S., Hamill, Thomas M. 2002: "Ensemble Data Assimilation without Perturbed Observations" *Mon. Wea. Rev.*, **130**, 1913-1924

Whitaker, Jeffrey S, Gibert P. Compo, Xue Wei and Thomas M. Hamill, 2003: Reanalysis without radiosondes using ensemble data assimilation. Submitted to *Mon. Wea. Rev.*