The University of Reading

Using model reduction methods within incremental 4D-Var

A.S. Lawless, N.K. Nichols
School of Mathematics, Meteorology and Physics, The University of Reading, PO Box 220, Whiteknights, Reading RG6 6AX UK

and

C. Boess, A. Bunse-Gerstner
Zentrum fuer Technomathematik, Universitaet Bremen, Postfach 330440, D-28334 Bremen FRG

NUMERICAL ANALYSIS REPORT 10/06

Department of Mathematics

Using model reduction methods within incremental 4D-Var

A.S. Lawless, N.K. Nichols, C. Boess and A. Bunse-Gerstner

Abstract

Incremental four-dimensional variational assimilation is a method of data assimilation that requires the minimization of a series of simplified cost functions. These simplified functions are usually derived from a spatial or spectral truncation of the full system being approximated. In this paper we propose a new method for deriving these simplified problems, based on model reduction. We show how this method can be combined with incremental 4D-Var to give an assimilation that retains more of the dynamical information of the full system. Numerical experiments are used to illustrate the superior performance to standard truncation methods.

1 Introduction

Data assimilation forms an important component of all numerical weather prediction systems. Since the first discoveries of Lorenz in the 1960s it has been known that small errors in the initial state of a numerical model can lead to large errors in its forecasts [18]. Such an effect is seen in operational weather prediction, where large forecast errors can often be traced to errors in the initial conditions [24]. For this reason much effort has been put into the development of good observational systems and good data assimilation techniques to provide the best use of the observations.

Most early data assimilation techniques consisted of an approximate combination of a model state and a set of observations at a given point in time, considering observations from nearby times to have been made at that time [6, Section 1.6]. A disadvantage with such methods is that they do not use the evolution of the model dynamics as a constraint on the assimilation process. Thus it is not possible to extract information contained in a time series of observations. More recently, advanced assimilation methods have been developed which account for the time dimension of the system. Such methods fall into two catagories - variational methods and Kalman filter methods. In the variational methods, such as four-dimensional variational assimilation (4D-Var), the assimilation treats a set of observations over a given time window in one assimilation step. The problem then reduces to an optimization problem over this time window, where the optimization is constrained by the nonlinear dynamical model [23], [27], [28]. The Kalman filter methods on the other hand perform an assimilation step at each observation time. In these methods information from previous observations is carried forward in time by an explicit update of the background error covariance matrix [5], [12].

In practice approximations must be made to implement these advanced methods for a large numerical weather prediction system. In the case of the Kalman filter, there have been many efforts to develop simplified filters (for example [22], [29], [7]), but as yet there is no operational system using this method. For 4D-Var assimilation, operational implementation was made possible by the introduction of the incremental method [4]. In this method the minimization of the full nonlinear cost function is approximated by the minimization of a series of linearized cost functions, constrained by the linearization of the dynamical model. This linear model is then approximated, which allows a computationally efficient algorithm to be obtained. This method is currently operational in several forecasting centres, for example the European Centre for Medium-range Weather Forecasting, the Met Office and the Meteorological Service of Canada [25], [26], [14]. However, even with the approximations discussed, incremental 4D-Var assimilation is a major contribution to the computational effort required to produce a weather forecast.

A disadvantage with incremental 4D-Var as currently implemented is that the approximations in the linear model are made on the basis of practical considerations, without necessarily taking into account whether the most important parts of the system are being retained. In fact, usually the major simplification is to run the linear model at a lower spatial resolution or spectral truncation than the nonlinear model, where the resolution or truncation is chosen by what can be afforded computationally. With such a method it is difficult to quantify how much information is being lost through the approximation of the model. In this paper we propose a new method for deriving an approximate linear model for use in an incremental 4D-Var system. This method is based on the ideas of model reduction, which has been successfully used to approximate very large dynamical systems in the field of control theory [1], [8]. The advantage of our method is that it produces a lower order version of the original linear model and observation operator, while retaining their most important properties. Such model reduction methods have been applied to data assimilation in the context of the Kalman filter under certain simplifying assumptions [7]. However the method has not previously been used within incremental 4D-Var, where the use of a tangent linear model gives a natural context for model reduction techniques. In this paper we develop the theory of how model reduction may be used within incremental 4D-Var. Preliminary numerical results are then presented to illustrate the potential benefit of this technique. In simple experiments with a shallow water system we show that provided that the low order system is calculated correctly, with proper account taken of the observation operator and the background error covariance matrix, then we may obtain a better solution of the inner loop problem than if low resolution operators were used.

The paper is arranged as follows. In the next section we explain in detail the incremental 4D-Var method and indicate how approximate linear models are used. Section 3 then sets out the theory of model reduction and the particular method of balanced truncation which we use in this paper. In Section 4 we put together the ideas of incremental 4D-Var and reduced order modelling and derive the appropriate inner loop cost function. In Section 5 we present some numerical experiments which compare the reduced order approach with the low resolution approach. The importance of taking into account the background error covariance matrix is also illustrated. Finally in Section 6 we summarise our findings and indicate some of the questions that remain to be answered.

2 Incremental 4D-Var

We present the data assimilation problem in the context of a general nonlinear dynamical system. We write the discrete system equations for the state vectors $\mathbf{x}_i \in \mathbb{R}^n$ at time levels t_i in the form

$$\mathbf{x}_{i+1} = \mathcal{M}_i(\mathbf{x}_i), \tag{1}$$

where $\mathcal{M}_i$ is the nonlinear model operator that propagates the state from time t_i to time t_{i+1} for $i = 0, 1, \ldots, N-1$. We assume that we have imperfect observations $\mathbf{y}_i \in \mathbb{R}^{p_i}$ of the system that are related to the model state at times t_i by

$$\mathbf{y}_i = \mathcal{H}_i(\mathbf{x}_i) + \boldsymbol{\eta}_i, \tag{2}$$

where the operators $\mathcal{H}_i : \mathbb{R}^n \rightarrow \mathbb{R}^{p_i}$ map the system state to observation space. The observation errors $\boldsymbol{\eta}_i$ are assumed to be unbiased, serially uncorrelated, random Gaussian errors with known covariance matrices $\mathbf{R}_i$.

For the data assimilation problem we assume that we have an *a priori* or background estimate $\mathbf{x}^b$ of the expected value of the state $\mathbf{x}_0$ at the initial time t_0 with errors $\boldsymbol{\epsilon}^b$, so that

$$\mathbf{x}_0 - \mathbf{x}^b = \boldsymbol{\epsilon}^b. \tag{3}$$

The background errors $\boldsymbol{\epsilon}^b$ are assumed to be unbiased, Gaussian errors, described by a known covariance matrix $\mathbf{B}_0$. These errors are assumed to be uncorrelated with the observational errors. Then the problem of data assimilation is to find the maximum prior likelihood estimate of the expected value of $\mathbf{x}_0$, which we refer to as the analysis $\mathbf{x}^a$, given all the available information [19].

In a full nonlinear 4D-Var system this problem is solved by directly minimizing the cost function

$$\mathcal{J}[\mathbf{x}_0] = \frac{1}{2}(\mathbf{x}_0 - \mathbf{x}^b)^{\mathrm{T}}\mathbf{B}_0^{-1}(\mathbf{x}_0 - \mathbf{x}^b) + \frac{1}{2}\sum_{i=0}^{N}(\mathcal{H}_i[\mathbf{x}_i] - \mathbf{y}_i)^{\mathrm{T}}\mathbf{R}_i^{-1}(\mathcal{H}_i[\mathbf{x}_i] - \mathbf{y}_i) \tag{4}$$

with respect to $\mathbf{x}_0$, subject to the states $\mathbf{x}_i$ satisfying the discrete nonlinear forecast model (1). The incremental formulation of 4D-Var solves this data assimilation problem by a sequence of minimizations of convex quadratic cost functions linearized around the present estimate of the model state. Recently it has been shown that this procedure is equivalent to applying an inexact Gauss-Newton method to the nonlinear cost function (4), where the convex minimization problems are each solved approximately. If the exact Gauss-Newton method is locally convergent, then the incremental method will also be locally convergent to the solution of (4) provided that each successive minimization is solved to sufficient accuracy [17].

To formulate the incremental 4D-Var algorithm we first write the linearization of the nonlinear system (1) and (2) as

$$\delta\mathbf{x}_{i+1} = \mathbf{M}_i\delta\mathbf{x}_i, \tag{5}$$
$$\mathbf{d}_i = \mathbf{H}_i\delta\mathbf{x}_i, \tag{6}$$

where

$$\mathbf{d}_i = \mathbf{y}_i - \mathcal{H}_i[\mathbf{x}_i] \tag{7}$$

and $\mathbf{M}_i$ and $\mathbf{H}_i$ are the linearizations of the operators $\mathcal{M}_i$ and $\mathcal{H}_i$, respectively, around the state $\mathbf{x}_i$, and are referred to as the tangent linear operators. Then the algorithm is given by the following steps:

- Set first guess $\mathbf{x}_0^{(0)} = \mathbf{x}^b$.
- Repeat for $k = 0, \ldots, K-1$
  - Find linearization states $\mathbf{x}_i^{(k)}$ by integrating the nonlinear model (1) forward from initial state $\mathbf{x}_0^{(k)}$ and find innovations $\mathbf{d}_i^{(k)}$ using (7).
  - Minimize

$$\begin{aligned}\tilde{\mathcal{J}}^{(k)}[\delta\mathbf{x}_0^{(k)}] &= \frac{1}{2}(\delta\mathbf{x}_0^{(k)} - [\mathbf{x}^b - \mathbf{x}_0{}^{(k)}])^{\mathrm{T}}\mathbf{B}_0^{-1}(\delta\mathbf{x}_0^{(k)} - [\mathbf{x}^b - \mathbf{x}_0{}^{(k)}]) \\ &+ \frac{1}{2}\sum_{i=0}^{N}(\mathbf{H}_i\delta\mathbf{x}_i^{(k)} - \mathbf{d}_i^{(k)})^{\mathrm{T}}\mathbf{R}_i^{-1}(\mathbf{H}_i\delta\mathbf{x}_i^{(k)} - \mathbf{d}_i^{(k)})\end{aligned} \tag{8}$$

with respect to $\delta\mathbf{x}_0^{(k)}$, subject to the states $\delta\mathbf{x}_i^{(k)}$ satisfying the discrete tangent linear model (5).

  - Update $\mathbf{x}_0^{(k+1)} = \mathbf{x}_0^{(k)} + \delta\mathbf{x}_0^{(k)}$.

- Set analysis $\mathbf{x}^a = \mathbf{x}_0^{(K)}$.

In practice this algorithm is still computationally too expensive to use in an operational system and so a further simplification is made. We introduce linear restriction operators $\mathbf{U}_i^T \in \mathbb{R}^{r\times n}$ that restrict the model variables $\delta\mathbf{x}_i$ to the space $\mathbb{R}^r$ with $r < n$, and we define variables $\delta\hat{\mathbf{x}}_i \in \mathbb{R}^r$ such that $\delta\hat{\mathbf{x}}_i = \mathbf{U}_i^T\delta\mathbf{x}_i$. We also define prolongation operators $\mathbf{V}_i \in \mathbb{R}^{r\times n}$ that map from the lower dimensional space to the original space. We can then write a restricted version of the linear system (5), (6) in $\mathbb{R}^r$ of the form

$$\begin{aligned} \delta\hat{\mathbf{x}}_{i+1} &= \hat{\mathbf{M}}_i\delta\hat{\mathbf{x}}_i, &(9)\\ \hat{\mathbf{d}}_i &= \hat{\mathbf{H}}_i\delta\hat{\mathbf{x}}_i, &(10)\end{aligned}$$

with

$$\begin{aligned} \hat{\mathbf{M}}_i &= \mathbf{U}_i^T\mathbf{M}_i\mathbf{V}_i, &(11)\\ \hat{\mathbf{H}}_i &= \mathbf{H}_i\mathbf{V}_i. &(12)\end{aligned}$$

The simplified incremental 4D-Var algorithm is then defined such that the inner minimization is performed in the lower dimensional space. We obtain the following algorithm:

- Set first guess $\mathbf{x}_0^{(0)} = \mathbf{x}^b$.

- Repeat for $k = 0, \ldots, K-1$:

  - Find linearization states $\mathbf{x}_i^{(k)}$ by integrating the nonlinear model (1) forward from initial state $\mathbf{x}_0^{(k)}$ and find innovations $\mathbf{d}_i^{(k)}$ using (7).

  - Minimize

$$\begin{aligned} \hat{\mathcal{J}}^{(k)}[\delta\hat{\mathbf{x}}_0^{(k)}] &= \frac{1}{2}(\delta\hat{\mathbf{x}}_0^{(k)} - \mathbf{U}_0^T[\mathbf{x}^b - \mathbf{x}_0{}^{(k)}])^{\mathrm{T}}\hat{\mathbf{B}}_0^{-1}(\delta\hat{\mathbf{x}}_0^{(k)} - \mathbf{U}_0^T[\mathbf{x}^b - \mathbf{x}_0{}^{(k)}])\\ &+ \frac{1}{2}\sum_{i=0}^{N}(\hat{\mathbf{H}}_i\delta\hat{\mathbf{x}}_i^{(k)} - \mathbf{d}_i^{(k)})^{\mathrm{T}}\mathbf{R}_i^{-1}(\hat{\mathbf{H}}_i\delta\hat{\mathbf{x}}_i^{(k)} - \mathbf{d}_i^{(k)}) \qquad (13)\end{aligned}$$

    with respect to $\delta\hat{\mathbf{x}}_0^{(k)}$, subject to the states $\delta\hat{\mathbf{x}}_i^{(k)}$ satisfying the discrete linear model of the reduced space (9). Here the matrix $\hat{\mathbf{B}}_0 = \mathbf{U}_0^T\mathbf{B}_0\mathbf{U}_0$ models the background error statistics in the reduced space.

  - Update $\mathbf{x}_0^{(k+1)} = \mathbf{x}_0^{(k)} + \mathbf{V}_0\delta\hat{\mathbf{x}}_0^{(k)}$.

- Set analysis $\mathbf{x}^a = \mathbf{x}_0^{(K)}$.

We note that the restriction operators $\mathbf{U}_i^T$ are examples of the simplification operators of incremental 4D-Var, as presented in [11], with $\mathbf{V}_i$ being the corresponding generalised inverses. In practice the restriction operators are usually defined as projections to lower spatial resolution for finite-difference models, or as spectral truncations for spectral models. In the remainder of this paper we propose a new method of choosing the restriction operators $\mathbf{U}_i^T$ and the prolongation operators $\mathbf{V}_i$, that takes account of the properties of the underlying dynamical model and assimilation system. First, we introduce the basic theory of model reduction on which our method is based, concentrating on the technique of balanced truncation which we use in this study.

3 Model reduction using balanced truncation

In this section we give a short introduction to model reduction as it is used for linear dynamical systems. The aim is to find a low order model that accurately approximates the output response of the system to the input data over a full frequency range. The response of the system is represented by its Hankel matrix [1]. We focus here on the balanced truncation method [21] for finding the reduced order model. This method ensures that the first singular values of the Hankel matrix of the reduced system exactly match the corresponding singular values of the full system Hankel matrix. A global error bound on the expected error between the frequency responses of the full and reduced systems, based on the neglected Hankel singular values, then exists [1]. The quality of the approximation found by the balanced truncation method is usually very good and the method is therefore appropriate for investigating the potential benefit from using model reduction techniques in data assimilation. Here we describe the method for time-invariant systems, but the method can be extended directly to linear time-varying systems [3].

We consider the discrete-time linear model

$$\begin{aligned} \mathbf{z}_0 &= 0, \\ \mathbf{z}_{i+1} &= \mathbf{M}\mathbf{z}_i + \mathbf{G}\mathbf{B}_0^{\frac{1}{2}}\mathbf{w}_i, \\ \mathbf{d}_i &= \mathbf{H}\mathbf{z}_i \end{aligned} \tag{14}$$

over the time window $[t_0, t_N]$, where $\mathbf{z}_i \in \mathbb{R}^n$ and $\mathbf{d}_i \in \mathbb{R}^p$ are the state and output (observation) vectors at time t_i, respectively, and $\mathbf{w}_i \in \mathbb{R}^n$ are uncorrelated white noise inputs, normally distributed with mean zero and covariance matrix equal the identity. The matrix $\mathbf{B}_0 \in \mathbb{R}^{n\times n}$ represents the covariance of the random inputs $\mathbf{u}_i = \mathbf{B}_0^{\frac{1}{2}}\mathbf{w}_i$, and the matrices $\mathbf{M} \in \mathbb{R}^{n\times n}$, $\mathbf{G} \in \mathbb{R}^{n\times n}$ and $\mathbf{H} \in \mathbb{R}^{p\times n}$ are system matrices describing the dynamics, input and output behaviour of the system. We remark that this is not a unique description of the system. By a change of co-ordinate variables the system can be transformed into an equivalent system represented by different system matrices. The response of the system is not altered by such a transformation.

The aim of the model reduction is to design a model of order $r < n$ of the form

$$\begin{aligned} \hat{\mathbf{z}}_0 &= 0, \\ \hat{\mathbf{z}}_{i+1} &= \hat{\mathbf{M}}\hat{\mathbf{z}}_i + \hat{\mathbf{G}}\mathbf{B}_0^{\frac{1}{2}}\mathbf{w}_i, \\ \hat{\mathbf{d}}_i &= \hat{\mathbf{H}}\hat{\mathbf{z}}_i, \end{aligned} \tag{15}$$

with inputs $\{\mathbf{w}_i\}$, outputs (observations) $\{\hat{\mathbf{d}}_i\}$ and model matrices $\hat{\mathbf{M}} \in \mathbb{R}^{r\times r}$, $\hat{\mathbf{G}} \in \mathbb{R}^{r\times n}$, $\hat{\mathbf{H}} \in \mathbb{R}^{p\times r}$, such that the expected value of the distance between the original observations and the reduced order model observations, written as

$$\lim_{i\to\infty} \mathcal{E}\left\{\|\hat{\mathbf{d}}_i - \mathbf{d}_i\|_2^2\right\} = \lim_{i\to\infty} \mathcal{E}\left\{\left[\hat{\mathbf{d}}_i - \mathbf{d}_i\right]^T \left[\hat{\mathbf{d}}_i - \mathbf{d}_i\right]\right\}, \tag{16}$$

is minimized over all inputs of normalized unit length, with $\lim_{i\to\infty} \mathcal{E}\left\{\frac{1}{n}\|\mathbf{w}_i\|_2^2\right\} = 1$, where $\mathcal{E}\{\cdot\}$ denotes the expected value.

Necessary conditions for such a minimum are established in [2]. It is not practicable to find the optimal reduced model matrices that satisfy these conditions, however, as large systems of nonlinear equations must be solved. Instead the method of balanced truncation is used here, which gives an approximation to the optimal solution. The difference between the optimal output error (16) and the output error of the approximate reduced system is bounded in terms of the Hankel singular values of the full system [1] and the approximate solution is expected to be close to optimal.

In the balanced truncation method the model is directly reduced by removing or 'truncating,' those states that are least influenced by the inputs and those that have least effect on the outputs, that is, those states which are least correlated through the inputs and which are least correlated through the outputs. In general these states do not coincide and it is necessary to transform the co-ordinate variables so that the states to be eliminated are the same in both cases. This is achieved by a 'balancing' transformation.

The balancing transform simultaneously diagonalizes the state covariance matrices $\mathbf{P}$ and $\mathbf{Q}$ associated with the inputs and outputs, respectively. These symmetric positive-definite matrices satisfy the two Stein equations

$$\begin{aligned} \mathbf{P} &= \mathbf{M}\mathbf{P}\mathbf{M}^T + \mathbf{G}\mathbf{B}_0\mathbf{G}^T, &(17)\\ \mathbf{Q} &= \mathbf{M}^T\mathbf{Q}\mathbf{M} + \mathbf{H}^T\mathbf{H}. &(18) \end{aligned}$$

The non-singular balancing transformation $\boldsymbol{\Psi} \in \mathbb{R}^{n\times n}$ is such that $\boldsymbol{\Psi}^{-1}\mathbf{P}\boldsymbol{\Psi}^{-T} = \boldsymbol{\Psi}^T\mathbf{Q}\boldsymbol{\Psi} = \boldsymbol{\Sigma}$ is diagonal and $\boldsymbol{\Psi}^{-1}\mathbf{P}\mathbf{Q}\boldsymbol{\Psi} = \boldsymbol{\Sigma}^2$. We remark that the transformation $\boldsymbol{\Psi}$ is thus given by the matrix of eigenvectors of $\mathbf{PQ}$ and the diagonal of $\boldsymbol{\Sigma}$ contains the Hankel singular values of the full system.

To obtain the reduced order model, the system (14) is first transformed into balanced form and then the last $n-r$ states of the balanced system, corresponding to the smallest singular values of the transformed covariance matrices, are eliminated. The reduced system state $\hat{\mathbf{z}}$ is then defined to be $\hat{\mathbf{z}} = \mathbf{U}^T\mathbf{z}$ and the reduced order system matrices are given by

$$\hat{\mathbf{M}} = \mathbf{U}^T\mathbf{M}\mathbf{V}, \qquad \hat{\mathbf{G}} = \mathbf{U}^T\mathbf{G}, \qquad \hat{\mathbf{H}} = \mathbf{H}\mathbf{V}, \tag{19}$$

where

$$\mathbf{U}^T = [\mathbf{I}_r, \mathbf{0}]\,\boldsymbol{\Psi}^{-1}, \qquad \mathbf{V} = \boldsymbol{\Psi}\begin{bmatrix} \mathbf{I}_r \\ \mathbf{0} \end{bmatrix}. \tag{20}$$

The restriction and prolongation operators $\mathbf{U}^T$ and $\mathbf{V}$ satisfy $\mathbf{U}^T\mathbf{V} = \mathbf{I_r}$ and $\mathbf{V}\mathbf{U}^T$ is a projection operator. Efficient and accurate numerical techniques are available for finding the restriction and prolongation operators in both time-invariant and time-varying systems of moderately large size [10],[15],[3]. For very large systems Krylov subspace methods [8] or approximate balanced truncation (rational interpolation) methods are available [9].

We now explain how these ideas can be used to design restriction and prolongation operators for application in incremental 4D-Var.

4 Combining model reduction with incremental 4D-Var

In order to apply a model reduction method to the inner loop of incremental 4D-Var we have to identify an appropriate dynamical system of the form (14). From Section 2 we see that the inner loop is solved subject to the linear dynamical system given by (5) and (6). The initial perturbation state $\delta\mathbf{x}_0$ is assumed to be normally distributed white noise with mean zero and covariance $\mathbf{B}_0$. Thus there exists a normally distributed white noise $\boldsymbol{\omega} \in \mathbb{R}^n$ with mean zero and covariance identity such that $\delta\mathbf{x}_0 = \mathbf{B}_0^{\frac{1}{2}}\boldsymbol{\omega}$. The dynamical system (5)–(6) that constrains incremental 4D-Var may therefore be written equivalently in the form

$$\begin{aligned} \delta\mathbf{x}_{-1} &= 0, \\ \delta\mathbf{x}_{i+1} &= \mathbf{M}_i\delta\mathbf{x}_i + \mathbf{B}_0^{\frac{1}{2}}\mathbf{w}_i, \\ \mathbf{d}_i &= \mathbf{H}_i\delta\mathbf{x}_i \end{aligned} \tag{21}$$

with white noise inputs $\{\mathbf{w}_i\}$ satisfying

$$\mathbf{w}_i := \begin{cases} \boldsymbol{\omega} \sim \mathcal{N}(0, \mathbf{I}_n), & \text{for } i = -1 \\ 0, & \text{for } i \geq 0. \end{cases} \tag{22}$$

The balanced truncation method may then be applied to the system (21) to obtain restriction and prolongation matrices $\mathbf{U}_i^T$ and $\mathbf{V}_i$ that may be used to reduce the system to the form (9)–(10) for use in a simplified incremental 4D-Var scheme. The error between $\mathbf{d}_i$ and $\hat{\mathbf{d}}_i$, as defined in (16), will then be small for all possible inputs $\{\mathbf{w}_i^{(k)}\}$, and thus the error will also be small for our special input (22). Since here the restriction and prolongation operators are calculated using dynamical information from the full system, we may expect a more accurate solution to the assimilation problem than that obtained from schemes based on other simplifications.

In the time-invariant case, the model and observation matrices $\mathbf{M}_i =: \mathbf{M}$, $\mathbf{H}_i =: \mathbf{H}$, for $i = -1, \ldots, N-1$, are all constant. The restriction and prolongation operators $\mathbf{U}^T$ and $\mathbf{V}$ determined by the balanced truncation procedure are also constant and the reduced order model matrices are given by $\hat{\mathbf{M}} = \mathbf{U}^T\mathbf{M}\mathbf{V}$, $\hat{\mathbf{G}} = \mathbf{U}^T\mathbf{I}_n$, $\hat{\mathbf{H}} = \mathbf{H}\mathbf{V}$. The restricted state variables are defined by $\delta\hat{\mathbf{x}}_i = \mathbf{U}^T\delta\mathbf{x}_i$. The reduced order model (9)–(10) is then given by

$$\begin{aligned} \delta\hat{\mathbf{x}}_{i+1} &= \mathbf{U}^T\mathbf{M}\mathbf{V}\delta\hat{\mathbf{x}}_i, & (23) \\ \hat{\mathbf{d}}_i &= \mathbf{H}\mathbf{V}\delta\hat{\mathbf{x}}_i, & (24) \end{aligned}$$

(where the input is defined by (22)), and on the inner loop of the incremental 4D-Var method we minimize

$$\begin{aligned} \hat{\mathcal{J}}^{(k)}[\delta\hat{\mathbf{x}}_0^{(k)}] &= \frac{1}{2}(\delta\hat{\mathbf{x}}_0^{(k)} - \mathbf{U}^T[\mathbf{x}^b - \mathbf{x}_0{}^{(k)}])^{\mathrm{T}}(\mathbf{U}^T\mathbf{B}_0\mathbf{U})^{-1}(\delta\hat{\mathbf{x}}_0^{(k)} - \mathbf{U}^T[\mathbf{x}^b - \mathbf{x}_0{}^{(k)}]) \\ &+ \frac{1}{2}\sum_{i=0}^{N}(\mathbf{H}\mathbf{V}\delta\hat{\mathbf{x}}_i^{(k)} - \mathbf{d}_i^{(k)})^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{H}\mathbf{V}\delta\hat{\mathbf{x}}_i^{(k)} - \mathbf{d}_i^{(k)}), \end{aligned}$$

subject to the states $\delta\hat{\mathbf{x}}_i^{(k)}$ satisfying the reduced order linear model (23). The prolongation operator $\mathbf{V}$ is then used to lift the solution $\delta\hat{\mathbf{x}}_0^{(k)}$ back into the full space in the outer loop update step.

As derived in [17], the minimization problem is equivalent to the linear least squares problem

$$\left\| \begin{bmatrix} \mathbf{R}^{-\frac{1}{2}}\mathbf{H}\mathbf{V} \\ \mathbf{R}^{-\frac{1}{2}}\mathbf{H}\mathbf{V}\mathbf{U}^T\mathbf{M}\mathbf{V} \\ \vdots \\ \mathbf{R}^{-\frac{1}{2}}\mathbf{H}\mathbf{V}(\mathbf{U}^T\mathbf{M}\mathbf{V})^{N-1} \\ (\mathbf{U}^T\mathbf{B}_0\mathbf{U})^{-\frac{1}{2}} \end{bmatrix} \delta\hat{\mathbf{x}}_0^{(k)} - \begin{bmatrix} \mathbf{R}^{-\frac{1}{2}}\mathbf{d}_0^{(k)} \\ \mathbf{R}^{-\frac{1}{2}}\mathbf{d}_1^{(k)} \\ \vdots \\ \mathbf{R}^{-\frac{1}{2}}\mathbf{d}_{N-1}^{(k)} \\ (\mathbf{U}^T\mathbf{B}_0\mathbf{U})^{-\frac{1}{2}}(\mathbf{x}^b - \mathbf{x}_0^{(k)}) \end{bmatrix} \right\|_2 = \min!, \tag{25}$$

which can be solved numerically by linear algebraic techniques.

In the next sections we investigate the potential benefit of using model reduction in data assimilation for the special case of a time invariant system model. The aim is to determine whether the reduced order method can lead to more efficient assimilation methods than those currently used in practice.

5 Numerical experiments

We now perform some numerical experiments to illustrate the benefit obtained from using low order models within the inner loop of incremental 4D-Var. To do this we set up an inner loop least squares problem of the form (8) with a known solution. An approximate solution is then found by solving a simplified problem of the form (13) and using the prolongation operator to lift the solution back to the full space. The accuracy of the solution found using the standard restriction operator of a spatial interpolator is then compared with that found using the restriction operator derived from the balanced truncation approach. In all cases we solve the linear least squares problem via the QR factorization. We now set out the system we use for the experiments and the details of the experimental design.

5.1 Experimental design

The system we use for this study is the one-dimensional shallow water equations for the flow of a fluid over an obstacle in the absence of rotation. We define the problem on a domain $x \in [0, L]$ and let $\bar{h}(x)$ be the height of the orography, $u(x,t)$ be the velocity of the fluid and $\phi(x,t) = gh(x,t)$ be the geopotential of the fluid, where g is the gravitational constant and $h(x,t)$ is the height of the fluid above the orography. Then the system is described by the equations by

$$\frac{Du}{Dt} + \frac{\partial \phi}{\partial x} = -g\frac{\partial \bar{h}}{\partial x}, \tag{26}$$

$$\frac{D(\ln \phi)}{Dt} + \frac{\partial u}{\partial x} = 0, \tag{27}$$

with

$$\frac{D}{Dt} = \frac{\partial}{\partial t} + u\frac{\partial}{\partial x}. \tag{28}$$

The system is discretized using a semi-implicit semi-Lagrangian integration scheme as described in [16].

In order to apply the balanced truncation method we need the linearization of the discrete nonlinear model in matrix format rather than as an operator. Although we cannot derive an analytical expression for this, we are able to calculate it numerically for a given linearization state. We first find the tangent linear model from the discrete nonlinear model using the standard automatic adjoint techniques, as described in [16]. The matrix of the linear operator can then be calculated from n runs of the tangent linear model using the unit vectors as input, where n is the dimension of the model. To see this we note that if we define $\mathbf{e}_k$ to be k^{th} unit vector, then given any matrix $\mathbf{M}$ the operation $\mathbf{M}\mathbf{e}_k$ picks out the k^{th} column of $\mathbf{M}$. Hence applying the tangent linear model to the unit vectors $\mathbf{e}_1$ to $\mathbf{e}_n$ will give the n columns of the linear model system matrix. Although this method would be impractical for a large system, in practice other model reduction methods would be used in such cases, as was mentioned in Section 3. However the method we use here is sufficient to illustrate the potential benefit of combining model reduction techniques with incremental 4D-Var.

The initial data for the linearization state are taken from Case II of [17]. These data consist of a developing shock solution in the wind and height fields at initial time. The model domain is defined over 200 grid points, separated by a spatial step of 0.01 m. The time step is 0.0092 s and the gravitational constant is set to $g = 10$ ms^{-1}. The remaining model parameters are set as in [17]. For the experiments performed in this study, observations are taken every 10 time steps over a 50 time step window. The matrix $\mathbf{M}$ is therefore obtained by running the tangent linear model for 10 time steps. We assume that this matrix remains constant for successive 10 time step windows, so that effectively each 10 time step window is linearized around the same nonlinear state. This avoids the need to recalculate the low order models for each 10 time step period.

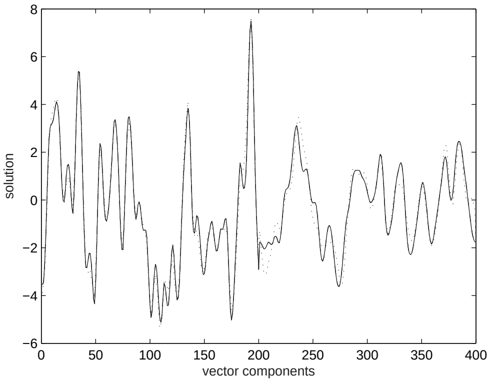

Figure 1: Solution to least squares problem lifted back to full state space. The solid line is the true solution, the dashed line is from the reduced order approach and the dotted line is from the low resolution approach.

We define the true solution of the linear least squares problem to be the difference between the linearization state and this state shifted by 0.5 m. The innovation vectors $\mathbf{d}$ are then the observations for this problem, which are generated from the true solution. Where imperfect observations are used, then Gaussian random noise is added to the true solution, with standard deviations of 0.1 ms^{-1} for the u field and 0.2 m^2s^{-2} for the ϕ field, corresponding to approximately 10% of the mean field values. The observation error covariance matrix $\mathbf{R}$ is then defined as a diagonal matrix of these variances. In order to generate a sensible background error covariance matrix we use the approach of [13] and define the inverse covariance matrix using a second-derivative smoothing operator with a length scale of 0.2 m.

5.2 Comparison of low order and low resolution inner loop

We begin the numerical experiments with a comparison of the low resolution and reduced order approaches using perfect observations. For the low resolution approach the lower spatial resolution is taken to be half that of the full resolution. Hence the low resolution grid has a total of 100 values of u and of ϕ, making the low order system of order 200. In this case the restriction operator is defined by mapping every second grid point of the high resolution grid onto the low resolution grid, while the prolongation operator is defined by a linear interpolation. We compare the solution to the linear least squares problem with that found using the reduced order approach, where the reduced order system is also taken to be of size 200, so that the low resolution and reduced order systems are of the same size. For the experiments of this section observations are taken to be at every second grid point of the full resolution grid, corresponding to every grid point on the low resolution grid.

In Figure 1 we plot the true solution of the least squares problem and the solutions from the low resolution and low order approaches, lifted back into the full order space of 200 grid points. In this plot and all similar plots the first 200 points of the solution vector correspond to values of the perturbation δu and the last 200 points correspond to values of $\delta\phi$. The error in these solutions, calculated as the difference from the true solution, is plotted in Figure 2. We see that for this problem the solution using the reduced order method is more accurate by approximately two orders of magnitude than the standard method of using a low resolution system of the same size.

Rather than considering how much more accurate the low order approach is for a

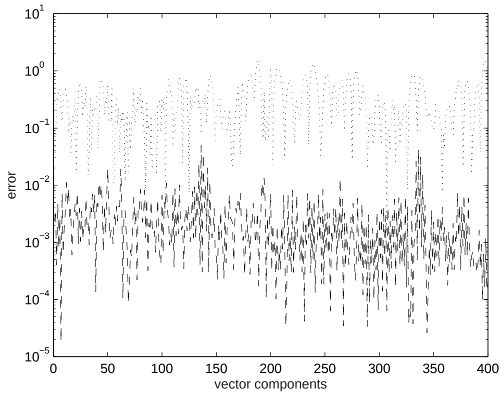

Figure 2: Error in solutions to least squares problem lifted back to full state space for reduced order approach (dashed line) and low resolution approach (dotted line).

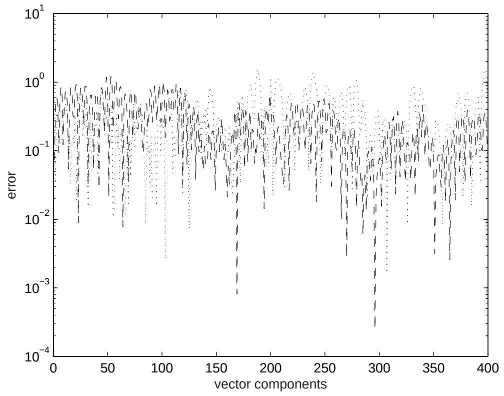

Figure 3: Error in solutions to least squares problem lifted back to full state space for reduced order approach of size 80 (dashed line) and low resolution approach of size 200 (dotted line).

given size of reduced system, we may consider the question of how small we can make the reduced order system and still match the accuracy of the low resolution approach. To test this the least squares problem was solved with low order models of various sizes. In Table 1 the error norms of the solutions from these tests are summarized. We find that even with a reduced order system of size 80 the error norm of the solution is less than that using the low resolution model of size 200. In Figure 3 we plot the error field in the lifted solution from these two experiments. We see that the errors obtained using the low resolution system and the much smaller low order system are of comparable magnitude in all components of the solution vector. Thus for this experiment, using the low order approach allows the use of a much smaller system than the low resolution approach to obtain a given level of accuracy.

In order to test whether the same conclusions hold when the observations contain errors, we add random Gaussian noise to the observations, as described in Section 5.1. We compare the solution of the simplified linear least squares problem using the low resolution approach with that obtained using the low order model of the same size. The errors, calculated as the difference from the exact solution of the problem with these observations, are shown in Figure 4. We see that, as for the case with perfect observations, the model

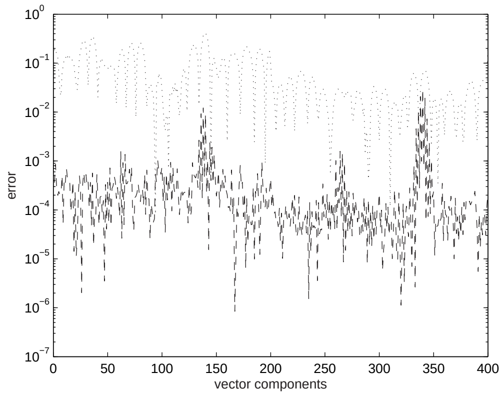

Figure 4: Error in solutions to least squares problem with imperfect observations lifted back to full state space for reduced order approach (dashed line) and low resolution approach (dotted line).

reduction approach gives a more accurate answer by two orders of magnitude. Again we find that if the reduced order model is reduced to size 80, the solution is still as accurate as with the low resolution model of size 200.

In order to understand why the low order approach shows such a benefit when compared with the low resolution approach, we examine the eigenstructure of the low order and low resolution model matrices of size 200. In Figure 5 we compare the eigenvalues of these two matrices with the eigenvalues of the full unapproximated model matrix. We see that the structure of the eigenvalues is approximated much more accurately by the low order matrix than by the low resolution matrix. Hence it appears that the generation of the simplified system by model reduction acts in such a way as to preserve characteristics of the eigenstructure of the original matrix, which is not the case in the low resolution approach. This preservation of eigenstructure allows a solution closer to the original problem to be obtained.

	reduced order	low resolution
l=200	0.0027	0.2110
l=150	0.0134	—
l=100	0.0623	—
l=90	0.1015	—
l=80	0.1726	—
l=70	0.2327	—

Table 1: Comparison of error norms for the low resolution and the reduced order method

5.3 Incorporation of the background covariance in the model reduction procedure

In the derivation of the balanced truncation method of Section 3 we started from a dynamical system with white noise inputs including their covariance. This leads to the incorporation of the background covariance matrix in the Stein equation (17). We now consider how important this is for the model reduction procedure. We repeat the perfect observation experiment of the previous section using a low order system of size 200, but this time the balanced truncation is performed without incorporating the covariance

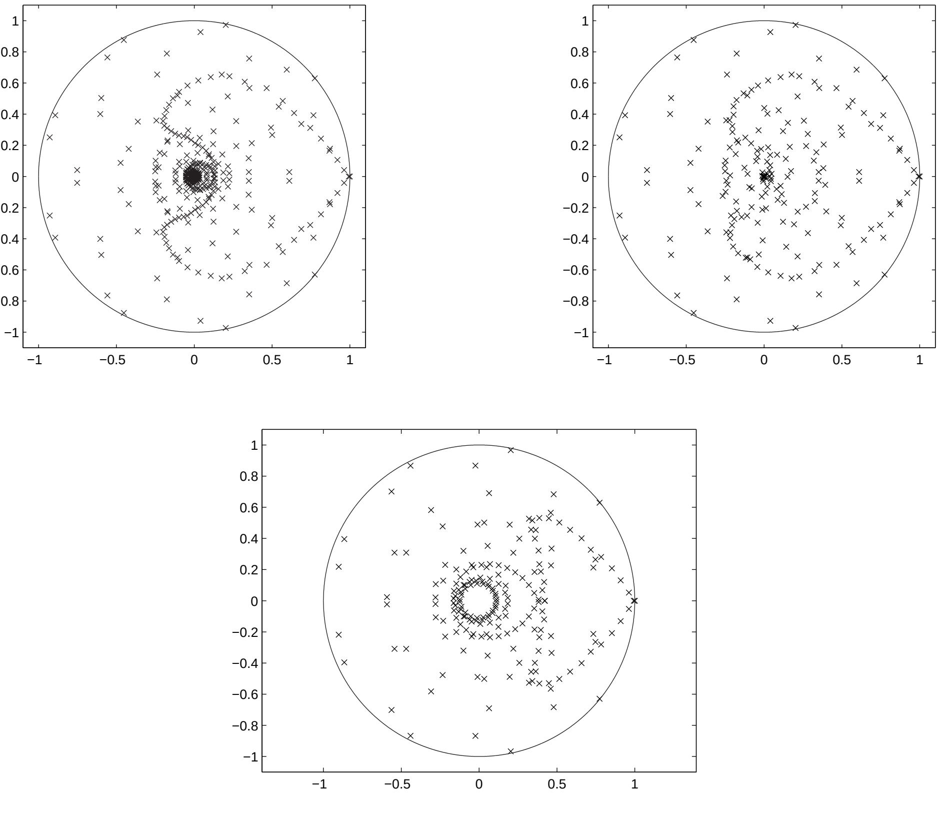

Figure 5: Eigenvalues of full matrix (top left), reduced order matrix (top right) and low resolution matrix (bottom).

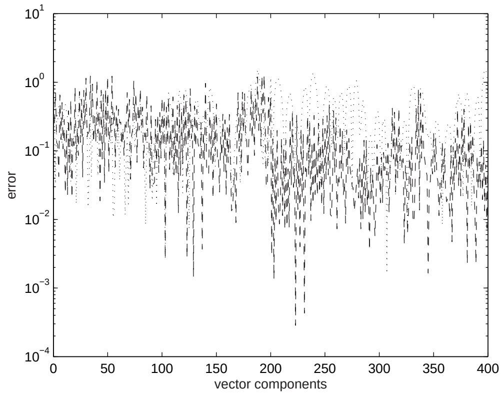

Figure 6: As Figure 2, but without incorporating the covariance matrix into the model reduction procedure.

matrix in the Stein equation, i.e. instead of (17), (18) we solve

$$\begin{aligned}\mathbf{P} &= \mathbf{M}\mathbf{P}\mathbf{M}^T + \mathbf{G}\mathbf{G}^T,\\ \mathbf{Q} &= \mathbf{M}^T\mathbf{Q}\mathbf{M} + \mathbf{H}^T\mathbf{H}.\end{aligned}$$

The error covariance matrix $\mathbf{B}_0$ in the least squares problem remains the same as in Section 5.2; the modification is only in the calculation of the reduced order system.

In Figure 6 we compare the errors in the final solution from this experiment with the errors from the solution using the low resolution approach. We see that now the errors using the two approaches are of the same magnitude. A comparison with Figure 2 shows that not incorporating the covariance $\mathbf{B}_0$ in the balanced truncation procedure has increased the error in the solution from the reduced order method by approximately two orders of magnitude. Thus the numerical results support the theory that it is important to incorporate the covariance information in the reduction process.

5.4 Different observation positions

In the experiments described so far we have assumed that observations of δu and $\delta\phi$ are available at every second grid point. We now test whether the above conclusions continue to hold when the observing network is changed. We first consider a case in which imperfect observations of the δu field are taken at every grid point of the full resolution grid, but no observations of $\delta\phi$ are taken. In Figure 7 we show the errors in the computed solutions of the least squares problem using the low resolution approach and the model reduction approach, where both low order systems are of size 200. Recalling that the first half of the state vector corresponds to values of δu we see that the solution of the δu variable is much more accurate using the reduced order model than using the low resolution model. For the $\delta\phi$ variable, which is not observed, the error in the low order model solution is higher than for the δu variable, but it is still lower than that found using the low resolution approach.

When imperfect observations are taken of $\delta\phi$ only, then the errors are of more similar magnitude for both the δu and $\delta\phi$ variables, but with slightly higher errors in the unobserved δu field. The error plot for this experiment is shown in Figure 8. As for all the previous experiments, the low order approach to solving the problem yields a much more accurate solution than the low resolution approach. Thus we conclude that the results of Section 5.2 remain valid when the observation network is changed.

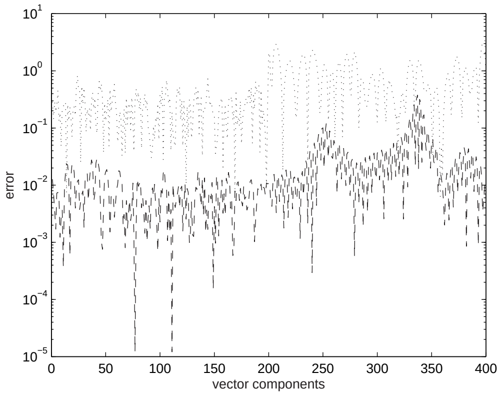

Figure 7: Error in solutions to least squares problem lifted back to full state space with observations of δu only, for reduced order approach (dashed line) and low resolution approach (dotted line).

6 Conclusions

When incremental 4D-Var data assimilation is applied to large-scale systems a simplification of the inner loop problem is usually necessary. In this work we have proposed a new method of simplifying this problem using model reduction ideas from control theory. This approach is designed to approximate the full dynamical system while retaining its essential properties. We have shown how this method naturally fits into the theory of incremental 4D-Var with an alternative definition of the restriction and prolongation operators. In the numerical experiments performed we have demonstrated that the reduced order approach to incremental 4D-Var is more accurate than the low resolution approach for the same size of reduced system. This conclusion has been shown to hold for perfect and noisy observations, and for different observation configurations. However, as expected from the theory, the accuracy depends on the correct inclusion of the covariance information in the model reduction procedure. If care is not taken to include this, then the results may not improve on the reduced resolution approach.

This paper has presented only a preliminary study of combining model reduction and incremental 4D-Var, and many questions remain to be answered before the method can be applied to an operational assimilation system. The model reduction approach of balanced truncation used in this study is not appropriate for such large scale systems and other more appropriate reduction methods need to be investigated. Efficient methods for including the variation of the system in time, as well as between outer loop iterates, also need to be studied in detail. Nevertheless the results from this initial study are encouraging and indicate that reduced order incremental 4D-Var has the potential to give an improvement over existing approaches.

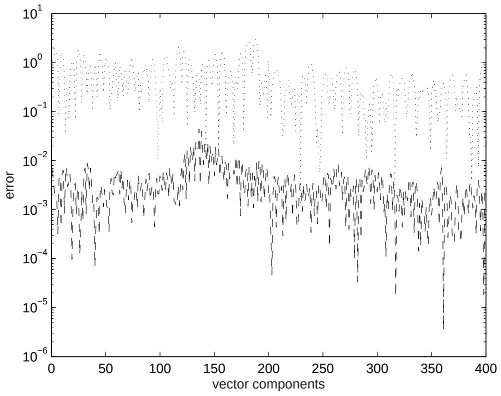

Figure 8: Error in solutions to least squares problem lifted back to full state space with observations of $\delta\phi$ only, for reduced order approach (dashed line) and low resolution approach (dotted line).

Acknowledgements

This research was supported in part by an Academic Research Collaboration Grant from the British Council and the German Academic Exchange Service (DAAD) and by the UK Natural Environment Research Council.

References

[1] Antoulas, A.C., 2005: Approximation of large-scale dynamical systems. SIAM Publisher, Philadelphia.

[2] Bernstein, D.S., Davis, L.D., Hyland, D.C., 1986: The Optimal Projection Equation for Reduced-Order, Discrete-Time Modeling, Estimation, and Control. *J. Guidance, Control, and Dynamics*, 9:288–293.

[3] Chahlaoui, Y., Van Dooren, P., 2005: Model Reduction of Time-Varying Systems. *Dimension Reduction of Large-Scale Systems.* Eds. Benner, P., Mehrmann, V., Sorensen, D., Springer Verlag.

[4] Courtier, P., Thépaut, J-N. and Hollingsworth, A., 1994: A strategy for operational implementation of 4D-Var, using an incremental approach. *Quarterly Journal of the Royal Meteorological Society*, 120:1367–1387.

[5] Cohn, S., 1997: An introduction to estimation theory. *J. Met. Soc. Japan*, 75:257–288.

[6] Daley, R., 1991: Atmospheric data analysis. Cambridge University Press.

[7] Farrell, B.F. and Ioannou, P.J., 2001: State estimation using a reduced-order Kalman filter. *J. Atmos. Sci.*, 58:3666–3680.

[8] Freund, R.W., 2003: Model reduction methods based on Krylov subspaces. *Acta Numerica*, 12:267–319.

[9] Gugercin, S., Sorensen, D.C. and Antoulas, A.C., 2003: A modified low-rank Smith method for large-scale Lyapunov equations. *Numerical Algorithms*, 32:27–55.

[10] Hammarling, S.J., 1982: Numerical solution of the stable, non-negative definite Lyapunov equation. *IMA J. Numerical Analysis*, 2:303–323.

[11] Ide, K., Courtier, C., Ghil, M., and Lorenc, A.C., 1997: Unified notation for data assimilation: Operational, sequential and variational. *J. Met. Soc. Japan*, 1B:181–189.

[12] Jazwinski, A.H., 1970: Stochastic processes and filtering theory. Academic Press.

[13] Johnson, C., Hoskins, B.J. and Nichols, N.K., 2005: A singular vector perspective of 4D-Var: Filtering and interpolation. *Quarterly Journal of the Royal Meteorological Society*, 131:1–20.

[14] Laroche, S., Gauthier, P., Tanguay, M., Pellerin, S., Morneau, J., Koclas, P. and Ek, N., 2005: Evaluation of the operational 4D-Var at the Meteorological Service of Canada. Preprints, *Proceedings of the Fourth WMO International Symposium on Assimilation of Observations in Meteorology and Oceanography*, Prague, WMO, 139.

[15] Laub, A.J., Heath, M.T., Paige, C.C., and Ward, R.C., 1987: , Computation of system balancing transformations and other applications of simultaneous diagnolization algorithms, *IEEE Trans. Automat. Control*, AC-32:115–122.

[16] Lawless, A.S., Nichols, N.K. and Ballard, S.P., 2003: A comparison of two methods for developing the linearization of a shallow-water model. *Quarterly Journal of the Royal Meteorological Society*, 129:1237–1254.

[17] Lawless, A.S., Gratton, S. and Nichols, N.K., 2005: An investigation of incremental 4D-Var using non-tangent linear models. *Quarterly Journal of the Royal Meteorological Society*, 131, 459–476.

[18] Lorenz, E.N., 1963: Deterministic nonperiodic flow. *J. Atmos. Sci*, 20, 130–141.

[19] Lorenc, A.C, 1986: Analysis methods for numerical weather prediction. *Quarterly Journal of the Royal Meteorological Society*, 112, 1177–1194.

[20] Lorenc, A.C, Ballard, S.P., Bell, R.S., Ingleby, N.B., Andrews, P.L.F., Barker, D.M., Bray, J.R., Clayton, A.M., Dalby, T., Li, D., Payne, T.J. and Saunders, F.W., 2000: The Met. Office global 3-dimensional variational data assimilation scheme. *Quarterly Journal of the Royal Meteorological Society*, 126, 2991–3012.

[21] Moore, B.C., 1981: Principal component analysis in linear systems: Controllability, observability and model reduction. *IEEE Trans. Automatic Control*, 26:17-32.

[22] Pham, D.T., Verron, J. and Roubaud, M.C., 1998: A singular evolutive extended Kalman filter for data assimilation in oceanography. *J. Marine Sys.*, 16, 323–340.

[23] Rabier, F. and Courtier, P., 1992: Four-dimensional assimilation in the presence of baroclinic instability. *Quarterly Journal of the Royal Meteorological Society*, 118, 649–672.

[24] Rabier, F., Klinker, E., Courtier, P. and Hollingsworth, A., 1996: Sensitivity of forecast errors to initial conditions. *Quarterly Journal of the Royal Meteorological Society*, 122, 121–150.

[25] Rabier, F., Jarvinen, H., Klinker, E., Mahfouf, J.-F. and Simmons, A., 2000: The ECMWF operational implementation of four-dimensional variational assimilation. I: Experimental results with simplified physics. *Quarterly Journal of the Royal Meteorological Society*, 126, 1143–1170.

[26] Rawlins, R., 2005: Operational implementation of 4D-Var in global model of the Met Office, U.K. Preprints, *Proceedings of the fourth WMO international symposium on assimilation of observations in meteorology and oceanography*, Prague, WMO, 138.

[27] Talagrand. O. and Courtier. P., 1987: Variational assimilation of meteorological observations with the adjoint vorticity equation. I: Theory. *Quarterly Journal of the Royal Meteorological Society*, 113, 1311–1328.

[28] Thépaut, J.N. and Courtier, P., 1991: Four-dimensional variational data assimilation using the adjoint of a multilevel primitive-equation model. *Monthly Weather Review*, 117, 1225–1254.

[29] Verlaan, M. and Heemink, A.W., 2001: Nonlinearity in data assimilation applications: A practical method for analysis. *Monthly Weather Review*, 129, 1578–1589.