TECHNICAL MEMORANDUM

469

Incremental 4D-Var Convergence Study

Yannick Trémolet

Research Department

July 6, 2005

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

Series: ECMWF Technical Memoranda

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

Contact: library@ecmwf.int

European Centre for Medium-Range Weather Forecasts
Shinfield Park, Reading, RG2 9AX, England

The information within this publication is given in good faith and considered to be true, but ECMWF accepts no liability for error, omission and for loss or damage arising from its use.

Abstract

Since its operational implementation at ECMWF, incremental 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 nonlinear 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 minimisation algorithm. The influence of the choice of the inner loop minimisation 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 consists in constructing a suitable initial condition using the observations and the model. ECMWF uses the four-dimensional variational data assimilation (4D-Var) method as described by Le Dimet and Talagrand (1986). The principle of the method is to minimise 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 didn't 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.

As the system has evolved to higher resolution and will continue to do so in the coming years, more small scales are resolved and nonlinear phenomena in the model become more important. Nonlinear phenomena are also increasingly important as assimilation of new types of observations related to clouds and rain are being developed (Andersson et al. (2005)). The normal mode initialisation, 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 initialisation as described by Gauthier and Thépaut (2001). 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 deficiencies. Section 4 will detail some of the reasons for the behaviour exhibited in the previous section and, in section 5, we propose possible improvement for the current operational system.

2 The incremental 4D-Var minimisation problem

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

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

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, $\mathscr{H}$ is the nonlinear observation operator and J_c is an initialisation term used to control gravity waves. $\mathscr{H}$ 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:

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

where $H = \frac{\partial \mathscr{H}}{\partial x}$ is the linearised observation operator and $d = y - \mathscr{H}(x_b)$ is the departure from observations. In this notation, the tangent linear model is embedded in the linearised observation operator. The gradient of the cost function with respect to the initial condition is obtained using the adjoint model. A nonlinear integration provides the trajectory around which the tangent linear and adjoint models and the observation operators are linearised. It is called trajectory run. The departures d are also computed in this trajectory run.

The approximate minimisation problem thus defined is solved using an iterative algorithm: this is the inner loop of 4D-Var. Currently at ECMWF, a preconditioned Lanczos-conjugate gradient algorithm (`CONGRAD`) is used to solve the inner loop minimisation problem as described by Fisher (1998). After this minimisation, the departures and trajectory can be recomputed using the nonlinear model and a new linearised problem is defined. The process can be repeated: this is the outer loop of incremental 4D-Var. If the linearised problem is reasonably close to the nonlinear problem, as tested in Trémolet (2004), its solution should be an approximation of the solution of the nonlinear problem. At the next outer loop iteration, the starting point is closer to the solution, the first order approximation is more accurate and provides a more accurate solution of the full problem. The algorithm should converge to the solution of the nonlinear problem, although there is no general theoretical proof of convergence.

In order to reduce the computational cost of the assimilation, the inner loop is run at lower resolution than the forecast. Currently at ECMWF, two iterations of the outer loop are run, the first one at T95, the second one at T159 while the outer loops and forecast run at T511. For further reduction of the computational cost, simplified linear physics is used in the first inner-loop minimisation. Note that the incremental algorithm is independent of these further approximations. The incremental 4D-Var algorithm is shown schematically on figure 1.

Figures 2 and 3 show the evolution of the various components of the cost function and its gradient in a configuration similar to today's operational 4D-Var. The only two changes with respect to an operational IFS CY29R1 analysis were to run ten outer-loop iterations instead of two and to limit the number of inner loop iterations for each of the second and following minimisations to 25.

On figure 2 and all similar figures in this paper, solid lines represent the J_o component of the cost function, as seen in the inner loop, the bars represent the value of J_o as seen in the nonlinear 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 evaluation of the cost function, at the start and at the end of the minimisation. For the figures, they are joined by straight lines as intermediate information is not available. Also note that J_b has been inflated by a factor of 10 and J_c by a factor of 100 to make the figures more readable.

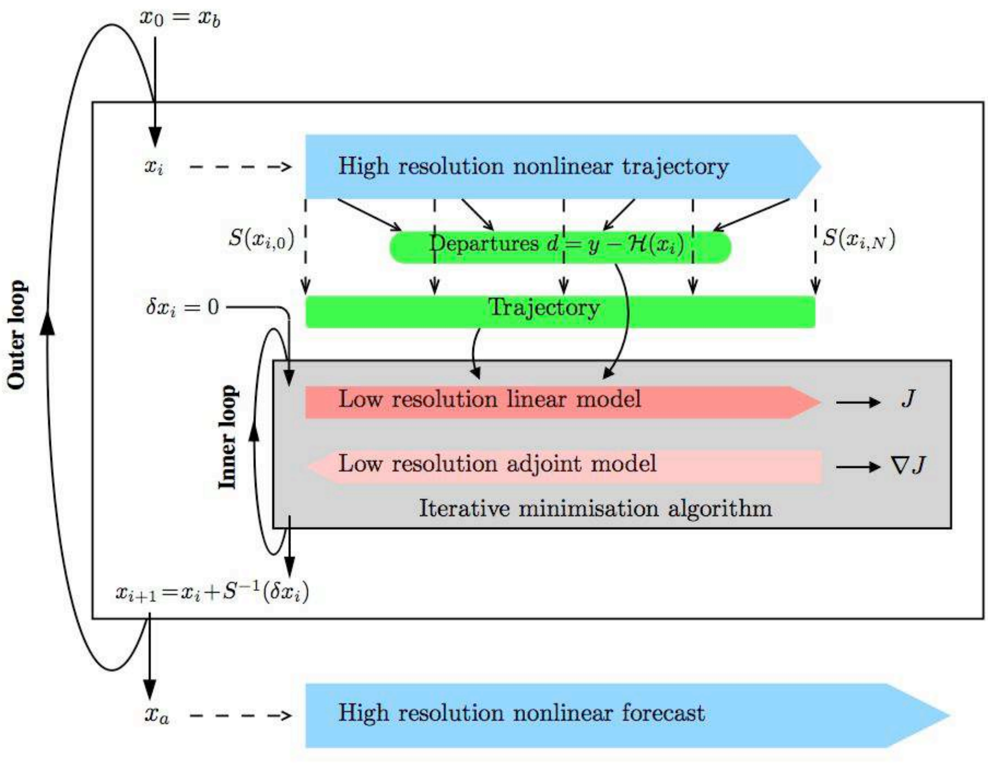

$x_0 = x_b$
x_i
High resolution nonlinear trajectory
$S(x_{i,0})$
Departures $d = y - \mathcal{H}(x_i)$
$S(x_{i,N})$
$\delta x_i = 0$
Trajectory
Outer loop
Inner loop
Low resolution linear model
J
Low resolution adjoint model
∇J
Iterative minimisation algorithm
$x_{i+1} = x_i + S^{-1}(\delta x_i)$
x_a
High resolution nonlinear forecast

Figure 1: Incremental 4D-Var algorithm. Departures d from observations y and trajectory are computed at high resolution. 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 current first guess. The process is repeated (outer loop, subscript i) until the analysis x_a is obtained.

The figure shows that the values of J_o at the end of a minimisation, in the following high resolution trajectory and at the starting point of the ensuing minimisation do not coincide. The jump in J_b between the end of the first minimisation and the begin of the second minimisation is due to the change of resolution and the implied change in the B matrix.

Figure 3 shows the evolution of the gradient of the cost function for the same experiment. The plain curve shows the gradient norm estimated by CONGRAD (the Lanczos-conjugate gradient minimisation algorithm used in the IFS) during the minimisation, the dashed curve shows the actual gradient computed before and after the minimisation, joined in a straight line. The increase of the gradient norm at the beginning of each minimisation is discussed in appendix A.

The first figure shows that the minimum value for J_o is obtained after four outer loops iterations. After that, J_o starts increasing slowly and 4D-Var diverges. The gradient does continue to decrease for two more outer loop iterations but then increases as well. In the remainder of this paper, we investigate this behaviour.

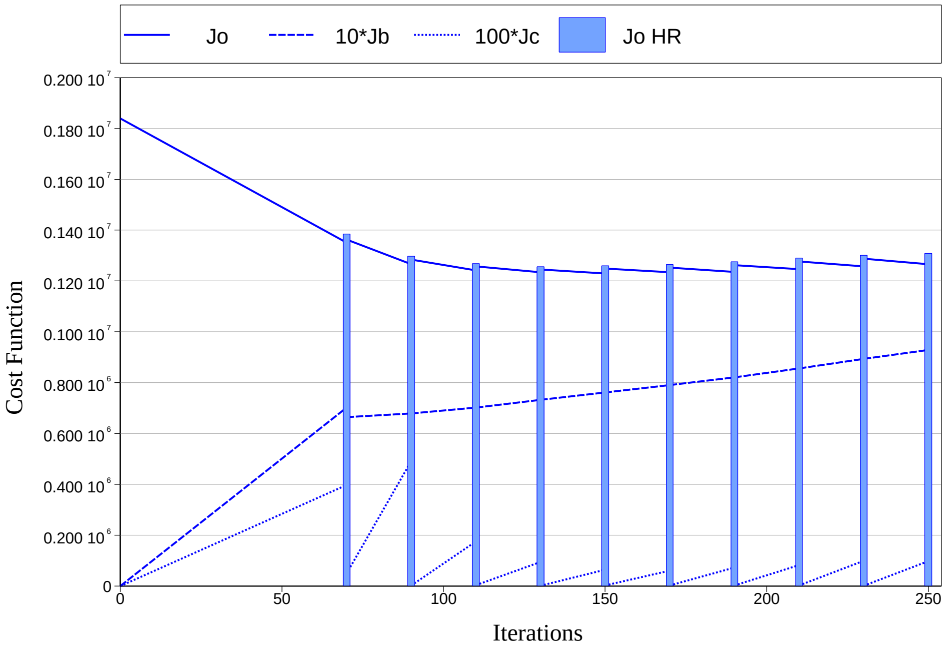


Figure 2: Evolution of the components of 4D-Var cost function as a function of the total number of inner loop iterations. The solid lines represent J_o as computed in the inner loop while the bars represent J_o as computed in the nonlinear trajectory runs. J_b (dashed lines) has been inflated by a factor of 10 and J_c (dotted lines) by a factor of 100 for readability.

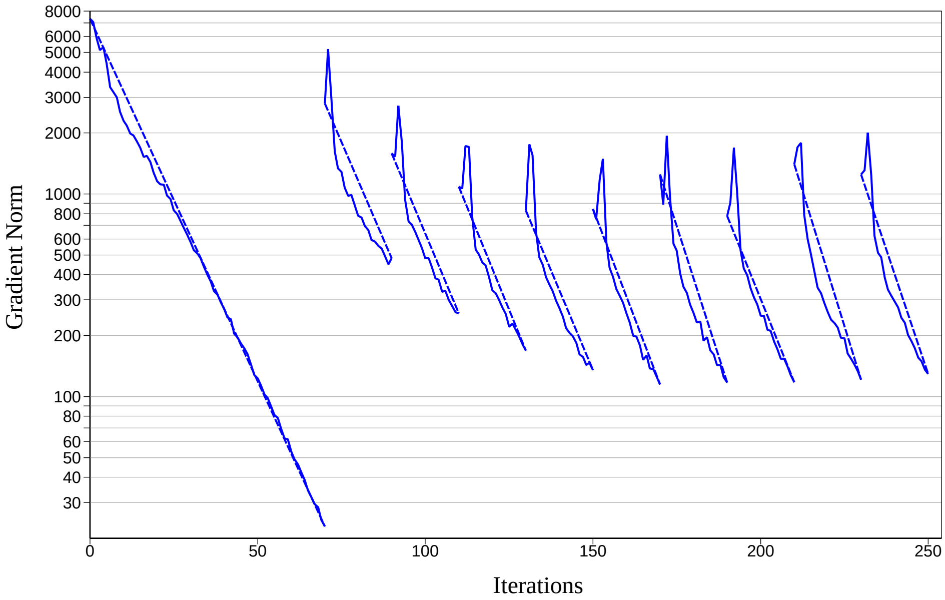


Figure 3: Evolution of the norm of the gradient of the cost function in 4D-Var as a function of the total number of inner loop iterations. The plain lines show the estimated gradient norm during the minimisation, the dashed lines show the actual gradient norm at the beginning and at the end of the minimisation.

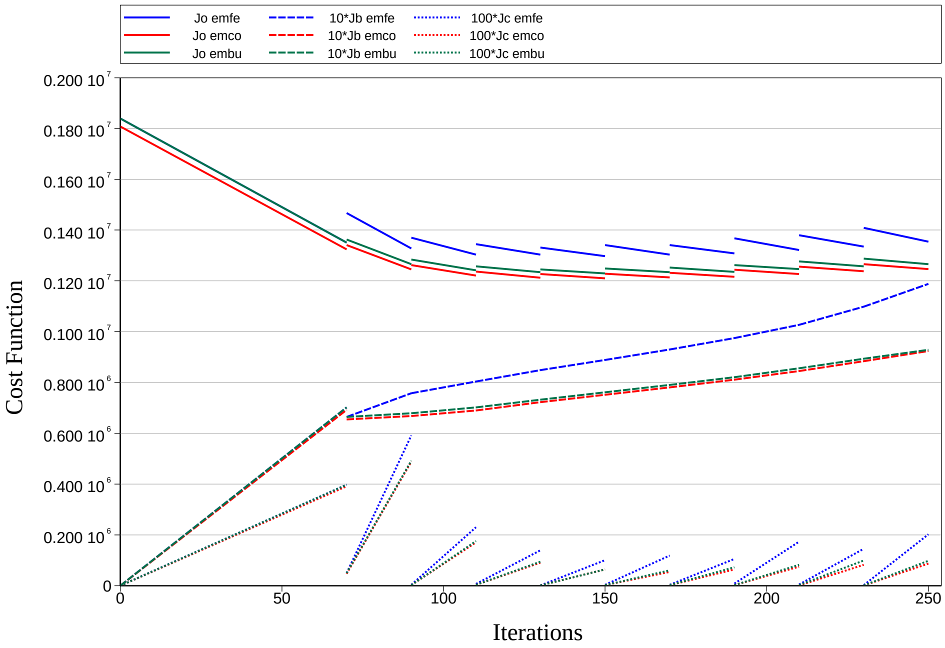


Figure 4: Evolution of 4D-Var cost function without VarQC (blue), without VarQC, scatterometer data or wave model coupling (red), the control experiment is shown in green.

3 Outer-loops convergence

3.1 Some discrepancies between inner and outer loops

Several differences exist between the high resolution outer loop trajectories and the inner loop minimisation in addition to the change of resolution. Variational quality control and ambiguous SCAT wind removal are applied at outer loop level, the coupled wave model is run only in the high resolution trajectories. All these have been removed, the results are presented on figure 4. It shows that 4D-Var diverges even more in that case, when some discrepancies have been removed and one would expect better agreement between inner and outer loops and thus better convergence. The mismatches in J_o between the inner and outer loops are also larger. The slightly lower J_o values are due to the removal of scatterometer data, the difference remains constant showing that this does not affect the convergence of 4D-Var. Other experiments were run where only one or several of these processes were deactivated. They indicate that removing the variational quality control (VarQC) is deteriorating convergence the most. VarQC rejects observations from which the solution is moving away. In doing that, it comforts the current estimate of the solution 4D-Var is producing. 4D-Var will not try to fit these difficult observations in the next minimisation. The problem becomes easier and convergence is improved. Overall, these slightly different settings between the inner and outer loops do not appear to be the cause of the lack of convergence of the minimisation.

3.2 Inner Loop Resolution

The impact of the difference in resolution between the inner and outer loops has been tested. In all experiments presented below, the outer loop was run at T255. This choice was made in order to be able to run the inner

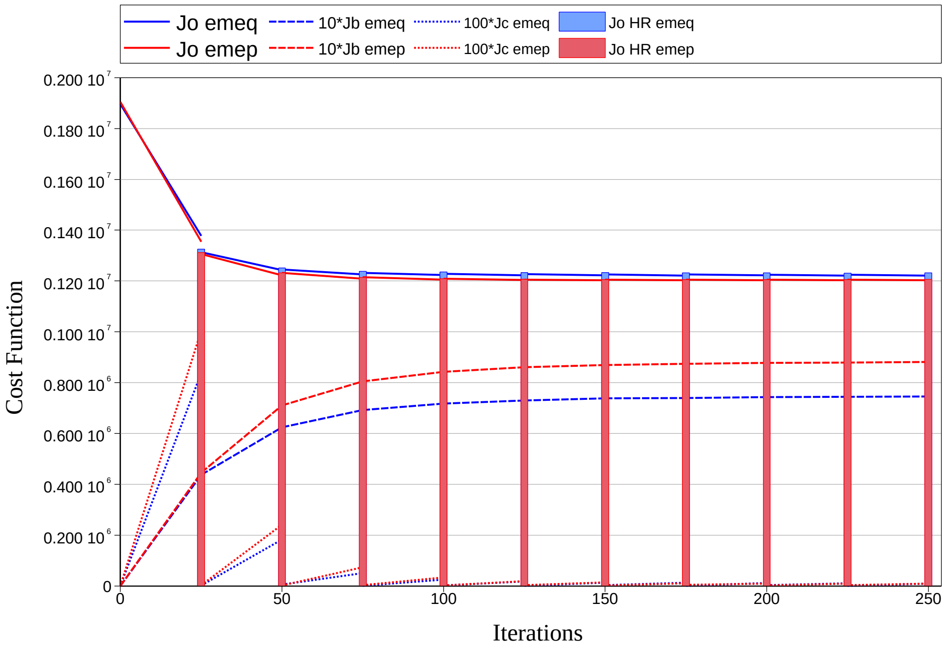


Figure 5: Evolution of T255/T159 (blue) and T255/T255 (red) 4D-Var cost functions.

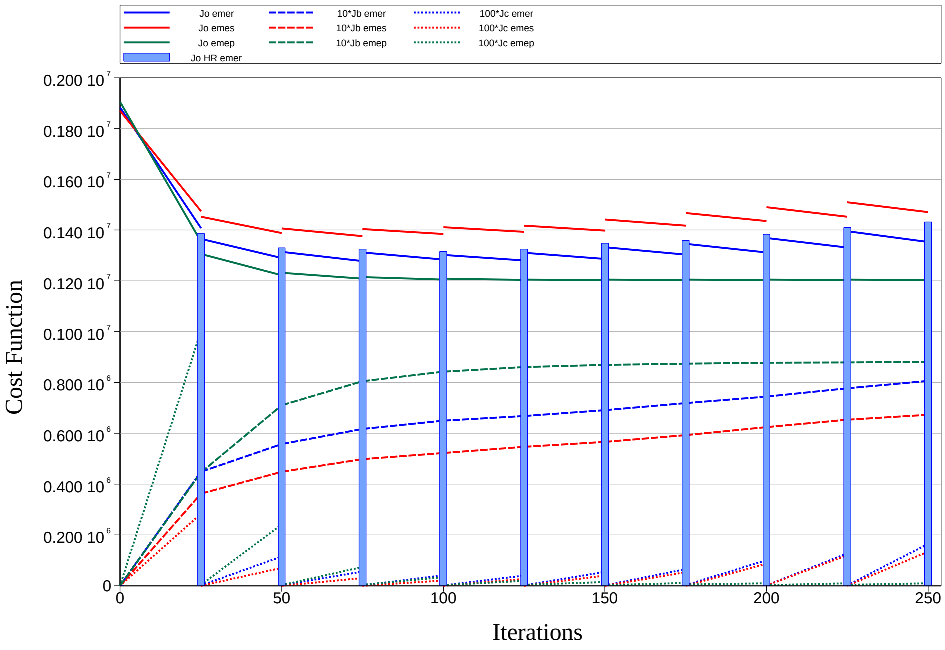


Figure 6: Evolution of T255/T42 (red), T255/T95 (blue) and T255/T255 (green) 4D-Var cost function.

loop at the same resolution[1]. All the minimisations are run with linearised physics (see IFS documentation [2], chapter II, section 2.6). The number of iterations is fixed to 25 per minimisation and the resolution is the same for all minimisations within each experiment. All other aspects are kept to the defaults[3] of IFS CY29R1. Figures 5 and 6 show the evolution of the cost function for inner loop resolutions of T255, T159, T95 and T42. When inner and outer loop resolutions are the same, figure 5 shows that 4D-Var does converge (this is also true for a T159/T159 experiment, not shown). 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 more and more. The discrepancies between J_o component of the cost function at the end of a minimisation, in the high resolution trajectory and at the beginning of the following minimisation disappear with the T159 and T255 inner loops and increase with the mismatch in resolution. The mismatch in resolution between inner and outer loops seems to be the most important factor for 4D-Var to converge or diverge. Notice that this does not prove that the solution is the correct one, even in the T255/T255 case.

3.3 Increments

The RMS values of the increments produced by the previous experiments with a T95 and T255 inner loops are shown as vertical profiles in figure 7 for the first two and last two minimisations. The increments for the first minimisations are the largest and similar in amplitude for both inner loop resolutions (the plain and dashed lines are similar). The increments from the last minimisations obtained with the T255 inner loops are smaller in amplitude (dashed green and magenta curves), indicating that the algorithm has properly converged and that no additional increments are needed.

The maps of temperature increments at level 49 (approx. 850hPa) (figures 8 and 9) show that the increments are also more localised with the higher resolution inner loop. The maps of surface pressure increments (figures 10 and 11) 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 plot, although not as clearly marked.

[1] It is now possible to run with T319 inner loops

[2] Available at `http://www.ecmwf.int/research/ifsdocs/`

[3] `LVERIFY_SCREEN` had to be set to false in order to get the correct J_o value in the last high resolution trajectory. This has no impact on the minimisation.

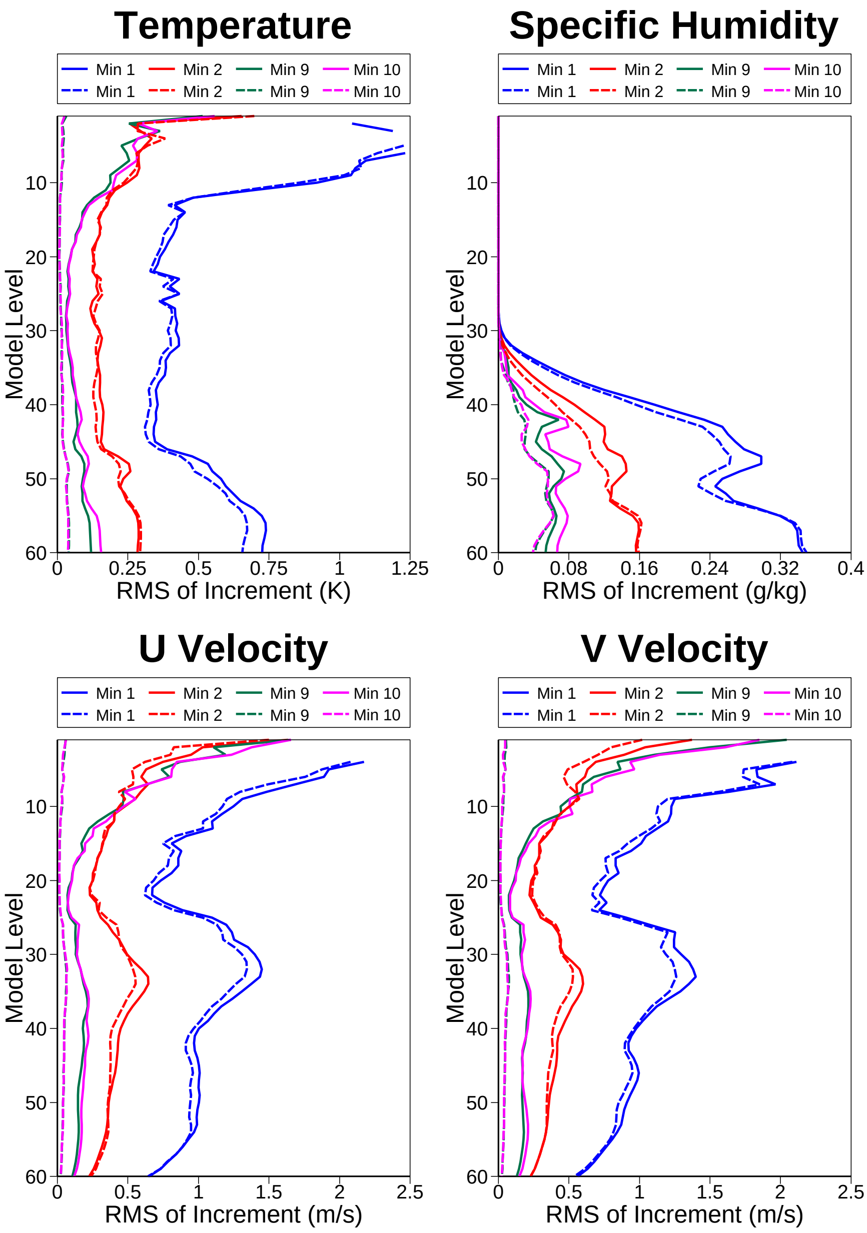


Figure 7: RMS of 4D-Var increments for minimisations 1, 2, 9 and 10 obtained with T95 (solid lines) and T225 (dashed lines) inner loops, both with T255 outer loops.

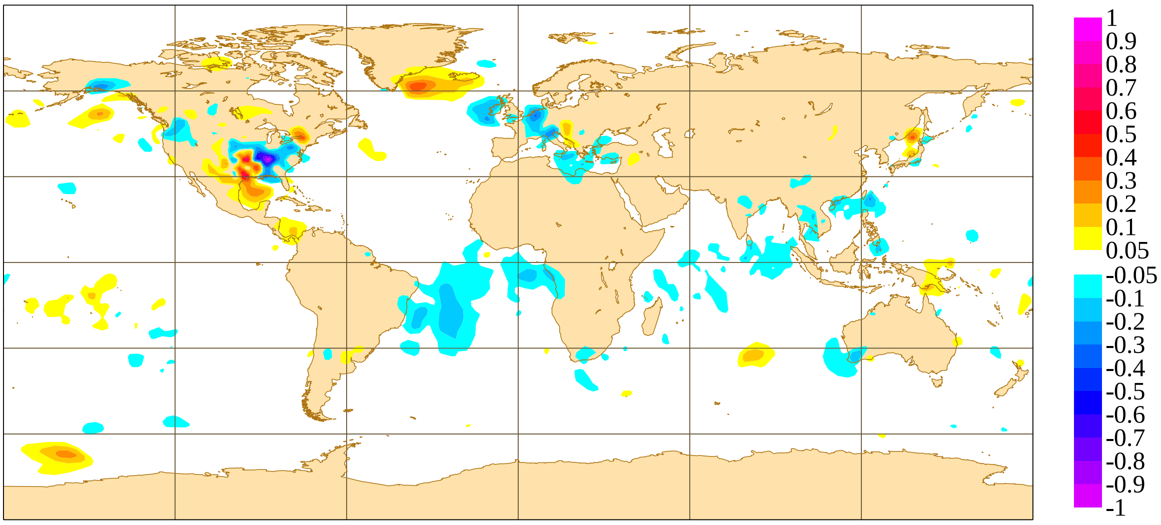


Figure 8: Temperature level 49 (approx 850hPa) increment at minimisation 7 for T255/T255 experiment.

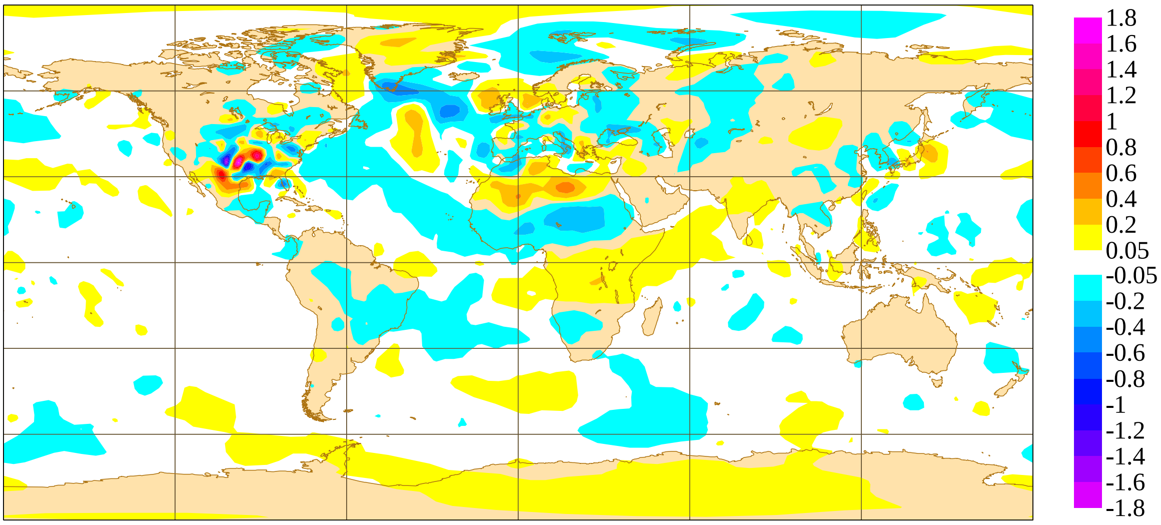


Figure 9: Temperature level 49 (approx 850hPa) increment at minimisation 7 for T255/T95 experiment.

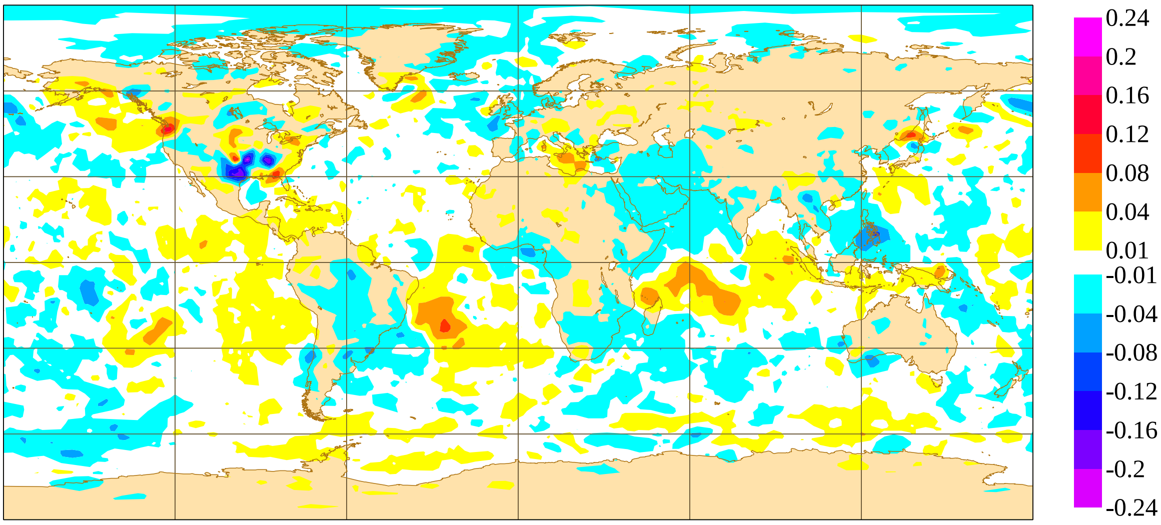


Figure 10: Surface pressure increment at minimisation 7 for T255/T255 experiment.

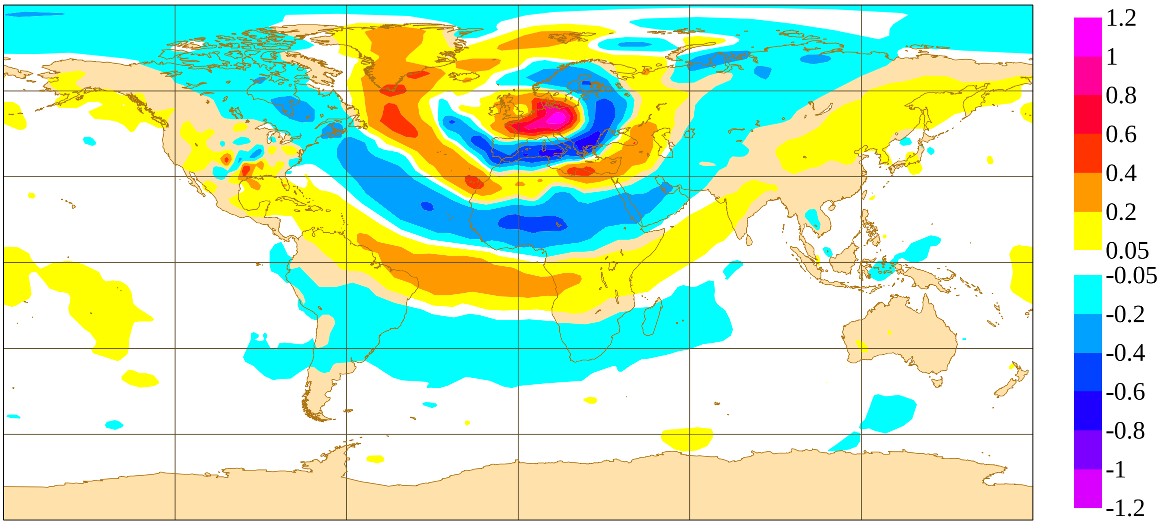


Figure 11: Surface pressure increment at minimisation 7 for T255/T95 experiment.

3.4 Time-step

The change of resolution between inner and outer loops implies a change of time-step. T255 and T159 inner loops experiments presented here were run with a time step of half an hour while T95 and T42 integrations were run with a one hour time-step. In order to separate the effect of spatial resolution from temporal resolution, figure 12 shows the 4D-Var cost function for a T255/T95 experiment using half an hour time-step in both inner and outer loops. We see that, with the shorter time step, most of the increase in J_o after iteration 100 has been eliminated. The magnitude of increments has reduced (not shown) in the later outer loop iterations. Figures 13 and 14 show that the spurious pattern in increments at minimisation 7 have disappeared as well. 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 (figure 5), 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.

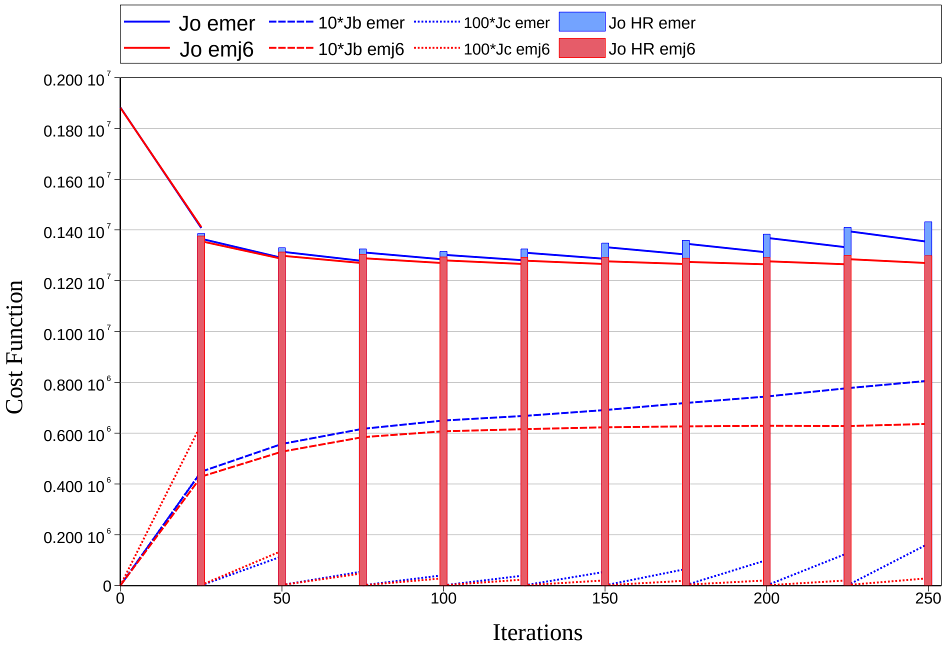


Figure 12: Evolution of 4D-Var cost function components for T255/T95 experiment with an inner loop time step of 1800 seconds (in red) and 3600 seconds (in blue).

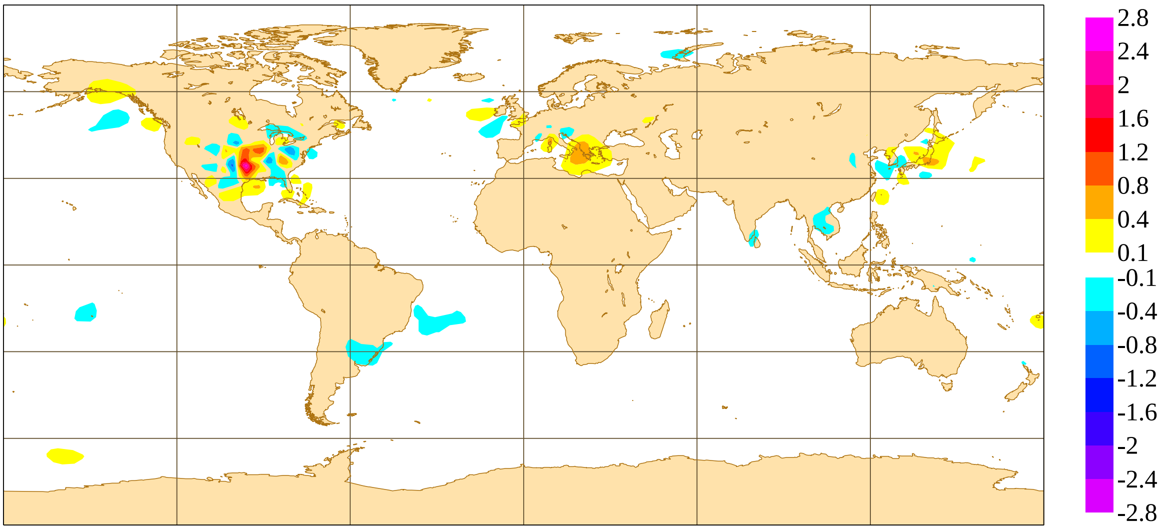


Figure 13: Temperature level 49 (approx 850hPa) increment at minimisation 7 for T255/T95 experiment with a time step of 1800 seconds in inner and outer loops, to be compared with figure 9

.

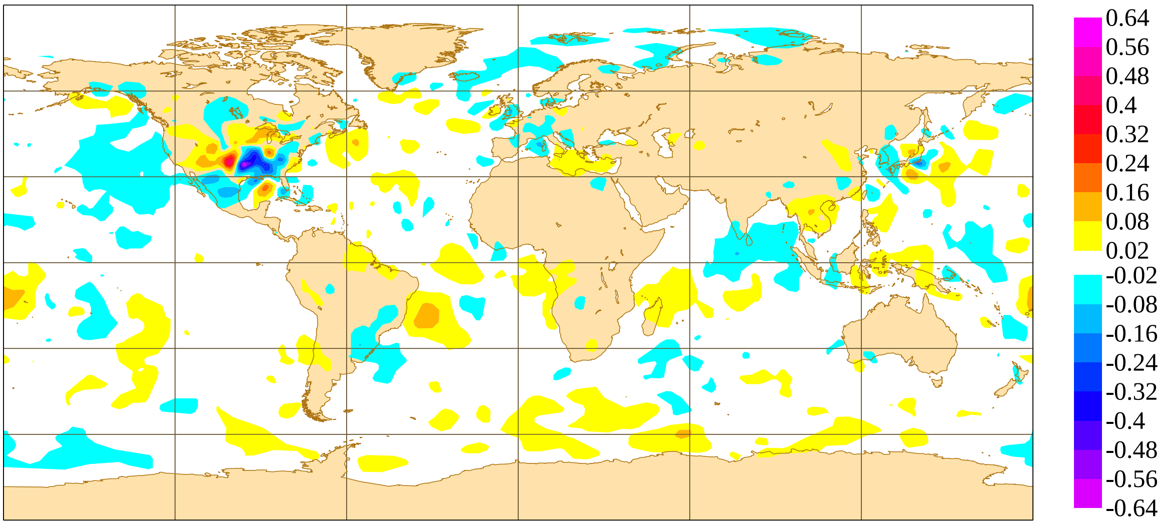


Figure 14: Surface pressure increment at minimisation 7 for T255/T95 experiment with a time step of 1800 seconds in inner and outer loops, to be compared with figure 11

.

3.5 Perfect solution case study

It is possible to create a *perfect solution* 4D-Var case (also called identical twin experiment) by replacing the observed values for all the observations used in a 4D-Var cycle by their model equivalent (run from the background). 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 and the background is the exact solution of the problem. 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 observation types matches exactly a real life situation.

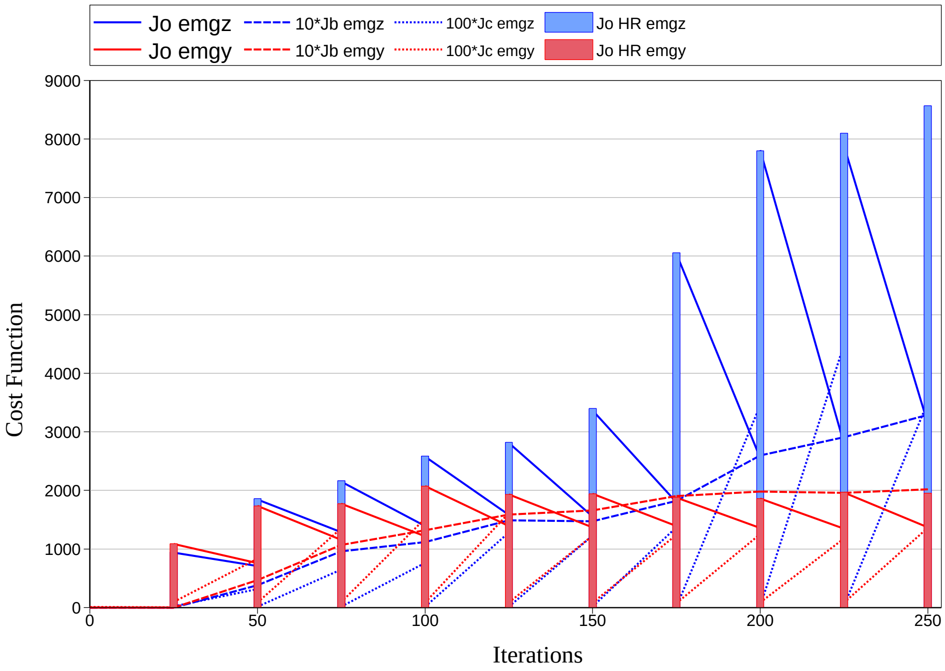


Figure 15: Evolution of T159/T159 (in red) and T159/T95 (in blue) 4D-Var cost function components in the perfect solution case where the correct solution should remain zero.

In the first identical twin experiment, inner and outer loops are run at T159, figure 15 shows that 4D-Var does produce a small increment. The value of J_o tends to stabilise at around 2000 ($J_o/n = 0.0008$, where n is the total number of observations, to be compared with values of the order of 1.0 for a real case) after a few iterations. 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 (in blue on the figure). Again, the mismatch in resolution and time-step seems to be important for 4D-Var convergence or divergence.

In the first nonlinear trajectory, J_o is exactly zero. In principle, the first minimisation should start with an initial increment also exactly equal to zero. That would imply that the initial gradient is zero and no increment would be generated. In the following nonlinear trajectory, because of the supersaturation check, J_o would be non-zero. From there on, a small increment would be generated and 4D-Var should reach an equilibrium between the super-saturation removal and the background. This is probably what we see in the T159/T159 experiment after the first minimisation. The first minimisation generates a non-zero increment because the background and the trajectory at initial time are not generated in the same way (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 by the red curve on figure 15 is thus deemed acceptable. The blue curve behaviour is pathological and will be investigated below.

Figure 16 shows the surface pressure increment generated in the last minimisation of the T159/T95 experiment. The pattern is the same as seen previously, although it appears later in the iterations and with a smaller amplitude. This shows that this pattern does not correspond to an actual increment since we know the correct solution is $\delta x = 0$ in this case. It is an artifact of the minimisation process.

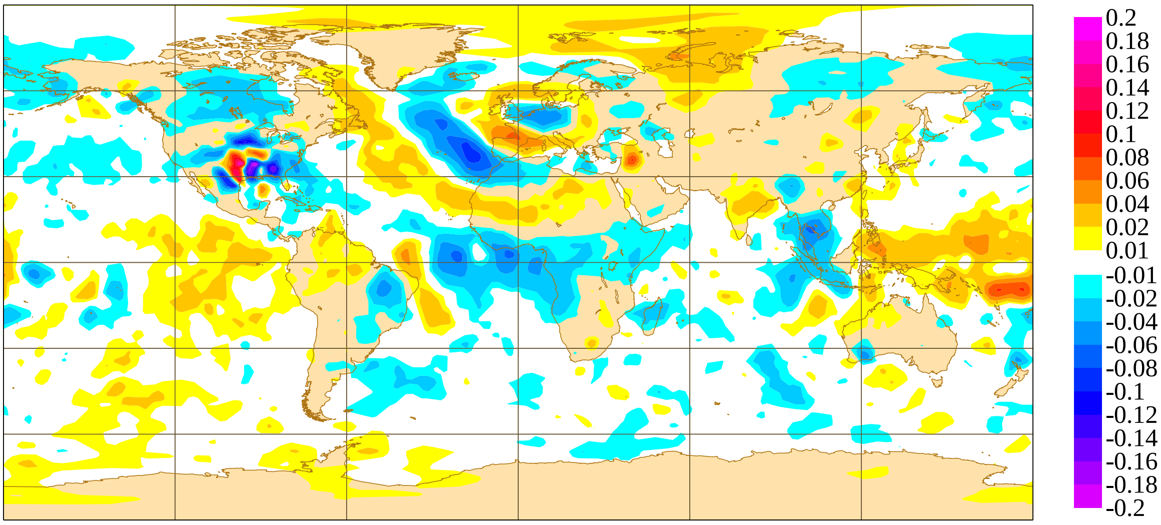


Figure 16: Surface pressure increment at minimisation 9 for T159/T95 perfect solution experiment.

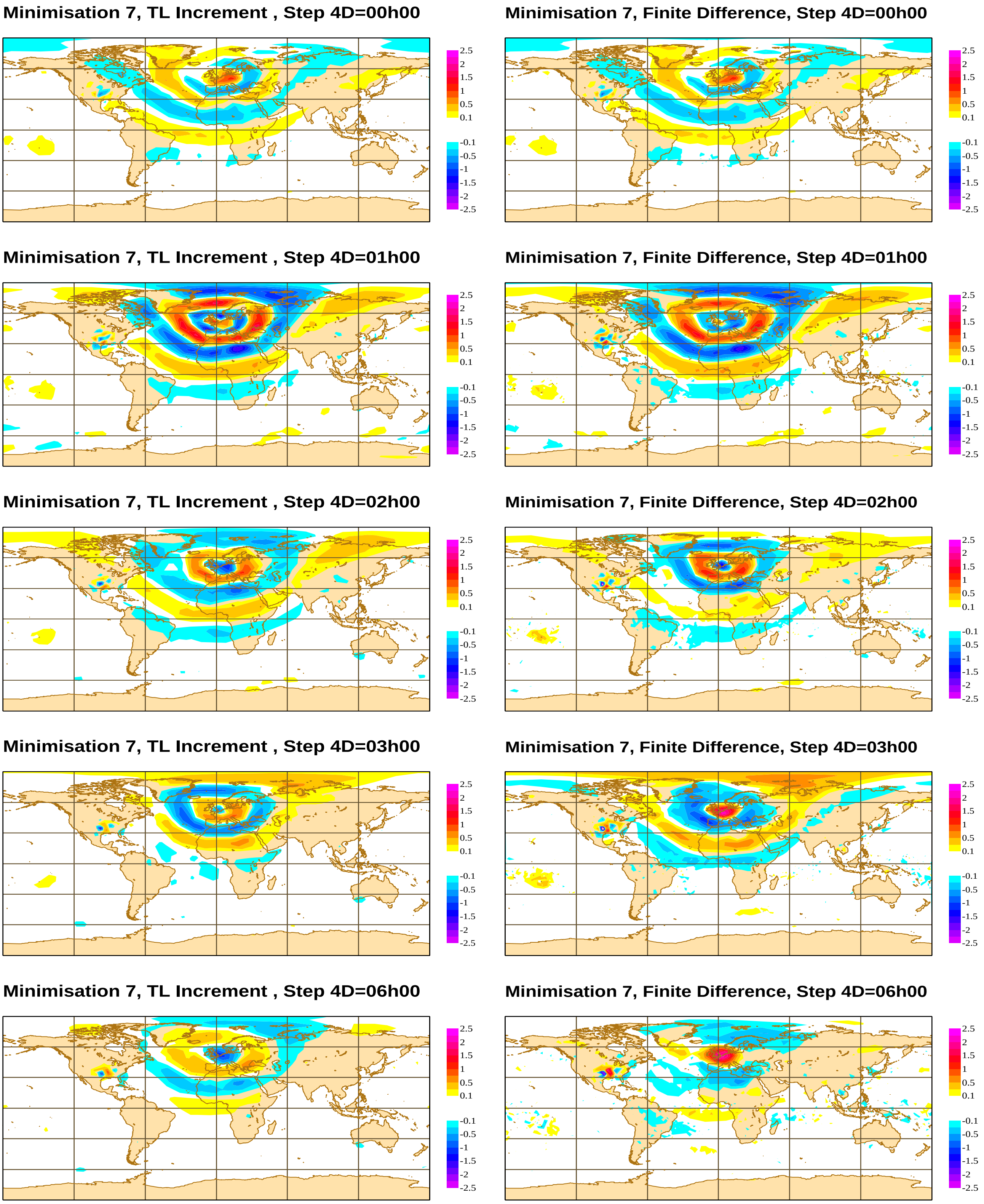
4 Study of poor convergence

4.1 Positive feedback

We have shown in section 3.2 that a 4D-Var assimilation experiment run using T255 outer loops with 30 minutes time steps and T95 inner loops with 1 hour time steps starts diverging after 4 outer loop iterations. The largest increments in the latest minimisations are in the surface pressure component of the control variable. Figure 17 shows the surface pressure increment from minimisation 7 and its evolution, in the inner loop with the tangent linear model in the left column and, in the outer loop with the nonlinear model in the right column. The rows show the increment at initial time and after 1, 2, 3 and 6 hours. The general pattern of the surface pressure increment is circular centred over northern Europe. The linear and nonlinear evolutions slowly diverge from each other. This is a gravity wave which propagates with different phase speeds in the linear and nonlinear integrations. After 6 hours, they are close to opposite phases, the increment at the centre of the pattern is positive in one case, negative in the other.

This can explain the divergence of 4D-Var. The inner loop minimisation fits the data in the area at the centre of the pattern. The increment is added to the first guess and evolves in the nonlinear integration. Because of the discrepancy between inner and outer loops, this new trajectory moves away from the data, in the opposite direction than intended by the inner loop. In the ensuing minimisation, an increment will be generated which will try again 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.

Gravity wave phase speed being highly dependent on time step, this explains why we do not observe that



Minimisation 7, TL Increment , Step 4D=00h00

Minimisation 7, Finite Difference, Step 4D=00h00

Minimisation 7, TL Increment , Step 4D=01h00

Minimisation 7, Finite Difference, Step 4D=01h00

Minimisation 7, TL Increment , Step 4D=02h00

Minimisation 7, Finite Difference, Step 4D=02h00

Minimisation 7, TL Increment , Step 4D=03h00

Minimisation 7, Finite Difference, Step 4D=03h00

Minimisation 7, TL Increment , Step 4D=06h00

Minimisation 7, Finite Difference, Step 4D=06h00

Figure 17: Surface pressure increment for minimisation 7 propagated by the T95 tangent linear model (left) and the T255 nonlinear model (right).

phenomena when both models use the same time step, regardless of spatial resolution. This has been tested this with several time step values and shows that the tangent linear model is able to reproduce the nonlinear model behaviour with respect to phase speed accurately for a range of time-step values. This is not a linearisation issue in that, given the same time step, the linear and nonlinear models behave similarly.

4.2 Hessian and increments

The previous paragraph shows how the increments generated after the fourth minimisation amplifies, resulting in divergence of 4D-Var. Figure 18 shows the partial surface pressure increments for the 10 minimisations of the T255/T95 experiment. Several experiments, run for various times of day, various times of the year and with several tangent linear model configurations, with or without physics, have shown similar increments in later outer loop iterations, as did the *perfect solution* experiment in section 3.5. 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 this eigenvector is driven by the density and accuracy of observations. They have shown that, in a simplified example with n observations in the same location, an approximation of the condition number of the minimisation problem is given by:

$$\kappa \approx 2n\frac{\sigma_b^2}{\sigma_0^2} + 1$$

where σ_b and σ_o are the background and observation errors. In the current system, surface pressure observations over Europe dominate as indicated by the leading eigenvector of the Hessian of the cost function (figure 19, top panel).

Two additional experiments were run where observation error for surface pressure observations was increased by 50% and 100% (doubling of error). Figure 19 shows that the leading eigenvectors of the Hessian for these three experiments are almost identical. However, the eigenvalues associated with those eigenvectors are respectively 5474, 2519 and 1502, in rough agreement with the simplified expression above. Figures 20 and 21 show the 10 partial increments when surface pressure observations error is increased by 50% and 100%. The amplitude of the spurious pattern is reduced in the first case, and disappears totally in the second case. Figure 22 shows that 4D-Var convergence is improved.

These experiments show that the eigen-structure of the Hessian of the cost function is a useful means to monitor the behaviour of incremental 4D-Var, in particular the largest eigenvalue. The leading eigenvectors indicate the directions where observations are going to be fitted the closest, regardless of the background. This means that an increment in those directions will inevitably be introduced by the minimisation. This effect is more pronounced as the associated eigenvalue becomes large. It seems particularly unfortunate that the pattern of the eigenvector associated to a large eigenvalue (determined by the data) is so sensitive and subject to a positive feedback effect (determined by 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 rate of convergence of 4D-Var.

4.3 Inner Loops Accuracy

Laroche and Gauthier (1998) have shown that the accuracy with which the inner loop minimisation is resolved has an impact on the overall convergence of incremental 4D-Var. More precisely, their results show that the best results are obtained when inner loop iterations are stopped early enough to avoid over-fitting observational noise, as this would introduce spurious gradients in the following trajectory integration. A set of experiments

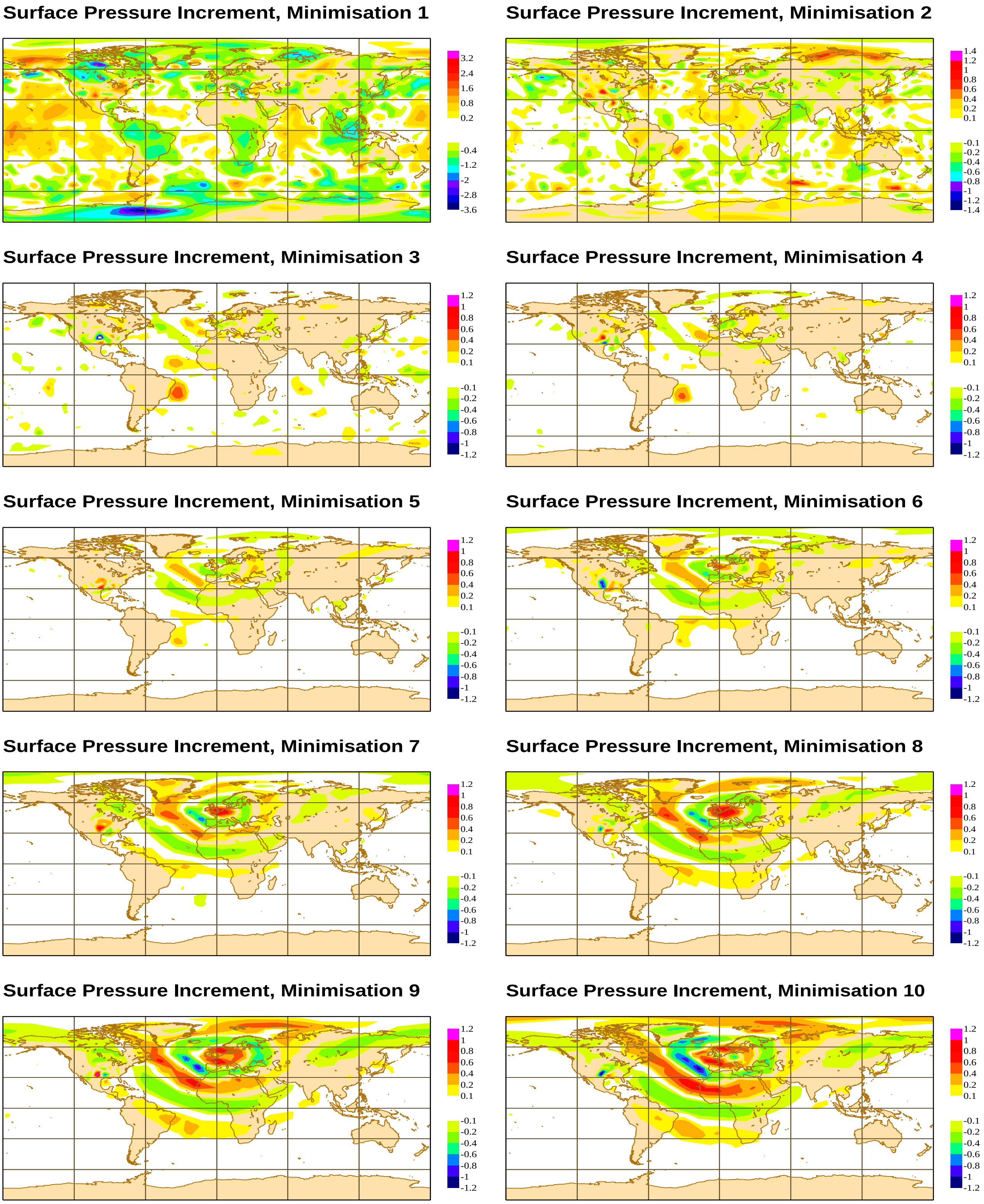

Surface Pressure Increment, Minimisation 1
3.2
2.4
1.6
0.8
0.2
-0.4
-1.2
-2
-2.8
-3.6
Surface Pressure Increment, Minimisation 2
1.4
1.2
1
0.8
0.6
0.4
0.2
0.1
-0.1
-0.2
-0.4
-0.6
-0.8
-1
-1.2
-1.4
Surface Pressure Increment, Minimisation 3
Surface Pressure Increment, Minimisation 4
Surface Pressure Increment, Minimisation 5
Surface Pressure Increment, Minimisation 6
Surface Pressure Increment, Minimisation 7
Surface Pressure Increment, Minimisation 8
Surface Pressure Increment, Minimisation 9
Surface Pressure Increment, Minimisation 10
1.2
1
0.8
0.6
0.4
0.2
0.1
-0.1
-0.2
-0.4
-0.6
-0.8
-1
-1.2

Figure 18: Partial surface pressure increments generated by the successive minimisations of the T255/T95 experiment, running from top left to bottom right. Note the the contour interval is different for the first two increments which have larger amplitude.

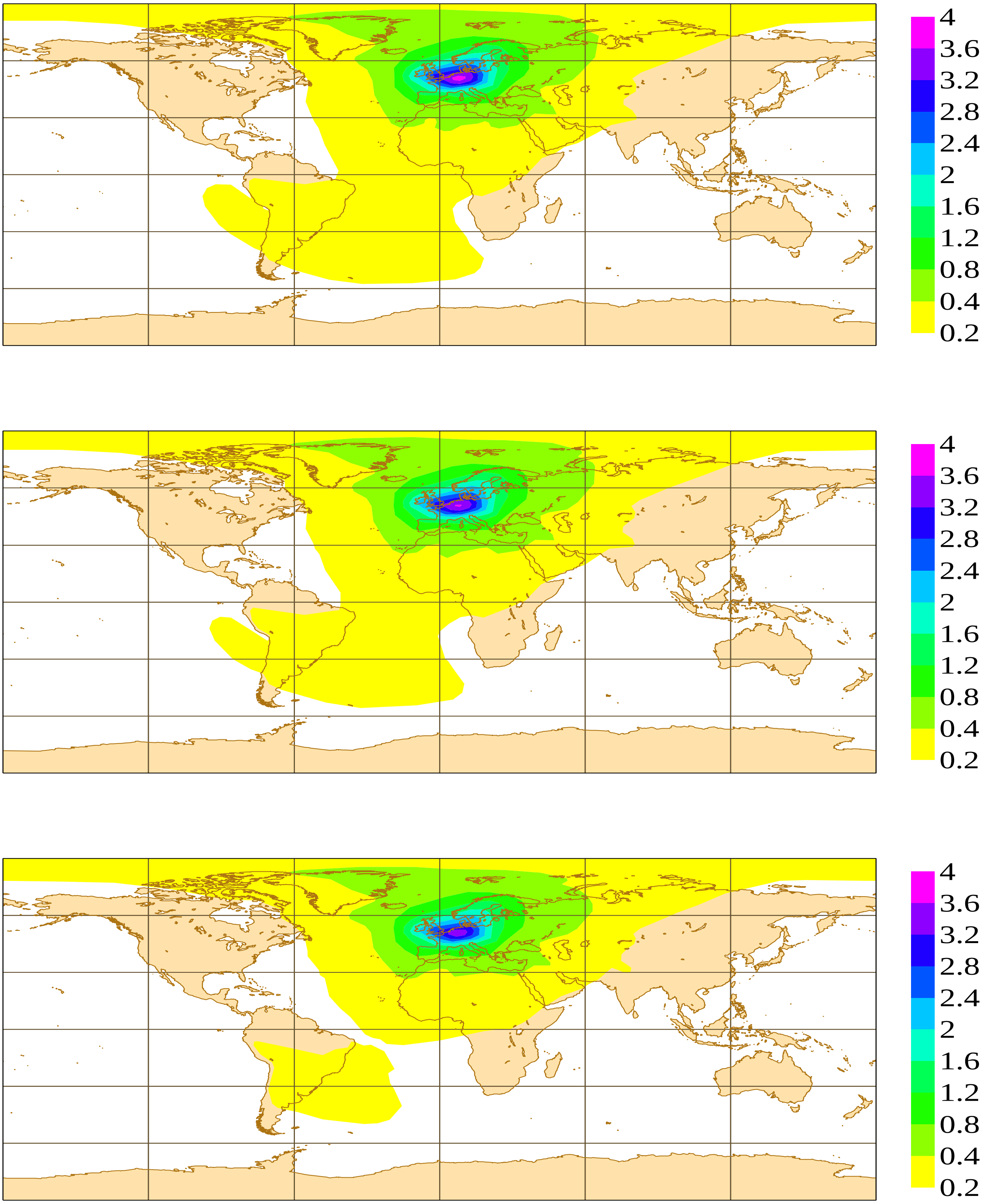
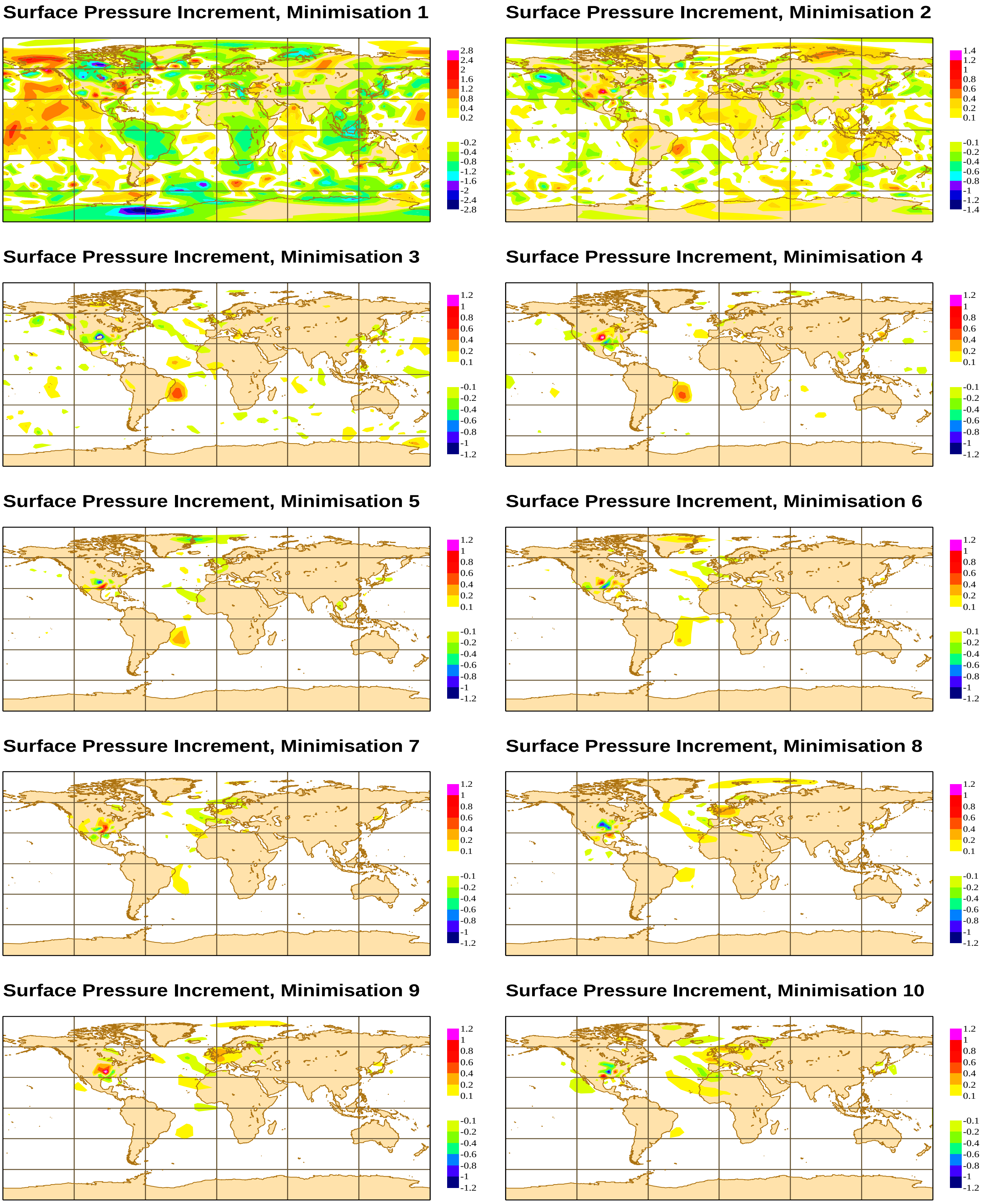


Figure 19: Surface pressure component of the leading eigenvector of the Hessian of the cost function for the T255/T95 reference experiment (top), with surface pressure observation error increased by 50% (middle) and by 100% (bottom).



Surface Pressure Increment, Minimisation 1

Surface Pressure Increment, Minimisation 2

Surface Pressure Increment, Minimisation 3

Surface Pressure Increment, Minimisation 4

Surface Pressure Increment, Minimisation 5

Surface Pressure Increment, Minimisation 6

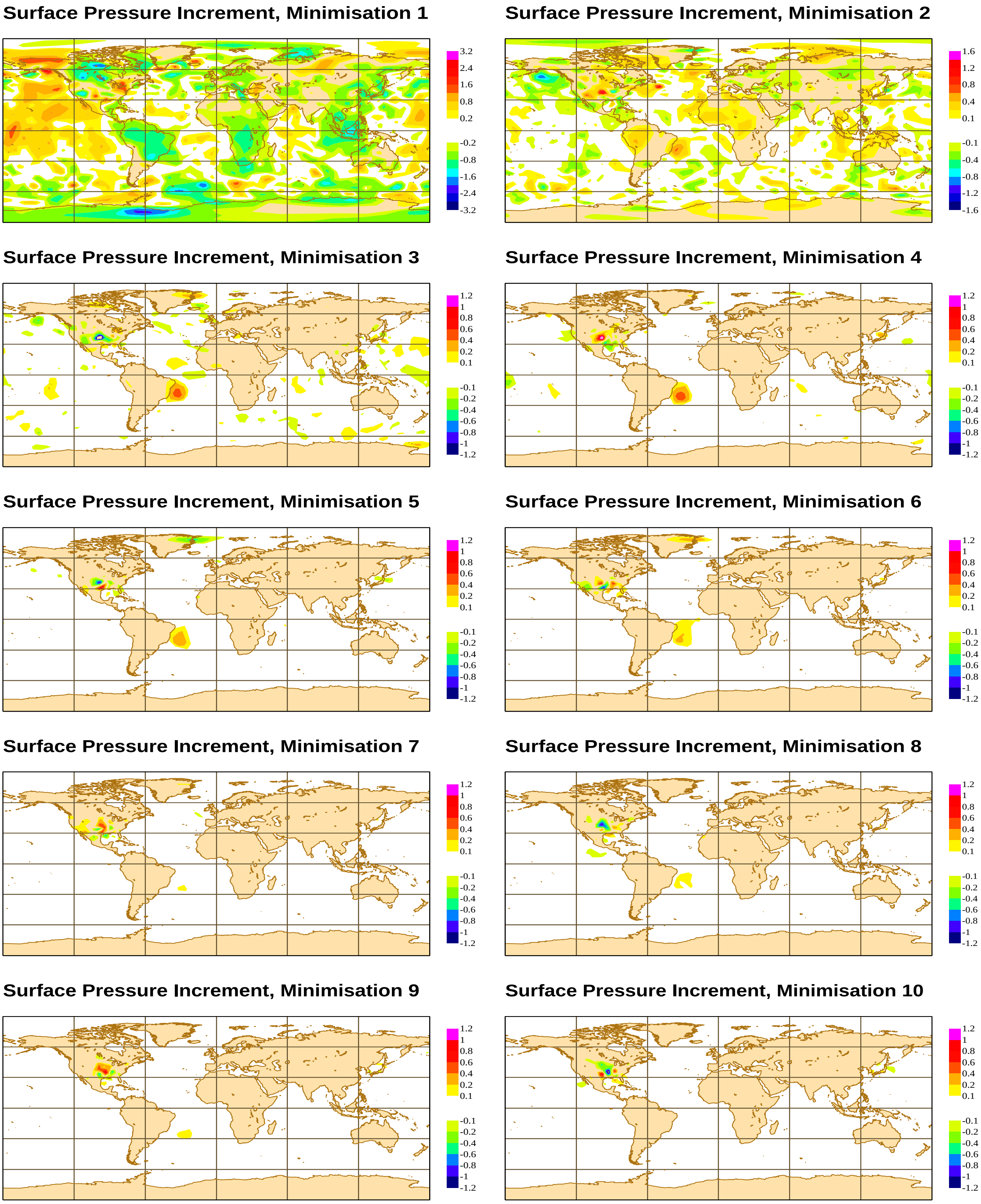
Surface Pressure Increment, Minimisation 7

Surface Pressure Increment, Minimisation 8

Surface Pressure Increment, Minimisation 9

Surface Pressure Increment, Minimisation 10

Figure 20: Partial surface pressure increments for T255/T95 experiment with surface pressure observation error inflated by 50%.



Surface Pressure Increment, Minimisation 1

Surface Pressure Increment, Minimisation 2

Surface Pressure Increment, Minimisation 3

Surface Pressure Increment, Minimisation 4

Surface Pressure Increment, Minimisation 5

Surface Pressure Increment, Minimisation 6

Surface Pressure Increment, Minimisation 7

Surface Pressure Increment, Minimisation 8

Surface Pressure Increment, Minimisation 9

Surface Pressure Increment, Minimisation 10

Figure 21: Partial surface pressure increments for T255/T95 experiment with doubled surface pressure observation error.

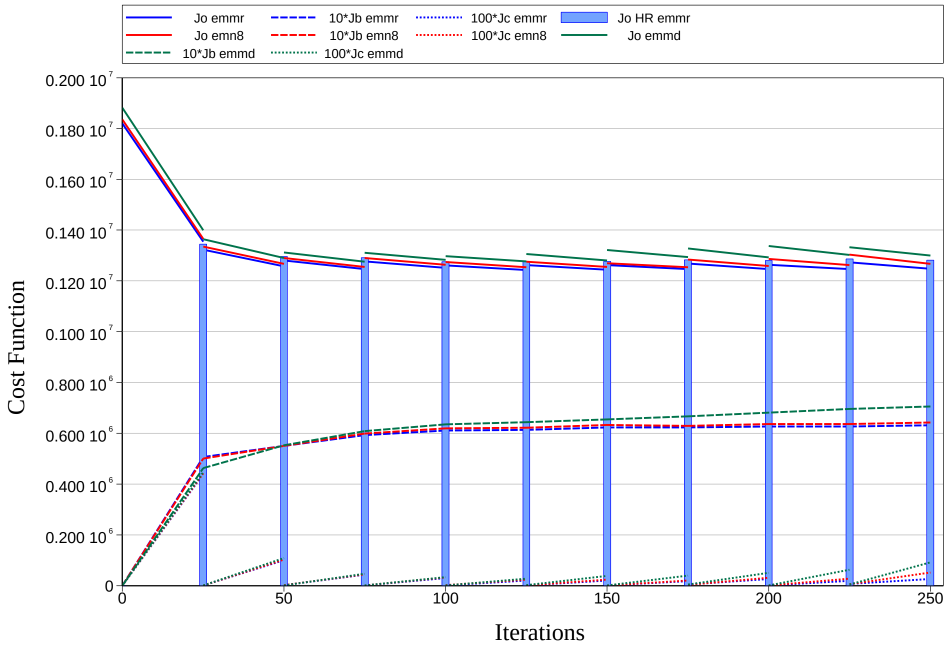


Figure 22: Evolution of 4D-Var cost function with when surface pressure observation error is increased by 50% (in red) and doubled (in blue) with respect to the control experiment (in green).

was performed where inner loop iterations were stopped when a given accuracy was reached. The stopping criterion used in these experiment is the reduction in the norm of the gradient of the cost function. The evolution of J_o in the outer loops is shown on figure 23 for stopping criterion values of 0.1, 0.25 and 0.4. It shows that solving the inner loop problems more accurately leads to the divergence of the algorithm at outer loop level. As explained above, this is not unexpected. The other values of the stopping criterion lead to similar J_o values, with a smaller J_b for the less accurate inner loop solution (figure 24). The horizontal axis shows the cumulated number of iterations in the inner loop but does not take into account the cost of the nonlinear trajectories. In practice, a stopping criteria value of 0.25 would give slightly faster convergence with few outer loops.

4.4 Quasi-Newton minimisation

In addition to the conjugate gradient-Lanczos algorithm (CONGRAD), the quasi-Newton minimisation algorithm (m1qn3) developed at INRIA by Gilbert and Lemaréchal (1989) is still available in the IFS. Figure 25 shows that which minimisation algorithm is used in the inner loop seems to have an impact on the overall convergence. The value of J_o in the outer loop converges with m1qn3 which was not the case with the conjugate gradient. The numerical values shows that J_o in the outer loops is in fact strictly decreasing, although it does not decrease as fast in the first iterations. Also notice that the value of J_b remains lower throughout with m1qn3 as is the case with the value of J_c after the third minimisation. The figure also shows results from another experiment where the first 6 minimisations were run using CONGRAD followed by 4 minimisations using m1qn3 (in green). Again, m1qn3 shows an improvement over CONGRAD for these 4 minimisations.

The two algorithms should give the same solution for the minimisation 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 are less sensitive to the eigenstructure of the cost function and not as aligned with the leading eigenvector even though they still have a

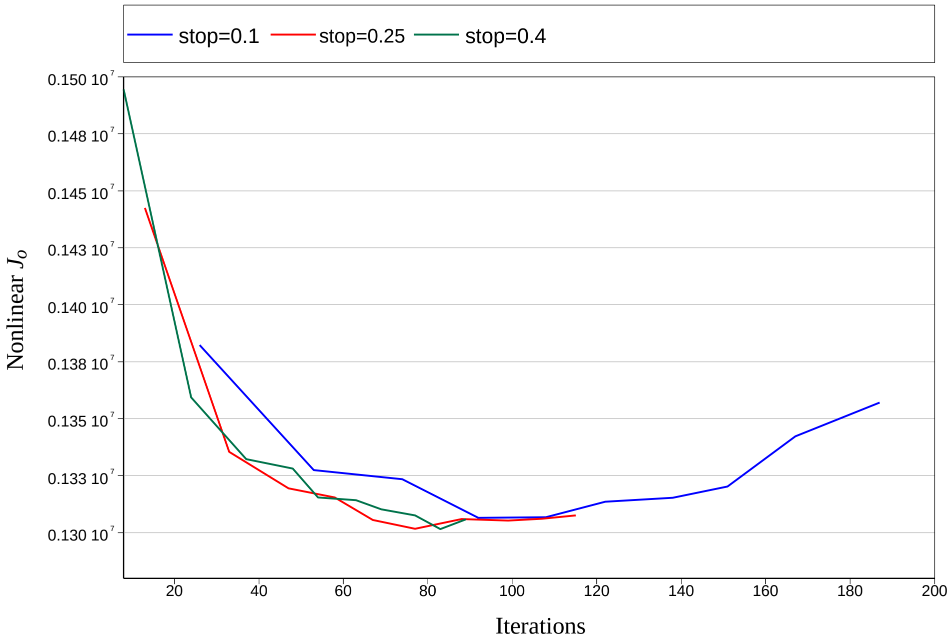


Figure 23: Evolution of J_o in outer loops for various stopping criteria in inner loop iterations.

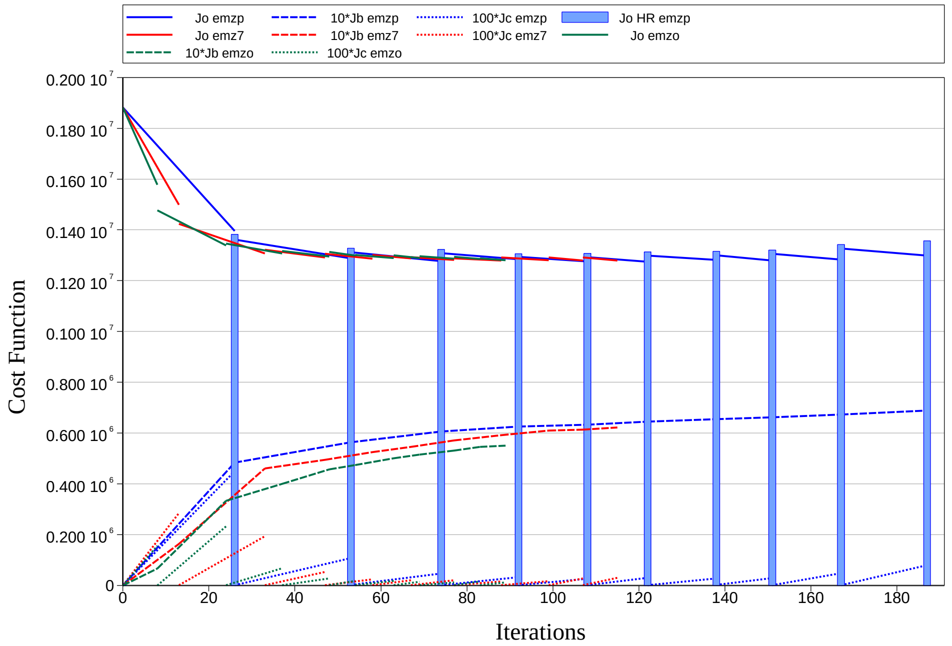


Figure 24: Evolution of 4D-Var cost function with stopping criteria of 0.1 (in blue), 0.25 (in red) and 0.4 (in green) for inner loop iterations.

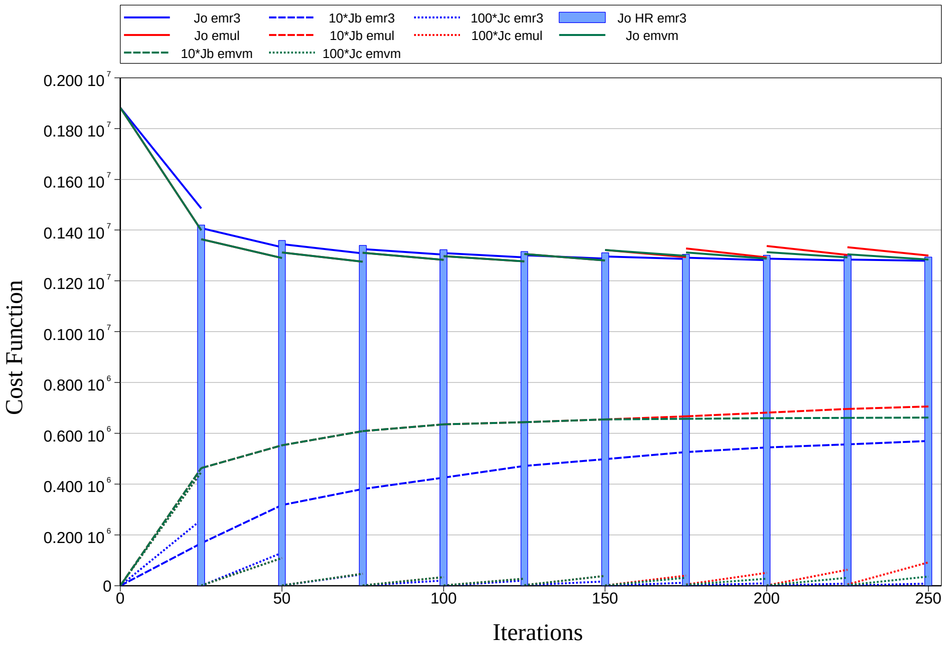


Figure 25: Evolution of 4D-Var cost function with `m1qn3` *(blue),* `CONGRAD` *(red) and 6 minimisations with* `CONGRAD` *followed by 4 minimisations with* `m1qn3` *(green).*

component along that direction as shown on figure 26.

The evolution of the gradient also shows a different behaviour (figure 27). The gradient jump between the end of a minimisation and the beginning of the next one has reduced with `m1qn3` with respect to `CONGRAD`. The sequence of gradient norms at the beginning of the successive minimisations is decreasing which is a good sign that each outer loop is improving the solution. If the increment produced by a given minimisation improves the solution of the overall nonlinear problem, the gradient at the new starting point should be better than the one at the previous starting point was, which is the case with the quasi-Newton algorithm.

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. 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 inner loop 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 0.25 for both algorithms. Figure 28 shows that, of both experiments, the lowest value of J_o at high resolution is obtained with `m1qn3` in the last nonlinear trajectory. It is worth noting that it is also obtained with a lower value of J_b ($J_b = 55620$) than the value of J_b when `CONGRAD` reaches its minimum ($J_b = 57055$), 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. This would mean that it is a better solution. The results presented here do not allow us to conclude on the quality of the ensuing forecast as running more cases would be necessary. However, because of the number of iterations required, this is not computationally affordable.

Preconditioning `m1qn3` should give the best of both worlds with fast convergence and better solution. It is possible to run the first minimisation with `CONGRAD` to compute the preconditioner and apply it in the remaining minimisations to precondition `m1qn3`. Figure 29 shows that in this case, the first minimisations

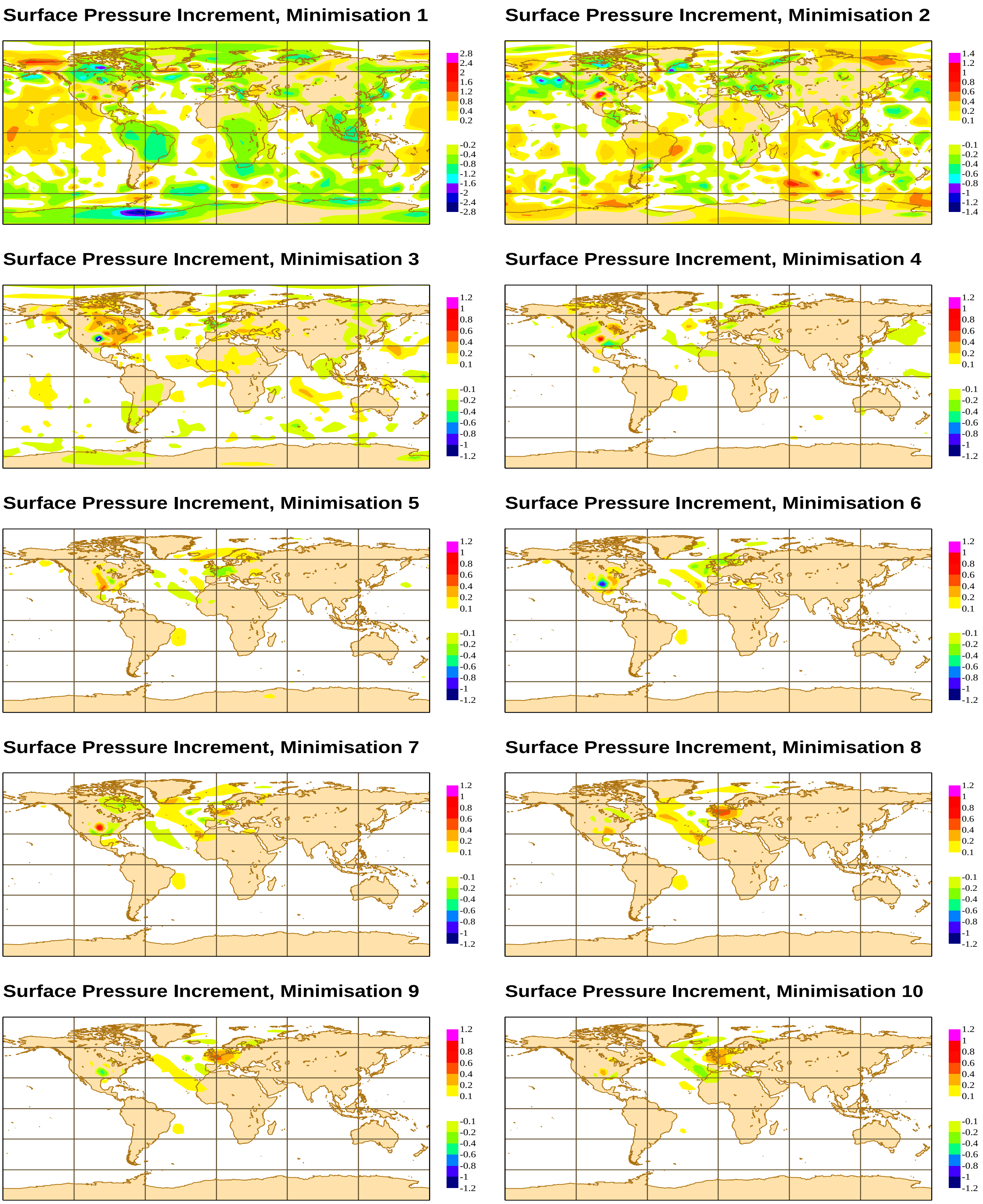

Surface Pressure Increment, Minimisation 1
2.8
2.4
2
1.6
1.2
0.8
0.4
0.2
-0.2
-0.4
-0.8
-1.2
-1.6
-2
-2.4
-2.8
Surface Pressure Increment, Minimisation 2
1.4
1.2
1
0.8
0.6
0.4
0.2
0.1
-0.1
-0.2
-0.4
-0.6
-0.8
-1
-1.2
-1.4
Surface Pressure Increment, Minimisation 3
Surface Pressure Increment, Minimisation 4
Surface Pressure Increment, Minimisation 5
Surface Pressure Increment, Minimisation 6
Surface Pressure Increment, Minimisation 7
Surface Pressure Increment, Minimisation 8
Surface Pressure Increment, Minimisation 9
Surface Pressure Increment, Minimisation 10
1.2
1
0.8
0.6
0.4
0.2
0.1
-0.1
-0.2
-0.4
-0.6
-0.8
-1
-1.2

Figure 26: Partial surface pressure increments for T255/T95 experiment with `m1qn3`.

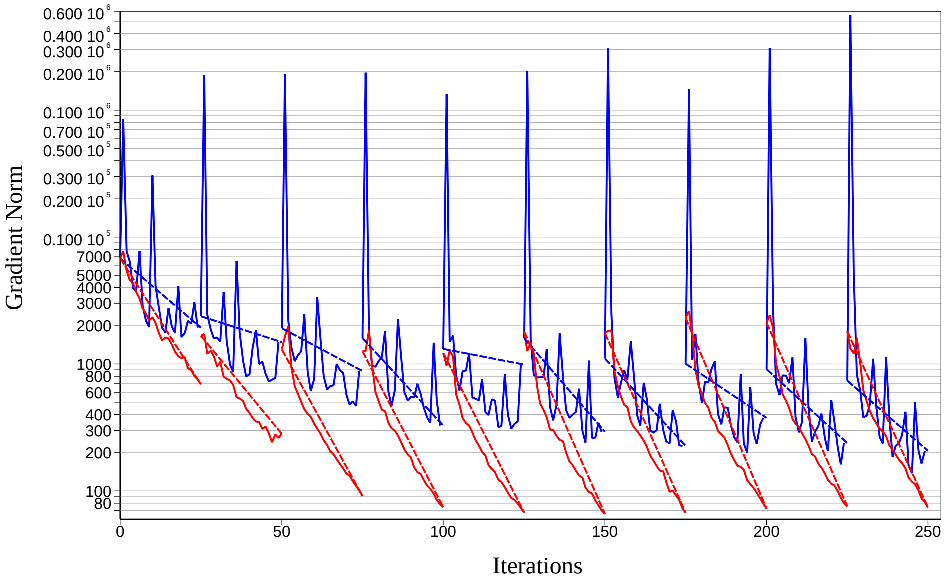


Figure 27: Evolution of the gradient norm with `m1qn3` *(blue) and* `CONGRAD` *(red).*

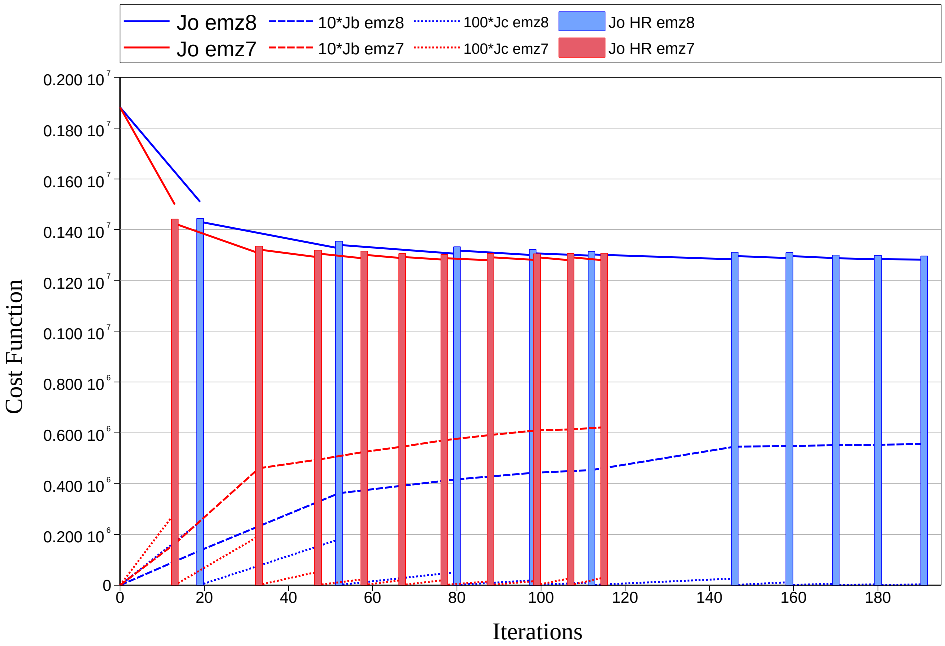


Figure 28: Evolution of 4D-Var cost function with `m1qn3` *(blue) and* `CONGRAD` *(red), both with stopping criterion set to 0.25.*

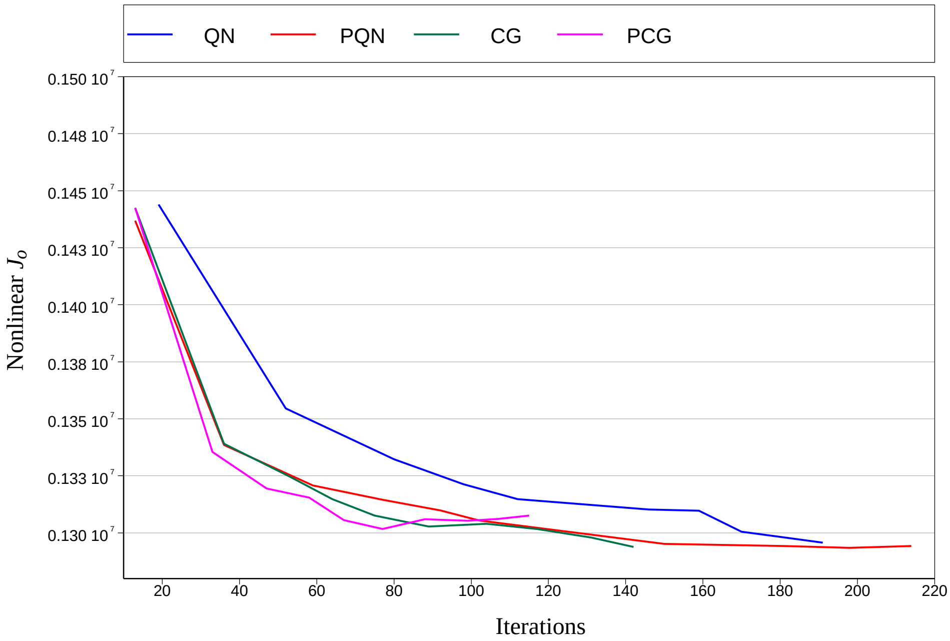


Figure 29: Evolution of J_o in outer loops for `m1qn3` *and* `CONGRAD`*, with and without preconditioning.*

converge faster but the last ones are slower to converge. It also changes the solution at it gives lower values of J_o but higher values of J_b. The same figure also shows the preconditioned and un-preconditioned conjugate gradient. As seen previously, the preconditioned conjugate gradient seems to diverge after a few iterations, this is not the case for the un-preconditioned algorithm. The accuracy with which the linearised problem is solved is not the only factor since all algorithms were stopped at the same accuracy. The preconditioner could lead to outer loop divergence.

Both the conjugate gradient and the conjugate gradient preconditioning are sensitive to the eigen-structure of the Hessian of the cost function. Using one, or both, tend to generate an increment that has a component in that direction. It then amplifies through the gravity wave positive feedback effect described in section 4.1.

5 Operational setup

5.1 Multi-resolution incremental 4D-Var

In order to reduce the computational cost of incremental 4D-Var, the resolution of the first minimisation is reduced in operations. It is possible to apply this idea with more than two outer loops and increase progressively the resolution in successive minimisations as experimented by Veersé and Thépaut (1998). Experiments have shown that Hessian eigenvectors are mostly large scale and can be computed at low resolution to form an effective preconditioner for higher resolution inner loops minimisations that follow (M. Fisher, personal communication) which makes this setup very attractive. Figure 30 shows the evolution of the cost function when the first three minimisations are run at T42, T95 and T159 respectively, the remaining minimisations being run at T255. The horizontal axis is the number of equivalent T255 iterations based on the computational cost of an iteration at each resolution as shown in the table below.

Resolution	Time per iteration (sec)
T42	11
T95	16
T159	57
T255	135

The figure shows that this setup does give the same final values of J_o and J_b as 4D-Var with T255 minimisations throughout. The increments produced in the last four minimisation in both setups are surprisingly similar (not shown). The multi-resolution incremental setup does allow to reach that solution significantly faster, thus saving computer time.

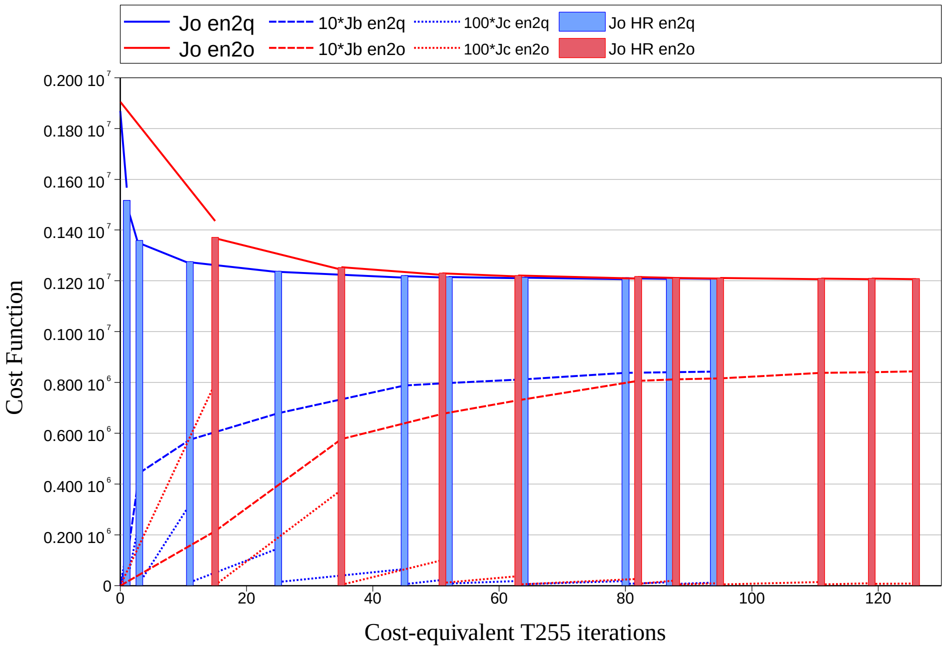


Figure 30: Evolution of T42/T95/T159/T255 multi-resolution incremental 4D-Var cost function (in blue) and T255 incremental 4D-Var (in red).

5.2 Operational setup

Because of the significant computational cost of the high resolution nonlinear trajectories, only a few outer loop iterations are affordable for operational data assimilation. Currently at ECMWF, 4D-Var is run with two outer loop iterations. The first minimisation is run at T95 for 70 iterations, the second one at T159 for, on average, 35 iterations. The number of iterations in the first minimisation is fixed and determined by the forecast error computation (Fisher and Courtier (1995)). From the results presented in this paper, it seems that running with reduced accuracy in the first minimisations and more nonlinear updates should be beneficial.

Figure 31 show the evolution of the components of the cost function for the current operational setup with two outer loop iterations, the first minimisation is run at T95 with a fixed number of iterations, the second minimisation at T159 and is stopped when the gradient norm has reduced by a factor of 0.05. It also shows another experiment with three outer loops where the first minimisation is unchanged and the following two minimisations are run at T159, with a stopping criterion of 0.2. It shows that, with the same total number of

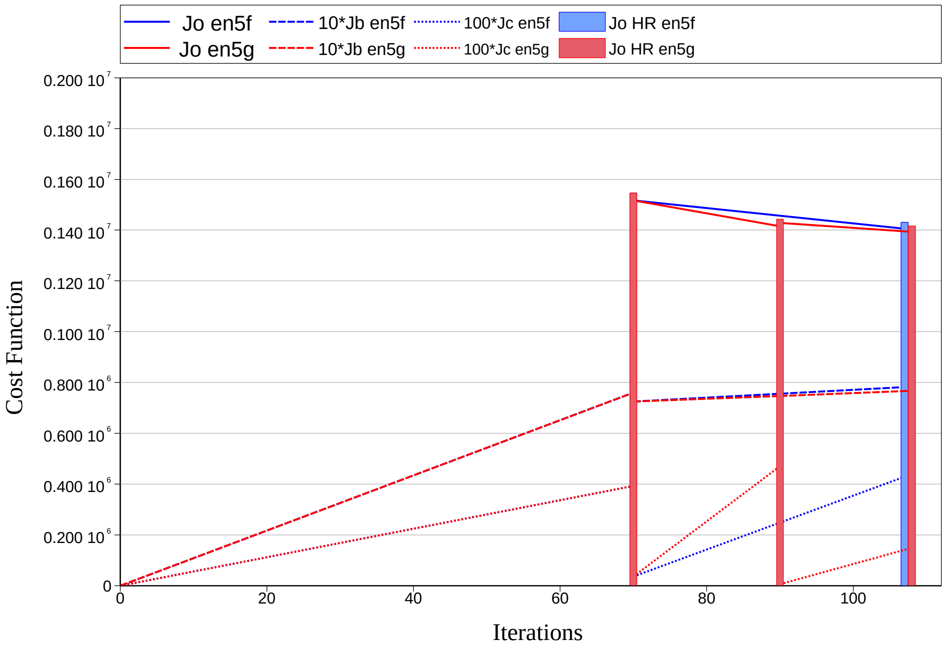


Figure 31: Evolution of 4D-Var cost function with the current operational setup (in blue) and with three outer loop iterations (in red).

iterations, slightly lower values of both J_o and J_b are obtained[4].

In terms of computational cost, the first minimisation at T95 is equivalent to 20 iterations at T159. The cost of each nonlinear trajectory is equivalent to approximately 8 T159 inner loop iterations. The total cost of 4D-Var is thus approximately 80 T159-equivalent iterations. Reducing the number of iterations in the first minimisation could compensate for the cost of another nonlinear trajectory. The additional minimisation would have to be a split of the current second one as tested above. Forecast error computation would then have to be done in the second or third minimisation, accumulating the eigenvectors computed in all the minimisations. That leaves very little room to manœuvre, at least at constant cost.

6 Conclusions

The convergence of incremental 4D-Var, as implemented at the moment in the IFS, with more outer loops is limited by a combination of two factors: as more iterations are run, the partial increments have a component along the direction of the leading eigenvector of the Hessian of the cost function. Currently, the shape of the cost function is such that this increment creates a gravity wave that propagates with a different phase speed in the inner and outer loops, leading to a positive feedback effect. Ultimately, the combination of these two factors leads to the divergence of the algorithm. 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 of 4D-Var.

The preconditioned Lanczos-conjugate gradient algorithm is very efficient and allows for fast convergence of the inner loop minimisations. However, it might over-emphasise the direction of the leading eigenvectors of the Hessian of the cost function. Unfortunately, in the current 4D-Var, this direction is unstable. The quasi-Newton

[4] An experiment is starting to evaluate the impact on the forecast

minimisation algorithm seems less sensitive to the leading eigenvectors, and thus to the feedback effect, but convergence of the inner loop minimisation is slower.

We have also shown that incremental 4D-Var converges better when inner loop minimisation are solved with relatively low accuracy. This affects the rate of convergence as well as the final solution as we were able to find a solution with lower values of both J_o and J_b. That indicates that the solution has a better fit to observations and background at the same time. Over-solving the inner loop minimisation can trap the incremental algorithm in a local minimum in the presence of significant nonlinearities.

In the near future, it should be possible to run incremental 4D-Var with at least one more outer loop iteration. This could potentially lead to an improved analysis. In the longer term, more nonlinear trajectory updates might become even more critical as observations that depend on more nonlinear phenomena and more nonlinear observation operators are assimilated. With that prospect in view, it is important to understand the properties of convergence of the incremental algorithm.

Experiments have shown that increasing the weight given to the weak constraint digital filter in the J_c term of the cost function, in particular for the surface pressure component of the control variable, would reduce the gravity wave noise. More experimentation would be required to determine the appropriate weights.

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 data assimilation problem would likely reduce or eliminate the incremental 4D-Var convergence problems by reducing the global weight given to these observations. Further research work is required in that direction.

Acknowledgements

The author would like to thank Mike Fisher and Erik Andersson for their comments on earlier versions of this manuscript as well as Sami Saarinen for his advice on modifying with the ODB to make it possible to run with more outer loop iterations.

References

E. Andersson, P. Bauer, A. Beljaars, F. Chevallier, E. Hólm, M. Janisková, P. Kållberg, G. Kelly, P. Lopez, A. Mcnally, E. Moreau, A. Simmons, J.-N. Thépaut, and A. Tompkins. Assimilation and Modeling of the Atmospheric Hydrological Cycle in the ECMWF Forecasting System. *Bull. Amer. Meteorol. Soc.*, 86:387–402, 2005.

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

E. Andersson, M. Fisher, R. Munro, and A. McNally. 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, 2000.

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.

M. Fisher. Minimization algorithms for variational data assimilation. In *Seminar on Recent developments in numerical methods for atmospheric modelling*, ECMWF, pages 364–385, September 1998.

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

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

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

B. Hoskins and A. Simmons. A multi-layer spectral model and the semi-implicit method. *Q. J. R. Meteorol. Soc.*, 101:637–655, 1975.

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.

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

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

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.

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.

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

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

A Preconditioning

The conjugate gradient preconditioning is explained in details in section 6.6 of the IFS documentation where equation (6.1) defines the preconditioner as:

$$M = I + \sum_{i=1}^{k} (\mu_i - 1) w_i w_i^T$$

where $\{\mu_i, w_i\}$ are estimates of the leading k eigenvalues and eigenvectors of the Hessian of the cost function produced by the conjugate gradient-Lanczos algorithm. In theory, the leading eigenvalues of the preconditioned Hessian become λ_i/μ_i where the λ_i are the eigenvalues of the full Hessian. If $\lambda_i/\mu_i < \lambda_{k+1}$, the condition number becomes λ_{k+1}. If $\mu_i = \lambda_i$, the leading eigen-pairs are effectively removed from the problem. In practice, this is not exact and for numerical reasons (conditioning of the preconditioner itself), large values of μ_i are reduced to a fixed value `R_MAX_CNUM_PC`, currently set to 10. In the current setup, the first 25 eigenvectors are used in the preconditioner and because the first 25 eigenvalues are larger than 10, $\mu_i = 10$ for all of them. However, the spectrum of the Hessian in 4D-Var is such that the leading eigenvalue divided by 10 is still larger than the 26th eigenvalue. This means that the leading eigenvector of the full Hessian is still the leading eigenvector of the preconditioned problem, only with a reduced associated eigenvalue. Setting `R_MAX_CNUM_PC=100` eliminates this problem. Figure 32 shows that the increase of the gradient norm at the beginning of each minimisation has been eliminated. The convergence of 4D-Var is very slightly improved but the overall impact is extremely limited. This modification could improve the conditioning of the second minimisation in the current operational 4D-Var and save one or two iterations.

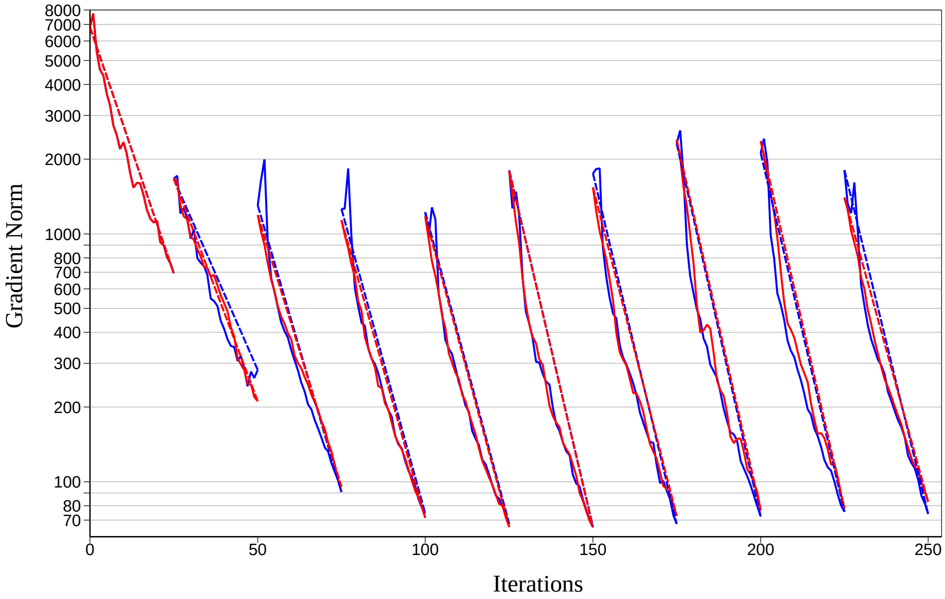


Figure 32: Evolution of the gradient of the cost function for `R_MAX_CNUM_PC=10` *(in blue) and* `R_MAX_CNUM_PC=100` *(in red).*

B Gravity wave speed in semi-implicit scheme

As pointed out by A. Hollingsworth, Hoskins and Simmons (1975) have shown that the speed of gravity waves in the semi-implicit scheme can be controlled by adjusting Δt in equation (17) of their paper. This has been tested in a 4D-Var experiment. Using the inner loop time step value for Δt is not possible, the forecast explodes in the first high resolution trajectory (before adding any increment). The value used here was the outer loop time step value (1/2 hour). Figure 34 shows that some gravity wave pattern is still present in the increments, although it is not as dominant. However, figure 33 shows that the convergence of the minimisation has worsened.

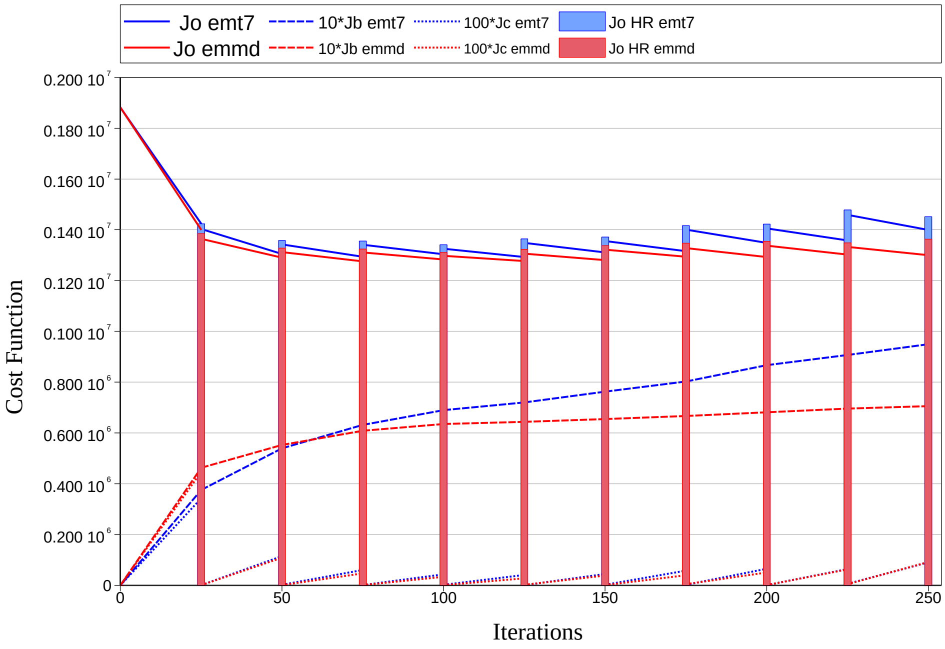
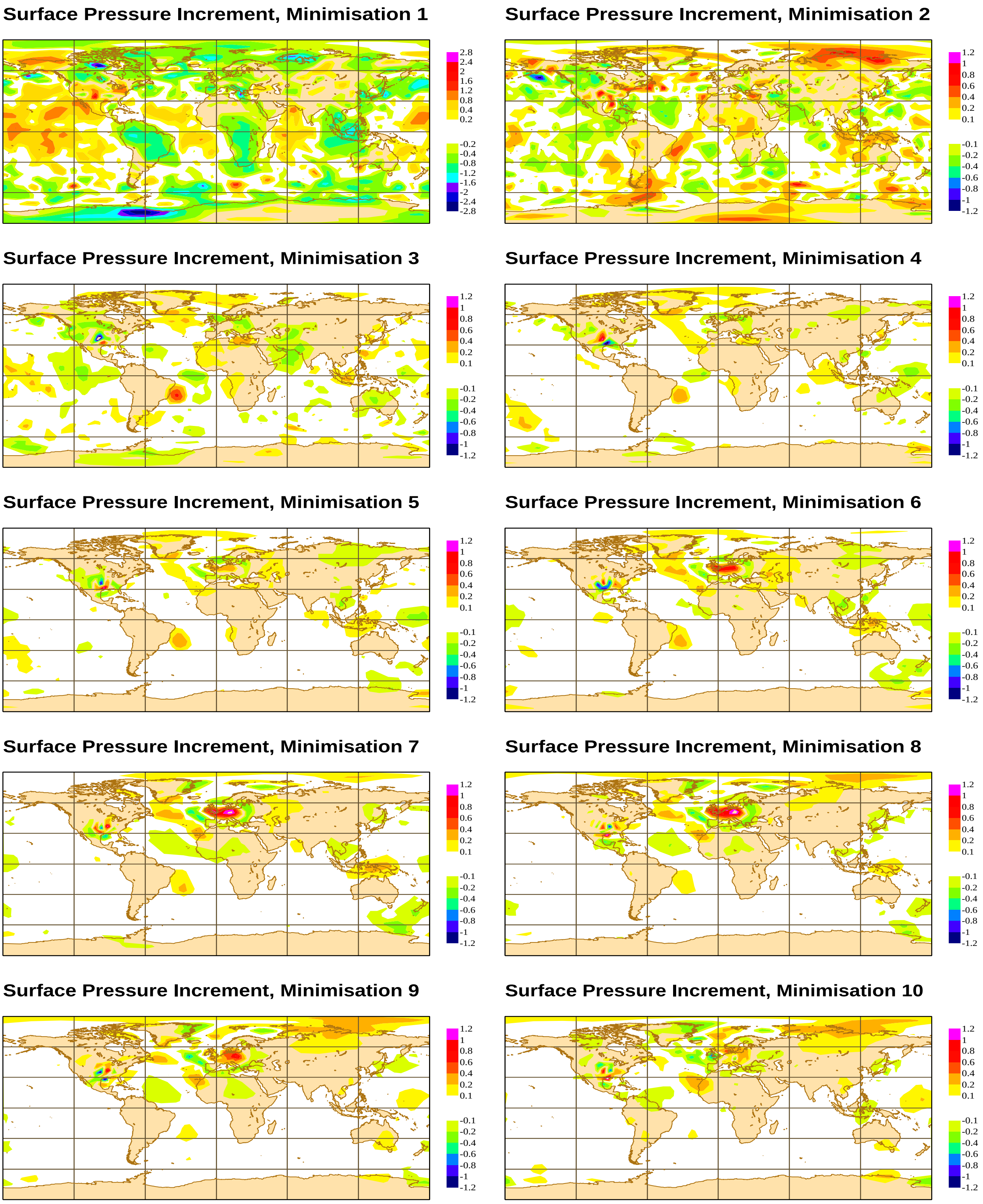


Figure 33: Evolution of 4D-Var cost function with $\Delta t = 1800$ *in the semi-implicit scheme (in blue), the reference experiment is shown in red.*

Modifying Δt in the semi-implicit scheme has some effect on the speed of gravity waves, it also changes the solution of the equation and might affect the overall consistency of the forecast. The nature of the discrepancy between inner and outer loops is different but it is still large enough to prevent convergence.


Surface Pressure Increment, Minimisation 1

Surface Pressure Increment, Minimisation 2

Surface Pressure Increment, Minimisation 3

Surface Pressure Increment, Minimisation 4

Surface Pressure Increment, Minimisation 5

Surface Pressure Increment, Minimisation 6

Surface Pressure Increment, Minimisation 7

Surface Pressure Increment, Minimisation 8

Surface Pressure Increment, Minimisation 9

Surface Pressure Increment, Minimisation 10

Figure 34: Partial surface pressure increments for T255/T95 experiment with $\Delta t = 1800$ *seconds in the semi-implicit scheme.*

C Digital filter initialisation

The weak constraint digital filter initialisation can be used to control the gravity waves generated by incremental 4D-Var. The weight α given to the J_c term in the cost function can be increased. In the current setup, the digital filter is only applied to the divergence part of the control variable, it is also possible to apply it to the surface pressure component of the control variable. Figure 35 shows that both methods can improve the convergence of incremental 4D-Var. More tests would be needed to determine the most appropriate settings for operational use with more outer loop iterations.

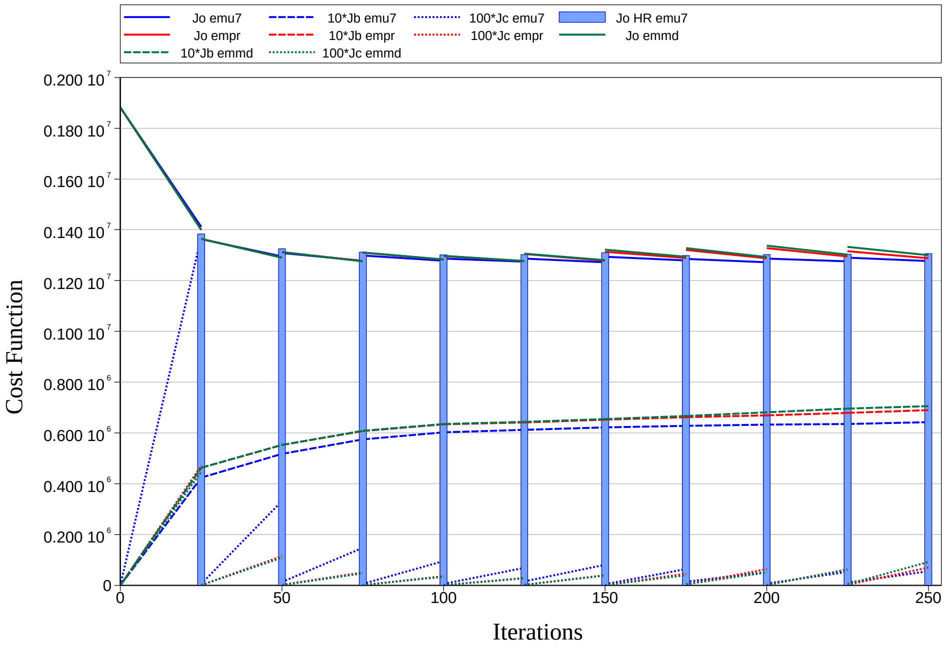


Figure 35: Evolution of 4D-Var cost function when J_c is applied to the surface pressure component of the control variable (in red) and for J_c applied to surface pressure and with increased weight (in blue). The control experiment is shown in green.