# Practical ensemble data assimilation

**P.L. Houtekamer and Herschel L. Mitchell**

*Service Météorologique du Canada / Meteorological Service of Canada,*
*Montreal, Québec, Canada*
*peter.houtekamer@ec.gc.ca, herschel.mitchell@ec.gc.ca*

## 1 Introduction

The ensemble Kalman filter (EnKF) is a 4-dimensional data-assimilation method that uses a Monte-Carlo ensemble of short-range forecasts to estimate the covariances of the forecast error (Evensen 1994; Houtekamer and Mitchell 1998, hereafter HM98).

It is a close approximation to the Kalman filter, which provides the least-squares solution in the case of small errors with Gaussian distribution (Maybeck 1979). The approximation becomes "automatically" more accurate as bigger ensembles are used. This is a desirable property in an environment where massively parallel computers become increasingly powerful.

The EnKF does not depend strongly on the validity of questionable hypotheses (linearity of the model dynamics) and is conceptually simple. It does not require the development or existence of an adjoint or tangent linear model. It is therefore an extremely attractive method.

An EnKF is being developed for operational implementation at the Meteorological Service of Canada (MSC). The different stages of this project have been well documented (HM98; Mitchell and Houtekamer 2000, hereafter MH; Houtekamer and Mitchell 2001, hereafter HM01; Mitchell et al. 2002, hereafter MHP, and Houtekamer et al. 2003, hereafter HEA). We shall summarize some key results from the above studies. The main point of this paper is that the development of an EnKF is feasible for an operational centre with perhaps limited resources. A more complete overview of the literature on the EnKF algorithm can be found in Evensen (2003) and Tippett et al. (2003).

In section 2, we shall discuss the link between the EnKF and the traditional Kalman filter. Next, in section 3, we briefly present the experimental environment that is used for most of the examples of this paper. We discuss the localization of covariances in section 4 and the simulation of model error in section 5. These two components are of critical importance for implementations of the EnKF that assimilate real observations and that use realistic atmospheric models. The dominant role of the model-error term in the error dynamics is highlighted in section 6. A comparison with an operational 3-dimensional data-assimilation algorithm is presented in section 7. We conclude the paper with a summary.

## 2 Link with the Kalman filter

In this section, we consider the Kalman filter equations. The standard technique of sequentially assimilating batches of observations greatly reduces the otherwise potentially prohibitive numerical cost of the matrix inversions. The use of ensembles reduces the storage requirements and the cost of covariance integration as compared to the Kalman filter. The use of small ensembles may lead to an unrealistically small ensemble spread. This problem is alleviated using a configuration with a pair of ensembles (HM98, Fig. 3; HM01, Fig. 3).

## 2.1 Kalman filter equations

The Kalman filter equations provide the optimal (minimum variance) solution to the data assimilation problem in the case of linear dynamics and Gaussian error statistics (Maybeck, 1979):

$$\begin{aligned}
\Psi^a(t) &= \Psi^f(t) + \mathrm{K}(o - \mathrm{H}\Psi^f(t)), && (1)\\
\Psi^f(t+1) &= M(\Psi^a(t)), && (2)\\
\mathrm{K} &= \mathrm{P}^f(t)\mathrm{H}^T(\mathrm{H}\mathrm{P}^f(t)\mathrm{H}^T + \mathrm{R})^{-1}, && (3)\\
\mathrm{P}^a(t) &= (\mathrm{I} - \mathrm{K}\mathrm{H})\mathrm{P}^f(t), && (4)\\
\mathrm{P}^f(t+1) &= \mathrm{M}\mathrm{P}^a(t)\mathrm{M}^T + \mathrm{Q}. && (5)
\end{aligned}$$

The symbols have the following definitions:

| | | | |
|---|---|---|---|
| $o$ | vector of observations, | H | forward interpolation matrix, |
| $\Psi^a$ | analysis, | $\Psi^f$ | first-guess field (also called prior), |
| $\mathrm{P}^a$ | covariance of the analysis error, | $\mathrm{P}^f$ | covariance of the forecast error, |
| R | covariance of the observational error, | K | gain matrix, |
| M | tangent linear operator, | Q | covariance of the model error, |
| $M$ | nonlinear forecast model, | $t, t+1$ | consecutive analysis times. |

The first two equations (Eqs. 1 and 2) also appear in 3-dimensional analysis algorithms (Lorenc 1986). The distinguishing feature of the Kalman filter approach is that the gain matrix, K, is computed using an evolving estimate of the forecast-error covariance matrix, $\mathrm{P}^f$. The gain matrix ensures an optimal weight, in Eq. 1, for the observations, $o$, and the prior estimate, $\Psi^f$. The improvement, due to the assimilation of new observations, is measured with Eq. 4. The resulting covariance estimate is integrated in time using Eq. 5. Note that we have used the nonlinear forecast model $M$ in Eq. 2 and the tangent linear model M in Eq. 5 as is done in the extended Kalman filter (Gauthier et al. 1993, Eq. 2.14).

## 2.2 sequential assimilation of observations

To avoid the storage and inversion (Eq. 3) of large matrices, the observations are organized into batches that are assimilated sequentially (Anderson and Moore 1979, pp. 142-146; HM01). This idea can be illustrated with a one-dimensional example in which we have a scalar prior estimate, $f$, and two corresponding independent observations, $o_1$ and $o_2$. All items have a standard normal distribution with zero expectation and unit variance. Not surprisingly, we obtain the same weight for the prior and the two observations.

$$\begin{aligned}
f &: \mathrm{N}(0,1), \quad o_1 : \mathrm{N}(0,1), \quad o_2 : \mathrm{N}(0,1),\\
a &= k_f f + k_1 o_1 + k_2 o_2,\\
k_f &= \frac{\sigma_f^{-2}}{\sigma_f^{-2} + \sigma_{o_1}^{-2} + \sigma_{o_2}^{-2}}, \qquad k_f = k_1 = k_2 = \frac{1}{3},\\
a &: \mathrm{N}(0, \frac{1}{3}).
\end{aligned}$$

The analysis has a normal distribution with zero expectation and variance $\frac{1}{3}$.

Alternatively, we can first assimilate observation $o_1$ to obtain an improved prior $f_2$. Subsequently, we combine $f_2$ with the second observation $o_2$. Working through the algebra, we notice that both procedures lead to the same resulting analysis $a$, which has zero expectation and variance $\frac{1}{3}$.

Step 1 of the sequential assimilation:

$$\begin{aligned} f_1 &: \mathrm{N}(0,1), \quad o_1 : \mathrm{N}(0,1), \\ f_2 &= k_{f_1} f_1 + k_1 o_1, \qquad k_{f_1} = k_1 = \frac{1}{2}, \\ f_2 &: \mathrm{N}(0,\frac{1}{2}). \end{aligned}$$

Step 2 of the sequential assimilation:

$$\begin{aligned} f_2 &: \mathrm{N}(0,\frac{1}{2}), \quad o_2 : \mathrm{N}(0,1), \\ a &= k_{f_2} f_2 + k_2 o_2, \qquad k_{f_2} = \frac{2}{2+1}, \qquad k_2 = \frac{1}{3}, \\ a &= \frac{1}{3}(f_1 + o_1 + o_2), \\ a &: \mathrm{N}(0,\frac{1}{3}). \end{aligned}$$

In atmospheric data-assimilation, we typically encounter $n_{obs} = O(10^5)$ or more observations. The cost of the matrix inversion in Eq. 3 is proportional to $n_{obs}^3$. The storage and inversion of matrices of order $O(10^5)$ would provide a computational challenge even on modern computers. At MSC, we use a sequential EnKF procedure with $O(10^3)$ batches of $O(100)$ observations each. With this batch size, the cumulative computational cost of the matrix inversions is still relatively small. We think that using a moderate batch size will facilitate the eventual implementation of a buddy-check quality-control procedure (Dee et al. 2001) and also the eventual accounting for correlated observations. Several groups propose square root EnKF algorithms that assimilate observations one at a time (Tippett et al. 2003).

### 2.3 evaluation of matrices

For the EnKF algorithm, the covariance matrices $\mathrm{P}^f$ and $\mathrm{P}^a$ need not be computed and stored. Instead, ensemble-based covariance-estimates for smaller matrices can be used. In fact, to compute the Kalman gain, we only need to estimate the terms $\mathrm{P}^f\mathrm{H}^T$ and $\mathrm{HP}^f\mathrm{H}^T$. These can be computed from the ensemble (HM98) as follows:

$$\mathrm{P}^f = \frac{1}{N-1}\sum_{i=1}^{N}(\Psi_i^f - \overline{\Psi_i^f})(\Psi_i^f - \overline{\Psi_i^f})^T \qquad \text{(not needed)} \tag{6}$$

$$\mathrm{P}^f\mathrm{H}^T = \frac{1}{N-1}\sum_{i=1}^{N}(\Psi_i^f - \overline{\Psi_i^f})(H(\Psi_i^f) - \overline{H(\Psi_i^f)})^T \tag{7}$$

$$\mathrm{HP}^f\mathrm{H}^T = \frac{1}{N-1}\sum_{i=1}^{N}(H(\Psi_i^f) - \overline{H(\Psi_i^f)})(H(\Psi_i^f) - \overline{H(\Psi_i^f)})^T. \tag{8}$$

Since $H$ is applied to each $\Psi_i^f$ individually on the right-hand side of Eqs. 7 and 8 (rather than to the covariance matrix $\mathrm{P}^f$), it is possible to use nonlinear forward interpolation operators (HM01). For example, $H$ can be a radiative transfer model, if radiance observations are available. The cost involved with the estimation of $\mathrm{P}^f\mathrm{H}^T$

and with the multiplication by $\mathrm{P}^f\mathrm{H}^T$ will likely dominate the other costs involved with the data-assimilation step (Eqs. 1 and 3). On modern computers, efficient algorithms for matrix multiplication are readily available. However, the developer will have to implement a parallelization strategy that allows for the distribution of operations involving the matrix $\mathrm{P}^f\mathrm{H}^T$ (HM01; Keppenne and Rienecker 2002).

The dimension of the full matrix $\mathrm{P}^f$, used in the Kalman filter, is the size of the phase space (order 1 000 000). However, the dimension of the ensemble-based estimate (Eq. 6) will be at most identical to $N-1$ (order 100). Similarly the term $\mathrm{P}^f\mathrm{H}^T$ will at most be of order $N-1$, and consequently the analysis increment will be in a space of dimension at most $N-1$. This implies that the (of order 100 000) observations will be projected onto a very low-dimensional space to form an analysis increment. The resulting loss of information is not acceptable.

The covariance localization, discussed in section 4, substantially reduces the dimensionality problem.

### 2.4 transport of covariance information

In the Kalman filter, the covariance information is transported using Eq. 5. To integrate the covariances, the tangent linear model has to be run once for each coordinate of the model. In practice, with a model-state vector of order $O(10^6)$, the cost of this operation is prohibitive. Instead, in the EnKF algorithm, we only have to perform as many integrations with the forecast model as we have members in the ensemble:

$$\Psi_i^f(t+1) = M(\Psi_i^a) + q_i, \qquad i = 1, \cdots, N \tag{9}$$

The replacement of the tangent linear model M by a model $M$ with full physics is an improvement. The full model will properly deal with saturation of errors. We also have additional flexibility to deal with model error. It can be sampled from a covariance matrix or also be simulated using, for instance, different realizations $M_i$ of the nonlinear forecast model.

With a sample, we may in principle transport information on higher moments that is lost when covariance matrices are used as in the Kalman filter. However, it is important to note that the reliable estimation of higher moments will require a relatively large ensemble. Also, the use of a gain matrix in the assimilation step is sub-optimal for non-Gaussian distributions of the prior. Nonlinear filtering methods have been proposed by Anderson and Anderson (1999) to deal with non-Gaussian prior distributions.

### 2.5 configuration

The EnKF can be considered to be a Monte Carlo method, where random input is used to obtain information about the accuracy of the system. The input to the data assimilation step consists of a prior and of a set of observations. The uncertainty in the prior is sampled using an ensemble of prior fields and the uncertainty in the observations is sampled using an ensemble of randomly perturbed observations, where the perturbations for each observation are consistent with its error statistics. The output consists of a set of analyses that reflects the uncertainty in the ensemble mean analysis.

In the EnKF, the set of prior fields is used not only to determine the weights used by the analysis but also to assimilate the observations, resulting in an analysis ensemble and hence an analysis-error-covariance matrix, $\mathrm{P}^a$. Thus the same sample is used to calculate the gain and estimate the error associated with using that gain. This duplicate use of the same information, also known as inbreeding, may lead to an unrealistically small spread in the ensemble (HM98). This problem is particularly important for ensembles with less than, say, 100 members. One approach to correct for this short-coming of the data-assimilation procedure and to arrive at a more realistic spread, is to inflate the ensemble spread by means of a “model error” term.

An alternative approach, known as the double EnKF or DEnKF, is displayed in Fig. 1. The ensemble is divided into a pair of two sub-ensembles where the gain used for one sub-ensemble is computed using the prior fields

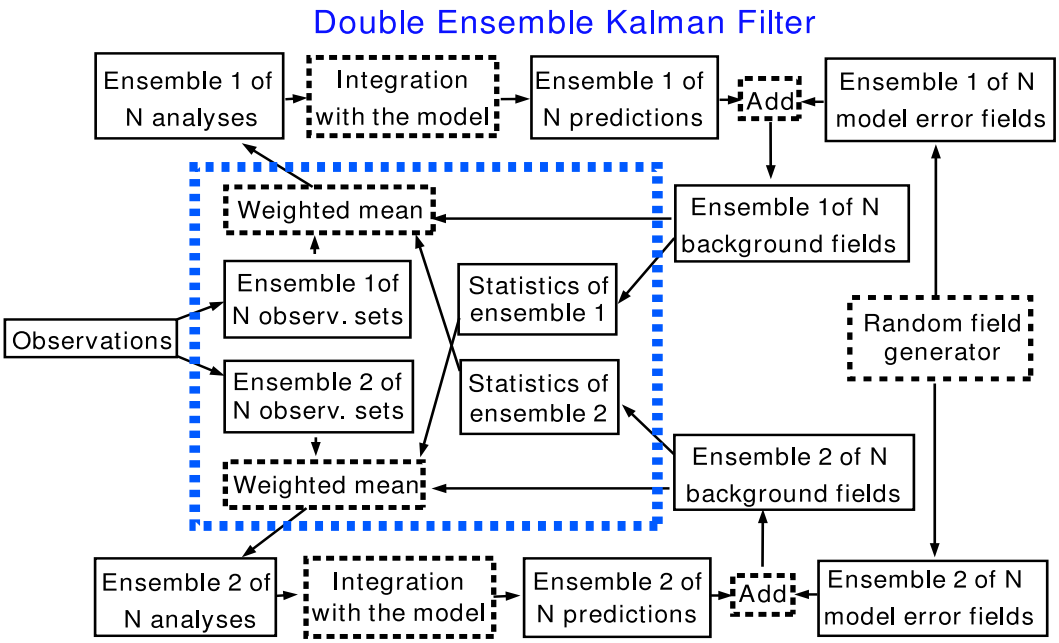


*Figure 1: The procedure used to perform a data-assimilation cycle with the double EnKF (from HEA).*

from the other sub-ensemble.

The advantage of this procedure is that the ensemble spread should now correspond with the actual ensemble mean error. This property facilitates the interpretation of experimental results. Consequently, the DEnKF is being used at MSC to arrive at an operational implementation. The disadvantage is that the covariance estimates are noisier due to the smaller size of the sub-ensembles (HM98).

It is not clear if the square-root technique can be applied in the case of a double EnKF and moderately sized batches of observations. More research can be done towards the optimal way of configuring the ensemble members, but this issue is probably not of critical importance when a total of, say, 100 members or more can be used. We shall be showing examples with ensemble sizes of $2 \times 48$ members and $2 \times 64$ members.

## 3 Experimental environment

A number of increasingly realistic experimental environments have been used in our sequence of studies on the EnKF algorithm. Because we aim at an eventual operational implementation at MSC, we now try to use the same forecast model and the same observational network as are used for our centre's deterministic forecast.

A number of data-assimilation cycles have been performed, starting 0000 UTC 19 May 2002 and ending 1200 UTC 2 June 2002 (HEA). The first four days of the experiment are usually discarded to permit the stabilization of the ensemble statistics. The verification is against a particularly reliable subset of the radiosonde network.

### 3.1 the model

We use the Global Environmental Multiscale (GEM) Model of MSC (Côté et al. 1998). The model is much like the version used to produce the higher resolution medium-range deterministic forecast at our centre. The main modifications are in the scripts submitting the model, to allow a large number of short integrations to be run simultaneously on the parallel computer. There is no need in the EnKF algorithm for the tangent linear and adjoint versions of the forecast model.

To facilitate experiments, we use a lower ($240 \times 120$) horizontal resolution than in the deterministic operational forecast. A timestep of 60 minutes can be used at this resolution. As in the operational configuration, we use 28 levels with the model top located at $p_{top} = 10$ hPa.

An $\eta$ vertical coordinate is used:

$$\eta \equiv \frac{p - p_{top}}{p_{surface} - p_{top}}.$$

The model starts from the fields of $u, v, T$, and specific humidity $q$ at each $\eta-$level and the surface pressure field $p_{surf}$. At the end of the 6-h integration, the model outputs the same fields as well as the skin temperature.

### 3.2 the observations

We try to use all observations that are assimilated by the deterministic analysis (3d-var) at our centre (Gauthier et al. 1999b). We thus benefit from the operational quality control procedure that consists of a "background check" and a "variational quality control".

Because we use a different (lower-resolution) orography, we verify that surface observations are not too far from the model surface and that upper air observations are not too close to the model surface.

We use the same error statistics for the observations as the 3d-var.

Currently we assimilate the following elements from:

- radiosondes: $u, v, T, q, p_{surface}$,
- aircraft: $u, v, T$,
- satellite: cloud track winds $u, v$, and AMSU-A microwave radiances,
- surface observations: $T, p_{surface}$.

We do not yet assimilate surface observations of wind and humidity. For surface wind observations, we did perform some preliminary experiments that did not show a positive impact. For observations of surface humidity, we have yet to implement the interpolation operator.

At a later stage, we plan to perform the "background check" using the ensemble mean prior field. We envision implementing a "buddy check" procedure in the EnKF. We would also like to investigate and partly relax the approximation that all observational errors are independent.

## 4 Localization of covariances

The EnKF provides an analysis increment in those directions where the ensemble indicates there is uncertainty. In the other directions there is no apparent need to use observational information. Consequently, if we have 100 ensemble members, the analysis increment will be in the space spanned by these members only. If the atmosphere has more than 100 degrees of freedom, as everyone believes, this will be a problem.

Artificial measures, not suggested by the Kalman filter equations, will be necessary to inflate the dimensionality of the ensemble. The result will necessarily be that the analysis increment will no longer be in the space spanned by the ensemble members. One has a trade-off between (i) producing nicely balanced analyses that remain far from the O(100 000) observations and (ii) having less balanced analyses that fit all observations pretty well (MHP).

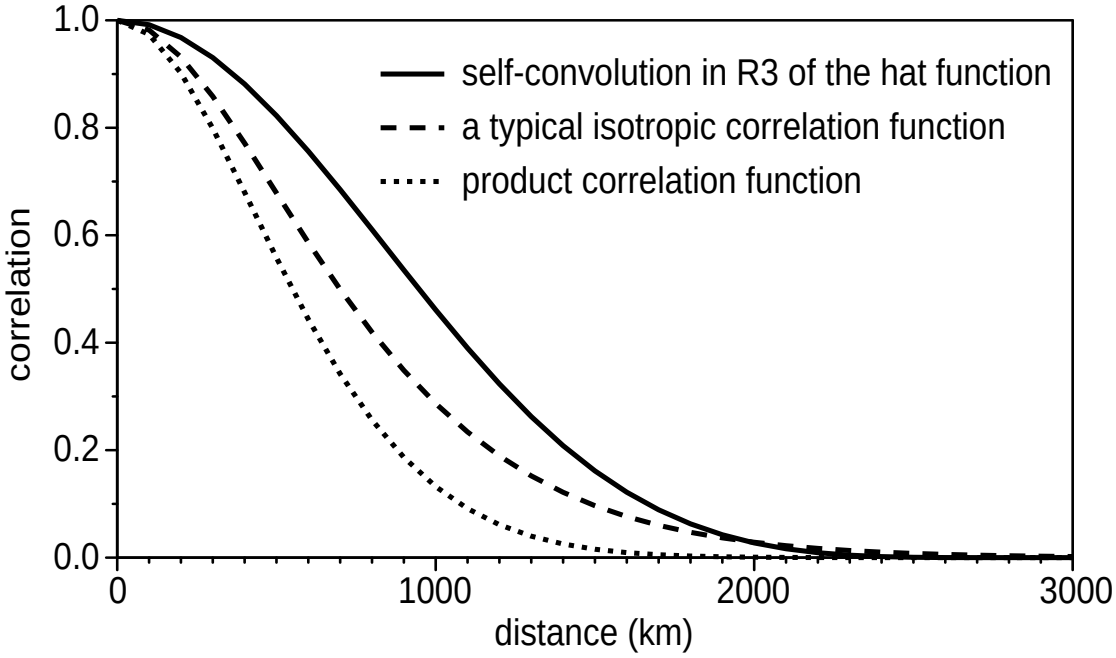


*Figure 2: The solid line is for a correlation function that becomes exactly zero at a distance of 2800 km. The dashed line is for a correlation function as typically used in 3-d algorithms. The dotted line is the product of the first two functions.*

### 4.1 horizontal localization

A Hadamard product (Gaspari & Cohn 1999), which does an element-wise product of two matrices, can be used to localize ensemble based covariances. This leads to a positive definite matrix $\mathrm{P}^f$:

$$P^f(r_i, r_j) = P^f_{ensemble}(r_i, r_j)\rho(r, L).$$

The ensemble based covariances $P^f_{ensemble}(r_i, r_j)$ contain dynamically relevant information at short distance. The multiplying function $\rho(r, L)$ (Fig. 2) is close to unity at short distance but drops to exactly zero at a distance at which the ensemble based covariance estimates are dominated by sampling error (HM98, Fig. 6). The resulting function $P^f(r_i, r_j)$ still contains the relevant information at short distance but drops to zero at a distance $L$. As we get bigger ensembles, and therefore smaller sampling error, we can make $L$ longer. This increases the degree of balance of the analyses (MHP, Fig. 2).

To investigate the impact of the localization, we performed one data-assimilation cycle with an enforced zero-impact at 3400 km and one with a more severe localization at 2300 km. For both experiments, twice 48 members were used. For the last 5 days of the 14-day experiment, the innovation statistics are compared in Fig. 3. The more severe localization has a positive impact on the standard deviations of the winds, the temperature and the dewpoint depression.

Based on these results, one would use a strong localization with enforced zero-impact at 2300 km. Alternatively, one could consider using more than $2 \times 48$ members. Some concern was expressed by MHP about the possible lack of balance with localizations narrower than 2800 km. An additional cycle using localization at 2800 km showed little difference with the 2300 km run for the standard deviations. We now perform most experiments with correlations forced to zero at 2800 km.

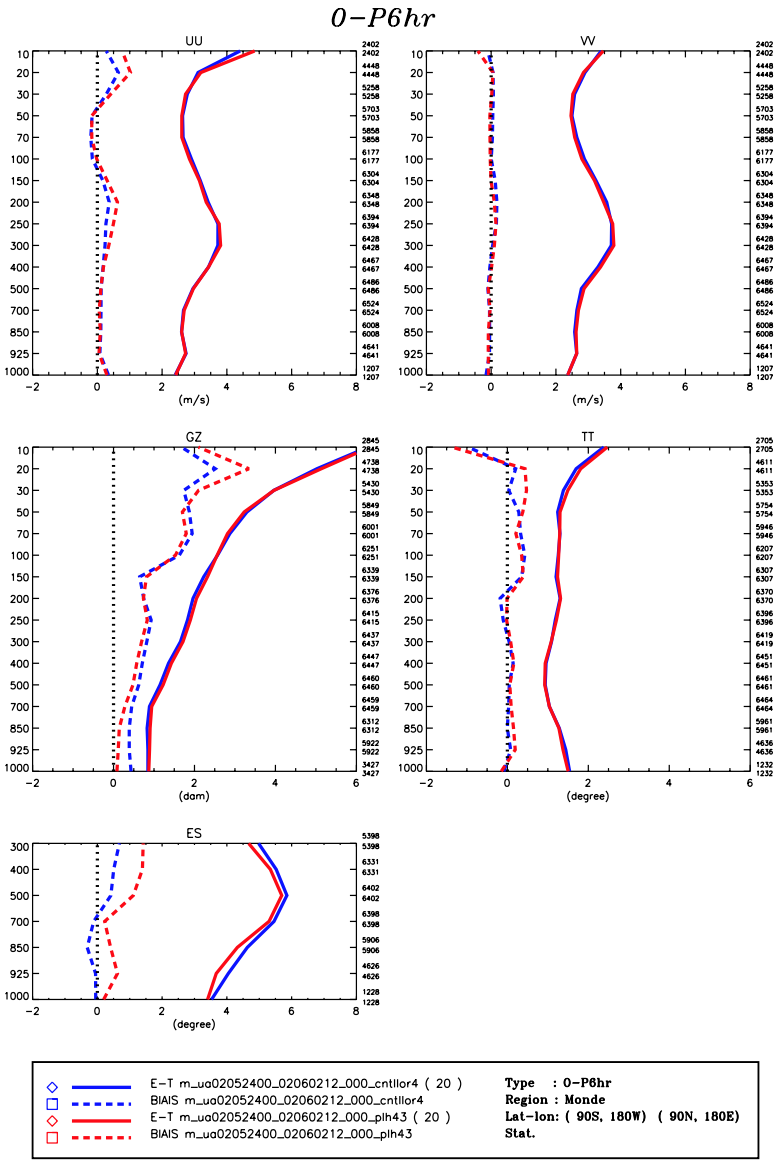


*Figure 3: Comparison of verification scores, obtained over a 5-day experimental period, with an enforced zero-impact at 3400 km (in blue) and at 2300 km (in red). The dashed lines show the bias and the solid lines show the standard deviation. The 5 panels are for the two wind components (top panels), geopotential height, temperature (middle panels) and dewpoint depression (below).*

## 4.2 vertical localization

A localization of covariance information can also be applied in the vertical (Keppenne and Rienecker 2002; Whitaker et al. 2003; HEA). The objective is again to filter distant covariance estimates that are dominated by sampling error. As in HEA, for the filtering, we measure vertical distance in units of ln(pressure). A zero-impact is enforced beyond 2 units of ln(pressure).

To confirm the need for a vertical localization, we performed two data-assimilation cycles. The reference experiment does not use a localization in the vertical. In the second run, correlations are forced to zero at 2 units of ln(pressure).

From Fig. 4, a positive impact of the localization is observed for all variables. In particular for the wind and temperature above 200 hPa, the standard deviations of the innovations are significantly smaller. This is likely

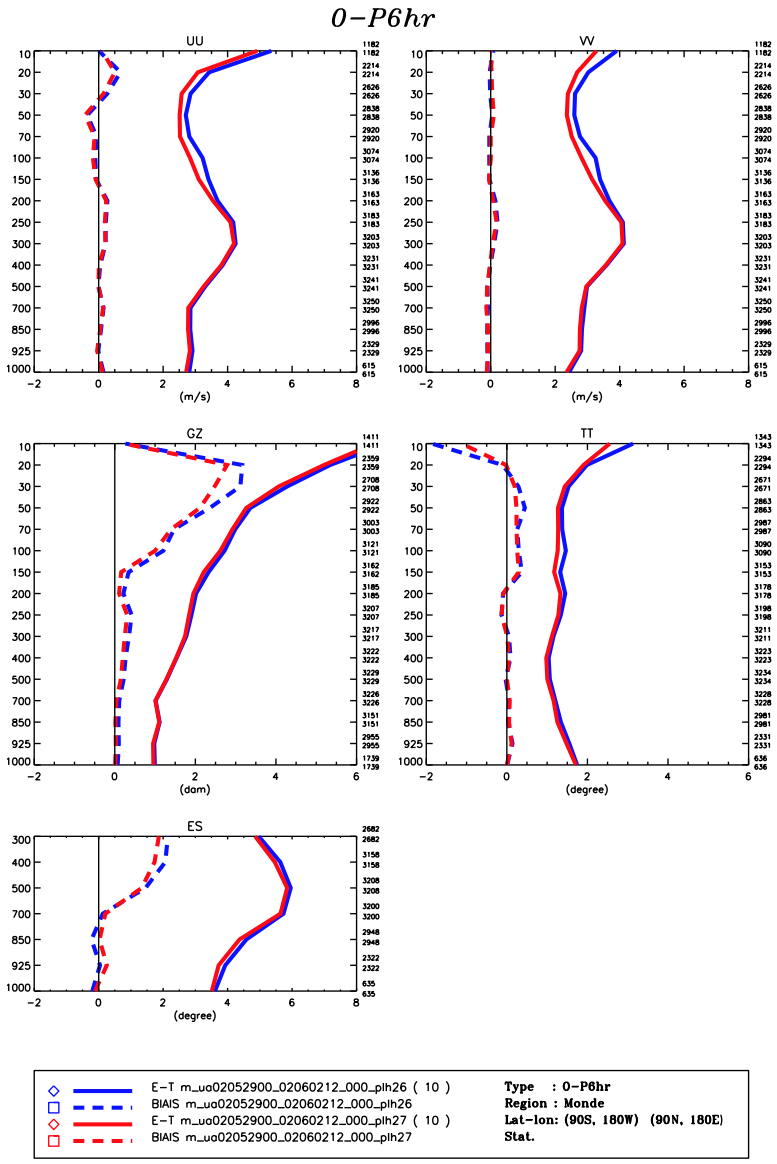


*Figure 4: Comparison of verification scores, obtained over a 5-day experimental period, obtained without vertical localization (in blue) and with correlations forced to zero at 2 units of ln (pressure). The dashed lines show the bias and the solid lines show the standard deviation. In both experiments, the number of ensemble members is twice 48.*

related to the existence of narrow vertical structures at these altitudes (HEA, Sect. 2).

## 5 Model error simulation

The EnKF provides an approximation to the Kalman filter that is computationally feasible on a modern parallel supercomputer. This does not necessarily imply that we will also obtain good results. For this it is essential to take proper account of the model-error term (Dee, 1991). In the worst case, neglecting model error entirely, we may obtain a very small spread of the ensemble. Such a small spread will cause the analysis to disregard new observations and diverge from reality. This condition is know as filter divergence.

### 5.1 isotropic model error

Little is known about the statistical properties of the model-error component (Dee, 1995). Our current hypothesis is that the model error is similar in structure to the forecast error as described in our centre's 3d-var algorithm (Gauthier et al. 1999a):

$$\mathrm{Q} = 0.25\mathrm{P}_{\mathrm{3dvar}}.$$

For each member, we use a random field generator to obtain a random model-error field $q_i$, for use in Eq. 9, that has isotropic error statistics as prescribed in the 3d-var (but with amplitudes that are smaller by a factor 0.5). The mean over the ensemble of the model-error fields $q_i$ is zero by construction.

Currently the model-error term includes:

- A balanced component that is introduced for streamfunction. After a transformation of variables, we obtain a balanced model error for the wind components $(u, v)$, the temperature and the surface pressure.
- An unbalanced temperature component that is significant near the surface, in the tropics and near the top of the model.

The rationale for this choice of Q is that the 3d-var of our centre is known to work well. Lacking any knowledge about model error, we feel that it is safe to choose a statistical description that has already been used in the 3d-var. With more knowledge about the model error, we will hopefully be able to implement more appropriate, less conservative, model-error descriptions. The main concern is that the use of realistic, highly unbalanced, model-error components could excite oscillations in the data-assimilation cycle.

### 5.2 verification of innovation amplitudes

Innovation statistics can be used to tune the model-error term (Dee 1995; MH). The covariance of the innovations can be decomposed into three components:

$$\langle \nu \nu^T \rangle = \mathrm{HPH}^T + \mathrm{HQH}^T + \mathrm{R}. \tag{10}$$

The innovation $\nu$ is computed as the difference between the interpolated ensemble mean state and the observations. The prediction-error covariance, P, is available from the ensemble and an estimate of the observation-error covariance, R, is available from the data assimilation algorithm. The remaining term, Q, can thus be adjusted such that the two sides of Eq. 10 match.

For the current study, we only considered the diagonal of Eq. 10. We performed a number of preliminary experiments to tune the amplitude and the length scale of the model-error component. The result of this procedure is a fairly nice agreement (Fig. 5 and HEA) between the observed and predicted innovation amplitudes (i.e. the left- and right-hand sides of Eq. 10).

The predicted innovations, in Fig. 5, appear to be too large for winds near the model top. However, we are not sure that these discrepancies are bigger than the uncertainty in the experimental procedure. For instance, our use of a subset of particularly reliable radiosonde stations for the verification may have caused some of the differences, as these stations, by selection, have smaller observational errors. The predicted innovations are too small for humidity, which may be explained by the absence of a term for humidity in the model-error parametrization.

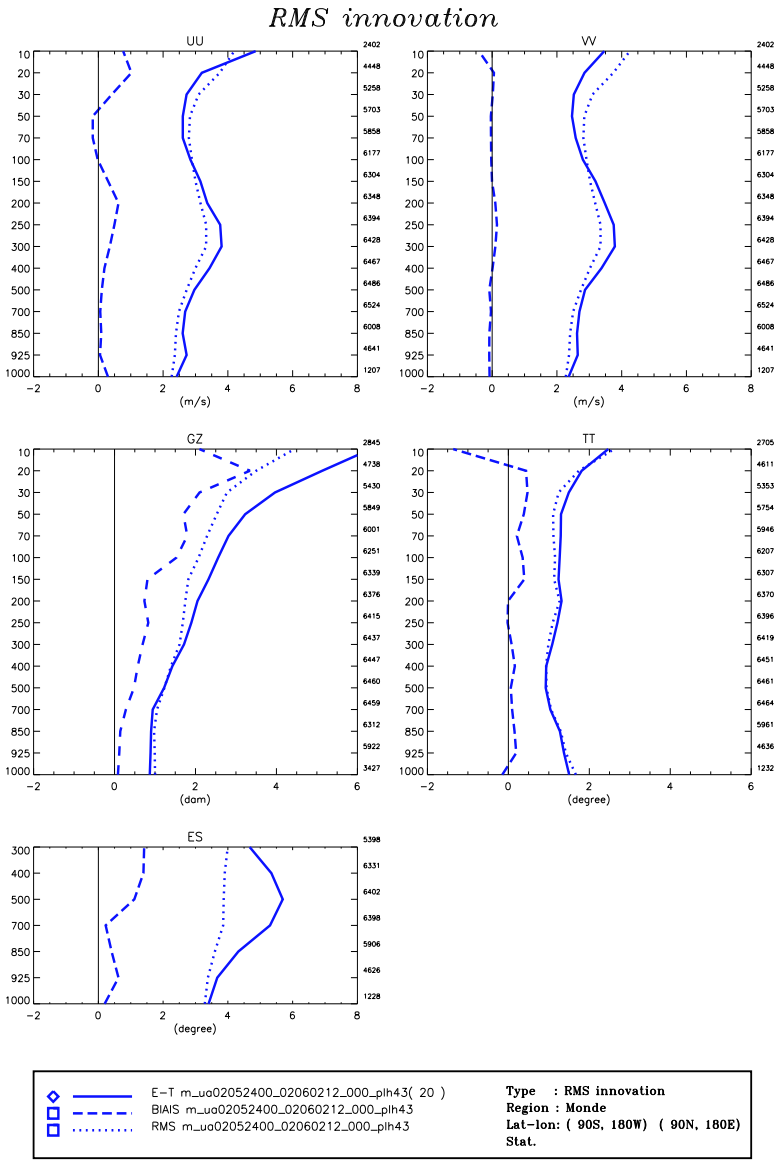


*Figure 5: Comparison of error amplitudes that have been averaged over a 10-day experimental period. The predicted innovation std dev (dotted) should match the observed innovation std dev (solid). The innovation bias is shown as a dashed line.*

## 6 Error dynamics

The uncertainty in a data-assimilation cycle is generally expected to grow during the 6-h forecast as a consequence of the chaotic dynamics of the atmosphere. Errors will also grow as a result of imperfections in the forecast model, which inevitably cannot simulate atmospheric dynamics and physics with zero error. Later, uncertainty will decrease due to the assimilation of new observations. Eventually, we obtain a balance between error growth and error reduction. In this section, we shall examine to what extent this a priori understanding applies to synoptic scale atmospheric data assimilation with the current EnKF implemented at MSC.

In Fig. 6, we display a snapshot of the error amplitudes after a few days of cycling with the EnKF. Generally, we observe large amplitudes over relatively data-sparse areas like the Atlantic, Pacific and Arctic Oceans and smaller amplitudes over well-observed areas like North America, central Europe and China. In addition, over in particular the Atlantic, we observe some intriguing details that may well have been shaped by unstable dynamics.

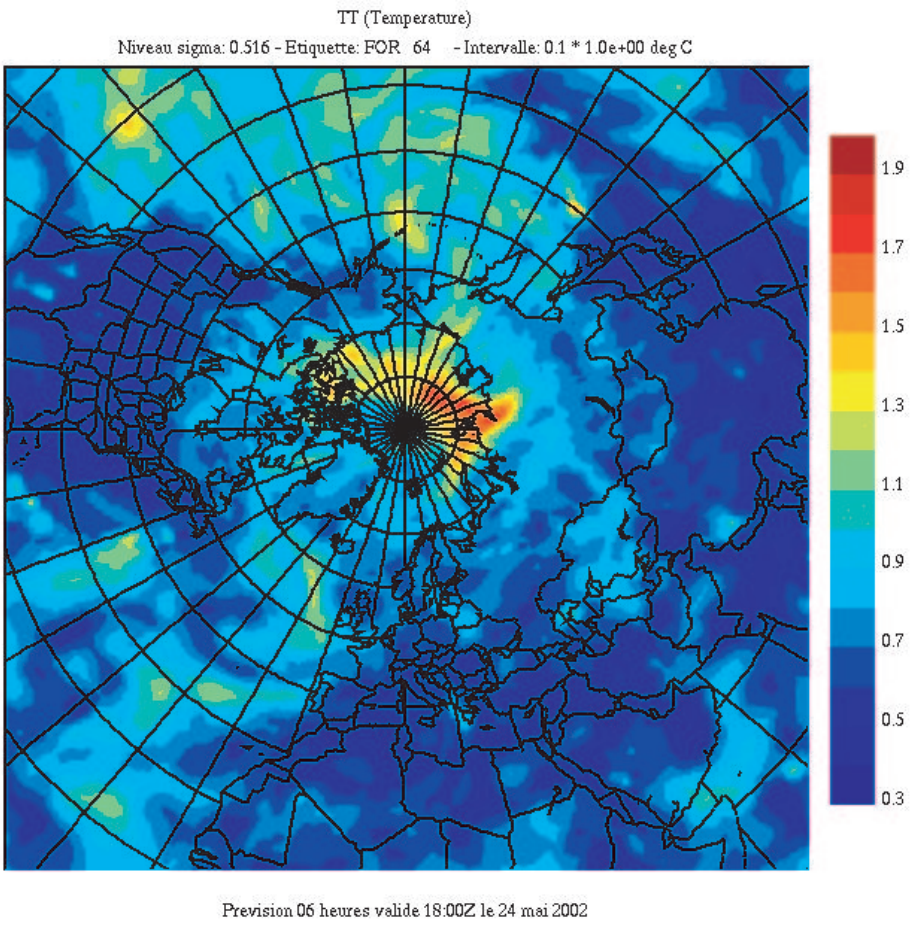


*Figure 6: The rms ensemble spread for temperature at level $\eta = 0.516$ is shown for an ensemble of guess fields, valid at 1800 UT 24 May 2002.*

To measure the growth of error structures, one may use an energy norm (Ehrendorfer and Errico, 1995; MHP):

$$E = \frac{1}{2S}\int_S\int_0^1 [u^2 + v^2 + \frac{c_p}{T_r}T^2 + R_a T_r(\frac{p_s}{p_r})^2]d\eta dS,$$

where $c_p$ is the specific heat at constant pressure, $R_a$ is the gas constant of dry air, $T_r$ is a reference temperature and $p_r$ is a reference pressure. The error norm is global and integrates over the depth of the atmosphere. We shall look separately at results for winds, temperature and surface pressure.

In Fig. 7, we observe a spin-up period of about 4 days after which there is a fairly stable recurring pattern. For wind and temperature, errors levels decay for 6 h with the dynamics of the model. The subsequent instantaneous increase of error levels is due to the addition of parameterized model error. Error levels decrease, as expected, due to the assimilation of observations. For surface pressure (Fig. 7b), the error increase due to model error and the error reduction due to data assimilation are of almost the same amplitude. The growth due to the dynamics of the model is relatively insignificant. For humidity, the qualitative behavior is similar for all mid-tropospheric levels. As is shown here for humidity at level $\eta = 0.631$, the error levels remain remarkably constant over the data-assimilation cycle. This is explained by the absence of a specific model-error component for humidity and by the paucity of humidity related observations in our data-assimilation system.

The remarkable feature of Fig. 7 is the absence of rapid error growth of unstable perturbations due to the dynamics of the model in our implementation of the EnKF. This contrasts with the classical picture of the analysis

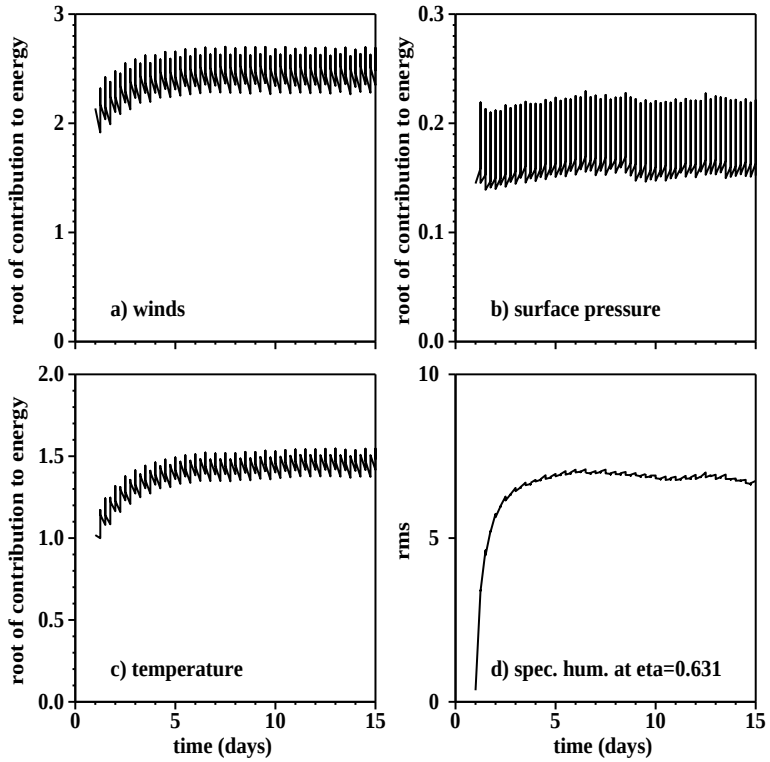


*Figure 7: The rms ensemble spread for the analysis, prediction and forecast, every 6 h for the period between 0000 UTC 19 May 2002 and 0000 UTC 2 June 2002. The results for winds (a), surface pressure (b), and temperature (c) are expressed in terms of a total energy norm. For specific humidity at level* $\eta = 0.631$ *(d), the global rms value is shown in units of* $10^{-4}$ *kg/kg.*

cycle being a "breeding ground" for fast growing modes (Toth and Kalnay 1993). Some further examination (HEA) has shown that error decays at all levels, but most significantly for the top levels of the model. This would be consistent with the co-existence of diffusive model dynamics near the top of the model and significant model error due to the difficulty of dealing with the top boundary.

At this stage, we do not know if the lack of error growth due to the dynamics, and the corresponding need for a large-amplitude model error, are a consequence of some aspect of our implementation of the EnKF, such as the use of a severe horizontal and vertical localization, or if instead this behavior is more generally present in data-assimilation cycles that have enough observations to control the amplitude of synoptic scale errors at low amplitudes. We remark that the "model error", as obtained here using a tuning procedure, has contributions from all terms that are not directly or not perfectly simulated in the EnKF (HEA). These terms include imperfections such as: i) errors in the interpolation operator; ii) the working hypothesis that all observational errors are independent; iii) imperfections in the forecast model; and, iv) errors due to imperfectly known surface fields. For a better simulation of the model error, we will first need to develop a better understanding of its nature.

## 7 Comparison with 3d-var

At MSC an operational 3d-variational procedure is used for the deterministic global analysis. It is our centre's standard of quality for operational data assimilation. A comparison of the EnKF, which is still under development, with the 3d-var has been performed (HEA). For the purpose of this comparison, the 3d-var and the EnKF have been run with exactly the same forecast model (resolution, physical parameterizations, etc) and

with exactly the same observational network. The same observations are assimilated, using the same statistics for the observational error. The quality control decisions (background check and variational quality control), to be used in both experiments, were taken from a prior 3d-var data assimilation cycle that would flag each observation as either accepted or rejected.

The innovation statistics, with respect to radiosondes, are compared in Fig. 8 for the period 00 UTC 24 May - 12 UTC 2 June 2002 (a 10-day period).

For winds and temperature, the EnKF and the 3d-var have remarkably similar innovation statistics. For humidity, the EnKF has a bigger bias but a smaller standard deviation. Generally the scores are very similar. One interpretation is that the impact of the 4D aspect is small. However, looking in more detail at the results, we found (HEA) that the 3d-var generally draws closer to the observations than the EnKF. This is consistent with generally bigger covariance estimates for the forecast error in the 3d-var. However, it would also be consistent with an inability of the EnKF to closely fit the large amount of data.

To investigate to what extent the analysis quality is limited by the number of directions available in the ensemble, we compared the performance with twice 48 and with twice 64 members. The results in Fig. 9 do not show a big improvement for the innovation statistics in the bigger ensemble. This would seem to suggest that the results, *with fixed localization*, have converged for ensemble size. We hope to investigate this further when significantly more computational resources become available.

## 8 Summary

We have shown that the EnKF is a computationally feasible approximation to the Kalman filter. The approximation is mainly in the evaluation of error covariances with a fairly small ensemble. The estimation of covariances from a small ensemble, instead of using a full covariance matrix, makes it necessary to perform both a horizontal and a vertical localization in computing the Kalman gain.

To represent model error, we use an isotropic parameterization that is based on the parameterization of the forecast error that is used in our centre's 3d-variational algorithm. Adjusting some of the parameters, we obtain innovation amplitudes that correspond fairly well to the ensemble-based prediction.

Looking in more detail at the error dynamics, we observe an approximate balance between error growth due to "model error" and error reduction due to the assimilation of new observations. The growth of unstable perturbations during the assimilation cycle does not dominate the error dynamics. In fact, error amplitudes generally decay during the 6-h integration of the ensemble of states.

With all the above components, we are able to obtain ensemble mean scores, as verified against radiosondes, that are surprisingly similar to the scores obtained with a 3d-var algorithm.

We feel that many of the decisions taken so far in this project merit further justification. Consequently we intend to focus future research on: the formulation of the model error; the dynamics of errors in the data-assimilation cycle; and the choice between either higher resolution or larger ensemble size with less severe localization.

## Acknowledgements

We would like to thank our co-workers Gérard Pellerin, Martin Charron, Lubos Spacek and Bjarne Hansen for the collaboration towards an operational ensemble Kalman filter. Because of the complexity of implementing a new data-assimilation methodology, we are grateful for the productive interactions with the people working on the forecast model and on the data-assimilation systems at our centre. This concerns both the exchange of experience, the brainstorming of new ideas and the transfer and adaptation of software. We acknowledge help from Mark Buehner, Bernard Dugas, Luc Fillion, Jacques Hallé, Pierre Koclas, Stéphane Laroche, Richard

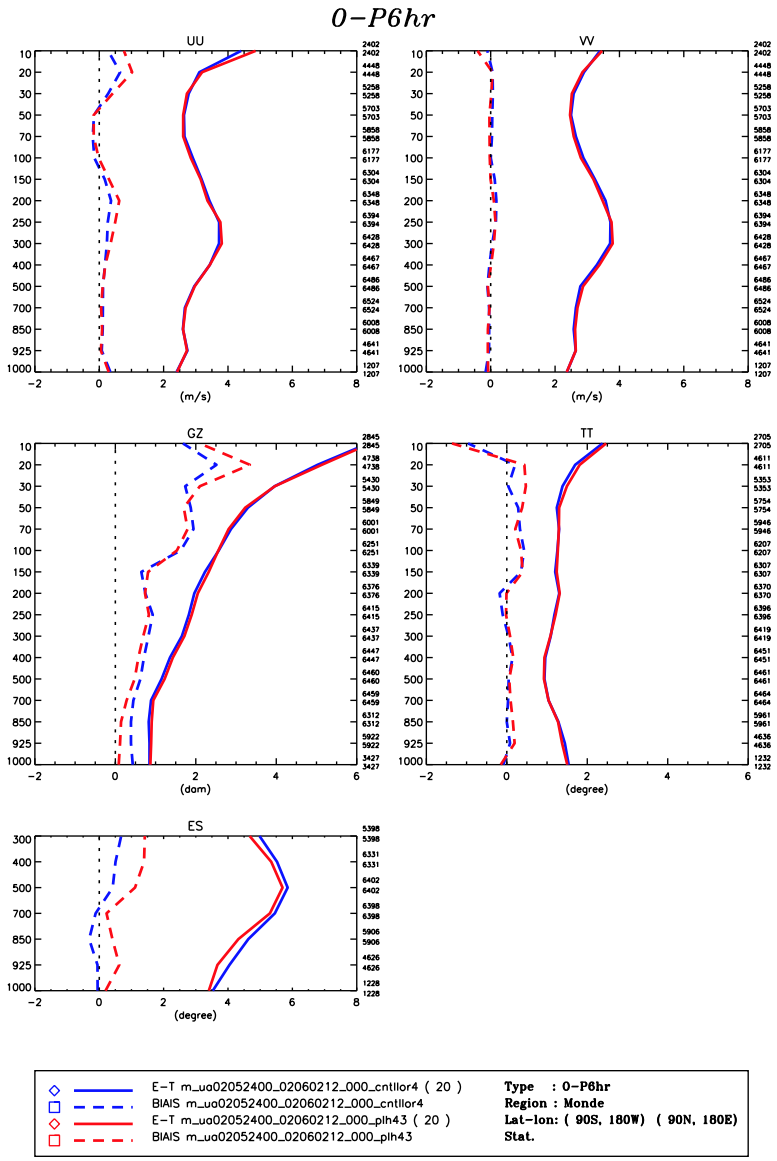


*Figure 8: Comparison of bias (dashed) and standard deviation (solid) for the EnKF (red) and the 3d-var (blue). The number of ensemble members for the EnKF is twice 64 members.*

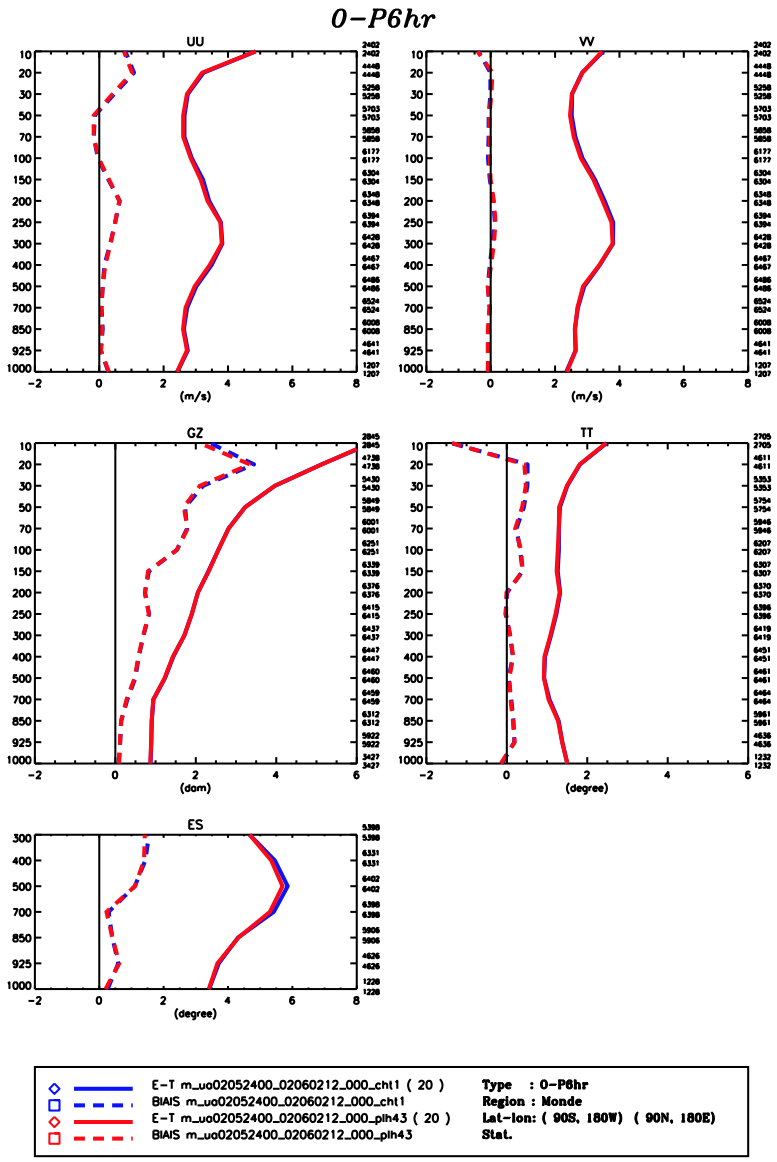


*Figure 9: Comparison of bias (dashed) and standard deviation (solid) for an experiment with twice 48 (blue) and twice 64 (red) members.*

Ménard, Josée Morneau, Alain Patoine, Michel Roch, Michel Valin, and Gilles Verner.

## REFERENCES

Anderson, B.D.O., and J.B. Moore, 1979: *Optimal Filtering*. Prentice-Hall, 357 pp.

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.

Côté, J., S. Gravel, A. Méthot, A. Patoine, M. Roch, and A. Staniforth, 1998: The operational CMC-MRB Global Environmental Multiscale (GEM) model. Part I: Design considerations and formulation. *Mon. Wea. Rev.*, **126**, 1373–1395.

Dee, D.P., 1991: Simplification of the Kalman filter for meteorological data assimilation. *Quart. J. Roy. Meteor. Soc.*, **117**, 365–384.

Dee, D.P., 1995: On-line estimation of error covariance parameters for atmospheric data assimilation. *Mon. Wea. Rev.*, **123**, 1128–1145.

Dee, D.P., L. Rukhovets, R. Todling, A.M. da Silva, and J.W. Larson, 2001: An adaptive buddy check for observational quality control. *Quart. J. Roy. Meteor. Soc.*, **127**, 2451–2471.

Ehrendorfer, M., and R.M. Errico, 1995: Mesoscale predictability and the spectrum of optimal perturbations. *J. Atmos. Sci.*, **52**, 3475–3500.

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), 10143–10162.

Evensen, G., 2003: The ensemble Kalman filter: theoretical formulation and practical implementation. *Ocean Dynamics*. In press.

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

Gauthier, P., P. Courtier, and P. Moll, 1993: Assimilation of simulated wind Lidar data with a Kalman filter. *Mon. Wea. Rev.*, **121**, 1803–1820.

Gauthier, P., M. Buehner, and L. Fillion, 1999a: Background-error statistics modelling in a 3D variational data assimilation scheme: Estimation and impact on the analyses. Proceedings, ECMWF Workshop on diagnosis of data assimilation systems. ECMWF, 131–145. [Available from ECMWF, Shinfield Park, Reading, Berkshire RG2 9AX, United Kingdom.]

Gauthier, P., C. Charette, L. Fillion, P. Koclas, and S. Laroche, 1999b: Implementation of a 3d variational data assimilation system at the Canadian Meteorological Centre. Part I: the global analysis. *Atmos.-Ocean*, **37**, 103–156.

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., and H.L. Mitchell, 2001: A sequential ensemble Kalman filter for atmospheric data assimilation. *Mon. Wea. Rev.*, **129**, 123–137.

Houtekamer, P.L., H.L. Mitchell, G. Pellerin, M. Buehner, M. Charron, L. Spacek and B. Hansen, 2003: Atmospheric data assimilation with the ensemble Kalman filter: Results with real observations. *Mon. Wea. Rev.* Submitted.

Keppenne, C.L., and M.M. Rienecker, 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., 1986: Analysis methods for numerical weather prediction. *Quart. J. Roy. Meteor. Soc.*, **112**, 1177–1194.

Maybeck, P.S., 1979: *Stochastic Models, Estimation and Control.* Vol. 1. Academic Press, 423 pp.

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, balance, and model-error representation in an ensemble Kalman filter. *Mon. Wea. Rev.*, **130**, 2791–2808.

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**, 1485–1490.

Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc.*, **74**, 2317–2330.

Whitaker, J.S., G.P. Compo, X. Wei, and T.M. Hamill, 2003: Reanalysis without radiosondes using ensemble data assimilation. *Mon. Wea. Rev.* Submitted.