TECHNICAL MEMORANDUM

477

Accounting for an imperfect model in 4D-Var

Yannick Trémolet

Research Department

This is a preprint of:
Y. Trémolet. Accounting for an imperfect model in 4D-Var.
Q. J. R. Meteorol. Soc., 132:2483-2504, 2006

October 2006

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

Abstract

In most operational implementations of 4D-Var, it is assumed that the model used in the data assimilation process is perfect or, at least, that errors in the model can be neglected when compared to other errors in the system. In this paper, we study how model error could be accounted for in four dimensional variational data assimilation.

We present three approaches for the formulation of weak constraint 4D-Var: estimating explicitly a model error forcing term, estimating a representation of model bias or, estimating a four dimensional model state as the control variable. The consequences of these approaches with respect to the implementation and the properties of 4D-Var are discussed.

We show that 4D-Var with an additional model error representation as part of the control variable is essentially an initial value problem and that its characteristics are very similar to that of strong constraint 4D-Var. Taking the four dimensional state as the control variable however leads to very different properties. In that case, weak constraint 4D-Var can be interpreted as a coupling between successive strong constraint assimilation cycles. A possible extension towards long window 4D-Var and possibilities for evolutions of the data assimilation system are presented.

1 Introduction

Data assimilation comprises combining all available sources of information about the atmosphere to produce the best possible forecast. Sources of information are of two types: observations of the atmosphere and the physical laws governing its evolution. ECMWF uses a four dimensional variational data assimilation system as described by Rabier *et al.* (2000), Mahfouf and Rabier (2000), Klinker *et al.* (2000) and Andersson *et al.* (2004). Within each assimilation window, this system assumes that the numerical model representing the evolution of the atmospheric flow is perfect, or at least that model errors can be neglected when compared to other errors in the data assimilation system. The goal of this paper is to study how model error could be accounted for in a four dimensional variational data assimilation system.

Weak constraint 4D-Var theory was introduced by Sasaki (1970). The main underlying idea is that since the model's equations are not exact, it is sufficient to satisfy them only approximately: they can be imposed as a weak constraint in the optimisation problem. Weak constraint 4D-Var has never been implemented fully with a realistic forecast model because of the computational cost and because of the lack of information to define the model error covariance matrix required to solve the problem. However, even with important approximations, in the representation of model error itself, and of the model error covariance matrix, good results have been obtained by several authors such as Derber (1989), Wergen (1992), Zupanski (1993) or Bennett *et al.* (1996) with atmospheric models and Vidard *et al.* (2004) with an ocean model.

We present in this paper several approaches for the implementation of weak constraint 4D-Var. It is organised as follows: in section 2 we recall briefly the theoretical formulation of variational data assimilation and show that 4D-Var and 3D-Var are approximations of the more general weak constraint 4D-Var. Section 3 explains how systematic model error, or model bias, can be taken into account and estimated. Section 4 gives the details of a realistic implementation of weak constraint 4D-Var, explicitly taking model error forcing as part of the control variable. In section 5 we show how weak constraint 4D-Var can be implemented using the four dimensional model state as the control variable. In these sections, we highlight the consequences of these choices with respect to the implementation and the fundamental differences with strong constraint 4D-Var. Finally, in section 6, we discuss and compare some aspects of the various possible implementations of 4D-Var accounting for imperfect model described earlier and outline possible future extensions.

2 Variational data assimilation

2.1 Probabilistic formulation

In a very general sense, the knowledge of a physical system falls into two categories: theoretical knowledge, usually in the form of equations representing the laws of physics, and experimental knowledge consisting of observations of the system. Most physical systems are continuous, but for practical purposes, they are represented by a discrete state variable $\mathbf{x}$ defined in model space. Observations of the system are represented by the vector $\mathbf{y}$ in observation space.

In most applications, theoretical knowledge of the system is such that it can be decomposed into, on one hand equations governing the physical state of the system and, on the other hand, equations relating the state of the system to observations. These two sets of equations can be written as:

$$\begin{aligned} \mathcal{G}(\mathbf{x}) &= 0, \\ \mathcal{H}(\mathbf{x}) &= \mathbf{y}, \end{aligned}$$

where $\mathcal{G}(\mathbf{x})$ represents theoretical knowledge of the system and $\mathcal{H}(\mathbf{x})$ represents knowledge of what the observations should be given the true state of the system. In meteorological applications, $\mathcal{G}$ can include the equations governing the evolution of the flow as well as additional constraints such as balance equations or prior knowledge about the state of the system.

Since both sources of information, theoretical and experimental, are incomplete and subject to errors, it is necessary to resort to a statistical description of the system. A natural way to describe available information in that context is to define the probability distribution for the parameters of the problem. Taking into account the uncertainties, the equations take the form:

$$\begin{aligned} \mathcal{G}(\mathbf{x}) &= \varepsilon_g, \\ \mathcal{H}(\mathbf{x}) - \mathbf{y} &= \varepsilon_o. \end{aligned}$$

$\mathcal{H}(\mathbf{x})$ represents the imperfect knowledge of what the observations should be given the state of the atmosphere, ε_o includes observation errors and uncertainties in $\mathcal{H}$ and $\mathcal{G}(\mathbf{x})$ represents imperfect theoretical knowledge about the atmosphere, it is affected by errors ε_g. Depending on available knowledge of the problem, the image of $\mathcal{G}$ and thus the space is which ε_g is defined may vary.

In order to proceed further, we assume that all errors present in the problem are unbiased (or that the bias has been removed) and can be represented by zero mean Gaussian distributions. These can be defined by the covariance matrices associated to each of the errors defined above. Assuming that errors in the observations and in $\mathcal{G}$ are uncorrelated and using Bayes theorem, it can be shown that, combining the two sources of information, the posterior probability distribution for $\mathbf{x}$ given the observations $\mathbf{y}$ is:

$$P(\mathbf{x}|\mathbf{y}) = \alpha \exp\left(-\frac{1}{2}[\mathcal{H}(\mathbf{x}) - \mathbf{y}]^T \mathbf{R}^{-1} [\mathcal{H}(\mathbf{x}) - \mathbf{y}] - \frac{1}{2}\mathcal{G}(\mathbf{x})^T \mathbf{C}_g^{-1} \mathcal{G}(\mathbf{x})\right)$$

where $\mathbf{C}_g$ is the error covariance matrix for uncertainties in $\mathcal{G}$, $\mathbf{R}$ is the observation error covariance matrix accounting for uncertainties in the observations and in the observation operator and α is a normalisation coefficient.

In variational data assimilation the most probable state is sought, corresponding to the maximum of the posterior distribution. This is done by iteratively minimising $-\log[P(\mathbf{x}|\mathbf{y})]$, which is typically written as a cost function:

$$J(\mathbf{x}) = \frac{1}{2}[\mathcal{H}(\mathbf{x}) - \mathbf{y}]^T \mathbf{R}^{-1} [\mathcal{H}(\mathbf{x}) - \mathbf{y}] + \frac{1}{2}\mathcal{G}(\mathbf{x})^T \mathbf{C}_g^{-1} \mathcal{G}(\mathbf{x}).$$

More details on this result are presented for example by Jazwinski (1970), Lorenc (1986) or Rodgers (2000). One particular source of information which is available in meteorology is a prior estimate of the state of the system. This could be, for instance, a climatological mean for the season. In practice, in operational weather forecasting centres, it is a forecast from the most recent analysis. This represents prior knowledge we have about the state of the system without resorting to the current observations $\mathbf{y}$. The prior estimate of the mean of the state variable is represented by $\mathbf{x}_b$ and called background, with background error covariance matrix $\mathbf{B}$. We will separate it from the other constraints in $\mathcal{G}$ and assume background error is uncorrelated with other errors in the problem. The cost function becomes:

$$\begin{aligned} J(\mathbf{x}) &= \frac{1}{2}(\mathbf{x}-\mathbf{x}_b)^T\mathbf{B}^{-1}(\mathbf{x}-\mathbf{x}_b)+\frac{1}{2}[\mathcal{H}(\mathbf{x})-\mathbf{y}]^T\mathbf{R}^{-1}[\mathcal{H}(\mathbf{x})-\mathbf{y}] \\ &+ \frac{1}{2}\mathcal{F}(\mathbf{x})^T\mathbf{C}_f^{-1}\mathcal{F}(\mathbf{x}) \end{aligned} \tag{1}$$

where $\mathcal{F}$ represents the remaining theoretical knowledge after background information has been accounted for and $\mathbf{C}_f$ is the associated error covariance matrix.

The background term of the cost function J_b is important for many reasons. In meteorological data assimilation, observations are not regularly distributed in time or space, some areas are not observed during the assimilation window. The covariance matrix $\mathbf{B}$ will determine how information is extrapolated from observed areas to unobserved regions and is therefore crucial for the quality of the analysis in areas where few observations are available. In mathematical terms, the problem would be under-determined in those areas without the background term. The second term in the definition of J is called the observation term denoted J_o. This term can be interpreted as a measure of the discrepancy between the observations and their equivalent obtained from the estimated state of the atmosphere $\mathbf{x}$. The cost function is a weighted measure of these discrepancies, weighted by the inverse of the error covariance matrices. This gives data (experimental or theoretical) a weight inversely proportional to the variance of errors affecting them, giving more weight to accurate information and less weight to inaccurate ones.

This short overview of the theoretical basis for variational data assimilation shows that, up to this point, no assumption has been made regarding the space over which $\mathbf{x}$ is defined, in particular whether it is a three or four dimensional space. According to the choice of control vector $\mathbf{x}$ and of the constraint $\mathcal{F}$, the approach presented here will lead to several classes of variational assimilation methods.

2.2 Four dimensional problem

As described very early on by Sasaki (1970), a very general approach is to consider $\mathbf{x}$ as a four dimensional representation of the atmosphere. In that case, the components of $\mathbf{x}$ are the physical variables describing the atmosphere (e.g. temperature, wind, humidity and surface pressure) discretised over the three spatial dimensions of the model's domain and the temporal dimension over the period for which observations are available. The assimilation window $[0,T]$ is discretised into $n+1$ time steps $\{t_i\}_{i=0,\dots,n}$. The state vector $\mathbf{x}_i$ represents the three dimensional state of the atmosphere at time t_i. The observation operator will use the components of the state variable at the appropriate time to evaluate the observation term of the cost function and will make the most accurate use of available observations.

It is also important to include the fact that the atmosphere evolves according to the general laws of physics. The values of $\mathbf{x}$ at successive times are not independent: a relation between successive values of $\mathbf{x}_i$ exists and is represented by the forecast model. As this model is not perfect, we can write:

$$\mathbf{x}_i = \mathcal{M}_i(\mathbf{x}_{i-1}) + \boldsymbol{\eta}_i \tag{2}$$

where $\mathcal{M}_i$ represents the model describing the evolution of the atmospheric flow from time t_{i-1} to time t_i and $\boldsymbol{\eta}_i$ is the three dimensional model error in that time step. The constraint $\mathcal{F}$ can be defined by:

$$\mathcal{F}_i(\mathbf{x}) = \mathbf{x}_i - \mathcal{M}_i(\mathbf{x}_{i-1}).$$

In this formulation, the atmospheric model is only imposed as a weak constraint since the minimising solution $\mathbf{x}$ is not an exact solution of the model. This formulation is known as weak constraint 4D-Var. $\mathbf{C}_f$ is the model error covariance matrix and is usually denoted by $\mathbf{Q}$ and the associated term in the cost function by J_q.

2.3 Reduction of the control variable

At the current operational resolution of the ECMWF data assimilation system, implementing 4D-Var as described above would require using a control variable with a dimension of the order of 10^9. The corresponding model error covariance matrix would have of the order of 10^{18} elements. This is more than the total number of observations of the atmosphere taken since routine meteorological observations started in the 1940s. Since approximately 6×10^6 observations are available each day, it would take 250 million years to gather as many observations as there are parameters in $\mathbf{Q}$. To gather meaningful statistics, it would require orders of magnitude more data, assuming that one could separate model error from other sources of error. There is not enough information available to define this problem without important simplifications.

Assuming a simplifying model can be found for the model error covariance matrix, as is already the case for the background error covariance matrix, the number of parameters defining $\mathbf{Q}$ would be reduced and could be determined.

Approximations are necessary in order to solve the variational data assimilation problem. In most operational variational data assimilation implementations, model error is assumed to be small enough to be neglected compared to initial condition error. The atmospheric model is assumed to be exact and is imposed as a strong constraint:

$$\mathbf{x}_i = \mathcal{M}_i(\mathbf{x}_{i-1}).$$

The evolution of the atmosphere is then entirely determined by the initial condition $\mathbf{x}_0$ and the control variable reduces to a three dimensional state. This reduction of the control variable, combined with the adjoint technique to compute the gradient of the cost function required by most minimisation algorithms, was introduced by Le Dimet and Talagrand (1986) and is usually referred to as strong constraint 4D-Var or simply 4D-Var. The control variable is defined over a three dimensional space but the time dimension of the information brought by the observations and the forecast model is taken into account. 4D-Var should perhaps more appropriately be known as $3\frac{1}{2}$D-Var. The size of the control variable and the elimination of the model error covariance matrix make this algorithm achievable operationally with today's supercomputers.

It is possible to simplify the problem even further and choose the control variable $\mathbf{x}$ as a three dimensional representation of the state of the atmosphere at analysis time. The components of $\mathbf{x}$ are the same physical variables describing the atmosphere discretised over the same spatial domain as in the four dimensional case. The observation operator computes the observation equivalent for that given state, ignoring the fact that all observations do not occur at analysis time. In that case, the forecast model is not used in the data assimilation process. This approach, called three dimensional variational data assimilation, or 3D-Var, is or has been used by several operational centres.

4D-Var	4D-Var$_\beta$ (Section 3)	4D-Var$_\eta$ (Section 4)	4D-Var$_x$ (Section 5)
$\mathbf{x}_0$	$\mathbf{x}_0, \boldsymbol{\beta}$	$\mathbf{x}_0, \boldsymbol{\eta}$	$\mathbf{x}$
$\mathbf{x}_i = \mathcal{M}_i(\mathbf{x}_{i-1})$	$\mathbf{x}_i = \mathcal{M}_{i,0}(\mathbf{x}_0) + \boldsymbol{\beta}_i$	$\mathbf{x}_i = \mathcal{M}_i(\mathbf{x}_{i-1}) + \boldsymbol{\eta}_i$	$\mathbf{x}_i \approx \mathcal{M}_i(\mathbf{x}_{i-1})$
$\Downarrow$	$\Downarrow$	$\Downarrow$	$\Downarrow$
3D Initial Condition	3D I.C. + Model Bias	3D I.C. + Model Error Forcing	4D Atmospheric State

Table 1: Summary of control variable and simplifying model assumptions in four dimensional variational data assimilation. Each formulation is described in detail in the sections indicated in the table.

2.4 Four dimensional control variable

We have shown that the well known 3D-Var and 4D-Var data assimilation algorithms are approximations of the more general weak constraint 4D-Var problem. In the remainder of this paper, we show how the four dimensional nature of the control variable can be retained. Equation (2) defines a change of variable which allows the computation of $\{\mathbf{x}_i\}_{i=0,\ldots,n}$ knowing $\mathbf{x}_0$ and $\boldsymbol{\eta} = \{\boldsymbol{\eta}_i\}_{i=1,\ldots,n}$ and vice versa. This defines two possible choices of control variable that remain four dimensional. Model error can also be defined as the difference between the perfect model trajectory and the state at each time step over the length of the assimilation period for a given initial condition. It is represented by $\boldsymbol{\beta} = \{\boldsymbol{\beta}_i\}_{i=1,\ldots,n}$, satisfying:

$$\mathbf{x}_i = \mathcal{M}_{i,0}(\mathbf{x}_0) + \boldsymbol{\beta}_i \tag{3}$$

where $\mathcal{M}_{i,0}$ represents the forecast model integrated from time t_0 to time t_i. This equation defines another change of variable and a third potential control variable for the weak constraint 4D-Var problem. This list of possible control variables is not exhaustive. For example, non-additive representations of model error could be used, such as multiplicative factors for model tendencies as often used in ensemble forecasting. We denote 4D-Var$_\eta$, 4D-Var$_\beta$ and 4D-Var$_x$ the weak constraint 4D-Var formulations with respectively $(\mathbf{x}_0, \boldsymbol{\eta})$, $(\mathbf{x}_0, \boldsymbol{\beta})$ and $\mathbf{x}$ for control variable. Table 1 summarises these possible choices of control variable and the simplifying assumptions that can be made in variational data assimilation. The meaning and implications of each of these approaches is examined in the following sections.

3 Model bias control variable

3.1 Control variable definition

It is a growing area of research to try to determine and eliminate biases in data assimilation systems, as summarised by Dee (2005). In particular, estimation of observation biases has become an active area of research in recent years with methods like variational observation bias correction introduced by Derber and Wu (1998). Biases are the mathematical expectation of the errors, or ensemble average of the errors. This assumes many realizations of the system are accessible for a proper estimation. In meteorology, this is not possible and the bias is usually assumed ergodic: forecast error statistics are accumulated over long periods of time and the time-averaged error, or systematic error, is used for model bias. This is a good representation of systematic or slowly varying errors. For example, it should represent well errors in the model that vary on a seasonal timescale, like misrepresentation of sea-ice (particularly in the spring and autumn) or errors affecting the stratosphere of polar regions in winter, described by McNally (2003). With this assumption, it is possible to estimate model bias with a weak constraint 4D-Var formulation.

In this section, we choose $\mathbf{x}_0$ and $\boldsymbol{\beta}$, as defined by equation (3), as the control variable. In that case, the constraint associated with model error is:

$$\mathcal{F}_i(\mathbf{x}) = \mathbf{x}_i - \mathcal{M}_{i,0}(\mathbf{x}_0).$$

For the reasons related to the size of the problem already given in section 2.3, it is necessary to make approximations in the representation of $\boldsymbol{\beta}$. In the following, it is assumed constant over the assimilation window. This is a good representation of model bias under the ergodicity assumption. The cost function which defines 4D-Var$_\beta$ is:

$$\begin{aligned} J(\mathbf{x}_0, \boldsymbol{\beta}) &= \frac{1}{2}(\mathbf{x}_0 - \mathbf{x}_b)^T \mathbf{B}^{-1}(\mathbf{x}_0 - \mathbf{x}_b) + \frac{1}{2}\boldsymbol{\beta}^T \mathbf{Q}^{-1} \boldsymbol{\beta} \\ &+ \frac{1}{2}\sum_{i=0}^{n}[\mathcal{H}_i(\mathcal{M}_{i,0}(\mathbf{x}_0) + \boldsymbol{\beta}) - \mathbf{y}_i]^T \mathbf{R}_i^{-1}[\mathcal{H}_i(\mathcal{M}_{i,0}(\mathbf{x}_0) + \boldsymbol{\beta}) - \mathbf{y}_i]. \end{aligned} \tag{4}$$

where $\mathbf{y}_i$ represents the observations at time t_i, $\mathbf{R}_i$ and $\mathcal{H}_i$ are the observation error covariance matrix and observation operator at time t_i and observation error is assumed uncorrelated in time.

3.2 Incremental formulation

At ECMWF and in other operational implementations, the incremental formulation of 4D-Var, introduced by Courtier *et al.* (1994), is used in strong constraint 4D-Var. In this approach, the nonlinear problem is treated as a succession of quadratic problems, approximating the nonlinear problem around a guess $\mathbf{x}^g$. This is achieved by using the initial condition increment $\delta\mathbf{x}_0 = \mathbf{x}_0 - \mathbf{x}_0^g$ as the control variable for the minimisation problem and by linearising the model and observation operator around the guess. This allows the use of very efficient minimisation algorithms such as the preconditioned conjugate gradient method developed by Fisher (1998). The incremental strong constraint 4D-Var cost function is:

$$\begin{aligned} J(\delta\mathbf{x}_0) &= \frac{1}{2}(\delta\mathbf{x}_0 + \mathbf{b})^T \mathbf{B}^{-1}(\delta\mathbf{x}_0 + \mathbf{b}) \\ &+ \frac{1}{2}\sum_{i=0}^{n}(\mathbf{H}_i\mathbf{M}_{i,0}\delta\mathbf{x}_0 + \mathbf{d}_i)^T \mathbf{R}_i^{-1}(\mathbf{H}_i\mathbf{M}_{i,0}\delta\mathbf{x}_0 + \mathbf{d}_i) \end{aligned}$$

where $\mathbf{M}_{i,0}$ represents the tangent linear model integrated from time t_0 to time t_i, i.e. $\mathbf{M}_{i,0} = \mathbf{M}_i\mathbf{M}_{i-1}\ldots\mathbf{M}_1$ and $\mathbf{M}_{0,0} = \mathbf{I}$, $\mathbf{M}_i$ and $\mathbf{H}_i$ are the tangent linear model and observation operator at time t_i, $\mathbf{d}_i = \mathcal{H}_i(\mathbf{x}_i^g) - \mathbf{y}_i$ and $\mathbf{b} = \mathbf{x}_0^g - \mathbf{x}_b$. The gradient of this cost function with respect to the initial condition increment is:

$$\nabla J_0 = \mathbf{B}^{-1}(\delta\mathbf{x}_0 + \mathbf{b}) + \sum_{i=0}^{n}\mathbf{M}_{i,0}^T\mathbf{H}_i^T\mathbf{R}_i^{-1}(\mathbf{H}_i\mathbf{M}_{i,0}\delta\mathbf{x}_0 + \mathbf{d}_i). \tag{5}$$

In practice, the rightmost terms in the equation (5), $\tilde{\mathbf{y}}_i = \mathbf{H}_i\mathbf{M}_{i,0}\delta\mathbf{x}_0 + \mathbf{d}_i$, are computed sequentially in one forward integration of the tangent linear model and stored. The adjoint variable is defined recursively by:

$$\delta\mathbf{x}_i^* = \mathbf{M}_{i+1}^T\delta\mathbf{x}_{i+1}^* + \mathbf{H}_i^T\mathbf{R}_i^{-1}\tilde{\mathbf{y}}_i \;\text{ with }\; \delta\mathbf{x}_n^* = \mathbf{H}_n^T\mathbf{R}_n^{-1}\tilde{\mathbf{y}}_n.$$

Using the linearity of the adjoint model, one shows that $\delta\mathbf{x}_0^*$ is equal to the sum in equation (5). The approximate cost function and its gradient are obtained by one integration of the tangent linear model and one backward integration of the adjoint model.

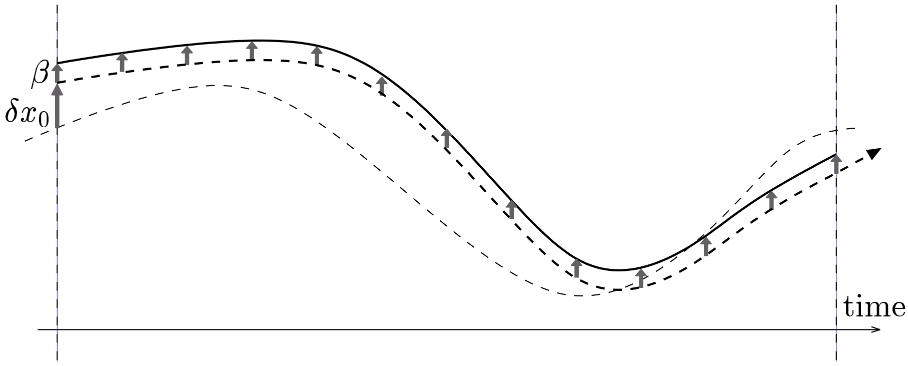

Figure 1: Weak constraint 4D-Var with model bias control variable. The bias $\boldsymbol{\beta}$ is added to the model state (thick dashed line) to provide a bias corrected state (thick solid line) which is then compared to observations. In the schematic example presented here, the bias is constant over the assimilation window.

The incremental 4D-Var formulation can easily be extended to the model bias control variable. The gradient of the quadratic approximation to the cost function with respect to the initial condition increment and to the bias increment become respectively:

$$\nabla J_0 = \mathbf{B}^{-1}(\delta\mathbf{x}_0 + \mathbf{b}) + \sum_{i=0}^{n} \mathbf{M}_{i,0}^T \mathbf{H}_i^T \mathbf{R}_i^{-1}[\mathbf{H}_i(\mathbf{M}_{i,0}\delta\mathbf{x}_0 + \delta\boldsymbol{\beta}) + \mathbf{d}_i]$$

$$\nabla J_\beta = \mathbf{Q}^{-1}(\delta\boldsymbol{\beta} + \boldsymbol{\beta}^g) + \sum_{i=0}^{n} \mathbf{H}_i^T \mathbf{R}_i^{-1}[\mathbf{H}_i(\mathbf{M}_{i,0}\delta\mathbf{x}_0 + \delta\boldsymbol{\beta}) + \mathbf{d}_i]$$

where $\boldsymbol{\beta}^g$ is the guess for the bias, $\delta\boldsymbol{\beta}$ the bias increment and where the nonlinear departures from observations $\mathbf{d}_i$ are computed taking the guess for the bias into account. The gradient with respect to the initial condition is identical to the gradient in strong constraint 4D-Var given by equation (5), except for the fact that, in the forward integration, $\delta\boldsymbol{\beta}$ is added to the evolved increment before entering the observation operator. The adjoint model does not appear in the gradient with respect to the bias: the output of the adjoint of the observation operators is added directly to the gradient of the cost function with respect to $\delta\boldsymbol{\beta}$. The tangent linear and adjoint models and observation operators are unchanged. Because the contributions to the gradient with respect to the bias are intermediate results of the contributions to the gradient with respect to the initial condition, the cost of computing the cost function and its gradient are similar to strong constraint 4D-Var.

Figure 1 gives an overview of the implementation of this formulation of the weak constraint 4D-Var problem. The background is obtained from a forecast from an earlier analysis (thin dashed line). The minimisation algorithm gives new estimates of the initial condition and of the model bias. The model is integrated from the new initial condition over the whole assimilation window (thick dashed line) before this forecast is bias corrected (thick solid line) and compared with observations. The ensuing forecast will be issued from the initial condition produced by the algorithm and bias corrected at post-processing stage if required.

Overall, the characteristics of this optimisation problem are very similar to that of strong constraint 4D-Var. Information is propagated between the initial condition and observations by the tangent linear and adjoint models. Model bias is represented by additional parameters in the optimisation problem without entering directly the model equations. In that respect, it is similar to the case where, for example, observation bias correction parameters are estimated.

4 Model error forcing control variable

4.1 Control variable definition

Equation (2) defines a change of variable which allows the computation of $\mathbf{x}$ given $\mathbf{x}_0$ and $\boldsymbol{\eta}$ and vice versa. In this section, we choose to use the initial condition $\mathbf{x}_0$ and model error forcing $\boldsymbol{\eta}$ as the control variable. This formulation of weak constraint 4D-Var is denoted 4D-Var$_\eta$. There have been several attempts to take model error into account in variational data assimilation and this is the choice most authors have made. As shown in section 2.3, a realistic compromise in the representation of model error is necessary to implement weak constraint 4D-Var.

Derber (1989) introduced the variational continuous assimilation where the control variable is systematic model error forcing rather than the initial condition. In this case, the model error control variable is the same size as the state variable and the size of the problem is unchanged. Zupanski (1993) defined model error as $\boldsymbol{\eta}_i = \lambda_i \boldsymbol{\Phi}$ where $\boldsymbol{\Phi}$ is a three dimensional field and the λ_i are predefined coefficients defining the evolution in time of model error. Zupanski (1997) defined model error as a first order Markov variable in which the random component could be defined on a coarser resolution in time or space. Griffith and Nichols (1998) propose using a spectral representation of model error of the form:

$$\boldsymbol{\eta}_i = \boldsymbol{\gamma}_0 + \boldsymbol{\gamma}_1 \sin(i\Delta t/\tau) + \boldsymbol{\gamma}_2 \cos(i\Delta t/\tau)$$

where $\boldsymbol{\gamma}_0$, $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$ are three dimensional fields, Δt is the model time step and τ is a constant depending on the timescale on which the model error is expected to vary, for example 24h. Errors with longer timescale are represented by the constant term over the assimilation period.

Another relatively simple approach, chosen in the remainder of this section, is to consider model error to be constant by intervals. The length of the interval can vary, from being as long as the assimilation window, which means that model error is constant for the whole assimilation window, to intervals as short as a model time step, which is the full four dimensional problem.

The weak constraint 4D-Var cost function is defined in the most general form by equation (1). It can be written more explicitly as a function of the components of the 4D-Var$_\eta$ control variable:

$$\begin{aligned} J(\mathbf{x}_0, \boldsymbol{\eta}) &= \frac{1}{2}(\mathbf{x}_0 - \mathbf{x}_b)^T \mathbf{B}^{-1} (\mathbf{x}_0 - \mathbf{x}_b) + \frac{1}{2}\sum_{i=1}^{n} \boldsymbol{\eta}_i^T \mathbf{Q}_i^{-1} \boldsymbol{\eta}_i \\ &+ \frac{1}{2}\sum_{i=0}^{n} [\mathcal{H}_i(\mathbf{x}_i) - \mathbf{y}_i]^T \mathbf{R}_i^{-1} [\mathcal{H}_i(\mathbf{x}_i) - \mathbf{y}_i] \end{aligned} \quad (6)$$

where observation errors and model errors are assumed uncorrelated in time and $\mathbf{x}_i = \mathcal{M}_i(\mathbf{x}_{i-1}) + \boldsymbol{\eta}_i$ is the forced model solution.

4.2 Incremental formulation

The incremental 4D-Var formulation can be extended to the model error forcing part of the control variable $\boldsymbol{\eta}$. The control variable for the incremental 4D-Var$_\eta$ is the departure $(\delta\mathbf{x}_0, \delta\boldsymbol{\eta})$ from a first guess $(\mathbf{x}_0^g, \boldsymbol{\eta}^g)$. For small perturbations $\delta\mathbf{x}_{i-1}$ of $\mathbf{x}_{i-1}^g$ and $\delta\boldsymbol{\eta}_i$ of $\boldsymbol{\eta}_i^g$, the model can be linearised and the perturbation evolves according to:

$$\delta\mathbf{x}_i = \mathbf{M}_i \delta\mathbf{x}_{i-1} + \delta\boldsymbol{\eta}_i = \mathbf{M}_{i,0}\delta\mathbf{x}_0 + \sum_{j=1}^{i} \mathbf{M}_{i,j}\delta\boldsymbol{\eta}_j \quad (7)$$

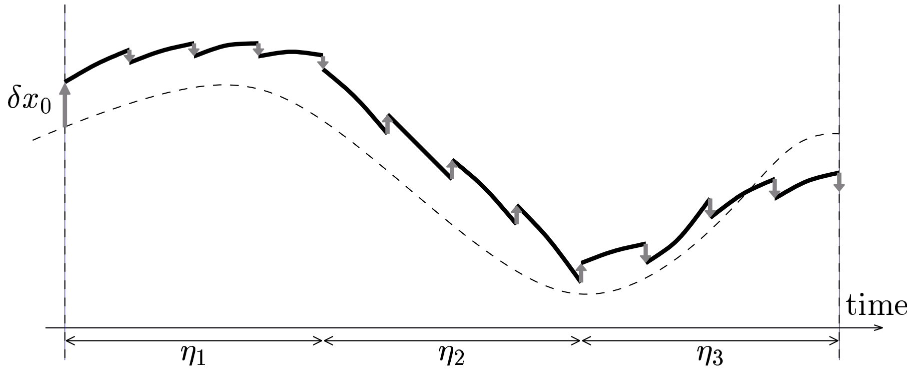

Figure 2: Weak constraint 4D-Var with model error forcing control variable. In the schematic example presented here, the forcing is constant for each of three intervals in the assimilation window (denoted $\boldsymbol{\eta}_1$, $\boldsymbol{\eta}_2$ and $\boldsymbol{\eta}_3$). It is applied at every time step, for illustrative purposes four times in each interval.

where $\mathbf{M}_{i,j}$ represents the tangent linear model integrated from time t_j to time t_i, i.e. $\mathbf{M}_{i,j} = \mathbf{M}_i\mathbf{M}_{i-1}\ldots\mathbf{M}_{j+1}$ and $\mathbf{M}_{i,i} = \mathbf{I}$. The gradients of the quadratic approximation of the cost function with respect to the initial condition increment and to the forcing increment at time t_i are respectively:

$$\nabla J_0 = \mathbf{B}^{-1}(\delta\mathbf{x}_0 + \mathbf{b}) + \sum_{j=0}^{n} \mathbf{M}_{j,0}^T \mathbf{H}_j^T \mathbf{R}_j^{-1}(\mathbf{H}_j\delta\mathbf{x}_j + \mathbf{d}_j)$$

$$\nabla J_i = \mathbf{Q}_i^{-1}(\delta\boldsymbol{\eta}_i + \boldsymbol{\eta}_i^g) + \sum_{j=i}^{n} \mathbf{M}_{j,i}^T \mathbf{H}_j^T \mathbf{R}_j^{-1}(\mathbf{H}_j\delta\mathbf{x}_j + \mathbf{d}_j)$$

where the nonlinear departures from observations $\mathbf{d}_j$ are evaluated from the guess using the forced nonlinear model. The gradient with respect to the initial condition is the same as in strong constraint 4D-Var except for the fact that the forward integration is carried out with the forced linear model defined by equation (7). The observation term gradient with respect to the forcing increment at time t_i has the same form as the gradient with respect to the initial condition increment, with the difference that the contributions to the sum are restricted to the steps $j \geq i$. The gradient with respect to the forcing increment at time t_i results from the backward integration of the adjoint from the end of the window to time t_i. The total gradient of the cost function is still obtained by one forward integration of the tangent linear model and one backward integration of the adjoint, as was the case in strong constraint 4D-Var. This requires only limited modifications to an adjoint model designed for use in a strong constraint 4D-Var system, as already pointed out by Derber (1989).

Figure 2 gives an overview of the incremental implementation of this formulation of the weak constraint 4D-Var problem. The background is obtained from a forecast from an earlier analysis (dashed line). The minimisation gives a new estimate of the initial condition and of the model error forcing. The solution (thick solid line) seems discontinuous when the forcing is applied. The forcing is applied at every time step and should be interpreted as a source term in the equations to correct for errors in each time step. The solution is not more discontinuous in principle than any discrete solution of the model's equations. In practice, the solution may present small amplitude discontinuities if the forcing is not applied at every time step but they should remain small enough not to create spin-up problems. Furthermore, balance constraints could be incorporated in the model error covariance matrix as is already done for the background error covariance matrix.

As with the model bias term, there is a direct propagation of information between the initial condition part of the control variable and all observations through the tangent linear and adjoint models. It is similar in that respect to strong constraint 4D-Var. Like the model bias, the model error forcing is estimated in addition to the initial condition. The main difference from the model bias formulation is that the forcing directly modifies the model state in the forward integration and, as a consequence, the adjoint model is present in the expression of the gradient with respect to the components of the model error forcing.

5 Model state control variable

5.1 Control variable definition

In this section, the four dimensional model state $\mathbf{x}$ is chosen as the control variable. We denote this formulation of weak constraint 4D-Var by 4D-Var$_x$. This definition is closer to the original formulation of the weak constraint 4D-Var problem as described in section 2 but has not been studied as much as the model error forcing formulation and has never been tested in an operational environment. The cost function can be written as a function of the components of the 4D-Var$_x$ control variable:

$$\begin{aligned} J(\mathbf{x}) &= \frac{1}{2}(\mathbf{x}_0-\mathbf{x}_b)^T\mathbf{B}^{-1}(\mathbf{x}_0-\mathbf{x}_b)+\frac{1}{2}\sum_{i=0}^{n}[\mathcal{H}_i(\mathbf{x}_i)-\mathbf{y}_i]^T\mathbf{R}_i^{-1}[\mathcal{H}_i(\mathbf{x}_i)-\mathbf{y}_i] \\ &+ \frac{1}{2}\sum_{i=1}^{n}[\mathbf{x}_i-\mathcal{M}_i(\mathbf{x}_{i-1})]^T\mathbf{Q}_i^{-1}[\mathbf{x}_i-\mathcal{M}_i(\mathbf{x}_{i-1})] \end{aligned} \tag{8}$$

where observation errors and model errors are assumed uncorrelated in time. In this formulation, the model does not appear in the observation term of the cost function but only in the model error constraint term.

As already mentioned, solving the four dimensional problem is not possible without some approximations. We choose to define the control variable as the model state at the start of several intervals in the assimilation window. Within each interval, or sub-window, the forecast model is used to define the state. This is a partial reduction of the control variable where the model is used as a strong constraint between the times when the components of the control variable are defined. We assume that the control variable is defined at m regularly spaced times in addition to the initial time. To fix the notations, the assimilation window is split into regular intervals of p time steps each, starting at steps $\{k_i = i\times p\}_{i=0,\dots,m}$. The cost function becomes:

$$\begin{aligned} J(\mathbf{x}) &= \frac{1}{2}(\mathbf{x}_0-\mathbf{x}_b)^T\mathbf{B}^{-1}(\mathbf{x}_0-\mathbf{x}_b) \qquad (9)\\ &+ \frac{1}{2}\sum_{i=0}^{m}\sum_{j=0}^{p-1}[\mathcal{H}_{k_i+j}(\mathcal{M}_{k_i}^{j}(\mathbf{x}_{k_i}))-\mathbf{y}_{k_i+j}]^T\mathbf{R}_{k_i+j}^{-1}[\mathcal{H}_{k_i+j}(\mathcal{M}_{k_i}^{j}(\mathbf{x}_{k_i}))-\mathbf{y}_{k_i+j}] \\ &+ \frac{1}{2}\sum_{i=1}^{m}[\mathbf{x}_{k_i}-\mathcal{M}_{k_{i-1}}^{p}(\mathbf{x}_{k_{i-1}})]^T\mathbf{Q}_{k_i}^{-1}[\mathbf{x}_{k_i}-\mathcal{M}_{k_{i-1}}^{p}(\mathbf{x}_{k_{i-1}})] \end{aligned}$$

where $\mathcal{M}_i^j = \mathcal{M}_{i+j-1}\circ\cdots\circ\mathcal{M}_i$ represents the model integrated for j steps from time t_i, where $\circ$ denotes the composition of functions and $\mathcal{M}_i^0$ is the identity.

The expression of the cost function is slightly more complicated than with other choices of control variable but can still be computed without difficulties. With one sub-window ($m=0$, $p=n+1$), the control variable reduces to the initial condition and this expression is the strong constraint 4D-Var cost function. When the control variable is defined at every step ($m=n$, $p=1$), this expression is the full weak constraint cost function as defined by equation (8). The choice of the interval length provides some control over the balance between the size of the control variable and the accuracy of the solution. With this approximation, the model is present in the observation term of the cost function, as in strong constraint 4D-Var.

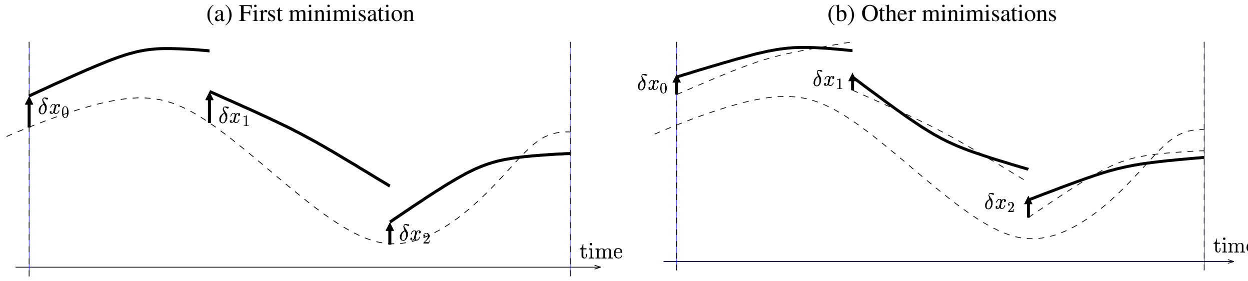

Figure 3: Incremental implementation of weak constraint 4D-Var with model state control variable. The control variable is defined with respect to the trajectory and independent from the state at the end of the previous sub-window.

5.2 Incremental formulation

In the incremental formulation, the control variable is the correction $\delta\mathbf{x}$ to a guess $\mathbf{x}^g$. The gradient of the quadratic approximation of the cost function with respect to the component of the control variable at step k_i is:

$$\begin{aligned}\nabla J_i &= \sum_{j=0}^{p-1}(\mathbf{M}_{k_i}^j)^T\mathbf{H}_{k_i+j}^T\mathbf{R}_{k_i+j}^{-1}(\mathbf{H}_{k_i+j}\mathbf{M}_{k_i}^j\delta\mathbf{x}_{k_i}+\mathbf{d}_{k_i+j})\\ &+ \mathbf{Q}_{k_i}^{-1}(\delta\mathbf{x}_{k_i}-\mathbf{M}_{k_{i-1}}^p\delta\mathbf{x}_{k_{i-1}}+\mathbf{q}_{k_i}^p)\\ &- (\mathbf{M}_{k_i}^p)^T\mathbf{Q}_{k_{i+1}}^{-1}(\delta\mathbf{x}_{k_{i+1}}-\mathbf{M}_{k_i}^p\delta\mathbf{x}_{k_i}+\mathbf{q}_{k_{i+1}}^p)\end{aligned}$$

where $\mathbf{q}_{k_i}^p=\mathbf{x}_{k_i}^g-\mathcal{M}_{k_{i-1}}^p(\mathbf{x}_{k_{i-1}}^g)$ and $\mathbf{M}_i^j$ represents the tangent linear model integrated for j steps from time t_i. This expression is valid for $1\le i\le m-1$. For $i=0$, the second term should be replaced by the gradient of the background term and, for $i=m$, the last term is not present.

In the incremental formulation, the control variable for a given minimisation is the departure from the current guess. In the first minimisation, it is usually the departure from the background forecast as shown in figure 3(a). In the following minimisations, it is the departure from the trajectory resulting from the output of the previous minimisation which is now discontinuous as shown by dashed lines in figure 3(b). It is also the trajectory around which the tangent linear and adjoint models are linearised. This is very different from the case when a forcing term is chosen as the control variable as shown in figure 2. In 4D-Var$_\eta$, a forcing term is added to the current evolved increment, whereas in 4D-Var$_x$, an increment replaces the evolved state and the evolution from that time onwards does not depend on the increment in the previous interval. The forward integrations on the various sub-windows can run independently. The expression of the gradient shows that the adjoint model can be integrated backwards independently on each sub-window. Knowing the trajectory and current estimate of the increment at steps k_i and k_{i+1} allows the computation of the components of J_q and their gradients in parallel. The consequence is that the scalar products and update of the current control vector in the minimisation algorithm are the only point where communication between the various sub-problems is required. This weak constraint formulation brings another possible dimension for parallelising 4D-Var. The tangent linear and adjoint models used in strong constraint 4D-Var can be used without modification.

In the 4D-Var$_x$ setup, weak constraint 4D-Var can be interpreted as solving several strong constraint 4D-Var cycles coupled in one single optimisation problem. Because the tangent linear and adjoint models start from independent states, the tangent linear approximation only needs to be valid within each sub-window which is less restrictive than with strong constraint 4D-Var, or weak constraint 4D-Var with a model error control variable, where it has to be valid over the whole assimilation window. This additional parallelism and relaxed linearity assumption should allow for the use of longer assimilation windows and higher resolution models that include more nonlinear phenomena.

5.3 Hessian properties

The Hessian of the cost function determines some of the properties of the optimisation problem arising from 4D-Var. An approximation of the Hessian of the incremental strong constraint 4D-Var cost function is:

$$J'' = \mathbf{B}^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}$$

where, for this paragraph, $\mathbf{H}$ is the global linearised observation operator and includes the tangent linear model. The change of variable $\chi = \mathbf{B}^{-1/2}\delta\mathbf{x}$ is applied, where $\mathbf{B}^{-1/2}$ is an inverse square root of $\mathbf{B}$ as explained by Fisher (2003). The Hessian of the preconditioned minimisation problem becomes:

$$\hat{J}'' = \mathbf{I} + \mathbf{B}^{T/2}\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}\mathbf{B}^{1/2}. \tag{10}$$

The second term is positive semi-definite, its eigenvalues are non-negative which implies that the eigenvalues of $\hat{J}''$ are all greater than or equal to one. As already stressed in section 2, in meteorology the number of observations is much smaller than the dimension of the state vector, the rank of $\mathbf{R}^{-1}$ and of the second term of the expression above is smaller than the dimension of the identity matrix in the first term. Even if there were more observations, poorly observed areas would likely remain: some components of the control variable are not observed and the rank of the second term remains smaller than the rank of the first term. This illustrates the requirement for the background term as explained in section 2.1, and implies that there are many unit eigenvalues in the spectrum and that the condition number is equal to the largest eigenvalue.

An approximation of the local leading eigenvalue of the strong constraint 4D-Var optimisation problem is given by Andersson *et al.* (2000) as:

$$\lambda_{max} \approx 1 + 2n_{obs}(\sigma_b/\sigma_o)^2$$

where n_{obs} is representative of the number of uncorrelated observations at each grid point and σ_b and σ_o are the background and observation error standard deviations as implied by $\mathbf{B}$ and $\mathbf{R}$. The maximum eigenvalue is obtained for the direction where $n_{obs}(\sigma_b/\sigma_o)^2$ is maximum, i.e. the direction where the most numerous (large n_{obs}) and accurate (small σ_o) observations and large background error (large σ_b) coincide. Currently in the ECMWF data assimilation system, the maximum is reached for the lowest model level over Europe because of the very dense and accurate observation network in that part of the world, as shown for example by Trémolet (2005).

In weak constraint 4D-Var, the approximate Hessian of the cost function is the sum of the Hessian of the observation term of the cost function and of the Hessian of $J_b + J_q$:

$$\begin{pmatrix}
\mathbf{B}^{-1} + \mathbf{M}_1^T\mathbf{Q}_1^{-1}\mathbf{M}_1 & -\mathbf{M}_1^T\mathbf{Q}_1^{-1} & & & 0 \\
-\mathbf{Q}_1^{-1}\mathbf{M}_1 & \mathbf{Q}_1^{-1} + \mathbf{M}_2^T\mathbf{Q}_2^{-1}\mathbf{M}_2 & & & \\
& -\mathbf{Q}_2^{-1}\mathbf{M}_2 & \ddots & -\mathbf{M}_{m-1}^T\mathbf{Q}_{m-1}^{-1} & \\
& & & \mathbf{Q}_{m-1}^{-1} + \mathbf{M}_m^T\mathbf{Q}_m^{-1}\mathbf{M}_m & -\mathbf{M}_m^T\mathbf{Q}_m^{-1} \\
0 & & & -\mathbf{Q}_m^{-1}\mathbf{M}_m & \mathbf{Q}_m^{-1}
\end{pmatrix}$$

This matrix is not block diagonal: the off-diagonal terms provide the propagation of the signal in time, across the sub-windows boundaries. This problem is fully four dimensional, in contrast with the model error formulations which are essentially initial condition problems with additional parameter estimation and where the additional parameters happen to represent model error. This changes the properties of 4D-Var and of the associated optimisation problem.

5.4 Simplified case

Assuming that the model state doesn't evolve too much in one time step, the approximation $\mathbf{M}_i \approx \mathbf{I}$ can be used for illustrative purposes. For this simplified case study, we also assume that the model error covariance matrix is invariant in time and equal to the background error covariance matrix: $\mathbf{B} = \mathbf{Q}_0 = \mathbf{Q}_i$ for $i = 1, \ldots, n$. A change of variable similar to the one used in strong constraint 4D-Var can be applied:

$$\chi = \mathbf{Q}^{-1/2} \delta \mathbf{x}$$

where, for this illustrative case, $\mathbf{Q}$ is the block diagonal matrix whose blocks are $\{\mathbf{Q}_i\}_{i=0,\ldots,n}$. The Hessian of the preconditioned minimisation problem becomes:

$$\hat{J}" = \mathbf{A} + \mathbf{Q}^{T/2} \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H} \mathbf{Q}^{1/2} \tag{11}$$

where $\mathbf{A}$ is the Hessian of the preconditioned $J_b + J_q$ component of the cost function:

$$\mathbf{A} = \begin{pmatrix} 2\mathbf{I} & -\mathbf{I} & & & 0 \\ -\mathbf{I} & 2\mathbf{I} & -\mathbf{I} & & \\ & \ddots & \ddots & \ddots & \\ & & -\mathbf{I} & 2\mathbf{I} & -\mathbf{I} \\ 0 & & & -\mathbf{I} & \mathbf{I} \end{pmatrix}.$$

The fact that this simplified matrix $\mathbf{A}$ is not block diagonal is one of the major differences with strong constraint 4D-Var where it is the identity matrix, as shown by equation (10). Here, it is very similar to the Laplacian operator, acting as a smoother, thus spreading and filtering information. Its eigenvalues are bounded by 0 and 4. Based on the assumptions of Andersson *et al.* (2000), this implies that the largest eigenvalue when the observation term is added is of the order of:

$$\lambda_{max} \approx 4 + 2 n_{obs} / m \, (\sigma_q / \sigma_o)^2$$

where the density of observations is reduced by a factor m compared to the strong constraint case assuming that observations are well distributed in time. This is a result from the fact that only the observations within each sub-window contribute to the observation term of the cost function for that sub-window. The maximum eigenvalue is similar to the maximum eigenvalue of the Hessians of strong constraint 4D-Var for the sub-windows.

In strong constraint 4D-Var, the minimum eigenvalue of the Hessian of the cost function is one. This simplified example shows that it is not true in this weak constraint formulation. The minimum eigenvalue is smaller than one and tends to zero as m increases. In the illustrative example given above, the matrix $\mathbf{A}$ can be formed explicitly and its minimum eigenvalue can be computed with a standard linear algebra package[1]. The results show that it decreases with m as $1/m^2$. The condition number of the full Hessian would thus be of the order of:

$$\kappa \approx 2 m \, n_{obs} \, (\sigma_q / \sigma_o)^2$$

This example shows that the condition number of 4D-Var$_x$ can be higher than the condition number of strong constraint 4D-Var and that it can increase with the number of sub-windows in the assimilation window. It seems reasonable that the number of iterations required to solve the minimisation problem, i.e. the condition number, increases with m since it takes m iterations to propagate information across the whole window through the J_q

[1] The `eigenvectors` method from the python package `numarray` was used here. More information is available at `http://www.stsci.edu/resources/software_hardware/numarray`.

coupling terms. This could create difficulties for the implementation of this formulation of weak constraint 4D-Var, in particular in the case of long windows as proposed by Fisher *et al.* (2005).

The simplified matrix $\mathbf{A}$ is a sparse matrix for which an inverse square root can easily be computed. It can be used to provide a better preconditioner for the minimisation. Because the inverse square root of $\mathbf{A}$ is also a sparse matrix, such a preconditioner is affordable, both in terms of storage and computations. Preconditioning the weak constraint 4D-Var optimisation problem with an inverse square root of $\mathbf{A}$ bears resemblance with preconditioning the strong constraint 4D-Var with an inverse square root of $\mathbf{B}$: it amounts to preconditioning the minimisation using an approximation of the Hessian of the cost function where the observation term has been neglected. Even though the condition number of the weak constraint 4D-Var might seem higher than for strong constraint 4D-Var, it is expected that with appropriate preconditioning, the condition number, and the computational cost that depends on it, could be reduced to similar levels as in strong constraint 4D-Var. Our intention here is only to highlight some differences between strong and weak constraint 4D-Var and preconditioning is an aspect that will require further investigation for an efficient implementation of this formulation of weak constraint 4D-Var.

Without a background term in the cost function, the matrix $\mathbf{A}$ would not be positive definite. This is an illustration in the weak constraint framework of the fact that without a background term the 4D-Var problem is not necessarily well posed.

5.5 Comments

This very simplified study of the Hessian of the weak constraint 4D-Var cost function illustrates some of the differences with strong constraint 4D-Var. However, because it is very simplified, some aspects were not taken into account. For example, because of the similarity of the simplified Hessian with the Laplacian operator, the eigenvectors associated with the lowest eigenvalues should be slowly varying in time and might be well observed. They would project onto eigenvectors of the Hessian of J_o for large eigenvalues, increasing the lowest eigenvalues and reducing the condition number. Other improvements are possible. For example, Jukes (2005) suggested treating weak constraint 4D-Var as an elliptical problem, and using a multi-grid approach to solve it. The ill-conditioned, slowly varying components might be resolved cheaply on the coarser grids while more rapidly varying aspect would be resolved on the finer grids. Another factor which has been ignored here is that, for a given assimilation window, as m increases, the length of the sub-windows and thus σ_q should decrease. This might reduce the condition number, although the exact dependence of σ_q on m is difficult to estimate.

When model error is correlated in time, the Hessian of the cost function is not block tri-diagonal: it becomes a full matrix. The approximation given above is no longer valid. The cost of computing the J_q terms at each iteration might become significant in the overall cost of 4D-Var. It also becomes more difficult to find an approximation of the Hessian of the cost function that can be used for preconditioning without further hypothesis on the form of the correlations. The property that the tangent linear and adjoint models can be integrated independently within each sub-window remains. Computing J_q by blocks of the form $[\mathcal{M}_i(\mathbf{x}_{i-1}) - \mathbf{x}_i]^T \mathbf{Q}_{i,j}^{-1} [\mathcal{M}_j(\mathbf{x}_{j-1}) - \mathbf{x}_j]$ might allow enough parallelism to make the computation affordable. Determining the appropriate correlation statistics would however remain a major difficulty.

An important aspect of this formulation of weak constraint 4D-Var is that a four dimensional atmospheric state is sought, not model error itself. This is closer to the original estimation problem presented in section 2 and does not rely explicitly on any particular form of model error. Assumptions will only be required to define the model error covariance matrix $\mathbf{Q}$. In that respect, it is more general than the other two formulations of weak constraint 4D-Var presented here.

6 Discussion

6.1 Bias and systematic error

The choices of $\mathbf{x}$, $(\mathbf{x}_0, \boldsymbol{\eta})$ or $(\mathbf{x}_0, \boldsymbol{\beta})$ as control variable are equivalent in the sense that they can be used indiscriminately to represent the full four dimensional solution. However, the approximations that $\boldsymbol{\eta}$ is constant in time, or that $\boldsymbol{\beta}$ is constant in time, are not equivalent and they lead to different optimisation problems.

The approach comprising estimation of a representation of model bias in 4D-Var has several interesting properties. The first one is that the model state is not directly modified and the model equations are not disturbed by this component of the control variable, avoiding in particular spin-up and balance issues. The bias is estimated in addition to the initial condition. It can be added to the analysis or to the forecast fields at post-processing stage if bias-corrected fields are required. It should not be added to the analysis before starting the forecast as it would evolve according to the model equations and not be preserved over time as intended in the analysis. This recognises the fact that the model is biased and prefers biased states.

This is different from a formulation where systematic model error forcing is estimated. In that case, a stationary source term is added to continually force the model towards an unbiased state which does not correspond to the natural state of equilibrium of the model. The forcing term is intended to prevent the drift of the model away from the true state. In contrast, the bias term is intended to estimate the discrepancy between the preferred state of the model and the true state without correcting the model.

The model bias determined in the 4D-Var$_\beta$ formulation is not the actual bias of the model with respect to the true atmospheric state; it is the bias of the model with respect to the available information i.e. a weighted mean of all observations. It cannot correct for bias in one set of observations against the model or against other observations but it could complement methods like variational observation bias correction (VarBC) by accounting for the global bias of the system if the sources of bias can be properly separated.

One danger of this approach is that the bias term might absorb the signal brought by observations and it is important to separate the two aspects. The first tool for this is to specify appropriate statistics in the covariance matrix as it filters the bias from the remaining signal. It might also prove necessary to add a background term for the bias component of the control variable to the cost function to allow only for slow variations of the model bias. This seems reasonable since the model is the same every day and systematic error should not vary rapidly from day to day. This would however allow for seasonal variations. The covariance matrix filters the bias in space while a bias background term would filter it in the time dimension.

This should be put in relation with some of the pitfalls of bias correction described by Dee (2005). In comparison to figure 6 of that paper, the present approach would provide an analysis consistent with the model, on a path parallel to observations, as shown on figure 1, but providing an estimate of the distance between model and observations. The issue of finding an unbiased reference state is still present as the estimation of the bias is only relative to observations.

The statistical theory behind variational data assimilation assumes that all errors in the system are unbiased. If the bias as described in the 4D-Var$_\beta$ is a true bias, it could be estimated in addition to the random component of model error. It is, in principle at least, possible to envisage an implementation of weak constraint 4D-Var where a slowly evolving bias, as defined in the 4D-Var$_\beta$ formulation, is estimated in addition to the unbiased four dimensional state of the atmosphere, as defined in the 4D-Var$_x$ formulation of weak constraint 4D-Var.

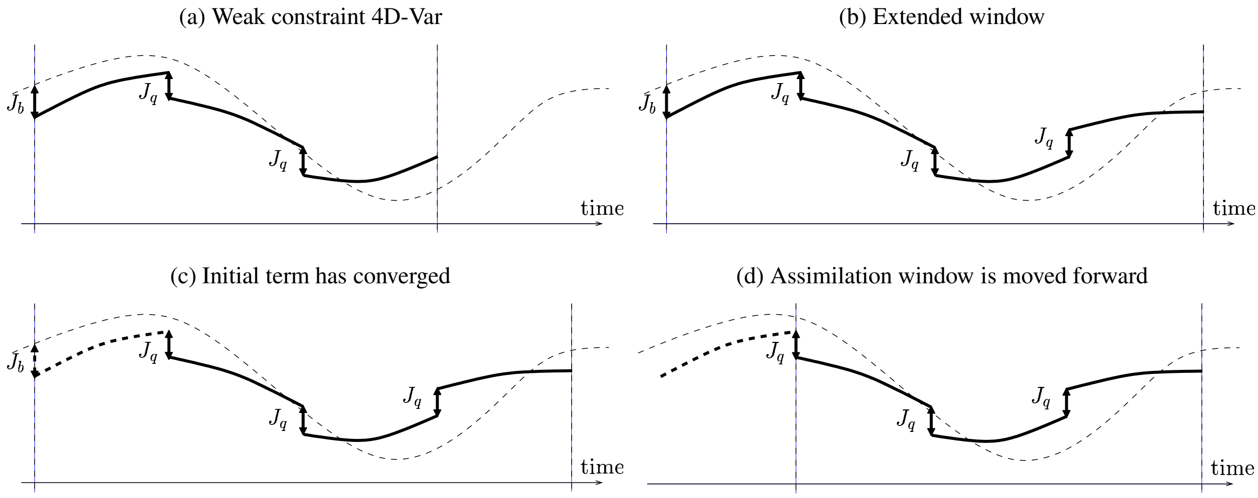

Figure 4: Long window weak constraint 4D-Var implemented with a sliding window. See text for a detailed explanation.

6.2 Long window 4D-Var

As shown by Fisher *et al.* (2005), an additional motivation to develop weak constraint 4D-Var is that when the assimilation window becomes long enough, it is equivalent to a full rank Kalman smoother. Another approach to the implementation of long window weak constraint 4D-Var is presented below.

Consider a weak constraint 4D-Var data assimilation system with model state control variable defined at regular intervals over the assimilation window as described in section 5 (4D-Var$_x$) and represented in figure 4(a). When a given assimilation cycle has converged, instead of the usual cycling implemented currently in 4D-Var, it is possible to extend the assimilation window by one sub-window as shown in figure 4(b). This could be repeated but the assimilation window would soon become too long for the problem to be manageable. However, as already pointed out by Fisher *et al.* (2005), after a certain number of cycles, the solution in the first sub-window is already very accurate and should not evolve any more. There are two reasons for this: the solution for that sub-window has converged because it is already the result of many iterations of the minimisation algorithm and the additional observations at the end of the overall window do not have an impact on the solution far enough in the past because J_q acts as a forgetting factor. Thus, it is possible to consider that the solution over the first sub-window is fixed. The J_b term of the cost function becomes a constant and the first J_q component is computed from a constant $\mathcal{M}_1(\mathbf{x}_0)$. This is shown in figure 4(c) where the dashed lines represent constant values. Since J_b is constant, it can be removed from the cost function in the new minimisation problem. The component of the control variable at initial time can be removed from the minimisation problem. The only required information from the first sub-window is the now fixed final state $\mathcal{M}_1(\mathbf{x}_0)$. In practice, one sub-window has been removed at the beginning of the assimilation period and one added at the end: the assimilation window has shifted by one sub-window period. The J_b term has been replaced by a J_q term computed from the fixed value of the solution at the end of the previous sub-window. This is shown schematically in figure 4(d).

The process can be repeated with an analysis window which progressively moves forward. The J_b term only needs to be present in a warm-up phase, when initiating the cycling. It is then progressively forgotten. As explained in section 2, the theoretical formulation of 4D-Var imposes that the background state is independent of the observations. It comes from an earlier analysis cycle where none of the observations in the current window were used. With assimilation windows of the order of 5 to 10 days, the background state might become inaccurate and affect the quality of the analysis. Here, the initial J_q term ensures that each approximate problem is well posed and the solution consistent in time without being subject to that requirement. It eliminates a potential difficulty from the long-window 4D-Var system as described by Fisher *et al.* (2005). Another consequence is that the simplified Hessian matrix $\mathbf{A}$ introduced in the previous section is always positive definite and might be used as a preconditioner. It is necessary that the assumption made above is verified and that the total window is long enough for the additional observations not to have an impact on the solution before the start of the window, otherwise, the J_b term should be used.

This algorithm is an approximation of weak constraint 4D-Var with an assimilation window that extends for as long as the assimilation has run. This seems particularly interesting in a reanalysis context where one could get a consistent four dimensional estimate of the atmospheric state for very long periods of time and where each analysis has been influenced by future as well as past observations.

As noted above, when the assimilation window becomes long enough, weak constraint 4D-Var becomes equivalent to a full rank Kalman smoother. The method outlined here might provide an efficient algorithm for implementing it.

6.3 Applications

A simple application of weak constraint 4D-Var with 4D state control variable would comprise splitting the assimilation window into a limited number of sub-intervals. The simplest case would be to use the current ECMWF operational 12 hours assimilation window split into two 6 hours sub-windows. Such a system should be better than a 6 hours strong constraint 4D-Var system because it solves two cycles at once, taking into account information from the adjacent sub-window through the J_q coupling. It is also potentially better than a strong constraint 12 hours system because it takes into account the information that the model is not perfect through the extra degrees of freedom available to solve the problem and through the model error covariance matrix. With such a restrictive approximation, this can be interpreted as coupling two cycles of 6 hours strong constraint 4D-Var as much as a weak constraint 4D-Var. This interpretation would justify using a 6 hours background error covariance matrix as an approximation for the model error covariance matrix.

At the other extreme, when the control variable is defined as the model state at every time step, each sub-window reduces to an instantaneous problem: 3D-Var. Weak constraint 4D-Var is then equivalent to a set of 3D-Var problems defined at each time step in the assimilation window and solved together as one coupled optimisation problem. This is of course not possible in practice but it can lead to another approximation of weak constraint 4D-Var. The control variable being defined at regular intervals in the assimilation window, the assimilation problem within each sub-window can be solved as a 3D-Var problem. This is equivalent to neglecting the model in the observation term of the cost function defined by equation (9). The coupling through the J_q terms would propagate the signal and provide consistency in time. This allows an implementation of an approximation of weak constraint 4D-Var where the observations part of the code (J_o computation) and the forecast model (J_q computation) can be kept independent of each other. This reduces greatly the complexity of the code and can be considered as an advantage. It also has the advantage over 3D-Var that the time dimension is taken into account, alleviating the main weakness of 3D-Var. However, and unlike in 3D-Var, the adjoint model is needed to compute the gradient of J_q.

In an incremental implementation where the assimilation window is moved forward when the solution has converged in the initial sub-window, the shifting of the window would be a natural time to re-linearise the problem and start the next outer loop iteration. This would imply that each sub-window sees as many outer loops as there are sub-windows in the overall assimilation window. In the short term, with a small number of sub-windows, this seems reasonable in light of the results obtained by Trémolet (2005) on outer-loop convergence.

Once a weak constraint 4D-Var system is in its stable regime, because the first sub-window of a given analysis cycle has already seen many iterations, it should converge very quickly. Only a few iterations should be needed before it converges and the window can be moved forward again. The concept of cycling becomes less obvious as the assimilation moves forward in time in a quasi-continuous mode if the sub-windows are short enough. In an operational context, the window could be shifted forward as soon as the observations are available, in shorter steps than the current cycling. This could provide almost continuously an up to date analysis and could be very useful for short range forecasting data assimilation applications.

7 Conclusion

Weak constraint 4D-Var is a generalisation of the more widely developed strong constraint 4D-Var where one simplifying assumption, namely the assumption that the forecast model is perfect over the length of the assimilation window, is removed. In addition to lifting a questionable assumption, model error is valuable information which can be used in several ways. It can be added as forcing in the forecast model or at the post-processing stage to correct the forecast. It is hoped that it will also help identify model deficiencies thereby leading to model improvement.

Three alternatives have been presented to remove or alleviate this approximation, using model bias, model error forcing or model state for control variable as summarised in table 1. The main conclusion is that when using model bias (4D-Var$_\beta$) or model error forcing (4D-Var$_\eta$) as the control variable, the problem is very similar in nature to strong constraint 4D-Var: it is essentially an initial condition problem with parameter estimation where the additional parameters represent model error. The information is propagated directly between the initial condition and the observations by the tangent linear and adjoint models. The cost function is four dimensional but the solution over the whole assimilation window depends directly on the initial condition component of the control variable which is only defined over a three dimensional space. On the other hand, when model state is used as the control variable (4D-Var$_x$), information is propagated in time through the model error term of the cost function and filtered by the model error covariance matrix. The model error term acts both as a coupling term propagating information between the sub-windows and as a forgetting factor through the filtering imposed by the covariance matrix and the fact that the model equations are not strictly verified. In that case, the problem is fully four dimensional. Another advantage of this implementation is that the tangent linear and adjoint models are only integrated for the interval between the successive components of the control variable. This has three important consequences: the tangent linear approximation is more accurate, the forward and backward integrations can be run in parallel for each interval and there is no need to choose a particular representation for model error.

The three options described in this paper for taking into account model imperfections in 4D-Var have been developed and are being evaluated at ECMWF, in operational and reanalysis environments. The initial configurations being tested are relatively simple: constant model error forcing, constant model bias and model state control variable with two to four components within an assimilation cycle. These will provide results for a realistic comparison of the algorithms described here and a testbed for evaluating various formulations of the model error covariance matrix, since this remains a major question for a successful implementation of weak constraint 4D-Var. In the near future, a system where the assimilation window could be moved forward by one sub-window could be envisaged, for example with a 12 hours window moved forward by 3 or 6 hours steps and where the problem would be re-linearised when the window is moved forward. At this stage, a background term would be necessary as the window would be too short to ignore the influence of observations at the end of the window on the initial state. Progressively, longer windows and shorter sub-windows could be considered as experience is gained with respect to the weak-constraint algorithms, long windows, outer loop convergence and as more computer power becomes available.

The approach comprising coupling sub-windows to solve a larger data assimilation problem can be generalised. First, it can be extended to spatial sub-domains in addition to temporal ones. This has been experimented in a simple case by Trémolet and Le Dimet (1996). Then, one can consider another generalisation where the model in each sub-window or sub-domain is not the same. A practical application would be coupling of an atmospheric model with an ocean model. As long as the coupling is expressed through a weak constraint term, the only requirement is for synchronisation between iterations of the minimisation algorithm. The tangent linear and adjoint of each model can be integrated independently, without the need for explicit coupling. Such an interface could in principle be added relatively easily to existing atmospheric and ocean assimilation systems. In a similar fashion, this might be applied to couple atmospheric analysis with surface fields analysis, including sea surface temperature. This could improve the consistency between the lowest atmospheric levels and the surface and bring additional information from the atmosphere into the surface fields and vice versa. Finally, another generalisation could be envisaged where the domains overlap but for models of a different nature. Several applications are possible. A global assimilation system could be coupled with a regional system, a weak constraint term being defined as the distance of the two solutions over the whole overlapping domain, not only the boundary condition. Another example would be the coupling of a numerical weather prediction model which an atmospheric chemistry model. Combinations of the above are also possible, at least from an algorithmic point of view. Of course, as with most data assimilation applications, success will depend on the ability to determine appropriate error statistics for the coupling constraints.

Acknowledgements

The author would like to thank Mike Fisher and Dick Dee for their very useful comments on earlier versions of this manuscript. The comments of two anonymous reviewers helped to greatly improve this manuscript.

References

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.

A. Bennett, B. Chua, and L. Leslie. Generalized Inversion of a Global Numerical Weather Prediction Model. *Meteorol. Atmos. Phys.*, 60:165–178, 1996.

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.

D. Dee. Bias and data assimilation. *Q. J. R. Meteorol. Soc.*, 131:3323–3343, 2005.

D. Dee and A. Da Silva. Data assimilation in the presence of forecast bias. *Q. J. R. Meteorol. Soc.*, 124: 269–295, 1998.

J. Derber. A variational continuous assimilation technique. *Mon. Wea. Rev.*, 117:2437–2446, 1989.

J. Derber and W.-S. Wu. The use of TOVS cloud-cleared radiances in the NCEP SSI analysis system. *Mon. Wea. Rev.*, 126:2287–2299, 1998.

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, M. Leutbecher, and G. Kelly. On the equivalence between Kalman smoothing and weak-constraint four-dimensional variational data assimilation. *Q. J. R. Meteorol. Soc.*, 131:3235–3246, 2005.

A. Griffith and N. Nichols. Adjoint methods for treating model error in data assimilation. In *Numerical Methods for Fluid Dynamics*, volume VI of *Institute for CFD, Oxford*, pages 335–344, 1998.

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

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

A. Lorenc. Analysis methods for numerical weather prediction. *Quart. J. R. Meteor. Soc.*, 112:1177–1194, 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.

A. McNally. The assimilation of stratospheric satellite data at ECMWF. In *ECMWF/ SPARC Workshop on modelling and assimilation for the stratosphere and tropopause*, pages 103–106, June 2003.

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.

C. Rodgers. *Inverse Methods for Atmospheric Sounding*. World Scientific, 2000.

Y. Sasaki. Some basic formalisms in numerical variational analysis. *Mon. Wea. Rev.*, 98:875–883, 1970.

Y. Trémolet. Incremental 4D-Var Convergence Study. Tech. Memo. 469, ECMWF, July 2005.

Y. Trémolet and F.-X. Le Dimet. Parallel algorithms for variational data assimilation and coupling models. *Parallel Computing*, 22:657–674, 1996.

P. A. Vidard, A. Piacentini, and F.-X. Le Dimet. Variational data analysis with control of the forecast bias. *Tellus A*, 56:177–188, 2004.

W. Wergen. The effect of model errors in variational assimilation. *Tellus*, 44A:297–313, 1992.

D. Zupanski. A General Weak Constraint Applicable to Operational 4DVAR Data Assimilation Systems. *Mon. Wea. Rev.*, 125:2274–2291, 1997.

M. Zupanski. Regional four-dimensional variational data assimilation in a quasi-operational forecasting environment. *Mon. Wea. Rev.*, 121:2396–2408, 1993.