Using Model Reduction Methods within Incremental Four-Dimensional Variational Data Assimilation

A. S. Lawless and N. K. Nichols
Department of Mathematics, University of Reading, Reading, United Kingdom

C. Boess and A. Bunse-Gerstner
Zentrum für Technomathematik, Universität Bremen, Bremen, Germany

(Manuscript received 22 November 2006, in final form 9 July 2007)

ABSTRACT

Incremental four-dimensional variational data assimilation is the method of choice in many operational atmosphere and ocean data assimilation systems. It allows the four-dimensional variational data assimilation (4DVAR) to be implemented in a computationally efficient way by replacing the minimization of the full nonlinear 4DVAR cost function with the minimization of a series of simplified cost functions. In practice, these simplified functions are usually derived from a spatial or spectral truncation of the full system being approximated. In this paper, a new method is proposed for deriving the simplified problems in incremental 4DVAR, based on model reduction techniques developed in the field of control theory. It is shown how these techniques can be combined with incremental 4DVAR to give an assimilation method that retains more of the dynamical information of the full system. Numerical experiments using a shallow-water model illustrate the superior performance of model reduction to standard truncation techniques.

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 (Lorenz 1963). Such an effect is seen in operational weather prediction, where large forecast errors can often be traced to errors in the initial conditions (Rabier et al. 1996). 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 (Daley 1991, section 1.6). A disadvantage with such methods is that they do not use the evolution of the model dynamics as an explicit constraint on the assimilation process. Thus it is not possible to extract significant information contained in a time series of observations. More recently, advanced assimilation methods have been developed that account explicitly for the time dimension of the system. Such methods fall into two categories: variational methods and Kalman filter methods. In the variational methods, such as four-dimensional variational data assimilation (4DVAR), 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 (Courtier and Talagrand 1990; LeDimet and Talagrand 1986; Rabier and Courtier 1992; Talagrand and Courtier 1987; Thépaut and Courtier 1991). 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 (Cohn 1997; Jazwinski 1970).

In practice, approximations must be made to imple-
ment 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 (e.g., Cohn and Todling 1996; Verlaan and Heemink 1997; Pham et al. 1998; Farrell and Ioannou 2001; Evensen 2003; Tippett et al. 2003), but as yet there is no operational system using this method. For 4DVAR assimilation, operational implementation was made possible by the introduction of the incremental method (Courtier et al. 1994; Laroche and Gauthier 1998). 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 centers, for example the European Centre for Medium-Range Weather Forecasts, the Met Office, and the Meteorological Service of Canada (Rabier et al. 2000; Rawlins 2005; Laroche et al. 2005). However, even with the approximations discussed, incremental 4DVAR assimilation is a major contribution to the computational effort required to produce a weather forecast.

A disadvantage with incremental 4DVAR 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 data assimilation 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 low-order data assimilation system for use in an incremental 4DVAR procedure. 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 (Antoulas 2005; Freund 2003). 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 (Farrell and Ioannou 2001) and in the context of 4DVAR using a proper orthogonal decomposition method (Cao et al. 2007). However the method has not previously been used within incremental 4DVAR, 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 4DVAR to develop a near-"optimal" reduced-order data assimilation system. 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 and observation error covariances, then we may obtain a better solution of the inner loop problem than if low-resolution operators are used.

The paper is arranged as follows. In the next section we explain in detail the incremental 4DVAR 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 that we use in this paper. In section 4 we put together the ideas of incremental 4DVAR and reduced-order modeling and derive the appropriate inner loop cost function. In section 5 we present some numerical experiments that compare the reduced-order approach with the low-resolution approach. The importance of taking into account the background error covariance matrix is also illustrated. In section 6 we summarize our findings and indicate some of the questions that remain to be answered.

2. Incremental 4DVAR

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

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

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 \( \mathbf{y}_i \in \mathbb{R}^p \) of the system that are related to the model state at times \( t_i \) by

\[
\mathbf{y}_i = \mathcal{H}_i(\mathbf{x}_i) + \mathbf{\eta}_i,
\]

where the operators \( \mathcal{H}_i : \mathbb{R}^n \to \mathbb{R}^p \) map the system state to observation space. The observation errors \( \mathbf{\eta}_i \) 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 \( \mathbf{x}_0^b \) of the expected value of the state \( \mathbf{x}_0 \) at the initial time \( t_0 \) with errors \( \mathbf{e}_0^b \), so that

\[
\mathbf{x}_0 - \mathbf{x}_0^b = \mathbf{e}_0^b.
\]
The background errors \( e^b \) are assumed to be unbiased Gaussian errors, described by a known covariance matrix \( B_0 \). 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 \( x_0 \), which we refer to as the analysis \( x^a \), given all the available information (Lorenc 1986).

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

\[
j(x) = \frac{1}{2} (x - x^b) ^T B_0 ^{-1} (x - x^b) + \frac{1}{2} \sum_{i=0}^{N} [g_i(x_i) - y_i] ^T R_i ^{-1} [g_i(x_i) - y_i],
\]

subject to the states \( x_i \) satisfying the discrete nonlinear forecast model (1). The incremental formulation of 4DVAR 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 (Lawless et al. 2005a). If the exact Gauss–Newton method is locally convergent, the incremental method will also be locally convergent to the solution of (4), provided that each successive minimization is solved to sufficient accuracy (Lawless et al. 2005b; Gratton et al. 2007).

To formulate the incremental 4DVAR algorithm, we first write the linearization of the nonlinear system (1) and (2) about the model state \( x_i \) as

\[
\delta x_{i+1} = M_i \delta x_i \quad \text{and} \quad d_i = H_i \delta x_i,
\]

where

\[
d_i = y_i - g_i(x_i),
\]

and the variable \( \delta x \) denotes a perturbation about the linearization state. The operators \( M_i \) and \( H_i \) are the linearizations of \( M \) and \( H \), respectively, around the state \( x_i \), and are referred to as the tangent linear operators. Then the algorithm is given by the following steps:

1) Set first guess \( x_0^{(0)} = x^b \).

2) Repeat for \( k = 0, \ldots, K - 1 \):

1) Find linearization states \( x_i^{(k)} \) by integrating the nonlinear model (1) forward from initial state \( x_i^{(k)} \) and find innovations \( d_i^{(k)} \) using (7).

2) Minimize

\[
j^k[\delta x_0^{(k)}] = \frac{1}{2} \{ \delta x_0^{(k)} - [x^b - x_0^{(k)}] ^T B_0 ^{-1} \times [\delta x_0^{(k)} - [x^b - x_0^{(k)}]]
\]

\[
+ \frac{1}{2} \sum_{i=0}^{N} [H_i \delta x_i^{(k)} - d_i^{(k)}] ^T R_i ^{-1} [H_i \delta x_i^{(k)} - d_i^{(k)}]
\]

subject to \( \delta x_0^{(k)} \), subject to the states \( \delta x_i^{(k)} \) satisfying the discrete tangent linear model (5).

3) Update \( x_0^{(k+1)} = x_0^{(k)} + \hat{d}_k^{(k)} \).

3) Set analysis \( x^a = x_0^{(K)} \).

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 \( U_i ^T \in \mathbb{R}^{r \times n} \) that restrict the model variables \( \delta x_i \) to the space \( \mathbb{R}^r \) with \( r < n \), and we define variables \( \hat{\delta} x_i \in \mathbb{R}^r \) such that \( \delta x_i = U_i ^T \hat{\delta} x_i \). We also define prolongation operators \( V_i \in \mathbb{R}^{n \times r} \) that map from the lower-dimensional space to the original space with the properties that \( U_i ^T V_i = I \), and \( V_i U_i ^T \) forms a projection operator. We can then write a restricted version of the linear system (5)–(6) in \( \mathbb{R}^r \) of the form

\[
\hat{\delta} x_{i+1} = \hat{M}_i \hat{\delta} x_i \quad \text{and} \quad \hat{d}_i = \hat{H}_i \hat{\delta} x_i,
\]

where \( \hat{M}_i \in \mathbb{R}^{r \times r} \) and \( \hat{H}_i \in \mathbb{R}^{n \times r} \) are simplified dynamic and observation operators such that \( V_i \hat{M}_i U_i ^T \) and \( V_i \hat{H}_i U_i ^T \) approximate the full dynamic and observation tangent linear operators \( M_i \) and \( H_i \).

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

1) Set first guess \( x_0^{(0)} = x^b \).

2) Repeat for \( k = 0, \ldots, K - 1 \):

1) Find linearization states \( x_i^{(k)} \) by integrating the nonlinear model (1) forward from initial state \( x_i^{(k)} \) and find innovations \( d_i^{(k)} \) using (7).

2) Minimize

\[
j^k[\delta x_0^{(k)}] = \frac{1}{2} \{ \delta x_0^{(k)} - [x^b - x_0^{(k)}] ^T B_0 ^{-1} \times [\delta x_0^{(k)} - [x^b - x_0^{(k)}]]
\]

\[
+ \frac{1}{2} \sum_{i=0}^{N} [H_i \delta x_i^{(k)} - d_i^{(k)}] ^T R_i ^{-1} [H_i \delta x_i^{(k)} - d_i^{(k)}]
\]
with respect to \( \delta \mathbf{x}_i^{(k)} \), subject to the states \( \delta \mathbf{x}_i^{(k)} \) satisfying the discrete linear model of the reduced space (9). Here the matrix \( \tilde{\mathbf{B}}_0 \) models the background error statistics in the reduced space.

3) Update \( \mathbf{x}_i^{(k+1)} = \mathbf{x}_i^{(k)} + \mathbf{V}_0 \delta \mathbf{x}_i^{(k)} \).

3) Set analysis \( \mathbf{x}^* = \mathbf{x}_0^{(k)} \).

We note that the restriction operators \( \mathbf{U}_i^T \) are examples of the simplification operators of incremental 4DVAR, as presented in Ide et al. (1997), with \( \mathbf{V}_i \) being the corresponding generalized inverses. In practice the simplified dynamics and observation operators are usually defined as low-resolution spatial operators 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 simplified system equations. The reduced-order dynamics and observation operators are taken to be exact projections of the full-order tangent linear operators and are thus given by \( \tilde{\mathbf{M}}_i = \mathbf{U}_i^T \mathbf{M} \mathbf{V}_i \) and \( \tilde{\mathbf{H}}_i = \mathbf{H} \mathbf{V}_i \), where the restriction operators \( \mathbf{U}_i^T \) and the prolongation operators \( \mathbf{V}_i \) are determined by “optimal” model reduction techniques that take 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 that 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 (Antoulas 2005). We focus here on the balanced truncation method (Moore 1981) 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 (Antoulas 2005). 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 (Chahlaoui and Van Dooren 2005).

We consider the discrete-time linear model

\[
\mathbf{z}_0 = 0,
\]

\[
\mathbf{z}_{i+1} = \mathbf{M}\mathbf{z}_i + \mathbf{G}\mathbf{u}_i,
\]

and

\[
\mathbf{d}_i = \mathbf{H}\mathbf{z}_i.
\]

over the time window \([t_0, t_N]\), where \( \mathbf{z}_i \in \mathbb{R}^p \) and \( \mathbf{d}_i \in \mathbb{R}^p \) are the state and output (observation) vectors at time \( t_i \), respectively, and \( \mathbf{u}_i \in \mathbb{R}^n \) are correlated random noise inputs, normally distributed with mean zero and covariance given by matrix \( \mathbf{B}_0 \in \mathbb{R}^{n \times n} \). The inputs may be written in the form \( \mathbf{u}_i = \mathbf{B}_0^{1/2}\mathbf{w}_i \), where \( \mathbf{w}_i \in \mathbb{R}^n \) are uncorrelated white noise inputs, normally distributed with mean zero and covariance matrix \( \mathbf{B}_0 \) equal to the identity. The matrices \( \mathbf{M} \in \mathbb{R}^{n \times r}, \mathbf{G} \in \mathbb{R}^{r \times n} \) and \( \mathbf{H} \in \mathbb{R}^{p \times r} \) are system matrices describing the dynamics, input and output behavior of the system. We remark that this is not a unique description of the system. By a change of coordinate 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

\[
\hat{\mathbf{z}}_0 = 0,
\]

\[
\hat{\mathbf{z}}_{i+1} = \hat{\mathbf{M}}\hat{\mathbf{z}}_i + \hat{\mathbf{G}}\hat{\mathbf{u}}_i = \hat{\mathbf{M}}\hat{\mathbf{z}}_i + \hat{\mathbf{G}}\mathbf{B}_0^{1/2}\mathbf{w}_i,
\]

and

\[
\hat{\mathbf{d}}_i = \hat{\mathbf{H}}\hat{\mathbf{z}}_i.
\]

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

\[
\lim_{i \to \infty} \mathbb{E}[(\hat{\mathbf{d}}_i - \mathbf{d}_i)^T \mathbf{R}^{-1}(\hat{\mathbf{d}}_i - \mathbf{d}_i)],
\]

is minimized over all inputs of normalized unit length, with \( \lim_{i \to \infty} \mathbb{E}[(1/n)\|\mathbf{w}_i\|_2^2] = 1 \), where \( \mathbb{E}[\cdot] \) denotes the expected value. The weighting matrix \( \mathbf{R} \) is taken to be symmetric positive definite.

Necessary conditions for such a minimum are established in Bernstein et al. (1986). 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 weighted difference between the optimal output error (14) and the output error of the approximate reduced system is bounded in terms of the Hankel singular values of the full system (Antoulas 2005) 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 that are least correlated through the inputs and are least correlated through the outputs. In general these states do not coincide and it is necessary to transform the coordinate 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 \( P \) and \( Q \) associated with the inputs and outputs, respectively. These symmetric positive-definite matrices satisfy the two Stein equations:

\[
P = MPM^T + GB_0G^T \quad \text{and} \quad Q = M^TQM + H^TR^{-1}H.
\]

(15) (16)

The nonsingular balancing transformation \( \Psi \in \mathbb{R}^{n \times n} \) is defined such that \( \Psi^{-1}P\Psi^{-T} = \Psi^TQ\Psi = \Sigma \) is diagonal and \( \Psi^{-1}PQ\Psi = \Sigma^2 \). We see that the transformation \( \Psi \) is thus given by the matrix of eigenvectors of \( PQ \) and that the diagonal of \( \Sigma \) then contains the Hankel singular values of the full system.

To obtain the reduced-order model, the system (12) 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 \( \hat{z} \) is then defined to be \( \hat{z} = U^Tz \) and the reduced-order system matrices are given by

\[
\hat{M} = U^TMV, \quad \hat{G} = U^TG, \quad \hat{H} = HV, \quad \text{where}
\]

(17) (18)

The restriction and prolongation operators \( U^T \) and \( V \) satisfy \( U^TV = I_r \), and \( VU^T \) is a projection operator.

Efficient and accurate numerical techniques are available for finding the restriction and prolongation operators by balanced truncation in both time-invariant and time-varying systems of moderately large size (Hammarling 1982; Laub et al. 1987; Chahlaoui and Van Dooren 2005). For very large systems, Krylov subspace (Lanczos or Arnoldi) methods (Freund 2003) or approximate balanced truncation (rational interpolation) methods are available (Gugercin et al. 2003). Recently effective new techniques based on proper orthogonal decomposition (POD) methods have been derived for determining balanced reductions (Willcox and Peraire 2002).

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

4. Combining model reduction with incremental 4DVAR

To apply a model reduction method to the inner loop of incremental 4DVAR we have to identify an appropriate dynamical system of the form (12). 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 \( \delta x_0 \) is assumed to be normally distributed white noise with mean zero and covariance identity such that \( \delta x_0 = B_0^{1/2} \omega \). The dynamical system (5)–(6) that constrains incremental 4DVAR may therefore be written equivalently in the form

\[
\delta x_{i-1} = 0, \\
\delta x_i = M_i \delta x_i + w_i, \quad \text{and} \\
\delta u_i = H_i \delta x_i,
\]

(19)

where \( u_i = B_0^{1/2} w_i \) and \( \{ w_i \} \) are white noise inputs satisfying

\[
w_i = \begin{cases} 
\omega \sim \mathcal{N}(0, I_{n_i}), & \text{for } i = -1 \\
0, & \text{for } i \geq 0.
\end{cases}
\]

(20)

The balanced truncation method may then be applied to the system (19) to obtain restriction and prolongation matrices \( U_i^T \) and \( V_i \) that may be used to reduce the system to the form (9)–(10) for use in a simplified incremental 4DVAR scheme. The error between \( d_i \) and \( \tilde{d}_i \), as defined in (14), will then be small for all possible inputs \( \{ w_i \} \), and thus the error will also be small for our special input (20). Since here the restriction and prolongation operators are calculated using 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 \( M_i = M, \quad H_i = H, \quad \text{for } i = -1, \ldots, N-1 \), are all constant. The restriction and prolongation operators \( U^T \) and \( V \) determined by the balanced truncation procedure are also constant, and the reduced-order model matrices are given by \( \hat{M} = U^T MV, \hat{G} = U^T I_{n_r}, \text{and} \hat{H} = HV \). The restricted-state variables are defined by \( \delta x_i = U^T \delta x_i \), and the background error covariance in the restricted space is given by \( \hat{B}_0 = U^T B_0 U \). Since the input
is given by (20), the reduced-order model (9)–(10) is then given by
\[ \hat{\delta x}_{i+1} = U^TMV\hat{\delta x}_i \] and
\[ \hat{d}_i = HV\hat{\delta x}_i, \]
and on the inner loop of the incremental 4DVAR method we minimize
\[ \hat{J}_{i} \approx \frac{1}{2} \| \left\{ \delta \hat{x}^{(k)}_0 - U^T(\hat{x}^b - \hat{x}_0^{(k)}) \right\}^T(U^TB_0U)^{-1} \]
\[ \times \left\{ \delta \hat{x}^{(k)}_0 - U^T(\hat{x}^b - \hat{x}_0^{(k)}) \right\} \]
\[ + \frac{1}{2} \sum_{i=0}^{N} [HV\delta \hat{x}^{(k)}_i - \hat{d}^{(k)}_i]^T R^{-1} [HV\delta \hat{x}^{(k)}_i - \hat{d}^{(k)}_i], \]
subject to the states \( \delta \hat{x}^{(k)} \) satisfying the reduced-order linear model (21). The prolongation operator \( V \) is then used to lift the solution \( \delta \hat{x}^{(k)}_0 \) back into the full space in the outer loop update step.

As derived in Lawless et al. (2005a), the minimization problem (23) is equivalent to a linear least squares problem that can be solved numerically by linear algebraic techniques. We note that this reduced-order data assimilation system is now near ‘optimal’ in the sense that we are matching the observations to predictions from a near ‘optimal’ reduced-order model, as defined by (14).

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 reduced-order models within the inner loop of incremental 4DVAR. 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 (11) and using the prolongation operator to lift the solution back to the full space. The accuracy of the solution found using the standard simplification 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 (23) in the form given by Lawless et al. (2005a) directly via the QR factorization. We now set out the system we use for the experiments and the details of the experimental design.

a. 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 \( h(x) \) 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
\[ \frac{Du}{Dt} + \frac{\partial \phi}{\partial x} = -g \frac{\partial h}{\partial x} \] and
\[ \frac{D(\ln\phi)}{Dt} + \frac{\partial u}{\partial x} = 0, \]
with
\[ \frac{D}{Dt} = \frac{\partial}{\partial t} + u \frac{\partial}{\partial x}. \]
The system is discretized using a semi-implicit semi-Lagrangian integration scheme as described in Lawless et al. (2003). The tangent linear model is derived from the discrete nonlinear model using the standard linearization process.

The initial data for the linearization state are taken from case II of Lawless et al. (2005a). 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 0.00092 s and the gravitational constant is set to \( g = 10 \text{ m s}^{-2} \). The remaining model parameters are set as in Lawless et al. (2005a). For the experiments performed in this study, observations are taken every 10 time steps over a 50-time-step window. The full-resolution matrix \( M \) is obtained by running the tangent linear model for 10 time steps. A time-invariant model is then defined by assuming that this matrix remains constant for successive 10-time-step windows.

The low-resolution model is obtained by the same procedure, running the discrete linearized model on a domain defined over 100 grid points with a separation of 0.02 m, so that the spatial resolution is halved. The time step for the low-resolution model is doubled to 0.0184 s, to ensure that the Courant number is the same in the high- and low-resolution runs. The low-resolution model matrix is then obtained by running the low-resolution tangent linear model for 5 time steps and the low-resolution model is defined by assuming
this matrix remains constant for successive time windows. The linearization state for the low-resolution tangent linear model is taken from a restriction of the high-resolution trajectory of the nonlinear model. This is the same procedure as used in operational centers.

The reduced-order models are obtained by applying the balanced truncation method of section 3 to the full-resolution linear system with model matrix $M$.

The true solution of the linear least squares problem is chosen to be of the form $x_0 = B_0^{1/2} \omega$, as described in section 4, where $B_0$ is the background covariance matrix and $\omega \in \mathbb{R}^n$ is normally distributed white noise with mean zero and covariance identity. To generate a sensible background error covariance matrix we use the approach of Johnson et al. (2005) and define the inverse covariance matrix using a second-derivative smoothing operator with a length scale of 0.04 m, or four grid lengths. The background error standard deviations are set to be 3.5 m s$^{-1}$ for the $u$ field and 1.75 m$^2$ s$^{-2}$ for the $\phi$ field. Further details of the construction of this covariance matrix are provided in the appendix.

The innovation vectors $d$ are the observations for this problem and are generated by applying the time-invariant linear model to the true solution. Where imperfect observations are used, then Gaussian random noise is added to the true solution using the same standard deviations as assumed in the background errors. The observation error covariance matrix $R$ is then defined as a diagonal matrix of the appropriate variances; otherwise, it is taken to be the identity matrix.

\subsection*{b. Comparison of reduced-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 half that of the full resolution. Hence the low-resolution grid has a total of 100 values of $u$ and of $\phi$, 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 low-resolution solution to the linear least squares problem with the solution 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 Fig. 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 $\delta u$ 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 Fig. 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 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 \( r \). 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 Fig. 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.

To test whether the same conclusions hold when the observations contain errors, we add random Gaussian noise to the observations, as described in section 5a. 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 Fig. 4. We see that, as for the case with perfect observations, the model 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.

To understand why the low-order approach shows such a benefit when compared with the low-resolution approach, we examine the Hankel singular vector structure and the eigenstructure of the low-order and low-resolution system matrices of size 200. The reduced system exactly matches the leading Hankel singular values of the full system, and hence the reduced system is expected to replicate the observed optimal growth structures of the full system. This is not the case in the low-resolution system, where the leading Hankel singular values underestimate those of the full system. In Fig. 5 the eigenvalues of these two models are compared 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 model matrix than by the low-resolution model matrix. Hence it appears that the generation of the simplified system by model reduction also acts in such a way as to preserve the normal mode characteristics of the original system, which is not the case in the low-resolution approach. The preservation of both the Hankel singular vector structure and the eigenstructure allows a solution closer to the original analysis to be obtained by the reduced-order assimilation system than by the low-resolution approach.

c. 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 in (15). We now consider how

### Table 1. Comparison of error norms for the low-resolution and the reduced-order method.

<table>
<thead>
<tr>
<th>Reduced order</th>
<th>Low resolution</th>
</tr>
</thead>
<tbody>
<tr>
<td>( r = 200 )</td>
<td>0.0027</td>
</tr>
<tr>
<td>( r = 150 )</td>
<td>0.0134</td>
</tr>
<tr>
<td>( r = 100 )</td>
<td>0.0623</td>
</tr>
<tr>
<td>( r = 90 )</td>
<td>0.1015</td>
</tr>
<tr>
<td>( r = 80 )</td>
<td>0.1726</td>
</tr>
<tr>
<td>( r = 70 )</td>
<td>0.2327</td>
</tr>
<tr>
<td>( r = 60 )</td>
<td>—</td>
</tr>
<tr>
<td>( r = 50 )</td>
<td>—</td>
</tr>
</tbody>
</table>

![Fig. 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).](image)

![Fig. 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).](image)
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 matrix in the Stein equation; that is, instead of (15) and (16) we solve

\[
P = MPM^T + GG^T \quad \text{and} \quad \quad (27)
\]
\[
Q = M^TQM + H^TR^{-1}H. \quad \quad (28)
\]

The error covariance matrix \( B_0 \) in the least squares problem remains the same as in section 5b; the modification is only in the calculation of the reduced-order system.

In Fig. 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 Fig. 2 shows that not incorporating the covariance matrix \( B_0 \) 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.

We remark that we have also carried out similar tests in which the observation operator \( H \) is replaced by the identity operator \( I \) in the Stein equation in (16) and separate tests in which the correct observation covariance matrix \( R \) is not incorporated in (16). The least squares problem again remains the same and it is only the calculation of the reduced system that is altered. In

---

**Fig. 5.** Eigenvalues of (top) full matrix, (middle) reduced-order matrix, and (bottom) low-resolution matrix.

**Fig. 6.** As in Fig. 2, but without incorporating the covariance matrix into the model reduction procedure.
all these tests the accuracy of the analysis produced by
the reduced-order data assimilation system deteriorates, indicating that the observation operator and ob-
servation covariances, as well as the background cova-
riances, are all significant factors in obtaining the best
reduced-order approximation to the full assimilation
problem.

d. Different observation positions

In the experiments described so far we have assumed
that observations of $\delta u$ and $\delta \phi$ are available at every
second grid point. We now test whether the above con-
cclusions continue to hold when the observing network
is changed. We first consider a case in which imperfect
observations of the $\delta u$ field are taken at every grid
point of the full-resolution grid, but no observations of
$\delta \phi$ are taken. In Fig. 7 we show the errors in the com-
puted solutions of the least squares problem using the
low-resolution approach and the model reduction ap-
proach, where both low-order systems are of size 200.
Recalling that the first half of the state vector corre-
sponds to values of $\delta u$ we see that the solution of the $\delta u$
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 $\delta u$ 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 a more similar magnitude for
both the $\delta u$ and $\delta \phi$ variables, but with slightly higher
errors in the unobserved $\delta u$ field. The error plot for this
experiment is shown in Fig. 8. As for all the previous
experiments, the reduced-order approach to solving the
problem yields a much more accurate solution than the
low-resolution approach. Thus we conclude that the re-
sults of section 5b remain valid when the observation
network is changed.

6. Conclusions

When incremental 4DVAR data assimilation is ap-
plied 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 data assimila-
tion system in a restricted space while retaining its es-
tential properties. We have shown how this method
naturally fits into the theory of incremental 4DVAR
with an alternative definition of the restriction and pro-
longation operators. In the numerical experiments per-
fomed we have demonstrated that the model reduction
approach to incremental 4DVAR is more accurate than
the low-resolution approach for the same size of re-
duced 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 inclu-
sion of the covariance information and also the obser-
vation operator in the model reduction procedure. If
care is not taken to include these, then the results may
not improve on the reduced-resolution approach.

This paper has presented only a preliminary study of
combining model reduction techniques and incremental
4DVAR, 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 appropri-
ate for such large-scale systems and other more appro-
priate reduction methods need to be investigated. Effi-
cient methods for including the variation of the system

![Fig. 7. Error in solutions to least squares problem lifted back to
full-state space with observations of $\delta u$ only, for reduced-order
approach (dashed line) and low-resolution approach (dotted line).]

![Fig. 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).]
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 4DVAR has the potential to give an improvement over existing approaches.

Acknowledgments. 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.

APPENDIX

Background Error Covariance Matrix

The background error covariance matrix $B_0$ used in the numerical experiments of this paper is based on the formulation of Johnson et al. (2005), which defines the inverse matrix based on a second-derivative smoother. We define $B_0 = \Sigma^{1/2}C\Sigma^{1/2}$, where $\Sigma$ is a diagonal matrix of variances and $C$ is a correlation matrix. It is assumed that there is no correlation between the background errors of $u$ and $\phi$, so that $C$ is block diagonal. Then the inverse correlation matrix is defined by

$$
C^{-1} = \gamma \begin{pmatrix}
1 + \frac{I^4}{2} (L_{xx})^2 & 0 \\
0 & 1 + \frac{I^4}{2} (L_{xx})^2
\end{pmatrix},
$$

(A1)

where $I$ is the correlation length scale and $\gamma$ is a scalar parameter that is specified so that the maximum magnitude of the elements of $C^{-1}$ is equal to 1. The matrix $L_{xx}$ represents a second-derivative finite-difference operator incorporating periodic boundary conditions of the form

$$
L_{xx} = \frac{1}{\Delta x^2} \begin{pmatrix}
-2 & 1 & 0 & \cdots & 0 & 1 \\
1 & -2 & 1 & 0 & \cdots & \vdots \\
0 & 1 & -2 & \cdots & \vdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\
0 & \cdots & \cdots & \cdots & 1 \\
1 & 0 & \cdots & 0 & 1 & -2
\end{pmatrix},
$$

(A2)

where $\Delta x$ is the spatial step size. When the matrix $C^{-1}$ is inverted, the rows of $C$ give smooth fields that depend on the length scale $I$.

REFERENCES


