TECHNICAL MEMORANDUM

399

Diagnostics of linear and incremental approximations in 4D-Var

Yannick Trémolet

Research Department

February 2003

Submitted to Q. J. R. Meteorol. Soc.

European Centre for Medium-Range Weather Forecasts
Europäisches Zentrum für mittelfristige Wettervorhersage
Centre européen pour les prévisions météorologiques à moyen terme

For additional copies please contact

The Library
ECMWF
Shinfield Park
Reading
RG2 9AX
library@ecmwf.int

Series: ECMWF Technical Memoranda

A full list of ECMWF Publications can be found on our web site under:
http://www.ecmwf.int/publications/

Abstract

The validity of the tangent linear approximation and of the incremental 4D-Var formulation is evaluated in the operational context. In ECMWF's operational system, the linear and adjoint models are run at a lower resolution than the nonlinear model. Furthermore, the physics is simpler in the tangent linear and adjoint than in the nonlinear model. Comparisons are made between the output of the linear model and the finite difference obtained by running the nonlinear model twice, with and without adding the analysis increment. The accuracy of the linear model is assessed with respect to resolution, linearised physics and the length of the assimilation window.

The most striking results are that linearisation errors are larger than expected and that large errors appear very early in the assimilation window. It is also shown that higher resolution 4D-Var will require more accurate linear physics than currently available. A modification of the computation of the trajectory around which the problem is linearised is shown to improve the accuracy of the linearisation.

The results and diagnostic tools presented should provide guidance for further developments of the incremental 4D-Var, with respect to resolution changes and improvements in linearised physics and dynamics.

1 Introduction

Meteorological forecasts are based on observations of the atmosphere and on models of the evolution of the atmospheric flow. In order to integrate a model and produce a forecast, an initial condition which describes the atmosphere at the initial time of the forecast is required. Observations of the atmosphere do not constitute a satisfactory initial condition because of their irregular distribution in time and space and because of the measurement errors. The data assimilation problem consists of constructing a suitable initial condition using the observations and the model. In this paper, we will focus on the four-dimensional variational data assimilation (4D-Var) method as described by LeDimet and Talagrand (1986). The principle of the method is to minimise a cost function which measures the gap between observations over a given period of time, and the solution of the model during the same period. This technique takes into account the distribution of data both in time and space. To perform the minimisation, the gradient of the cost function with respect to the initial condition is obtained using the adjoint model.

The 4D-Var assimilation system implemented at ECMWF is described by Rabier et al. (2000), Mahfouf and Rabier (2000) and Klinker et al. (2000). Because of the computational cost of 4D-Var, some approximations are made. In particular, the ECMWF implementation is based on the incremental formulation described by Courtier et al. (1994). The minimisation is performed at lower resolution than the main forecast, with simpler physics, and using a cost function evaluated by integrating the tangent linear model rather than the full nonlinear model.

There have been several studies discussing the validity of the linear assumption in weather forecasting, such as those by Errico et al. (1993), Vukicevic and Errico (1993), Mahfouf (1999), Janisková et al. (1999), Errico and Raeder (1999) or Gilmour et al. (2001). As summarised by Pires et al. (1996), the general conclusion is that for large scale flows, the linear approximation is valid for periods of two to three days and for mesoscale flows for periods of the order of 36 hours. In these studies, the authors consider small perturbations or in some cases perturbations of the magnitude of analysis error and the resolutions they use do not match today's operational forecast resolution. Typically, studies are done with a T63 model, using perturbations with a maximum amplitude of 1K for temperature and 1 m/s for winds with the exception of Mahfouf (1999) and Janisková et al. (1999) who use analysis increments.

In this study, we evaluate the linear assumption in the conditions in which it is currently used in incremental 4D-Var: the perturbations we consider are analysis increments and their low resolution linear evolution will be evaluated with respect to the operational high resolution forecast model.

It is expected that in the near future higher density data will be available and a more accurate assimilation system will be necessary to extract all the potential information from it. In the same way, new types of data are becoming available, such as rain and cloud measurements, which will require more accurate representation of the associated phenomena i.e. more physics to be included in the assimilation. Finally, in order for 4D-Var to better take into account dynamical aspects of the atmospheric flow, it would be beneficial to increase the length of the assimilation window. However, this requires the linear approximation to be valid for the whole length of the assimilation period. As well as diagnosing the current system, we also aim to evaluate the possibility for future developments. We will therefore study the impact of higher resolution minimisation (T255), physical processes included in the linear model, and the length of the assimilation window.

The outline of the paper is as follows: section 2 describes incremental 4D-Var and the method to evaluate the system. It gives an outline of the 4D-Var algorithm, based on *inner-loop* iterations with a low resolution linear model embedded within *outer* iterations with the full nonlinear model. Section 3 investigates the error introduced in the initial condition between the inner and outer loops while section 4 presents several aspects of the errors in the linear propagation of increments in the inner loop of 4D-Var. The following section presents recent modifications introduced at ECMWF to improve the consistency between 4D-Var inner and outer loops. Section 6 provides a linearity study of the 4D-Var problems which sheds some light on whether it can be solved accurately by a linear approximation. Although results are presented for the ECMWF assimilation system, some conclusions apply to any high resolution 4D-Var which resolves similar scales.

2 Experimental framework

2.1 Incremental 4D-Var

4D-Var consists in minimising the discrepancy between observations of the atmosphere and a forecast over a period of time called the assimilation window. The control variable of the problem is the initial condition of the model. The cost function which is minimised includes three terms and can be written as:

$$J(x) = (x - x_b)^T B^{-1}(x - x_b) + (H(x) - y)^T R^{-1}(H(x) - y) + J_c$$

where x is the control variable, x_b is the background state, y is the vector of observations, B is the background error covariance matrix, R is the observation error covariance matrix, H is the nonlinear observation operator and J_c is an initialisation term used to control gravity waves. The J_c term will be omitted in the remaining of this paper for simplicity. H computes the observation equivalent at the correct location and time and includes the forecast model.

In its incremental formulation, the minimisation problem is written as a function of the departure from the background $\delta x = x - x_b$. At the minimum, δx will be the analysis increment. A first order approximation of the cost function is given by:

$$J(\delta x) = \delta x^T B^{-1} \delta x + (H\delta x - d)^T R^{-1}(H\delta x - d)$$

where $H = \frac{\partial H}{\partial x}$ is the linearised observation operator and $d = y - H(x_b)$ is the departure from observations. In this notation, the tangent linear model is embedded in the linearised observation operator.

The minimisation problem is solved using an iterative algorithm (conjugate gradient or quasi-Newton algorithms). This is the inner loop of 4D-Var. In order to reduce the computational cost of the assimilation, the inner loop is run at lower resolution than the forecast. However, in order to retain the maximum information from the observations, the departures d are computed at high resolution. The starting point for the minimisation (first guess) is interpolated to the inner loop resolution using an operator S. After the minimisation, the departures are recomputed at high resolution and the process is repeated. This is the outer loop of incremental

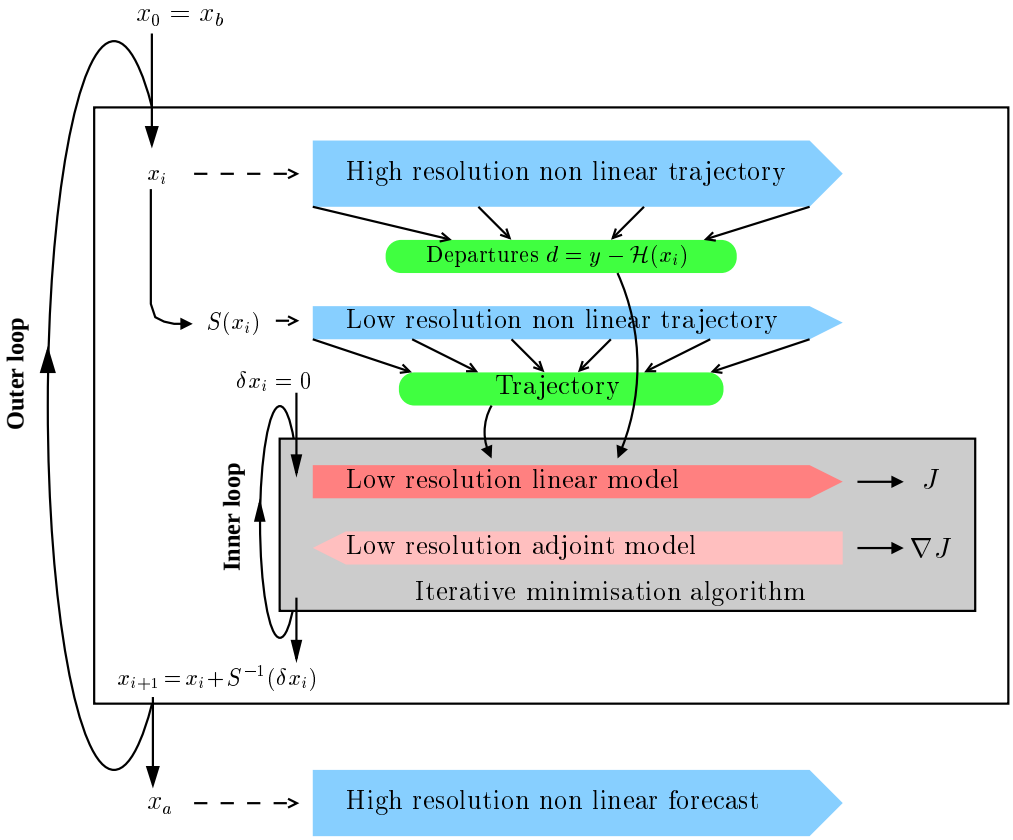

Figure 1: Incremental 4D-Var algorithm: departures from observations d are computed at high resolution, the first guess is interpolated to low resolution using the operator S and the nonlinear trajectory is computed from that state, the cost function is minimised at low resolution using an iterative algorithm (inner loop), the resulting increment δx_i is interpolated back to high resolution (S^{-1}) and added to the first guess, the process is repeated (outer loop, subscript i, currently two iterations at ECMWF) until the analysis is obtained.

4D-Var. The high resolution nonlinear runs also define the state around which the observation operator is linearised and are called trajectory runs. In the current setup, the low resolution nonlinear model is also integrated to provide the trajectory around which the tangent linear and adjoint models are linearised. The incremental 4D-Var algorithm is shown schematically in figure 1. Currently at ECMWF, two iterations of the outer loop are run. For further reduction of the computational cost, the linear physics is omitted in the first inner-loop minimisation.

Because of this setup, it is possible for the fit between an observation and the high and low resolution states to be different. In some cases, this difference can be large enough for the two departures from a given observation to be of the opposite sign. The increment which is computed at low resolution would then lead to a deterioration of the fit to that observation at high resolution. A phenomenon of oscillation between the inner and outer loop could appear which would prevent the proper convergence of the algorithm. Because of these errors, it was found necessary to introduce an incremental convergence check as described by Bouttier (2001): if the discrepancy between the two departures is too large then the observation is rejected. This reduces the problem of the convergence of the algorithm but could lead to the rejection of good data.

2.2 Diagnostics

Tangent linear models are usually validated by comparing the output of a linear run with the finite difference between two nonlinear runs of the corresponding model using the ratio:

$$r = \frac{M\delta x}{M(x+\delta x) - M(x)}$$

where r is evaluated for each component of the model output. According to the Taylor formula, when the size of the perturbation tends to zero, the finite difference and the linear model should behave similarly. In practice, this is limited by machine precision. The ratio between the output of the linear model and the value of the finite difference gets linearly close to 1 and then diverges when machine precision is reached. This test should be true for any perturbation and in practice a set of random perturbations are used. In order for this test to be valid, both models are run at the same resolution, with the same physical processes included and the same values for all parameters.

However, in operational data assimilation, these conditions are not satisfied. Because of computational cost, the minimisation is run at lower resolution than the forecast, the linear model does not contain all the physical processes which are present in the forecast model and, in the ECMWF system, humidity is represented in spectral form in the linear model while it is only present in grid point form in the nonlinear model. When adding the low resolution increment to the first guess, the operator S^{-1} is not in practice the inverse of S but only a pseudo-inverse: S being a truncation operator it is not invertible. Furthermore, a super saturation check is applied to the updated state vector after the increment has been added. Finally, the perturbation is not arbitrary in size or direction: it is an analysis increment.

In order to diagnose the resulting errors in data assimilation, the output of the linear model used in ECMWF 4D-Var assimilation system and the difference between two runs of the forecast model are compared in this paper. An analysis increment will be used as initial perturbation. The typical maximum amplitude of the perturbation will be of the order of 3K and 12 m/s.

4D-Var is an iterative process in which several integrations of both the linear and nonlinear models are integrated in the inner and outer loops respectively. Consequently, all the necessary information to perform a test of the linearisation is already being computed. For the experiments presented here, the output of the last integration of the linear model in a given minimisation is saved as well as the output of the two high resolution forecasts surrounding it. The difference between the two high resolution runs is then compared with the output of the linear run. This approach has two advantages: the added computational cost is negligible when data assimilation is running and most importantly, the linear and high resolution models are run exactly as used in data assimilation which could be difficult to enforce in any other way.

In the following sections, the relative error

$$r = \frac{\|M(x_i + S^{-1}\delta x_i) - M(x_i) - M_i\delta x_i\|}{\|M(x_i + S^{-1}\delta x_i) - M(x_i)\|}$$

will be presented where x_i is the first guess, δx_i the analysis increment, M the nonlinear forecast model, M_i the tangent linear model linearised around x_i and S^{-1} is the pseudo-inverse of the simplification operator as used in 4D-Var (see figure 1 for more details on the notations). The output fields are interpolated to inner loop resolution in order to compute the errors. In this study M was run at T511 resolution which is the current operational resolution at ECMWF. M is currently run at T159 resolution and this resolution will be used in the tests presented here unless noted otherwise in the text. The diagnostics presented here are globally averaged errors computed in grid-point space on the model's reduced Gaussian grid. Both models were run with the operational 60 levels vertical resolution.

3 Initial conditions

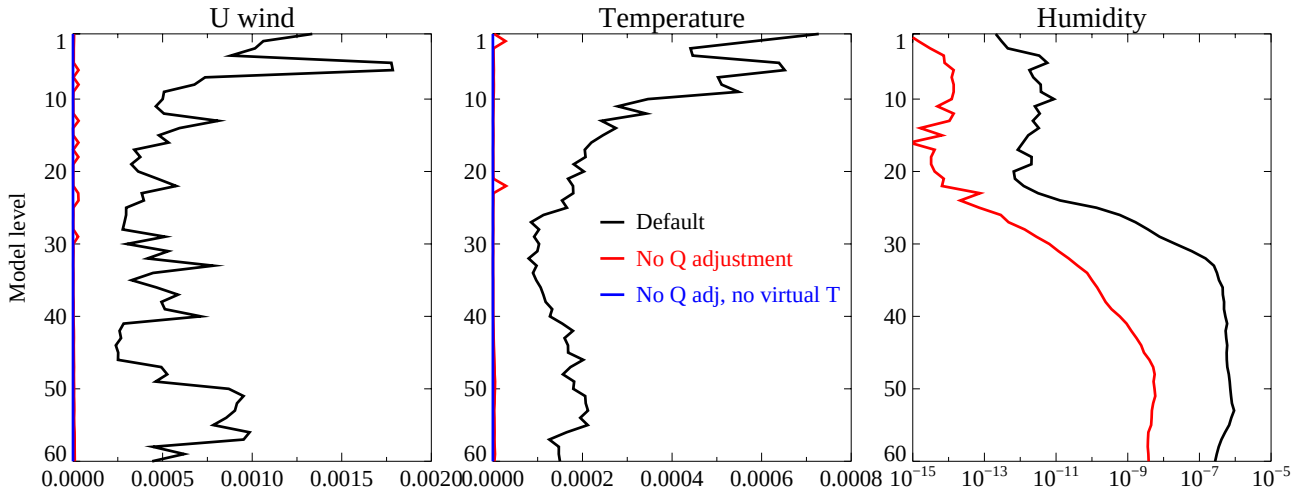

Figure 2: Average difference after 6h between successive T159 nonlinear trajectories when increments are set to zero. The blue curve for humidity is not visible because the values are smaller than 10^{-15}.

In this section, errors due to the difference between two successive nonlinear trajectory runs are assessed. It is expected that when the increments are set to zero these two runs should be identical. This was initially done purely as a sanity check before going on to the main investigation in the following section. Figure 2 shows the difference between the two nonlinear runs. The figure indicates that the difference is far from small, contrary to expectations.

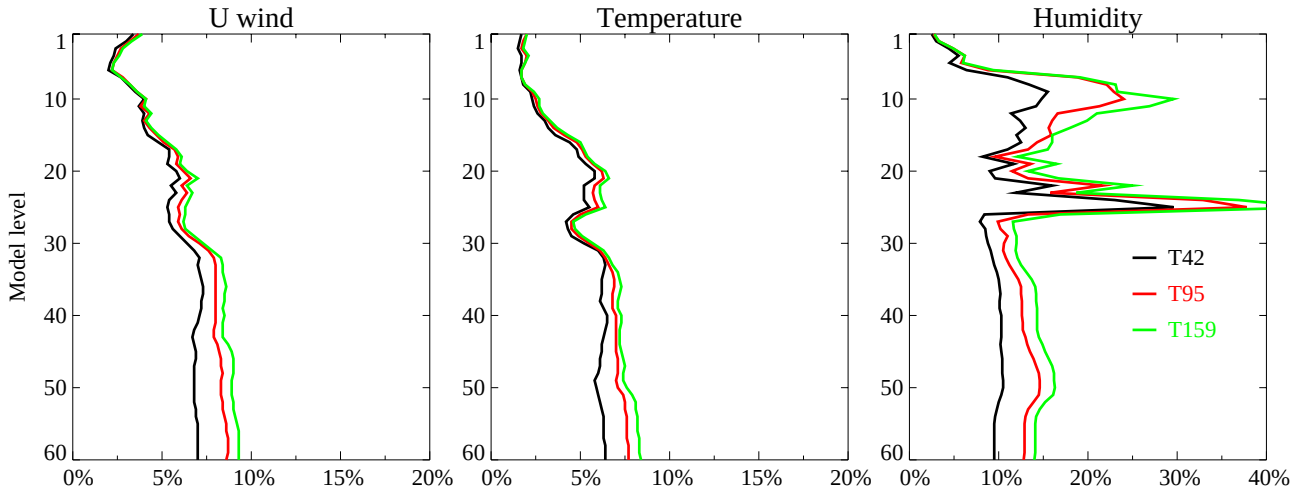

Figure 3: Relative error in the initial condition for several inner loop resolutions as indicated by the legend compared to T511 outer loop.

Several discrepancies between the model integrations were identified. The most important one is due to super-saturation removal. By default, the super-saturation check is not applied to the first trajectory but only to the subsequent ones when an increment has been added. This experiment shows that super-saturation is in fact already present in the background state even though it is not checked for or removed. Another, much smaller, error is introduced by the transformation of the temperature to virtual temperature. When the background is read into the model in a given trajectory, the temperature is immediately converted to virtual temperature. In order

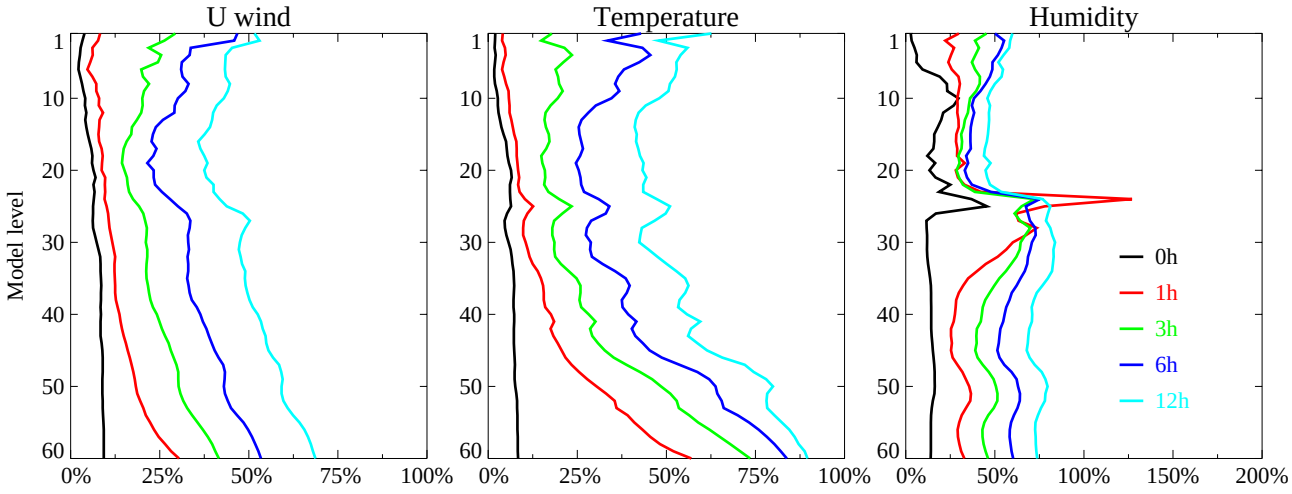

Figure 4: Evolution of the relative error in the T159 tangent linear model with respect to the T511 forecast model over the length of the assimilation window.

to add the analysis increment, virtual temperature is converted back to real temperature, then the increment is added and another conversion to virtual temperature is performed. When the increment is set to zero, the effect of this is to introduce an extra conversion to real temperature and back together with associated round-off errors. Although very small, this difference makes the trajectory not reproducible.

Switching off the humidity adjustment and suppressing the extra virtual temperature conversion makes the two trajectories exactly identical. Having established this basic result, we can now proceed with the study of linear and incremental approximations.

Figure 3 shows the error $r = \delta x_i / S^{-1} \delta x_i$ in the initial condition between the linear model and the difference between the two nonlinear runs at initial time. In this case, the error does not come from the linearisation but rather from the approximations due to our implementation of incremental 4D-Var. The first source of error is again the super-saturation check which is not applied in the linear model. The other source of error is the difference in resolution between the fields. In order to start 4D-Var, the background is needed at the two resolutions of the inner and outer loops. An interpolation has to be used for that purpose. At ECMWF, an interpolation with an adjustment to the low resolution orography is used. However, when the low resolution increment is added to the high resolution background, no adjustment back to the high resolution orography is made in the vertical. In the case of humidity which has a very strong gradient near the tropopause, the slight shift in the vertical introduces a large relative error which is the peak visible at around level 25 on figure 3.

These inconsistencies in the initial conditions for the assimilation system have values of the order of 5 to 10 percent, and cannot be ignored when analysing 4D-Var performance.

4 Errors due to incremental approximations

4.1 Time evolution of the error

In the previous section, the initial condition error has been studied. In this section, we investigate the evolution of the error as the linear and nonlinear models are integrated. Figure 4 shows how the relative error evolves through the period of the assimilation window, currently 12 hours. Three aspects are worth noticing.

First, the relative error can come close to 100% after 12 hours. This might indicate a potential limitation to the

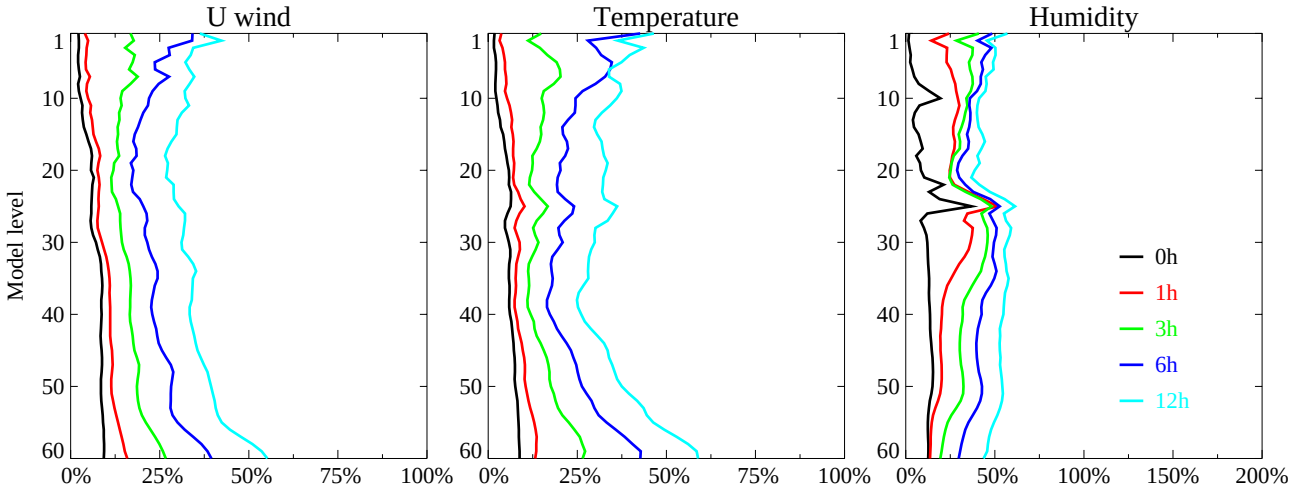

Figure 5: Same as figure 4 but for adiabatic tangent linear and forecast models.

extension of the 4D-Var window beyond 12 hours, at least with the current linearised model.

Second, the humidity field has the largest relative error. Note that the scale for humidity is different from that for other variables. This will be the case for most subsequent figures. The peak around level 25 (which corresponds roughly to the tropopause) is due to the large gradient in the humidity field at that level: a small error in the position of the tropopause will create a very large relative error. All this is true for most of the figures presented in this paper and shows the greater sensitivity of this field. Another discrepancy in the representation of humidity is the fact that it is treated as a grid point field in the forecast model and as a spectral field in the inner loop.

Finally, another large error is located near the surface and grows very rapidly. It is already more than 50% for the temperature field after only one hour. This might look erroneous but this behavior has been confirmed by other independant tests (M. Janisková, personnal communication). Possible reasons for this error could include the lack of some physical processes in the linear model, or inconsistencies in the treatment of the surface fields. Currently, the surface fields are analysed separately from the upper air fields, after the main analysis. Surface fields could be added to the 4D-Var control variable and become active fields in the linear model which would probably improve the consistency of the system near the surface as some aspects of the cost function might be better reduced by changing the surface fields rather than the upper air fields. This could be investigated in the future but would require other issues to be resolved, such as the mixture of grid point and spectral fields in the control variable or the definition of the background term for such fields.

An adiabatic test has been run where both the tangent linear and the nonlinear model were run without physics. Figure 5 shows that in this case, the large error early in the forecast almost disappears which leads us to the hypothesis that the error comes from the physical processes. The error is thus not intrinsic to the incremental implementation of 4D-Var. One will also notice that in the adiabatic case, humidity error is of the same order of magnitude as that of the other fields, which is consistent with the fact that humidity is then a passive variable.

4.2 Physics impact

In ECMWF's operational 4D-Var, two successive minimisations are run, the first one without physics in the tangent linear model to reduce the computational cost, the second one with linearised physics for better accuracy. Figure 6 shows the error with (in black) and without (in green) physics in the tangent linear model for the same increment. The error of the adiabatic linear model with respect to the adiabatic nonlinear model is also

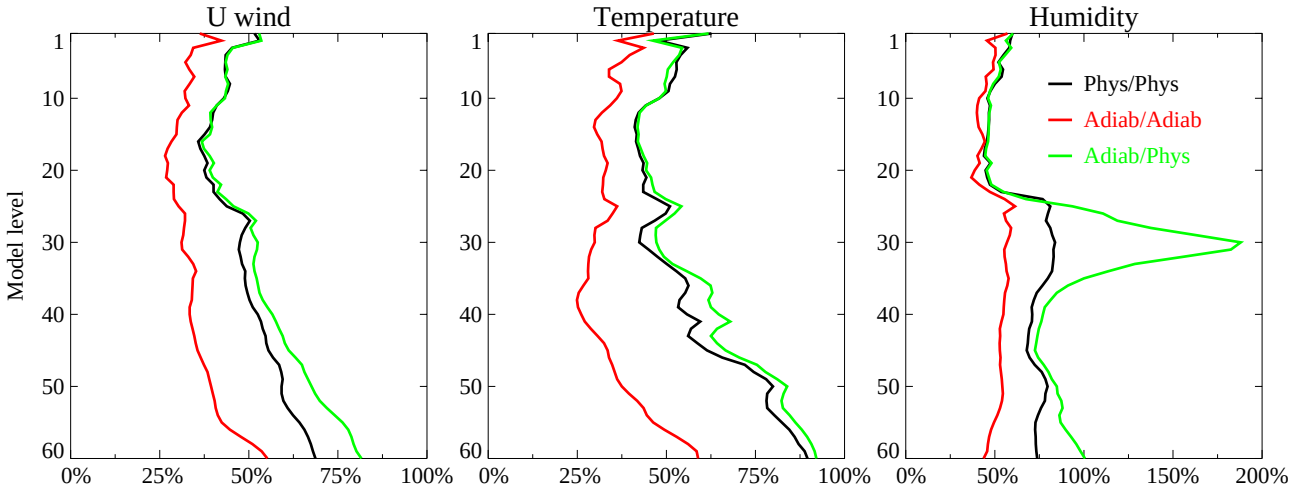

Figure 6: Relative error of the T159 linear model with physics (in black) and without physics (in green) with respect to the T511 nonlinear model and of the adiabatic linear model with respect to the adiabatic nonlinear model (in red) after 12h.

shown (in red). The largest impact of the linear physics is on the humidity field. The impact also increases at the lowest levels of the atmosphere which is where the physics is more active. However, the relative error with physics is still larger than in the fully adiabatic case.

The data assimilation scheme is currently being developed for the assimilation of cloud and precipitation observations (Hólm et al. (2002)). The observation operators for these types of data will require that the moist physical processes are resolved in the inner loop for a proper assimilation to be possible. More accurate linearised physics will therefore become even more important in the future than it already is.

4.3 Resolution impact

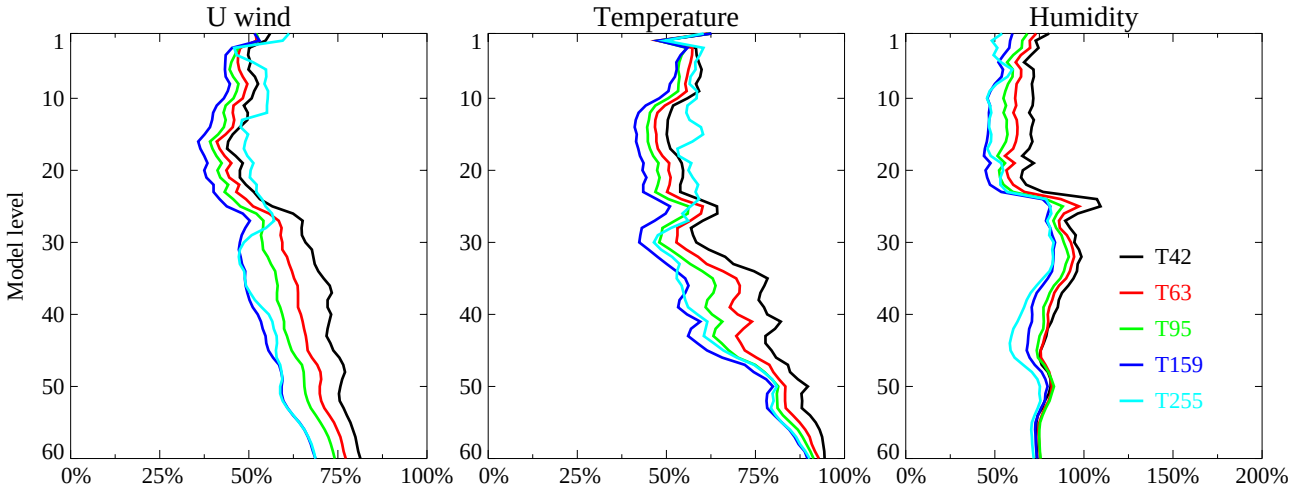

Figure 7: Relative error of the tangent linear model for various resolutions with respect to the T511 nonlinear, diabatic, model after 12h.

The need for better weather forecasts combined with the increased available computer power pushes operational centres to increase the resolution of their forecasts. At ECMWF, it is planned to increase the resolution of the

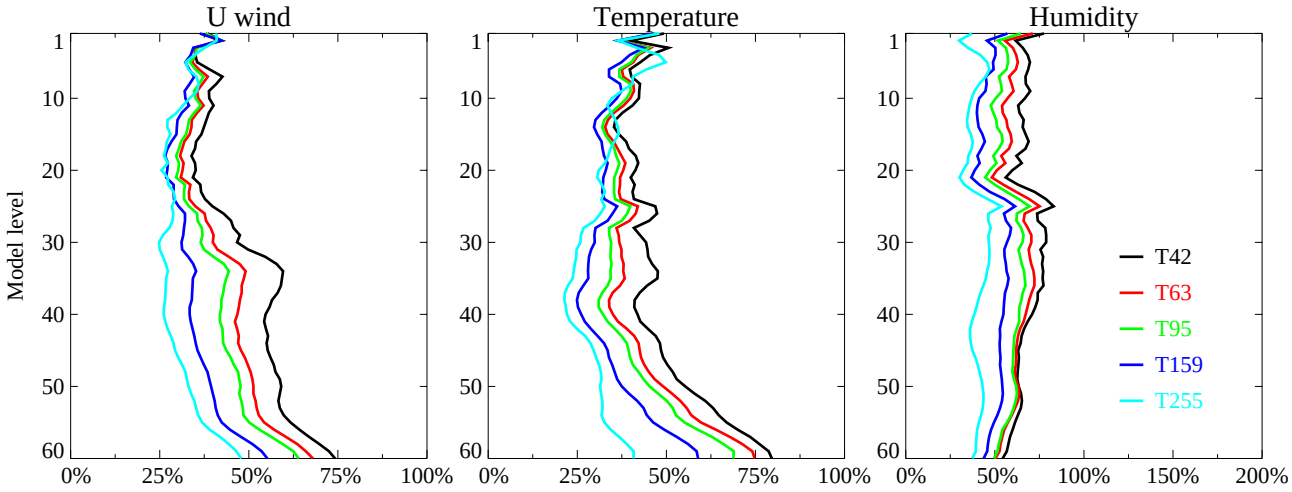

Figure 8: Same as figure 7 but for adiabatic tangent linear and nonlinear models.

main forecast to T719 or T799 in the coming years. In order to maintain consistency in the system, it is expected that the resolution of the inner loop of 4D-Var will also need to be increased.

It is also expected that in the near future, new types of data will require higher resolution inner loops. There are two main reasons for this: the resolution of the data and the quantities measured. At present, many satellites provide data at higher resolution than the assimilation system can ingest, and these have to be thinned. It would be desirable to use current and future data at their full resolution. New data types will require more accurate observation operators and linear physics which in turn require higher resolution inner loop to resolve the scales at which these phenomena occur. It is therefore important to study the behaviour of the tangent linear model as we go to higher resolutions.

The impact of inner loop resolution can be studied with the diagnostic proposed here. Figure 7 shows the performance of the tangent linear model for resolutions varying from T42 to T255. As expected, the quality of the linear model in general improves with increasing resolution. However, the quality for wind and temperature is degraded when going from T159 to T255.

An adiabatic experiment can help understand where the problem lies. Figure 8 shows that the impact of higher resolution in the adiabatic context is beneficial for all resolutions including T255. It shows that the tangent linear model can resolve some nonlinear phenomena up to that scale provided they are correctly represented (the tangent linear version of the adiabatic model is expected to be correct) and points to insufficient linear physics in the current system for T255 application. This is consistent with some preliminary 4D-Var experiments which did not show any forecast improvement with a T255 inner loop in the current system.

4.4 Small scales

An atmospheric model involves a variety of dynamical and physical processes which do not all affect the atmosphere on the same scale. It is therefore interesting to study how various scales are represented in the assimilation system as it will affect the assimilation of certain data types which are sensitive to local conditions and have a high spatial variability.

For a given resolution of the linear model, increments of varying scales are not propagated with equal accuracy. Figure 9 shows the relative error when propagating various increments with a T255 linear model with respect to the T511 nonlinear finite difference. The relative error when propagating an increment increases with the

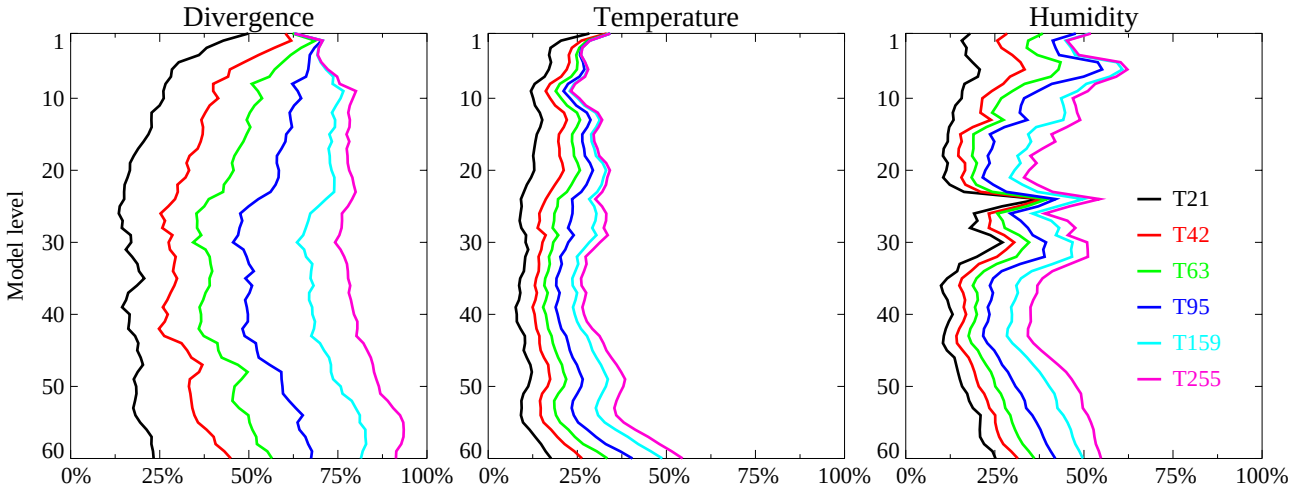

Figure 9: Relative error when propagating increments of various resolution as indicated in the legend with the same T255 linear model relative to the T511 forecast model.

truncation number of that increment for a given resolution of the linear model. This is another way of showing that small scale phenomena are not well described by the linear model. It is consistent with figures 8 and 7 which show that small scale dynamics are treated correctly but better physics is needed for the smaller scales.

4.5 Outer loops

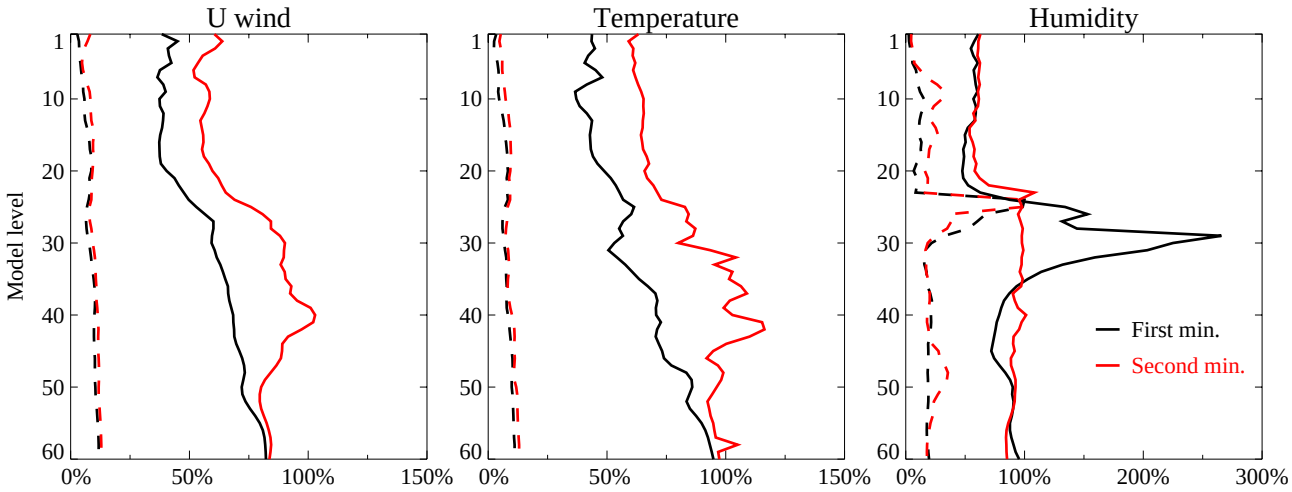

Figure 10: Relative error of the tangent linear model in the first and second 4D-Var minimisations. Note that in this figure the analysis increments are representative of successive minimisations and are thus different, in magnitude and scale. The plain lines are for the end of the 4D-Var window (12 hours), the dashed lines for the initial condition.

Currently, two iterations of the outer loop are performed. In the first minimisation the linear physics is not used because of the computational cost. In the second minimisation, the linear physics is activated. Figure 10 shows the relative error of the tangent linear model in both minimisations. The plain lines are for the end of the 4D-Var window (12 hours), the dashed lines for the initial condition. Note the difference with figure 6 where the tangent linear model was tested with and without physics with the same analysis increment. In successive minimisations as presented here, the partial increments are not the same. The increment in the second minimisation is known to be of smaller amplitude and to include more small scale features. These

two factors have opposite effects on the linear approximation: the smaller amplitude means that the linear approximation should be better justified while the smaller scale features are usually related to more nonlinear processes and would degrade the accuracy of the linear model. The figure shows that in that case, even though linear physics is used, the accuracy is degraded for the wind and temperature fields. It is however improved for the humidity field. This result is in agreement with the fact that incremental 4D-Var is not proven to converge at outer loop level even though it works in practice with few outer loop iterations: as the increments become smaller in amplitude and are made of smaller scale features, the tangent linear approximation breaks and the inner and outer loops become too different for the algorithm to converge at outer loop level.

4.6 A look at the maps

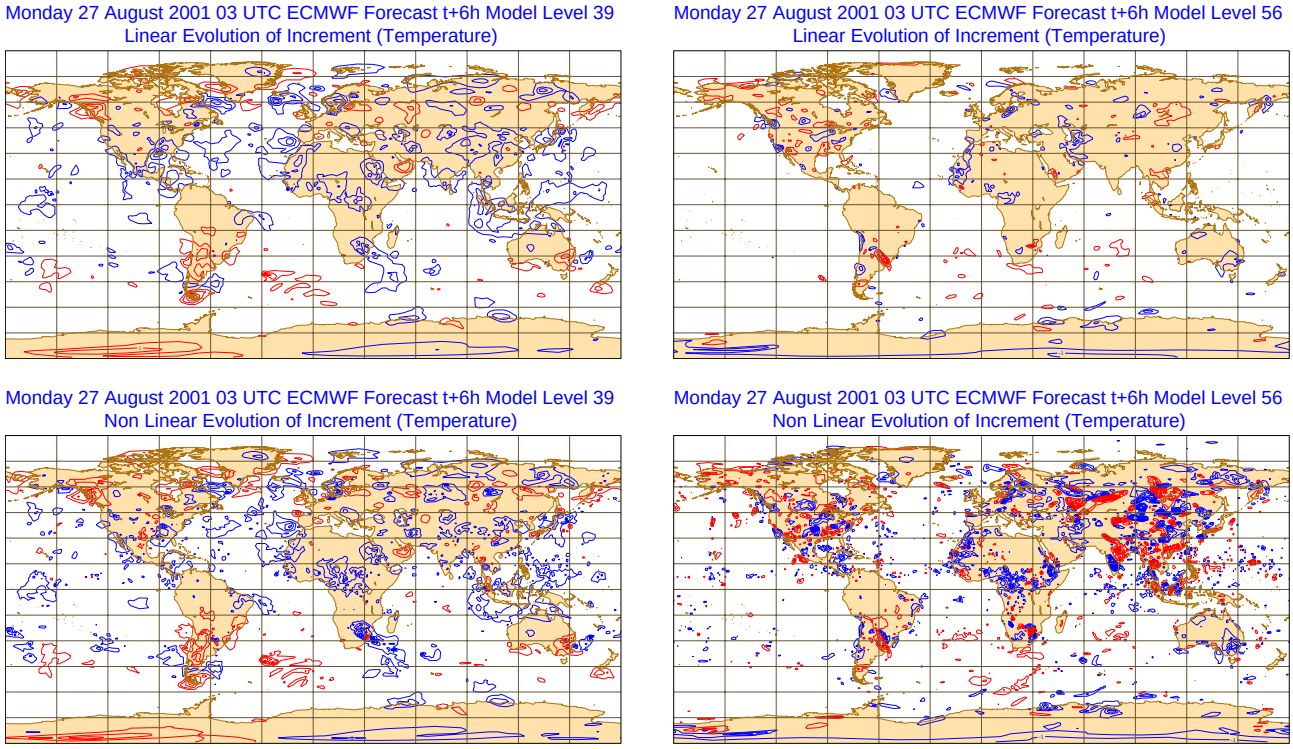

Figure 11: 6h evolution of an increment propagated by the tangent linear model (top panels) and by the nonlinear model (bottom panels). Temperature increments are shown with a contour interval of 0.5K. Although not identical, the increments at level 39 (500 hPa) show some similarity (left panels) while near the surface (level 56, right panels) they are very different.

Although all the figures presented so far give a global view of the tangent linear model error, it is worth looking in more details at a particular case. Figure 11 shows how an increment is propagated by the linear (top panels) and nonlinear models (bottom panels). The temperature increment in both cases is shown after a 6h evolution at model levels 39 and 56 (roughly 500 hPa and 990 hPa). The average relative error for both fields in this case are 49.7% and 87.7% respectively. In the free atmosphere, although not identical, the linearly and nonlinearly evolved increments have similar patterns (left panels). The right panels show that near the surface, this is not the case, the linearly evolved increment being much smaller and missing most of the features present in the nonlinear one. One would expect from these maps that useful information is retained at level 39 but very little information remains at the lower level. This seems particularly true over land areas. It confirms that the global averages presented here are meaningful in the meteorological context.

5 Interpolated trajectory

In the operational 4D-Var, the trajectory for the integration of the tangent linear and adjoint models was computed and stored by running the nonlinear model at the beginning of the execution of each minimisation, at the same low resolution as the minimisation. It has been shown that in some cases, this trajectory can diverge quickly from the high resolution forecast, even for forecasts as short as 12 hours (the length of the assimilation window). Figure 12 shows the error which occurred between the high and low resolution runs in the case of the storms over France in December 1999 for the T319/T63 system operational before October 2000 (left) and the T511/T159 system operational thereafter (right). One consequence was that the increments which were computed relative to the low resolution trajectory were not appropriate when added to the high resolution forecast. As shown by Bouttier (2001), this leads to misuse of correct data since the increment computed at low resolution brings the forecast away from observations in some locations when used in the higher resolution forecast. At ECMWF, to prevent this, an incremental convergence check was used: if the relative difference between the high and low resolution increments $d = \|H\delta x - H(\delta_x)\| / \|H(\delta_x))\|$ was greater than a certain threshold, the assigned observation error was increased and the data was effectively ignored. This can lead to the rejection of correct data if the high and low resolution trajectories are too inconsistent.

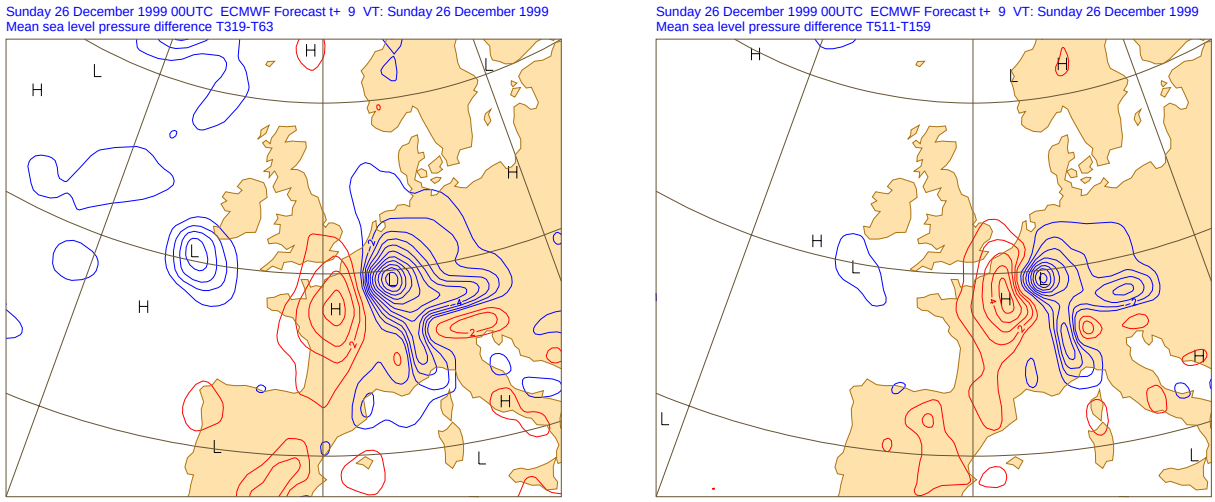

Figure 12: The left panel shows the MSLP difference between T319 and T63 forecasts 9 hours into the assimilation window for the December 1999 storm over France. The dipole pattern shows a shift in the position of the storm between the two forecast. The right panel shows the difference between T511 and T159 forecasts: although the difference is slightly smaller, the increase in operational inner and outer loop resolution did not eliminate the problem.

In order to improve the accuracy of the low resolution linear model, an interpolation procedure was developed, that truncates the high resolution trajectory to the resolution of the inner loops. The spectral components of the trajectory are truncated while a bilinear interpolation is used for grid point surface fields. Secondary trajectory fields such as those used in the physics and semi-Lagrangian parts of the code are recomputed at low resolution from the basic state which was interpolated from the high resolution trajectory.

There is a choice of the time interval between two trajectory states which allows the user to make a choice between the accuracy of the trajectory and the amount of memory used to store it. More memory can also be saved by storing the trajectory values with a lower precision than that used for computation. In the experiment presented here the trajectory is saved every hour and with half the precision of the computations (32 bits reals instead of 64). It may seem as though the accuracy with which the trajectory is stored is not important. This would contradict the reason why the possibility to interpolate the trajectory from the high resolution model was

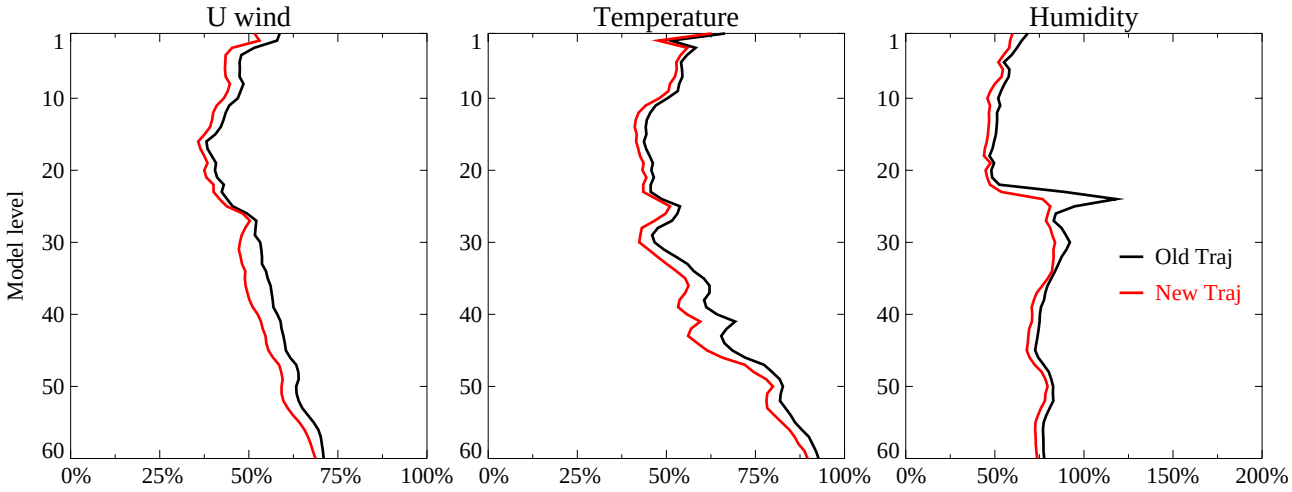

Figure 13: T159 tangent linear model relative error after 12 hours with respect to T511 nonlinear finite difference using the low resolution trajectory (in black) and using the interpolated trajectory (in red).

developed in the first place. However, one has to remember that in the tangent linear and adjoint models, the trajectory is not used as a starting point to carry forward an integration but rather as a reference state. It is thus important that it reproduces the evolution of the flow all along the assimilation window with reasonable and relatively constant accuracy. This is why it has to be computed from a high resolution run, but each state doesn't need to be as accurate as the initial condition to integrate the model needs to be. In other words, it is better to compute the trajectory accurately and then truncate the output rather than truncating the initial condition and integrating the trajectory from that truncated state.

Figure 13 shows the relative error of the tangent linear model after 12 hours with respect to the full nonlinear model for the IFS (cycle CY24R3). The use of the interpolated trajectory brings a small but consistent improvement for all the variables and all the levels in the model.

6 Linearity of the assimilation problem

In addition to the approximations which are necessary to make 4D-Var affordable, intrinsic nonlinearity in the laws governing the atmospheric flow also limits the use of the linear approximation in data assimilation. This issue has been studied in other contexts, but it is worth looking at it from the data assimilation point of view.

To test the linearity of the problem, the full forecast model is run three times, once from the background state, once from the background plus the analysis increment and once from the background minus the analysis increment. If the data assimilation problem was linear, the evolutions of the analysis increment and of its opposite should stay opposite over the length of the assimilation window.

A measure of the relative nonlinearity is given by Gilmour et al. (2001):

$$\theta = \frac{|\delta^+ + \delta^-|}{(|\delta^+| + |\delta^-|)/2}$$

where δ^+ is the evolved perturbation and δ^- is evolved from the opposite of the same perturbation.

As for the other results presented so far, it can be averaged for each variable at each model level. Figure 14 shows the values obtained with the T511 nonlinear model with the initial perturbation being an analysis increment. This figure has many similarities with the tangent linear relative error presented above. The errors

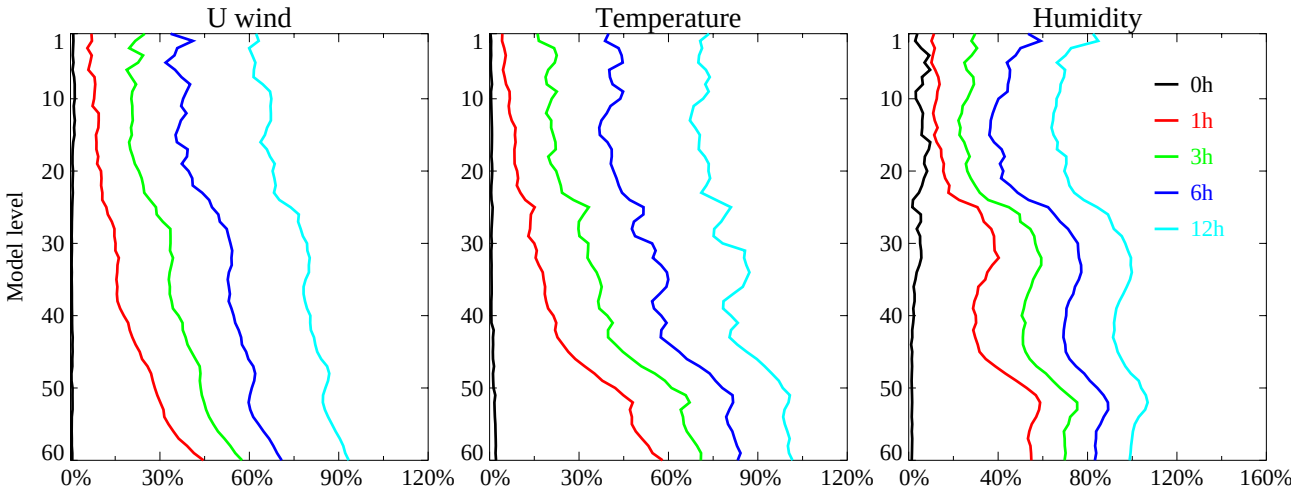

Figure 14: Relative nonlinearity of the T511 forecast model for an initial perturbation set to an analysis increment.

grow to close to 100% after 12 hours and the error near the surface grows very rapidly at the beginning of the assimilation window. The humidity error however has a different shape: the sharp maximum at the tropopause has disappeared. This was to be expected since here the comparison is between runs at the same resolution without interference from adjustment to orography.

This may seem a very discouraging result since it means that the assimilation problem at the scale and with the amplitudes at which we treat it is nonlinear in nature and that no linear approximation can be used to resolve it with adequate accuracy. This tentative conclusion should be modulated by the fact that the model is affected by a spin-down phenomenon during the first 12 to 24 hours of the forecast and the characteristics of the model in that range do not necessarily reflect that of the underlying problem, this is another inconsistency of the system. It would, in theory at least, be possible to create perturbations which would evolve to be opposite after 12 or 24 hours in the forecast and study the behaviour of the model from that point onwards. This would require more technical development than is possible in the scope of this study and would also take us away from the original idea which was to study the behaviour of 4D-Var as implemented operationally at ECMWF.

7 Conclusion

The validity of the tangent linear approximation in the context of incremental 4D-Var has been tested. The evolution of analysis increments propagated by the tangent linear model as used in the 4D-Var minimisation was compared with the high resolution forecast from the analysis generated by these increments. The impact of several factors such as physical processes, resolution and length of the assimilation window have been studied.

The first general result is that the relative error which is made by using a low resolution linear model rather than the full resolution nonlinear model in the inner loop is larger than expected. This at first can seem very alarming for 4D-Var. However, from ECMWF operational analysis performance, one can note that this method can be used to produce good quality analysis and forecasts. From the maps presented in figure 11 and from operational practice, it seems that relative errors of the order of 50% in the inner loop relative to the outer loop are acceptable in practice even though a lower error would improve the accuracy of 4D-Var.

For computational reasons, the 4D-Var inner loop has to be run at lower resolution than the forecast. The resolution difference introduces a discrepancy in the analysis system. The discrepancy can be reduced by increasing the resolution of the inner loop. However, the experiments with a T255 inner loop have shown that

the current implementation of the linear model is not sufficient for use at such high resolution.

It was also shown that the presence of linearised physics in the inner loop improves its accuracy, most noticeably near the surface and for the humidity fields. This could be expected since physics is more active in these areas. Work is currently under way at ECMWF to develop more accurate linear physics packages, additional processes such as radiation and convection are currently being linearised (Janisková et al. (2002)) and it is expected that this will reduce further the error in the linear model. The representation of humidity in the inner loop is also currently being revised. Allowing the use of grid point humidity in the tangent linear and adjoint models will reduce inconsistencies and should improve further the performance of the system in that respect. More generally, the humidity analysis is being reviewed and a new control variable has recently been proposed by Hólm et al. (2002).

The length of the 4D-Var window might have to be reduced to allow for higher resolution inner loop and compensate for more nonlinear phenomenon to be included, although this does not seem to be the critical factor at the moment. We have shown that the most rapid error growth occurs during the very first time-steps of the model. Genuine nonlinear processes might be the reason for this behaviour but we also know that the forecast model is affected by spin-down during the first 12 to 24 hours which is the range in which 4D-Var uses the model. It is therefore not possible at this time to determine whether the problem is really nonlinear, in which case no linear approximation can be expected to be accurate, or if we see another inconsistency generated by the spin-down of the model.

As a first step towards improving 4D-Var consistency, a new algorithm has been implemented in which the trajectory around which the tangent linear and adjoint models are run is interpolated from a high resolution model. It was tested using the diagnostics presented in this paper which showed a small and consistent improvement of the linear approximation accuracy and was implemented in the operational system in January 2003.

An increase in inner loop accuracy is important to reduce the discrepancy with the outer loop, but also for the use of higher resolution data, since 4D-Var can only assimilate data at resolutions that are resolved by the inner loop. In the coming years, data assimilation systems will face new challenges with the arrival of more data that describe the atmosphere with higher resolution both in space and time. It is also expected that new types of observations can be assimilated, such as rain or cloud observations. It is known that the phenomena involving rain and clouds can be very local and very nonlinear. They will have to be described in 4D-Var inner loop, both in terms of resolution and of the physical processes involved. This will push the current algorithm to its limits in terms of resolution and linearity.

All the components of 4D-Var are constantly evolving, through improvements of the nonlinear forecast model, of the linear and adjoint models or of the formulation of the cost function. The performance of the system depends on all its components. For example, in recent years, a lot of effort has been put into improving the treatment of humidity and moist processes across the system. Improvements in the forecast model should provide a better first guess for 4D-Var. This implies a better trajectory and smaller increments. This will in turn affect the validity of the linear approximation which should be re-evaluated regularly. The diagnostics presented in this paper are easy to run and should help assess the adequacy of the linear approximation for data assimilation in the future.

Acknowledgements

The author would like to thank the participants to the Workshop on Applications of Adjoint Models in Dynamic Meteorology (April 2002) where some of these results were first presented for their comments as well as E. Andersson, A. Simmons, M. Janisková and other colleagues at ECMWF for their comments on earlier versions of this paper.

References

F. Bouttier. The Development of 12 hourly 4D-Var. Tech. Memo. 348, ECMWF, September 2001.

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

R. Errico and K. Raeder. An examination of the accuracy of the linearization of a mesoscale model with moist physics. *Q. J. R. Meteorol. Soc.*, 125:169–195, 1999.

R. Errico, T. Vukicevic, and K. Raeder. Examination of the accuracy of a tangent linear model. *Tellus*, 45A: 462–477, 1993.

I. Gilmour, L. A. Smith, and R. Buizza. On the duration of the linear regime: Is 24 hours a long time in weather forecasting? Tech. Memo. 328, ECMWF, January 2001.

E. Hólm, E. Andersson, A. Beljars, P. Lopez, J.-F. Mahfouf, A. Simmons, and J.-N. Thépaut. Assimilation and Modelling of the Hydrological Cycle: ECMWF's Status and Plans. Tech. Memo. 383, ECMWF, September 2002.

M. Janisková, J.-F. Mahfouf, J.-J. Morcrette, and F. Chevallier. Linearized radiation and cloud schemes in the ECMWF model: Development and evaluation. *Q. J. R. Meteorol. Soc.*, 128:1505–1527, 2002.

M. Janisková, J.-N. Thépaut, and J.-F. Geleyn. Simplified and regular physical parametrizations for incremental four-dimensional variational assimilation. *Mon. Wea. Rev.*, 127:26–45, 1999.

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

F.-X. LeDimet and O. Talagrand. Variational algorithms for analysis and assimilation of meteorological observations : Theorical aspects. *Tellus*, 38A:97–110, 1986.

J.-F. Mahfouf. Influence of physical processes on the tangent-linear approximation. *Tellus*, 51A:147–166, 1999.

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

C. Pires, R. Vautard, and O. Talagrand. On extending the limits of variational assimilation in nonlinear chaotic systems. *Tellus*, 48A:96–121, 1996.

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

T. Vukicevic and R. Errico. Linearization and adjoint of parameterized moist diabatic processes. *Tellus*, 45A: 493–510, 1993.