Tellus (2007), 59A, 706–718
Printed in Singapore. All rights reserved

TELLUS

Incremental 4D-Var convergence study

By YANNICK TRÉMOLET*, *European Centre for Medium-Range Weather Forecasts, Shinfield Park, Reading RG2 9AX, United Kingdom*

(Manuscript received 23 February 2007; in final form 2 July 2007)

ABSTRACT

Since its operational implementation at European Centre for Medium-Range Weather Forecasts, the incremental four-dimensional variational data assimilation system (4D-Var) has run with two outer loop iterations. It has been shown in the past that more outer loop iterations were leading to the divergence of the algorithm. We re-evaluate here the convergence of 4D-Var at outer loop level with the current system.

Experimental results show that 4D-Var in its current implementation does diverge after four outer loop iterations. Various configurations are tested and show that convergence can be obtained when inner and outer loops are run at the same resolution, or at least with the same time-step. This is explained by the presence of gravity waves which propagate at different speeds in the linear and non-linear models. It is shown that these gravity waves are related to the shape of the leading eigenvector of the Hessian of the 4D-Var cost function which is determined by surface pressure observation and which controls the behaviour of the minimization algorithm. The influence of the choice of the inner loop minimization algorithm and preconditioner is also presented. Finally, some directions for possible future configurations of incremental 4D-Var are given.

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 measurement errors. The data assimilation problem comprises constructing a suitable initial condition accounting for the observations and the model. The European Centre for Medium-Range Weather Forecasts (ECMWF) uses four-dimensional variational data assimilation (4D-Var). The basic principle of the method is to minimize a cost function which measures the gap between observations and the solution of the model during the assimilation window. The control variable of the problem is the initial condition of the model.

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). A complete description of ECMWF 4D-Var assimilation system was given by Rabier et al. (2000), Mahfouf and Rabier (2000), Klinker et al. (2000) and updated by Andersson et al. (2004). As incremental 4D-Var was developed, Rabier et al. (2000) tested various configurations of the algorithm and found that it did not converge satisfactorily when more than two trajectory updates were used. They promised at that time that 'We shall re-examine the impact of the number of outer loops in 4D-Var later'. Since 4D-Var has been introduced in operations in November 1997, the system has evolved considerably but no systematic attempt has been made to re-assess the possible use of more outer loop iterations.

The normal mode initialization, shown to be the main reason for the lack of convergence of 4D-Var at outer loop level by Rabier et al. (2000) has been replaced in June 2000 by a weak constraint digital filter initialization as described by Gauthier and Thépaut (2001). As the system has evolved to higher resolution and will continue to do so in the coming years, more small scales are resolved and non-linear phenomena in the model are becoming more important. Non-linear phenomena are also increasingly important as assimilation of new types of observations related to clouds and rain are being developed (Andersson et al., 2005). In incremental 4D-Var, non-linearities can only be taken into account in the outer loops and, to better resolve them, it is important to be able to run more non-linear outer iterations. In this paper, we re-evaluate the convergence of incremental 4D-Var at the outer loop level in this new context.

The outline of the paper is as follows: in the next section, we described the incremental 4D-Var algorithm as implemented at ECMWF. In the following section, we show some diagnostics of convergence of the current system and highlight some of its

*Correspondence.
e-mail: y.tremolet@ecmwf.int
DOI: 10.1111/j.1600-0870.2007.00271.x

deficiencies. In Section 4, we detail some of the reasons for the behaviour exhibited in the previous section and we propose possible improvements for the current operational system.

2. The incremental 4D-Var minimization problem

The cost function which is minimized in 4D-Var includes three terms and can be written as:

$$J(\mathbf{x}) = (\mathbf{x} - \mathbf{x}_b)^T \mathbf{B}^{-1} (\mathbf{x} - \mathbf{x}_b) + [\mathcal{H}(\mathbf{x}) - \mathbf{y}]^T \mathbf{R}^{-1} [\mathcal{H}(\mathbf{x}) - \mathbf{y}] + \mathcal{J}_c,$$

where $\mathbf{x}$ is the control variable, $\mathbf{x}_b$ is the background state, $\mathbf{y}$ is the vector of observations, $\mathbf{B}$ is the background error covariance matrix, $\mathbf{R}$ is the observation error covariance matrix, $\mathcal{H}$ is the non-linear observation operator and $\mathcal{J}_c$ is an initialization term. $\mathcal{J}_c$ limits the growth of gravity waves by adding a penalty to rapidly varying oscillations in the divergence field as described by Gauthier and Thépaut (2001). In the strong constraint formulation of 4D-Var used in this study, the control variable is the initial condition at the beginning of the assimilation window. In this notation, the forecast model is embedded in the observation operator $\mathcal{H}$ which computes the observation equivalent at the correct time and location. The first term in the cost function is called the background term, denoted J_b and the second term is called the observation term, denoted J_o.

In its incremental formulation, the minimization problem is written as a function of the correction $\delta\mathbf{x}$ from a first guess $\mathbf{x}_g$. At the minimum, $\delta\mathbf{x}$ will be the analysis increment. A quadratic approximation of the cost function is given by:

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

where $\mathbf{b} = \mathbf{x}_g - \mathbf{x}_b$, $\mathbf{d} = \mathcal{H}(\mathbf{x}_g) - \mathbf{y}$ is the departure from observations, $\mathbf{H}$ is the linearized observation operator, and J_c is the linearized digital filter initialization term. $\mathbf{x}_g$ is the starting point of the minimization, it is usually equal to the background in the first minimization. The tangent linear model is embedded in the linearized observation operator. The gradient of the cost function with respect to the initial condition is obtained using the adjoint model. A non-linear integration from the guess $\mathbf{x}_g$ provides the trajectory around which the tangent linear and adjoint models and the observation operators are linearized. It is called the trajectory run. The departures from observations $\mathbf{d}$ are also computed in this trajectory run.

This approximate quadratic minimization problem is solved using an iterative algorithm: this is the inner loop of 4D-Var. At the time of this study at ECMWF, a preconditioned Lanczos-conjugate gradient algorithm (CONGRAD) is used to solve the inner loop minimization problem as described by Fisher (1998). After this minimization, the departures and trajectory can be recomputed using the non-linear model and a new linearized problem is defined around an improved guess. The process can be repeated: this is the outer loop of incremental 4D-Var. If the linearized problem is reasonably close to the non-linear problem, as tested by Trémolet (2004), its solution should be an approximation of the solution of the non-linear problem. At the next outer loop iteration, the starting point is closer to the solution, the quadratic approximation is more accurate and should in principle provide a more accurate solution to the non-linear problem. The algorithm should converge to the solution of the non-linear problem, although there is no formal proof of convergence in the most general case. Relying on particular assumptions, proofs of convergence can be obtained, as shown, for example, by Gratton et al. (2007). Unfortunately, these assumptions involve quantities that cannot be computed in a realistic system and thus cannot be tested.

In order to reduce the computational cost of the data assimilation process, the inner loop minimizations are run at lower resolution than the forecast. ECMWF's Integrated Forecasting System (IFS) uses a spectral forecast model and an assimilation window of 12 h. Two inner loop minimizations are run, the first one at the spectral truncation of T95, the second one at T255 while the outer loops and forecast are run at T799. For further reduction of the computational cost, only very simplified linear physics are used in the first inner loop minimization. The incremental 4D-Var algorithm is shown schematically on Fig. 1.

An initial experiment was run where the only differences with respect to an operational IFS analysis were to run 10 outer loop iterations instead of two and to set the number of inner loop iterations for each of the second and following minimizations to 25. The evolution of the components of the cost function in this configuration is shown on Fig. 2.

On this figure, and all similar figures in this paper, solid lines represent the J_o component of the cost function, as evaluated in the inner loop, the vertical bars represent the value of J_o as evaluated in the non-linear trajectory runs. The figure also shows the evolution of the other components of the cost function: J_b (dashed lines) and J_c (dotted lines). These are only computed at low resolution, in the inner loop. The values of J_o, J_b and J_c in the inner loops are obtained from the first and last evaluations of the cost function, at the start and at the end of the minimization. For the figures, they are joined by straight lines as intermediate information is not available. J_b has been inflated by a factor of 10 and J_c by a factor of 100 to improve the readability of the figures.

The figure shows that the values of J_o at the end of a minimization, in the following high resolution trajectory and at the starting point of the ensuing minimization do not coincide exactly. This is because they are evaluated at different resolutions. The discrepancy in J_b between the end of the first minimization and the beginning of the second minimization is due to the change of resolution between the minimizations and the implied change in the $\mathbf{B}$ matrix. J_c is a digital filter initialization term, it is only

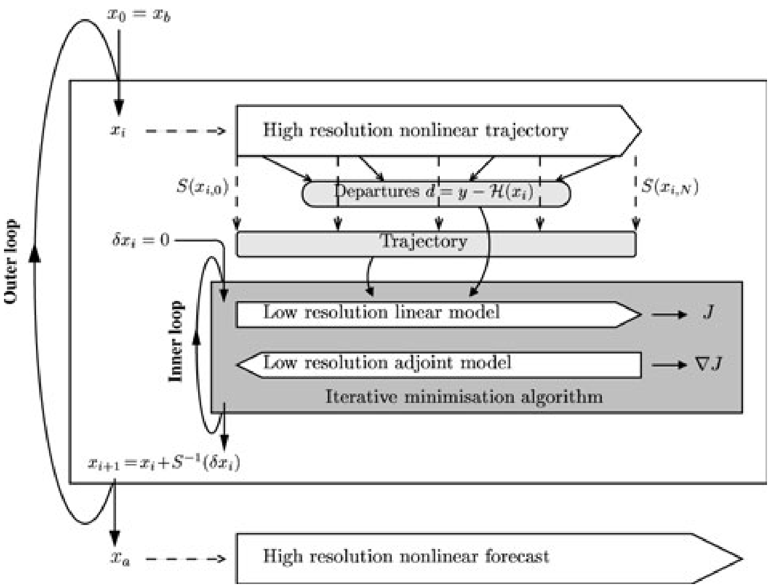

Fig. 1. Incremental 4D-Var algorithm. The trajectory and the departures **d** from the observations **y** are computed at high resolution and the trajectory is interpolated to low resolution (S). The linearized cost function is minimized at low resolution using an iterative algorithm (inner loop). The resulting increment $\delta \mathbf{x}_i$ is interpolated back to high resolution (S^{-1}) and added to the current first guess. The process is repeated (outer loop, subscript i) until the analysis $\mathbf{x}_a$ is obtained.

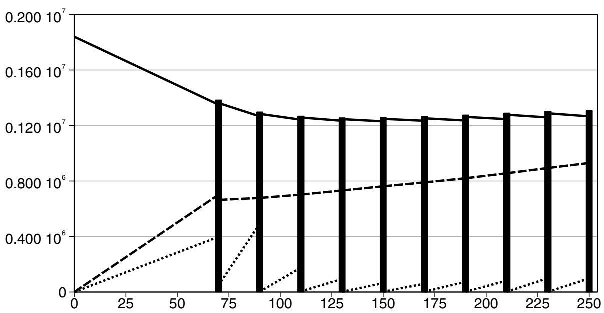

Fig. 2. Evolution of the components of the 4D-Var cost function as a function of the cumulated number of inner loop iterations. The solid line represents J_o as computed in the inner loop while the bars represent J_o as computed in the non-linear trajectory runs. The dashed lines represent J_b and the dotted lines J_c. For readability of the figure, J_b has been inflated by a factor of 10 and J_c by a factor of 100.

applied to the partial increment within each minimization which explains why its initial value is zero for each minimization.

This figure shows that the minimum value for J_o is obtained after four outer loops iterations. Then, J_o starts increasing slowly and 4D-Var diverges. The norm of the gradient of the cost function evolves in a similar manner. As was the case when 4D-Var was first introduced at ECMWF, the incremental algorithm still diverges when more than a few outer loop iterations are run. In the remainder of this paper, we investigate this behaviour.

3. Outer loops convergence experiments

3.1. Inner loop resolution

The impact of the difference in resolution between the inner and outer loops is examined first. In all experiments presented below, the outer loop was run at T255. This resolution was chosen in order to be able to run the inner loops at the same resolution. All the minimizations are run with all available linear physics (see IFS documentation[1]). The number of iterations is fixed to 25 per minimization and the resolution is the same for all minimizations within each experiment. All other aspects are kept identical to the operational configuration at the time of this study (IFS CY29R1). Figure 3 shows the evolution of the cost function for inner loop resolutions of T42, T95, T159 and T255. The first result from that experiment is that when inner and outer loop resolutions are the same, 4D-Var does converge. With T159 inner loops and T255 outer loops, 4D-Var still converges but the final J_o is slightly higher than with T255 inner loops. As the inner loop resolution goes down to T95 and T42, 4D-Var diverges. The discrepancy between the J_o component of the cost function at the end of a minimization and at the beginning of the following minimization increases with the mismatch in resolution. The mismatch in resolution between inner and outer loops seems to be an important factor for 4D-Var to converge or diverge. However, convergence of the algorithm does not prove that the solution is a minimum of the non-linear cost function, even in the T255/T255 case.

The maps of temperature increments at model level 49 (approximately 850 hPa) (Fig. 4, left column, first two rows) show that the increments are more localized with the higher resolution inner loop. The maps of surface pressure increments (Fig. 4, right column, first two rows) show a very large-scale increment with a ring pattern centred over Europe with the T95 inner loops. A similar pattern is visible on the temperature increment map, although not as clearly marked. The maps, shown here for

[1]Chapter II, section 2.6, available at http://www.ecmwf.int/research/ifsdocs/

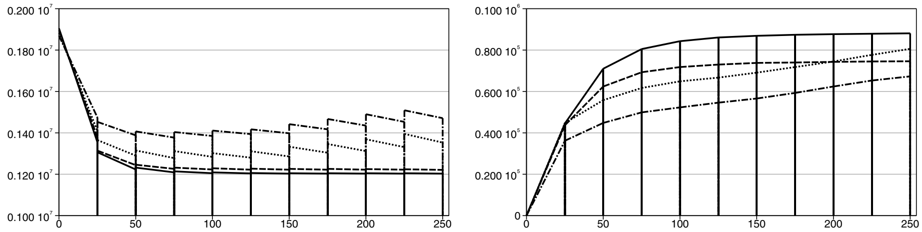

Fig. 3. Evolution of J_o (left-hand panel) and J_b (right-hand panel) components of the 4D-Var cost function for inner loop resolutions of T42 (dot–dashed), T95 (dotted), T159 (dashed) and T255 (solid), all with T255 outer loops.

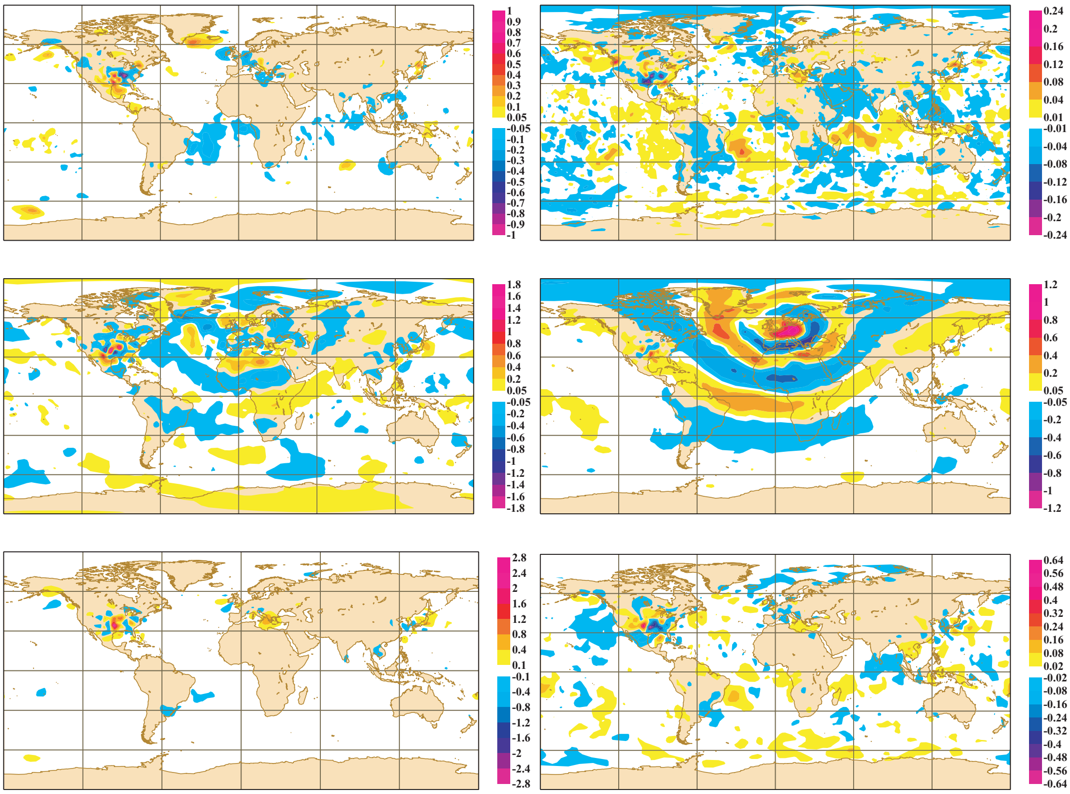

Fig. 4. Increments for temperature at model level 49 (≈850 hPa, left column, K) and surface pressure (right-hand column, hPa) at minimization 7 for T255/T255 experiment (top), T255/T95 experiment (middle) and for T255/T95 experiment with a time-step of 1800 s in inner and outer loops (bottom).

minimization 7, are representative of all minimizations after the first four or five.

3.2. *Time-step*

The change of resolution between inner and outer loops implies a change in the default time-step used in the IFS. The T159 and T255 inner loops of the experiments were run with a 30-min time-step while the T95 and T42 inner loops were run with a 1 h time-step. In order to separate the effect of spatial resolution from temporal resolution, Fig. 5 shows the evolution of the 4D-Var cost function for a T255/T95 experiment with a 30-min time-step in both inner and outer loops. With the shorter time step, most of the increase in J_o after iteration 100 has been eliminated. The amplitude of the increments has reduced (not shown) in the later outer loop iterations and the spurious pattern of the increments at minimization 7 has disappeared (Fig. 4, bottom). More than spatial resolution, it seems that temporal resolution is a key element in the convergence of 4D-Var. This is probably the key factor explaining the convergence of the T255/T159 experiment where inner and outer loops were run with the same time-step (Fig. 3), more than the smallest spatial resolution difference. The discrepancy between high and low resolution J_o values is also reduced when the same time-step is used.

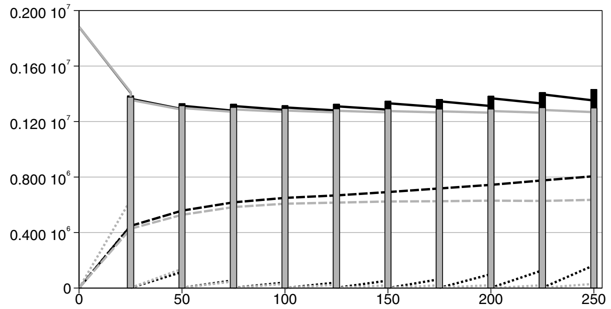

Fig. 5. Evolution of 4D-Var cost function components for T255/T95 experiment with an inner loop time step of 3600 s (in black) and 1800 s (in grey). The outer loop time step is 1800 s.

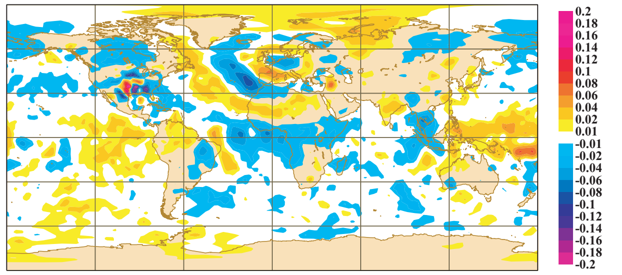

Fig. 7. Surface pressure increment at minimization 10 for T159/T95 idealized experiment (hPa).

3.3. Idealized case study

It is possible to create an idealized 4D-Var test case by replacing the observed values for all the observations used in an assimilation cycle by their background equivalent, that is, by the output of the observation operator computed from the background $\mathcal{H}(\mathbf{x}_b)$. This can easily be done when the departures from observations are computed in the first trajectory run. In that case, the initial value of J_o is exactly zero, the background is the exact solution of the minimization problem and 4D-Var should not produce any increment. One advantage of generating simulated observation with this method is that the distribution of observations in time, space and types exactly matches a real case, avoiding the issue of specifying a simplified but realistic observation network.

In the first idealized experiment, inner and outer loops are all run at T159. Figure 6 shows that 4D-Var does produce a small increment. The value of J_o stabilizes at around 2000 after a few iterations which is three orders of magnitude smaller than in a real case but not zero as expected. In the first non-linear trajectory, J_o is exactly zero. In principle, the first minimization should start with an initial increment also exactly equal to zero. This implies that the initial gradient is zero and no increment should be generated. In the following non-linear trajectory, because a supersaturation removal procedure is applied and because the background is supersaturated, J_o is non-zero. From there on, a small increment is generated and 4D-Var reaches an equilibrium between the super-saturation removal and the background constraint. This is visible in the experiment after the first minimization. In practice, the first minimization also generates a very small but non-zero increment because the background and the trajectory at initial time are not generated with the same procedure[2] (different interpolation and GRIB packing) which means J_b is non-zero in the initial evaluation of the cost function and generates a non-zero gradient. The behaviour shown on Fig. 6 is thus deemed acceptable.

When the inner loop is run at T95, 4D-Var produces an increment and, in this case, the values of J_o and J_b keep increasing as more iterations are performed. Again, the mismatch in resolution and time-step seems to be important for 4D-Var convergence or divergence, even in this idealized case. Figure 7 shows the surface pressure increment generated in the last minimization of the idealized T159/T95 experiment. The pattern is the same as seen previously, although it appears later in the iterations and with a smaller amplitude. This pattern does not correspond to an actual increment since we know that in this case, the correct solution is $\delta x = 0$. It is an artefact of the minimization process. This behaviour is pathological and is investigated below.

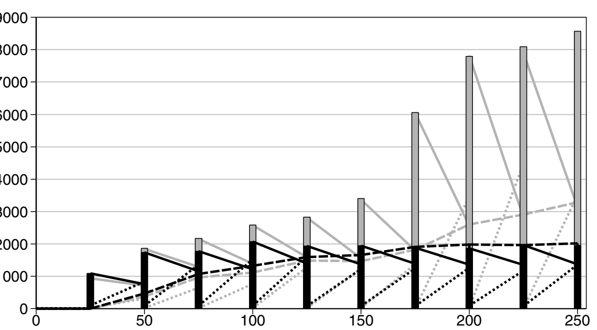

Fig. 6. Evolution of T159/T159 (in black) and T159/T95 (in grey) 4D-Var cost function components in the idealized case where the exact solution should remain zero.

3.4. Positive feedback

We have shown in Section 3.1 that a 4D-Var assimilation experiment run using T255 outer loops with a 30-min time-step and T95 inner loops with a 1 h time step starts diverging after four outer loop iterations. The largest increments in the latest minimizations are in the surface pressure component of the control variable. Figure 8 shows the surface pressure increment from minimization 7 and its evolution in the first 6 h of the assimilation window, propagated by the tangent linear model in the inner loop on the left and by the non-linear model in the outer loop on the right. The general pattern of the surface pressure increment is circular centred over northern Europe. The linear and non-linear evolutions slowly diverge from each other. This

[2]This has been modified since this study was performed.

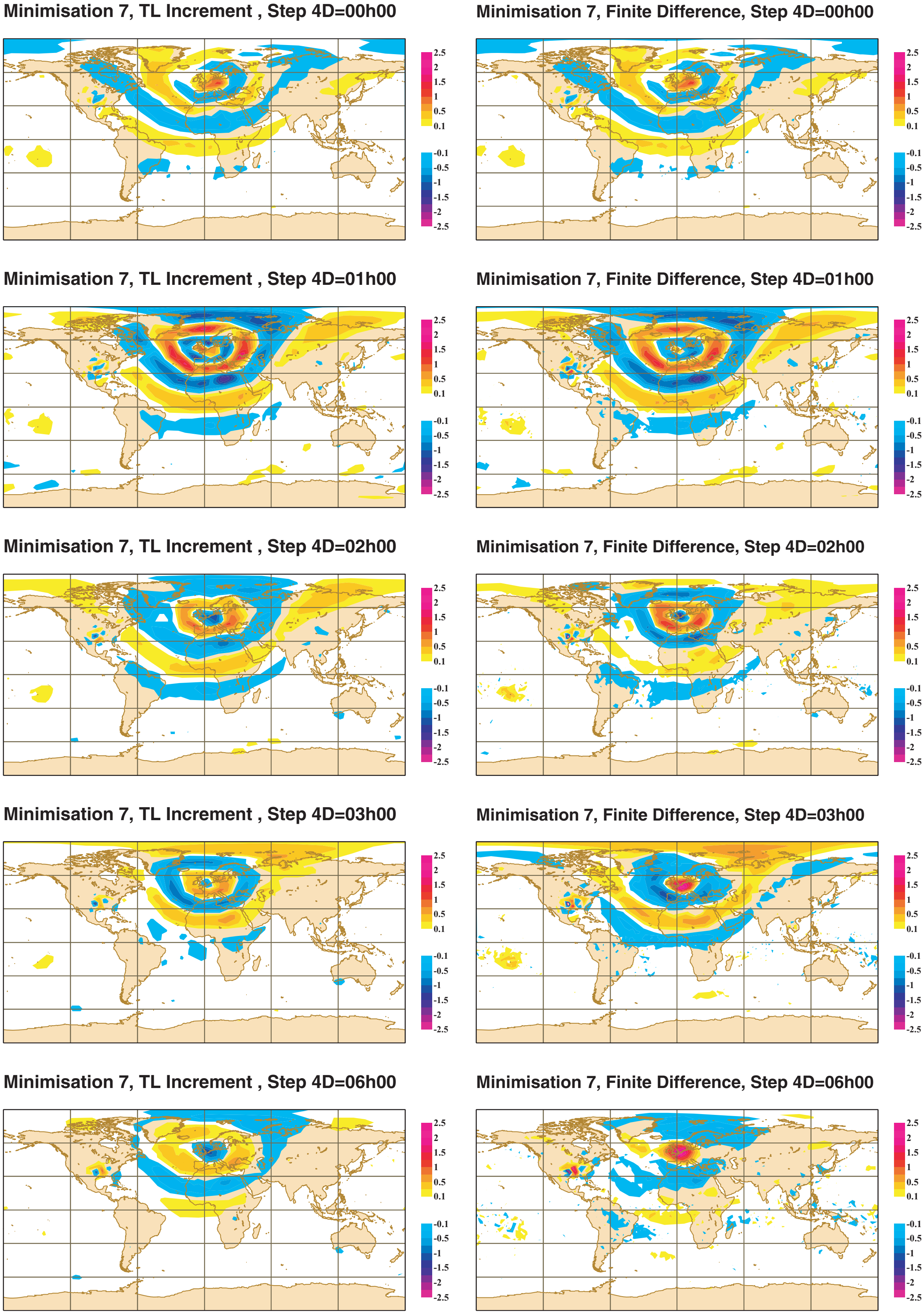

Fig. 8. Surface pressure increment for minimization 7 propagated by the T95 tangent linear model (left-hand panel) and the T255 non-linear model (right-hand panel) at initial time and at various times in the assimilation window (hPa).

is a gravity wave which propagates with different phase speeds in the linear and non-linear integrations. After 6 h, they are close to opposite phases, the increment at the centre of the pattern is positive in one case, negative in the other.

What is happening is that the inner loop minimization attempts to fit the observations in the area at the centre of the pattern. The increment is then added to the first guess and evolves in the non-linear integration. Because of the discrepancy between inner and outer loops, this new trajectory moves away from the data, in the direction opposite to the one intended by the minimization. In the ensuing minimization, an increment will be generated which will again try to fit the data, by adding another increment, with the same pattern as the previous one and even larger amplitude to compensate for the new high resolution departure. There is a positive feedback between the inner and outer loops, the increment keeps growing and 4D-Var eventually diverges. Although the experiments presented here were run with a 12 h assimilation window, the fact that the gravity waves in the inner and outer loops are in opposite phase after 6 h shows that the same phenomenon could affect 4D-Var with a 6 h window.

Gravity wave phase speed being highly dependent on time step, this explains why we do not observe that phenomena when both models use the same time step, regardless of spatial resolution. This has been tested with several time step values and shows that the tangent linear model is able to reproduce the non-linear model behaviour with respect to phase speed accurately for a range of time-step values. This is not a linearization issue since, given the same time step, the linear and non-linear models behave similarly. It is a consequence of the other approximation made in operational 4D-Var to reduce the computational cost: the reduction of resolution, both spatial and temporal, in the inner loops.

In a semi-implicit scheme such as the one used in the IFS, the phase speed of gravity waves depends on the time step being used. No simple modification of the semi-implicit scheme was found in the framework of this study that would eliminate the discrepancy between the inner and outer loops without changing significantly the behaviour of the model. It is however an aspect that could be investigated further in the future.

4. Explanation of poor convergence

4.1. Hessian and increments

The previous section shows how the increments generated after the fourth minimization amplifies, resulting in divergence of 4D-Var. Figure 9 shows the partial surface pressure increments for the 10 minimizations of the T255/T95 experiment. In order to understand the origin of this pattern, several experiments were run, for various times of day, various times of the year and with several tangent linear model configurations, with or without physics. All have shown similar increments in later outer loop iterations, as did the idealized experiment in Section 3.3.

The pattern of the increment is related to the shape of the leading eigenvector of the Hessian of the 4D-Var cost function. Andersson et al. (2000) have shown that the eigenstructure of the Hessian of the cost function is driven by the density and accuracy of observations. They showed in a simplified case that with n observations at the same location, an approximation of the largest eigenvalue and condition number of the minimization problem is given by:

$$\kappa \approx 2n(\sigma_b/\sigma_o)^2 + 1,$$

where σ_b^2 and σ_o^2 are the variances of background and observation errors. In the current ECMWF system, the dense and accurate network of surface pressure observations over Europe dominate as indicated by the leading eigenvector of the Hessian of the cost function (Fig. 10, left-hand panel).

An additional experiment was run where observation error for surface pressure observations was doubled. Figure 10 shows that the leading eigenvectors of the Hessian of the cost function for this experiment (right-hand panel) is almost unchanged. However, the eigenvalue associated with this eigenvector is 1502 against 5474 for the original experiment, in rough agreement with the simplified expression above. Figure 11 shows the 10 partial increments when surface pressure observations error is doubled. The spurious pattern has disappeared and Fig. 12 shows that 4D-Var convergence is improved.

This experiment shows that surface pressure observations over Europe are controlling the convergence of 4D-Var and that the eigenstructure of the Hessian of the cost function is a useful means to monitor the behaviour of incremental 4D-Var, in particular the largest eigenvalue and the eigenvector associated to it. It seems particularly unfortunate that the pattern of the eigenvector associated to the largest eigenvalue (a consequence of the observing system) is so sensitive and subject to an unstable positive feedback effect (a consequence of the model). It is the combination of observations, through their distribution and error characteristics, and of the model, through the speed of gravity waves, that currently determines the convergence or divergence of 4D-Var.

4.2. Inner loop minimization algorithm

In incremental strong constraint 4D-Var, the smallest eigenvalue of the minimization problem is always 1 and the condition number is equal to the largest eigenvalue. Without any other preconditioning, the condition number gives a measure of the difficulty to find the minimum, regardless of the actual background and observation values but depending on their error characteristics. The conjugate gradient-Lanczos minimization algorithm (CONGRAD) used in the IFS is based on an spectral preconditionner that counters this effect by modifying the apparent leading eigenvalues thus reducing the apparent condition number. Since the eigenstructure of the cost function is at the root of the divergence

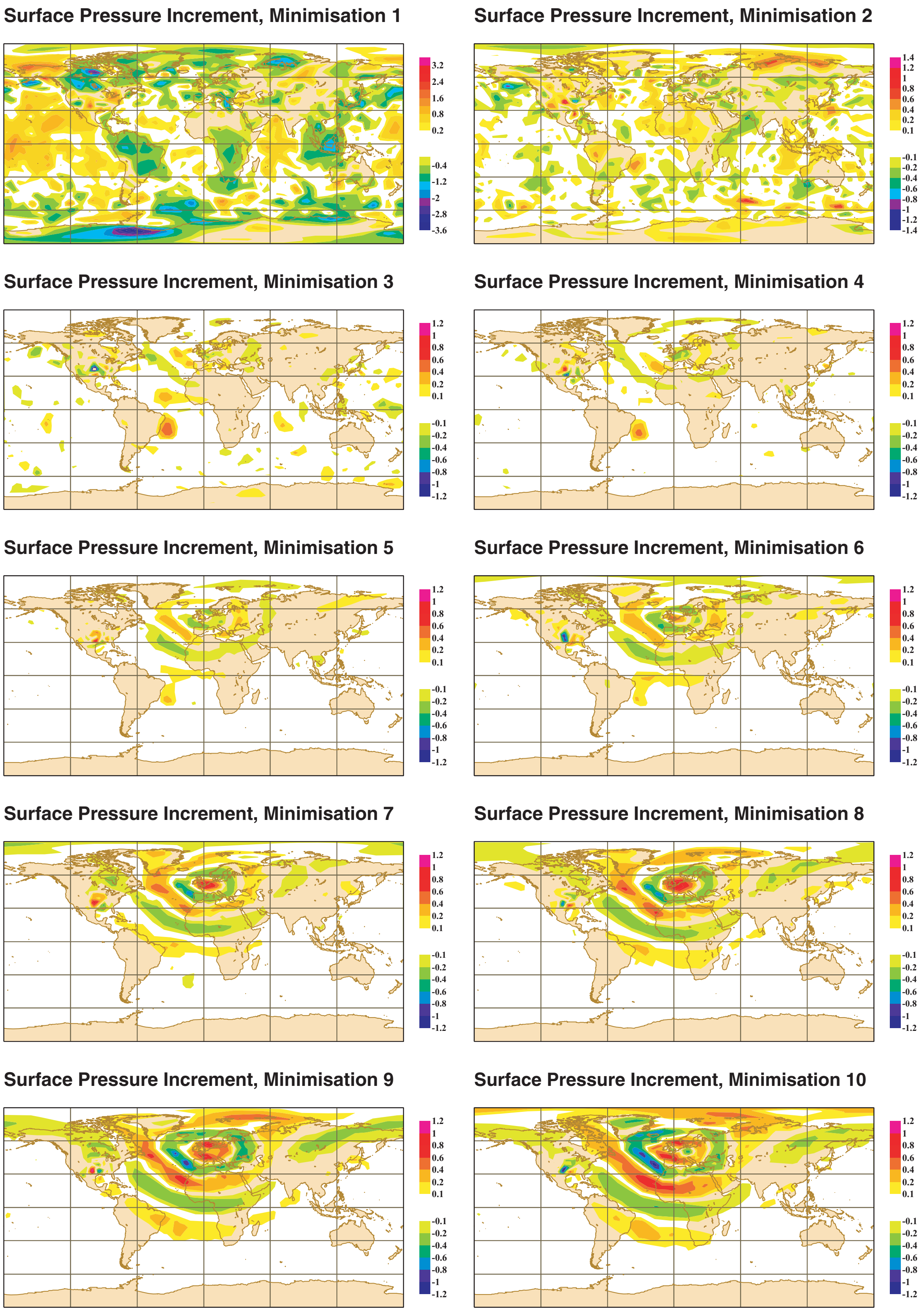

Fig. 9. Partial surface pressure increments (hPa) generated by the successive minimizations of the T255/T95 experiment, running from top left to bottom right. The contour interval is different for the first two increments which have larger amplitude.

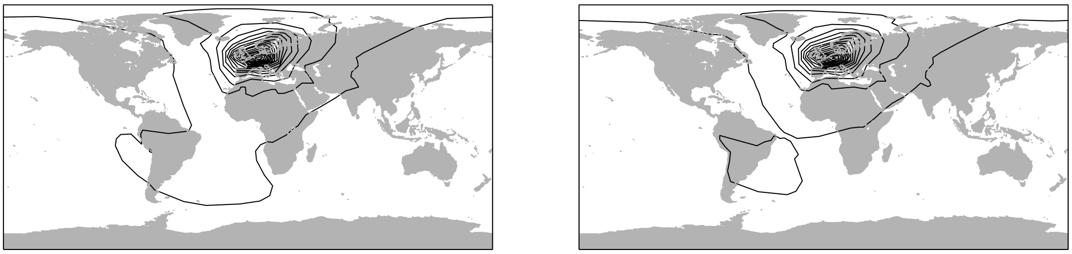

Fig. 10. Surface pressure component of the leading eigenvector of the Hessian of the cost function for the T255/T95 reference experiment (left-hand panel) and with doubled surface pressure observation error (right-hand panel). Contour interval is 0.2.

of 4D-Var, it is worth testing the outer loop convergence with another inner loop minimization algorithm.

The quasi-Newton minimization algorithm (m1qn3) developed by Gilbert and Lemaréchal (1989) is available in the IFS and experiments show that the minimization algorithm used in the inner loop has an impact on the overall convergence (Fig. 13). The value of J_o in the outer loop converges with m1qn3 which was not the case with the conjugate gradient. The numerical values show that J_o in the outer loops is strictly decreasing, although it does not decrease as rapidly in the first iterations. The value of J_b also remains lower throughout with m1qn3 as is the case with the value of J_c after the third minimization. Another experiment where the first six minimizations were run using CONGRAD followed by four minimizations using m1qn3 was performed and showed that m1qn3 improves over CONGRAD for these latter four minimizations.

The two algorithms should in principle give the same solution for the quadratic minimization problem at inner loop level. However, the problem is not solved to convergence and since the two algorithms take a different path towards the solution, the partial increments differ. The partial increments obtained from m1qn3 seem less prone to the positive feedback effect associated with the leading eigenvector of the Hessian of the cost function and to the positive feedback effect associated with it.

In this comparison between conjugate gradient and quasi-Newton algorithms, the number of iterations for each inner loop has been kept fixed to 25 for both algorithms. Laroche and Gauthier (1998) and Lawless et al. (2005) have shown that the accuracy with which the inner loop minimization is resolved has an impact on the overall convergence of incremental 4D-Var with imperfect observations. More precisely, their results show that the best results are obtained when inner loop iterations are stopped early enough to avoid overfitting observational noise, as this would introduce spurious gradients in the following trajectory integration. Since we know that the preconditioned conjugate gradient converges faster, it is legitimate to ask whether it is the accuracy of the solution in each minimization that has an impact on the overall incremental 4D-Var convergence.

Another comparison between quasi-Newton and conjugate gradient algorithms was performed where inner loop iterations are stopped when the norm of the gradient has been reduced by a factor of 4 for both algorithms. Figure 14 shows that the lowest overall value of J_o at high resolution is obtained with m1qn3 in the last non-linear trajectory. It is worth noting that it is also obtained with a lower value of $J_b (J_b = 55\,620)$ than the value of J_b when CONGRAD reaches its minimum ($J_b = 57\,055$), at iteration 6 in that case. Thus, in addition to the fact it does not diverge, m1qn3 seems to provide a better fit to observations with a smaller increment, which should in principle be a better solution. Because of the total number of iterations performed here and the cost of the outer loops, it is not computationally affordable to run enough cases to reliably assess the quality of the ensuing forecasts.

The accuracy with which the linearized problem is solved is not the only factor to consider since in this latest experiment both algorithms were stopped with the same convergence criterion. One characteristic of a preconditioned conjugate gradient algorithm such as CONGRAD is that the directions of the leading eigenvectors are emphasized. In order to reach a given convergence criterion, the algorithm solves the problem very accurately in those directions. The consequence is that it will fit the observations in these directions very closely. This is the over-fitting phenomenon already mentioned even if, in the present case, it might be limited to the subspace spanned by the leading eigenvectors. Unfortunately, this subspace is unstable because of the gravity wave positive feedback effect described in Section 3.4. In contrast, m1qn3 does not emphasize this particular direction and tends to reduce the gradient more evenly, thus alleviating the positive feedback effect. Several factors can contribute to this: the fact that approximate line searches are used and the fact that only the most recent curvature information is retained in m1qn3 could both play a role by limiting the accuracy of the solution. It is expected that running m1qn3 to convergence would fit all observations more closely and, passed a certain threshold, would produce the same positive feedback effect. This threshold was not reached in the experiments performed in the course of this study.

Gratton et al. (2007) gave theoretical convergence criteria for incremental 4D-Var seen as an approximate Gauss-Newton method. These criteria cannot be evaluated for a real atmospheric system but the authors showed that when they are not satisfied, the algorithm can converge to a point which is not a minimum of the non-linear cost function. This is likely to be the case here,

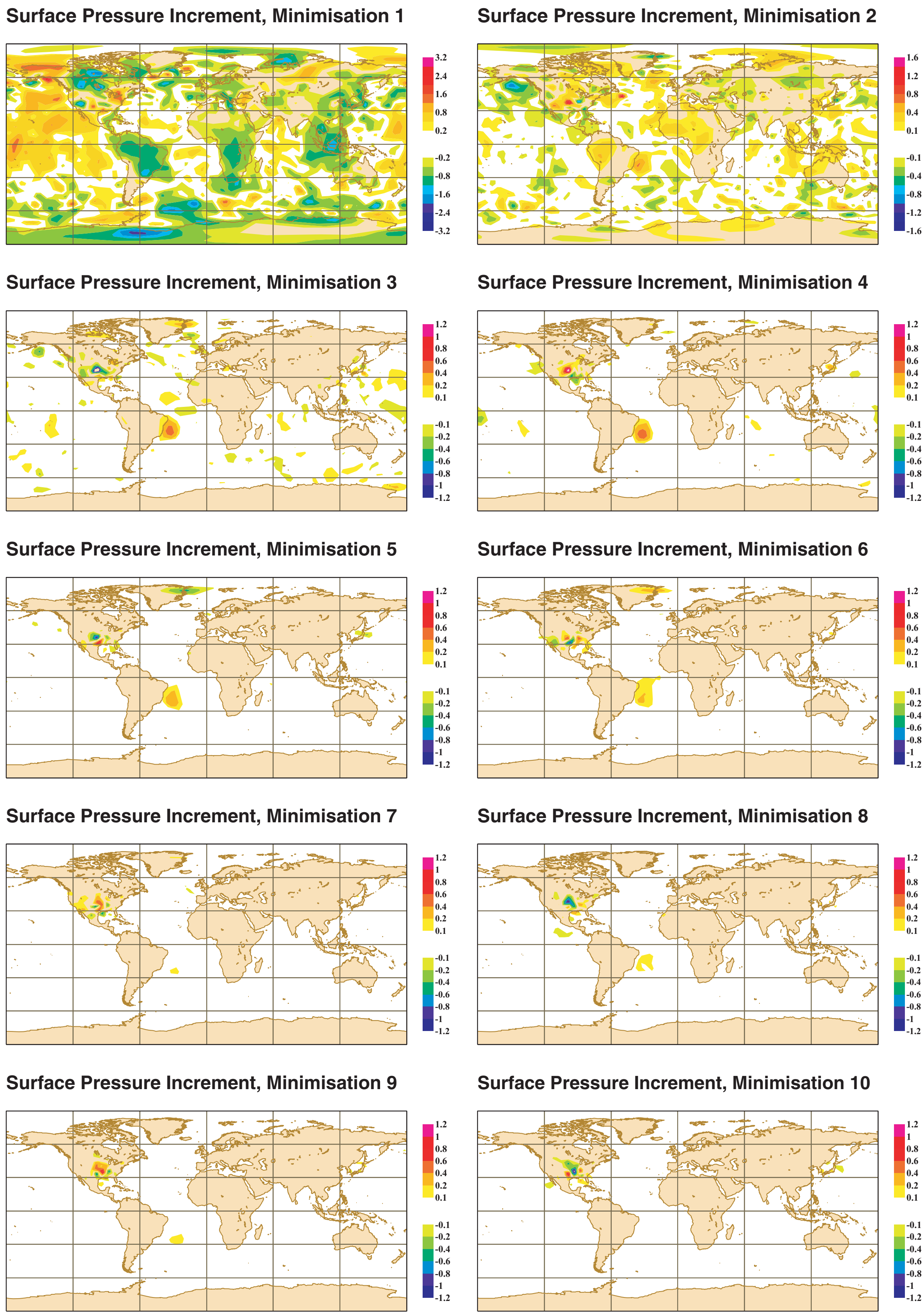

Fig. 11. Partial surface pressure increments for T255/T95 experiment with doubled observation error for surface pressure observations (hPa).

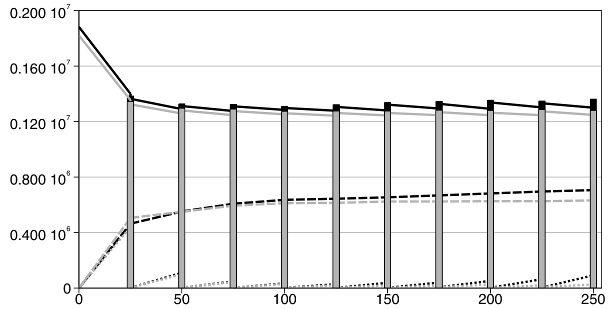

Fig. 12. Evolution of T255/T95 4D-Var cost function with when surface pressure observation error is doubled (grey) with respect to the control experiment (black).

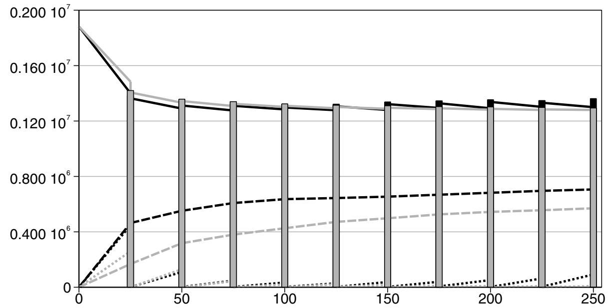

Fig. 13. Evolution of T255/T95 4D-Var cost function with m1qn3 (grey) and CONGRAD (black), both with 25 iterations per minimization.

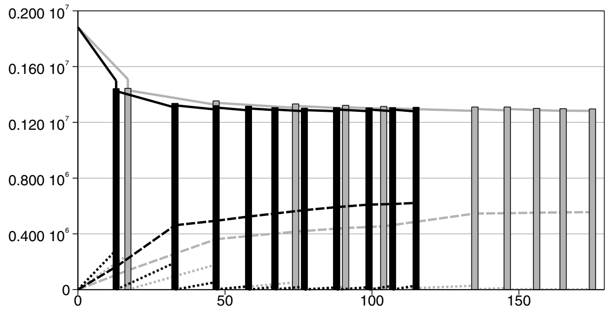

Fig. 14. Evolution of T255/T95 4D-Var cost function with m1qn3 (grey) and CONGRAD (black), both with a stopping criterion of 0.25.

although, in 4D-Var, the cost function which includes an atmospheric forecast model and other non-linear components, can have multiple local minima. The present study does not prove whether the various configurations of the algorithm find different local minima of the cost function or converge to points which are not always local minima of the non-linear cost function.

Currently, because of the computational cost of high resolution trajectories, too few outer iterations are performed in operational weather forecasting for the result to be sensitive to the minimization algorithm used in the inner loop. The preconditionned conjugate gradient algorithm can be used in that context since it converges faster thus saving computer time. More extensive investigation of this issue should be carried out in the future, studying the impact of other minimization algorithms and other preconditionners.

Care should be taken in evaluating preconditionners as the instability highlighted in this study depends on factors such as the distribution of observations and gravity wave propagation which are not easily reproduced in simplified systems. Experiments such as the idealized case presented in Section 3.3 or experiments in an operational context are very useful in that respect and complement theoretical and simplified studies.

4.3. Multi-incremental algorithm

In operational practice, incremental 4D-Var is run with two outer iterations. In order to reduce the computational cost, the resolution in the first minimization is lower than in the second minimization. It is possible to apply this idea with more than two outer loops and increase progressively the resolution in successive minimizations as experimented by Veersé and Thépaut (1998). Fisher (2003) showed that Hessian eigenvectors are mostly large-scale and can be computed at low resolution to form an effective preconditioner for the higher resolution inner loop minimizations that follow. This makes a multi-incremental setup where the resolution is increased in successive minimizations very attractive. This also has the effect of limiting the accuracy of the first inner loop minimizations thus reducing the introduction of noise and spurious gradients resulting from overfitting observations.

Experiments were run where the first three minimizations were run at T42, T95 and T159, respectively, the remaining minimizations being run at T255. Figure 15 shows the evolution of the cost function as a function of the number of equivalent T255 iterations based on the computational cost at each resolution. This setup does give the same final values of J_o and J_b as 4D-Var with T255 minimizations throughout. The increments produced in the last four minimization in both setups are surprisingly similar (not shown) as are the number of iterations for each of these four minimizations (visible on Fig. 15). The multiresolution incremental setup does allow that solution to be reached significantly faster, thus saving computer time.

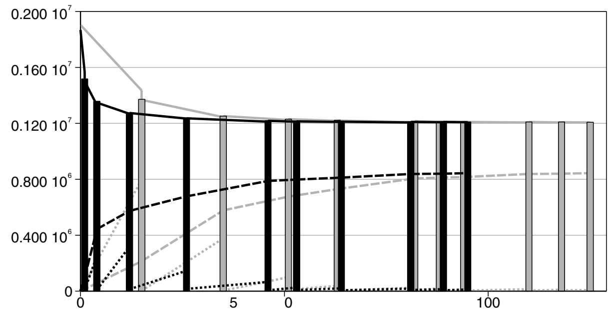

Fig. 15. Evolution of T42/T95/T159/T255 multiresolution incremental 4D-Var cost function (black) and T255 incremental 4D-Var (grey), both with T255 outer loops.

Table 1. Average final cost function values with two and three outer 4D-Var iterations for the month of May 2006

O.L.	J_o	J_o/p	J_b
2	1360 887	0.2967	78 804
3	1346 637	0.2943	73 968

At operational resolution, because of the significant computational cost of the T799 high resolution non-linear trajectories, only a few outer loop iterations are affordable. Currently at ECMWF, 4D-Var is run with two outer loop iterations. The first minimization is run at T95 for 70 iterations, the second one at T255 for, on average, 35 iterations. The number of iterations in the first minimization is fixed and determined by the forecast error computation implemented by Fisher and Courtier (1995). From the experimental results presented above, it is expected that more non-linear updates with reduced accuracy in the first minimizations could be beneficial.

An experiment at operational resolution was performed with three outer loops. The first minimization remained unchanged and the following two minimizations were run at T159 and T255 with a relaxed stopping criterion chosen to maintain the overall computational cost constant on average. The facts that the second minimization is run at the intermediate resolution of T159 instead of T255 and that the number of iterations in each minimization is reduced compensate for the cost of the additional high resolution non-linear trajectory. With the same total computational cost, slightly lower values of both J_o and J_b are obtained (Table 1). This algorithm yields a better solution, providing a better fit to observations with a smaller increment. This configuration with three outer iterations also showed improved forecast skills and has become operational at ECMWF in June 2007.

No evidence is available to prove whether the three outer iteration configuration finds different local minima or whether the two outer iteration configuration stopped at points that were not local minima of the non-linear cost function. However, the fact that the improvement is systematic tends to be in favour of the second hypothesis since two algorithms being trapped in different local minima are more likely to produce a more random distribution of cases and, because the first minimization being the same, it should steer the minimization towards the same local minimum.

5. Conclusions

Experiments presented in this paper have shown that it is the combination of observations, through their distribution and error characteristics (that determine the leading eigenvectors of the minimization problem), and of the use of the model in the assimilation algorithm, through the phase speed of gravity waves, that currently determines the convergence or divergence of 4D-Var.

The preconditioned Lanczos-conjugate gradient algorithm is very efficient and allows for fast convergence of the inner loop minimizations. However, it emphasizes the direction of the leading eigenvectors of the Hessian of the cost function and fits the data very accurately in those directions. Independent studies have shown that oversolving the inner loop minimization can trap the incremental algorithm away from the true minimum by fitting observational noise and introducing spurious gradients. The minimization algorithm used in the IFS can lead to the same phenomenon by oversolving the inner loop in the subspace spanned by the leading eigenvectors. Moreover, in the IFS operational configuration, the shape of the cost function is such that these increments, in the direction of the leading eigenvector, generate gravity waves that propagate with different phase speeds in the inner and outer loops. Ultimately, that leads to a positive feedback effect and to the divergence of the algorithm.

We have shown that increasing observation error for surface pressure observations reduces the eigenvalue associated with the leading eigenvector and the overfitting of these observations, leading to the convergence of the algorithm. Another possibility to avoid introducing an increment in the unstable direction of the leading eigenvector is to use another minimization algorithm which does not emphasize the direction of the leading eigenvectors such as the quasi-Newton algorithm. This algorithm is less likely to overfit the observations and produce an increment in that unstable direction but at the cost of a slower convergence in the inner loop minimizations.

We have highlighted two issues regarding the convergence of incremental 4D-Var, both related to overfitting of observations. The first difficulty is that overfitting observations in the direction of the leading eigenvector of the minimization problem generates gravity waves in the linear and non-linear models that lead to the divergence of the incremental algorithm. The second more general issue is that overfitting observations can cause incremental 4D-Var to converge to a point which is not the minimum of the non-linear cost function. Solving the first issue will be necessary if more outer iterations are to be used in the current implementation of incremental 4D-Var. The consequences of the second aspect of the problem are less ubiquitous and might be overlooked. However, ensuring that the solution of incremental 4D-Var is as close as possible to the minimum of the non-linear cost function is important and can lead to improved forecast performance.

We have shown in an operational context that running three outer loop iterations with reduced accuracy, for the same global computing cost as the current two outer iterations, leads to a solution with lower values of both J_o and J_b. The solution has a better fit to observations and background at the same time and is more likely to be close to the minimum of the non-linear cost function. It has also shown improved forecast skills and has recently become operational.

Independently of the convergence properties, running more outer loops provides more opportunities for 4D-Var to take non-linear effects into account. This will become more important in the future as resolution increases and observations that depend on more non-linear observation operators are assimilated. With that prospect in view, it is important to understand the properties of convergence of the incremental algorithm.

In the specific case we have identified, the leading eigenvector of the Hessian of the 4D-Var cost function is driven by accurate and dense surface pressure observations in Europe. At the current resolution of the data assimilation system, many of these observations are in the same model grid-box. Representativeness error and observation error correlations should reflect that fact, however, at the moment, these are ignored in the IFS. A better specification of the statistical properties and error correlation of these observations would reduce the global weight given to these observations and alleviate the incremental 4D-Var convergence problems. Further research work aimed at accounting for observation error correlation is on-going and could affect the convergence of incremental 4D-Var.

Since this particular case of divergence is partly due to gravity waves, existing techniques to eliminate gravity waves from the data assimilation system can be applied. Increasing the weight given to the weak constraint digital filter in the J_c term of the cost function, in particular applying it to the surface pressure component of the control variable, could dampen the external gravity waves seen in this study. This would alleviate the positive feedback effect described in this paper and should be pursued if more outer iterations were to be used.

Finally, research should continue into minimization algorithms and preconditionners which could provide faster convergence and be less prone to generate increments in unstable directions. Currently, the unstable directions are related to external gravity waves and surface pressure observations. Future changes in the observing system could have an impact on the leading eigenvectors of the Hessian of the cost function and modify the convergence properties of incremental 4D-Var. More research would be necessary to determine an algorithm that could remain stable regardless of this type of changes. Until then, it will be necessary to re-examine the convergence of incremental 4D-Var at outer loop level periodically.

6. Acknowledgments

The author would like to thank colleagues at ECMWF for their comments on earlier versions of this manuscript, especially Mike Fisher for discussions about CONGRAD and Sami Saarinen for his advice on modifying the ODB to make it possible to run with many outer loop iterations.

References

Andersson, E., Bauer, P., Beljaars, A., Chevallier, F., Hólm, E. and co-authors. 2005. Assimilation and modeling of the atmospheric hydrological cycle in the ECMWF forecasting system. *Bull. Amer. Meteorol. Soc.* **86**, 387–402.

Andersson, E., Cardinali, C., Fisher, M., Hólm, E., Isaksen, L. and co-authors. 2004. Developments in ECMWF's 4D-Var System. In *AMS Symposium on Forecasting the Weather and Climate of the Atmosphere and Ocean*. Available from the American Meteorological Society, http://ams.confex.com/ams/84Annual/20WAF16NW/program.htm.

Andersson, E., Fisher, M., Munro, R. and McNally, A. 2000. Diagnosis of background errors for radiances and other observable quantities in a variational data assimilation scheme, and the explanation of a case of poor convergence. *Q. J. R. Meteorol. Soc.* **126**, 1455–1472.

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.

Fisher, M. 1998. Minimization algorithms for variational data assimilation. In: *Seminar on Recent Developments in Numerical Methods for Atmospheric Modelling*, ECMWF, pages 364–385.

Fisher, M. 2003. Background error covariance modelling. In *Workshop on Recent developments in data assimilation for atmosphere and ocean*, ECMWF, pages 45–63.

Fisher, M. and Courtier, P. 1995. Estimating the covariance matrices of analysis and forecast error in variational data assimilation. Tech. Memo. 220, ECMWF.

Gauthier, P. and Thépaut, J.-N. 2001. Impact of the Digital Filter as a weak constraint in the preoperational 4DVAR assimilation system of Meteo-France. *Mon. Wea. Rev.* **129**, 2089–2102.

Gilbert, J.-C. and Lemaréchal, C. 1989. Some numerical experiments with variable storage quasi-Newton algorithms. *Math. Program.* **45**, 407–435.

Gratton, S., Lawless, A. and Nichols, N. 2007. Approximate gauss–newton methods for nonlinear least squares problems. *SIAM J. Opt.*, **18**(1), 106–132.

Klinker, E., Rabier, F., Kelly, G. and Mahfouf, J.-F. 2000. 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.

Laroche, S. and Gauthier, P. 1998. A validation of the incremental formulation of 4D variational data assimilation in a nonlinear barotropic flow. *Tellus* **50A**, 557–572.

Lawless, A., Gratton, S. and Nichols, N. 2005. An investigation of incremental 4D-Var using non-tangent linear models. *Q. J. R. Meteorol. Soc.* **131**, 459–476.

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

Rabier, F., Järvinen, 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.

Trémolet, Y. 2004. Diagnostics of linear and incremental approximations in 4D-Var. *Q. J. R. Meteorol. Soc.* **130**, 2233–2251.

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.