TECHNICAL MEMORANDUM

479

Will the 4D-Var approach be defeated by nonlinearity?

E. Andersson, M. Fisher, E. Hólm, L. Isaksen, G. Radnóti, and Y. Trémolet

Research Department

September 2005

This paper has not been published and should be regarded as an Internal Report from ECMWF.
Permission to quote from it should be obtained from the ECMWF.

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

Series: ECMWF Technical Memoranda

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

Contact: library@ecmwf.int

© Copyright 2005

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

Literary and scientific copyrights belong to ECMWF and are reserved in all countries. This publication is not to be reprinted or translated in whole or in part without the written permission of the Director. Appropriate non-commercial use will normally be granted under the condition that reference is made to ECMWF.

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.

Summary

Some of the issues involved in extending the current 4D-Var algorithm to higher resolution and/or longer assimilation window in the presence of nonlinearity, are investigated and discussed. Tests have been conducted within the context of the global numerical weather prediction (NWP) system at ECMWF. The accuracy of the 4D-Var solution algorithm and its convergence are investigated with respect to nonlinearities and their dependence on analysis resolution and length of the assimilation window. From a literature review and some new results in an idealized setting (the Lorenz-1995 40 parameter model) we conclude that a long-window weak-constraint 4D-Var has exciting prospects, alleviating the severity of the linearity assumption compared to the strong-constraint formulation. In weak-constraint 4D-Var a sequence of model states are estimated (rather than just the initial state), with the consequence that the tangent linear assumption is relied upon only for the shorter time segment between successive state estimates, and not for propagation of information throughout the assimilation window. This property makes a weak-constraint 4D-Var an attractive prospect that should be pursued as a main avenue of research in the coming years, with the goal to lengthen the assimilation window. Conclusions are drawn with respect to the ECMWF research plans.

1. Introduction

The 4D-Var data assimilation system is an integral part of ECMWF's operational forecast suite. Its development has contributed to the improvements seen in analysis and forecast accuracy in recent years (Simmons and Hollingsworth 2002). Several key elements in the current four-year research plan and the (proposed) long-term strategy build on the existing 4D-Var system, and rely on its assumed continued development for their success. The planned increases in resolution of the analysis are ambitious and computationally expensive. It is thus important to further enhance the efficiency of the 4D-Var solution algorithm (the minimisation) and to maintain high accuracy. The main challenges arise from an expected increase in the nonlinearity of the estimation problem, requiring minimisation of an increasingly complex cost-function with potential for multiple minima. These challenges have been identified as carrying risk (SAC 2004), as relatively little is known about the behaviour of a large-scale 4D-Var system in the presence of significant nonlinearities. In this paper we summarize and discuss a selection of results from two accompanying papers: *Diagnostics of linear and incremental approximations in 4D-Var, revisited for higher resolution*, by Radnóti et al. (2005) and *Incremental 4D-Var convergence study*, by Trémolet (2005), and draw conclusions with respect to the ECMWF research plans. We refer the reader to the two papers for their more comprehensive sets of results and complete descriptions, while at the same time endeavouring to keep the present paper self-contained. We investigate aspects of the current 4D-Var system from the viewpoint of nonlinear estimation, as this is considered a deciding factor for dealing with e.g. cloud and rain assimilation (Marécal and Mahfouf 2003; Moreau et al. 2004; Mahfouf et al. 2005), dealing with non-Gaussian error distributions (Lorenc 2003b) and higher-resolution analysis (Bouttier 2001; 2003).

The accuracy and efficiency of the currently adopted 4D-Var solution algorithm depend on the validity of the tangent-linear assumption (Bouttier 2001; Trémolet 2004). The requirements increase with analysis resolution and with the length of the assimilation window. A tangent-linear model of atmospheric dynamics and physics, linearized around the current best estimate of the atmospheric state, is used in 4D-Var to evolve the analysis increments over the assimilation window (currently 12 hours). Similarly, tangent-linear observation operators are used for the comparison between model and observations. Linearity is a prerequisite for the most efficient iterative minimisation algorithms, such as conjugate gradient (Fisher, 1998). Nonlinearities are accounted for by re-linearizing around the state incremented by the output of the linear iterations. This re-linearization can be repeated, thus obtaining the combined 'inner/outer' solution algorithm (Laroche and Gauthier, 1998; Lawless et al., 2005) consisting of linear inner loops nested within

nonlinear outer-loops; see Fisher (1998) and Andersson et al. (2004) for further information on the ECMWF implementation. The schematic in Figure 1, reproduced from Laroche and Gauthier (1998), shows a succession of quadratic inner-loop cost functions (dashed) approximating the non-quadratic cost function (full line) of the nonlinear problem. This approach will, however, break down (1) if the linearization errors in the inner loops are too large, or (2) if the outer-loop iterations diverge or are slow to converge. These two issues have most recently been investigated by Radnóti et al. (2005) and Trémolet (2005), respectively.

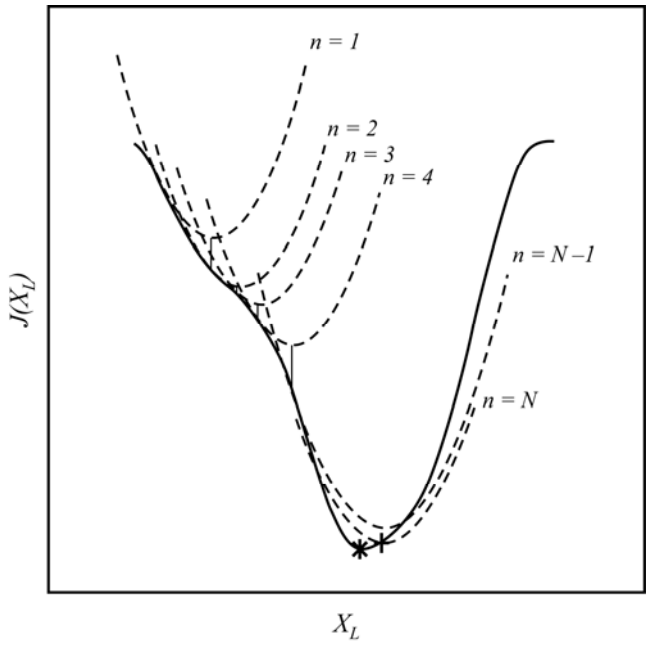

Figure 1. Schematic of the inner/outer solution algorithm, in which the full nonlinear problem, with a non-quadratic cost-function (full line) is solved through a succession of N linear problems, with quadratic cost-functions (dashed), approximating the original problem (from Laroche and Gauthier, 1998). The N re-linearizations constitute the outer-loop, while iterative minimisation of each of the linear problems constitutes the inner loops.

A literature review of recent work on nonlinear estimation (presented here in Section 2) has provided some evidence to support the contention that the 4D-Var method, at least in principle, is a practical method for significantly nonlinear problems. Moreover, the review pointed out that the problems posed by nonlinearity are not unique to the 4D-Var method. Alternative approaches, such as the ensemble Kalman filter, make different but equally stringent assumptions about the linearity of the underlying estimation problem. Based on the review, discussions in Fisher et al. (2005) and further, new results presented herein, we suggest that one way in which the challenges could be met is through a 'weak-constraint' formulation (accounting for model error) of 4D-Var with a combined 'inner/outer' solution algorithm (Section 5).

In the ECMWF operational system (Rabier et al. 2000), the tangent-linear $\mathbf{M}$ (and adjoint $\mathbf{M}^{\mathrm{T}}$) models are run at a lower resolution than the nonlinear model M. Furthermore, the physics is simpler in the tangent-linear and adjoint than in the nonlinear model. Trémolet (2004) investigated the errors introduced by these approximations, by comparing the output of the linear model $\mathbf{M}(\delta\mathbf{x})$ (i.e. an evolved analysis increment) with the finite difference $M(\mathbf{x}+\delta\mathbf{x})-M(\mathbf{x})$ obtained by running the nonlinear model twice, with and without addition of the analysis increment $\delta\mathbf{x}$. Using the spring 2002 version of ECMWF's integrated

forecasting system (IFS), Trémolet (2004) found that the linearization errors (at the time) were larger than expected, and that large errors appeared very early in the assimilation window. An important conclusion was that increasing the 4D-Var (inner-loop) analysis resolution beyond T159 (120 km) required more accurate tangent-linear physics. Recently, Radnóti et al. (2005) revisited the Trémolet (2004) diagnostics, using the improved, spring 2005 versions of the 4D-Var system, with a view to the forthcoming increase in analysis resolution to T255 (80 km). It was shown that the current 4D-Var system could derive benefit from analysis resolution increases up to at least T319 (65 km). On that background, we seek to address two questions raised by the SAC (2004):

- When will the linearization hypotheses (fundamental to 4D-Var) be defeated? There is a scientific risk carried in the proposed resolution increases, longer assimilation window and cloud/rain assimilation.
- What is the optimal ratio between outer and inner-loop resolutions?

In the various configurations of 4D-Var that have been operational at ECMWF the ratio between outer and inner loops has varied: starting from ratio=3.4 in T213/T63 (since November 1997) to ratio=5.1 in T319/T63 (April 1998) improving to ratio=3.2 in the T511/T159 implementation (November 2000). The T799/T255 system currently in pre-operational tests has ratio=3.1.

These two questions are investigated in the context of the current 4D-Var system for a range of resolutions up to T799/T319, assimilating the full set of observations operationally used. We recognize that the behaviour of the system may change in the future due to further developments affecting the consistency between nonlinear and tangent-linear models (Bouttier 2001) and observation operators (Chevallier et al. 2004; Andersson et al. 2005), making it necessary to regularly revisit these issues. A highly nonlinear long-window 4D-Var (Fisher et al. 2005) has been studied in an idealized context (the Lorenz 1995 40-parameter model), comparing strong and weak-constraint formulations. We show that nonlinearity is reduced in weak-constraint 4D-Var, with the implication that the linear assumption is significantly less severe when model error is accounted for in the assimilation.

Our literature review on nonlinear estimation in meteorology and oceanography is presented in Section 2. The main results of the two accompanying papers, with implications for higher resolution and longer-window assimilation, are summarized in Sections 3 and 4, respectively. In Section 5, we present results from simulations with a long-window 4D-Var in the 40-parameter Lorenz (1995) model, contrasting strong and weak-constraint formulations. Conclusions and perspectives are given in Section 6.

2. Brief literature review on nonlinear estimation

2.1 Non-Gaussian estimation

Linear estimation assumes error statistics are Gaussian (Lorenc 1986). The issue of nonlinearity is thus closely linked to non-Gaussian distributions of error. The non-Gaussian pdfs can be specified inputs to the assimilation (through statistical models for background and/or observation errors) or they arise through application of a nonlinear atmospheric model M or nonlinear observation operators H to otherwise Gaussian error covariances. In variational systems, either case will result in a non-quadratic cost function, with potential for multiple minima.

Nearly all statistical analysis methods that have been proposed for numerical weather prediction assume that background and observation errors have essentially Gaussian distributions. This is clearly the case for all current flavours of ensemble Kalman filter (e.g. Houtekamer and Mitchell, 1998; 2001; Hamill et al. 2001; Kalnay et al. 2002; Ott et al. 2004; Whitaker and Hamill, 2002) where, despite the use of the full nonlinear model to propagate the ensemble, the analysis step is linear and completely determined by second-order (covariance) statistics. In other words: the forecast ensemble members may represent a discrete non-Gaussian pdf, but only its Gaussian approximation is accounted for in the analysis step of the filter.

In 4D-Var, it is possible to make the contributions to the cost function non-quadratic, corresponding to a non-Gaussian pdf. This is already done for the observation cost function to implement variational quality control (Andersson and Järvinen 1999), and to account for the directional ambiguity of scatterometer winds (Stoffelen and Anderson 1997). In the current operational humidity analysis (Hólm *et al.* 2002) the non-Gaussian behaviour of the humidity background error near saturation and due to the absolute limit at zero has been accounted for through a nonlinear variable-transformation at the outer-loop level.

In practice, the difficulties associated with minimising a complicated cost function (Pires *et al.* 1996) together with the computational requirement to use efficient incremental solution methods (Laroche and Gauthier 1998; Lawless *et al.* 2005; Fisher 1998) impose serious restrictions on the degree to which fully general non-Gaussian background error statistics could be incorporated into 4D-Var. Furthermore, the usefulness of the maximum-likelihood solution is questionable in the case of a multi-modal pdf (Lorenc 2003b). In that case, the analysis can not usefully be represented by one single realisation, but a sample of likely outcomes is required to represent the analysis pdf (next subsection). But, it is worth noting that in current global NWP there is little evidence to suggest that background errors are significantly non-Gaussian (even for humidity after the suitable transformation, Hólm *et al.* 2002). It is thought, however, that non-Gaussian behaviour becomes more prominent at higher resolution and for precipitation (Errico et al. 2000), in particular on the convective scale (SRNWP 2004; Bouttier 2003).

2.2 Monte-Carlo approaches to data assimilation

Fully abandoning the assumption of Gaussian error statistics requires a Monte Carlo approach. The theory behind such methods (known variously as particle filtering, sequential importance sampling (SIS), bootstrap filtering, condensation, etc.) has developed rapidly over recent years (see, e.g. Doucet *et al.* 2001). The ideas have been applied with success to a variety of highly nonlinear, low dimension estimation problems (see for example: Vlassis, *et al.* 2001; Gustafsson *et al.* 2002; van Leeuwen 2003).

The essence of the Monte Carlo approach is to assign a weight to each member of an ensemble of forecasts in such a way that the sum of all the weights is equal to one. The weighted ensemble defines a discrete pdf that may be thought of as an approximation to the true, continuous pdf. In the fully general case, the discrete pdf is the best approximation we can hope for, since it is impossible (without making assumptions about the shape of the pdf) to know the probability associated with points in phase space that do not coincide with a member of the ensemble. The theory of the particle filter is described in the Appendix.

The Particle Filter has several attractive properties. It is statistically well founded, and is easy to implement, requiring no matrix inversions, adjoints or iterative solution methods. A particular difference from more conventional approaches is that observations are not actually assimilated. That is, they are not used to modify the model's state vector, which remains unperturbed. The observations are used only to determine the weight allocated to each member. Thus, the delicate dynamical and physical balances of the model remain undisturbed.

Despite these attractive properties, the particle filter approach also has some serious difficulties, which may prevent practical application for problems with large dimension. The approximation of the continuous pdf of background error by the discrete pdf of weighted ensemble members is clearly crude unless the ensemble is huge. For reasonable ensemble sizes, the expected mean-square error is likely to be large in comparison with other analysis methods for NWP, and decreases slowly as the number of members is increased (Crisan and Doucet, 2002). The necessary re-sampling step is an additional source of random error.

In part, the inefficiency of the particle filter for large-dimension systems results from the reduction of entire model states to single probabilities (weights). It is known that the ensemble size required for an ensemble Kalman filter may be drastically reduced by adopting a local analysis strategy. It is possible that a similar local approach may be developed in the future for the particle filter. However, at present it is unclear how this might be achieved. Furthermore, the stitching together of local analyses that would be entailed would seriously undermine one of the particle filter's chief advantages, namely that the analysis procedure does not perturb the model.

In our view there is no proven practical data assimilation methodology available for the convective scale, unless nonlinearity can be limited by shortening the analysis period: Using the same 2 km resolution cloud-resolving model Zhang et al. (2004) developed an ensemble Kalman filter with 5-minute cycling, and Sun (2005) constructed a strong-constraint 4D-Var with 10-minute cycling, for assimilation of radar data. For weakly nonlinear (i.e. weakly non-Gaussian) problems the 4D-Var algorithm still applies. Weak-constraint 4D-Var (next subsection) can extend the validity of 4D-Var to longer assimilation periods, even in the presence of nonlinearities, as it can shorten the retention of information on the less-predictable scales through careful specification of the model error term.

2.3 Long-window assimilation and weak-constraint 4D-Var

The weak-constraint 4D-Var system (Zupanski 1997; Trémolet 2003) is truly four-dimensional in the sense that the atmospheric state vector is determined at a succession of times within the assimilation window. In the strong-constraint 4D-Var, on the other hand, covariance evolution is deterministically obtained by evolving the background-error covariance matrix with the tangent linear model (Andersson 2003). This strongly constrains the covariance at every time within the window (Lorenc 2003a) and forces the observational information at any time to be linked with perturbations (or increments) at initial time.

The equivalence between weak-constraint 4D-Var and the fixed-interval Kalman smoother was brought to the attention of the meteorological community by Ménard and Daley (1996), although the proof of this equivalence dates back to the work of Rauch *et al.* (1965) and Bryson and Frazier (1963). Further proofs are given by Li and Navon (2001). The equivalence states that, for a linear model, and given the same covariance matrices of background error (at the start of the analysis interval), model error and observation error, weak constraint 4D-Var produces the same estimate of the state throughout the analysis interval as is given by the fixed-interval Kalman smoother. Of particular interest for NWP is the solution at the end of the analysis interval, which is identical to that of the Kalman filter.

This result is perhaps unsurprising. It extends to imperfect models a result already well known for the perfect model case. What makes it interesting is the realization by Fisher *et al.* (2005) that it provides a means to implement an un-approximated, full-rank Kalman filter, using variational methods. Fisher *et al.* (op cit.) argued that the influence of the initial covariance matrix of background error on the analysed state at the end of the Kalman smoother analysis interval decreases as the length of the interval is increased. This is a consequence of the presence of both growing and decaying error structures, and of the action of the

covariance matrix of model error, which acts as a "forgetting factor". For a sufficiently long interval, the Kalman smoother analysis at the end of the analysis interval is effectively independent of the initial state, the initial covariance matrix, and of the observations near the beginning of the interval.

Exact equivalence between the Kalman smoother and weak-constraint 4D-Var holds for a linear model. Fisher *et al.* (op cit.) argued that the incremental formulation of 4D-Var may be regarded as a global iteration method, as defined by Jazwinski (1970) for solving the nonlinear smoothing problem. The length of analysis window required to implement a full Kalman filter, using the variational approach, in a full NWP system was estimated as between four and ten days. Note, however, that this is the length of interval required to give complete independence of the final analysed state from the background (and its error covariance matrix). It is likely that analysis quality close to that of the Kalman filter could be achieved with a much shorter analysis interval, given a well-constructed covariance matrix of background error.

There has been relatively little work on long-window, variational data assimilation. Evensen (1997) and Evensen and Fario (1997) applied a gradient descent algorithm to minimize a weak-constraint functional, and showed that the method worked well for the classical three-variable Lorenz equations. It is encouraging that the analysis interval in this case was long enough for the model to exhibit highly nonlinear behaviour. The interval encompassed roughly 50 orbits of the lobes of the attractor, and 14 lobe transitions. Swanson *et al.* (1998) applied strong-constraint 4D-Var to a three-level quasi-geostrophic model. They found no difficulty in locating the global minimum of the cost function for intervals as long as five days.

3. Towards higher resolution

3.1 Diagnostics on linear and incremental approximations - methodology

In this section, we first recall the methodology of Trémolet (2004) and Radnóti et al. (2005). The linear assumption is evaluated in the conditions that are most relevant for the incremental 4D-Var: the perturbations we consider are analysis increments, and their low-resolution linear evolution is evaluated with respect to the operational high-resolution forecast model.

4D-Var consists in minimizing the discrepancy between observations of the atmosphere and a forecast over a period of time called the assimilation window (currently 12 hours, Bouttier 2001). The control variable of the strong-constraint problem is the initial condition of the model, i.e. the atmospheric state at the initial time of the assimilation window. The cost function which is minimized (Lorenc 1986) includes three terms and can be written:

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

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, H is the nonlinear observation operator and J_c is an initialization term used to control gravity waves**.** The latter term will be omitted in the remaining of this paper for simplicity. H computes the observation equivalent at the correct location and time and includes the forecast model in our notation. In its incremental formulation (Courtier et al. 1994), the minimization problem is written as a function of the departure from the background: $\delta\mathbf{x} = \mathbf{x}-\mathbf{x}_b$. At the minimum, $\delta\mathbf{x}$ will be the analysis increment. A first-order approximation of the cost function is given by:

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

where **H** is the linearized observation operator and $\mathbf{d} = \mathbf{y} - H(\mathbf{x}_b)$ is minus the departure from observations. In this notation, the tangent-linear model is **M** embedded in the linearized observation operator. The minimization problem is solved using an iterative algorithm (conjugate gradient or quasi-Newton algorithms): This is the inner loop of 4D-Var. In order to reduce the computational cost of the assimilation, the inner loop is run at lower resolution than the forecast. However, in order to retain the maximum information from the observations, the departures are computed at high resolution. The starting point for the minimization (first guess) is interpolated to the inner-loop resolution using an operator S. After the minimization, the analysis increments at low resolution are projected onto the high-resolution space, operator S^{-1}, and added to the reference state. Then the high-resolution trajectory and the departures are recomputed. This is the outer loop of incremental 4D-Var. The high-resolution nonlinear runs also define the state around which the observation operator is linearized, after interpolation to the inner-loop resolution.

The incremental 4D-Var algorithm is shown schematically in Figure 2. Currently at ECMWF, two iterations of the outer loop are run. For further reduction of the computational cost, the linear physics is omitted in the first inner-loop minimization. Moreover, the first inner-loop minimisation is run at T95 resolution, and the second at T159, following the multi-resolution incremental approach of Veersé and Thépaut (1998).

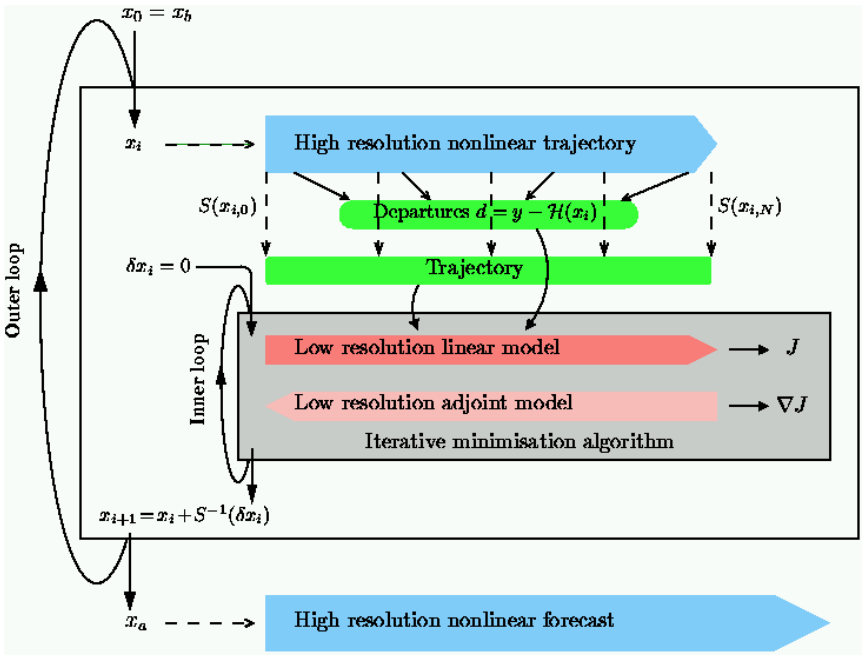

Figure 2: Incremental 4D-Var algorithm: The nonlinear trajectory and departures (d) from the observations (y) are computed at high resolution. The first guess (x_0) and the nonlinear trajectory are interpolated to low resolution using the operator S. The cost function is minimized at low resolution using an iterative algorithm (inner loop). The resulting increment δx is interpolated back to high resolution (symbolized by S^{-1}) and added to the current state x_i. The process is repeated (outer loop, subscript i, currently two iterations at ECMWF) until the analysis x_a is obtained.

In order to diagnose the resulting errors in data assimilation, the output of the linear model used in ECMWF 4D-Var assimilation system and the difference between two runs of the forecast model are compared. The accuracy of the linear model depends on the size and structure of the perturbation - here we use an analysis increment as initial perturbation. Its typical maximum amplitude will be of the order of 3 K and 12 m/s.

The relative error

$$r = \frac{\left\| M(x_i + S^{-1}\delta x_i) - M(x_i) - \mathbf{M}_i(\delta x_i) \right\|}{\left\| M(x_i + S^{-1}\delta x_i) - M(x_i) \right\|}$$

will be presented where x_i is the first guess, δx_i the analysis increment, M the nonlinear forecast model, $\mathbf{M}_i$ the tangent linear model linearized around x_i and S^{-1} is the pseudo-inverse truncation operator. The relative error r is computed for each field and each level. In this study, the nonlinear model was run at resolutions up to T799, which is the horizontal resolution of the next proposed upgrade of the ECMWF operational forecasting system. The linear model is run at a range of resolutions up to T319. The diagnostics presented here are globally averaged errors computed in grid-point space on the model's T255 reduced Gaussian grid. Both models were run with the current operational 60-level vertical resolution.

In most of the experiments presented below, we use a single outer-loop 4D-Var to determine δx_i. This setting is different from the operational one that presently uses two outer-loop iterations. However, this change is appropriate as it provides a single increment, its linear evolution and the corresponding evolution of the nonlinear model finite difference, from which absolute and relative errors that are characteristic of the 4D-Var procedure can easily be computed.

3.2 Tangent linear accuracy as a function of inner-loop resolution

For a given outer-loop resolution it is expected that the relative error r will decrease with increasing inner-loop resolution. However, Trémolet (2004) showed that with the January 2002 version of the forecasting system, improvements levelled off at T159, and no further improvement was seen going from T159 to T255 inner loops. This was a clear indication that the accuracy of the tangent-linear model at the time was not sufficient to support inner-loop resolutions beyond T159.

Carrying out similar tests with the spring 2005 version of the assimilation system, Radnóti et al. (2005) showed that the situation has improved significantly: for T511 outer-loop resolution r for 12-hour evolved increments decreases over the whole range of tested inner-loop resolutions from T42 to T319. The results are shown here in Figure 3, as profiles for wind, temperature and humidity. Additionally, the fourth panel in Figure 3 shows the evolution of surface pressure errors as a function of time within the assimilation window. We see that in the present experiments, a gradual improvement with the resolution increase continues across the whole range of inner-loop resolutions, even up to T319.

This very promising result is the positive outcome of a combination of developments that took place in the last three years, notably in the tangent-linear physics. It is conceivable also that changed structure, or reduced magnitude, of the increments themselves, following changes in the background-error formulation and observation data usage, may have contributed to the result.[1] Attempts so far to identify the main contributor to the improvement have been inconclusive; there appears to be a combination of interacting factors.

[1] This emphasises the fact that these tests have to be carried out in the 4D-Var context

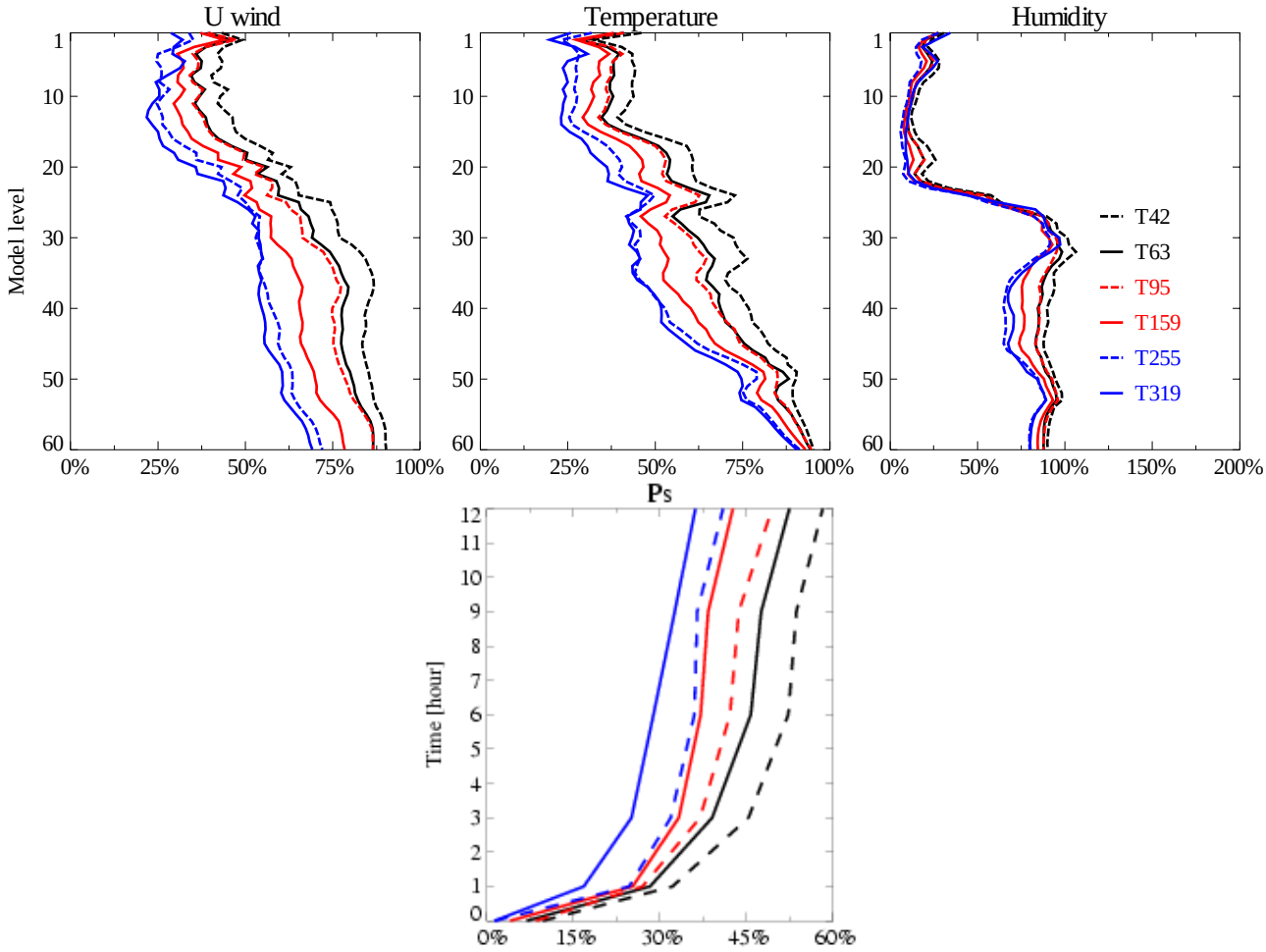

Figure 3: Relative error of the tangent linear model for various resolutions with respect to T511 nonlinear, diabatic model after 12h for the 3 dimensional variables and for the whole 4D-Var window for surface pressure. Diagnostics are computed on the T255 resolution grid. From Radnóti et al. (2005).

3.3 Outer/inner resolution ratio

It is planned that the next version of the forecasting system will run at T799 resolution. Radnóti et al. (2005) showed that the T799 system behaves similar to T511 in terms of linearization errors and their evolution during the 12-hour assimilation window. However, it was shown that in order to preserve the level of relative error *r* of the current T511/T159 system the inner-loop resolution would need to be increased: T799/T159 with a resolution ratio of 5.0 performed clearly worse than T511/T159 with ratio=3.2. As previously noted by Lars Isaksen in Appendix B of Bouttier (2001) it appears that the inner-loop resolution ought to be increased at the same time as the outer-loop resolution, in order to retain or improve the assimilation performance. Is the outer/inner resolution ratio the important factor?

Figure 4 shows *r* for three combinations of outer and inner loop resolutions with very similar ratios: T319/T95 (ratio=3.4), T511/T159 (ratio=3.2) and T799/T255 (ratio=3.1). Compared to the wide range of results in Figure 3, these three curves are very similar to one another. It appears that constant outer/inner resolution ratio of approximately 3 corresponds to near-constant performance in terms of *r* for a range of resolutions. Combining this result with the result in Figure 3 for T319 inner loops, we can conclude that a T1000/T319 assimilation system with 12-hour window would be viable in the future, once the computational resources become available.

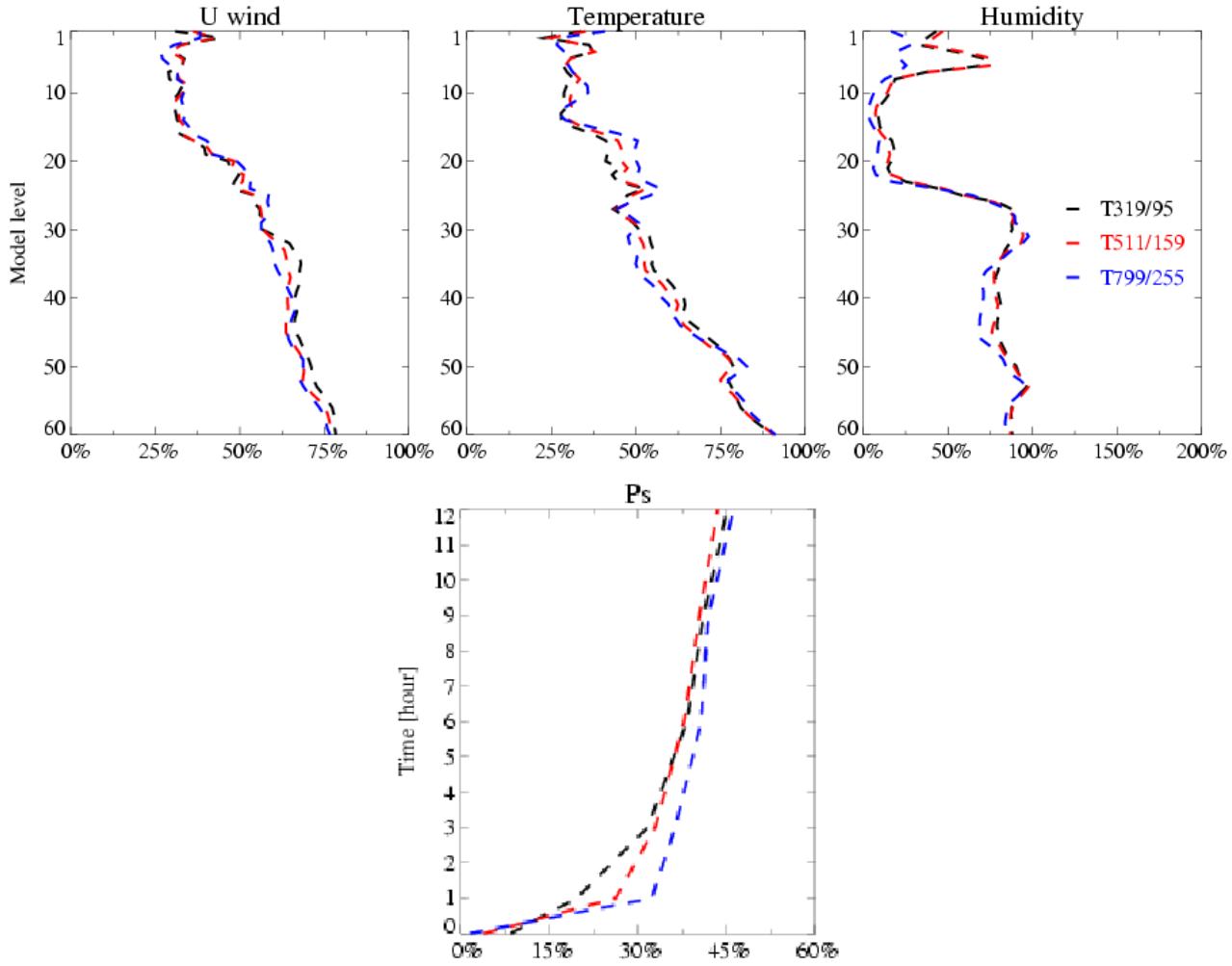

Figure 4 Relative error of the tangent linear model for three combinations of outer and inner resolutions (see legend) with similar resolution ratio close to 3, after 12 hours for the 3 dimensional variables and as a function of time within the assimilation window for surface pressure. Diagnostics are computed on the T255 resolution grid.

The evolution of *r* during the assimilation window is shown for surface pressure in the fourth panel of Figure 4. The plot shows that early on, the relative errors are larger in the higher-resolution systems, and they become more similar after a few (3 to 6) hours of model integration. This is due to a time-step dependence of the evolution of an external gravity wave (Trémolet 2005), which is a source of error in surface pressure as different time-steps are used in inner and outer loops corresponding to each model resolution. We will return to this issue in Section 4.3 below.

3.4 Accuracy of the incremental approach - discussion

The results presented in this section are based on diagnostics that measure the consistency between nonlinear and tangent-linear model, which depends on a number of factors:

- **The nonlinearity of the analysis problem.** This in turn depends on resolution (Radnóti et al. 2005), length of assimilation window (Trémolet 2004) and size and structure of the increments. The size of increments has reduced significantly in recent years (Simmons 2003), as more observations are assimilated (Thépaut and Andersson, 2004), the model has improved, and the background state has

become more accurate. This trend is likely to continue as important aspects of the assimilation system continue to improve. However, the L91 extension of the analysis domain to the lower mesosphere has revealed that analysis increments there can be very large and violate the linearized advection equation - resulting in noisy analyses and unrealistic humidity (currently being studied).

- **The outer/inner resolution ratio** (as demonstrated in the figures above). However, the optimal resolution ratio is difficult to determine. A ratio closer to 1.0 would be the most accurate, but not the most computationally efficient choice. The choice of ratio depends on what is affordable on the available computers. There is a trade-off against other costs, such as additional outer-loop iterations (Trémolet 2005) and/or improved, costlier, tangent-linear physics.

- **The accuracy of tangent linear physics**, as demonstrated by the results of Trémolet (2004) and Radnóti et al. (2005). Newly developed components of the tangent-linear physics (Janisková 2002; Tompkins and Janisková 2004; Lopez and Moreau 2004) are gradually being introduced into 4D-Var, and further improved accuracy can be expected. The additional computational cost is non-negligible.

- **Consistency between inner and outer loop dynamics.** The semi-Lagrangian advection of the nonlinear model is formulated in terms of virtual temperature, with humidity and ozone as grid-point variables (Temperton et al. 2001). However, in the tangent-linear model, spectral temperature, humidity and ozone representation have been used. This inconsistency has recently been removed and preliminary results are shown in Radnóti et al. (2005). Further illustration of its impact is shown here in terms of background and analysis fit to SSM/I data (Figure 5). The closer observation fit to both background and analysis is due to improved consistency between inner and outer loop models, which should be beneficial to outer-loop convergence.

- **The linearization state**, that is, the 12-hour 'trajectory'. The accuracy of the linearization state was improved by Trémolet (2004); the trajectory is now created from the outer-loop high-resolution forecast through interpolation to the inner-loop resolution (Figure 2), rather than by running a low-resolution forecast. This is essential to ensure that weather features evolve in exactly the same manner and location in inner and outer loop models. The trajectories for humidity and ozone have been improved by retaining grid-point representation, throughout (described by Radnóti et al. 2005). Cloud field trajectories are currently created via a low-resolution forecast forced by the trajectory fields, but it is planned that this will be replaced in the near future by grid-point interpolated values from the high-resolution forecast trajectory. These measures should be sufficient to provide accurate linearization states for direct assimilation of cloud and precipitation data (Lopez, pers. comm.).

- **The accuracy of the interpolation method.** Interpolation going from high resolution to low (the truncation operator S in Figure 2) is required for trajectory and background fields. Interpolation in the opposite direction, S^{-1}, is required for the analysis increments, before they are added to the background state to form the analysis. Options for conserving interpolation have been developed and initial tests are promising (see Radnóti et al. 2005).

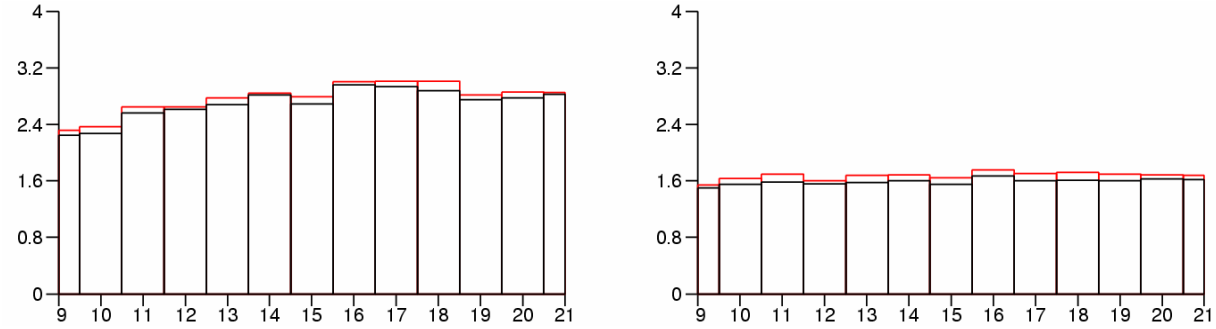

Figure 5 Hourly SSM/I data fits to the background (left) and analysis (right) during the 12 hour 4D-Var assimilation period, comparing spectral (red) with grid-point (black bars) humidity in the assimilating tangent-linear and adjoint models. The statistics are based on eight days of T799/T255 assimilation. Data is presented only for channel 3 in the tropics, but similar results are obtained for other channels and other humidity sensitive satellite and terrestrial data.

The diagnostics presented here show relative errors around 45 % for surface pressure, in the range 45-75 % for temperature and wind in the troposphere, and 70-100% for humidity. Maps such as Figure 6 below (temperature near 500 hPa) and Figure 15 in Radnóti et al. (2005) (humidity near 980 hPa), indicate that the errors generally are quite small, and the globally averaged, point-wise RMS measure used here (that is, *r*) may be dominated by local features and a relatively small number of large (extreme) values. The spectra shown in Figure 7 confirm that the errors (red line) in 12-hour evolved increments are relatively largest for the smallest scales. The known result that the inner and outer-loop analyses fit the observations almost equally well demonstrates that the incremental 4D-Var approach nevertheless works well - this will become evident also in Figure 8 below from the small jumps in *Jo* cost-function between inner and outer loops. We can conclude that the effect of nonlinearity (and the other errors contributing to the above diagnostics) must be rather limited for most of the observations currently assimilated. The presence of larger linearization errors, primarily in the boundary layer and in the vicinity of convective systems (e.g. the Caribbean, Figure 6), is likely to be of some detriment to the analysis, locally. The possibility of improving the representation of nonlinear effects by iterating the outer-loop closer to convergence is investigated in the next section.

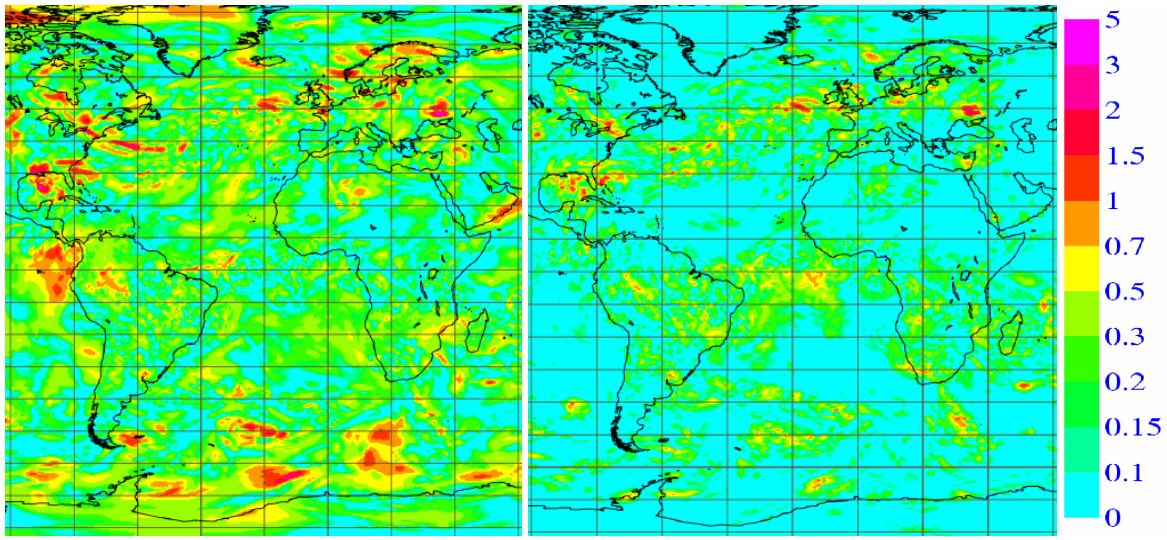

Figure 6 Twelve-hour evolved temperature analysis increment (K) at model level 39 (~500 hPa), 1 February 2004, 03 UTC, as obtained by finite difference with the nonlinear model (left), and the corresponding error (right) in the tangent-linear evolution, i.e. its difference with respect to the panel on the left. Shading intervals are logarithmic, see legend.

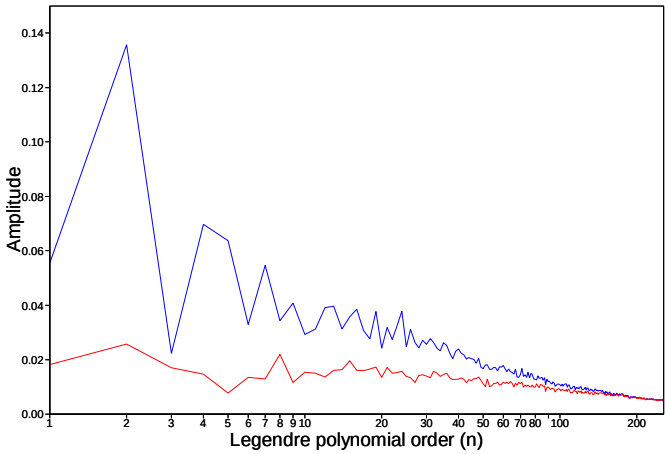

Figure 7 Global spectra corresponding to the two fields shown in Figure 6, that is, temperature at level 39 12-hour nonlinearly evolved increments (blue) and the error due to the tangent-linear approximation (red).

4. Improving the representation of nonlinear effects

4.1 Outer-loop convergence

In order to extend the current 4D-Var approach to the more nonlinear analysis problem expected in the coming years, it is likely that it will become necessary to perform additional outer-loop iterations to re-linearize the problem more frequently (c.f. Figure 1). In the current implementation of 4D-Var, and ever since its original implementation in 1997, two outer iterations are performed: the tangent linear model and observation operators are thus re-linearized once, only. It was noted by Rabier et al. (2000) that additional outer iterations were detrimental. Bouttier (2001) also noted that additional outer-loop iterations led to divergence away from the correct solution. A detailed investigation of this problem has now been carried out (Trémolet 2005) and the outer-loop convergence properties of 4D-Var have been re-evaluated using the spring 2005 version of the forecasting system.

4.2 Experiments at operational resolution

Figure 8 and Figure 9 show the evolution of the various components of the cost function and its gradient in a configuration similar to today's operational 4D-Var at T511/T159 resolution. 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.

In Figure 8, solid lines represent the *Jo* component of the cost function, as computed in the inner loop, that is, from departures $\mathbf{d}_i = \mathbf{y} - H(\mathbf{x}_i) - \mathbf{H}(\delta\mathbf{x}_i)$. The bars represent the value of *Jo* as computed in the nonlinear trajectory runs from departures $\mathbf{d}_{i+1} = \mathbf{y} - H(\mathbf{x}_i + \delta\mathbf{x}_i)$. The figure also shows the evolution of the other components of the cost function: *Jb* (dashed lines) and *Jc* (dotted lines). These are only accessible at low resolution, in the inner loop. The values of *Jo*, *Jb* and *Jc* 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 *Jb* has been inflated by a factor of 10 and *Jc* by a factor of 100 to make the figures more readable. The jump in *Jb* between the end of the first minimisation and the beginning of the second minimisation is due to the change of resolution and the implied change in the **B** matrix.

Figure 9 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, Fisher 1998) during the minimisation, the dashed curve shows the actual gradient computed before and after the minimisation, joined in a straight line. The reason for the increase of the gradient norm at the beginning of each minimisation is discussed in appendix A of Trémolet (2005).

From Figure 8 we see that the minimum value for *Jo* is obtained after four outer-loop iterations. After that, *Jo* starts increasing slowly and 4D-Var diverges. The gradient (Figure 9) continues to decrease for two more outer loop iterations but then increases as well.

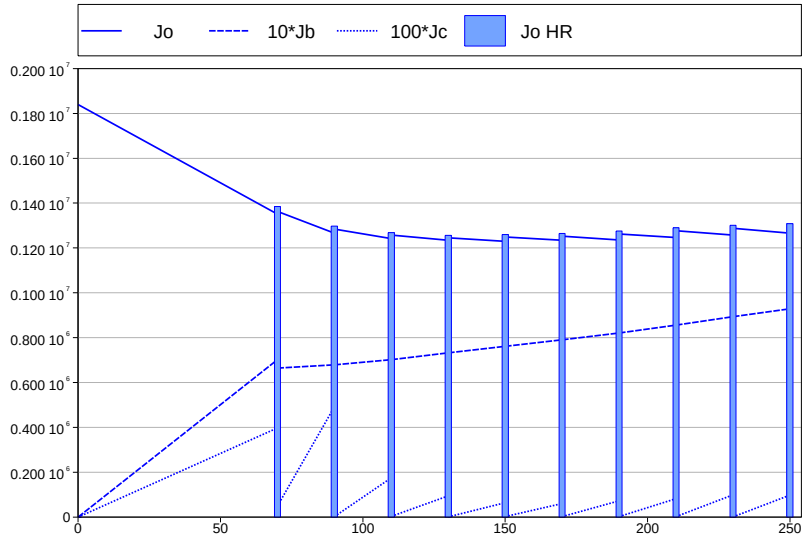

Figure 8 Evolution of the components of the 4D-Var cost function at T511/T159 resolution, as a function of the total number of inner loop iterations. The solid lines represent Jo as computed in the inner loop, while the bars represent Jo as computed in the nonlinear trajectory runs. Jb (dashed) has been inflated by a factor of 10, and Jc (dotted lines) by a factor 100 for readability. From Trémolet (2005).

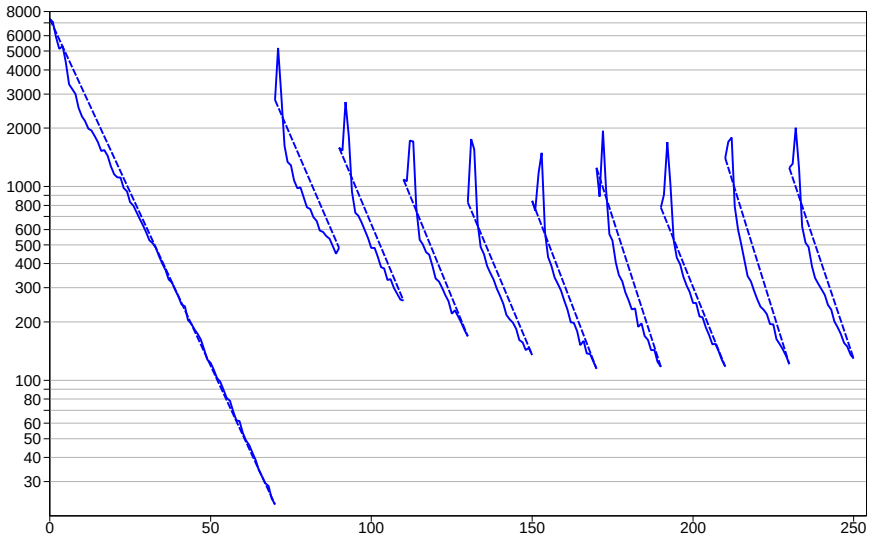

Figure 9 Evolution of the norm of the gradient of the cost function in 4D-Var as a function of the total number of inner loop iterations. The plain lines show the estimated gradient norm during the minimisation, the dashed lines show the actual norm at the beginning and at the end of the minimisation. From Trémolet (2005).

4.3 Resolution and time-step dependence

Is the lack of convergence due to the change in resolution between inner and outer loops? This has been tested in a set of experiments at T255 outer loop resolution, with varying inner-loop resolution. All the minimisations were run with linearized physics. Figure 10 shows the evolution of the cost function for inner loop resolutions of T255, T95 and T42.

When inner and outer loop resolutions are the same (T255/T255), Figure 10 shows that 4D-Var does converge (this is also true for a T159/T159 experiment, not shown). With the T159 inner loop, 4D-Var still converges but the final *Jo* is slightly higher. As the resolution goes down to T95 and T42, 4D-Var diverges. The discrepancies between the *Jo* component of the cost function at the end of a minimisation, in the high-resolution trajectory and at the beginning of the following minimisation diminish with the T159 and T255 inner loops and increase with the mismatch in resolution. The RMS profiles of the increments produced by the converging T255 and the diverging T95 experiments are shown on Figure 11 for the first, second and final two minimisations. The increments for the first and second minimisations are the largest and similar in amplitude for both T255 and T95 inner loop resolutions (the plain and dashed lines are similar). The increments from the 9th and 10th minimisations obtained with the T255 inner loops are very small in amplitude (dashed green and magenta curves), indicating that the algorithm has properly converged and that no additional increments are needed. At T95, however, the 9th and 10th increments are non-zero, indicative of the convergence problem.

The corresponding map (T255/T95 10th iteration) of surface pressure increments (Figure 12) shows a very large-scale ring pattern centred over Europe. The mismatch in resolution between inner and outer loops seems to be an important factor for outer-loop divergence. However, each resolution is associated with its own model time step. In order to separate the time-step dependence from the resolution dependence a further set of experiments was performed. It was found that, as long as the model time step is the same in both inner and outer loop, the outer loop converges (not shown).

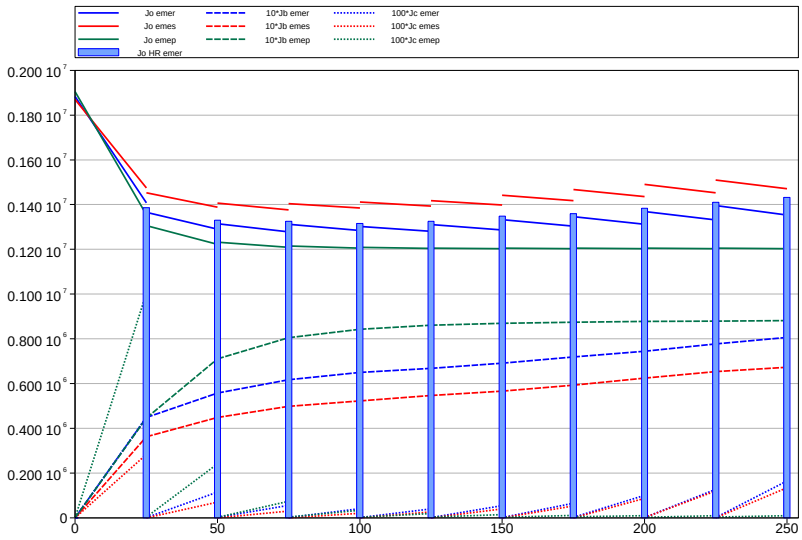

Figure 10 Evolution of T255/T42 (red), T255/T95 (blue) and T255/T255 (green) 4D-Var cost functions, otherwise like Figure 8. From Trémolet (2005).

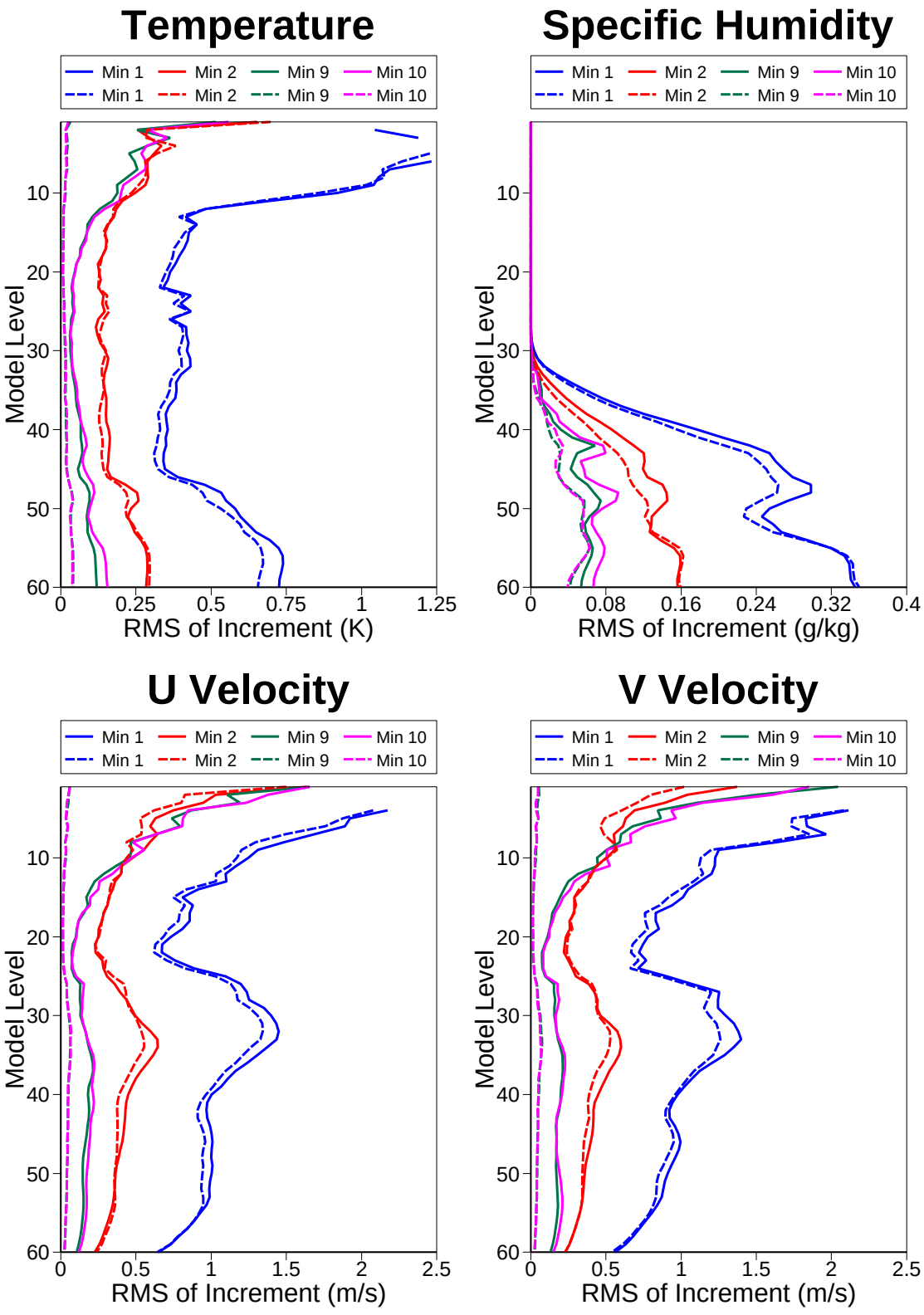

Figure 11 RMS of the 4D-Var increments for minimisations 1 (blue), 2 (red), 9 (green) and 10 (magenta) obtained with T95 (solid lines) and T255 (dashed lines) inner loops, both with T255 outer loops. The T255/T255 analysis increments are small at iterations 9 and 10 which indicates that it has converged, in contrast to T255/T95. From Trémolet (2005).

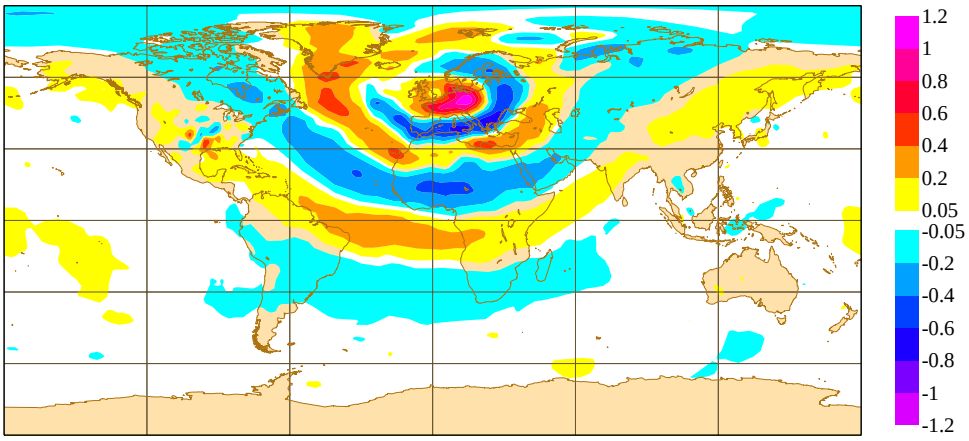

Figure 12 Surface pressure increments (hPa) at outer-loop iteration 7, for the T255/T95 experiment. From Trémolet (2005).

4.4 Discussion and a proposal for a third outer iteration

The general pattern of the surface pressure increment (Figure 12) is near-circular centred over northern Europe. Trémolet (2005) demonstrated that the inner and outer-loop evolutions of this pattern slowly diverge from each other over time within the assimilation window. This is a gravity wave, which propagates with different phase speeds in the two 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 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 nonlinear model behaviour with respect to phase speed accurately for a range of time-step values. This is not a linearization issue in that, given the same time step, the linear and nonlinear models behave similarly.

Options for how to resolve the problem are extensively discussed by Trémolet (2005). Using the same time step in inner and outer loops is not an option in practice due to prohibitive computational cost. Instead, we will seek to reduce the dominance of the surface pressure pattern over Europe, by thinning (or combining) some of the densest data in the affected area - as a pragmatic solution. The addition of a third or fourth outer iteration is then doable. A reduced number of inner iterations combined with the addition of a third outer iteration is proposed for evaluation and possible operational implementation in 2006.

5. Long window assimilation

In this section, we consider the different impacts of nonlinearity on strong-constraint and weak-constraint 4D-Var, for a simplified system. We use the analysis system for the Lorenz (1995) model described by Fisher *et al*. (2005). The model consists of a set of 40 coupled ordinary differential equations, and was devised as a simple analogue of mid-latitude tropospheric dynamics. It was used to study observation targeting by Lorenz and Emanuel (1998).

5.1 Weak and strong-constraint cost functions

The solid line in Figure 13 shows a cross-section of the observation cost function for a hypothetical strong-constraint 4D-Var for the Lorenz (1995) model:

$$J_o\left(\mathbf{x}_0\right)=\sum_{k=0}^{K}\left(\mathbf{y}_k-M_{t_0\to t_k}\left(\mathbf{x}_0\right)\right)^{\mathrm{T}}\mathbf{R}_k^{-1}\left(\mathbf{y}_k-M_{t_0\to t_k}\left(\mathbf{x}_0\right)\right) \qquad (1)$$

The cost function is non-incremental, and uses the full nonlinear model, $M_{t_0\to t_k}$, to propagate the initial state at time t_0 to the observation times, t_k. The analysis window in this case was 20 days, and observations were specified at 3 out of every 5 grid-points, and at 6-hourly intervals. The observation error, σ_o, and the parameters of the model were identical to those used by Fisher et al. (2005). The cross-section was taken along the line $\mathbf{x}_0=\mathbf{x}_0^a+\varepsilon\mathbf{v}_0$ for $-\sigma_o\le\varepsilon\le\sigma_o$, where $\mathbf{v}_0$ is a vector of independent random numbers with unit variance, and $\mathbf{x}_0^a$ is the state at the start of the analysis window analysed by the Fisher et al. (2005) weak-constraint 4D-Var.

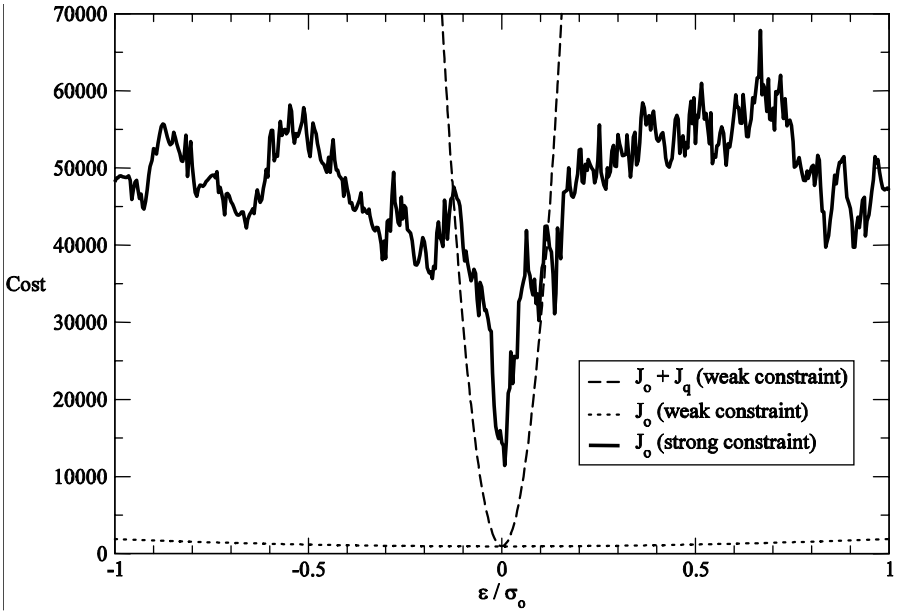

Figure 13: Cross sections in random directions of the observation cost function, J_o, for strong-constraint 4D-Var (solid) and weak-constraint 4D-Var (dotted), and of the total weak-constraint cost function, J_o+J_q. In all cases, the assimilation window is 20 days, and the standard configuration of the Lorenz (1995) model was used.

It is clear from Figure 13 that the nonlinear propagation of initial perturbations over 20 days results in a cost function with many minima. A conventional, gradient-based minimization method (such as the quasi-Newton method) would be unable to locate the global minimum of the cost function given a starting point just a small fraction of σ_o away from the minimum. The cost function cannot be approximated by a quadratic function, implying that the model is not well-approximated by its tangent linearization for an analysis window of this length.

In contrast, the dotted line in Figure 13 shows the observation cost function for the corresponding weak-constraint 4D-Var system, as used by Fisher et al. (2005). This cost function is quadratic by construction. In a weak-constraint analysis, the control variable consists of the entire four-dimensional state, and the model is not required to propagate the state from an initial time to the time of the observation. The observation cost function is simply:

$$J_o^{\text{weak}}\left(\mathbf{x}_0,\ldots,\mathbf{x}_K\right)=\sum_{k=0}^{K}\left(\mathbf{y}_k-\mathbf{x}_k\right)^{\mathrm{T}}\mathbf{R}_k^{-1}\left(\mathbf{y}_k-\mathbf{x}_k\right) \tag{2}$$

and is plotted along the line $\mathbf{x}_k=\mathbf{x}_k^a+\varepsilon\mathbf{v}_k$, where, for $k=0\ldots K$, the vectors $\mathbf{v}_k$ are independent vectors of independent, unit-variance random numbers; and where $\mathbf{x}_k^a$ is the four-dimensional analysis produced by the weak-constraint system.

Weak-constraint 4D-Var incorporates an additional term in the cost function that is not present in strong-constraint 4D-Var. The term penalizes departures of the four-dimensional analysis from a model trajectory. In the Fisher et al. (2005) system, this "model-error" term is expressed in terms of departures from an initial (four-dimensional) guess, $\mathbf{x}_k^g$, as:

$$J_q\left(\delta x_0,\ldots\delta x_K\right)=\sum_{k=1}^{K}\left(\mathbf{q}_k+\delta\mathbf{x}_k-\mathbf{M}_{t_{k-1}\to t_k}\delta\mathbf{x}_{k-1}\right)^{\mathrm{T}}\mathbf{Q}_k^{-1}\left(\mathbf{q}_k+\delta\mathbf{x}_k-\mathbf{M}_{t_{k-1}\to t_k}\delta\mathbf{x}_{k-1}\right) \tag{3}$$

where $\mathbf{q}_k=\mathbf{x}_k^g-M_{t_{k-1}\to t_k}\left(\mathbf{x}_{k-1}^g\right)$, and where $\mathbf{M}_{t_{k-1}\to t_k}$ represents the tangent-linear model integrated from time t_{k-1} to time t_k.

Note that the model-error term is quadratic here. This was achieved by linearizing the model operator $M_{t_{k-1}\to t_k}$ about the first guess, assuming quasi-linear propagation of perturbations for the interval $t_{k-1}\le t\le t_k$. Such linearization was not possible for the strong-constraint cost function (Eq. 1), where the model operator appearing in the cost function represents integration of the model over the (typically) much longer interval $t_0\le t\le t_k$.

The dashed line in Figure 13 shows the sum of J_o and J_q for the weak-constraint 4D-Var. The curve has a single, global minimum that is guaranteed to be found by any conventional minimization algorithm. (Note that the Fisher et al. (2005) analysis has no background term, so that the dashed curve in Figure 13 represents the full cost function for the system. Incorporation of a background term would add a further quadratic term to the cost function.)

5.2 Cycling of long-window, weak-constraint 4D-Var

Cycling of a long-window, weak-constraint 4D-Var analysis cannot be done in the same way as for a short-window system. In particular, the analysis can no longer be regarded as providing only an initial state at the start of the analysis window, from which states at later times can be calculated by integrating the numerical model. For an analysis window length of several days, such an integration will depart significantly from reality towards the end of the window. This affects two areas of the cycling algorithm in particular: construction of the first guess, and the incremental formulation.

Construction of an accurate first guess[2] is crucial in reducing the effects of nonlinearity in the analysis. An inaccurate first guess (for example a first guess provided by a model integration over several days) will provide a poor state about which to linearize the model, and will require large increments. Fisher et al. (2005) addresses this problem by using overlapping analysis windows. For any given analysis cycle, the first guess is provided largely by the previous analysis, with just a short final section of the first guess coming from a forecast. The method is shown schematically in Figure 14. An alternative method was proposed by Pires et al. (1996). (Note that Figure 14 does not show the cycling of background fields. Statistical considerations dictate that the background field for an analysis cycle must be independent of the observations used during that cycle, for example, from an earlier, non-overlapping analysis cycle).

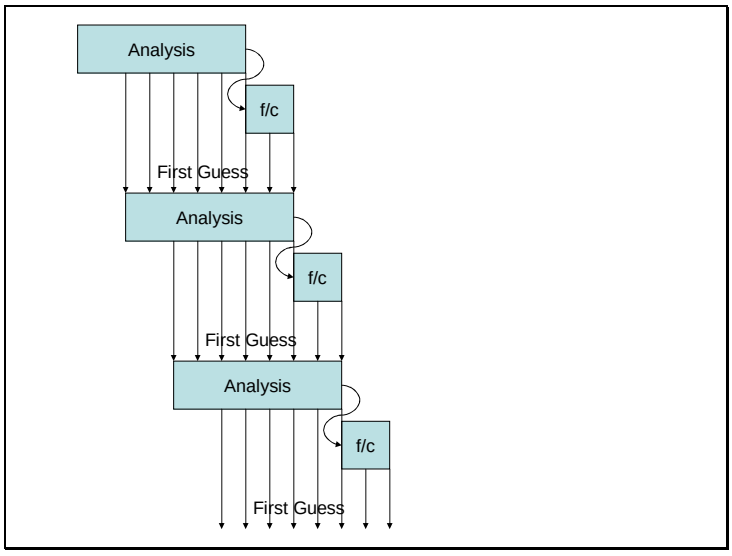

Figure 14: Cycling of the first guess in long-window 4D-Var For any analysis cycle, the first guess is provided by the previous analysis cycle, augmented by a short forecast to cover the non-overlapping section of the analysis window.

For a linear system, and in the limit of an infinitely-long analysis window, there is an exact correspondence between the Kalman filter and weak-constraint 4D-Var (Ménard and Daley, 1996). For weakly nonlinear systems, both 4D-Var and the extended Kalman filter (EKF) provide different approaches for handling nonlinearity. The EKF linearizes the model used to propagate the covariances about the most recent first guess. However, being a sequential method, the linearization is performed only for the most recent analysis cycle. By contrast, 4D-Var re-linearizes the problem for the entire analysis window at every analysis cycle. If overlapping analysis windows are used, the first guess (i.e. linearization state) for a timestep in the middle of

[2] Note that we make a distinction here between the first guess and the background. The former is the four-dimensional state that defines the starting point for the minimization. The latter is defined only at the start of the analysis window, and provides *a priori* information about the state at the start of the window.

the analysis window is taken from an analysis that has already used most of the observations that are used in the current cycle, and is consequently highly accurate. It is interesting to note that, for sufficiently long analysis windows, the errors of the 4D-Var analyses performed by Fisher et al. (2005), and subsequent analyses performed using the same Lorenz (1995) system (not shown), are marginally smaller than those of the corresponding EKF. It is likely that this slight superiority of 4D-Var over the EKF results from the use of more accurate linearization states in 4D-Var.

In the current ECMWF analysis system, the incremental algorithm (Figure 2) is implemented by updating intermediate high-resolution trajectories at the start of the analysis window, using the low-resolution increment provided by the minimization. The trajectory for the rest of the analysis window is constructed by integrating the high-resolution model. This approach cannot be used for long analysis windows. Rather, the minimization must be regarded as providing increments throughout the analysis window. The updated high-resolution state at a given time-step is constructed without integrating the model, by adding the low resolution increment for the time-step to the corresponding, contemporaneous first guess fields. In fact, this approach is already used for the final incremental update in the current ECMWF analysis.

6. Conclusions and perspectives

Issues affecting the feasibility of the 4D-Var approach in the presence of nonlinearity have been investigated and discussed, in the context of global NWP with ECMWF's 4D-Var data assimilation system. Nonlinearity is considered a deciding factor for dealing with e.g. direct assimilation of cloud and precipitation data, dealing with non-Gaussian error distributions and higher-resolution analysis.

In the accompanying paper by Radnóti et al. (2005), it has been shown that increased inner-loop resolution brings the tangent linear model closer to the finite differences of the nonlinear model in the whole range of tested inner-loop resolutions from T42 to T319. This is an important improvement with respect to the earlier model versions, where no improvement was found beyond T159. This will benefit the forthcoming resolution increases to T799/T255, and it makes further increased inner-loop resolution conceivable. It has been found that the T799/255 higher-resolution configuration (currently in pre-operational testing) shows slightly better linearity error characteristics than the operational T511/159 configuration. It appears outer/inner resolution ratio ≈ 3 provides satisfactory results. We conclude that at least a T1000/T319 assimilation system will be viable in the future, once the computational resources become available.

The reason for outer-loop divergence in the current 4D-Var system has been investigated and explained. Options for how to resolve the problem are extensively discussed in the accompanying paper by Trémolet (2005). Using the same time step in inner and outer loops is not an option in practice due to prohibitive computational cost. Instead, we will seek to reduce the dominance of the identified surface pressure pattern over Europe, by thinning (or combining) some of the densest data in the affected area - as a pragmatic solution. Provided the computational cost can be afforded, additional outer loop iterations could then be introduced, providing more frequent re-linearizations of the tangent-linear model and observation operators and a robust framework for a nonlinear 4D-Var can be established. The benefits of adding a third outer loop iteration to the operational system are being investigated in the short term.

Results with a long-window strong-constraint 4D-Var in the Lorenz (1995) system, shows a very convoluted cost function with multiple minima that would be difficult or impossible to minimise. However, the weak-constraint counterpart is quadratic. In weak-constraint 4D-Var a sequence of model states are estimated (rather than just the initial state), with the consequence that the tangent linear assumption is relied upon only for the shorter time segment between successive state estimates, and not for propagation of information

throughout the assimilation window. This property makes a weak-constraint 4D-Var an attractive prospect that should be pursued as a main avenue of research in the coming years, with the goal to lengthen the assimilation window. There are substantial challenges involved, not least in constructing a suitable statistical representation of model error. Unpredictable features, and those that are highly nonlinear for perturbations the size of a fraction of an analysis increment will nevertheless be poorly analysed, e.g. convective scale features.

A long-window extension of ECMWF's 4D-Var system will rely on the model error term to provide the appropriate temporal retention of information (memory) at all spatial scales: several days for synoptic scales, down to a few hours for the smallest resolved scales. This will be an important subject for research in the coming years. In the storm-track regions, it is thought that the model error source of variance will result in analysis perturbations that quickly evolve into singular-vector like structures to fit the available observations; the precise covariance structure of the model error might itself not be a vital concern. In the tropics, a long-window 4D-Var offers the prospect of projecting the observational information onto the slowly evolving large-scale equatorial modes of divergent motion (Kelvin waves and mixed Rossby-gravity waves) (Žagar *et al.* 2004; 2005).

Diagnostics on linearity will be regularly reassessed, in particular once observational data and observation operators for direct assimilation of cloud (Tompkins and Janisková 2004; Lopez and Moreau 2004) and precipitation (Moreau et al. 2004; Bauer et al. 2005a; 2005b) become available, and are implemented in the IFS. The configuration of the operational 4D-Var system, in terms of window-length, resolutions and timeliness is subject to operational constraints, and depends on the availability of computing resources. Combinations of long-window with long data cut-off times, with shorter window 4D-Var with short data cut off, can be seen as a natural extension to the 'Early Delivery Suite' (Haseler 2004) in current operations.

References

Andersson, E. and H. Järvinen, 1999: Variational quality control. *Q. J. R. Meteorol. Soc.*, **125,** 697—722.

Andersson, E., 2003: Modelling the temporal evolution of innovation statistics. Proc. ECMWF Seminar on "Recent Developments in Data Assimilation for Atmosphere and Ocean", 8—12 Sept 2003, Reading, U.K., 153—164.

Andersson, E., C. Cardinali, M. Fisher, E. Hólm, L. Isaksen, Y. Trémolet and A. Hollingsworth, 2004: Developments in ECMWF's 4D-Var system. Proc. AMS Symposium on "Forecasting the Weather and Climate of the Atmosphere and Ocean". Seattle, Washington State, 11-15 Jan. 2004. Available from the American Meteorological Society, http://ams.confex.com/ams/84Annual/20WAF16NW/program.htm

Andersson, E., P. Bauer, A. Beljaars, F. Chevallier, E. Hólm M. Janisková, P. Kållberg, G. Kelly, P. Lopez, A. McNally, E. Moreau, A. J. Simmons, J.-N. Thépaut and A. M. Tompkins, 2005: Assimilation and modeling of the atmospheric hydrological cycle in the ECMWF forecasting system. *Bull. Amer. Meteorol. Soc.*, **86,** 387—402.

Bauer, P., P. Lopez, A. Benedetti, D. Salmond and E. Moreau, 2005a: Implementation of 1D+4D-Var assimilation of microwave radiances in precipitation at ECMWF. Part I: 1D-Var. To appear as ECMWF Tech Memo.

Bauer, P., P. Lopez, A. Benedetti, D. Salmond, S. Saarinen and M. Bonazzola, 2005b: Implementation of 1D+4D-Var assimilation of microwave radiances in precipitation at ECMWF. Part II: 4D-Var. To appear as ECMWF Tech Memo.

Bouttier, F., 2001: The development of 12-hourly 4D-Var. *ECMWF Tech. Memo.* **348,** pp 21.

Bouttier, F., 2003: The AROME mesoscale project. Proc. ECMWF Seminar on "Recent Developments in Data Assimilation for Atmosphere and Ocean", 8-12 Sept 2003, Reading, U.K. 433—447.

Bryson, A. E. and M. Frazier, 1963: Smoothing for Linear and Nonlinear Dynamic Systems. Proc. Optimum Sys. Synthesis Conf., U.S. Air Force Tech. Rept. ASD-TDR-63-119.

Chevallier, F., P. Lopez, A. M. Tompkins, M. Janisková and E. Moreau, 2004: The capability of 4D-Var systems to assimilate cloud affected satellite infrared radiances. *Q. J. R. Meteorol. Soc.*, **130,** 917—932.

Courtier, P., J.-N. Thépaut and A. Hollingsworth, 1994: A strategy for operational implementation of 4D-Var, using an incremental approach. *Q. J. R. Meteorol. Soc*. **120,** 1367—1388.

Crisan D. and A. Doucet, 2002: A Survey of Convergence Results on Particle Filtering Methods for Practitioners. IEEE Transactions on Signal Processing, **50,** 736—746.

Doucet A., J. F. G. de Freitas and N. J. Gordon, 2001: An introduction to sequential Monte Carlo methods. In *Sequential Monte Carlo Methods in Practice*, Eds. A. Doucet, J. F. G. de Freitas and N. J. Gordon, Springer Verlag, New York.

Errico, R. M., L. Fillion, D. Nychka and Z.-Q. Lu, 2000: Some statistical considerations associated with the data assimilation of precipitation observations. *Q. J. R. Meteorol. Soc*. **126,** 339—359.

Evensen, G., 1997: Advanced data assimilation in strongly nonlinear dynamics. *Mon. Wea. Rev.*, **125,** 1342—1354.

Evensen, G. and N. Fario, 1997: A weak constraint variational inverse for the Lorenz equations using substitution methods. *J. Meteor. Soc. Japan*, **75 (1B),** 229—243

Fisher, M., 1998: Minimization Algorithms for Variational Data Assimilation. Proc. ECMWF Seminar on "Recent Developments in Numerical Methods for Atmospheric Modelling", Reading 7-11 September 1998, 364—385.

Fisher, M., 2003: Background error covariance modelling. Proc. ECMWF Seminar on "Recent Developments in Data Assimilation for Atmosphere and Ocean", 8-12 Sept 2003, Reading, U.K. 45—64.

Fisher M., M. Leutbecher and G. Kelly, 2005: On the Equivalence Between Kalman Smoothing and Weak-Constraint Four-Dimensional Variational Data Assimilation. (*Q. J. R. Meteorol. Soc.*, Accepted). Available as *ECMWF Tech. Memo.*, **447,** pp 14.

Gustafsson F., F. Gunnarsson, N. Bergman, U. Forssell, J. Jansson, R. Karlsson and P.-J. Nordlund, 2002: Particle filters for positioning, navigation and tracking. IEEE Transactions on Signal Processing.

Hamill, T. M., J. Whitaker and C. Snyder, 2001: Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter. *Mon. Wea. Rev.*, **129**, 2776-2790.

Haseler, J., 2004: Early delivery suite. *ECMWF Tech. Memo*., **454,** pp 35.

Hólm, E., E. Andersson, A. Beljaars, P. Lopez, J.-F. Mahfouf, A. J. Simmons and J.-N. Thépaut, 2002: Assimilation and modeling of the hydrological cycle: ECMWF's status and plans. *ECMWF Tech. Memo*., **383,** pp 55.

Houtekamer, P. L. and H. L. Mitchell, 1998: Data assimilation using an ensemble Kalman filter technique. *Mon. Wea. Rev.*, **126,** 796—811.

Houtekamer, P. L. and H. L. Mitchell, 2001: A sequential ensemble Kalman filter for atmospheric data assimilation. *Mon. Wea. Rev.*, **129,** 123—137.

Janisková, M., J.-F. Mahfouf, J.-J. Morcrette and F. Chevallier, 2002: Linearized radiation and cloud schemes in the ECMWF model: Development and evaluation. *Quart. J. Roy. Meteor. Soc.*, **128,** 1505—1527.

Jazwinski, A. H., 1970: *Stochastic Processes and Filtering Theory*. Academic Press.

Kalnay, E., M. Corazza, S.-C. Yang, B. Hunt, E. Kostelich, E. Ott, D. J. Patil, I. Szunyogh, J. Yorke and A. Zimin, 2002: Data assimilation via local ensemble Kalman filtering. Proc. ECMWF Seminar on "Predictability of weather and climate", Reading, 9-13 September 2002, 41—45.

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

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

van Leeuwen P. J., 2003: A Variance-Minimizing Filter for Large-Scale Applications. *Mon. Wea. Rev.*, **131**, 2071—2084.

Li, Z. and I. M. Navon, 2001: Optimality of 4D-Var and its relationship with the Kalman filter and Kalman smoother. *Q. J. R. Meteorol. Soc.*, **127,** 661—684

Lopez, P. and E. Moreau, 2005: A convection scheme for data assimilation: Description and initial tests. *Q. J. R. Meteorol. Soc.*, **131,** 409—436.

Lorenc, A. C., 1986: Analysis methods for numerical weather prediction. *Q. J. R. Meteorol. Soc.*, **112,** 1177—1194.

Lorenc, A. C., 2003a: Modelling of error covariances by 4D-Var data assimilation. *Q. J. R. Meteorol. Soc.*, **129,** 3167 —3182.

Lorenc, A. C., 2003b: The potential of the ensemble Kalman filter for NWP – a comparison with 4D-Var. *Q. J. R. Meteorol. Soc.*, **129,** 3183—3203.

Lorenz, E. N., 1995: Predictability: a problem partly solved. Proc. ECMWF Seminar on "Predictability", ECMWF, Reading, 4—8 September 1995, Vol I, 1—18.

Lorenz E. N. and K. A. Emanuel, 1998: Optimal sites for supplementary weather observations: Simulation with a small model. *J. Atm. Sci.*, **55,** 399—414.

Mahfouf, J.-F., P. Bauer and V. Marécal, 2005: The assimilation of SSM/I and TMI rainfall rates in the ECMWF 4D-Var system. *Q. J. R. Meteorol. Soc.*, **131,** 437—458.

Marécal, V. and J.-F. Mahfouf, 2003: Experiments on 4D-Var assimilation of rainfall data using an incremental formulation. *Q. J. R. Meteorol. Soc.*, **129,** 3137—3160.

Ménard, R. and R. Daley, 1996: The application of Kalman smoother theory to the estimation of 4DVAR error statistics. *Tellus*, **48A,** 221—237.

Moreau, E., P. Lopez, P. Bauer, A. M. Tompkins, M. Janisková and F. Chevallier, 2004: Variational retrieval of temperature and humidity profiles using rain rates versus microwave brightness temperatures. *Q. J. R. Meteorol. Soc.*, **130,** 827—852.

Ott, E. and Coauthors, 2004: A local ensemble Kalman filter for atmospheric data assimilation. *Tellus*, **56A,** 415—428.

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

Rabier, F., H. Järvinen, E. Klinker, J.F. Mahfouf and A. Simmons, 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.

Rauch, H. E., F. Tung and C. T. Striebel, 1965: Maximum Likelihood Estimates of Linear Dynamic Systems, *AIAA Journal*, **3,** 1445—1450.

Radnóti, G., Y. Trémolet, E. Andersson, L. Isaksen E. Hólm and M. Janisková, 2005. Diagnostics of linear and incremental approximations in 4D-Var revisited for higher resolution analysis. *ECMWF Tech. Memo.*, **467,** pp 20.

SAC, 2004: Report of the 33rd session of the Scientific Advisory Committee held at ECMWF 4—6 October 2004. ECMWF/SAC/04/33/R1, www.ecmwf.int/about/committees/sac, pp 21.

Simmons, A. J. and A. Hollingsworth, 2002: Some aspects of the improvement in skill of numerical weather prediction. *Q. J. R. Meteorol. Soc.*, **128,** 647—677.

Simmons, A. J., 2003: Observations, assimilation and the improvement of global weather prediction - Some results from operational forecasting and ERA-40. Proc. ECMWF Seminar on "Recent Developments in Data Assimilation for Atmosphere and Ocean", 8-12 Sept 2003, Reading, U.K. 1—28.

SRNWP, 2004: SRNWP/MetOffice/HIRLAM Workshop on "High resolution data assimilation: towards 1-4 km resolution", 15-17 November 2004, Exeter, UK, pp 35.

Stoffelen, A. and D. Anderson: 1997: Ambiguity removal and assimilation of scatterometer data. *Q. J. R. Meteorol. Soc.*, **123,** 491—518.

Sun, J., 2005: Initialization and numerical forecasting of a super-cell storm observed during STEPS. *Mon. Wea. Rev.*, **133,** 793—813.

Swanson, K., R. Vautard and C. Pires, 1998: Four-Dimensional Variational Assimilation and Predictability in a Quasi-Geostrophic Model. *Tellus*, **50A,** 369—390.

Temperton, C., M. Hortal and A. Simmons, 2001: A two-time-level semi-Lagrangian global spectral model. *Q. J. R. Meteorol. Soc.*, **127,** 111—127.

Thépaut, J.-N. and E. Andersson, 2003: Assimilation of high-resolution satellite data. *ECMWF Newsletter*, **97,** 6—12.

Tompkins, A. M., and M. Janisková, 2004: A cloud scheme for data assimilation purposes: Description and initial tests. *Q. J. R. Meteorol. Soc.*, **130**, 2495-2517

Trémolet, Y., 2003: Model error in variational data assimilation. Proc. ECMWF Seminar on "Recent Developments in Data Assimilation for Atmosphere and Ocean", 8-12 Sept 2003, Reading, U.K, 361—368.

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

Trémolet, Y., 2005: Incremental 4D-Var convergence study. *ECMWF Tech. Memo.*, **469,** pp 34.

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

Vlassis N., B. Terwijn and B. Krose, 2002: Auxiliary particle filter robot localization from high-dimensional sensor observations. Proc. IEEE Int. Conf. on Robotics and Automation, Washington.

Whitaker, J. S. and G. P. Compo, 2002: An ensemble Kalman smoother for reanalysis. Proc. Symp. on Observations, Data Assimilation and Probabilistic Prediction, Orlando, FL, Amer. Meteor. Soc., 144—147.

Whitaker J. S. and T. H. Hamill, 2002: Ensemble data assimilation without perturbed observations. *Mon. Wea. Rev.*, **130**, 1913—1924.

Žagar, N., N. Gustafsson and E. Källén, 2004: Variational data assimilation in the tropics: the impact of a background error constraint. *Q. J. R. Meteorol. Soc.*, **130,** 103-125.

Žagar, N., E. Andersson and M. Fisher, 2005: Balanced tropical data assimilation based on study of equatorial waves in ECMWF short-range forecast errors. *Q. J. R. Meteorol. Soc.*, **131,** Accepted. Also available as *ECMWF Tech. Memo.*, **437,** pp 28.

Zhang, F., C. Snyder and J. Sun, 2004: Impacts of initial estimate and observation availability on convective-scale data assimilation with an ensemble Kalman filter. *Mon. Wea. Rev.*, **132,** 1238—1253.

Zupanski, D., 1997: A gener al weak constraint applicable to operational 4DVar data assimilation systems. *Mon. Wea. Rev.*, **125,** 2274—2291.

A The Particle Filter

The theory behind the particle filter is as follows. Consider a large ensemble of states $\mathbf{x}_k^i$ drawn from a p.d.f. $q\left(\mathbf{x}_k \middle| \mathbf{x}_{k-1}^i, \mathbf{y}_k\right)$, known as the "importance distribution". Here, the superscript *i* denotes the ensemble member and the subscript *k* denotes timestep. The p.d.f., *q*, is conditional on the state at the preceding timestep, and on a set of observations, $\mathbf{y}_k$, but is not necessarily the p.d.f. of one-timestep forecasts from $\mathbf{x}_{k-1}^i$ (although this is the optimal choice). Rather, *q* is a p.d.f. that is easily sampled.

Suppose that at time t_{k-1}, each ensemble member has associated weight, w_{k-1}^i, such that $\sum_i w_{k-1}^i = 1$, and that the discrete p.d.f. $p\left(\mathbf{x}_{k-1} \middle| \mathbf{y}_1 \ldots \mathbf{y}_k\right)$ is approximated by:

$$p\left(\mathbf{x}_{k-1} \middle| \mathbf{y}_1 \ldots \mathbf{y}_k\right) \approx \sum_i w_{k-1}^i \delta\left(\mathbf{x}_{k-1} - \mathbf{x}_{k-1}^i\right) \tag{A.1}$$

It can be shown that with this approximation, the weights associated with the states at time t_k are given by (see, for example, Doucet *et al*., 2001):

$$w_k^i \propto w_{k-1}^i \frac{p\left(\mathbf{y}_k \middle| \mathbf{x}_k^i\right) p\left(\mathbf{x}_k^i \middle| \mathbf{x}_{k-1}^i\right)}{q\left(\mathbf{x}_k^i \middle| \mathbf{x}_{k-1}^i, \mathbf{y}_k\right)} \tag{A.2}$$

The two terms in the numerator of equation (A.2) are readily evaluated. They are the observation-error and model-error probabilities. The denominator is the prescribed importance distribution.

A common, albeit sub-optimal choice for the importance distribution is: $q = p\left(x_k^i \middle| x_{k-1}^i\right)$. With this choice, equation (A.2) takes on a particularly simple form, and the importance distribution is easily sampled by adding random forcing, with the statistical characteristics of model error, during the model integrations from timestep t_{k-1} to t_k.

Equation (A.2) allows the particle filter to be cycled. At each cycle, the ensemble members are integrated forward in time. The importance distribution is then sampled, and the weights associated with each member are updated using equation (A.2). However, this basic particle filter algorithm suffers from a severe problem of "degeneracy", whereby the weights associated with a few members become large while those associated with most of the members become vanishingly small. To avoid this problem, it is necessary periodically to replace the ensemble with an ensemble drawn at random from the discrete distribution given by equation (A.1). In this re-sampling step, members with large associated weight are more likely to be picked (and may be picked several times) whereas those with very small weights are unlikely to be picked. After re-sampling, all members are equally likely (since each is a random draw from the discrete p.d.f.) and the weight associated with each member is reset to 1/*N*, where *N* is the size of the ensemble.