Practical ensemble data assimilation

Peter Houtekamer, Herschel Mitchell
Meteorological Service of Canada

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

It is a close approximation to the Kalman filter. The Kalman filter provides the least-squares solution in the case of small errors with Gaussian distribution. The approximation becomes "automatically" more accurate as bigger ensembles are used.

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 an adjoint or tangent linear model. It is therefore an extremely attractive method.

Overview

- Introduction
- Kalman filter equations
- Setting up an experimental environment
- Localization of covariances
- Model error simulation
- Error dynamics
- Comparison with 3d-var
- Conclusions

Results will be mainly from the Canadian EnKF developed by:

Peter Houtekamer, Herschel Mitchell, Gérard Pellerin, Bjarne Hansen, Lubos Spacek, Chantal Côté and Martin Charron.

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, Stochastic Models, Estimation and Control).

$$
\begin{aligned}
\Psi^a &= \Psi^f + \mathsf{K}(o - \mathsf{H}\Psi^f) \\
\mathsf{K} &= \mathsf{P}^f\mathsf{H}^T(\mathsf{H}\mathsf{P}^f\mathsf{H}^T + \mathsf{R})^{-1} \\
\mathsf{P}^a &= (\mathsf{I} - \mathsf{K}\mathsf{H})\mathsf{P}^f \\
\mathsf{P}^f(t+6) &= \mathsf{M}\mathsf{P}^a\mathsf{M}^T + \mathsf{Q}
\end{aligned}
$$

o	vector with observations,	Ψ^f	first guess field,
Ψ^a	analysis,	P^f	Covariance of the forecast error,
P^a	Covariance of the analysis error,	H	forward interpolation matrix,
R	Covariance of the observational error,	K	Gain-matrix,
M	tangent linear operator,	Q	Covariance of the model error.

Sequential Assimilation

Example:

$$
\begin{aligned}
f &: N(0,1), \quad o1 : N(0,1), \quad o2 : N(0,1) \\
a &= k_f f + k_1 o1 + k_2 o2 \\
k_f &= \frac{\sigma_f^{-2}}{\sigma_f^{-2} + \sigma_{o1}^{-2} + \sigma_{o2}^{-2}}, \qquad k_f = k_1 = k_2 = \frac{1}{3}
\end{aligned}
$$

Step 1 of the sequential assimilation:

$$
\begin{aligned}
f1 &: N(0,1), \quad o1 : N(0,1) \\
f2 &= k_{f1} f1 + k_1 o1, \qquad k_{f1} = k_1 = \frac{1}{2} \\
f2 &: N(0, \frac{1}{2})
\end{aligned}
$$

Step 2 of the sequential assimilation:

$$
\begin{aligned}
f2 &: N(0, \frac{1}{2}), \quad o2 : N(0,1) \\
a &= k_{f2} f2 + k_2 o2, \qquad k_{f2} = \frac{2}{2+1}, \qquad k_2 = \frac{1}{3} \\
a &= \frac{1}{3}(f1 + o1 + o2)
\end{aligned}
$$

Differences with the Kalman filter: evaluation of matrices

$$
\begin{aligned}
\mathsf{P}^f &= \frac{1}{N-1}\sum_{i=1}^{N}(\Psi_i^f - \overline{\Psi_i^f})(\Psi_i^f - \overline{\Psi_i^f})^T \\
\mathsf{P}^f\mathsf{H}^T &= \frac{1}{N-1}\sum_{i=1}^{N}(\Psi_i^f - \overline{\Psi_i^f})(\mathsf{H}\Psi_i^f - \overline{\mathsf{H}\Psi_i^f})^T \\
\mathsf{H}\mathsf{P}^f\mathsf{H}^T &\equiv \frac{1}{N-1}\sum_{i=1}^{N}(\mathsf{H}(\Psi_i^f) - \overline{\mathsf{H}(\Psi_i^f)})(\mathsf{H}(\Psi_i^f) - \overline{\mathsf{H}(\Psi_i^f)})^T
\end{aligned}
$$

A random ensemble is used to estimate error covariances.

The dimension of the full matrix, used in the Kalman filter, will be the size of the phase space (order 1 000 000).

The dimension of the ensemble-based matrix is identical to $N-1$ (order 100). The analysis increment will be in the space of dimension $N-1$. The of order 100 000 observations will thus be projected to a very low-dimensional space.

One may use covariance localization to deal with the dimensionality problem.

Differences with the Kalman filter: transport equation

In the Kalman filter, the covariances are transported using:

$$\mathrm{P}^f(t+6) = \mathrm{M}\mathrm{P}^a\mathrm{M}^T + \mathrm{Q}$$

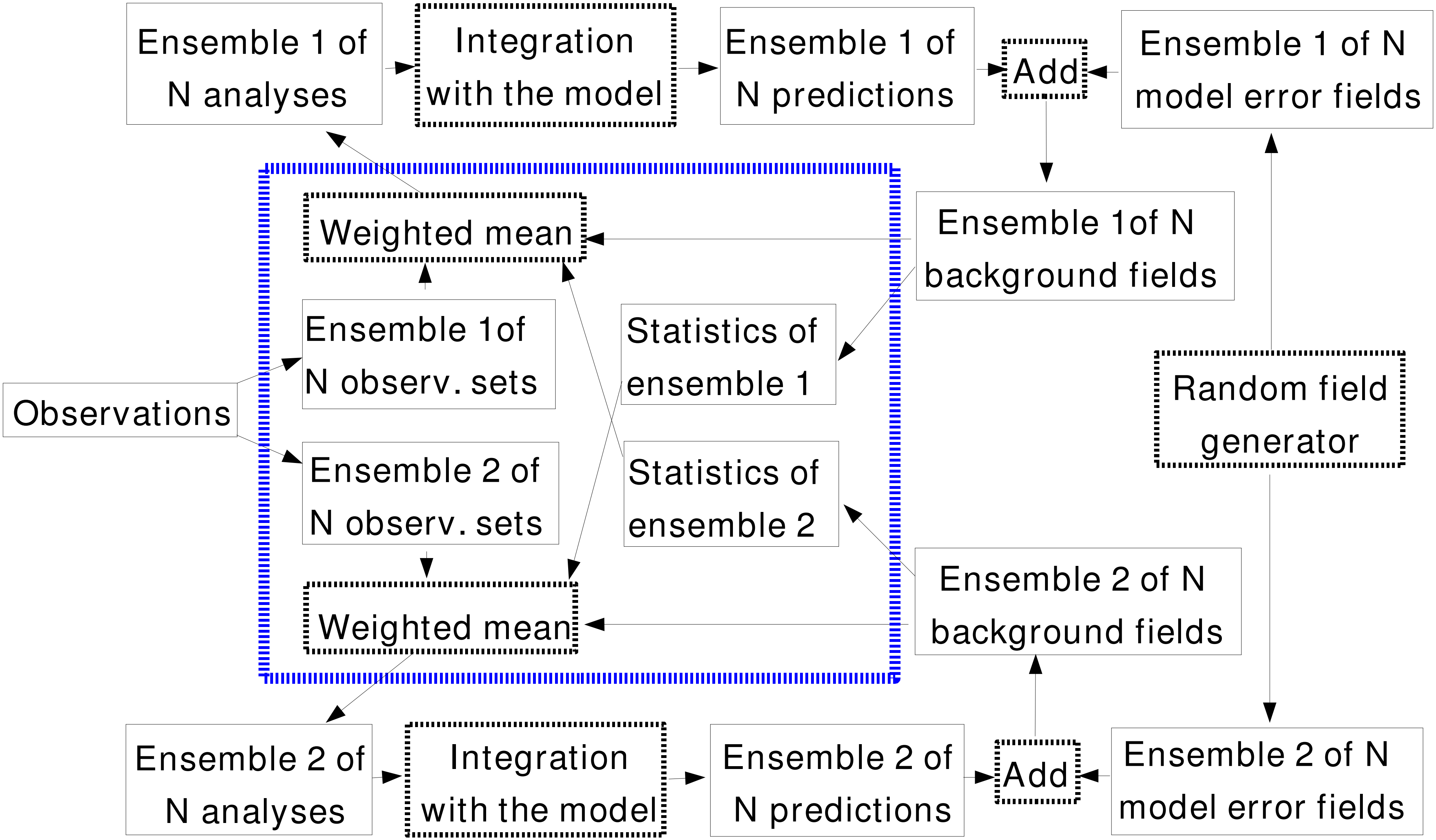
In the EnKF, one uses the full model to integrate an ensemble of analyses:

$$\Psi_i^f(t+6) = M(\Psi_i^a) + q_i$$

The replacement of the tangent linear model by a model with full physics would seem to be an improvement. The full model will properly deal with saturation of errors. One also has additional flexibility to deal with model error. It can be sampled from a covariance matrix or also using for instance different realizations of the model error.

With a sample one may in principle transport information on higher moments that is lost when covariance matrices are used as in the Kalman filter.

Double Ensemble Kalman Filter



Setting up the experimental environment

The operational implementation of a new data-assimilation methodology is a complex project. It is necessary to interact with the people working on the forecast model and on the data-assimilation system. This concerns both the exchange of experience, the brainstorming of new ideas and the transfer and adaptation of software.

We acknowledge help from:

- Luc Fillion (early development of the algorithm),
- Richard Ménard (localization operator),
- Gilles Verner and Pierre Koclas (treatment of observations),
- Josée Morneau (validation software),
- Stéphane Laroche and Jacques Hallé (interpolation operators),
- Mark Buehner (model error description),
- Michel Roch (use of the forecast model),
- Our computer support group (an excellent computational environment).

Experimental environment: the model

We use the Global Environmental Model (GEM) of MSC. The model is much like the version used to produce the higher resolution medium-range deterministic forecast at our centre.

The poles of the model are located at the geographical poles.

Grid: 240×120.
vertical resolution: 28 levels.

top of the model: 10 hPa.

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

timestep: 60 minutes.

model variables: u, v, T, q, and $p_{surface}$.

ensemble size: 2×48 members and 2×64 members.

Experimental environment: observations

We try to use all observations that are assimilated by the deterministic analysis (3d-var) at our centre. 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:

- radiosondes: $u, v, T, q, p_{surface}$
- aircraft: u, v, T
- satellite: u, v,TOVS 1b radiances
- surface observations: $T, p_{surface}$

We do NOT YET assimilate surface observations of wind and humidity.

Localization

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

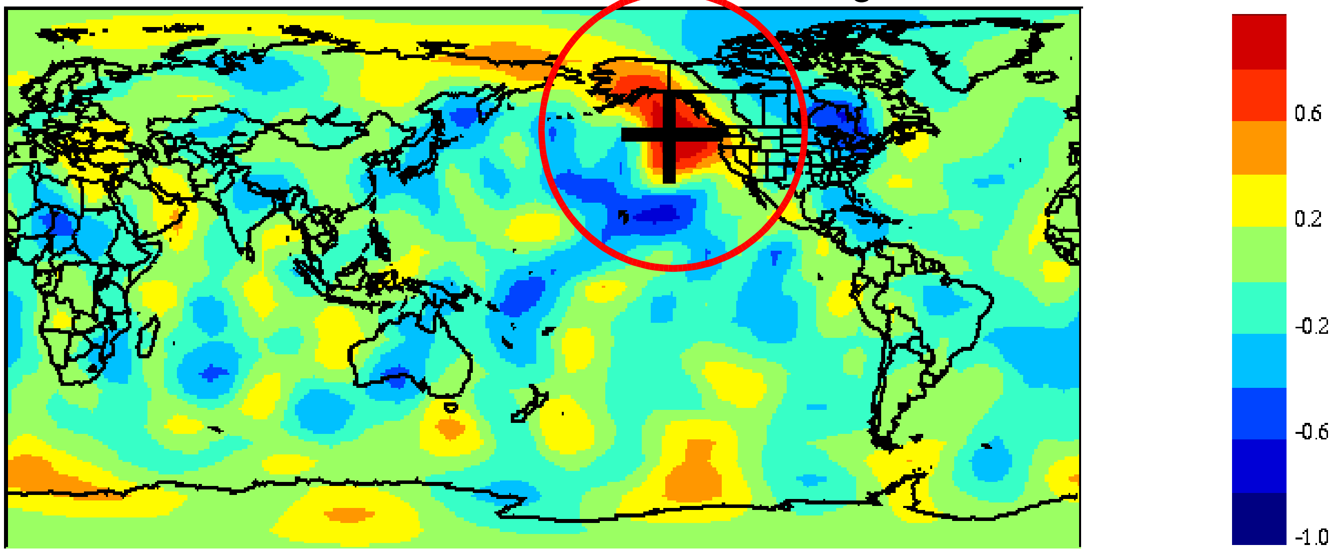
Artificial measures, not suggested by the Kalman filter equations, will be necessary to inflate the dimensionality of the ensemble.

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 producing nicely balanced analyses that remain far from the O(100 000) observations and between having fairly unbalanced analyses that fit all observations pretty well.

Need for localization

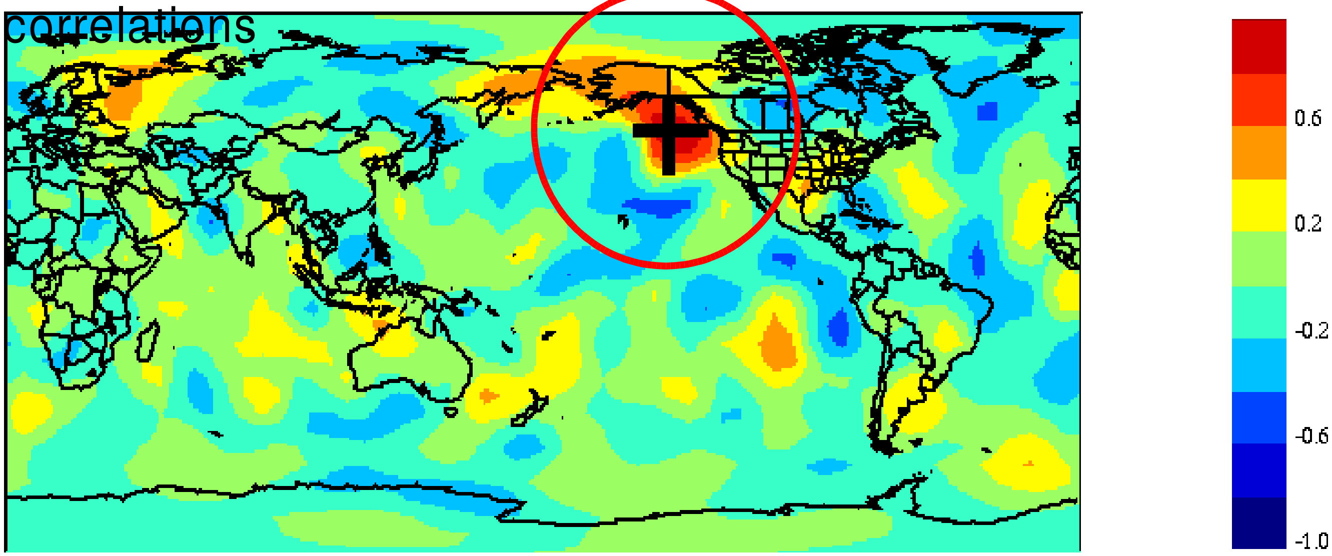
32 members, ensemble 1, 500 hPa height correlations



Correlations are shown with respect to a central point.

The correlations are seen to mostly agree inside the red circle.

32 members, ensemble 2, 500 hPa height correlations

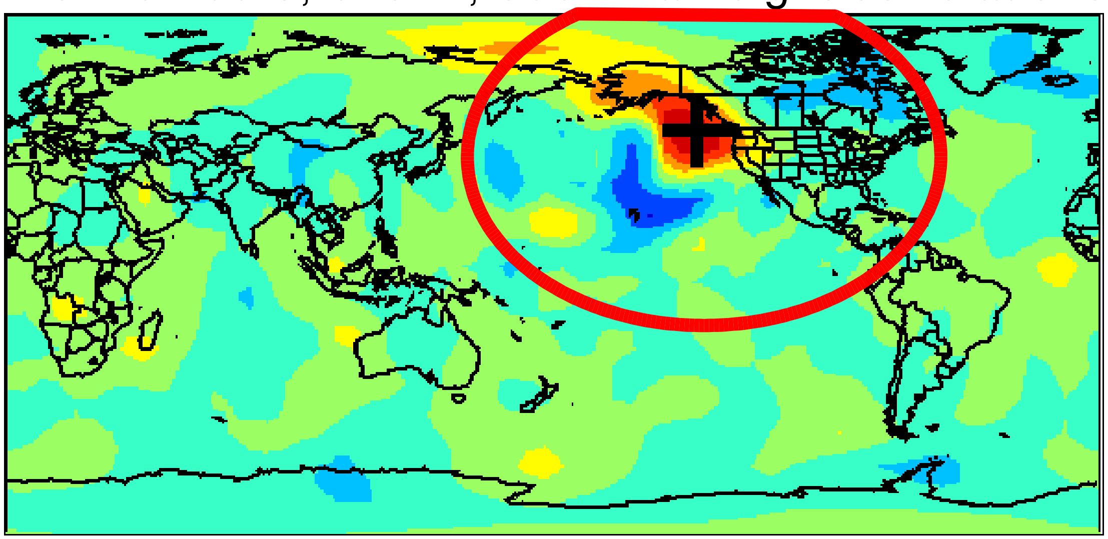


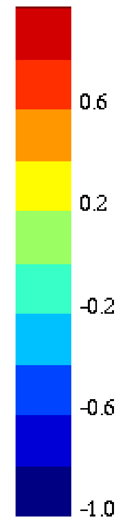
Distant correlation estimates do not correspond between the two ensembles.

Results from Houtekamer and Mitchell, 1998, Mon. Wea. Rev.

Impact of using a bigger ensemble

128 members, ens. 1, 500 hPa height correlations

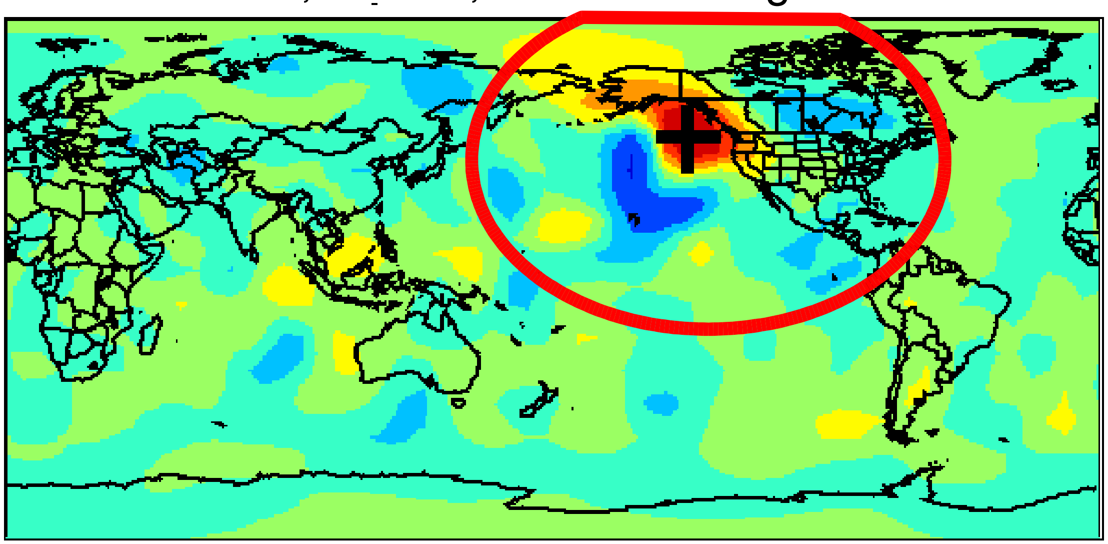


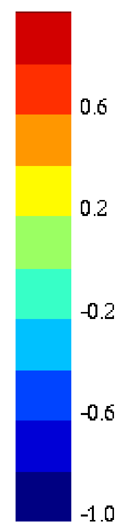


Using more members we have agreement out to a bigger distance.

The correlation estimates at large distance still disagree but they are smaller by about a factor of 2.

128 members, ens. 2, 500 hPa height correlations





The estimation error reduces as the root of the ensemble size.

For big ensembles we do not need to localize covariances.

Results from HM

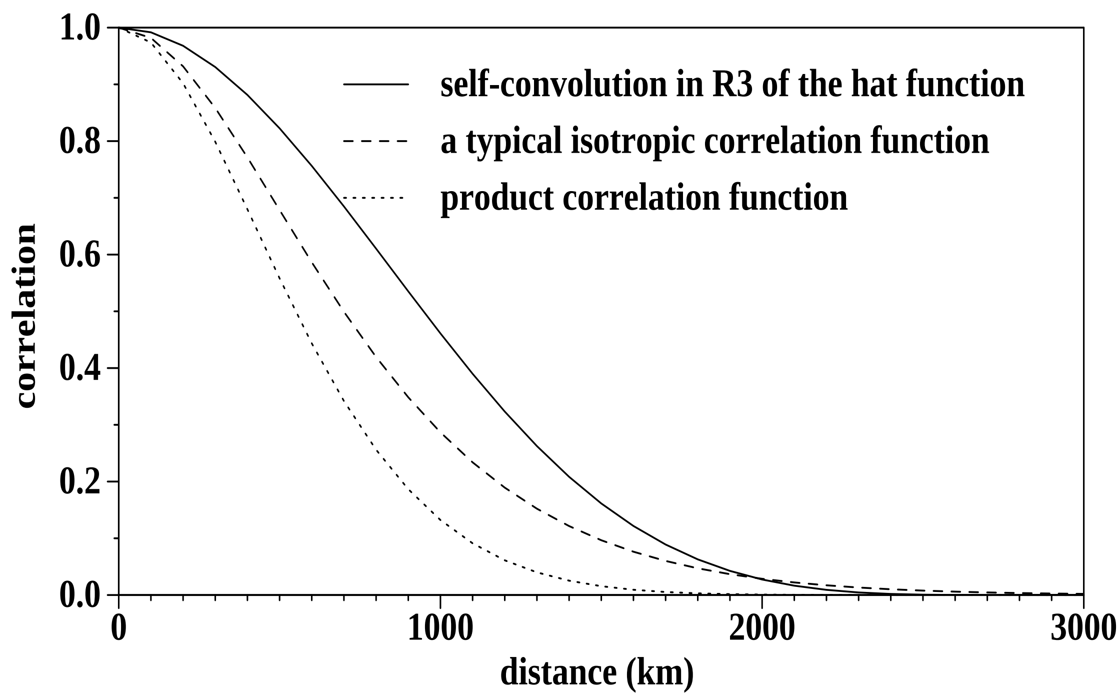
Horizontal Hadamard product

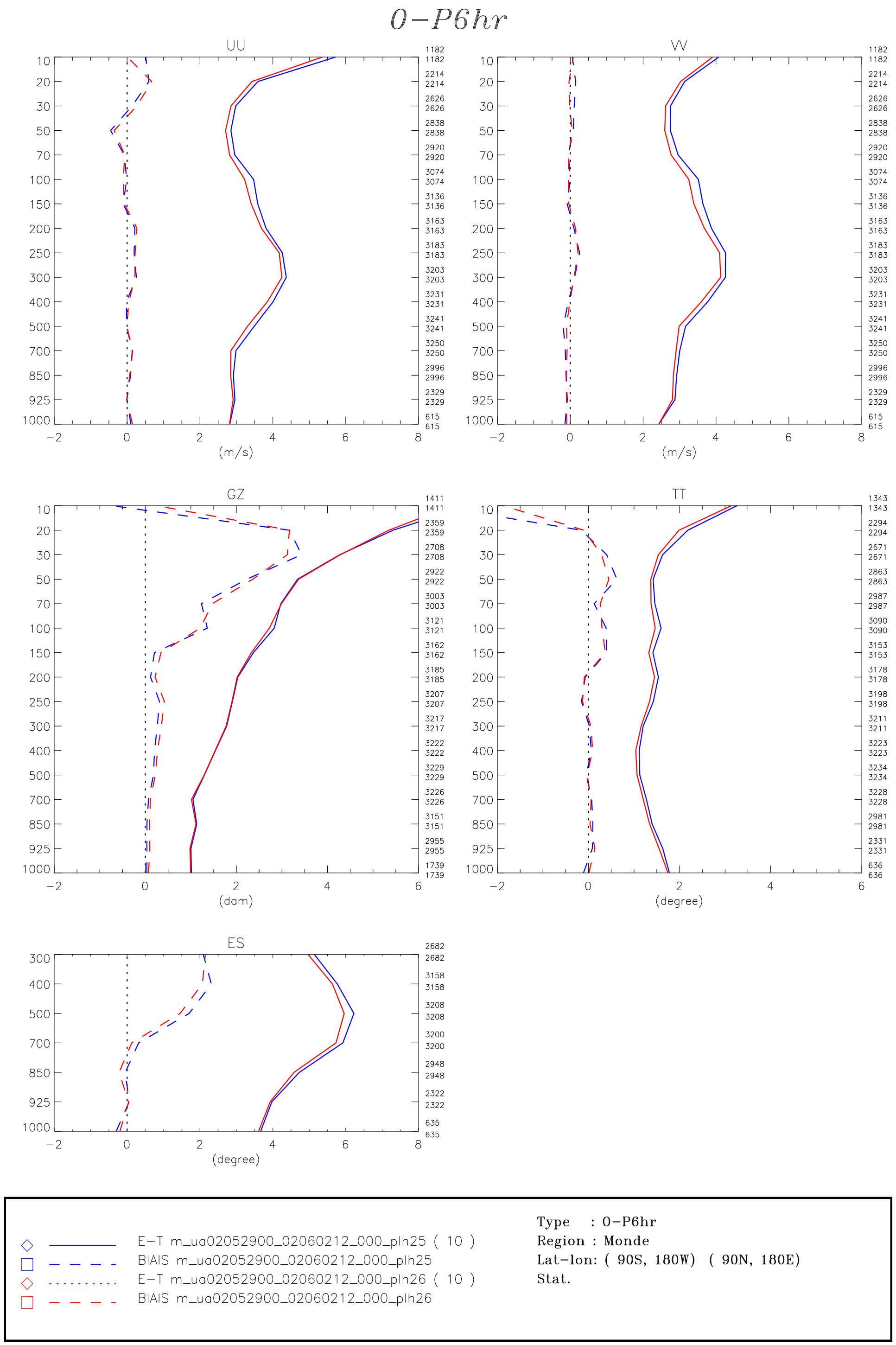
Idea (obtained from Richard Ménard) (Gaspari & Cohn, 1999, QJRMS, 723-757):

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

To filter covariances at long distances we use a Hadamard product (which does a point-wise product of two matrices). This leads to a positive definite matrix P^f.

As we get bigger ensembles, we can make L longer.





Impact of the horizontal Hadamard product

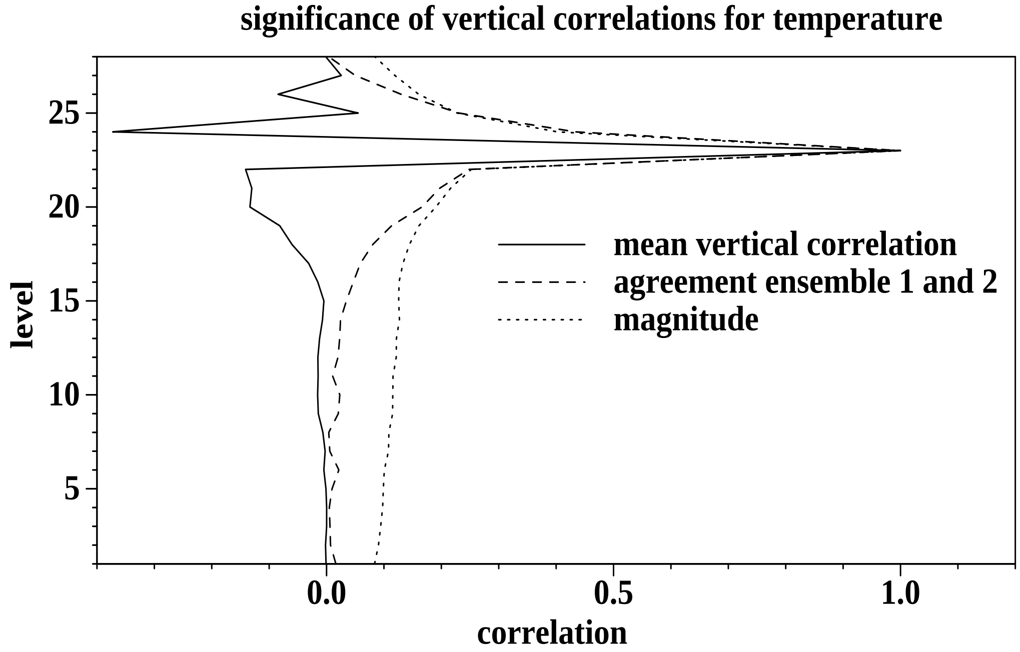
In the reference experiment correlations are forced to zero at a distance of 3400 km.

Smaller errors are obtained when forcing the errors to zero at a distance of 2300 km.

Based on these results one would use a strong localization with enforced zero-impact at 2300 km. Alternatively one could consider using more than 2×48 members.

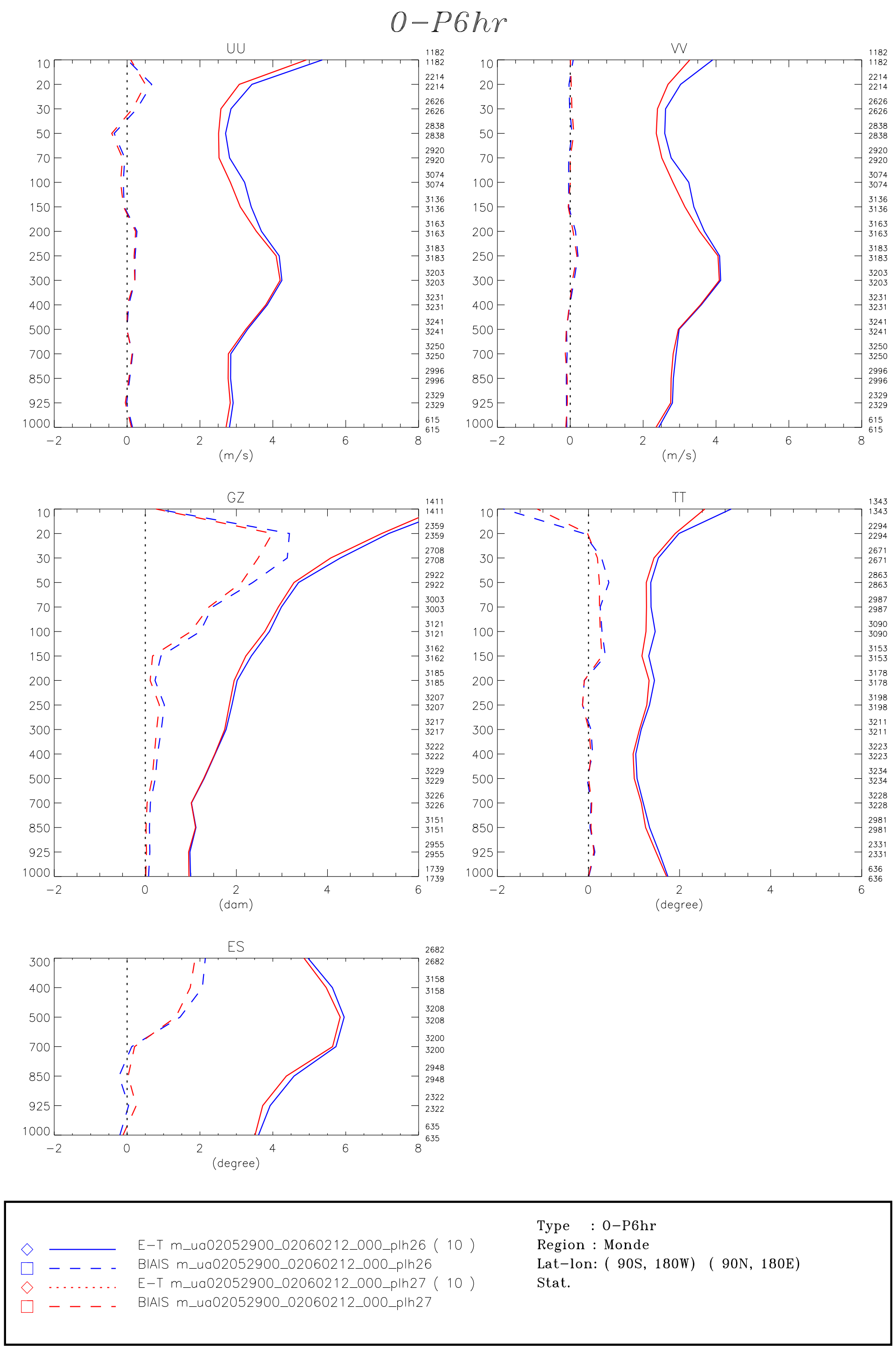
We now perform most experiments with localization at 2800 km (quality is close to that using 2300 km).

Example of narrow vertical correlations



For a pair of 48 member ensembles the vertical correlations were computed with respect to level 23 (at about 100 hPa). The global mean correlations are negative at levels 22 and 24! Above level 26 and below level 19 the two 48 member ensembles tend to disagree. Estimated vertical correlations are at the noise level and should therefore be filtered. This is done with a Hadamard product in the vertical.

Paper by Houtekamer et al. to be submitted to Mon. Wea. Rev.

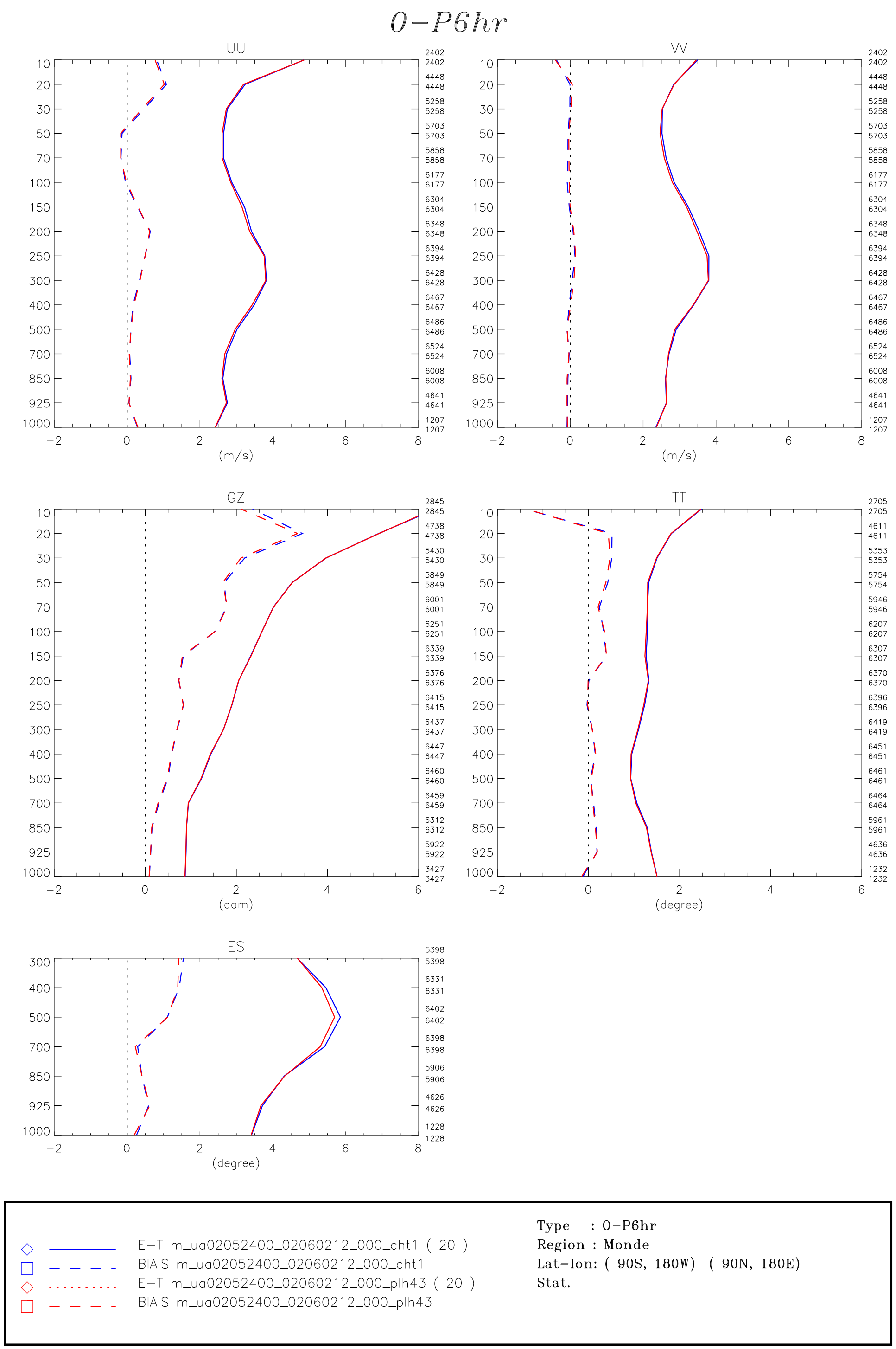


Impact of the vertical Hadamard product

The reference experiment has no localization in the vertical

It is compared with an experiment where correlations are forced to zero in 2 units of ln (pressure).

The impact is very positive for in particular the upper levels where vertical correlations were seen to be very narrow.



Impact of ensemble size

The reference experiment has 2 $\times$ 48 members.

It is compared with an experiment using 2 $\times$ 64 members.

Results are slightly better with 2 $\times$ 64 members. It would appear that the EnKF has converged for ensemble size. Nevertheless we would like to perform some experiments with say O(1000) members just to be sure.

Model error simulation

Our current hypothesis is that the model error is similar in structure to the forecast error as described in our centre's 3d-var scheme.

$$
\begin{aligned}
\mathsf{P}^f(t+6) &= \mathsf{M}\mathsf{P}^a\mathsf{M}^T + \mathsf{Q} \\
\mathsf{P}^f(t+6) &= \mathsf{M}\mathsf{P}^a\mathsf{M}^T + 0.25\mathsf{P}_{\text{3dvar}}
\end{aligned}
$$

For each member we obtain a random model error field that has isotropic error statistics as prescribed in the 3d-var (but smaller).

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 on 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.

Using innovations to tune the model error term Q

One may use innovation statistics to tune the model error term:

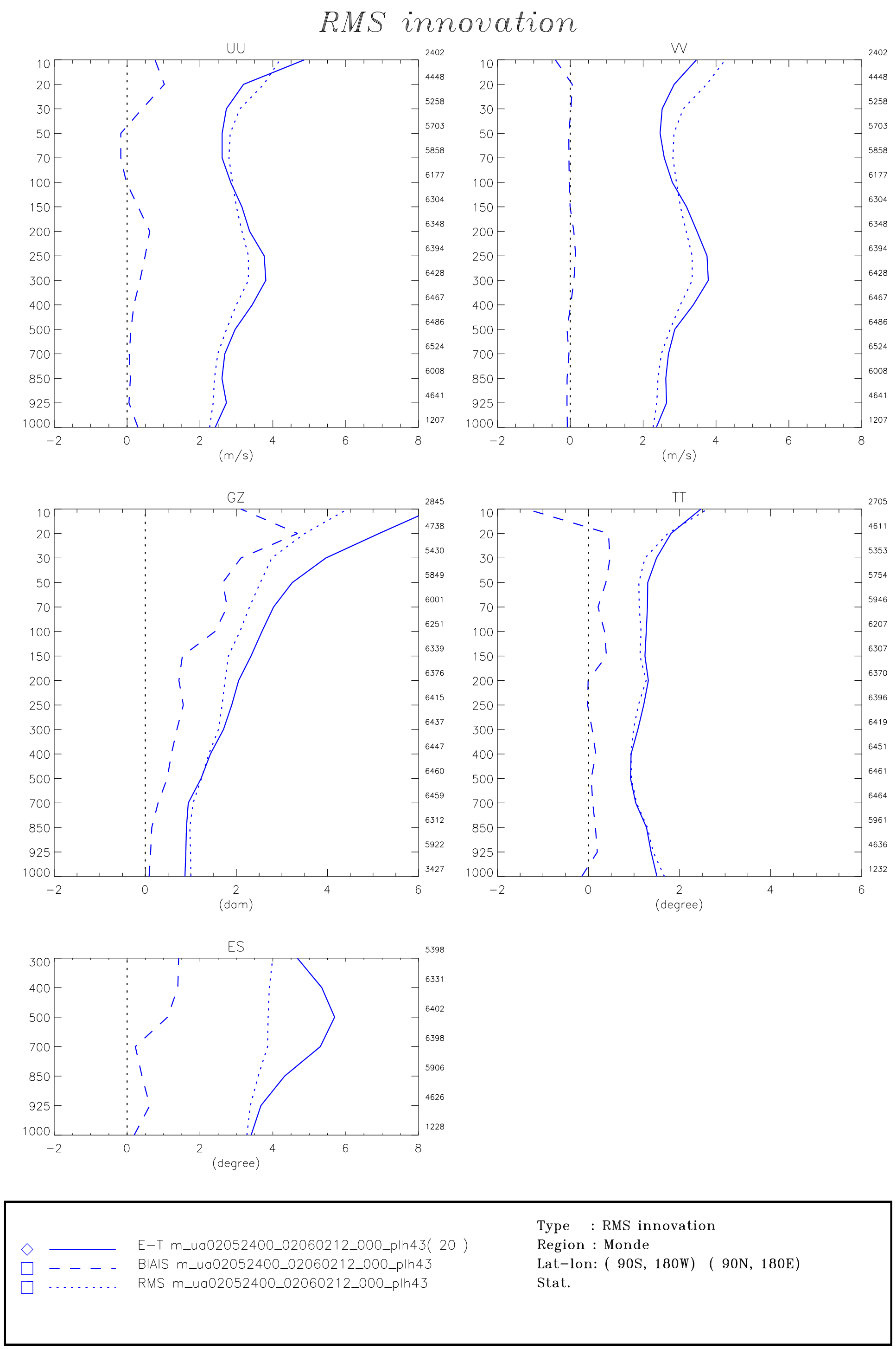
- Dee, 1995, On-line estimation of error covariance parameters for atmospheric data assimilation. Mon. Wea. Rev. 1128-1145.
- Mitchell and Houtekamer, 2000, An adaptive ensemble Kalman filter. Mon. Wea. Rev. 416-433.

Basic equation:

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

The innovation ν is computed as the difference between the interpolated ensemble mean state and the observation. The interpolated forecast error covariance P is available from the ensemble. An estimate of the observation error covariance R is available from the data assimilation algorithm. The remaining term Q can thus be estimated.

In the current study we only consider the diagonal of the basic equation.



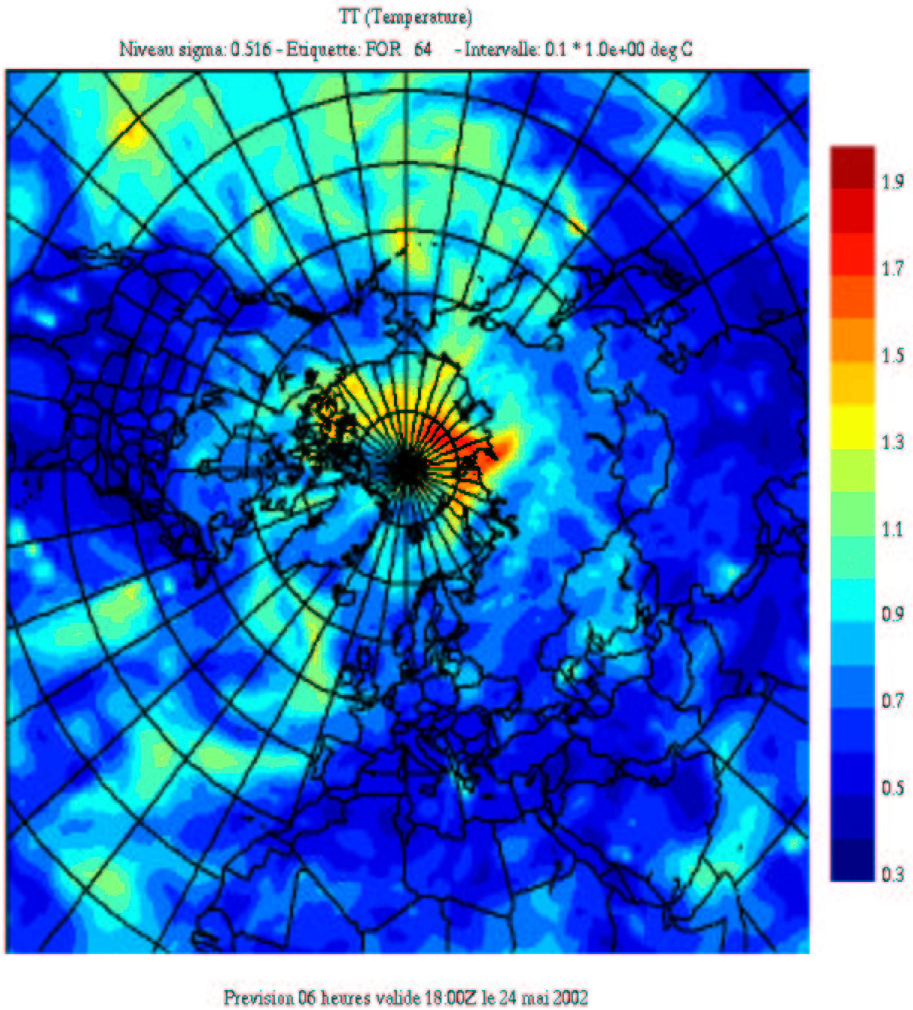
Quality of error statistics

The solid line is the rms amplitude of the innovations for radiosonde observations.

The dotted line is the ensemble based prediction of the innovation amplitude. It is the root of the sum of the observational variance and the ensemble spread.

There is excellent agreement for the temperature. For winds the ensemble spread is too large near the model top. The spread is too small for humidity.

Standard deviation in a background field



For the temperature at level $\eta = 0.516$, we look at the standard deviation of the background field (24 May 2002 18 UT).

The assimilation has managed to reduce the errors over the continents.

One also notes dynamically looking structures over the oceans.

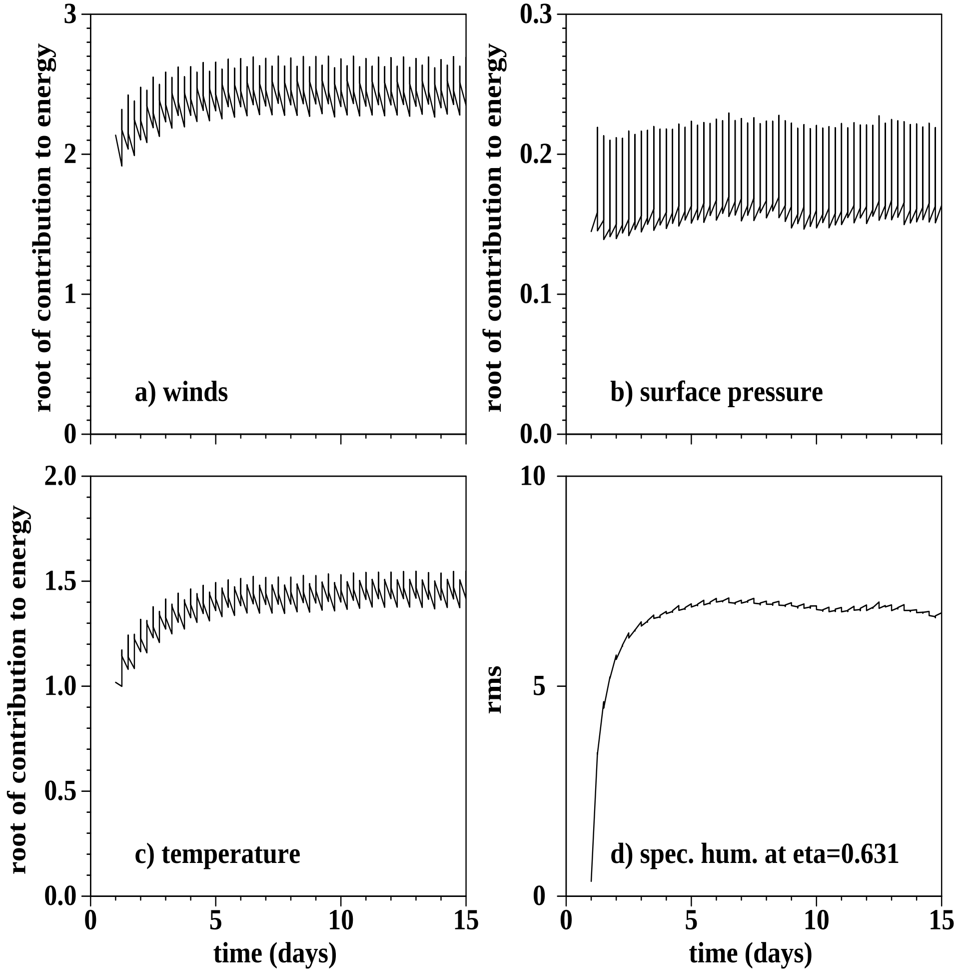
Error dynamics

To measure the growth of error structures one may use an energy norm (Ehrendorfer and Errico, 1995, J. Atmos. Sci., 3475-3500):

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

The error norm is global and integrates over the depth of the atmosphere. We will look separately at results for winds, temperature and surface pressure. For winds and temperature one may also look at the contribution per level.

Error dynamics during the data assimilation cycle



Error amplitudes for winds and temperature decay during the 6 hour integration with the forecast model.

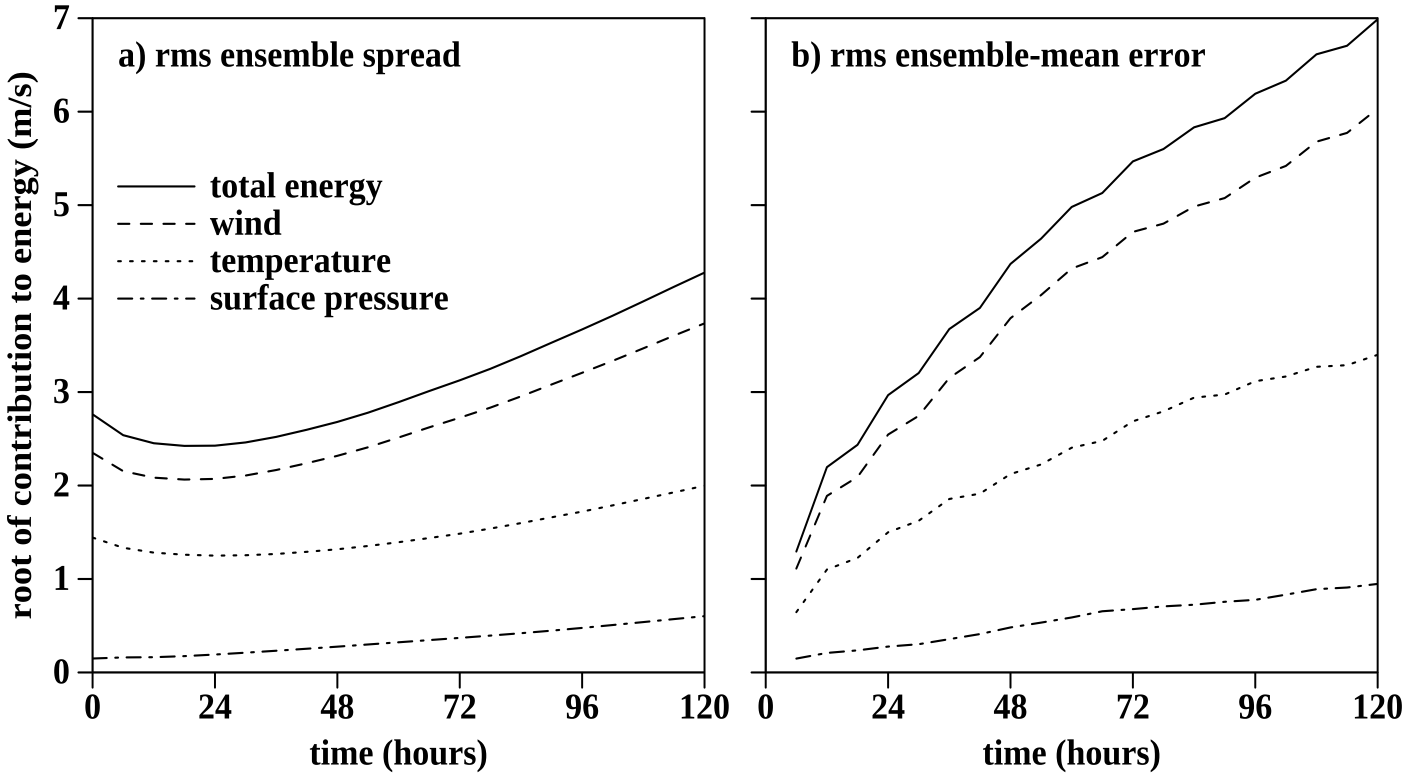
Error amplitudes subsequently increase due to the addition of parameterized model error.

Error amplitudes decrease thanks to the assimilation of new observations.

It would appear that the rapid growth of unstable perturbations is insignificant in the experiments performed here. Instead, analysis errors seem to obtain their amplitude due to “model error”.

Error dynamics during a 120-h forecast

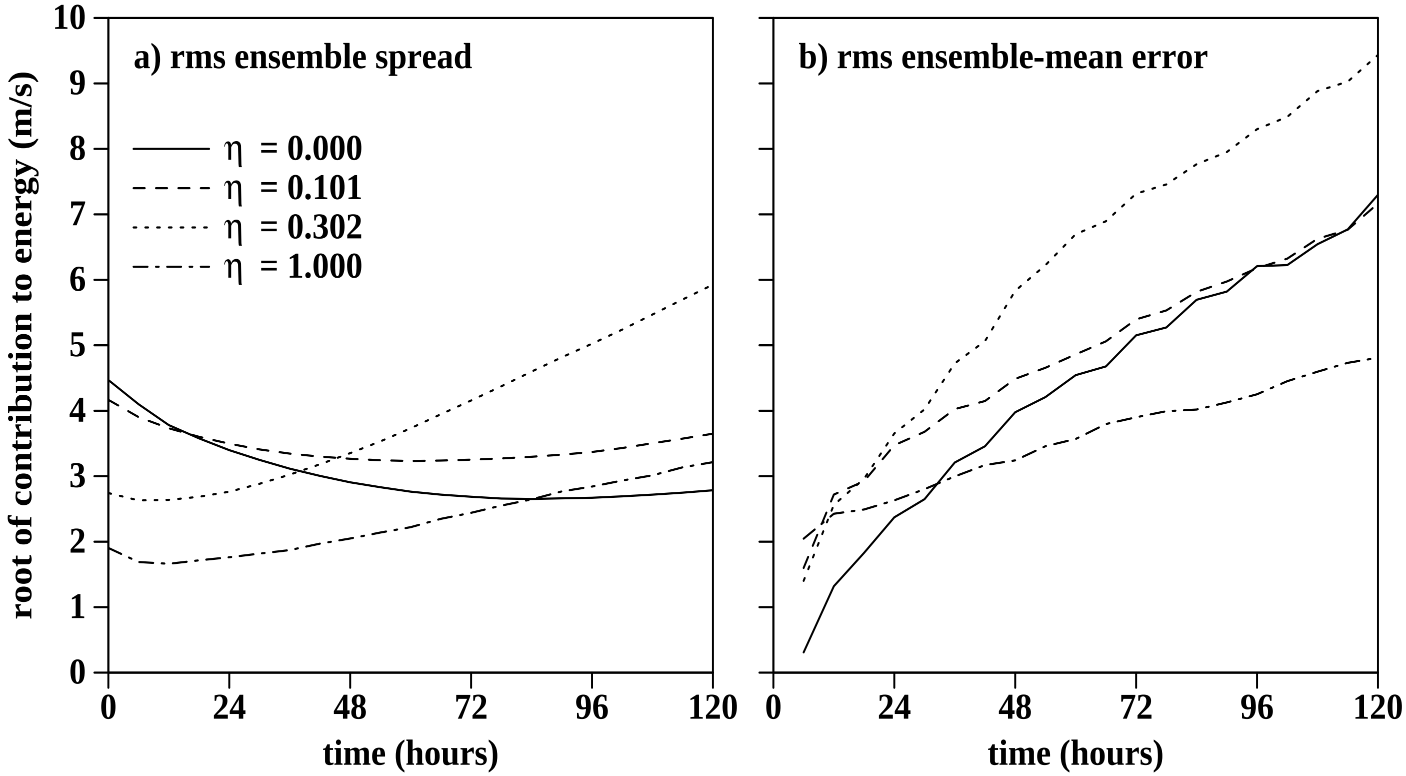
An ensemble of 120-h forecasts has been initiated from an ensemble of initial conditions provided by the EnKF. The ensemble spread decays for about 24 h. The ensemble mean forecasts are validated against the subsequently performed ensemble mean analyses. The growth rate of the actual error is higher than the growth rate of the simulated error.



Results from Houtekamer et al. 2003. To be submitted to Mon. Wea. Rev.

Error dynamics per level

To better understand the initial decrease of the ensemble spread we look at the error energy at the levels $\eta = 0.0$ (model top), 0.101, 0.302 and 1.0 (model surface). The ensemble spread at the model top decreases for about 4 days (perhaps due to model diffusion). The true error increases at all levels. This suggests that the model error is significant near the model top.



Results from Houtekamer et al. 2003. To be submitted to Mon. Wea. Rev.

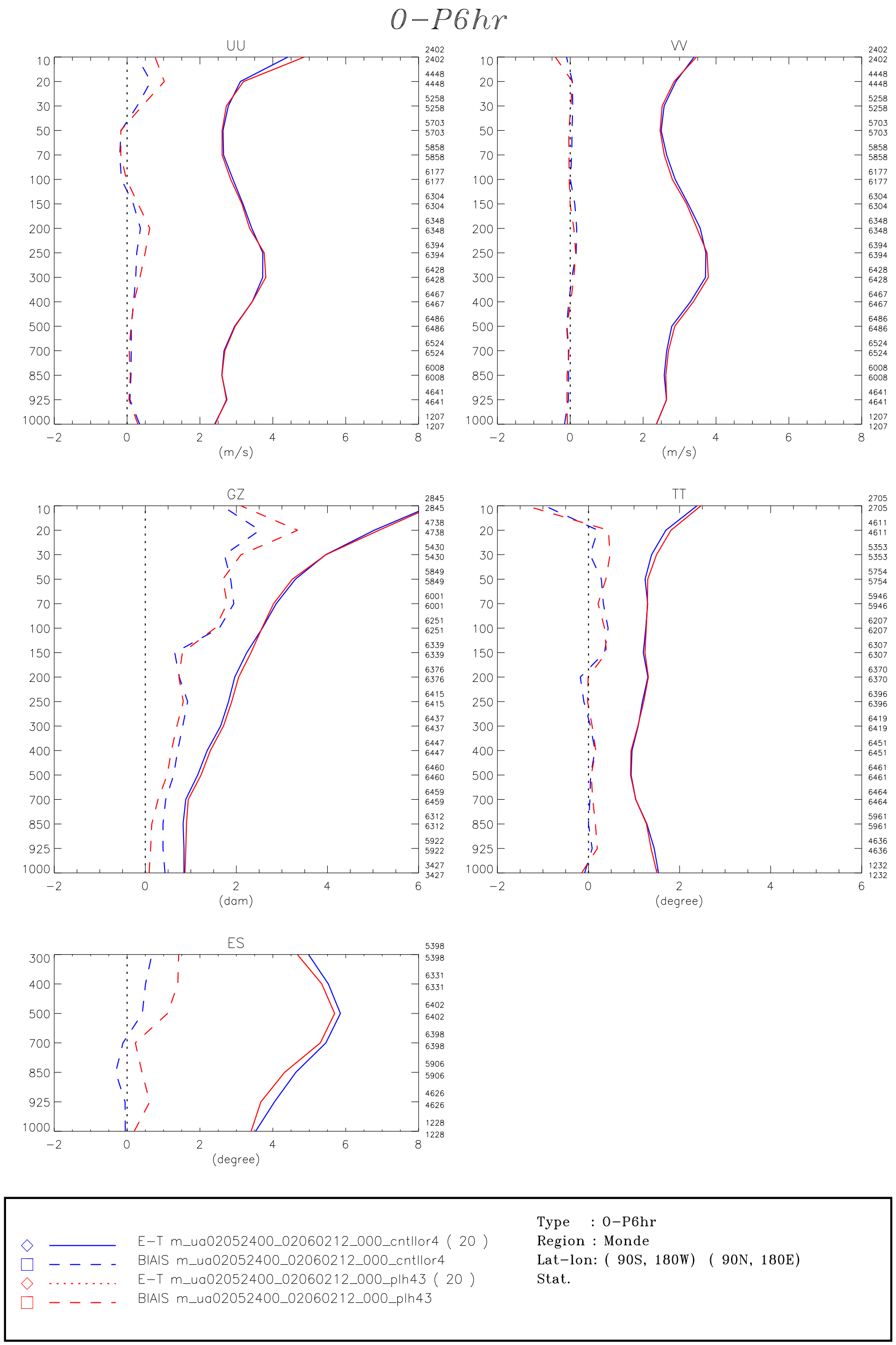
Comparison of 3D-VAR and EnKF

A comparison of 3D-VAR and EnKF has been performed
(manuscript by Houtekamer et al. to be submitted to MWR).

The 3D-VAR and the EnKF have been used with exactly the same forecast model (resolution, physical parameterizations, etc) and with exactly the same observational network. The same data are assimilated. The same error statistics are used for the observations. The same quality control procedure (background check and variational quality control) is used.

Data assimilation cycles were started on 00 UTC May 19 2002.

The innovation statistics are compared for the period 00 UTC May 24 - 12 UTC June 2 2002 (a 10 day period). Innovation statistics were computed with respect to an extremely reliable subset of the radiosonde network.



Agreement between the EnKF and the 3D-VAR

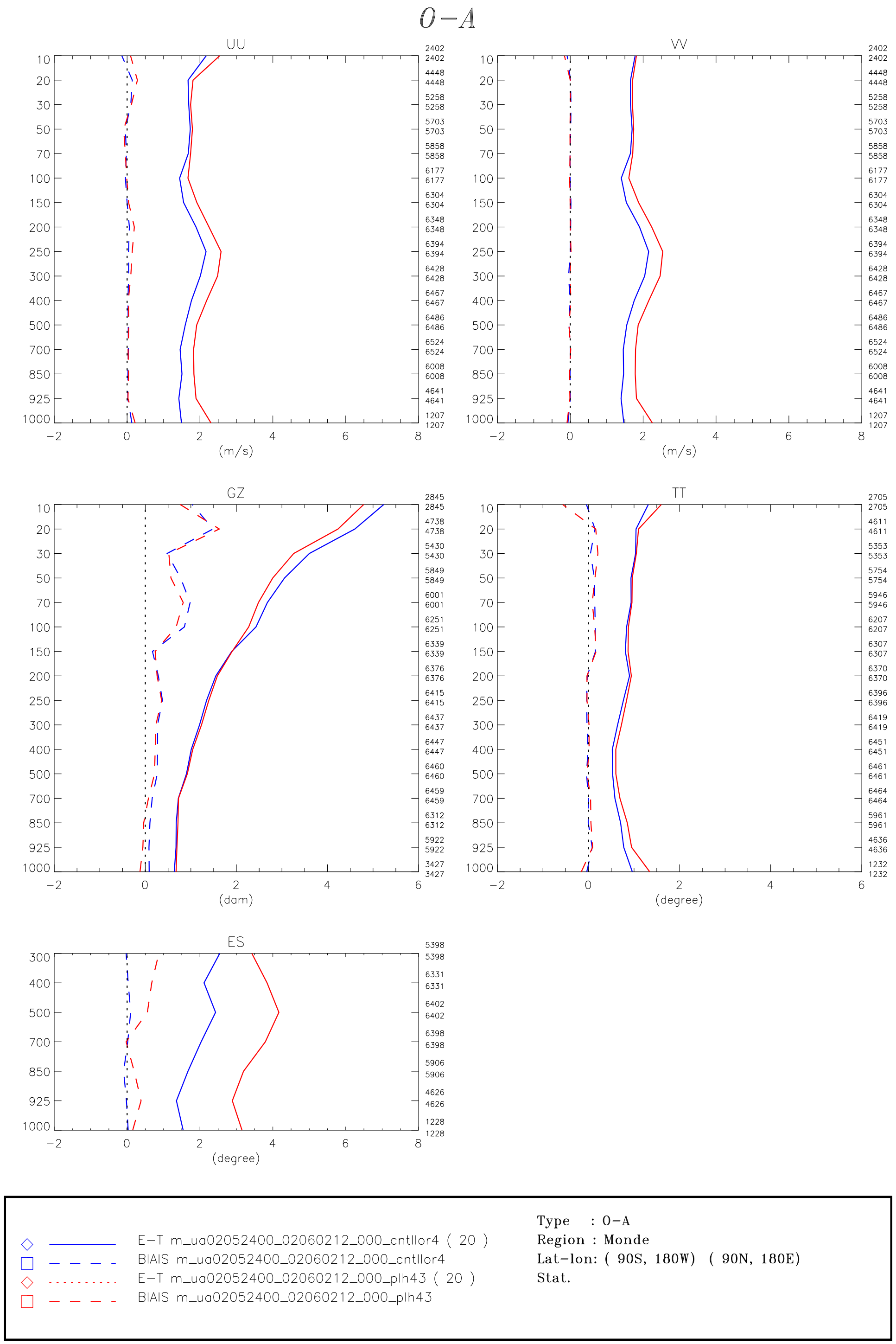
3D-VAR is in blue.

EnKF is in red.

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 rms error.

Generally the scores are very similar. It would appear that the impact of the 4D aspect is small.



Difference between the EnKF and the 3D-VAR

3D-VAR is in blue.

EnKF is in red.

The 3D-VAR analysis draws much closer to the observations than the EnKF analyses. However, at the time of the 6 hour forecasts the two systems are of about the same quality.

Comparison of the EnKF and the 3D-VAR

A detailed comparison of 3D-VAR and the EnKF is currently in progress at MSC.

Innovation statistics with respect to independent radiosonde observations are fairly similar. This implies that the quality of the two analysis systems is fairly similar.

The Observation-Analysis statistics are fairly different. Apparently the schemes give different weights to the observations in spite of the similar quality of the background fields. This implies that either the 3D-VAR or the EnKF can be improved by means of a retuning of some of its parameters.

The comparison is not just of academic interest. It is also a means to validate the new EnKF algorithm.

As we look in more different ways at the EnKF results the quality of these results tends to improve. The extension of the diagnostic package for the EnKF is currently a priority.

Conclusion

The EnKF can be used for atmospheric data-assimilation.

Suggested areas for research:

- Formulation of the model error,
- Dynamics of errors in the data-assimilation cycle,
- Choice between high horizontal or vertical resolution or large ensemble size and less severe localization.