Assimilation Techniques (4): 4dVar
April 2001

By **Mike Fisher**

European Centre for Medium-Range Weather Forecasts.

Table of contents

1. Introduction

This lecture has two aims. The first aim is to describe the differences between the ECMWF 4dVar and 3dVar analysis systems, and to summarize the main features of the current ECMWF operational configuration of 4dVar. (The reader is referred to Rabier *et al.* 2000, Mahfouf and Rabier 2000, and Klinker *et al.* 2000 for a more complete description of the ECMWF 4dVar.)

The second aim is to demonstrate the way in which 4dVar responds to observational data by considering the increments generated by a single isolated observation.

2. Comparison between the ECMWF 3dVar and 4dVar systems

The ECMWF 4dVar system is, from the technical point of view, very similar to the 3dVar system. Both systems use the same unix scripts, and share much of the Fortran code. The same background error covariance matrix and observation operators are used, and most of the peripheral tasks such as fetching and archiving of fields and observations are the same for the two analysis systems.

The main differences between the ECMWF 3dVar and 4dVar analyses (apart from the obvious difference that 4dVar includes integrations of the tangent linear and adjoint models during the minimization) are as follows:

- In 4dVar we perform two incremental updates. The tangent linear and adjoint models used during the first update include only very simple parameterizations of physical processes (Buizza 1994). A more complete package of physical parameterizations (Mahfouf and Rabier, 2000) is included for the second update. However, the parameterizations are computationally expensive. To reduce computational expense, the second update carries out fewer iterations of minimization than the first.
- For historical reasons, 3dVar approximates the tangent linear observation operators using a finite difference:

$$H'(\delta x) \approx H(Sx_b + \delta x) - H(Sx_b) \tag{1}$$

whereas 4dVar uses the true tangent linear operators.

- In 3dVar, all observations for the 6-hour window centred on the analysis time are collected together and are compared with the high-resolution trajectory at the nominal analysis time. The incremental cost function is:

$$\begin{aligned} J(\delta x_n) &= \frac{1}{2}(Sx_{n-1} + \delta x_n - Sx_b)^T B^{-1}(Sx_{n-1} + \delta x_n - Sx_b) \\ &+ \frac{1}{2}(H(x_{n-1}) + H'\delta x_n - y)^T R^{-1}(H(x_{n-1}) + H'\delta x_n - y) \end{aligned} \tag{2}$$

In 4dVar, the observations are divided into 1-hour timeslots and compared with the trajectory at the appropriate time:

$$\begin{aligned} J(\delta x_n) &= \frac{1}{2}(Sx_{n-1}(t_0) + \delta x_n(t_0) - Sx_b)^T B^{-1}(Sx_{n-1}(t_0) + \delta x_n(t_0) - Sx_b) \\ &+ \frac{1}{2}\sum_i (H(x_{n-1}(t_i)) + H'\delta x_n(t_i) - y_i)^T R^{-1}(H(x_{n-1}(t_i)) + H'\delta x_n(t_i) - y_i) \end{aligned} \tag{3}$$

3. The current operational configuration of 4dVar

The current ECMWF operational 4dVar assimilation system is as follows:

- Outer resolution: T511 L60.
- Resolution of increments: T159 L60
- Two incremental updates:
  - First update: 50 iterations with very simple physical parameterizations.
  - Second update: 25 iterations with comprehensive physics.
- Physical parameterizations during the second update:
  - Vertical diffusion.
  - Sub-grid scale orographic drag.
  - Large scale precipitation.
  - Longwave radiation.
  - Deep moist convection.
- Separate analyses of some surface quantities (sea ice, sea surface termperature, soil wetness, snow, 2m temperature and 2m humidity).

4. Increments from a single observation

In my previous lecture, I demonstrated that for the simple case of a single observation of a model variable at a gridpoint, the analysis increment in 3dVar is proportional to a column of the background error covariance matrix. It is instructive to consider how 4dVar responds to the same observation.

As before, consider the non-incremental formulation, and suppose that the observation is at the gridpoint corresponding to the k^{th} element of the analysis vector. In 4dVar, we must also specify the time of the observation. Denote this by t_i.

The 4dVar cost function is:

$$\mathrm{J}(\mathrm{x}(t_0)) \;=\; \frac{1}{2}(\mathrm{x}(t_0)-\mathrm{x}_b)^{\mathrm{T}}\mathrm{B}^{-1}(\mathrm{x}(t_0)-\mathrm{x}_b)+\frac{1}{2}(\mathrm{H}\mathrm{x}(t_i)-\mathrm{y})^{\mathrm{T}}\mathrm{R}^{-1}(\mathrm{H}\mathrm{x}(t_i)-\mathrm{y}) \tag{4}$$

Note that the cost function is regarded as a function of $\mathrm{x}(t_0)$, whereas the observation is compared to the gridpoint value at time t_i. We can eliminate $\mathrm{x}(t_i)$ by noting that $\mathrm{x}(t_i)$ is the result of a model integration with initial conditions $\mathrm{x}(t_0)$. Let us write this as:

$$\mathrm{x}(t_i) \;=\; M_{t_0\to t_i}(\mathrm{x}(t_0)) \tag{5}$$

The cost function is then:

$$\mathrm{J}(\mathrm{x}(t_0)) \;=\; \frac{1}{2}(\mathrm{x}(t_0)-\mathrm{x}_b)^{\mathrm{T}}\mathrm{B}^{-1}(\mathrm{x}(t_0)-\mathrm{x}_b)+\frac{1}{2}(\mathrm{H}M_{t_0\to t_i}(\mathrm{x}(t_0))-\mathrm{y})^{\mathrm{T}}\mathrm{R}^{-1}(\mathrm{H}M_{t_0\to t_i}(\mathrm{x}(t_0))-\mathrm{y}) \tag{6}$$

The gradient of the cost function for the analysis is zero, giving (if we ignore 2^{nd} order and higher derivatives):

$$\mathrm{B}^{-1}(\mathrm{x}_a(t_0)-\mathrm{x}_b)+\mathrm{M}^{\mathrm{T}}_{t_0\to t_i}\mathrm{H}^{\mathrm{T}}\mathrm{R}^{-1}(\mathrm{H}M_{t_0\to t_i}(\mathrm{x}_a(t_0))-\mathrm{y}) \;=\; 0 \tag{7}$$

where $\mathrm{M}^{\mathrm{T}}_{t_0\to t_i}$ represents the adjoint of the model integration from time t_0 to t_i.

Multiplying through by B and rearranging gives, as for the 3dVar example, an expression for the analysis increment (at the start of the 4dVar assimilation window, time t_0). Since we have just one observation, the expression $\mathrm{R}^{-1}(\mathrm{H}M_{t_0\to t_i}\mathrm{x}_a(t_0)-\mathrm{y})$ is simply a scalar, and we find that, whereas in 3dVar, the analysis increment was proportional to a column of B, in 4dVar, it is proportional to a column of $\mathrm{B}\mathrm{M}^{\mathrm{T}}_{t_0\to t_i}$:

$$\mathrm{x}_a(t_0)-\mathrm{x}_b \;=\; \left(\frac{\mathrm{y}-\mathrm{x}_a(t_i)_k}{\sigma_o^2}\right)\begin{bmatrix}(\mathrm{B}\mathrm{M}^{\mathrm{T}}_{t_0\to t_i})_{1k}\\(\mathrm{B}\mathrm{M}^{\mathrm{T}}_{t_0\to t_i})_{2k}\\\cdots\\(\mathrm{B}\mathrm{M}^{\mathrm{T}}_{t_0\to t_i})_{Nk}\end{bmatrix} \tag{8}$$

A somewhat more interesting equation results if we multiply both sides of Eq. (7) to the left by $\mathrm{M}_{t_0\to t_i}\mathrm{B}$ before rearranging:

$$\mathrm{M}_{t_0\to t_i}(\mathrm{x}_a(t_0)-\mathrm{x}_b) \;=\; \left(\frac{\mathrm{y}-\mathrm{x}_a(t_i)_k}{\sigma_o^2}\right)\begin{bmatrix}(\mathrm{M}_{t_0\to t_i}\mathrm{B}\mathrm{M}^{\mathrm{T}}_{t_0\to t_i})_{1k}\\(\mathrm{M}_{t_0\to t_i}\mathrm{B}\mathrm{M}^{\mathrm{T}}_{t_0\to t_i})_{2k}\\\cdots\\(\mathrm{M}_{t_0\to t_i}\mathrm{B}\mathrm{M}^{\mathrm{T}}_{t_0\to t_i})_{Nk}\end{bmatrix} \tag{9}$$

In this case, we note that the expression on the left hand side is:

$$\mathrm{M}_{t_0\to t_i}(\mathrm{x}_a(t_0)-\mathrm{x}_b) \;=\; M_{t_0\to t_i}(\mathrm{x}_a(t_0))-M_{t_0\to t_i}(\mathrm{x}_b)=\; \mathrm{x}_a(t_i)-M_{t_0\to t_i}(\mathrm{x}_b) \tag{10}$$

(Remember, we are ignoring second order and higher derivatives.)

So, the left hand side of equation 9 is the difference between the analysis and the forecast for time t_i with initial conditions given by the background at time t_0. In other words, the left hand side of Eq. (9) is the difference between the analysis trajectory and the background trajectory at the time of the observations. This difference is proportional to a column of $\mathrm{M}_{t_0 \to t_i} \mathrm{B} \mathrm{M}^{\mathrm{T}}_{t_0 \to t_i}$.

Now, if B is the covariance matrix for errors in the background at time t_0, then the matrix $\mathrm{M}_{t_0 \to t_i} \mathrm{B} \mathrm{M}^{\mathrm{T}}_{t_0 \to t_i}$ is the covariance matrix for errors in a forecast from time t_0 to t_i with initial conditions equal to the background. This is easy to see:

Under the tangent linear assumption, and for a perfect model, the background errors at time t_0 and t_i are related by:

$$\varepsilon_b(t_0) = \mathrm{M}_{t_0 \to t_i} \varepsilon_b(t_i) \tag{11}$$

So, the covariance matrix for background errors at time t_i is:

$$\overline{\varepsilon_b(t_i)(\varepsilon_b(t_i))^{\mathrm{T}}} = \mathrm{M}_{t_0 \to t_i} \overline{\varepsilon_b(t_0)(\varepsilon_b(t_0))^{\mathrm{T}}} \mathrm{M}^{\mathrm{T}}_{t_0 \to t_i} = \mathrm{M}_{t_0 \to t_i} \mathrm{B} \mathrm{M}^{\mathrm{T}}_{t_0 \to t_i} \tag{12}$$

To summarize. In 4dVar, the analysis increment at the time of the observation is given by the column of the *evolved* covariance matrix. This matrix describes errors in the background trajectory at the time of the observation. The covariance matrix is *implicitly* evolved by 4dVar using the dynamics of the tangent linear model. As a consequence, both the covariance matrix at the observation time and the analysis increments are *flow-dependent*.

For an observation at the beginning of the 4dVar assimilation window (i.e. for $t_i = t_0$) the matrix $\mathrm{M}_{t_0 \to t_i}$ is the identity matrix. (Integration of the tangent linear model for no timesteps does nothing!) In this case, equation 9 is identical to the 3dVar case. *The analysis increments for an observation at the beginning of the 4dVar assimilation window are the same as would be produced in 3dVar.* This illustrates the main shortcoming of 4dVar. Namely, that at each cycle of assimilation the initial covariance matrix is the fixed, static and flow-independent matrix B. The flow-dependent covariances which are used implicitly during the assimilation are not propagated to the next cycle. To propagate these covariances, we must turn to the Kalman filter. This will be the subject of my next lecture.

4.1 Examples

Fig. 1 shows analysis increments for 3 separate analyses for the same date. Each analysis had just a single observation of geopotential at 850hPa, 40N, 60W. In Fig. 1 (b), the observation was placed at the beginning of the 4dVar assimilation window, and a cross-section of the analysis increment at the beginning of the assimilation window is shown. The analysis increment in this case is determined entirely by the background error covariance matrix B, and is clearly almost symmetric and without any vertical tilt. (The slight asymmetry is probably due to the effects of normal model initialization of the increments.) The increment is the same as would be generated by 3dVar for this observation. By contrast. if the same observation is placed in the middle of the analysis window (Fig. 1 (c)) then the analysis increment, also in the middle of the window, shows a marked vertical tilt towards the jet. (Fig. 1 (a) shows a cross section of the background zonal wind.) The tilt is more marked if the observation is placed at the end of the assimilation window. Fig. 1 (d) shows the increment at the end of the analysis window in this case.

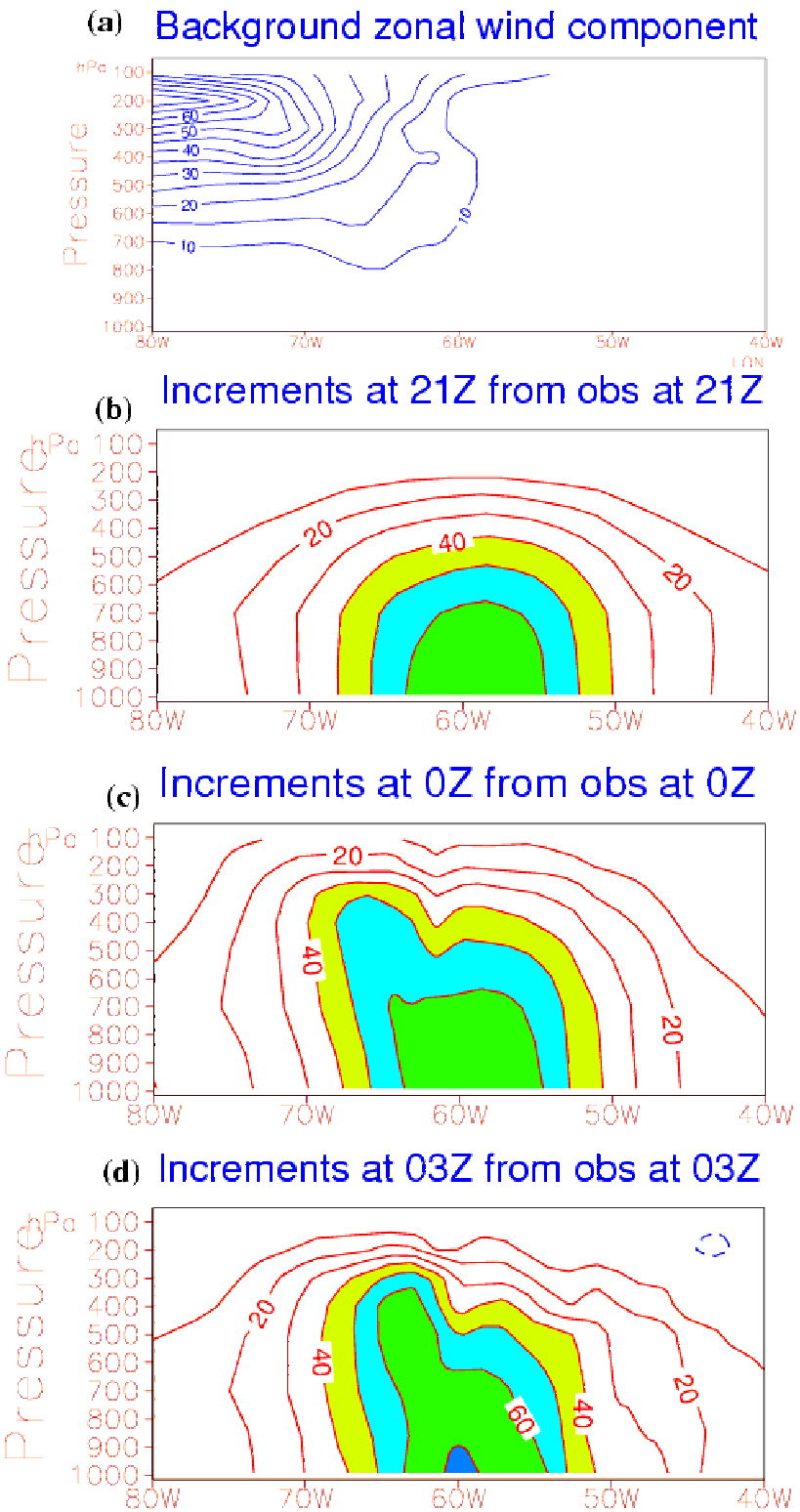

Figure 1. (a): Background zonal wind cross section. (b), (c) and (d): Increments at the beginning, middle and end of the 4dVar assimilation window for observations of 850hPa height at 40N, 60W at the beginning, middle and end of the window respectively.

The increments in Fig. 1 are plotted for the same time as the observations. Fig. 2 shows increments at the middle of the assimilation window (i.e. at the nominal analysis time, 0z) for single observations at the beginning, middle and end of the window. (Again, three separate analyses are shown. Each analysis used a single observation.) This illustrates that 4dVar spreads the infomation provided by the observations in a dynamically consistent way throughout the analysis window.

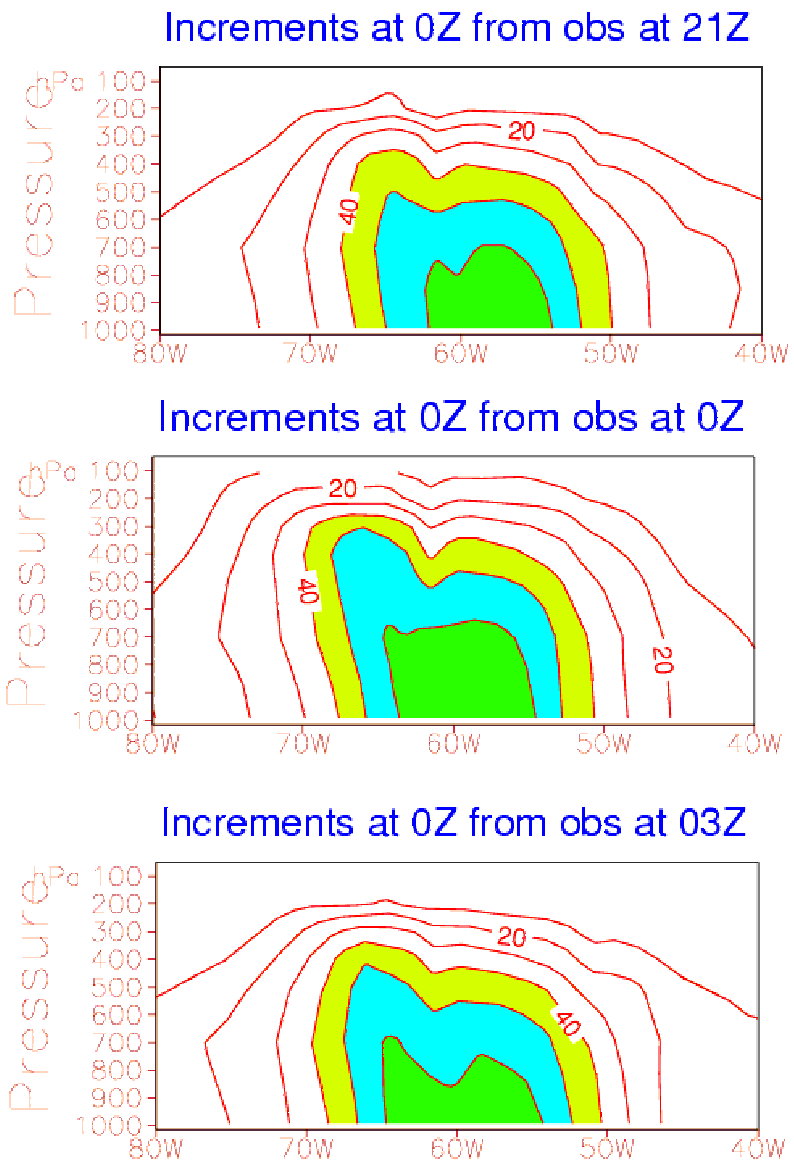

Figure 2. Geopotential analysis increments at the nominal analysis time, 0z, generated by observations of geopotential at 850hPa, 40N, 60W at the beginning, middle and end of the 4dVar assimilation window.

5. A CAUTIONARY EXAMPLE

In view of the results presented above, it is tempting to ascribe the superior performance of 4dVar (relative to 3dVar) to the dynamical propagation of the covariance matrix. In this section, a simple example will be presented

to show that *there is more to 4dVar than covariance propagation*.

Consider a 4dVar system with the following characteristics:

- M is an orthogonal matrix. (That is $\mathrm{M}^{-1} = \mathrm{M}^{T}$.)
- $\mathrm{B} = \sigma_b^2 \mathrm{I}$
- $\mathrm{R} = \sigma_o^2 \mathrm{I}$
- The entire state vector is observed at some single time t, so that $\mathrm{H} = \mathrm{I}$.

Now, the covariance matrix of analysis error at the beginning of the assimilation window ofa 4dVar system is the inverse of the Hessian matrix of the cost function (see, for example Rabier and Courtier, 1992).

For observations at a single time, a 4dVar analysis has:

$$\mathrm{P}_{4D}^{a}(t_0) = (\mathrm{B}^{-1} + \mathrm{M}^{T}\mathrm{H}^{T}\mathrm{R}^{-1}\mathrm{HM})^{-1} \tag{13}$$

Now, for our simple example, we have:

$$\mathrm{M}^{T}\mathrm{H}^{T}\mathrm{R}^{-1}\mathrm{HM} = \frac{1}{\sigma_o^2}\mathrm{M}^{T}\mathrm{M} = \frac{1}{\sigma_o^2}\mathrm{I} = \mathrm{R}^{-1} \tag{14}$$

In other words, the covariance matrix of analysis error is:

$$\mathrm{P}_{4D}^{a}(t_0) = (\mathrm{B}^{-1} + \mathrm{R}^{-1})^{-1} = \frac{1}{\sigma_b^2 + \sigma_o^2}\mathrm{I} \tag{15}$$

This is *identical* to the covariance matrix $\mathrm{P}_{3D}^{a}(t_0)$ of analysis error implied by the corresponding 3dVar analysis. At later times during the analysis window, the analysis error covariance matrix is dynamically propagated:

$$\mathrm{P}_{4D}^{a}(t) = \mathrm{M}^{T}(\mathrm{P}_{4D}^{a}(t_0))\mathrm{M} \tag{16}$$

However, since the initial covariance matrix is proportional to the identity matrix, it commutes with M, and since $\mathrm{M}^{T}\mathrm{M} = \mathrm{I}$ we find that the covariance matrix is constant throughout the analysis window, and equal to the corresponding 3dVar matrix. By the same argument, the covariance matrix of background error is also constant throughout the analysis window.

For the particular system described above, there is no covariance propagation. The covariance matrices of both analysis and background error are identical to those of the corresponding 3dVar system. It might be imagined that the 3dVar and 4dVar analyses are identical. This is not the case. Moreover, 4dVar is superior to 3dVar. To see this, let us rewrite the 4dVar analysis equation (assuming a linear model and linear observation operators) in the form:

$$\mathrm{x}_a(t_0) = \mathrm{x}_b(t_0) + \mathrm{P}_{4D}^{a}(t_0)\mathrm{M}_{t_0 \to t_i}^{\mathrm{T}}\mathrm{H}^{\mathrm{T}}\mathrm{R}^{-1}(\mathrm{y} - \mathrm{H}\mathcal{M}_{t_0 \to t_i}(\mathrm{x}_a(t_0))) \tag{17}$$

The corresponding 3dVar analysis equation corresponds to dropping both the model integration and the adjoint integration from Eq. (17), and replacing $\mathrm{P}_{4D}^{a}(t_0)$ by $\mathrm{P}_{3D}^{a}(t_0)$:

$$\mathrm{x}_a(t_0) = \mathrm{x}_b(t_0) + \mathrm{P}_{3D}^{a}(t_0)\mathrm{H}^{\mathrm{T}}\mathrm{R}^{-1}(\mathrm{y} - \mathrm{H}\mathrm{x}_a(t_0)) \tag{18}$$

Comparing Eqs. (17) and (18), we see that the function of the forward model integration in 4dVar is to allow comparison of the observation with the corresponding model state at the correct time. By contrast, 3dVar compares the observation, which is valid at time t, with the model state at time t_0. Similarly, the function of the adjoint integration in 4dVar is to propagate the information from the observation back in time to the beginning of the analysis window. This is missing in 3dVar.

Fig. 3 shows an example in which the state vector is a single wind vector $(u,v)^T$ and the model dynamics corresponds to rotation of the vector through an angle $\theta(t)$. The correct analysis at time t lies between the observation and the background at time t. I.e., the analysis should increase the wind speed, but should not alter its direction.

The 3dVar and 4dVar analyses at time t_0 are also shown. 4dVar is optimal. Also shown is a so-called 3dFGAT analysis, which compares the observation and background at the correct time, but does not propagate the increment back in time to t_0. 3dFGAT is superior to 3dVar, but is not optimal. The ECMWF 40-year reanalysis project uses a 3dFGAT system.

It is worth reiterating that for this simple example, the differences between 4dVar, 3dFGAT and 3dVar are not due to covariance propagation, since all three systems have the same covarince matrices of background and analysis error. Rather, the differences are due to the different ways in which the model state is propagated to the time of the observation, and the increment is propagated to the time of the analysis. Of course, in a real 4dVar analysis it is likely that covariance propagation plays at least some part in explaining the differences between 3dVar and 4dVar.

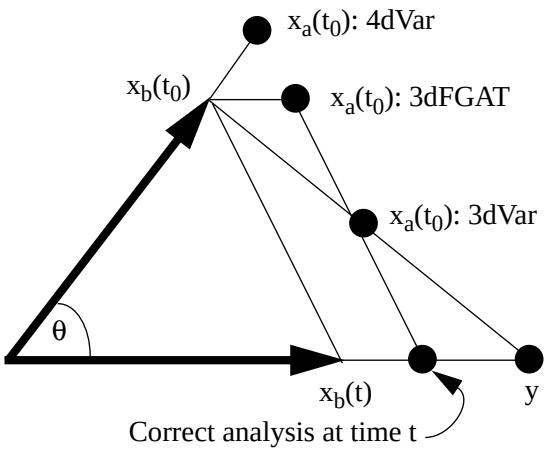

Figure 3. An simple example, showing the differences between 4dVar, 3dVar and 3dFGAT for an idealized system.

REFERENCES

Rabier F. and P. Courtier, 1992, Four dimensional assimilation in the presence of baroclinic instability. Q. J. R. Met. Soc., Vol. 118, pp649-672.

Rabier F., H. Järvinen, E. Klinker, J.-F. Mahfouf and A. Simmons, 2000, The ECMWF operational implementation of four-dimensional variational assimilation. I: Experimental results with simplified physics. Q. J. R. Met. Soc., Vol. 126, No. 564, pp 1143.

Mahfouf J.-F. and F. Rabier, 2000, The ECMWF operational implementation of four-dimensional variational assimilation. II: Experimental results with improved physics. Q. J. R. Met. Soc., Vol. 126, No. 564, pp 1171.

Klinker E., F. Rabier, G. Kelly, and J.-F. Mahfouf, 2000, The ECMWF operational implementation of four-dimensional variational assimilation. III: Experimental results and diagnostics with operational configuration. Q. J. R. Met. Soc., Vol. 126, No. 564, pp 1191.

Buizza R., 1994, Sensitivity of optimal unstable structures. Q. J. R. Met. Soc., Vol. 120, pp429-451.