Developments of Variational Data Assimilation

Andrew C. Lorenc

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

ABSTRACT

I go back to fundamental Bayesian principles to discuss the design of four-dimensional variational data assimilation methods, taking account of what has been found important over the past 30 years and what is likely to increase in importance as bigger computers allows higher resolutions and more sophisticated algorithms. Deterministic 4D-Var, statistical incremental 4D-Var, and developments in covariance modelling are all discussed, as are the impact on 4D-Var of chaos and the butterfly effect, and concepts of spin-up. Finally, 4D-Ensemble-Var is outlined as a potential method for keeping most of 4D-Var (but not the perturbation and adjoint models) at high resolution on future massively parallel computers.

1 Introduction

Four-dimensional variational data assimilation (4D-Var) is a mature technology – it has been the method of choice for most major global numerical weather prediction (NWP) centres for the past decade. The predictive skill of NWP has been improving steadily over the past 3 decades; 4D-Var has contributed, but most of the improvement is due to better forecast models. Some improvement is due to better observations, although more is due to better use of all observations. Our challenging goal is to continue this rate of improvement, using increases in computer power, more observations and better mathematical techniques, without giving up these past successes. In section 2 I briefly review the causes of past improvements. NWP systems are very complex – a successful stratagem has been to base each development on scientific insight and mathematical analysis of a component of the problem, so in section 3 I review the theoretical basis of 4D-Var. Section 4 covers improvements that are being developed in the statistical Bayesian framework of 4D-Var, in particular to the representation of the prior PDF which summarises the accuracy of the information which has been carried forward by the ongoing data assimilation process. Section 5 looks more from a dynamical systems viewpoint; the atmosphere is nonlinear and chaotic and these properties are becoming increasingly important as we move to higher resolutions. Finally, section 6 mentions some worries about the efficiency of 4D-Var on future massively parallel computers, and the 4D-Ensemble-Var which is being developed in several centres as a possible way round them.

2 Historical background – What aspects are important to retain from current NWP systems?

Figure 1 shows one of the longest available series on NWP verification statistics and, like many similar results, demonstrates a steady improvement of about a day's predictive skill per decade. Attributing the improvement is less straightforward: it coincides with my career at the Met Office, but I do not claim I caused it! The impact of each upgrade to an operational system is measured in pre-operational trials; I

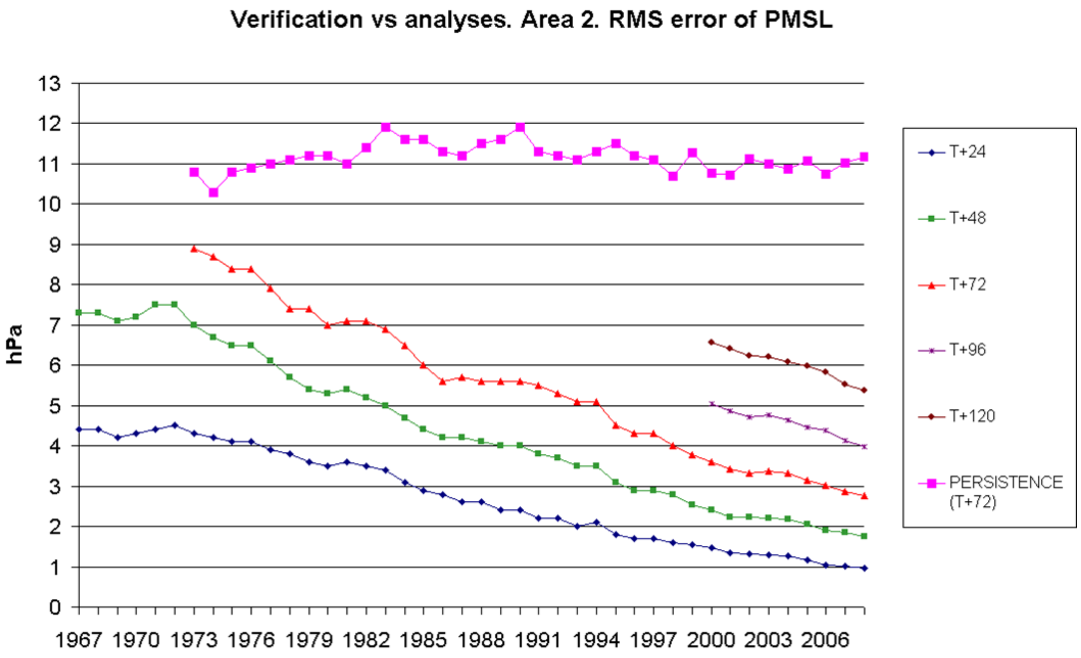

Figure 1: Record of Met Office forecast RMS surface pressure error over the N. Atlantic & W. Europe.

was involved in the Met Office's and saw many others reported at WGNE meetings (WMO (2007)). But some changes interact – often the full benefits are not seen until a subsequent change re-tunes what was previously a compensating error. The largest improvements are due to better forecast models, especially higher resolution. The computer power available to NWP has doubled about every 18 months, giving a 10^6 increase over the last three decades. Most of this has been spent on resolution, with only a factor of 10 on improved DA algorithms and model formulation. The trend to higher resolutions continues, although in the coming decades some power will also be deployed on ensembles. The improvements were studied by Simmons and Hollingsworth (2002). They concluded that a major element was better usage of satellite data, especially the variational assimilation of radiances. They also pointed out that since about 2000 the average one-day forecast errors were less than those of observations such as radiosondes. This is understandable: as assimilation improves, the forecasts effectively incorporate information from the last few days of observations, and hence contain more useful information than the current batch of observations. Revised background error terms in variational assimilation, taking account of this, have been a major contributor to the continuing decrease in errors. 4D-Var implements an implicit four-dimensional background error covariance model (Lorenc (2003a)), so is the natural extension to this trend. The impacts of three recent 4D-Var implementations were compared in Rawlins *et al.* (2007); all gave significant improvements in their year of implementation. In the comparison of verification results organised by CBS, most of the top global forecast centres now use 4D-Var.

It has been common to attribute a major part of the improvement to satellites, but in fact improvements in usage have been more important than improvements in the observation themselves. A review of recent observing system experiment (OSE) studies (WMO (2008)) estimated that most satellite systems contribute up to about 6 hours in skill. This is consistent with Met Office OSEs (Dumelow (2009)) which show that even with all satellite data omitted (or indeed all sondes), northern hemisphere scores were better than those from an "All Data" OSE run 6 years previously. In other words, model resolution improvements and 4D-Var implementation in that period were more important than all the satellites. Comparison of re-analyses using modern systems with operational results at the time shows a similar story (e.g. Onogi *et al.* (2007)) – about a quarter of the improvement over the last 3 decades came from improvements to the observing system, three quarters came from improvements to the NWP and data assimilation systems.

In summary: NWP systems are improving by 1 day of predictive skill per decade. Over the past 3 decades this has been due, in order of importance, to:

1. Model improvements, especially resolution.
2. Careful use of forecast and observations, allowing for their information content and errors; achieved by variational assimilation, e.g. of satellite radiances.
3. 4D-Var.
4. Better observations.

Another important lesson from practical experience is in coping with complexity. NWP systems are very large – too large to be analysed as an entity and improved with confidence. They are also very expensive to test (more about this in section 5.1) – too expensive for all developments to be thoroughly tuned and tested in the full NWP system. Because of this complexity, the only consistent way of making improvements is to base each on scientific insight and mathematical analysis of a component of the problem, with the belief (checked by testing) that theoretically better parts will ultimately lead to a better whole, while at least not harming shorter term performance.

3 Derivation of 4D-Var

3.1 Deterministic – fitting a model evolution to observations

I first summarise[1] the traditional way of deriving 4D-Var. Notation follows Ide *et al.* (1997) with an extension that, to avoid explicit summations over time, underlined variables include the time-dimension and underlined operators produce underlined variables. For the time being I assume a perfect forecast model, so that knowledge of initial conditions $\mathbf{x}$ defines a four-dimensional trajectory $\underline{\mathbf{x}} = \underline{M}(\mathbf{x})$. We want to find the best fit of this trajectory to observations distributed in time ($\underline{\mathbf{y}}^o$) and a prior estimate of $\mathbf{x}$ ($\mathbf{x}^b$). We could simply define best to be a minimum variance solution, but to link with later I prefer a Bayesian derivation. An expression for the probability distribution function (PDF) is:

$$P\left(\mathbf{x}|\underline{\mathbf{y}}^o\right) \propto P(\mathbf{x})\, P\left(\underline{\mathbf{y}}^o|\mathbf{x}\right). \tag{1}$$

We assume the prior PDF for model state $\mathbf{x}$ is a Gaussian with mean $\mathbf{x}^b$ and covariance $\mathbf{B}$:

$$P(\mathbf{x}) \propto \exp\left(-\tfrac{1}{2}\left(\mathbf{x}-\mathbf{x}^b\right)^T \mathbf{B}^{-1}\left(\mathbf{x}-\mathbf{x}^b\right)\right). \tag{2}$$

The observations in the time-window, $\underline{\mathbf{y}}^o$, are usually assumed to have Gaussian errors in observation-space, with covariance $\underline{\mathbf{R}}$, uncorrelated with background errors:

$$P\left(\underline{\mathbf{y}}^o|\mathbf{x}\right) = P\left(\underline{\mathbf{y}}^o|\underline{\mathbf{y}}\right) \propto \exp\left(-\tfrac{1}{2}\left(\underline{\mathbf{y}}-\underline{\mathbf{y}}^o\right)^T \underline{\mathbf{R}}^{-1}\left(\underline{\mathbf{y}}-\underline{\mathbf{y}}^o\right)\right), \tag{3}$$

where $\underline{\mathbf{y}}$ represents the estimate of the observations calculated from $\underline{\mathbf{x}}$ using observation operator $\underline{H}$:

$$\underline{\mathbf{y}} = \underline{H}\left(\underline{M}(\mathbf{x})\right) \tag{4}$$

[1] based on a fuller presentation in Lorenc and Payne (2007)

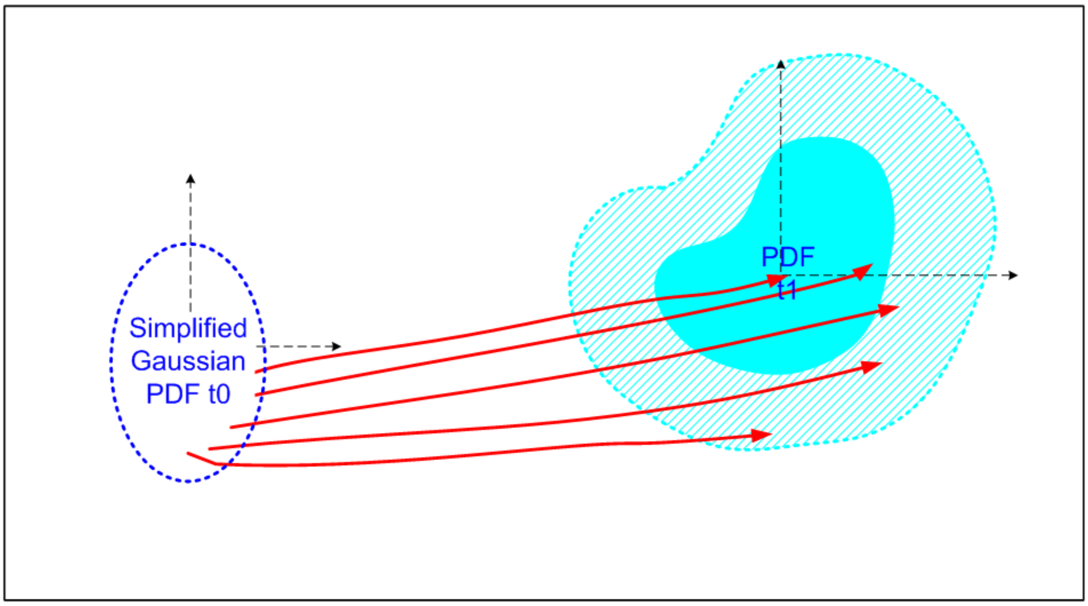

Figure 2: Deterministic 4D-Var. The initial PDF is approximated by a Gaussian. The descent algorithm only explores a small part of the PDF, on the way to a local minimum. The 4D analysis is a trajectory of the full model, optionally augmented by a model error correction term.

$\underline{M}$ and $\underline{H}$ are nonlinear, so the PDF obtained by substituting ((2)), (3) and (4) into (1) is not Gaussian. It is not practicable to evaluate this full PDF for an NWP system, so it is convenient to assume that the desired estimate is the **x** which maximises $P\left(\mathbf{x}|\underline{\mathbf{y}}^o\right)$, or equivalently minimises

$$J(\mathbf{x}) = \tfrac{1}{2}\left(\mathbf{x}-\mathbf{x}^b\right)^T \mathbf{B}^{-1}\left(\mathbf{x}-\mathbf{x}^b\right) + \tfrac{1}{2}\left(\underline{\mathbf{y}}-\underline{\mathbf{y}}^o\right)^T \underline{\mathbf{R}}^{-1}\left(\underline{\mathbf{y}}-\underline{\mathbf{y}}^o\right) \tag{5}$$

subject to (4). This is commonly called the 4D-Var penalty function. The gradient of J with respect to **x** is given by

$$\nabla_{\mathbf{x}} J(\mathbf{x}) = \mathbf{B}^{-1}\left(\mathbf{x}-\mathbf{x}^b\right) + \underline{\mathbf{M}}^* \underline{\mathbf{H}}^* \underline{\mathbf{R}}^{-1}\left(\underline{\mathbf{y}}-\underline{\mathbf{y}}^o\right) \tag{6}$$

where $\underline{\mathbf{M}}^*$ and $\underline{\mathbf{H}}^*$ are the adjoints of the jacobians of $\underline{M}$ and $\underline{H}$, taken at point **x**. Equations (5) and (6) are used in an iterative descent algorithm illustrated schematically in figure 2. This is affordable as only a small part of the full PDF is explored.

In practice this method cannot be applied to current NWP models. The issues are more fundamental than the use of “IF” tests in most NWP models $\underline{M}$; they are due to the physical processes being represented and the principle discussed at the end of section 2 that each component of the complex NWP system should be based on physical understanding. The atmosphere has many processes which often do not give a gradient (6) pointing towards the minimum of (5):

Thermostats: Fast processes which are modulated to maintain a longer-time-scale “balance” (e.g. boundary layer fluxes).

Limits to growth: Fast processes which in a nonlinear model are limited by some available resource (e.g. evaporation of raindrops).

Butterflies: Fast processes which are not predictable over a long 4DVar time-window (e.g. eddies with short space- & time-scales).

Observations of intermittent phenomena: If something (e.g. a cloud or rain) is missing from a state, then the gradient does not say what to do to make it appear.

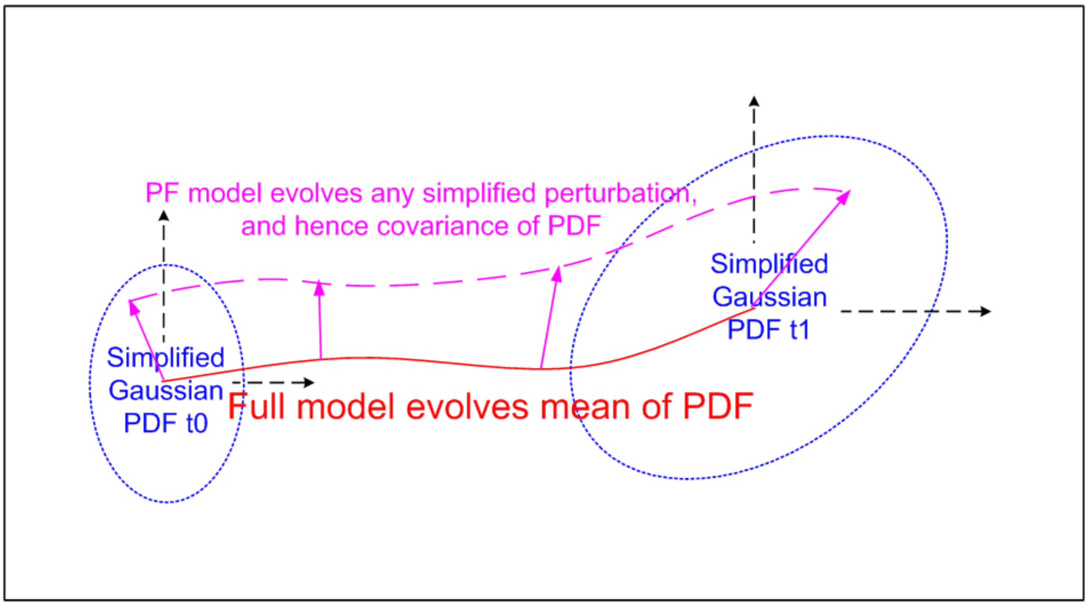
PF model evolves any simplified perturbation,
and hence covariance of PDF
Simplified Gaussian PDF t1
Simplified Gaussian PDF t0
Full model evolves mean of PDF

Figure 3: Statistical, incremental, 4D-Var approximates entire PDF by a Gaussian. The 4D analysis increment is a trajectory of the PF model, optionally augmented by a model error correction term.

These are fundamental atmospheric processes – it is impossible to write a good NWP model, following the principle that each component is based on a physical understanding, without representing them.

3.2 Statistical 4D-Var: the extended Kalman filter

The problem discussed in the previous section can be avoided if we seek a more appropriate "best" estimate. To define what is best, it is necessary to specify a cost function measuring the additional costs incurred by issuing an imperfect forecast. The simplest is a quadratic cost: the expected root mean square error (RMSE) is minimised by the mean of the PDF:

$$\mathbf{x}^a = \int \mathbf{x} P\left(\mathbf{x} | \underline{\mathbf{y}}^o\right) \mathbf{dx}. \tag{7}$$

This integral requires evaluation of the whole PDF, rather than the relatively few function and derivative evaluations needed in the descent algorithm of deterministic 4D-Var. Eq. (7) is not amenable to calculation for PDFs as complicated as (1) for a full NWP model. However, rather than minimising the wrong equation (5), it may be better to find an approximate solution to (7). Lorenc (2003a) suggested one way to do it in the context of 4D-Var, based on the fact that the mode of a PDF is also its mean, as long as the distribution is Gaussian. It is the nonlinear evolution of the prior PDF $P(\mathbf{x})$ by $\underline{M}$ which makes (1) non-Gaussian and (5) non-quadratic. So let us instead approximate that evolution by $\underline{\bar{M}}$, which predict the evolution of the mean, and a perturbation forecast (PF) model $\underline{\tilde{\mathbf{M}}}$ which gives linear best estimates of the evolution of finite perturbations about this mean (Fig. 3).

If necessary, a similar strategy can be applied to H, $\bar{H}$ and $\tilde{\mathbf{H}}$. (Note that we can easily extend the approach to models with errors, by augmenting the model variables to contain also parameters describing the errors. I do not show these terms here – they are in Lorenc (2003a).) The whole PDF (1) (rather than its behaviour near its mode) is then approximated using an incremental approach about a guess $\mathbf{x}^g$ of the ensemble mean (Courtier *et al.* (1994)):

$$\mathbf{x} = \mathbf{x}^g + \delta\mathbf{x}. \tag{8}$$

$$
\begin{aligned}
P\left(\delta\mathbf{x}|\mathbf{y}^{o}\right) \propto \exp & \left(-\tfrac{1}{2}\left(\delta\mathbf{x}-\left(\mathbf{x}^{b}-\mathbf{x}^{g}\right)\right)^{T}\mathbf{B}^{-1}\left(\delta\mathbf{x}-\left(\mathbf{x}^{b}-\mathbf{x}^{g}\right)\right)\right) \\
& \times \exp\left(-\tfrac{1}{2}\left(\underline{\mathbf{y}}-\underline{\mathbf{y}}^{o}\right)^{T}\underline{\mathbf{R}}^{-1}\left(\underline{\mathbf{y}}-\underline{\mathbf{y}}^{o}\right)\right)
\end{aligned}, \tag{9}
$$

$$
\underline{\mathbf{y}} = \underline{\tilde{\mathbf{H}}\tilde{\mathbf{M}}}\delta\mathbf{x} + \underline{H}\left(\underline{\bar{M}}\left(\mathbf{x}^{g}\right)\right) \tag{10}
$$

Since (10) is linear in $\delta\mathbf{x}$, (9) is Gaussian. So if we use the same descent algorithm approach to efficiently approximate the minimum of

$$
\begin{aligned}
J\left(\delta\mathbf{x}\right) = & \tfrac{1}{2}\left(\delta\mathbf{x}-\left(\mathbf{x}^{b}-\mathbf{x}^{g}\right)\right)^{T}\mathbf{B}^{-1}\left(\delta\mathbf{x}-\left(\mathbf{x}^{b}-\mathbf{x}^{g}\right)\right) \\
& +\tfrac{1}{2}\left(\underline{\mathbf{y}}-\underline{\mathbf{y}}^{o}\right)^{T}\underline{\mathbf{R}}^{-1}\left(\underline{\mathbf{y}}-\underline{\mathbf{y}}^{o}\right)
\end{aligned} \tag{11}
$$

subject to (10), we are also approximating the mean of (9). By construction (11) is quadratic, so we avoid the minimisation problems which plague (5).

Unfortunately this approach is not fully satisfactory either. As discussed in section 2, it has been found to be beneficial to use the best available, high-resolution, model to carry information forward, so that we retain information from past observations as accurately as possible. Components of NWP models are developed based on physical principles. We do not have an accurate model $\underline{\bar{M}}$ to predict the evolution of the mean of the PDF (a non-physical field), as required by (10). So in practice we use the normal NWP model $\underline{M}$. To reduce the effect of linearisation errors, (11) can be iterated a few times in an outer-loop, updating $\mathbf{x}^{g}$. This gives a compromise between the "deterministic" and "statistical" approaches; we are trying to fit the full model to the observations, but the method used to do this considers an approximation to the full PDF rather than just seeking a local mode.

It is worth highlighting how large in practice are the approximations in (10). We can measure this using the linearisation error $\varepsilon_{lin} = \mathbf{M}\delta\mathbf{x} - \left(M\left(\mathbf{x}+\delta\mathbf{x}\right) - M\left(\mathbf{x}\right)\right)$, calculated for typical perturbations $\delta\mathbf{x}$ of similar size to the analysis increment. Then the relative error is a norm $\|.\|$ is given by

$$
R = \frac{\|\varepsilon_{lin}\|}{\|M\left(\mathbf{x}+\delta\mathbf{x}\right) - M\left(\mathbf{x}\right)\|} \tag{12}
$$

If $R > 1$ for any norm relevant to observations, it is likely that the analysed correction $\delta\mathbf{x}$ will not improve the fit of the full model integration to those observations. For some variables (e.g. humidity) and forecast lengths of 6 hours or more, R is much closer to 1 than zero (Radnóti *et al.* (2005)).

4 Developments to covariance modelling

4.1 Flow dependence

I stressed in section (2) that an important part of the improvement in forecasts was due to the correct (Bayesian) combination of information from the current observations with the forecast prior (which summarises the information from previous observations). Much important and detailed work is needed to correctly characterise the error PDF of each type of observation – the variance, the bias and the probability of gross error (e.g. talks by Rabier and Desroziers this seminar). I will concentrate on the prior or background error distribution. An NWP model is very large, and the errors for different variables are strongly related, so the best we can normally attempt is to model the error covariance **B**. The first operational 3D multivariate statistical analysis method (Lorenc (1981)) made the following assumptions about the **B** which characterizes background errors, all of which are wrong!

Stationary – time & flow invariant

Balanced – predefined multivariate relationships exist

Homogeneous – same everywhere

Isotropic – same in all directions

3D separable – horizontal correlation independent of vertical levels or structure & vice versa.

Since then many valiant attempts have been made to address them individually, but with limited success because of the errors remaining in the others. The most attractive ways of addressing them all are long-window 4D-Var or hybrid ensemble-VAR. Thépaut *et al.* (1996) showed that a constant "climatological" covariance evolved into plausible flow-dependent patterns in a 24 hour 4D-Var window, Zhang *et al.* (2007) showed that even random structures grow to similar patterns in 24-36 hours, while Fisher *et al.* (2005) showed that there is no advantage to going beyond 5 days. Fisher (this seminar) discusses how such a long window might be affordable. Alternatively, ensemble Kalman filters allow a sample of error patterns to grow over a long window. As well as the many flavours of ensemble Kalman filter, there are two alternative methods of using these patterns in a variational covariance model. Firstly, one can estimate parameters of the covariance model from the ensemble. Bonavita (this seminar) describes such a system. This approach retains the proven benefits of the existing variational method, but it is difficult to address all the weaknesses listed above. For instance Montmerle and Berre (2010) demonstrated a situation dependence to the inter-variable correlations, and many studies have shown non-isotropic flow-dependent correlations – neither are easy to parametrise. The second approach is to use the ensemble perturbations, after localisation, directly to augment (or even replace) the traditional covariance model. The Met Office has just implemented such a hybrid ensemble-4D-Var scheme giving a considerable benefit (Barker, this seminar). There is potential for this approach to even replace the linearised perturbation model within 4D-Var; I say more about this in section 6.

4.2 Non-Gaussianity

It is not possible to represent many aspects of a PDF for as many variables as an NWP model; we usually only consider mean errors and covariances. But in some cases it has been found advantageous to change to variables with more Gaussian errors. Several centres (e.g. ECMWF, HIRLAM and the Met Office) have implemented a nonlinear humidity transform to compensate for the non-Gaussian errors of humidity forecasts (Hólm (2003), Gustafsson *et al.* (2011), Ingleby *et al.* in preparation). The largest cause of the non-Gaussianity is the physical limits to humidity – it must always be positive and seldom goes very super-saturated.

Let us assume that our forecast model is unbiased, in that its distribution of model humidity values is the same as that of the atmosphere mapped into model space. This "truth" state is the goal of our assimilation process and background errors are in principle measured from it. As we do not know it, we have to study background errors using a proxy – in this section for illustrative purposes I use a large set of radiosonde observations mapped to model levels. Figure 4, adapted from Lorenc (2007), show the joint distribution of background and observed (proxy true) values from the Met Office global assimilation. It is close to symmetric about the diagonal, showing that our assumption is reasonable. Yet the distribution of true values conditional on any particular background value is biased, with mean value given by the dash-dot line. So, without considering observations, the minimum variance best estimate of the true RH would be obtained by bias correcting the background to this line. The resulting overall distribution would not be correct – no RH would be greater than 90% so there would be insufficient cloud and precipitation in the subsequent forecast. The method suggested by Hólm (2003) is to use probability distributions conditional on (RHb+RHa)/2. We can illustrate the effect by plotting the joint PDF of the difference in RH and the mean RH (not shown since it is equivalent to figure 4 rotated by 45 degrees). This is unbiased. However having the assumed probability distribution dependent on the analysed value makes

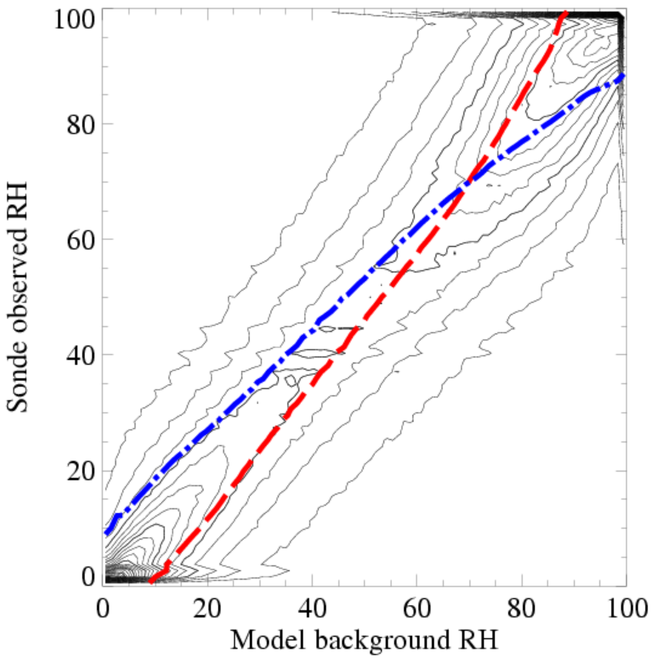

Figure 4: Joint PDF of collocated background RH from the Met Office global 6 hour forecast and observed RH mapped to model levels, for 740328 radiosonde soundings from December 2005 to July 2007 (Lorenc (2007)). The dash-dot line (blue) shows the mean sonde RH for each background RH bin, and the dashed (red) line shows the mean background RH for each sonde RH bin.

the problem implicit, requiring an iterative solution method. ECMWF and HIRLAM put this iteration in their outer-loops; the Met Office can solve it more accurately, in their non-quadratic inner loop.

This method performs well: there is no spurious bias; the background error standard deviation (which is a factor in the equation for the analysis increment) can be small when the background RH is near zero and the increment is negative, making negative analysed values unlikely, while for positive increments the standard deviation can be larger, making it possible to change near zero background RH to any positive value. It is interesting to consider precisely which assumptions about the prior distribution we need to make for this to be the correct Bayesian method:

- The distribution of values in the background, generated by the model, is close to correct – we have the right cloud cover on average.

- It is important to us to retain this correct distribution – more so than to reduce the expected RMS error at each point.

The Hólm transform constructs a (skewed) prior whose mode is the background. We rely on a minimisation which finds this mode (*not the mean*) and hence returns the model background unaltered in the absence of observations. I say more about this desire to rely on the model in the section 5.2.

5 Coping with Butterflies

5.1 Predictability and Chaos

Lorenz (1969) pointed out that the atmosphere has many scales of motion and that errors in small scales will quickly grow and affect larger scales. Revised in detail, this is now the accepted picture of the growth of errors in the spectral domain (Tribbia and Baumhefner (2004)), commonly known as the butterfly effect (e.g. Palmer (2005)). Lorenc and Payne (2007) showed that because of the butterfly effect, conventional deterministic 4D-Var will not work as model resolutions increase towards the unfiltered continuous limit. They suggested that a solution is to use statistical incremental 4D-Var, with a perturbation forecast model which is filtered to prevent the rapid growth of scales which would otherwise grow excessively over the time-window. This idea has been demonstrated in ocean 4D-Var by Hoteit *et al.* (2005); it is implicit in the designs of most operational NWP 4D-Var systems which are forced for computational reasons to use a low resolution linear model in the inner loop.

The butterfly effect is to do with the multiple scales; it means that we cannot necessarily expect 4D-Var algorithms to continue working as we move to higher resolutions on more powerful computers. At any given resolution an NWP forecast model is chaotic – chaos is different from the butterfly effect in that it can be exhibited in toy models with low resolution (e.g. Lorenz (1963)). Abarbanel *et al.* (2010), approaching data assimilation as synchronised chaos, say that there must be enough [observational] controls to move the positive conditional Lyapunov exponents on the synchronization manifold to negative values. This is normally quite easy to achieve in a low-resolution system with few chaotic Lyapunov vectors, but in modern practical NWP, with varying observational networks over the globe, it is much harder to ensure that everything which grows is sufficiently observed. It is quite easy to test if a data assimilation system (rather than the model it uses) is chaotic. We just run the entire system with identical inputs from initial conditions which differ by a very small perturbation, as in the original Lorenz (1963) demonstration of chaos. Figure 5 shows the result of such an experiment with the Met Office system. The initially small perturbation grows for several days before saturating on an attractor with RMS differences of order 0.5 m/s for wind components. (An exception is the top of the model, where differences drift to become increasing large due to the difficulty of controlling model errors and biases at these levels (Polavarapu this seminar)).

These differences are of course much smaller than those between two free model runs in a similar experiment, but they still represent an irreducible uncertainty in the analyses produced by this DA system. Deterministic 4D-Var, using the exact tangent-linear model to the NWP model used, would probably not work. The uncertainty is another reason why trialling of DA system changes is difficult; identical DA systems can still give apparently random different signals when verified against independent observations, necessitating a longer trial to get significant results about a real change.

5.2 Benefiting from the Attractor

The problems with long-window 4D-Var for a chaotic model are due to the continuing exponential growth of some infinitesimal perturbations, and hence of similar perturbations of any amplitude in the tangent linear model. In the full nonlinear model the growth slows as the amplitude increases, leading to saturation often at quite small amplitudes as seen in figure 5, so an ensemble Kalman Filter using the nonlinear model does not have the same problem. Because their predictions remain bounded, chaotic models have an attractor of states which they might pass through which has much lower dimension than the space which could be represented if all the model variables were independent. Meteorologists have long understood this behaviour and developed rules to describe plausible states: both in terms of balance, modes and power spectra, and also synoptically with conceptual models of fronts, cyclones, cloud-capped inversions, etc. This is important prior information which we should use in data assim-

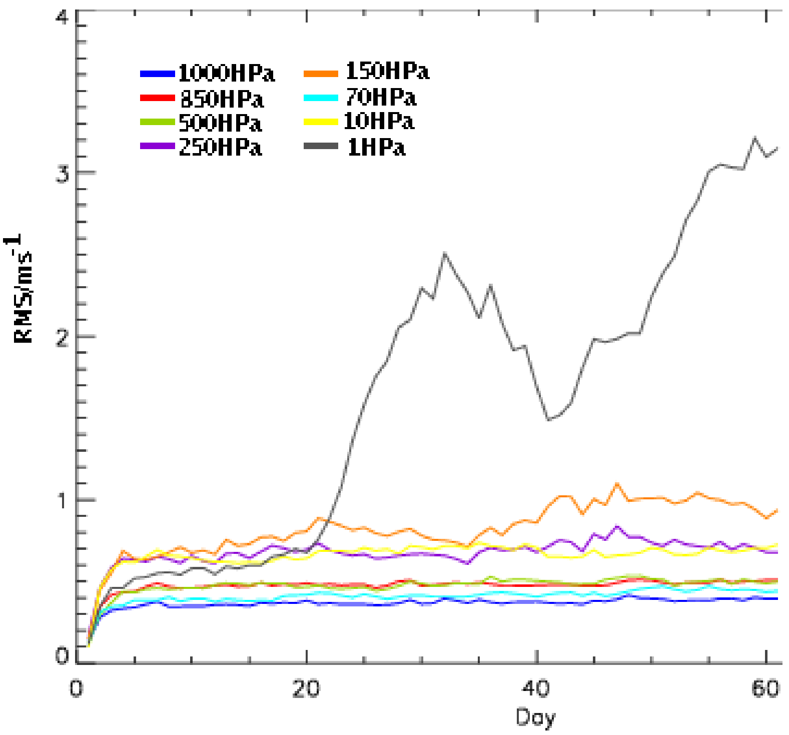

Figure 5: Global RMS differences between u-components in identical NWP assimilations, due to small initial perturbations in the background at day 0. (Peter Jermey, personal communication).

ilation. For instance a human meteorologist could use a conceptual model of a front to fit scattered observations and draw an analysis which could be used in an accurate forecast of the weather at a point ahead of the front. In NWP data assimilation we do not usually use this prior knowledge, we compensate by having a good background forecast, but it would be better to use both. Methods for developing flow dependent correlations are making a start, but the persistent structures are usually maintained by nonlinear processes and do not have Gaussian PDFs. For instance Lorenc (2007) showed that vertical covariances could not describe the errors associated with a cloud-capped inversion. In practice, the best way we have found of describing the attractor is as "states that the model likes". Data assimilation algorithms have regularly used the model in this way, in diabatic nonlinear normal mode initialisation, spin-up periods, and other approaches which often seemed at the time like simple trial and error engineering and tuning.

Modern incremental VAR methods use it by having a spun-up model state as background, and only altering it (in a smooth way) when there is clear observational evidence to do so. We used this concept already in section 4.2. As we move to higher resolutions, with models which are representing the complex (and poorly observed) structures of convection, the reliance on the model will grow – other simpler concepts of balance do not hold at the convective scale but there are still clear recognisable structures for convection (and many more structures which are unlikely ever to occur). So planning for the future we should select methods which allow the model to spin up and evolve states on the attractor. The particle filter is the extreme way of ensuring this (e.g. van Leeuwen this seminar). Another argument for an outer-loop and a long window is that together they seek a spun-up model trajectory which fits the observations. On the other hand the 4D control variable approach to 4D-Var (Fisher this seminar) deliberately avoids long model runs in the data assimilation, so may have problems of spin up. Some EnKF methods recentre the ensemble each cycle about the ensemble mean analysis. This may be undesirable because the ensemble mean is not on the attractor – a forecast from it would be expected to give a poor short-period forecast of “weather” such as cloud and precipitation.

6 4D-Ensemble-Var

It is expected that the computers available for NWP over the next decade will continue to get more powerful, but in the number of processors rather than the speed of each (Isaksen, this seminar). Using this power will require a more parallel DA algorithm; the current bottleneck in 4D-Var is the PF model. Fisher (this seminar) is discussing one approach to making 4D-Var more parallel by changing the control variable so that time-segments of each PF and adjoint integration can be run in parallel. Here I outline a more radical approach, doing away with the PF model completely.

The main idea is to extend in time the use of the ensemble perturbations, currently used in the Met Office operation hybrid ensemble-4D-Var (Barker, this seminar), so they are used to fit the observations in a time-window as 4D-Var does, but without the cost of iterating a PF and adjoint model. The potential of Ensemble Kalman Filters to do this has been recognised for some time (e.g. Lorenc (2003b)). Hunt *et al.* (2004) demonstrated it with an ensemble square root filter for the Lorenz96 model, Fertig *et al.* (2007) compared a 4D-LETKF with 4D-Var for the same model, and Harlim and Hunt (2007) applied 4D-LETKF to the SPEEDY model. The explicit documentation and testing in a VAR environment has been published by Liu *et al.* (2008), Liu *et al.* (2009), Buehner *et al.* (2010a) and Buehner *et al.* (2010b); Liu called the technique En4DVAR and Buehner En-4D-Var[2]. I prefer the name 4D-Ensemble-Var or 4D-En-Var since the key feature is the 4-dimensional use of the ensemble; it also is more consistent with the 4DEnKF terminology of Hunt *et al.* (2004). I reserve En-4D-Var to describe a component of the approach in our current hybrid 4D-Var: using the ensemble to estimate the background error covariance **B** at the beginning of the time window[3], with the fitting of observations distributed in time done as in 4D-Var using a PF and adjoint model. Buehner *et al.* (2010b) presented results from a near-operational-quality Canadian NWP system showing that 4D-En-Var is competitive with traditional 4D-Var and with En-4D-Var.

6.1 Basic 4D-En-Var Equations

I consider a 4-dimensional best fit to all the observations in an assimilation window from start time t_s to end time t_e. With the addition of the underline notation to denote the extra time-dimension, and the replacement of the climatological **B** by the predicted 4-dimensional $\underline{\mathbf{P}}$, this has identical form to the well known incremental 3D-Var. We seek to minimise:

$$J\left(\delta\underline{\mathbf{x}}\right) = \frac{1}{2}\delta\underline{\mathbf{x}}^T\underline{\mathbf{P}}^{-1}\delta\underline{\mathbf{x}} + \frac{1}{2}\left(\underline{H}\left(\underline{\mathbf{x}}^b + \delta\underline{\mathbf{x}}\right) - \mathbf{y}^o\right)^T\underline{\mathbf{R}}^{-1}\left(\underline{H}\left(\underline{\mathbf{x}}^b + \delta\underline{\mathbf{x}}\right) - \mathbf{y}^o\right) \tag{13}$$

As usual in NWP DA algorithms, we cannot actually handle (13); $\delta\underline{\mathbf{x}}$ is big and $\underline{\mathbf{P}}^{-\mathbf{1}}$ is much too big to manipulate! So we seek a representation of $\delta\underline{\mathbf{x}}$ in terms of a reduced set of control variables. The basic idea of 4D-En-Var is that $\delta\underline{\mathbf{x}}$ is made up as a locally weighted linear combination of perturbation trajectories $\underline{\mathbf{x}}'_i$ which are scaled (and perhaps transformed) differences between ensemble members and the ensemble mean:

$$\delta\underline{\mathbf{x}}_e = \sum_i \underline{\alpha}_i \circ \underline{\mathbf{x}}'_i \tag{14}$$

We assume the perturbations are independent, so that we can define each $\underline{\alpha}_i$ independently. To this we can add additional terms, designed to provide scope to correct model errors which are not sampled by

[2] More recently Mark Buehner says he prefers to call it simply En-Var.

[3] Buehner *et al.* (2010b) called this 4D-Var-Benkf.

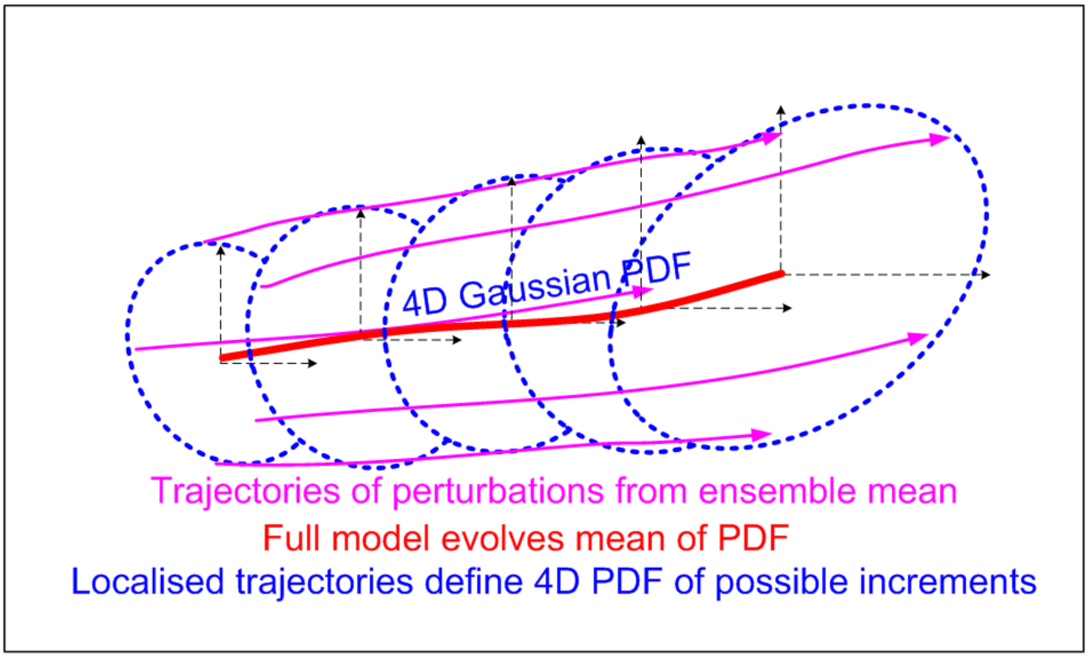
4D Gaussian PDF
Trajectories of perturbations from ensemble mean
Full model evolves mean of PDF
Localised trajectories define 4D PDF of possible increments

Figure 6: A schematic diagram of 4D-En-Var, for comparison with figure 3. The 4D analysis is a localised linear combination of model trajectories – it is not itself a model trajectory.

the trajectories, and generally to allow the use of hybrid methods to compensate for a small ensemble. I just show a time constant and one time-varying term for each

$$\delta\underline{\mathbf{x}} = \beta_{e0}\delta\underline{\mathbf{x}}_{e0} + \beta_{e1}\delta\underline{\mathbf{x}}_{e1} + \beta_{c0}\delta\underline{\mathbf{x}}_{c0} + \beta_{c1}\delta\underline{\mathbf{x}}_{c1} \tag{15}$$

The Met Office's existing hybrid ensemble-4D-Var only has hybrid weights β_e and β_c because it does not allow for model error. Note that the climatological term $\delta\underline{\mathbf{x}}_{c0}$ is constant over the time-window, since we are not using the PF model; in this aspect the new method is 3D-Var rather than 4D-Var. Probably this means that we will want to make more use of the $\delta\underline{\mathbf{x}}_{e0}$ term, which does allow for time evolution, by making the ensemble larger, and β_{c0} smaller than in hybrid ensemble-4D-Var. Weak constraint terms allowing for model error are included for completeness: the $\delta\underline{\mathbf{x}}_{c1}$ term allows for a constant model error tendency error and the ensemble term $\delta\underline{\mathbf{x}}_{e1}$ allows for the weights $\underline{\alpha}_i$ to vary in time.

We make the key error modelling assumption that the terms are independent from each other. We can then go on to define independent transforms ($\underline{\mathbf{U}}^{e0}$, $\underline{\mathbf{U}}^{e1}$, $\underline{\mathbf{U}}^{c0}$, $\underline{\mathbf{U}}^{c1}$) to diagonalise the control variables. Actually, to allow for non-Gaussian errors, we currently use a nonlinear parameter transform (section 4.2). To be correct, this has to transform the total increment. We also want to do "balance aware" localisation. So the Met Office's initial design will use

$$\delta\underline{\mathbf{x}} = U_p\left(\beta_{e0}\delta\underline{\mathbf{x}}_{e0} + \beta_{e1}\delta\underline{\mathbf{x}}_{e1} + \beta_{c0}\delta\underline{\mathbf{x}}_{c0} + \beta_{c1}\delta\underline{\mathbf{x}}_{c1}\right) \tag{16}$$

where the $\delta\underline{\mathbf{x}}_{e0}$ etc. terms are in transformed parameter space.

These transforms are constructed in the normal VAR way, so the transformed control variables are independent with unit variance. They are combined into a single control vector

$$\mathbf{v} = \left(\mathbf{v}^{c0}, \left(\mathbf{v}_i^{\alpha 0}\right)_{i=1,K}, \mathbf{v}^{c1}, \left(\mathbf{v}_i^{\alpha 1}\right)_{i=1,K}\right) \tag{17}$$

This gives us a new penalty function to replace (13):

$$J(\mathbf{v}) = \frac{1}{2}\mathbf{v}^T\mathbf{v} + \frac{1}{2}\left(\underline{H}\left(\underline{\mathbf{x}}^b + \delta\underline{\mathbf{x}}\right) - \mathbf{y}^o\right)^T \underline{\mathbf{R}}^{-1}\left(\underline{H}\left(\underline{\mathbf{x}}^b + \delta\underline{\mathbf{x}}\right) - \mathbf{y}^o\right) \tag{18}$$

6.2 Prospects

Above we described a smoother, giving a 4-dimensional $\delta\underline{\mathbf{x}}$ over the time-window. Work is needed to think more carefully about how to add this to the full model and start the next forecast, i.e. how to convert the smoother solution to an ongoing filter, taking account also of the considerations discussed in section 5.2. Plans are also being developed to apply the same 4D-En-Var algorithm (using the same ensemble perturbations) to each ensemble member, replacing the localised ETKF in the current MOGREPS system.

The 4D-Ensemble-Var approach replaces the costly, sequential, PF and adjoint model integrations by the use of pre-calculated perturbation trajectories, which have to be input and stored. Current indications are that, even allowing for this, it can be made at least as fast as 4D-Var on our current computer (IBM power6), while results such as Buehner *et al.* (2010b) give the expectation that it will be comparable in quality. The new algorithm has the advantage of being highly scalable, so it is more likely to work efficiently at higher resolution on future computers. It also removes the need for an adjoint model. This is otherwise a worry since independent developments of efficient NWP models for future computers may well lead to a radical restructuring and different grid – recoding of an adjoint for such a model is not an attractive prospect.

References

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

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, doi:10.1175/2009MWR3157.1.

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, doi:10.1175/2009MWR3158.1.

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.

Dumelow R. 2009. Global data denial experiments using 4D-Var. *Met Office Forecasting Research Tech. Rept.* **532**, URL `http://research.metoffice.gov.uk/research/nwp/publications/papers/technical_reports`.

Fertig EJ, Harlim J, Hunt BR. 2007. A comparative study of 4D-VAR and a 4D Ensemble Kalman Filter: perfect model simulations with Lorenz-96. *Tellus A* **59**(1): 96–100, doi:{10.1111/j.1600-0870.2006.00205.x}.

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.

Gustafsson N, Thorsteinsson S, Stengel M, Hólm E. 2011. Use of a nonlinear pseudo-relative humidity variable in a multivariate formulation of moisture analysis. *Q. J. R. Meteorol. Soc.* **137**(657): 1004–1018, doi:10.1002/qj.813, URL `http://dx.doi.org/10.1002/qj.813`.

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:{10.1111/j.1600-0870.2007.00255.x}.

Hólm E. 2003. Revision of the ecmwf humidity analysis: construction of a gaussian control variable. *ECMWF/GEWEX Workshop on Humidity Analysis* URL http://www.ecmwf.int/publications/library/do/references/list/17000.

Hoteit I, Cornuella B, Kohl A, Stammer D. 2005. Treating strong adjoint sensitivities in tropical eddy-permitting variational data assimilation. *Q. J. R. Meteorol. Soc.* **131**: 3659–3682.

Hunt B, Kalnay E, Kostelich E, Ott E, Patil D, Sauer T, Szunyogh I, Yorke J, Zimin A. 2004. Four-dimensional ensemble Kalman filtering. *Tellus A* **56**(4): 273–277.

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.

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, doi:10.1175/2008MWR2699.1.

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

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.

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**, URL http://www.metoffice.gov.uk/research/nwp/publications/papers/technical_reports/.

Lorenc AC, Payne T. 2007. 4D-Var and the butterfly effect: Statistical four-dimensional data assimilation for a wide range of scales. *Q. J. R. Meteorol. Soc.* **133**: 607–614.

Lorenz EN. 1963. Deterministic nonperiodic flow. *J. Atmos. Sci.* **20**: 130–141.

Lorenz EN. 1969. The predictability of a flow that possesses many scales of motion. *Tellus A* **21**: 289–307.

Montmerle T, Berre L. 2010. Diagnosis and formulation of heterogeneous background-error covariances at the mesoscale. *Q. J. R. Meteorol. Soc.* **136**(651, Part B): 1408–1420, doi:{10.1002/qj.655}.

Onogi K, Tslttsui J, Koide H, Sakamoto M, Kobayashi S, Hatsushika H, Matsumoto T, Yamazaki N, Kaalhori H, Takahashi K, Kadokura S, Wada K, Kato K, Oyama R, Ose T, Mannoji N, Taira R. 2007. The JRA-25 reanalysis. *J. Met. Soc. of Japan* **85**(3): 369–432, doi:{10.2151/jmsj.85.369}.

Palmer TN. 2005. Quantum reality, complex numbers, and the meteorological butterfly effect. *Bull. Am. Meteorol. Soc.* **86**: 519–530.

Radnóti G, Trémolet Y, Andersson E, Isaksen L, Hólm E, Janiskov M. 2005. Diagnostics of linear and incremental approximations in 4d-var revisited for higher resolution analysis. *ECMWF Technical Memorandum* **467**, URL http://www.ecmwf.int/publications/library/do/references/show?id=86713.

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

Simmons AJ, Hollingsworth A. 2002. Some aspects of the improvement in skill of numerical weather prediction. *Q. J. R. Meteorol. Soc.* **128**: 647–677.

Thépaut JN, Courtier P, Belaud G, Lemaître G. 1996. Dynamical structure functions in a four-dimensional variational assimilation: A case study. *Q. J. R. Meteorol. Soc.* **122**: 535–561.

Tribbia JJ, Baumhefner DP. 2004. Scale interactions and atmospheric predictability: An updated perspective. *Mon. Weather Rev.* **132**: 703–713.

WMO. 1994-2007. Working group on numerical experimentation (WGNE). meeting reports. URL "http://www.wmo.int/pages/about/sec/rescrosscut/resdept_wgne.html".

WMO. 2008. Fourth WMO workshop on the impact of various observing systems on NWP. Geneva, Switzerland. May 2008. Summary and Conclusions. URL http://www.wmo.int/pages/prog/www/OSY/Meetings/NWP-4-Geneva2008/Summary-Conclusions.pdf.

Zhang F, Bei N, Rotunno R, Snyder C, Epifanio CC. 2007. Mesoscale predictability of moist baroclinic waves: Convection-permitting experiments and multistage error growth dynamics. *J. Atmos. Sci.* **64**: 3579–3594.