tnerols

The University of Reading

Using model reduction methods within incremental 4D-Var

A.S. Lawless, N.K. Nichols

School of Mathematics, MeteorolSchoThe U25(eryThe)-349(ls)]F293.77.93.07TReteR

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 assim-

e±cient 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 O±ce 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 e®ort 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 simpli¯cation 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 a®orded computationally. With such a method it is di±cult 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 ¯eld 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 ¯lter 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

where the operators $H_i : \mathbb{R}^n \to \mathbb{R}^{p_i}$ map the system state to observation space. The observation errors η_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 ϵ^b, so that

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

The background errors ϵ^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

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

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

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

and $\mathbf{M}_i$ and $\mathbf{H}_i$ are the linearizations of the operators M_i and 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

    Set first guess $\delta\mathbf{x}_0^{(k)} = 0$

    $J[\delta\mathbf{x}_0^{(k)}] =$

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

- Update $x_0^{(k+1)} = x_0^{(k)} + \pm x_0^{(k)}$.

- 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 $\pm x_i$ to the space $\mathbb{R}^r$ with $r < n$, and we define variables $\pm\hat{x}_i \in \mathbb{R}^r$ such that $\pm\hat{x}_i = U_i^T \pm x_i$. We also define prolongation operators $V_i \in \mathbb{R}^{r \times n}$ 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

$$\pm\hat{x}_{i+1} = \hat{M}_i \pm \hat{x}_i; \tag{9}$$

$$\hat{d}_i = \hat{H}_i \pm \hat{x}_i; \tag{10}$$

with

$$\hat{M}_i = U_i^T M_i V_i; \tag{11}$$

$$\hat{H}_i = H_i V_i: \tag{12}$$

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 $x_0^{(0)} = x^b$.

- Repeat for $k = 0; : : : ; K - 1$:

  - Find linearization states $x_i^{(k)}$ by integrating the nonlinear model (1) forward from initial state $x_0^{(k)}$ and find innovations $d_i^{(k)}$ using (7).

  - Minimize

$$\mathcal{J}^{(k)}[\pm\hat{x}_0^{(k)}] = \frac{1}{2}(\pm\hat{x}_0^{(k)}$$

(14) $w_i \in \mathbb{R}^{n}$ and

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

$$\begin{aligned} z_0 &= 0, \\ z_{i+1} &= Mz_i + GB_0^{\frac{1}{2}} w_i, \\ d_i &= Hz_i \end{aligned} \tag{14}$$

over the time window $[t_0, t_N]$, where $z_i \in \mathbb{R}^n$ and $d_i \in \mathbb{R}^p$ are the state and output (observation) vectors at time t_i, respectively, and $w_i \in \mathbb{R}^n$ are uncorrelated white noise inputs, normally distributed with mean zero and covariance matrix equal the identity. The matrix $B_0 \in \mathbb{R}^{n \times n}$ represents the covariance of the random inputs $u_i = B_0^{\frac{1}{2}} w_i$, and

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 in°uenced by the inputs and those that have least e®ect 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 P and Q associated with the inputs and outputs, respectively. These symmetric positive-de¯nite matrices satisfy the two Stein equations

$$P = MPM^T + GB_0G^T; \tag{17}$$

$$Q = M^TQM + H^TH: \tag{18}$$

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

To obtain the reduced order model, the system (14) is ¯rst 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 de¯ned to be $\hat{z} = U^Tz$ and the reduced order system matrices are given by

$$\hat{M} = U^TMV; \qquad \hat{G} = U^TG; \qquad \hat{H} = HV; \tag{19}$$

where

$$U^T = [I_r, 0]\Psi^{-1}; \qquad V = \Psi \begin{bmatrix} I_r \\ 0 \end{bmatrix}. \tag{20}$$

The restriction and prolongation operators U^T and V satisfy $U^TV = I_r$ and VU^T is a projection operator. E±cient and accurate numerical techniques are available for ¯nding 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 $\pm x_0$ is assumed to be normally distributed white noise with mean zero and covariance B_0. Thus there exists a normally distributed white noise $\omega \in \mathbb{R}^n$ with mean zero and covariance identity such that $\pm x_0 = B_0^{\frac{1}{2}}\omega$. The dynamical system (5){(6) that constrains incremental 4D-Var may therefore be written equivalently in the form

$$\begin{aligned} \pm x_{i-1} &= 0; \\ \pm x_{i+1} &= M_i\pm x_i + B_0^{\frac{1}{2}}w_i; \\ d_i &= H_i\pm x_i \end{aligned} \tag{21}$$

with white noise inputs $\{w_i\}$ satisfying

$$w_i := \begin{cases} \omega \sim N(0; I_n); & \text{for } i = 1 \\ 0; & \text{for } i \neq 0: \end{cases} \tag{22}$$

The balanced truncation method may then be applied to the system (21) to obtain restriction and prolongation matrices

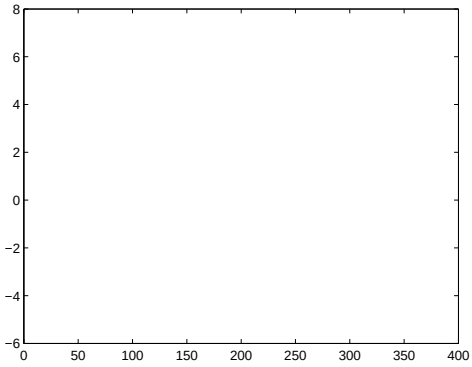

Figure 1: Solution to least squares problem lifted back to full state space. The solid line is the true solution, the dashed line is from the reduced order approach and the dotted line is from the low resolution approach.

We de¯ne the true solution of the linear least squares problem to be the di®erence 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 $ms^{i\,1}$ for the u ¯eld and 0:2 $m^2s^{i\,2}$ for the $Á$ ¯eld, corresponding to approximately 10% of the mean ¯eld values. The observation error covariance matrix $\mathbf{R}$ is then de¯ned as a diagonal matrix of these variances. In order to generate a sensible background error covariance matrix we use the approach of [13] and de¯ne 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 $Á$, making the low order system of order 200. In this case the restriction operator is de¯ned by mapping every second grid point of the high resolution grid onto the low resolution grid, while the prolongation operator is de¯ned 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 ¯rst 200 points of the solution vector correspond to values of the perturbation $±u$ and the last 200 points correspond to values of $±Á$. The error in these solutions, calculated as the di®erence 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

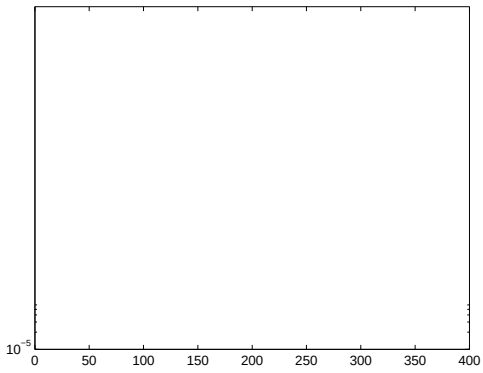

Figure 2: Error in solutions to least squares problem lifted back to full state space for reduced order approach (dashed line) and low resolution approach (dotted line).

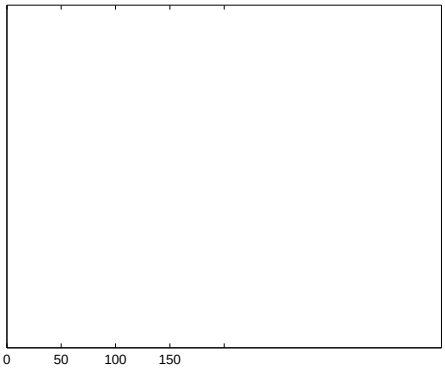

Figure 3: Error in solutions to least squares problem lifted back to full state space for reduced order approach of size 80 (dashed line) and low resolution approach of size 200 (dotted line).

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

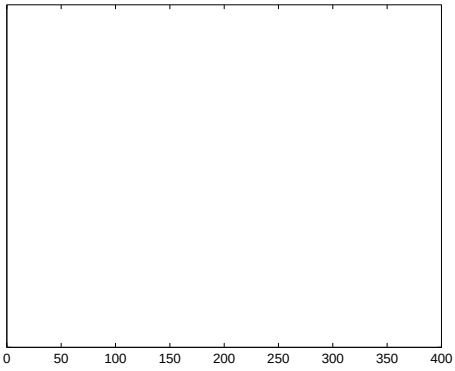

Figure 4: Error in solutions to least squares problem with imperfect observations lifted back to full state space for reduced order approach (dashed line) and low resolution approach (dotted line).

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 Incorpor6(t)27(w)77m310.38-2por6(t)27(w)77327(l)-3cm63orpor6(t)271iorp

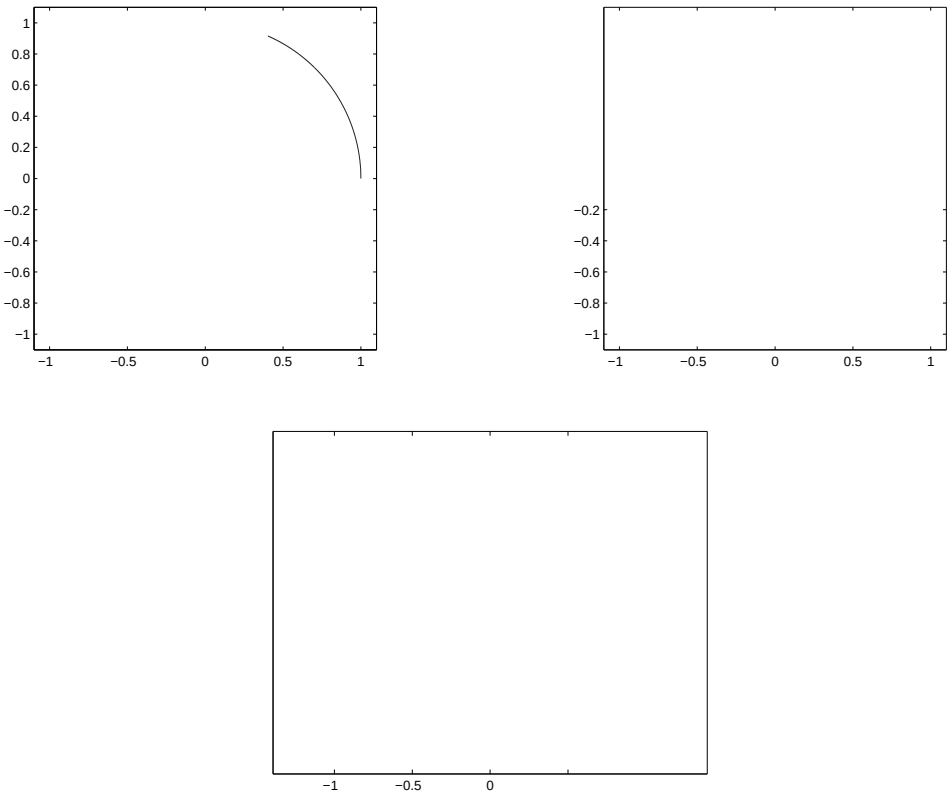

Figure 5: Eigenvalues of full matrix (top left), reduced order matrix (top right) and low resolution matrix (bottom).

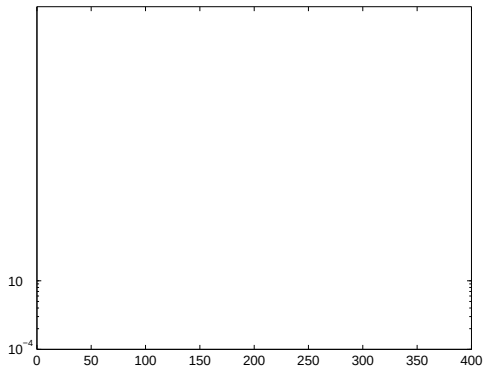

Figure 6: As Figure 2, but without incorporating the covariance matrix into the model reduction procedure.

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

$$P = MPM^T + GG^T;$$
$$Q = M^TQM + H^TH:$$

The error covariance matrix B_0 in the least squares problem remains the same as in Section 5.2; the modi¯cation is only in the calculation of the reduced order system.

In Figure 6 we compare the errors in the ¯nal 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 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 Di®erent observation positions

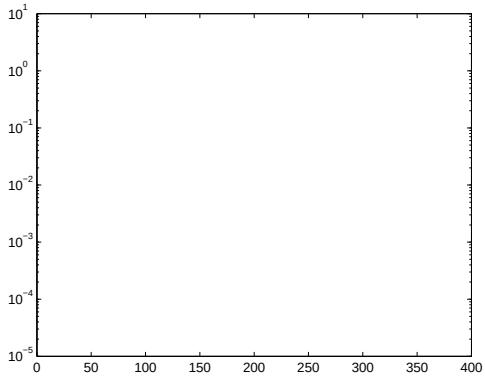

Figure 7: Error in solutions to least squares problem lifted back to full state space with observations of $\pm u$ only, for reduced order approach (dashed line) and low resolution approach (dotted line).

6 Conclusions

When incremental 4D-Var data assimilation is applied to large-scale systems a simpli¯cation 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 ¯ts into the theory of incremental 4D-Var with an alternative de¯nition 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 di®erent observation con¯gurations. 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. E±cient 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.

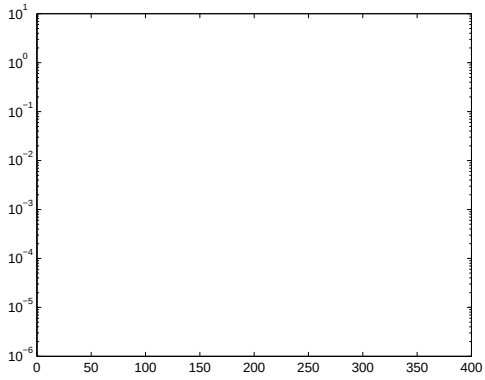

Figure 8: Error in solutions to least squares problem lifted back to full state space with observations of $\pm\acute{A}$ only, for reduced order approach (dashed line) and low resolution approach (dotted line).

Acknowledgements

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

[12] Jazwinski, A.H., 1970: Stochastic processes and ¯ltering 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

[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] Th¶epaut, 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.