Variational Data Assimilation: Theory and Overview

Florence Rabier and Zhiquan Liu*

Météo-France/CNRM/GMAP, Toulouse, France
** National Satellite Meteorological Center, Beijing, China*

ABSTRACT

Data assimilation is briefly described in its variational formulation. Both three- and four-dimensional variational assimilation are presented with an emphasis on their comparison, strengths and weaknesses. A toy model is used to illustrate the benefit of including the time dimension in the analysis, before summarizing a few operational results.

1 Introduction

Data assimilation is a major component of Numerical Weather Prediction. The data assimilation problem consists in using the available observations together with the model trajectory to provide an accurate description of the atmospheric state. This so-called "analysis" can then be used to initialize a forecast or on its own, for instance to help understand atmospheric properties or in the context of field experiments or re-analyses over long periods of time. There are mainly two different ways of performing data assimilation. The sequential way is using observations in small batches in time, as they become available. In contrast, the continuous way is working over time windows, using all the observations together. This is particularly well suited for re-analyses problems to obtain the best possible state of the atmosphere at time t using observations before and after this time. In general, for operational NWP, the time window of interest is typically 3 to 12 hours, due to the frequency of forecasts which are issued to the users. The most commonly used is the 6-hour assimilation window, which will be the basis of most of the discussions in this paper. There are various sources of information available within the assimilation window. The two main ones are the observations and a background vector (which usually comes from a short-range forecast). All this information characterizes the atmospheric state $\mathbf{x}$, but each source of information is characterized by various errors (mainly instrument or representativeness errors for the observations, forecast error for the background information). One cannot simply mix this information together to get an analysis, without taking these errors into account. There is a need for a statistical approach to set up the problem properly. In other words, one needs to find the best compromise between the various sources of information, trusting each of them according to their error statistics. The error statistics are assumed to be known in this paper, although this is the subject of active research: in practice, one can only evaluate or approximate a few of their components and assumptions are needed to define completely the relevant matrices. There is a particular technique, called variational assimilation, which solves the analysis problem through the optimisation of a given criterion (minimisation of a so-called cost-function). This allows to solve the global problem in one go, and it is now widely used in the meteorological community. There are various ways to account for the time dimension in this optimisation problem. The principle of four-dimensional variational (4D-Var) assimilation usually assumes implicitly that the forecast model is "perfect" within the assimilation window and looks for the model trajectory which best fits the data (background and observations) over the window. A straightforward approximation of 4D-Var is 3D-Var where there is no attempt to use the time dimension within the assimilation window: one looks for the best compromise between the background and all available observations as if they were at the analysis time. A more complex approximation of 4D-Var is 3D-FGAT (First-Guess at the Appropriate Time) in which one compares the model to observations using the model trajectory over the assimilation window, but performs the analysis in three-dimensions (at the central time of the window). These three variations of variational techniques will be further discussed in the various sections of this paper.

2 General description of variational assimilation

2.1 Statistical estimation: the viewpoint of least squares

With the observations

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \boldsymbol{\varepsilon}_\mathbf{r}$$

and the background vector (which usually comes from a short-range forecast)

$$\mathbf{x}_b = \mathbf{x} + \boldsymbol{\varepsilon}_\mathbf{b}$$

The least squares method for obtaining the estimation is to minimize the cost function

$$J(\mathbf{x}) = \tfrac{1}{2}(\mathbf{x}-\mathbf{x}_b)^{\mathrm{T}}\mathbf{B}^{-1}(\mathbf{x}-\mathbf{x}_b) + \tfrac{1}{2}(\mathbf{y}-\mathbf{H}\mathbf{x})^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{y}-\mathbf{H}\mathbf{x}) \tag{1}$$

where $\mathbf{B}$ is the covariance matrix of the background error ε_b and $\mathbf{R}$ the covariance matrix of the observation error ε_r which includes the instrument error and representativeness error. $\mathbf{H}$ is called the observation operator which maps the variable from the model space to the observation space. The minimisation of (1) requires the gradient of the cost function with respect to $\mathbf{x}$ to vanish, namely

$$\tfrac{\partial J(\mathbf{x})}{\partial \mathbf{x}} = \mathbf{B}^{-1}(\mathbf{x}-\mathbf{x}_b) - \mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{y}-\mathbf{H}\mathbf{x}) = 0 \tag{2}$$

The solution of (2) is given (e.g., Lorenc, 1986) by

$$\mathbf{x}_a = \mathbf{x}_b + \mathbf{K}(\mathbf{y}-\mathbf{H}\mathbf{x}_b) = (\mathbf{I}-\mathbf{K}\mathbf{H})\mathbf{x}_b + \mathbf{K}\mathbf{y} \tag{3}$$

where the optimal gain matrix $\mathbf{K}$ is given by

$$\mathbf{K} = \mathbf{B}\mathbf{H}^{\mathrm{T}}(\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}}+\mathbf{R})^{-1} = (\mathbf{B}^{-1}+\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1} \tag{4}$$

The corresponding analysis error is given by

$$\varepsilon_a = (\mathbf{I}-\mathbf{K}\mathbf{H})\varepsilon_b + \mathbf{K}\varepsilon_r \tag{5}$$

Note that the expressions for the analysis and its error have the same form. The analysis error covariance is

$$\mathbf{A} = (\mathbf{I}-\mathbf{K}\mathbf{H})\mathbf{B}(\mathbf{I}-\mathbf{K}\mathbf{H})^{\mathrm{T}} + \mathbf{K}\mathbf{R}\mathbf{K}^{\mathrm{T}} = (\mathbf{I}-\mathbf{K}\mathbf{H})\mathbf{B} \tag{6}$$

where the background and the observation errors have been assumed to be uncorrelated. Note that the simpler expression $\mathbf{A} = (\mathbf{I}-\mathbf{K}\mathbf{H})\mathbf{B}$ is valid only when the $\mathbf{B}$ and $\mathbf{R}$ matrices in $\mathbf{K}$ are correctly specified. In this case, the term "optimal interpolation" (OI) is used to name the analysis scheme. If that is not the case, the expression $\mathbf{A} = (\mathbf{I}-\mathbf{K}\mathbf{H})\mathbf{B}(\mathbf{I}-\mathbf{K}\mathbf{H})^{\mathrm{T}} + \mathbf{K}\mathbf{R}\mathbf{K}^{\mathrm{T}}$ is more general for any $\mathbf{K}$. The term "statistical interpolation" (SI) is preferable.

2.2 Statistical estimation: the viewpoint of minimum variance

The OI/SI scheme in the early literature was often derived from the viewpoint of the minimum analysis error variance. Moreover, the analysis equation is usually derived using the form of component and the observation operator $\mathbf{H}$ is not explicitly present in the equation (e.g., Lorenc, 1981), which often leads to a forest of superscripts and subscripts. Here we give a more compact derivation using the form of the vector-matrix. Because the problem is linear, the solution is expected to take the form

$$\mathbf{x}_a - \mathbf{x}_b = \mathbf{K}(\mathbf{y}-\mathbf{H}\mathbf{x}_b) \tag{7}$$

That is, the analysis increment is the linear combination of the innovation departure (also called observation increment $\mathbf{y}-\mathbf{H}\mathbf{x}_b$) with a matrix $\mathbf{K}$ to be determined as the weight of the linear combination. Assuming that the background error and the observation error are uncorrelated, the analysis error covariance is given by

$$\mathbf{A} = (\mathbf{I}-\mathbf{K}\mathbf{H})\mathbf{B}(\mathbf{I}-\mathbf{K}\mathbf{H})^{\mathrm{T}} + \mathbf{K}\mathbf{R}\mathbf{K}^{\mathrm{T}} \tag{8}$$

The solution of the minimum variance requires $\frac{\partial}{\partial \mathbf{K}} trace(\mathbf{A}) = 0$. Using the following two relations about the derivative of matrices

$$\frac{\partial}{\partial \mathbf{A}} trace(\mathbf{B}\mathbf{A}\mathbf{C}) = \mathbf{B}^{\mathrm{T}}\mathbf{C}^{\mathrm{T}},$$
$$\frac{\partial}{\partial \mathbf{A}} trace(\mathbf{A}\mathbf{B}\mathbf{A}^{\mathrm{T}}) = \mathbf{A}(\mathbf{B}+\mathbf{B}^{\mathrm{T}}),$$

one gets

$$\begin{aligned} &\tfrac{\partial}{\partial \mathbf{K}} trace(\mathbf{A}) \\ &= -(\mathbf{I}-\mathbf{K}\mathbf{H})(\mathbf{B}+\mathbf{B}^{\mathrm{T}})\mathbf{H}^{\mathrm{T}} + \mathbf{K}(\mathbf{R}+\mathbf{R}^{\mathrm{T}}) \\ &= -2(\mathbf{I}-\mathbf{K}\mathbf{H})\mathbf{B}\mathbf{H}^{\mathrm{T}} + 2\mathbf{K}\mathbf{R} \\ &= -2\mathbf{B}\mathbf{H}^{\mathrm{T}} + 2\mathbf{K}(\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}} + \mathbf{R}) \end{aligned} \tag{9}$$

and the optimal weight matrix $\mathbf{K}$ is then given by

$$\mathbf{K} = \mathbf{B}\mathbf{H}^{\mathrm{T}}(\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}} + \mathbf{R})^{-1} \tag{10}$$

As expected, the result is the same as that derived from the viewpoint of the least squares.

2.3 3D/4D-Var algorithm

The direct minimisation of J(x) as given in (1) is called 3D-Var. It has been described above for a linear version, for which a direct solution can also be obtained. More generally, the formulation can be generalized to conditions with a non-linear observation operator H. In this case, $\mathbf{H}\mathbf{x}$ in Eq. (1) should be replaced by $H(\mathbf{x})$. The solution can be obtained by an iterative minimization algorithm which requires the computation of the gradient of the cost function with respect to the state vector at each iterative step. Currently, the variational algorithms are algorithms that use the adjoint (AD) model (of the tangent linear model) of the non-linear model H to compute this gradient (Le Dimet and Talagrand, 1986; Lewis and Derber, 1985). Consequently, the variational algorithm is sometimes called the "adjoint method". The 4D-Var algorithm can be considered as an extension of the 3D-Var to the time dimension, for which the operator H also includes the dynamical models to evolve in time the atmospheric state. One of the advantages of the 4D-Var algorithm over 3D-Var is to better consider the time distribution of observations such as from remote-sensing instruments. The standard 4D-Var algorithm and the incremental algorithm (currently used in the operational implementations) are described below.

2.3.1 Standard algorithm of the 4D-Var

The atmospheric flow is governed by a number of dynamical and physical laws. One can symbolically write the NWP model as

$$\mathbf{x}_{i+1} = M_{i+1,i}(\mathbf{x}_i) \tag{11}$$

where $M_{i+1,i}$ is the non-linear NWP model from time t_i to t_{i+1}. A perturbation of the atmospheric state is evolved by the tangent linear (TL) model, namely

$$\delta\mathbf{x}_{i+1} = \mathbf{M}_{i+1,i}(\mathbf{x}_i)\delta\mathbf{x}_i \tag{12}$$

Note that the linearisation is performed in the vicinity of $\mathbf{x}_i$. In general, the NWP model has both systematic and random errors which should be considered in the algorithm. For the sake of simplicity, one does not treat them in the following description. That is, the model is used as a "strong constraint" according to the terminology in Sasaki (1970).

In the **standard formulation** of the 4D-Var (Le Dimet and Talagrand, 1986), the solution is found by minimizing a cost function $J(\mathbf{x}_0)$ which measures the distance between the model trajectory and the observations as well as the background at the initial time during a time window (i.e., 6h, 12h, 24h), given by

$$\begin{aligned} J(x_0) &= \frac{1}{2}(\mathbf{x}_0-\mathbf{x}_b)^{\mathrm{T}}\mathbf{B}^{-1}(\mathbf{x}_0-\mathbf{x}_b) \\ &\quad +\frac{1}{2}\sum_{i=0}^{N}[H_i(\mathbf{x}_i)-\mathbf{y}_i]^{\mathrm{T}}\mathbf{R}_i^{-1}[H_i(\mathbf{x}_i)-\mathbf{y}_i] \end{aligned} \tag{13}$$

where

$\mathbf{x}_0$ is the model state at time t_0

$\mathbf{x}_b$ is the background state at time t_0, which is often a 6h forecast from the previous analysis.

$\mathbf{B}$ is the background error covariance matrix

$\mathbf{y}_i$ is the observation vector at time t_i

H_i is the observation operator at time t_i

$\mathbf{x}_i = M_{i,0}(\mathbf{x}_0)$ is the model state at time t_i

$\mathbf{R}_i$ is the observation error covariance matrix at time t_i.

Substituting the dynamical constraint (11) into the cost function, the constrained minimization problem becomes unconstrained, and the control variable of the optimal control problem is reduced to the initial model state $\mathbf{x}_0$. The minimization of Eq. (13) requires its gradient with respect to the control variable $\mathbf{x}_0$ to vanish, namely

$$\nabla_{\mathbf{x}_0}J(\mathbf{x}_0) = \mathbf{B}^{-1}[\mathbf{x}_0-\mathbf{x}_b]+\sum_{i=0}^{N}\mathbf{M}_{i,0}^{\mathrm{T}}\mathbf{H}_i^{\mathrm{T}}\mathbf{R}_i^{-1}[H_i(\mathbf{x}_i)-\mathbf{y}_i] = 0 \tag{14}$$

where $\mathbf{H}_i$ is the tangent linear operator of the observation operator H_i ($\mathbf{H}_i^{\mathrm{T}}$ is the corresponding adjoint operator and is simply the complex-conjugate transpose of the tangent linear by definition) and

$$\mathbf{M}_{i,0}^{\mathrm{T}} = \mathbf{M}_{1,0}^{\mathrm{T}}\mathbf{M}_{2,1}^{\mathrm{T}}\cdots\mathbf{M}_{i,i-1}^{\mathrm{T}} \tag{15}$$

is usually called the adjoint model (ADM) and is a backward integration from time t_i to t_0. It can be programmed from the tangent linear model (TLM) of a NWP model and there is no need to derive analytically the adjoint equations (Talagrand and Courtier, 1987).

In general, the minimization of the cost function is attained by using a "descent algorithm" as follows:

- Initialize the first-guess as $\mathbf{x}^0(t_0) = \mathbf{x}_b(t_0)$.
- For iteration step $k = 1,\cdots,K$:

  (i) Compute and save the first-guess trajectory and the observation departures $[H_i(\mathbf{x}_i)-\mathbf{y}_i]$ by integrating the non-linear model $\mathbf{x}^k(t_i) = M(t_i,t_0)\ (\mathbf{x}^k(t_0))$.

(ii) Starting from the adjoint variable $\delta'\mathbf{x}^k(t_N) = 0$, integrate the adjoint model (which will use the first-guess trajectory from step (i)) backwards in time from t_N to t_0. If meeting the observations in the course of the integration, the observation forcing $\mathbf{H}_i^{\mathrm{T}}\mathbf{R}_i^{-1}[H_i(\mathbf{x}_i) - \mathbf{y}_i]$ is added to the currently computed $\delta'\mathbf{x}^k(t_i)$. The integrated final value $\delta'\mathbf{x}^k(t_0)$ plus the background term gradient $\mathbf{B}^{-1}[\mathbf{x}^k(t_0) - \mathbf{x}_b]$ is the gradient ∇J^k of cost function with respect to $\mathbf{x}^k(t_0)$. The cost function value is also computed in this step.

(iii) Find the optimal stepsize $\rho^k(>0)$. For a linear model, ρ^k should be computed from the Hessian of the cost function and can be calculated explicitly. For a non-linear model, ρ^k cannot be determined analytically and has to be estimated approximately (this is generally done by the minimization subroutine itself).

(iv) Update the first guess with

$$\mathbf{x}^{k+1}(t_0) = \mathbf{x}^k(t_0) - \rho^k \nabla J^k$$

(v) Check to see if some convergence criteria are met. If not, go to (i).

Note that in the standard 4D-Var algorithm as described above, the tangent linear model is not used in the computation and is only used to derive the adjoint model.

2.3.2 Incremental algorithm of the 4D-Var

The cost of a backward integration of the ADM is about twice that of a forward integration of the non-linear model. The cost of an iteration will be three times that of a model integration in the standard algorithm. Additionally, the nonlinearity of the model makes the convergence slower. Courtier *et al.* (1994) suggested an incremental 4D-Var algorithm with which the minimization can be carried out at a reduced model resolution and uses a forward integration of the TLM instead of non-linear one. This leads to an effective reduction of computational cost. The incremental 4D-Var algorithm is to minimize a cost function given by

$$J(\delta\mathbf{x}_0) = \frac{1}{2}\delta\mathbf{x}_0^{\mathrm{T}}\mathbf{B}^{-1}\delta\mathbf{x}_0 + \frac{1}{2}\sum_{i=0}^{N}(\mathbf{H}_i\delta\mathbf{x}_i - \mathbf{d}_i)^{\mathrm{T}}\mathbf{R}_i^{-1}(\mathbf{H}_i\delta\mathbf{x}_i - \mathbf{d}_i) \tag{16}$$

where

$\delta\mathbf{x}_0 = (\mathbf{x}_0 - \mathbf{x}_b)$ is called the increment at time t_0,

$\mathbf{d}_i = \mathbf{y}_i - H_i(\mathbf{x}_i)$ is the innovation departure or observation increment at time t_i.

The solution $\delta\mathbf{x}_0^a$ of the cost function minimization is added to the background to obtain the analysis at t_0,

$$\mathbf{x}_0^a = \mathbf{x}_b + \mathbf{S}^{-1}\delta\mathbf{x}_0^a \tag{17}$$

where $\mathbf{S}^{-1}$ is an operator transforming the field from low to high resolution. The process of minimization is similar to the standard algorithm except that the control variable is the increment at time t_0 and the increment trajectory is obtained by the integration of the TLM. The reference trajectory required by the TLM and ADM is from the background integration and is not updated at each iteration. This drawback can be partially overcome by an "outer loop" updating the high resolution reference trajectory and the observation departures. Correspondingly, the iterative procedure of minimizing the incremental control variable is called "inner loop". It is possible to use a successively increased inner loop resolution after each outer loop update (referred to as "multi-incremental" algorithm, Veersé and Thépaut, 1998). Obviously, the incremental algorithm can also be applied to the 3D-Var algorithm.

Figure 1 gives a schematic view of the incremental implementation of 4D-Var, as it had been introduced operationally at ECMWF.

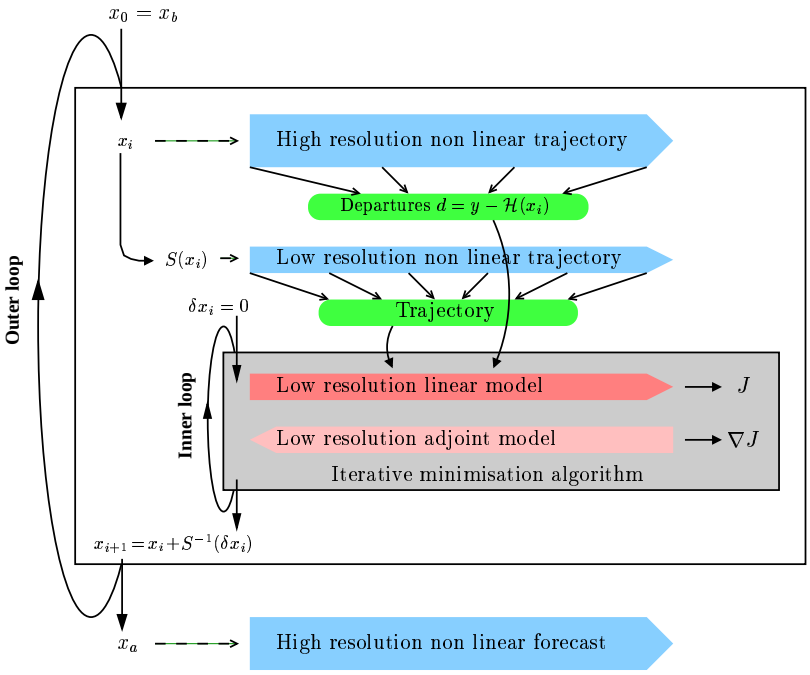


Figure 1: Representation of the incremental 4D-Var (courtesy of Y. Trémolet).

2.3.3 Quality of the analysis

The error covariance for the analysis at initial time t_0 of the assimilation window is equal to the inverse of the Hessian (the second derivative) of the 4D-Var cost function with respect to the initial state vector x_0 (control variable) given by

$$\mathbf{A}_0 = \left(\mathbf{B}^{-1} + \sum_{i=0}^{N} \mathbf{M}_i^{\mathrm{T}} \mathbf{H}_i^{\mathrm{T}} \mathbf{R}_i^{-1} \mathbf{H}_i \mathbf{M}_i \right)^{-1} \tag{18}$$

Minimization algorithms usually provide approximates of the inverse of the Hessian. However, what we want to obtain eventually is the analysis error covariance at the end of the assimilation window which could be used as the background error covariance for the next assimilation cycle. Suppose that the initial error ε_0 within the assimilation window is evolved by the TLM, namely

$$\varepsilon_f = \mathbf{M}\varepsilon_0 \tag{19}$$

where ε_f stands for the error at the end of time window and $\mathbf{M}$ is the TLM integration from the beginning time t_0 to the end time T of the assimilation window. Performing the mathematical expectation operation, one gets the error covariance for the analysis at the end time T

$$\mathbf{A}_f = < \varepsilon_f \varepsilon_f^{\mathrm{T}} > = < \mathbf{M}\varepsilon_0 (\mathbf{M}\varepsilon_0)^{\mathrm{T}} > = \mathbf{M}\mathbf{A}_0\mathbf{M}^{\mathrm{T}} = \mathbf{M}(\mathbf{M}\mathbf{A}_0)^{\mathrm{T}} \tag{20}$$

Note that obtaining $\mathbf{A}_f$ requires $2N$ (N is the model dimension) integrations of the TLM starting from each column of the initial covariance matrix $\mathbf{A}_0$. For a high dimensional NWP model with $N = 10^7$, it becomes an almost impossible task and is a common issue with the Extended Kalman Filter assimilation scheme described below.

2.4 Extended Kalman Filter: update of the forecast error covariance matrix

In general, 3D/4D-Var date assimilation is performed in a cycling manner with a fixed assimilation window. Each cycle uses a short-range (e.g., 6h) forecast from previous analysis as the new background. Naturally, we

expect the background error covariance **B** to depend on the atmospheric state and hence to be time-dependent. The 4D-Var algorithm uses implicitly flow-dependent structure functions (Thépaut *et al.*, 1996) within an assimilation window, but the algorithm itself does not provide the estimate of the background error covariance for the next cycle. The background error covariance is often estimated by the so-called NMC method (Parrish and Derber, 1992) at most NWP centres and is generally specified as constant. The proper update of the background error covariance is not done in an operational assimilation context and remains an active research field.

The Kalman filter theory (Kalman, 1960) provides a basis for updating the forecast error covariance. Standard Kalman Filter (KF) was derived for a linear dynamical system. An extension to non-linear models is referred to as the Extended Kalman Filter (EKF), which is derived by a number of authors (e.g., Daley 1991). EKF and KF formulations have no difference in the form, simply the linear operators (of model and observation) in the KF should be extended to the Tangent Linear (TL) operators in the EKF even though they can use the same notations. EKF can be regarded as a sequential optimal interpolation (OI) or 3D-Var analysis for which the observations are used once they are available. In addition, the forecast error covariance is explicitly computed and used as the background weight for the following analysis. For a perfect forecast model (without model error), the EKF equations are as follows:

- Analysis Step

$$\mathbf{x}_n^a = \mathbf{x}_n^f + \mathbf{K}_n[\mathbf{y}_n - H(\mathbf{x}_n^f)] \tag{21}$$

  with the Kalman gain matrix

$$\mathbf{K}_n = \mathbf{P}_n^f \mathbf{H}_n^{\mathrm{T}} [\mathbf{H}_n \mathbf{P}_n^f \mathbf{H}_n^{\mathrm{T}} + \mathbf{R}_n]^{-1} \tag{22}$$

  and the optimal analysis error covariance

$$\mathbf{P}_n^a = (\mathbf{I} - \mathbf{K}_n \mathbf{H}_n) \mathbf{P}_n^f \tag{23}$$

  or the suboptimal analysis error covariance

$$\mathbf{P}_n^a = (\mathbf{I} - \mathbf{K}_n \mathbf{H}_n)(\mathbf{P}_n^f)^{\mathrm{t}} (\mathbf{I} - \mathbf{K}_n \mathbf{H}_n)^{\mathrm{T}} + \mathbf{K}_n \mathbf{R}_n^{\mathrm{t}} \mathbf{K}_n^{\mathrm{T}} \tag{24}$$

- Forecast Step

$$\mathbf{x}_{n+1}^f = M_n(\mathbf{x}_n^a) \tag{25}$$

$$\mathbf{P}_{n+1}^f = \mathbf{M}_n \mathbf{P}_n^a \mathbf{M}_n^{\mathrm{T}} = \mathbf{M}_n (\mathbf{M}_n \mathbf{P}_n^a)^{\mathrm{T}} \tag{26}$$

Here $\mathbf{x}_n^f$ and $\mathbf{x}_n^a$ are the forecast and analyzed state vectors at time t_n; $\mathbf{y}_n$, H and $\mathbf{H}$ are respectively the observation vector, the non-linear observation operator and its TL at time t_n; $\mathbf{P}_n^f$ and $\mathbf{R}_n$ are the forecast and observation error covariances at t_n; M_n and $\mathbf{M}_n$ stand for the non-linear NWP model and its TL model integrating from time t_n to t_{n+1}. The main difficulty in using the EKF in NWP models is the large computational and storage requirements for the error covariance propagation equation (26), which requires $2N$ integrations of the TLM (N is the model dimension) and the storage of a $N \times N$ covariance matrix. Some simplified schemes based on the EKF have been proposed (e.g., Todling and Cohn 1994). It is also known that the 4D-Var and the EKF are equivalent in some circumstances (e.g., with a perfect model and using the same TL approximation, and the initial background as well as its error covariance are also the same for both). They obtain the same analysis and corresponding error at the end of a data assimilation window (Rabier *et al.*, 1993). Over the whole assimilation window, 4D-Var is equivalent to a smoother algorithm. 4D-Var can be extended to non-Gaussian errors (Lorenc and Hammon, 1988; Andersson and Järvinen, 1998). Another advantage of 4D-Var is that it can use a wide range of observations including those with a complex link to atmospheric variables (e.g. radiances; Andersson et al., 1994). It can also efficiently use asynoptic data as shown in Järvinen et al (1999).

3 1D Burgers' equation application with advanced assimilation schemes

In this section, we introduce a 1D non-linear advection-diffusion equation (Burgers' model). Some advanced assimilation algorithms (variational and Extended Kalman Filter) are implemented and compared.

3.1 1D Burgers' model and its tangent linear and adjoint

One considers a 1D non-linear advection-diffusion equation (Burgers' equation)

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = \nu\frac{\partial^2 u}{\partial x^2} \tag{27}$$

where $u(x,t)$ and ν are respectively wind speed and diffusive coefficient and x the coordinate of a 1D periodic domain, $-\pi a \leq x < \pi a$ with $a = 1250km$, as defined in Liu and Rabier (2002). The corresponding tangent linear equation is given by

$$\frac{\partial \delta u}{\partial t} + u\frac{\partial \delta u}{\partial x} + \delta u\frac{\partial u}{\partial x} = \nu\frac{\partial^2 \delta u}{\partial x^2} \tag{28}$$

where δu is the perturbation in the vicinity of the reference state u. In practice, one will take the trajectory of the non-linear model as the reference state. The pseudo-spectral method (Orszag 1970) is used to construct our numerical model and the model variable consists of real Fourier coefficients. A semi-implicit leapfrog integration initialized by a forward Euler step is used as the time discretization scheme. An Asselin time filter (Asselin, 1972) is used to remove the computational solution of the leapfrog scheme.

3.2 Frontal type dynamics

The sinusoidal initial condition

$$u(x,0) = -Usin(x/a) \tag{29}$$

with $U = 20m/s$ is used as the true field at the initial time t_0. It can produce frontal type dynamics (see Ménard, 1994 and references therein), which are commonly observed in the atmospheric wind field. A main property of the evolution of the initial error for the frontal type dynamics is the rapid growth of error variance near the front (convergence area). In the subsequent numerical experiments, one always works in the diffusive case with the Reynolds number $Re = 2\pi aU/\nu = 100$, for which the maximal development of the front at $x = 0$ (the strongest gradient $\partial u/\partial x$) is reached at a critical time $t_c = (a/U) * \pi/2 \approx 27$h.

For demonstrating the evolution of the initial error, one constructs an initial background field $u_b(x,0)$ by adding a correlated random noise to the true initial condition $u(x,0)$. The background error correlation is the same degenerate second-order autoregressive function with a correlation scale of $208km$ as in Liu and Rabier (2002, Eq. (6)). The homogeneous error standard deviation σ_b is equal to $2m/s$. The top panels in Figure 2 give the results of a 48h integration starting from the true (solid line) and background (dashed line) initial conditions. One clearly sees the decrease of the error in the divergence areas and the growth in the frontal area near $x = 0$. The consequence of this error growth is a misdetermination of the frontal position, such as the one observed in the forecast at 24h and 48h. The corresponding evolution of the background error field is presented in the same figure (bottom panels). The solid curves are the difference between two non-linear model integrations respectively with $u(x,0)$ and $u_b(x,0)$ as the initial conditions, and the dashed lines are the results of the propagation of the initial error by the TLM. The error at $x = 0$ grows up to about $8m/s$ at time t_c from the initial error of $-2m/s$ then decreases slowly during the frontolysis stage. The error obtained by the TLM is only slightly larger than that obtained by the non-linear model integration even until 48h integration. The precision of the tangent linear approximation will be high in the subsequent assimilation experiments with a 6h or 24h assimilation window.

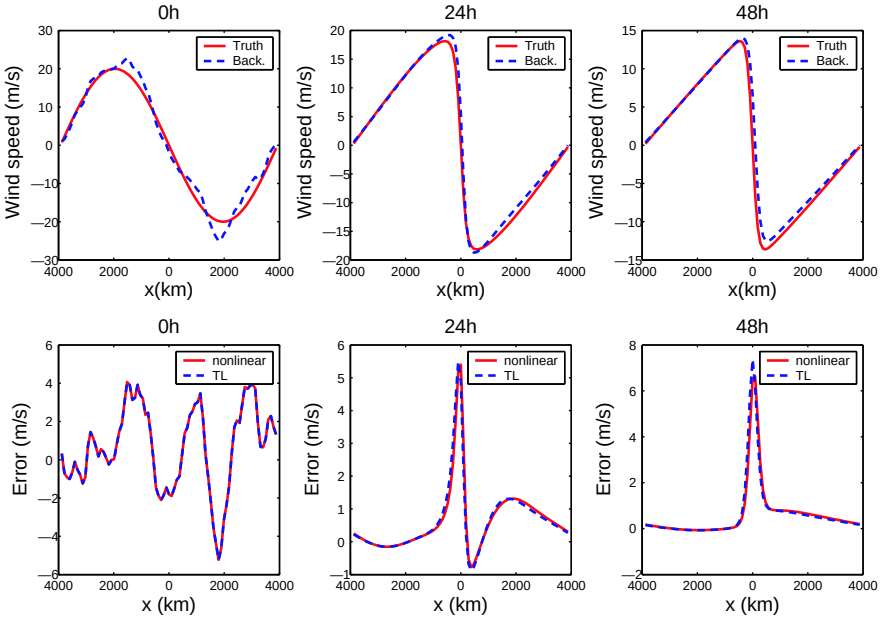


Figure 2: Frontal type dynamics. Top panels: evolution of the wind field, solid line for the true field and dashed line for the background field; Bottom panels: evolution of the error field, solid line for the difference between two non-linear model integrations, dashed line for the propagation of the initial error using the tangent linear model.

3.3 Comparison of several assimilation schemes

Although this is an academic situation, it can be helpful to understand the possible impact of the various assimilation schemes.

With the frontal type dynamics described above as the true meteorological situation, the following six experiments are compared:

(1) EKF that uses the equation (26) to propagate the forecast error covariance

(2) 3D-VAR-S6 that uses the standard 3D-Var algorithm with a 6h assimilation window

(3) 4D-VAR-S6 that uses the standard 4D-Var algorithm with a 6h assimilation window

(4) 4D-VAR-I6 that uses the incremental 4D-Var algorithm with a 6h assimilation window

(5) 4D-VAR-S24 that uses the standard 4D-Var algorithm with a 24h assimilation window

(6) 4D-VAR-I24 that uses the incremental 4D-Var algorithm with a 24h assimilation window.

The schemes (2) $\sim$ (6) use a constant background error covariance for every assimilation cycle. Note that the names "3D-Var" and "4D-Var" are used to denote the experiments, although our study only deals with a 1D spatial domain. The observation network is fixed and coincides with the grid points associated with the "spectral" model. The simulated observations are available every 6h. The assimilation is done from 00h to 48h. At 00h, only the background field is available with the error covariance described in section 3.2. This means that the observations are located at the end of the assimilation window for a 6h window. The observations are simulated by adding the uncorrelated random noise to the true field. The observation operator is a real Fourier (linear) interpolation which transforms directly $2K+1$ real Fourier coefficients to grid point values in the arbitrary N observation positions. All six experiments are carried out with a Fourier spectral truncation $T41$

(83 grid points), the incremental algorithm did not use a reduced resolution and the outer loop to update the trajectory.

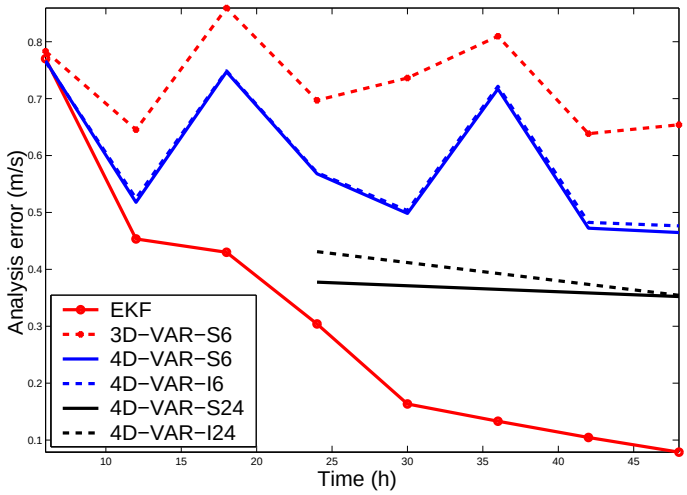


Figure 3: Analysis RMS error for 6 assimilation schemes (the different schemes are defined at the beginning of 3.3. The analysis errors are shown every 6 hours starting from 6h, except for 24-hr 4D-Var).

The analysis RMS error for 6 experiments is shown in Figure 3. As expected, the scheme **EKF** which correctly propagates the error covariance is obviously the best one. Note that the analysis errors for the 4D-Var experiments are those at the end of the assimilation window. The equivalence of the 4D-Var and the EKF can be seen from results at 06h by comparing the scheme **EKF** and the scheme **4D-VAR-S6**. This equivalence is determined by the consistency of the initial background error covariance and the TLM. **4D-VAR-I6** is only slightly worse than **4D-VAR-S6**. The analysis error is further reduced for **4D-VAR-S24** and **4D-VAR-I24** with a longer assimilation window. The difference between the incremental and standard 4D-Var algorithms is more obvious in the first cycle of **4D-VAR-S24** and **4D-VAR-I24** for which the update of the trajectory is more important. In the second cycle for **4D-VAR-S24** and **4D-VAR-I24**, small initial differences lead to the same analysis result at 48h, although both are worse than that of **EKF** due to the incorrect specification of the initial background (the analysis at 24) error covariance. Finally, one notes that **4D-VAR-S6/I6** is systematically better than **3D-VAR-S6**. The improvement comes from the advantage of the dynamical structure functions implied in the 4D-Var (Thépaut *et al.*, 1996; Rabier *et al.*, 2000).

Figure 4 shows the analysis error fields at 24h for the 6 experiments. One can see that the large error reduction for **EKF, 4D-VAR-S24** and **4D-VAR-I24** is located in the divergence areas where the correct propagation of the background error covariance implies that more weight is given to the accurate background at 24h (see Figure 2 for the evolution of the initial error). However, **4D-VAR-I24** has the largest error in the frontal area where nonlinearity of the flow is stronger and the validity of the incremental formulation is reduced. For **3D-VAR-S6** and **4D-VAR-S6/I6**, serious overestimation of the background error in the divergent areas implies that more weight is given to observations, which leads to less error reduction.

This academic example clearly shows that there will be some benefit in including explicitely the time dimension in the assimilation (4D-Var versus 3D-Var). However, one can notice that even 4D-Var is sub-optimal when cycled with a constant B matrix.

4 Operational results

After a few papers demonstrating the strengths of 4D-Var in global models on case studies or with simulated data, Rabier et al (1998) performed some experiments in a context close to the operational one. Results showed

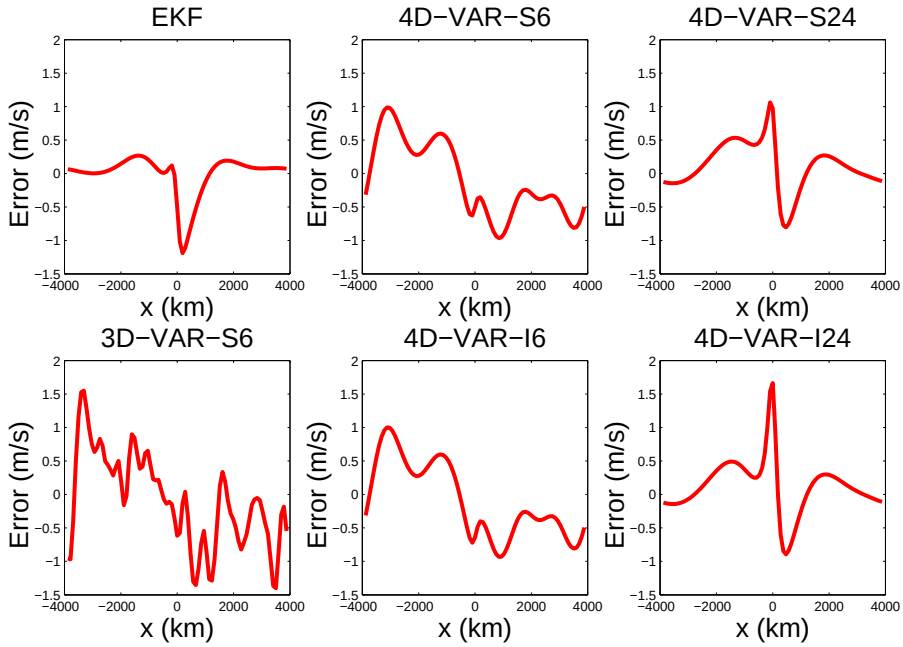


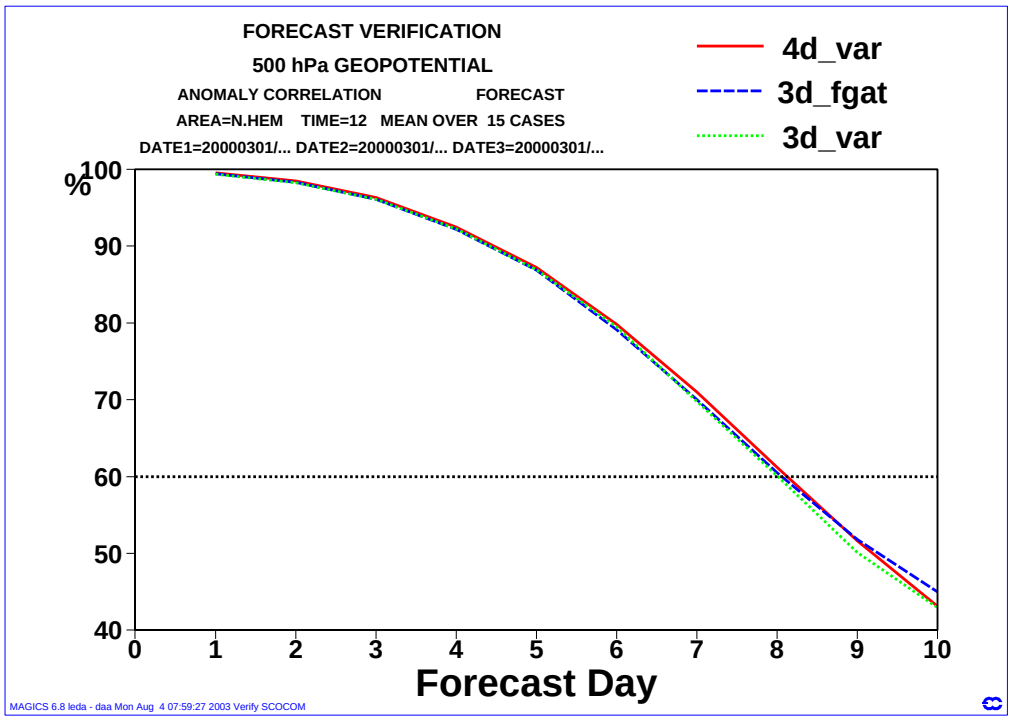
Figure 4: Analysis error fields at 24h for the 6 assimilation schemes.

that 4D-Var over 6 and 12-hour assimilation windows performed better (in cycling mode) than 3D-Var. The 6-hour 4D-Var was then further developed and tested to reach operational testing. An operational 4D-Var assimilation system has been respectively implemented on 25th November 1997 at ECMWF (Rabier *et al.*, 2000; Mahfouf and Rabier 2000; Klinker *et al.*, 2000) and on 20th June 2000 at Météo-France (Janisková *et al.*, 1999; Thépaut and Moll, 2000; Gauthier and Thépaut, 2001, Desroziers *et al.*, 2003). Results showed that, even with only very simplified physics in the minimization, 4D-Var outperforms 3D-Var in the extra-tropics. 4D-Var performs particularly well in dynamically active areas, such as the Atlantic storm-track (Rabier et al, 2000; Desroziers et al., 2003). The impact of more elaborate linearized physics in the minimisation are beneficial mostly in the Tropics. One notes that there is a better agreement between inner and outer loops when a more elaborate physics package is used in the minimisation (Mahfouf and Rabier, 2000). The overall performance of 4D-Var is slightly positive, at all ranges. Significance tests show that the improvement was significant, particularly at short ranges and in the Southern Hemisphere (Klinker *et al.*, 2000).

More recent results compared 3D-Var, 3D-FGAT and 6-hour 4D-Var over a two-week period in March 2000 (courtesy of E. Andersson). Scores are presented in Figures 5 and 6 for the anomaly correlation over the Northern and Southern Hemispheres, respectively. One can see from Figure 5 that the 4D-Var performance is slightly better than 3D-Var and 3D-FGAT which behave similarly in the Northern Hemisphere. In Figure 6, 4D-Var has a larger advantage over 3D-FGAT in the Southern Hemisphere, which itself performs better than 3D-Var. It is then confirmed that 4D-Var, even on a 6-hour window performs better than Three-dimensional algorithms (either 3D-Var or 3D-FGAT). The impact is more marked in the Southern Hemisphere. In this area, there is a slight benefit of using 3D-FGAT compared with 3D-Var, which comes from comparing the observations and the model at the observation time.

Concerning other operational developments, one might be interested to read about the extraction of temporal information through the use of frequent data (Järvinen *et al.*, 1999), developments in the background error covariance estimation (Derber and Bouttier, 1999), the implementation of 12-hour 4D-Var (Bouttier, 2001), the multi-incremental technique (Veersé and Thépaut, 1998), initialisation with the Digital Filter technique (Gauthier and Thépaut, 2001), or the continuous improvement of linearized physics (Janiskova, this volume).

There has also been some interesting developments in the context of limited-area models (eg Huang, 1999, Zou and Kuo, 1996, Zupanski, 1993, Gustafsson, this volume).



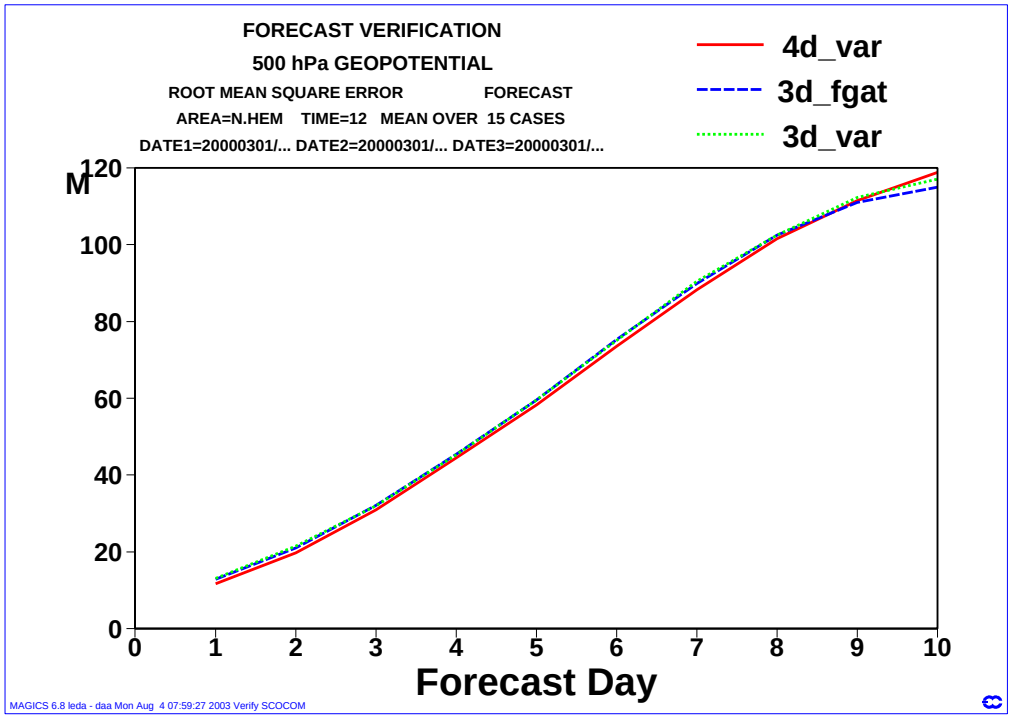


Figure 5: RMS of forecast error at various ranges over the Northern Hemisphere for the 3 assimilation schemes.

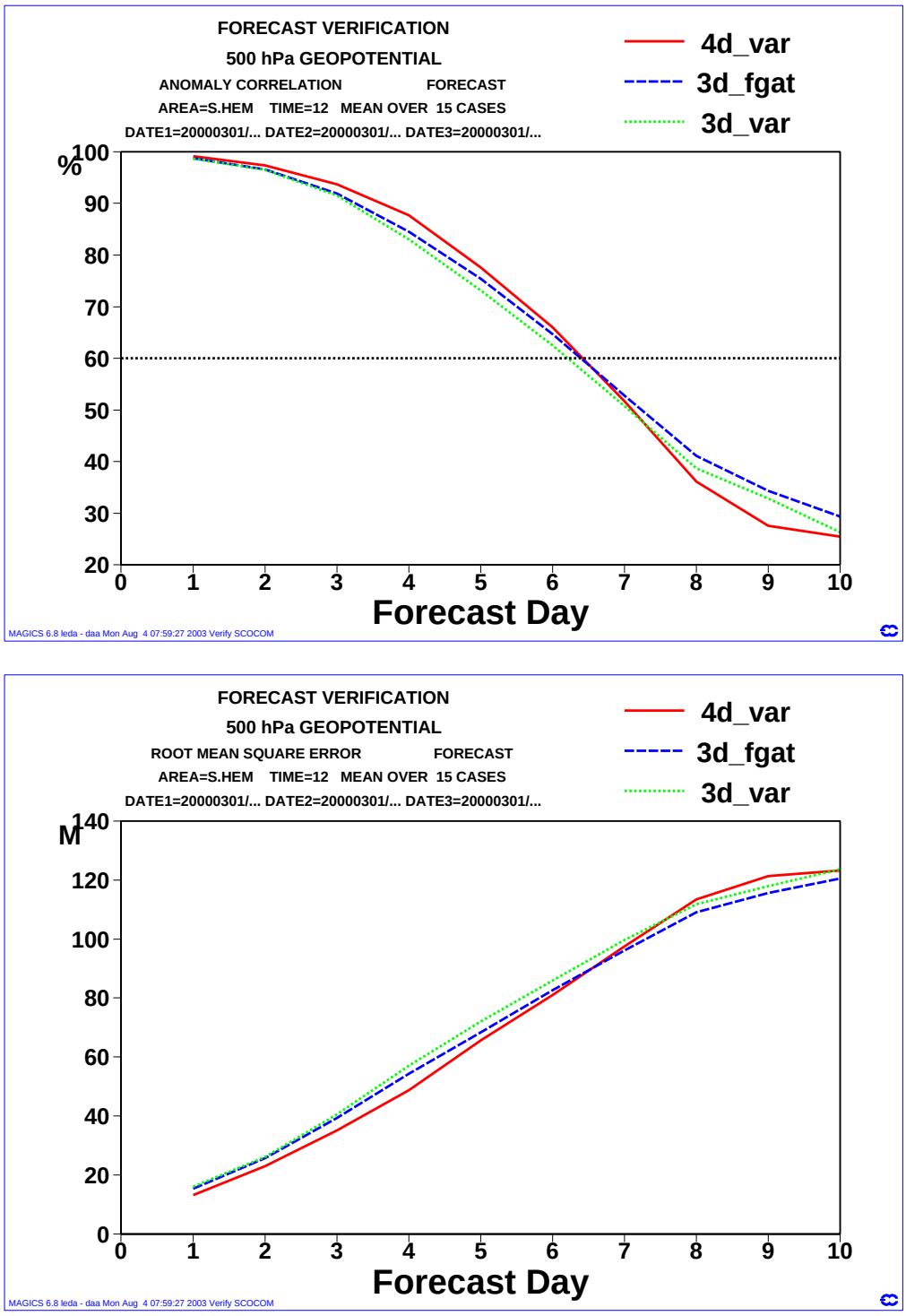


Figure 6: RMS of forecast error at various ranges over the Southern Hemisphere for the 3 assimilation schemes.

5 Limitations and perspectives

Although these variational techniques are both very powerful and elegant, one cannot deny that there are some obvious drawbacks. The major drawback is the amount of effort which is needed to implement these methods, in terms of heavy technical developments (coding of the adjoint operators in particular). Another limitation comes from the need for an incremental technique, which allows the problem to be used in high-dimension models but needs careful testing.

In terms of perspectives, there are some very challenging issues arising currently, after the successful implementation of 4D-Var has been completed. In terms of data assimilation theory, more effort has now to be devoted to the inclusion of model error terms in 4D-Var, which would alleviate the perfect model assumption (Trémolet, this volume), and to a combination with more probabilistic approaches to get better estimates of analysis error uncertainty (Lorenc, this volume). Regarding the data usage, two major perspectives deal with the need for a higher resolution analysis using high density satellite data, and with the new promising area of cloud and rain assimilation. Variational techniques certainly seem to be an appropriate framework to go forward in these various directions, in order to provide a more comprehensive system in the coming years.

References

[1] Andersson, E. and H. Järvinen, 1998: Variational quality control. *Technical Memorandum, ECMWF,* **250,** 31pp

[2] Andersson, E., J. Pailleux,J., Thépaut, J-N., Eyre, J., McNally, A.P., Kelly, G., and P. Courtier, 1994: Use of cloud-cleared radiances in three/four-dimensional variational data assimilation. *Q. J. R. Meteorol. Soc.*, **120,** 627-653

[3] Asselin, R.A., 1972: Frequency Filter for time integration. *Mon. Wea. Rev.*, **100,** 487-490

[4] Bouttier, F., 2001: The Development of 12-hourly 4D-Var. *Technical Memorandum, ECMWF,* **348,** 21pp

[5] Courtier, P., Thépaut, J.-N. and Hollingsworth, A., 1994: A strategy for operational implementation of 4D-Var, using an incremental approach. *Q. J. R. Meteorol. Soc.*, **120,** 1367-1387

[6] Daley, R., 1991: Atmospheric data analysis. *Cambridge University Press,* 457pp

[7] Derber, J., and F. Bouttier, 1999: A reformulation of the background error covariance in the ECMWF global data assimilation system. *Tellus.*, **51A,** 195-221

[8] Desroziers, G., G. Hello and J-N. Thépaut., 2003: A 4D-Var re-analysis of FASTEX. *Q. J. R. Meteorol. Soc.*, **129,** 1301-1315

[9] Gauthier, P. and Thépaut, J.-N., 2001: Impact of the digital filter as a weak constraint in the preoperational 4DVAR assimilation system of Météo-France. *Mon. Weather Rev.*, **129,** 2089-2102

[10] Huang, X-Y., 1999: A generalization of using an adjoint model in intermittent data assimilation systems. *Mon. Wea. Rev.*, **127,** 766-787

[11] Janisková, M., Thépaut, J.-N. and Geleyn, J. F., 1999: Simplified and regular physical parameterization for incremental Four-dimensional variational assimilation. *Mon. Wea. Rev.*, **127,** 26-44

[12] Järvinen, H., Andersson, E., and F. Bouttier, 1999: Variational assimilation of time sequences of surface observations with serially correlated errors. *Tellus.*, **51A,** 469-488

[13] Kalman, R.E., 2000: A new approach to linear filtering and prediction problems. *Trans. AMSE, Ser. D, J. Basic. Eng.*, **82,** 35-45

[14] Klinker, E., Rabier, F., Kelly G. and Mahfouf,J.-F., 2000: The ECMWF operational implementation of four-dimensional variational assimilation III: experimental results and diagnostics with operational configuration. *Q. J. R. Meteorol. Soc.*, **126,** 1191-1215

[15] Le Dimet, F.X., and Talagrand, O., 1986: Variational algorithms for analysis and assimilation of meteorological observation: theoretical aspects. *Tellus.*, **38A,** 97-110

[16] Lewis, J.M. and Derber, J. C.,: 1985: The use of adjoint equations to solve a variational adjustment problem with advective constraints. *Tellus,* **37A,** 307-322

[17] Liu,Z.-Q. and Rabier,F., 2002: The interaction between model resolution and observation resolution and density in data assimilation: A one-dimensional study. *Q. J. R. Meteorol. Soc.*, **128,** 1367-1386

[18] Lorenc, A. C., 1981: A global three-dimensional multivariate statistical interpolation scheme. *Mon. Wea. Rev.*, **109,** 701-721

[19] Lorenc, A. C., 1986: Analysis method for numerical weather prediction. *Q. J. R. Meteorol. Soc.*, **112,** 1177-1194

[20] Lorenc, A. C., and P. Hammon, 1988: Objective quality control of observations using Bayesian methods. Theory and a pratical omplementation. *Q. J. R. Meteorol. Soc.*, **114,** 515-543

[21] Mahfouf,J.-F. and Rabier, F., 2000: The ECMWF operational implementation of four-dimensional variational assimilation II: experimental results with improved physics. *Q. J. R. Meteorol. Soc.*, **126,** 1171-1190

[22] Ménard, R., 1994: Kalman filtering of Burgers' equation and its application to atmospheric data assimilation. *Ph. D. Thesis, McGill University. Stormy Weather Group scientic report*, **MW-100,** 211pp

[23] Orszag, S. A., 1970: Transform method for calculation of vector coupled sums: application to the spectral form of the vorticity equation. *J. Atmos. Sci.*, **27,** 890-895

[24] Parrish, F. and Derber, J. C., 1992: The national meteorological center's spectral statistical-interpolation analysis system. *Mon. Wea. Rev.*, **120,** 1747-1763

[25] Rabier, F., Courtier, P., Pailleux, J., Talagrand, O. and Vasiljevic, D., 1993: A comparison between four-dimensional variational assimilation and simplified sequential assimilation relying on the three-dimensional variational analysis. *Q. J. R. Meteorol. Soc.*, **119,** 845-880

[26] Rabier, F., J-N. Thépaut, and P. Courtier 1998: Extended assimilation and forecast experiments with a four-dimensional assimilation system. *Q. J. R. Meteorol. Soc.*, **124,** 1861-1887

[27] Rabier, F, Jarvinen, H., Klinker, E., Mahfouf,J.-F. and Simmons, A., 2000: The ECMWF operational implementation of four-dimensional variational assimilation Part I: experimental results with simplified physics. *Q. J. R. Meteorol. Soc.*, **126,** 1143-1170

[28] Sasaki Y., 1970: Some basic formulation in numerical variational analysis. *Mon. Wea. Rev.*, **98,** 875-883

[29] Talagrand, O. and Courtier, P., 1987: Variational assimilation of meteorological observations with the adjoint vorticity equation. I: Theory. *Q. J. R. Meteorol. Soc.*, **113,** 1311-1328

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

[31] Thépaut, J.-N. and Moll, P., 2000: L'assimilation variationnelle 4D opérationnelle à Météo-France. *Atelier de modélisation de l'atmosphère.*, **29 et 30 novembre 2000**

[32] Todling, R. and Cohn, S.E., 1994: Suboptimal schemes for atmospheric data assimilation based on the Kalman filter. *Mon. Wea. Rev.*, **122,** 2530-2557

[33] Veersé, F. and Thépaut, J.-N., 1998: Multiple-truncation incremental approach for four-dimensional variational data assimilation. *Q. J. R. Meteorol. Soc.*, **124,** 1889-1908

[34] Zou, X., and Y-H. Kuo, 1996: Rainfall assimilation through an optimal control of initial and boundary conditions in a limited-area mesoscale model. *Mon. Wea. Rev.*, **124,** 2589-2882

[35] Zupanski, M., 1993: Regional four-dimensional variational data assimilation in a quasi-operational forecasting environment. *Mon. Wea. Rev.*, **121,** 2396-2408