Developments in Ensemble Data Assimilation

Jeffrey S. Whitaker

NOAA Earth System Research Laboratory
Boulder, CO, USA

1. Background

Ensemble forecasts, in addition to providing raw material for producing probabilistic forecasts, provide the forecaster with a way to estimate the uncertainty in a single, deterministic forecast. The accuracy of quantitative estimates of forecast uncertainty derived from an ensemble forecast system depend on the ability of the system to sample all sources of forecast uncertainty, including uncertainties in the initial state and boundary conditions as well as uncertainties in the forecast model. The Ensemble Kalman Filter (EnKF) is a data assimilation technique designed specifically to meet this need. Unlike variational data assimilation systems, which produce a single deterministic analysis, the EnKF produces an ensemble of initial states designed to sample the probability distribution consistent with observations and prior forecast errors.

All data assimilation systems require "background-error covariances" ($\mathbf{P}^b$) which describe the error characteristics of the first-guess forecast - these determine how the first-guess forecast is blended with observations in the analysis. The unique aspect of the EnKF is that the $\mathbf{P}^b$ is estimated from an ensemble of first-guess, or background forecasts, and therefore adapts to the dynamics. This leads to an improvement in analysis quality (through a better blending of first-guess and observation information) and a more accurate, situation-dependent estimate of initial condition uncertainty for initializing ensemble forecasts.

Figure 1 shows some idealized examples of how flow dependent background-error covariances can drastically alter how observations are used in certain dynamical situations. In the hurricane example, an observation of meridional wind just east of the storm center can increment the entire axisymmetric circulation in the EnKF, since the $\mathbf{P}^b$ estimate contains information about that structure. When the $\mathbf{P}^b$ is flow-independent (in this case homogeneous and isotropic), the wind observation only affects the wind speed on one side of the storm, leading to a spurious asymmetry. Similarly, the dynamical information present in the EnKF estimate of $\mathbf{P}^b$ can allow a temperature observation ahead of a warm front to shift the position of the entire warm front, instead of locally bulging the warm front toward the observation. In both of these examples it is assumed that there is only one observation within the dynamical feature of interest. If they were many more observations sampling the entire feature, the benefits of flowdependent background-error covariances would be much less apparent. In other words, the benefits of flow-dependent background-error covariances are greatest in situations where there are sparsely observed, but dynamically coherent structures in the background forecast.

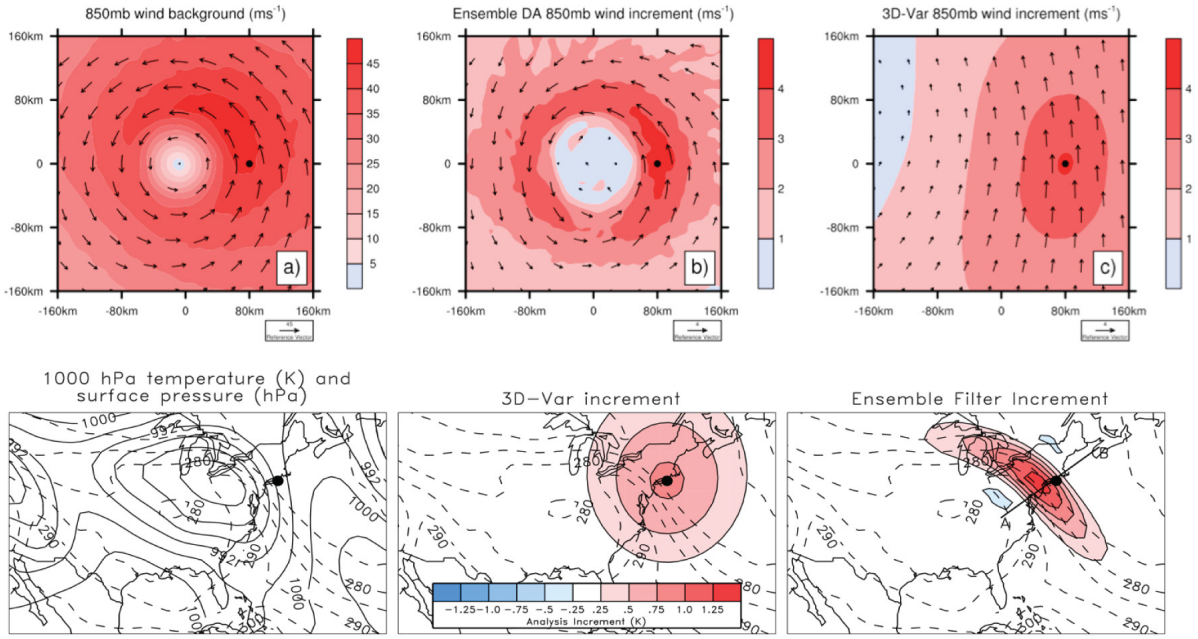

Figure 1: Idealized examples of the impact of flow dependent versus static (isotropic and homogeneous) background-error covariances on analysis increments. Top row: idealized hurricane with an observation of meridional wind at the location of the black dot. Bottom row: Idealized warm front with an observation of temperature at the location of the black dot. Left hand panels show the background forecast. Middle panels show the "3DVar" increment computed using homogeneous, isotropic flowindependent background error covariances. Right hand panels show increment computed using flow-dependent background-error covariances estimated from an ensemble of short-range forecasts.

2. Mathematical background and algorithmic details

Here we use the standard notation of Ide et al (1997), so that $\mathbf{y}$ is a vector of observations with expected error $\boldsymbol{\varepsilon}$, $\mathbf{x}$ is the model state vector (subscript **b** denotes background or prior, subscript **a** denotes analysis or posterior, subscript **t** denotes "truth"), $\mathbf{H}$ is the forward observation operator that converts the model state to observation space ($\mathbf{H}\mathbf{x}_t = \mathbf{y} + \boldsymbol{\varepsilon}$), $\mathbf{R}$ is the observation error covariance matrix ($= < \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T >$), $\mathbf{P}^b$ is the background-error covariance matrix ($=<\mathbf{x}_b\,\mathbf{x}_b^T>$, s.t. $\mathbf{x}_t = N\left(\overline{\mathbf{x}}_b, \mathbf{P}^b\right)$), <..> denotes expected value, the overbar ensemble mean, and prime ensemble perturbation (s.t. $\mathbf{x} = \overline{\mathbf{x}} = \mathbf{x}'$).

If we assume that background observation errors are Gaussian, i.e.

$$\mathbf{x}_t = N\left(\overline{\mathbf{x}}_b, \mathbf{P}^b\right) \text{ and } \varepsilon = N\left(0, \mathbf{R}\right) \tag{1}$$

then Bayes rule $p(\mathbf{x}|\mathbf{y})$ $p(\mathbf{y}|\mathbf{x})p(\mathbf{x})$ implies the Kalman Filter equations (Kalman and Bucy, 1961);

$$\overline{\mathbf{x}}_a = \overline{\mathbf{x}}_b + \mathbf{K}\left(\mathbf{y} - \mathbf{H}\overline{\mathbf{x}}_b\right); \mathbf{P}^a = \left(\mathbf{I} - \mathbf{K}\mathbf{H}\right)\mathbf{P}^b \tag{2a,b}$$

Where

$$\mathbf{K} = \mathbf{P^b}\,\mathbf{H^T}\left(\mathbf{H}\mathbf{P^b}\,\mathbf{H^T} + \mathbf{R}\right)^{-1} \tag{3}$$

is the Kalman gain. Solving the Kalman filter directly is impractical for high-dimensional models, since $\mathbf{P^b}$ is an N_x by N_x matrix (where N_x the dimension of the model state vector $\mathbf{x}$). The EnKF, first proposed by Evensen (1994) and Houtekamer and Mitchell (1998) approximates $\mathbf{P^b}$ with a sample estimate of dimension N_e, which is typically many orders of magnitude less than N_x. It can be shown that under assumptions of linearity and Gaussianity, the EnKF converges to the full Kalman Filter as $N_e \rightarrow N_x$. Simply using a sample estimate for $\mathbf{P^b}$ is not enough to make the EnKF practical – next I describe several other computational shortcuts that almost all current implementation use.[1]

2.1. Simplifying the Kalman Gain calculation

Instead of computing the N_x by N_x matrix $\mathbf{P^b}$ and then post-multiplying by the N_o by N_o matrix $\mathbf{H^T}$ (where N_o is the dimension of the observation vector $\mathbf{y}$), the N_x by N_o matrix $\mathbf{P^b}\,\mathbf{H^T}$ is computed directly in observation space via

$$\mathbf{P^b}\mathbf{H^T}\frac{1}{N_e-1}\sum_{j=1}^{N_e}\left(\mathbf{x_j^{b'}}\mathbf{H}\mathbf{x_j^{b'}}\right) \text{ and } \mathbf{H}\mathbf{P^b}\,\mathbf{H^T}\frac{1}{N_e-1}\sum_{j=1}^{N_e}\left(\mathbf{H}\,\mathbf{x_j^{b'}}\mathbf{H}\mathbf{x_j^{b'}}\right) \tag{4}$$

where the subscript j denotes ensemble member. This saves computation since N_o is typically much smaller than N_x. Furthermore, when covariance localization is applied (section 2.2), the number of points where this calculation is performed is reduced to the subset of model state elements that are "close" to each observation.

2.2. Covariance Localization

If a distance can be computed between each model prior element in $\mathbf{x_b}$ and each observation in $\mathbf{y}$, and it is assumed that the magnitude of the true covariance between model prior element in $\mathbf{x_b}$ and each observation prior element in $\mathbf{Hx_b}$ decreases with spatial separation, then the calculations in (4) need only be computed for "nearby" state and observation prior pairs. In other words, for a given ensemble size, the covariance between model state and observation priors becomes statistically indistinguishable from zero at some separation distance L. Typically, a smooth function, such as the one proposed by Gaspari and Cohn (1999) is used to smoothly taper the covariances from unity at zero separation distance to zero at separation distance L (Hamill et al 2001). By removing the sampling noise while retaining most of the using signal, covariance localization can improve the accuracy of the sample covariance estimate. This is illustrated in Figure 2, which shows the effect that covariance localization has on the 50-member sample estimate of an idealized "true" covariance.

[1] Much of the material in this section is covered in more detail in the review papers of Evensen (2003) and Hamill (2006).

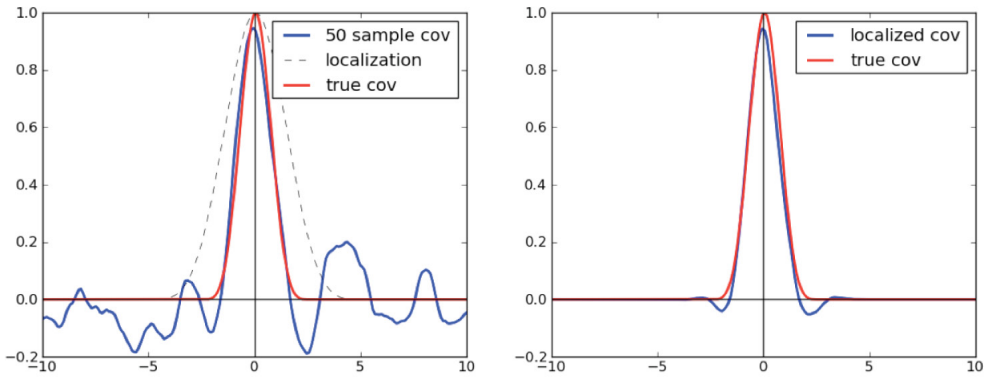

Figure 2: Effect of covariance localization on a sample covariance estimate using 50 ensemble members. Red curves show the exact covariance. The blue curve on the left is the sample estimate computing by randomly sampling the exact covariance matrix with 50 ensemble members. The dashed curve on the left shows the covariance localization function. The blue curve on the right is the localized sample covariance estimate computed by multiplying the blue and dashed curve show on the left.

Localization effectively increases the rank of the sample estimate of $\mathbf{P}^{\mathbf{b}}$. Without localization, the global analysis increment is required to project completely into the low-dimensional space spanned by the background ensemble. With localization, only the local increment (within a distance L of an observation) is constained to lie in the ensemble subspace *in that local region*. This means that, in a global sense, the EnKF can make corrections to the background in directions not contained in the subspace of the ensemble. In addition to dramatically reducing the computational load, covariance localization is crucial for avoiding the catastrophic condition known as "filter divergence" which occurs whenever the global subspace spanned by the EnKF analysis increments is smaller than the space spanned by the true $\mathbf{P}^{\mathbf{a}}$. Filter divergence causes the EnKF solution to drift farther and farther from the observations, so that the ensemble mean analysis errors grow while the ensemble variance decreases. It is impossible to overstate the importance of covariance localization in making the EnKF a viable solution for numerical weather prediction - it really is the "secret ingredient" that makes it all possible.

Localization implemented with the observation space estimate of $\mathbf{P}^{\mathbf{b}}\,\mathbf{H}^{\mathbf{T}}$ (equation 4) works well when the forward observation operator of H is nearly diagonal, as is the case when the observed variable is a model state variable and $\mathbf{H}$ is simply linear interpolation. However, when $\mathbf{H}$ is more complicated, such as is the case with satellite radiance observations, equation (4) is an approximation to $\mathbf{P}^{\mathbf{b}}\,\mathbf{H}^{\mathbf{T}}$, even for infinite ensemble size, since the covariance calculation and the forward observation operator do not commute (Campbell et al, 2010).

2.3. Serial observation processing

If the observation error covariance $\mathbf{R}$ is diagonal, that is errors in individual observations are uncorrelated with each other, then assimilating observations one at a time is equivalent to assimilating them all at once in the Kalman filter (Gelb 1974). For a single observation, the term $\mathbf{P}^{\mathbf{b}}\,\mathbf{H}^{\mathbf{T}}$ in the Kalman gain becomes a vector of length N_x and the innovation covariance $\mathbf{H}\mathbf{P}^{\mathbf{b}}\,\mathbf{H}^{\mathbf{T}} + \mathbf{R}$ becomes a

scalar. Therefore, the computation of the Kalman gain no longer involves matrix inversions and can be computed as a simple dot product. After a single observation is assimilated, the updated model state becomes the background for the assimilation of the next observation. To avoid having to recompute the forward observation operator for all unassimilated observations after each step, the EnKF update can be applied in observation space to update the model observation priors along with the model state priors. The serial processing approach is described in more detail in Whitaker and Hamill (2002) and Anderson (2001), and is implemented in the operational EnKF run for global weather prediction at the U.S. National Center for Environmental Prediction (NCEP) and the community Data Assimilation Research Testbed (DART) software (Anderson et al, 2009). An efficient approach for parallelizing the serial filter is described in Anderson and Collins (2007), and is implemented in both the NOAA/NCEP and DART systems. Instead of assimilating observations one at a time, they can be assimilated in small batches to keep the computation of the inverse of the innovation covariance matrix tractable. This is done in the current operational system at Environment Canada (Houtekamer et al, 2005). When using serial assimilation, it is important to remember that covariance localization destroys the formal equivalence between serial processing and batch processing of observations, and introduces a dependence on the order observations are assimilated (Whitaker and Hamill, 2002). However, experience so far has not shown any negative consequences associated with the combination of serial assimilation with covariance localization. If observation errors are correlated (i.e. $\mathbf{R}$ is not diagonal), serial processing can still be used after transforming the observations and observation priors into the eigenspace of $\mathbf{R}$.

2.4. Local implementations of the EnKF

Another way to reduce the size of the matrices involved in the solution of the EnKF is to perform the state update in local, overlapping regions of the model state space. For example, if all the N_L observations with the covariance localization radius L of a given model grid point are used to update that single model grid point, the dimension of the innovation covariance matrix becomes N_L and $\mathbf{P}^{\mathbf{b}}\mathbf{H}^{\mathbf{T}}$ is simply a vector of length N_L. The solution for each grid point can be computed independently, so the local approach is highly parallelizable. However, the price to be paid is lots of redundant computation, since adjacent model grid points have a large degree of overlap in their sets of nearby observations. There is also an approximation implicit in the local approach, since the covariances between observation priors outside and inside the local region are ignored. The most popular algorithm for solving the EnKF locally is the Local Ensemble Transform Kalman Filter (LETKF), described in detail by Hunt et al (2007). The local approach is similar to the "box" method of observation selection in optimal interpolation (Lorenc, 1981).

2.5. Stochastic versus Deterministic implementations

The EnKF as originally formulated by Houtekamer and Mitchell (1998) treats the observations as an ensemble by adding $N(0,\mathbf{R})$ noise to the observation vector. This is needed to prevent under-estimation of $\mathbf{P}^{\mathbf{a}}$ when every ensemble member updated with the same Kalman Filter update equations (Burgers et al, 1998). This "perturbed observation" approach has come to be known as the stochastic implementation of the EnKF. Whitaker and Hamill (2002) pointed out that the ensemble mean and ensemble perturbations can be updated separately to ensure that the analysis-error covariance $\mathbf{P}^{\mathbf{a}}$ is consistent with what is expected from the Kalman filter covariance update equation 2b, without having to add random noise to the observations. The class of algorithms that follow this approach are known as deterministic filters, and while they all use the Kalman filter update equation (2a) to update

the ensemble, they use different methods to update deviations from the ensemble mean. This is because there are many ways to construct an ensemble perturbation update that produces the expected $\mathbf{P}^a$ (Tippett et al, 2003). The most popular deterministic filter implementations are the LETKF (Hunt et al, 2007) and the serial "ensemble square-root filter" of Whitaker and Hamill (2002)[2]. The deterministic approach is more accurate for small ensemble sizes, since it avoids an extra source of sampling error in the estimation of the observation-error covariance matrix (Whitaker and Hamill, 2002), while the stochastic approach is more robust in the presence of outliers that result from non-gaussianity in the background ensemble (Lawson and Hansen, 2004).

3. What makes the EnKF suboptimal?

The EnKF is optimal in the least-squares sense if the following conditions are met:

i. Observation and background forecast errors are uncorrelated and Gaussian.

ii. Ensemble size large enough so that sampling errors are small ($N_e \rightarrow N_x$)

iii. All sources of forecast error are sampled by the ensemble, including model error.

The first condition applies to all techniques based on the Kalman filter, including variational methods. The second and third conditions are specific to ensemble data assimilation systems. EnKF development efforts are currently focused on developing better methods to deal with sampling errors, model errors and other un- or underrepresented sources of error in the background forecast ensemble. In addition to model and sampling error, other sources of background error that may be under-represented in EnKF systems include

i. mis-specification of observation errors (including correlated observation errors)

ii. errors in the forward observation operator **H**

iii. errors in boundary conditions (such as the sea and land surface).

Neglecting or under-estimating any of these sources of error in the ensemble forecast system will cause the assimilation to give too little weight to observations. For the remainder of this paper, I will describe how these missing sources of background error are currently parameterized, and some ongoing research efforts aimed at improving the representation of these errors within ensemble data assimilation and forecast systems.

3.1. Flow-dependent covariance localization

The simplest form of covariance localization, described in section 2.2, tapers the covariance between model state and observation priors to zero at a specified distance from the observation location. The taper distance is a single parameter, and is typically not a function of spatial location or the flow situation. As pointed out by Bishop and Hodyss (2007), this is a very crude approximation and there is

[2] Identical to the "ensemble adjustment filter" of Anderson (2001)

likely to be a high-degree of flow dependent variability in background-error covariance sampling errors. For example, Figure 3 shows the covariance between the background ensemble temperature at the location of a radiosonde over Australia and temperature at other locations in an EnKF data assimilation system. In the lower troposphere (850 hPa) the covariances are fairly compact, and the 1800 km covariance localization radius is fairly effective at removing the noisy far-field covariance while preserving the apparent signal close to the observation location. However, at 10 hPa, the covariances are much larger scale, and the fixed localization is likely removing signal as well as noise in the covariance estimate. This is because the dynamics of error evolution are much different in the stratosphere than they are in the lower troposphere, resulting in much longer length scales in the stratosphere. Another simple example illustrating how the assumptions used in simple covariance localization can break down involves the assimilation of observations separated in both time and space from the model state element being updated. If the observation time is prior to the analysis time, the maximum covariances will be downstream of the observation location, while the simple localization described in section 2.2 assumes that the covariance maximum occurs at the observation location. One simple way to deal with the problems illustrated in Figure 3 is to make the localization radius vertically varying, increasing upward in the stratosphere. Alternately, one can try to estimate localization function itself directly from the background ensemble, thereby taking into account flow-dependence (and implicitly accounting for the effect of temporal as well as spatial separation between observation and model state priors).

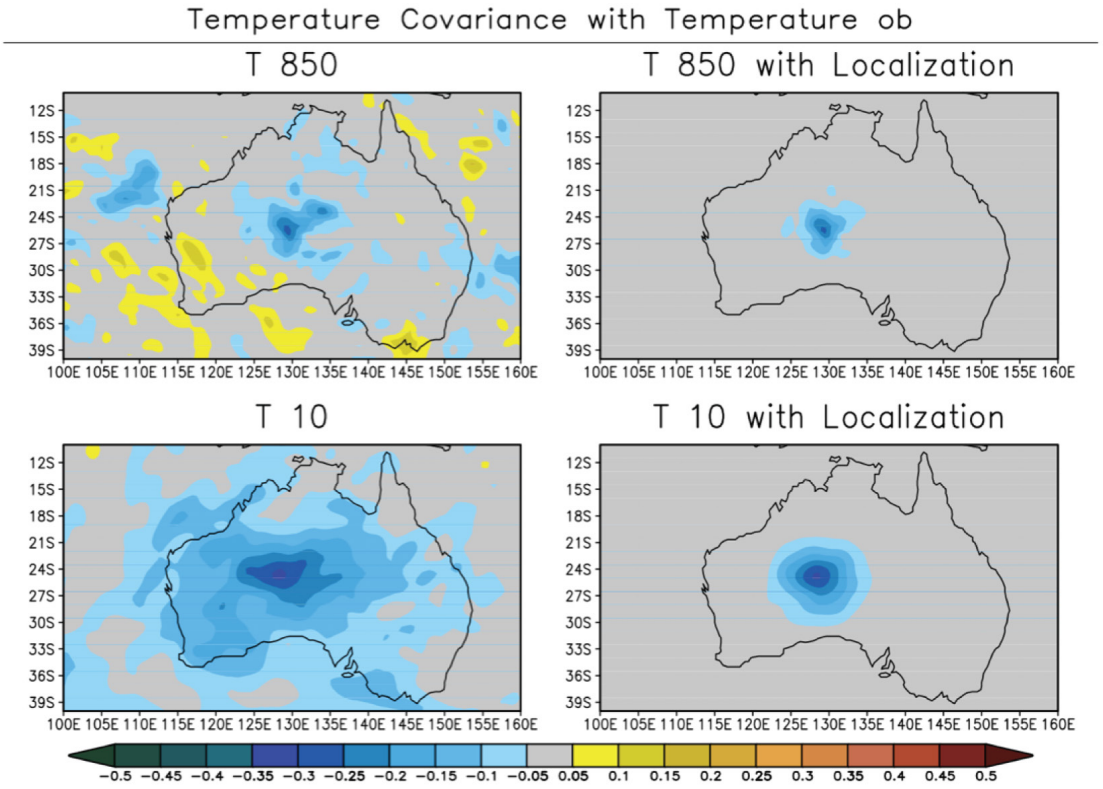

Figure 3: Covariances estimated in a EnKF between temperature at every grid point and a grid point over Australia at 850 hPa (top panels) and 10 hPa (bottom panels), before (left) and after (right) covariance localization is applied. The covariance localization function tapers to zero 1800 km from the observation location.

In a series of papers, Bishop and Hodyss (see e.g. 2007, 2011 and references therein) have devised a method for estimating a localization function from sample correlations computed smoothed, normalized ensemble perturbations. These localization functions vary with the flow situation in a continuous and physically plausible way. However, there is a significant extra cost incurred in applying this flow dependent localization, and it is not yet clear whether this extra computation results in analyses that are superior to what could be obtained by simply increasing the ensemble size for the same cost and using a fixed non-adaptive localization scheme with a larger localization radius.

Anderson (2007) proposed a different approach which he calls a "hierarchical ensemble filter", which involves running multiple parallel ensemble data assimilation cycles (effectively an ensemble of ensemble Kalman filters). The differences between the sample covariance estimates in the parallel cycles is used to construct a localization function - where the variability of the covariance estimates is much larger (smaller) than the mean covariance estimate, the localization function is set to a value near zero (one). This method is also very expensive, and the cost/benefit ratio relative to running a single larger ensemble is questionable. However, Anderson points out that the hierarchical approach could be run for only a short period, to derive a spatially varying but temporally fixed "climatological" localization function which can then be applied in single ensemble filter. Anderson (2011) also proposed a much simpler and cheaper scheme in which the localization function depends only on the sample correlation and ensemble size.

All covariance localization schemes will tend to destroy the underlying dynamical balances (such as geostrophy) inherent in the background ensemble. The more severe the localization, the larger the imbalances will be in the analysis ensemble. This means that post-analysis balance adjustments, such as digital filter finalization (Lynch and Huang, 1992), are needed to prevent gravity wave noise from building up as the system is cycled (Mitchell et al, 2002; Houtekamer and Mitchell, 2005). This motivated Kepert (2010) to design a localization scheme explicitly to maintain geostrophic balance.

3.2. Inflation: multiplicative and additive

Missing or under-represented sources of error in the background ensemble will inevitably lead to an underestimation of the background error variance, and hence to a under-utilization of observation information. Even if the forecast model is perfect, sampling error will lead to an underestimation of the variance. Multiplicative variance inflation, first proposed by Anderson and Anderson (1999), is the simplest way to deal with the inevitable deficiency in ensemble spread in EnKF systems. In its simplest form, deviations from the ensemble mean are simply inflated by a constant factor $r > 1$, thereby increasing the ensemble variance by a factor r^2. Sacher and Bartello (2008) noted that sampling errors in the estimate of the Kalman gain should be proportional to the Kalman gain itself, so that more variance inflation should be needed where the Kalman gain suggests observations should make a larger correction to the background. Whitaker and Hamill (2012) devised a simple multiplicative inflation scheme call "relaxation to prior spread' (RTPS) that inflates the posterior ensemble proportional to the amount that the ensemble variance was reduced by the assimilation of observations. The inflation factor has the form

$$r = \alpha\left(\frac{\sigma_b - \sigma_a}{\sigma_a} + 1\right) \tag{5}$$

Where σ_b is the background error standard deviation, σ_a is the analysis error standard deviation (prior to inflation), and α is a tunable parameter. If α is zero there is no inflation, and if $\alpha = 1$ the analysis variance is equal to the background variance. This scheme has the desirable property that regardless of the value of α, no inflation will occur where observations have no impact on the model state. Anderson (2009) proposed a Bayesian algorithm for estimating a spatially and temporally varying field of covariance inflation within the data assimilation. When run as part of an EnKF assimilation system using a global general circulation model with all "conventional" (i.e. non satellite radiance) observations, the Bayesian algorithm produces a spatial field of inflation that looks very similar to that implied by RTPS inflation (equation 5), with large values of inflation in regions of dense and/or accurate observations, like North America and Europe (Fig. 13 in Anderson et al. (2009)).

Instead of increasing the amplitude of existing ensemble perturbations, one can add new structures to the ensemble by adding random samples from a specified covariance distribution. This is known as additive inflation (Mitchell and Houtekamer 2000). Additive inflation is often intended to represent model error, so that the distribution the random samples are drawn from is assumed to represent the climatological statistics of the random (non-systematic) component of model error. Zhang et al (2004) proposed an algorithm that contains both multiplicative and additive aspects, which Whitaker and Hamill (2012) denote "relaxation-to-prior perturbations", or RTPP inflation. This method relaxes the posterior ensemble perturbations back to the prior perturbations via

$$\mathbf{x}'_{\mathbf{a}} \rightarrow (1-\alpha)\,\mathbf{x}'_{\mathbf{a}} + \alpha\,\mathbf{x}'_{\mathbf{b}} \tag{6}$$

Whitaker and Hamill (2012) found that RTPS inflation (equation 5) generally outperforms RTPP inflation (equation 6), while both outperform constant variance inflation, in idealized data assimilation experiments with a two-level primitive equation model including the effects of truncation model error. RTPP inflation does preserve the balances inherent in the background ensemble better, and leads to analysis ensemble perturbations that grow faster during the forecast. A combination of additive inflation and multiplicative RTPS inflation was found to work better than either alone when both model and sampling error were present, while RTPS inflation alone worked best in a perfect model environment and additive inflation alone worked best when sampling error was minimal. This is consistent with the notion that multiplicative inflation works best representing observation-network dependent sources of error in the assimilation system, such as sampling error, while additive inflation works best for representing sources of error not related to the assimilation system, such as model error. Since inflation is simple to implement, it is useful as a baseline for measuring the performance of more sophisticated methods for representing missing sources of error in the background ensemble, such as those discussed in the next section.

3.3. Stochastic and multi-model approaches

Several methods for accounting for model uncertainty within the forecast model itself have been developed in the context of ensemble prediction. These include stochastic energy backscatter (SKEB: Shutts 2005; Berner et al 2009), stochastically perturbed physics tendencies (SPPT: Buizza et al, 1999), and stochastic convection parameterization (Plant and Craig, 2008; Teixeira and Reynolds, 2008). The performance of SKEB and SPPT within the ECMWF ensemble prediction system was examined in Palmer et al 2009. An alternative approach to sampling model uncertainty is to use models with traditional deterministic physics parameterizations, but design an ensemble so that

different members are integrated with different models with different dynamical core formulations and/or physical parameterizations. If each of the component models is equally skillful, and the differences between the models spans the range of uncertainty in the models themselves, then the multi-model approach is a viable approach for representing model error. However, there may be a large development cost associated with maintaining multiple forecast models so that each has comparable forecast skill. These techniques and others where the subject of a recent ECMWF workshop (http://www.ecmwf.int/publications/library/do/references/list/201106), and the reader is referred to the workshop proceedings for a more detailed discussion of the state of the art in this field.

The performance of schemes for representing model uncertainty in ensemble prediction systems are typically evaluated using spread/skill consistency metrics or probabilistic verification scores. Evaluating these schemes in the context of ensemble data assimilation provides a more rigorous test, especially when their performance is measured relative to the inflation schemes mentioned in the previous section. This is because in order to improve upon inflation, a scheme must improve the representation of background-error covariances in the ensemble, whereas any scheme that increases the variance of a spread-deficient ensemble prediction system is likely to improve probabilistic skill scores. Houtekamer et al (2009) examined the impact of SKEB, SPPT and multi-physics ensembles in the operational Environment Canada EnKF data assimilation system, with additive inflation (using perturbations drawn from an isotropic, homogeneous covariance matrix) as a baseline. None of the candidate schemes performed as well as additive inflation, although the multi-physics ensemble in conjunction with additive inflation performed slightly better than additive inflation alone. They attributed the relatively disappointing impact of SKEB and SPPT in the data assimilation to the fact that they were tuned to perform well within the medium-range ensemble prediction system, not for the short-range forecasts required for data assimilation, and that other under-represented sources of error in the EnKF system (such as sampling error, mis-specification of observation errors, and forward observation operator errors) may be as large or larger than model uncertainty.

4. Summary

EnKF data assimilation systems have matured considerably since the last ECMWF Data Assimilation seminar. Several centers are now using them operationally, either as standalone systems are as part of a hybrid variational-ensemble data assimilation system. Both variational and ensemble-based data assimilation systems are approximate solutions to the Kalman Filter, but each makes different, perhaps complementary approximations. Therefore, it is likely that a combination of the two can produce an analysis that combines the “best of both worlds”, and performs better than either alone. Table 1 presents a summary of the relative strengths of the EnKF and 4DVar.

The subject of hybrid variational/ensemble data assimilation is covered in the article by Dale Barker in this volume.

Features from EnKF	Features from 4DVar
Extra flow-dependence of $\mathbf{P}^b$, propagation of flow-dependent information from one cycle to the next.	Covariance localization done correctly (in model space, not observation space).
More flexible treat of model error (within the ensemble).	Reduction in sampling error in the estimation of time-lagged covariances (through full-rank propagation of $\mathbf{P}^b$ through the assimilation window by the tangent linear model).
Automatic initialization of ensemble forecasts.	Flexibility to add extra constraints to the cost function (to preserve balance for example).

Table 1: Summary of the unique benefits offered by EnKF and 4DVar. Hybrid variational/ ensemble data assimilation systems offer the potential for incorporating all of these features together in one system, and hence to provide the "best of both worlds".

EnKF research has now shifted from the development of new algorithms for calculating analysis increments to improving the treatment of sampling error and model uncertainty, as well as other potentially under-sampled sources of error in the background ensemble. Efforts to develop practical methods for flow-adaptive covariance localization are ongoing, but have not yet been shown to be worth the extra computational effort - many centers instead are electing to use extra computation resources to run larger ensembles (Environment Canada has recently increased their EnKF ensemble size to 192 members). Indeed, now that ensemble-based and hybrid variational/ensemble data assimilation systems offer the prospect of improving even deterministic analyses and forecasts by increasing ensemble size, operational centers now may now have to think twice before simply increasing model resolution with each computer upgrade.

Acknowledgements

Thanks to Xuguang Wang and Tom Hamill for producing Figure 1, and Phil Pegion for producing Figure 3. Thanks to Massimo Bonavita, Lars Isaksen and Mats Hamrud for many illuminating discussions about ensemble data assimilation during my visits to ECMWF in 2010 and 2011.

References

Anderson, J. L. and S. L. Anderson, 1999: A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. *Mon. Wea. Rev.*, **127**, 2741–2758.

Anderson, J. L., 2001: An ensemble adjustment filter for data assimilation. *Mon. Wea. Rev.*, **129**, 2884-2903.

Anderson, J.L., and N. Collins, 2007: Scalable implementations of ensemble filter algorithms for data assimilation. *J. Atmos. Oceanic Tech*, **24**, 1452–1463.

Anderson, J. L., 2007: Exploring the need for localization in the ensemble data assimilation using a hierarchical ensemble filter, *Physica* D, **230**, 99–111.

Anderson, J. L., T. Hoar, K. Raeder, H. Liu, N. Collins, R. Torn, and A. Avellano, 2009: The data assimilation research testbed: A community facility. *Bull. Amer. Meteor. Soc*., **90**, 1283–1296.

Anderson, J. L., 2009: Spatially and temporally varying adaptive covariance inflation for ensemble filters. *Tellus* A, **61**, 72–83.

Anderson, J. L, 2011: Localization and Correlation in Ensemble Kalman Filters. *Presented at the 91st Annual Meeting of the AMS* in Seattle, WA, USA. (http://www.image.ucar.edu/pub/DART/2011/JLA_sec_seattle.pdf)

Berner, J., G. J. Shutts, M. Leutbecher, and T. N. Palmer, 2009: A spectral stochastic kinetic energy backscatter scheme and its impact on flow-dependent predictability in the ECMWF ensemble prediction system. *J. Atmos. Sci*., **66**, 603–626.

Bishop, C. H., and D. Hodyss, 2007: Flow-adaptive moderation of spurious ensemble correlations and its use in ensemble-based data assimilation, *Quart. J. Roy. Meteor. Soc*., **133**, 2029–2044.

Bishop, Craig H., Daniel Hodyss, 2011: Adaptive Ensemble Covariance Localization in Ensemble 4D-VAR State Estimation. *Mon. Wea. Rev*., **139**, 1241–1255.

Buizza, R. and T. N. Palmer, 1999: Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc*., **125**, 2887–2908.

Burgers, G., P. J. van Leeuwen, and G. Evensen, 1998: Analysis scheme in the ensemble Kalman filter. *Mon. Wea. Rev*., **126**, 1719-1724.

Campbell, W., Bishop, C. and D. Hodyss, 2010: Vertical Covariance Localization for Satellite Radiances in Ensemble Kalman Filters. *Mon. Wea. Rev*., **138**, 282–290.

Evensen, G., 1994: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics, *J. Geophys. Res*., **99**(C5), 143-162.

Evensen, G, 2003: The ensemble Kalman filter: Theoretical formulation and practical implementation, *Ocean Dynamics*, **53**, pp. 343--367.

Gaspari, G., and S. Cohn, 1999: Construction of correlation functions in two and three dimensions, *Quart. J. Roy. Meteor. Soc*., **125**,723-757.

Gelb, A., J. F. Kasper, R. A. Nash, C. F. Price, and A. A. Sutherland, 1974: *Applied Optimal Estimation*. M. I. T. Press, 374 pp.

Hamill, T. M., J. S. Whitaker, and C. Snyder, 2001: Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter, *Mon. Wea. Rev*., **129**, 2776–2790

Hamill, T. M., 2006: Ensemble-based atmospheric data assimilation. Chapter 6 of *Predictability of Weather and Climate*, Cambridge Press, 124-156.

Houtekamer, P. L., and H. L. Mitchell, 1998: Data assimilation using an Ensemble Kalman Filter technique, *Mon. Weather Rev*., **126**, 796–811.

Houtekamer, P.L and H. L. Mitchell, 1998: Data assimilation using an ensemble Kalman filter technique. *Mon. Wea. Rev*., **126**, 796-811.

Houtekamer, P. L., H. L. Mitchell, G. Pellerin, M. Buehner, M. Charron, L. Spacek, and B. Hansen, 2005: Atmospheric data assimilation with an Ensemble Kalman Filter: Results with real observations. *Mon. Wea. Rev*., **133**, 604–620.

Houtekamer, P. L., H. L. Mitchell, and X. Deng, 2009: Model error representation in an operational ensemble Kalman filter. *Mon. Wea. Rev*., **137**, 2126–2143.

Hunt, B. R., E. J. Kostelich, and I. Szunyogh, 2007: Efficient data assimilation for spatiotemporal chaos: a local ensemble transform Kalman filter. *Physica* D, **230**, 112-126.

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

Kalman, R., and R. Bucy, 1961: New results in linear filtering and prediction theory. Transactions of the ASME. Series D, *Journal of Basic Engineering*, **83**, 95-107.

Kepert, J. D, 2010: Balance-aware covariance localisation for atmospheric and oceanic ensemble Kalman filters. *Computational Geosciences*, **15**, 239-250.

Lawson, W. G., and J. A. Hansen, 2004: Implications of stochastic and deterministic filters as ensemble-based data assimilation methods in varying regimes of error growth, *Mon. Wea. Rev*., **132**, 1966–1981.

Lorenc, A., 1981: A Global Three-Dimensional Multivariate Statistical Interpolation Scheme. *Mon. Wea. Rev*., **109**, 701-721.

Lynch, P., and X. Y. Huang, 1992: Initialization of the HIRLAM model using a digital filter. *Mon. Wea. Rev*., **120**, 1019–1034.

Mitchell, H. L. and P. L. Houtekamer, 2000: An adaptive ensemble Kalman filter. *Mon.Wea. Rev*., **128**, 416–433.

Mitchell, H. L., P. L. Houtekamer, and G. Pellerin, 2002: Ensemble size, and model error representation in an Ensemble Kalman Filter, *Mon. Wea. Rev*., **130** , 2791–2808

Palmer, T. N., Buizza, R., Doblas-Reyes, F., Jung, T., Leutbecher, M., Shutts, G.J., Steinheimer M., & Weisheimer, A., 2009: Stochastic parametrization and model uncertainty. *ECMWF Research Department Technical Memorandum* no. 598, ECMWF, Shinfield Park, Reading, UK, pp. 42.

Plant, R.S. and Craig, G.C., 2008: A stochastic parameterization for deep convection based on equilibrium statistics. *J. Atmos. Sci*., **65**, 87-105.

Sacher, W. and P. Bartello, 2008: Sampling errors in ensemble Kalman filtering. part I: Theory. *Mon. Wea. Rev*., **136**, 3035–3049.

Teixeira, J., and C. A. Reynolds, 2008: Stochastic nature of physical parameterizations in ensemble prediction: A stochastic convection approach. *Mon. Wea. Rev*., **136**, 483–496.

Tippett, M. K., J. L. Anderson, C. H. Bishop, T. M. Hamill, and J. S. Whitaker, 2003: Ensemble square-root filters, *Mon. Weather Rev*., **131**, 1485–1490.

Whitaker, J. S. and T. M. Hamill, 2002: Ensemble data assimilation without perturbed observations. *Mon. Wea. Rev*., **130**, 1913–1924.

Whitaker, J. S. and T. M. Hamill, 2012: Evaluating methods to account for system errors in ensemble data assimilation, *Mon. Wea. Rev*. **140**, accepted.

Zhang, F., C. Snyder, and J. Sun, 2004: Impacts of initial estimate and observation availability on convective-scale data assimilation with an Ensemble Kalman Filter. *Mon. Wea. Rev*., **132**, 1238–1253.