Tellus (2005), 57A, 546–559
Printed in Singapore. All rights reserved

TELLUS

Development of NAVDAS-AR: formulation and initial tests of the linear problem

By LIANG XU[1]*, TOM ROSMOND[1] and ROGER DALEY[2], [1]*Marine Meteorology Division, Naval Research Laboratory, Monterey, CA 93943, USA;* [2]*Deceased*

(Manuscript received 20 July 2004; in final form 10 December 2004)

ABSTRACT

A 4-D implementation of an observation space variational data assimilation system is under development at the Marine Meteorology Division of the Naval Research Laboratory (NRL). The system is an extension of the current US Navy 3-D operational data assimilation system, the NRL Atmospheric Variational Data Assimilation System (NAVDAS). The new system, NAVDAS-AR, where AR stands for accelerated representer, is similar in many respects to the European Centre for Medium-Range Weather Forecasts (ECMWF) 4DVAR system. However, NAVDAS-AR is based on a weak constraint observation space, while the ECMWF system is based on a strong constraint model space. In this paper the formulation of NAVDAS-AR is described in detail and preliminary results with a perfect model assumption and comparisons with the operational NAVDAS are presented.

1. Introduction

The NRL Atmospheric Variational Data Assimilation System (NAVDAS), (Daley and Barker, 2000, 2001), is the US Navy's current operational 3-D variational (3DVAR) data assimilation system. NAVDAS-AR is a natural 4-D variational (4DVAR) extension of NAVDAS, where the AR stands for "accelerated representer". It is designed to be used for both regional and global atmospheric data assimilation applications.

The representer algorithm was introduced into oceanography by Bennett and McIntosh (1982) and Bennett and Thorburn (1992). The original representer algorithm (also known as the direct representer algorithm) requires a backward adjoint and a forward model integration to assimilate each observation. The computational cost of data assimilation using the direct representer is proportional to the number of observations. For a small number of observations this algorithm is an effective way to solve a generalized inverse problem—that is, to variationally minimize a 4-D cost function which measures the fit of the analysis to the observations, the forecast model, spatial boundaries and any existing initial state estimate. But it is very expensive to apply the original representer algorithm to assimilate the huge number of observations that are routinely available from various observation platforms in the atmosphere. However, one can accelerate the solution process by using an observation-space descent algorithm that requires computation equivalent to calculation of one representer per iteration and converges in far fewer iterations than the number of observations. This algorithm is described by Egbert et al. (1994), Amodei (1995), Courtier (1998) and Chua and Bennett (2001), and is sometimes referred to as the 4D-PSAS or the indirect representer algorithm.

As originally conceived by Bennett and McIntosh (1982), the algorithm was designed to perform a single minimization over a long time period, achieving an optimum ocean analysis using all of the observations (a rather small number) available for the period. The data assimilation problem in meteorology is somewhat different because of the need to make new forecasts every few hours. In typical meteorological operational practice, an analysis and forecast is run for a short cycle (say 6 h), with the forecast model being run to produce an initial state estimate and then the observations being assimilated in the analysis step to produce an improved state estimate. This process is carried on *ad infinitum*. Xu and Daley (2000) designed a cycling version of the representer algorithm and tested it on a 1-D constituent transport problem. The cycling representer algorithm performs a sequence of 4-D minimizations over short time periods. However, this requires that the error covariances (as well as the state estimates) be updated at the beginning of every cycle. In other words, the analysis and associated analysis error covariance at the analysis time in the current cycle become the initial background condition and the associated initial background error covariance in the next cycle, respectively. The representer mechanics were used in Xu and Daley (2000) to update the covariances.

Covariance updating in any data assimilation algorithm is very computationally expensive. Consequently an accelerated form of

*Corresponding author.
e-mail: xu@nrlmry.navy.mil

the cycling representer algorithm has been developed and is described by Xu and Daley (2002). In the accelerated representer algorithm the error covariances at the beginning of each cycle are specified (as they are in any 3DVAR algorithm and in the standard 4DVAR algorithm) rather than being updated. This is much more computationally tractable. We refer to the accelerated cycling representer as the accelerated representer in the rest of the paper. The accelerated representer algorithm is the *dual* of the standard 4DVAR algorithm when both the observation operator and the dynamic model are linear. That is, while the standard 4DVAR algorithm (as implemented at ECMWF) is a *model grid-space* algorithm, the accelerated representer algorithm is an *observation-space* algorithm (as is NAVDAS itself). Xu and Daley (2002) have examined the properties of the accelerated representer algorithm based on the numerics of the US Navy's Coupled Ocean/Atmosphere Mesoscale Prediction System (COAMPS™; Xu, 1995; Hodur, 1997) in the context of a barotropically unstable shallow water system. It was found that the much cheaper accelerated representer algorithm (i.e. the initial background error covariance being specified) gave satisfactory results as compared with the cycling representer algorithm.

With the successful completion of the shallow water study by Xu and Daley (2002), we began construction of the global application of NAVDAS-AR in the summer of 2001. The numerical weather prediction model used in NAVDAS is the Navy Operational Global Atmospheric Prediction System (NOGAPS), (Hogan and Rosmond, 1991). Briefly, the NOGAPS forecast model is a spherical harmonic spectral model similar in design to those run at several large operational numerical weather prediction (NWP) centres around the world. For simplicity we refer to the global application of NAVDAS-AR as NAVDAS-AR in the rest of the paper. NAVDAS-AR was designed to be a scalable system from the beginning and to be completely compatible with all of the quality control and other observation handling software developed for NAVDAS itself. The multivariate initial background error covariance in NAVDAS-AR is also specified exactly as in NAVDAS (Daley and Barker, 2000, 2001). To our knowledge no other atmospheric implementation of the accelerated representer data assimilation algorithm exists.

We have three objectives in this paper. The first objective is to describe the formulation of NAVDAS-AR with an emphasis on the solution to the linear problem (i.e. the so-called inner loop). The treatment of non-linearity in NAVDAS-AR will be reported in a subsequent companion paper. The second objective is to document the implementation of NAVDAS-AR on scalable computer architectures. The third objective is to demonstrate the potential of NAVDAS-AR for future operational 4-D data assimilation applications. The paper is organized as follows. In Section 2, we first present a generalized inverse problem in terms of minimizing a generalized quadratic cost function. After we obtain a set of non-linear governing equations for estimating the state where the cost function is minimum, we focus on a solution to the linear problem associated with the so-called inner loop. In Section 3, we deal with some of the practical issues related to the implementation of NAVDAS-AR, such as the forward and adjoint models of NOGAPS, the specification of background and observation error covariances, the observation operator and its Jacobian and adjoint, as well as the descent algorithms. In Section 4, we present our initial strong constraint 4-D variational data assimilation results with operational observations using NAVDAS-AR. We compare the analysis increments from NAVDAS-AR and NAVDAS using the same background forecast and the same observations. The summary and conclusions are given in Section 5.

2. Formulation of NAVDAS-AR

In this section, we define a generalized inverse problem formulated in terms of minimizing a generalized non-linear weighted least-square cost function. The fundamental assumption is that all errors are normally distributed. The generalized inverse problem is equivalent to solving a coupled non-linear Euler–Lagrange system, which is in general very difficult to solve directly. However, it is possible to recast the non-linear problem as a sequence of coupled linear problems using one of the iterative algorithms. In fact, we have successfully developed and tested such an iterative algorithm to treat the non-linearity in NAVDAS-AR. We will report the non-linear aspects of NAVDAS-AR in a subsequent companion paper. With various additional appropriate assumptions, the coupled non-linear Euler–Lagrange system can be reduced to a much simpler coupled linear problem from which NAVDAS-AR is constructed. All derivations are presented in discretized form following meteorological convention.

2.1. The generalized non-linear cost function

The generalized inverse problem can be posed as the minimization of the following generalized cost function:

$$J = J_0^{\mathrm{b}} + J^{\mathrm{q}} + J^{\mathrm{r}} \tag{1a}$$

$$J_0^{\mathrm{b}} = \frac{1}{2}\left[\mathbf{x}_0^{\mathrm{b}} - \mathbf{x}_0\right]^{\mathrm{T}}\left[\mathbf{P}_0^{\mathrm{b}}\right]^{-1}\left[\mathbf{x}_0^{\mathrm{b}} - \mathbf{x}_0\right], \tag{1b}$$

$$J^{\mathrm{q}} = \frac{1}{2}\sum_{n=1}^{N}\sum_{n'=1}^{N}\left[\mathbf{x}_n - \mathcal{M}\left(\mathbf{x}_{n-1}\right)\right]^{\mathrm{T}} \times \mathbf{Q}_{nn'}^{-1}\left[\mathbf{x}_{n'} - \mathcal{M}\left(\mathbf{x}_{n'-1}\right)\right], \tag{1c}$$

$$J^{\mathrm{r}} = \frac{1}{2}\sum_{n=0}^{N}\sum_{n'=0}^{N}\left[\mathbf{y}_n - \mathcal{H}\left(\mathbf{x}_n\right)\right]^{\mathrm{T}} \times \mathbf{R}_{nn'}^{-1}\left[\mathbf{y}_{n'} - \mathcal{H}\left(\mathbf{x}_{n'}\right)\right], \tag{1d}$$

for the time period $t_0 \le t_n \le t_N$. Notice that there are $N+1$ time bins and N time steps. J^{r} and J^{q} are the scalar cost functions for

the observation and model error, respectively, and J_0^{b} is the scalar cost function for the background forecast (or prior) error at the beginning of the period ($t = t_0$). Superscripts r, q, b and a refer to the observations, model, background and analysis, respectively. Superscript T refers to a matrix transpose. Subscripts, 0, n and $n-1$ refer to the time at $t = t_0$, $t = t_n$ and $t = t_{n-1}$, respectively. The n' is a time subscript indicating time correlated model or background errors. If the errors are not correlated, $n = n'$. $\mathcal{M}$ is the non-linear forecast model and has I grid points (or more precisely the number of grid points times the number of prediction variables). In the present development $\mathcal{M}$ is the NOGAPS forecast model mentioned above. $\mathbf{x}_n^{\mathrm{b}}$ and $\mathbf{x}_n$ are column vectors of length I which describe the atmospheric estimates at time $t = t_n$. $\mathbf{x}_n^{\mathrm{b}}$ is the prior background forecast and $\mathbf{x}_n$ the truth on the analysis grid. $\mathbf{Q}_{nn'}$ is an $I \times I$ block model error covariance matrix at $t = t_n$ and $t = t_{n'}$. $\mathbf{P}_0^b$ is an $I \times I$ symmetric, positive-definite background error covariance matrix at the initial time $t = t_0$. There are K observations during the time period, with k_n observations for the time t_n and $\sum_{n=0}^{N} k_n = K$. $\mathbf{y}_n$ is a vector of observations of length k_n at time t_n. $\mathcal{H}$ is an operator going from analysis space to observation space that produces k_n pseudo-observations from a given state vector $\mathbf{x}_n$ at time t_n. $\mathbf{R}_{nn'}$ is a $k_n \times k_n$ observation error covariance matrix between observations between $t = t_n$ and $t = t_{n'}$. It includes errors in the measurements and the observation operators.

Taking the first variation of eq. (1a) yields:

$$\begin{aligned}\delta J = & -[\delta \mathbf{x}_0]^{\mathrm{T}} \left[\mathbf{P}_0^{\mathrm{b}}\right]^{-1} \left[\mathbf{x}_0^{\mathrm{b}} - \mathbf{x}_0\right] \\ & + \sum_{n=1}^{N} \sum_{n'=1}^{N} [\delta \mathbf{x}_n - \mathbf{M}_{n-1} \delta \mathbf{x}_{n-1}]^{\mathrm{T}} \\ & \times \mathbf{Q}_{nn'}^{-1} [\mathbf{x}_{n'} - \mathcal{M}(\mathbf{x}_{n-1})] \\ & - \sum_{n=0}^{N} \sum_{n'=0}^{N} [\mathbf{H}_n \delta \mathbf{x}_n]^{\mathrm{T}} \mathbf{R}_{nn'}^{-1} [\mathbf{y}_{n'} - \mathcal{H}(\mathbf{x}_{n'})], \end{aligned} \tag{2}$$

where the matrix $\mathbf{M}_{n-1} = \partial \mathcal{M}(\mathbf{x}) / \partial \mathbf{x}$ evaluated at $\mathbf{x}_{n-1}$ is an $I \times I$ Jacobian matrix (also commonly known as the forward model in atmospheric sciences) corresponding to the (possibly) non-linear model $\mathcal{M}$. When $\mathcal{M}$ is linear we have $\mathcal{M}(\mathbf{x}_{n-1}) = \mathbf{M}_{n-1}\mathbf{x}_{n-1}$. The matrix $\mathbf{H}_n = \partial \mathcal{H}(\mathbf{x}) / \partial \mathbf{x}$ is evaluated at $\mathbf{x}_n$ and is an $k_n \times I$ Jacobian matrix corresponding to the (possibly) non-linear forward observation operator $\mathcal{H}$. When $\mathcal{H}$ is linear we have $\mathcal{H}(\mathbf{x}_n) = \mathbf{H}_n \mathbf{x}_n$. We next introduce an adjoint field, $\boldsymbol{\lambda}^{\mathrm{T}} = (\boldsymbol{\lambda}_1^{\mathrm{T}} \ldots \boldsymbol{\lambda}_n^{\mathrm{T}} \ldots \boldsymbol{\lambda}_N^{\mathrm{T}})$, as a long state vector of length $N \cdot I$, where $\boldsymbol{\lambda}_n$ (of length I) is defined as follows:

$$\boldsymbol{\lambda}_n \equiv \sum_{n'=1}^{N} \mathbf{Q}_{nn'}^{-1} [\mathbf{x}_{n'} - \mathcal{M}(\mathbf{x}_{n'-1})], \quad \text{for } 1 \le n \le N. \tag{3}$$

Let us also introduce an observation-space vector $\mathbf{h}^{\mathrm{T}} = (\mathbf{h}_0^{\mathrm{T}} \ldots \mathbf{h}_n^{\mathrm{T}} \ldots \mathbf{h}_N^{\mathrm{T}})$ of length K, where $\mathbf{h}_n$ (of length k_n) is defined as:

$$\mathbf{h}_n = \sum_{n'=0}^{N} \mathbf{R}_{nn'}^{-1} [\mathbf{y}_{n'} - \mathcal{H}(\mathbf{x}_{n'})], \quad \text{for } 0 \le n \le N. \tag{4}$$

Substituting eqs (3) and (4) into (2), we have:

$$\begin{aligned}\delta J = & -\delta \mathbf{x}_0^{\mathrm{T}} \left[\mathbf{P}_0^{\mathrm{b}}\right]^{-1} \left[\mathbf{x}_0^{\mathrm{b}} - \mathbf{x}_0\right] \\ & + \sum_{n=1}^{N} [\delta \mathbf{x}_n - \mathbf{M}_{n-1} \delta \mathbf{x}_{n-1}]^{\mathrm{T}} \boldsymbol{\lambda}_n \\ & - \sum_{n=0}^{N} [\mathbf{H}_n \delta \mathbf{x}_n]^{\mathrm{T}} \mathbf{h}_n. \end{aligned} \tag{5}$$

Rearranging some of the terms in (5) using integration by parts, we have:

$$\begin{aligned}\delta J = & -\delta \mathbf{x}_0^{\mathrm{T}} \left[\mathbf{P}_0^{\mathrm{b}}\right]^{-1} \left[\mathbf{x}_0^{\mathrm{b}} - \mathbf{x}_0\right] \\ & + \sum_{n=1}^{N} \delta \mathbf{x}_n^{\mathrm{T}} \boldsymbol{\lambda}_n - \sum_{n=0}^{N-1} \delta \mathbf{x}_n^{\mathrm{T}} \mathbf{M}_n^{\mathrm{T}} \boldsymbol{\lambda}_{n+1} \\ & - \sum_{n=1}^{N} \delta \mathbf{x}_n^{\mathrm{T}} \mathbf{H}_n^{\mathrm{T}} \mathbf{h}_n - \delta \mathbf{x}_0^{\mathrm{T}} \mathbf{H}_0^{\mathrm{T}} \mathbf{h}_0, \end{aligned} \tag{6}$$

or

$$\begin{aligned}\delta J = & -\delta \mathbf{x}_0^{\mathrm{T}} \left\{ \left[\mathbf{P}_0^{\mathrm{b}}\right]^{-1} \left[\mathbf{x}_0^{\mathrm{b}} - \mathbf{x}_0\right] + \mathbf{M}_0^{\mathrm{T}} \boldsymbol{\lambda}_1 + \mathbf{H}_0^{\mathrm{T}} \mathbf{h}_0 \right\} \\ & + \sum_{n=1}^{N-1} \delta \mathbf{x}_n^{\mathrm{T}} \left\{ \boldsymbol{\lambda}_n - \mathbf{M}_n^{\mathrm{T}} \boldsymbol{\lambda}_{n+1} - \mathbf{H}_n^{\mathrm{T}} \mathbf{h}_n \right\} \\ & + \delta \mathbf{x}_N^{\mathrm{T}} \left\{ \boldsymbol{\lambda}_N - \mathbf{H}_N^{\mathrm{T}} \mathbf{h}_N \right\}. \end{aligned} \tag{7}$$

At the minimum cost function $J = J_{\min}$, we have $\delta J = 0$, or

$$\boldsymbol{\lambda}_n^{\mathrm{a}} - \mathbf{M}_n^{\mathrm{T}} \boldsymbol{\lambda}_{n+1}^{\mathrm{a}} = \mathbf{H}_n^{\mathrm{T}} \mathbf{h}_n^{\mathrm{a}}, \quad \text{for } 1 \le n \le N-1, \tag{8a}$$

$$\boldsymbol{\lambda}_N^{\mathrm{a}} = \mathbf{H}_N^{\mathrm{T}} \mathbf{h}_N^{\mathrm{a}}, \tag{8b}$$

$$\left[\mathbf{x}_0^{\mathrm{a}} - \mathbf{x}_0^{\mathrm{b}}\right] = \mathbf{P}_0^{\mathrm{b}} \left\{ \mathbf{M}_0^{\mathrm{T}} \boldsymbol{\lambda}_1^{\mathrm{a}} + \mathbf{H}_0^{\mathrm{T}} \mathbf{h}_0^{\mathrm{a}} \right\}. \tag{8c}$$

Using eqs (3)–(4) and (8a)–(8c), we have:

$$\begin{aligned} & \boldsymbol{\lambda}_n^{\mathrm{a}} - \mathbf{M}_n^{\mathrm{T}} \boldsymbol{\lambda}_{n+1}^{\mathrm{a}} = \mathbf{H}_n^{\mathrm{T}} \sum_{n'=0}^{N} \mathbf{R}_{nn'}^{-1} \left[\mathbf{y}_{n'} - \mathcal{H}\left(\mathbf{x}_{n'}^{\mathrm{a}}\right)\right], \\ & \qquad \text{for } 1 \le n \le N-1, \\ & \text{subject to } \boldsymbol{\lambda}_N^{\mathrm{a}} = \mathbf{H}_N^{\mathrm{T}} \sum_{n'=0}^{N} \mathbf{R}_{Nn'}^{-1} \left[\mathbf{y}_{n'} - \mathcal{H}\left(\mathbf{x}_{n'}^{\mathrm{a}}\right)\right], \end{aligned} \tag{9}$$

and

$$\begin{aligned} & \mathbf{x}_n^{\mathrm{a}} - \mathcal{M}\left(\mathbf{x}_{n-1}^{\mathrm{a}}\right) \equiv \sum_{n'=1}^{N} \mathbf{Q}_{nn'} \boldsymbol{\lambda}_{n'}^{\mathrm{a}}, \\ & \qquad \text{for } 1 \le n \le N, \text{ subject to } \left[\mathbf{x}_0^{\mathrm{a}} - \mathbf{x}_0^{\mathrm{b}}\right] = \\ & \qquad \mathbf{P}_0^{\mathrm{b}} \left\{ \mathbf{M}_0^{\mathrm{T}} \boldsymbol{\lambda}_1^{\mathrm{a}} + \mathbf{H}_0^{\mathrm{T}} \sum_{n'=1}^{N} \mathbf{R}_{0n'}^{-1} \left[\mathbf{y}_{n'} - \mathcal{H}\left(\mathbf{x}_{n'}^{\mathrm{a}}\right)\right] \right\}. \end{aligned} \tag{10}$$

Equations (9)–(10) form a coupled non-linear Euler–Lagrange system, which is very difficult to solve. When the forecast model $\mathcal{M}$ and the observation operator $\mathcal{H}$ are linear, however, various representer methods can be used to decouple and solve the linear Euler–Lagrange system (see Bennett (2002), Chua and Bennett (2001) and Xu and Daley (2000, 2002)). A detailed derivation and discussion regarding the treatment of non-linearity in NAVDAS-AR will be given in a follow-on paper. In this paper, we only deal with the linear part of NAVDAS-AR, which is also commonly known as the inner loop.

2.2. *Solution to the coupled linearized Euler–Lagrange system*

When the forecast model and the observation operator are linear, we have $\mathcal{M}(\mathbf{x}_n) = \mathbf{M}_n\mathbf{x}_n$ and $\mathcal{H}(\mathbf{x}_n) = \mathbf{H}_n\mathbf{x}_n$. We further assume that model errors and observation errors are not correlated in time, so we have $\mathbf{Q}_{nn'} = \mathbf{0}$ and $\mathbf{R}_{nn'} = \mathbf{0}$ for $n \neq n'$ and $\mathbf{Q}_n = \mathbf{Q}_{nn'}$ and $\mathbf{R}_n = \mathbf{R}_{nn'}$ for $n = n'$. $\mathbf{Q}_n$ is a symmetric positive-definite block diagonal model error covariance matrix at each $t = t_n$. If we also assume that observation errors are uncorrelated in space, then $\mathbf{R}_n$ is a diagonal matrix of observation error variance at each $t = t_n$. For cycling operational numerical weather prediction applications, the observations at the beginning of a given data assimilation period, $t = t_0$, are usually already used in the previous data assimilation cycle, i.e. at $t = t_N$ of the previous cycle. Since observations should not be assimilated twice, we must exclude observations in the $t = t_0$ bin. which requires $\mathbf{h}_0 = \mathbf{0}$. As a result, we can now simplify eqs (3) and (9)–(10) as the following:

$$\boldsymbol{\lambda}_n^{\mathrm{a}} \equiv \mathbf{Q}_n^{-1}\left[\mathbf{x}_n^{\mathrm{a}} - \mathbf{M}_{n-1}\mathbf{x}_{n-1}^{\mathrm{a}}\right], \quad \text{for } 1 \leq n \leq N, \tag{11}$$

$$\begin{aligned}\boldsymbol{\lambda}_n^{\mathrm{a}} - \mathbf{M}_n^{\mathrm{T}}\boldsymbol{\lambda}_{n+1}^{\mathrm{a}} &= \mathbf{H}_n^{\mathrm{T}}\mathbf{R}_n^{-1}\left[\mathbf{y}_n - \mathcal{H}\left(\mathbf{x}_n^{\mathrm{a}}\right)\right], \\ &\quad \text{for } 1 \leq n \leq N-1, \\ \text{subject to } \boldsymbol{\lambda}_N^{\mathrm{a}} &= \mathbf{H}_N^{\mathrm{T}}\mathbf{R}_N^{-1}\left[\mathbf{y}_N - \mathcal{H}\left(\mathbf{x}_N^{\mathrm{a}}\right)\right],\end{aligned} \tag{12}$$

and,

$$\begin{aligned}\mathbf{x}_n^{\mathrm{a}} - \mathbf{M}_{n-1}\mathbf{x}_{n-1}^{\mathrm{a}} &\equiv \mathbf{Q}_n\boldsymbol{\lambda}_n^{\mathrm{a}}, \quad \text{for } 1 \leq n \leq N, \\ \text{subject to } \left[\mathbf{x}_0^{\mathrm{a}} - \mathbf{x}_0^{\mathrm{b}}\right] &= \mathbf{P}_0^{\mathrm{b}}\mathbf{M}_0^{\mathrm{T}}\boldsymbol{\lambda}_1^{\mathrm{a}}.\end{aligned} \tag{13}$$

The representer method is used to decouple the linear Euler–Lagrange system (12)–(13), yielding

$$\mathbf{x}^{\mathrm{a}} = \mathbf{x}^{\mathrm{b}} + \mathbf{P}^{\mathrm{b}}\mathbf{H}^{\mathrm{T}}[\mathbf{H}\mathbf{P}^{\mathrm{b}}\mathbf{H}^{\mathrm{T}} + \mathbf{R}]^{-1}[\mathbf{y} - \mathbf{H}\mathbf{x}^{\mathrm{b}}] \tag{14}$$

where,

$$\mathbf{P}^{\mathrm{b}} = \mathbf{M}\left[\mathbf{M}_0\mathbf{P}_0^{\mathrm{b}}\mathbf{M}_0^{\mathrm{T}} + \mathbf{Q}\right]\mathbf{M}^{\mathrm{T}} \tag{15}$$

and the prior state background estimate $\mathbf{x}^{\mathrm{b}}$ is from a non-linear model forecast,

$$\begin{aligned}\mathbf{x}_n^{\mathrm{b}} &= \mathcal{M}\left(\mathbf{x}_{n-1}^{\mathrm{b}}\right) \quad \text{for } 1 \leq n \leq N, \\ &\quad \text{subject to the initial condition } \mathbf{x}_0^{\mathrm{b}}.\end{aligned} \tag{16}$$

The details of the derivation can be found in Xu and Rosmond (2004).

$\mathbf{P}^{\mathrm{b}}$ is an $(N \cdot I) \times (N \cdot I)$ background error covariance in model grid space and is also known as the reproducing kernel (Bennett, 2002). $\mathbf{y}^{\mathrm{T}} = [\mathbf{y}_1^{\mathrm{T}} \ldots \mathbf{y}_n^{\mathrm{T}} \ldots \mathbf{y}_N^{\mathrm{T}}]$ is a vector of all observations of length K. $\mathbf{R}$ is the $K \times K$ block diagonal matrix with N blocks each of size $k_n \times k_n$. $\mathbf{H}^{\mathrm{T}} = [\mathbf{H}_1^{\mathrm{T}} \ldots \mathbf{H}_n^{\mathrm{T}} \ldots \mathbf{H}_N^{\mathrm{T}}]$ is a $(N \cdot I) \times K$ matrix, with $\mathbf{H}_n^{\mathrm{T}}$ being an $I \times k_n$ matrix. $\mathbf{H}^{\mathrm{T}}$ essentially brings all the observations to the model variables on the model grid points. $[\mathbf{x}^{\mathrm{b}}]^{\mathrm{T}} = [[\mathbf{x}_0^{\mathrm{b}}]^{\mathrm{T}} \ldots [\mathbf{x}_n^{\mathrm{b}}]^{\mathrm{T}} \ldots [\mathbf{x}_N^{\mathrm{b}}]^{\mathrm{T}}]$ and $[\mathbf{x}^{\mathrm{a}}]^{\mathrm{T}} = [[\mathbf{x}_0^{\mathrm{a}}]^{\mathrm{T}} \ldots [\mathbf{x}_n^{\mathrm{a}}]^{\mathrm{T}} \ldots [\mathbf{x}_N^{\mathrm{a}}]^{\mathrm{T}}]$ are vectors of length $[(N+1) \cdot I]$ of background and analysis state estimates, respectively. $\mathbf{H}\mathbf{P}^{\mathrm{b}}\,\mathbf{H}^{\mathrm{T}}$ is a $K \times K$ background error covariance in observation space and is the same as the representer matrix $\mathbf{R}_e$ as described in eq. (2.15) of Xu and Daley (2000). $\mathbf{P}^{\mathrm{b}}\mathbf{H}^{\mathrm{T}}$ is a $(N \cdot I) \times K$ background error covariance between model and observation-spaces. Equation (14) is formally like the observation-space solution of the 3DVAR NAVDAS problem in Daley and Barker (2000), where they first solve the problem $\mathbf{z} = [\mathbf{H}\mathbf{P}^{\mathrm{b}}\mathbf{H}^{\mathrm{T}} + \mathbf{R}]^{-1}[\mathbf{y} - \mathbf{H}\mathbf{x}^{\mathrm{b}}]$, the solver, and then multiply the solution $\mathbf{z}$ by $\mathbf{P}^{\mathrm{b}}\mathbf{H}^{\mathrm{T}}$, the post-multiplier. However, the formal similarity with NAVDAS is tempered by the presence of the $\mathbf{M}$ and $\mathbf{M}^{\mathrm{T}}$ operators, which explicitly introduce the time coordinate into $\mathbf{P}^{\mathrm{b}}$, making it a gigantic covariance matrix in both space and time that is usually unknown except for the $I \times I$ block $\mathbf{P}_0^{\mathrm{b}}$ that has been prescribed exactly as in NAVDAS (Daley and Barker, 2000, 2001).

From eq. (15), we see that the representer method provides a way to explicitly calculate the state/flow-dependent background error covariance, $\mathbf{P}^{\mathrm{b}}$. An example of explicitly constructed $\mathbf{P}^{\mathrm{b}}$ using both the adjoint and the forward observation operators can be found in Xu and Daley (2000) for a simple one-dimensional problem, where each column of $\mathbf{P}^{\mathrm{b}}$ is obtained by passing a unit impulse through one backward adjoint integration and a corresponding forward integration. Due to the high dimensionality of most atmospheric applications, however, the computational cost and storage requirements make it impractical to explicitly calculate $\mathbf{P}^{\mathrm{b}}$. Fortunately, we do not have to have $\mathbf{P}^{\mathrm{b}}$ explicitly if the purpose is only to obtain the solution of (14) using one of many decent algorithms, such as the conjugate gradient method, where results of multiplying $\mathbf{P}^{\mathrm{b}}$ with a state vector (i.e. a matrix/vector multiplication) are needed. In the following, we describe the accelerated representer procedure used in NAVDAS-AR to solve (14). Interested readers can find a detailed derivation and description of a matrix/vector multiplication using adjoint and forward models in eqs (1)–(6) of Xu and Daley (2002). The procedure consists of two components, namely, the solver and the post-multiplier.

2.2.1. The post-multiplication. $\mathbf{P}^{\mathrm{b}}\mathbf{H}^{\mathrm{T}}$ Although it may seem more natural to begin with the solver, because it is performed first, it is more instructive to begin with the post-multiplication:

$$\mathbf{P}^{\mathrm{b}}\mathbf{H}^{\mathrm{T}}\mathbf{z} \tag{17}$$

First define $\mathbf{z}^{\mathrm{T}} = [\mathbf{z}_1^{\mathrm{T}} \dots \mathbf{z}_n^{\mathrm{T}} \dots \mathbf{z}_N^{\mathrm{T}}]$, where $\mathbf{z}_n^{\mathrm{T}}$ is a vector of length k_n, corresponding to time t_n. Introduce two vectors of length I, $\mathbf{f}_n$ and $\mathbf{g}_n$, which are defined for each value t_n. Then calculate

$$\mathbf{f}_n = \mathbf{M}_n^{\mathrm{T}}\mathbf{f}_{n+1} + \mathbf{H}_n^{\mathrm{T}}\mathbf{z}_n, \quad \text{for } 1 \le n \le N-1, \quad \text{subject to } \mathbf{f}_N = \mathbf{H}_N^{\mathrm{T}}\mathbf{z}_N, \tag{18}$$

using the adjoint forecast model, $\mathbf{M}_n^{\mathrm{T}}$ in (18). This is referred to as the *backward sweep* and produces a vector $\mathbf{f}_0$. We follow this with the *forward sweep* using the forward model, $\mathbf{M}_n$ in the following equation, starting at time t_0,

$$\mathbf{g}_n = \mathbf{M}_{n-1}\mathbf{g}_{n-1} + \mathbf{Q}_n\mathbf{f}_n, \quad \text{for } 1 \le n \le N, \quad \text{subject to } \mathbf{g}_0 = \mathbf{P}_0^{\mathrm{b}}\mathbf{f}_0. \tag{19}$$

Combining (17)–(19) with (14), we finally have

$$\mathbf{x}_n^{\mathrm{a}} = \mathbf{x}_n^{\mathrm{b}} + \mathbf{g}_n, \quad \text{for } 0 \le n \le N. \tag{20}$$

If the forecast model is assumed to be perfect (strong constraint approximation), then $\mathbf{Q}_n = 0$ and eq. (19) becomes,

$$\mathbf{g}_n = \mathbf{M}_{n-1}\mathbf{g}_{n-1}, \text{ for } 1 \le n \le N, \quad \text{subject to } \mathbf{g}_0 = \mathbf{P}_0^{\mathrm{b}}\mathbf{f}_0. \tag{21}$$

Given that the solver, $[\mathbf{H}\mathbf{P}^{\mathrm{b}}\mathbf{H}^{\mathrm{T}} + \mathbf{R}]^{-1}[\mathbf{y} - \mathbf{H}\mathbf{x}^{\mathrm{b}}]$, has already been performed, eqs. (19)–(20) require one single backward and one single forward sweep to produce the new state estimate $\mathbf{x}^{\mathrm{a}}$ for any time during the period. However, unlike the cycling representer algorithm, no estimate of the analysis error covariance is produced.

2.2.2. The solver. $[\mathbf{H}\mathbf{P}^{\mathrm{b}}\mathbf{H}^{\mathrm{T}} + \mathbf{R}]^{-1}[\mathbf{y} - \mathbf{H}\mathbf{x}^{\mathrm{b}}]$ Here, we wish to obtain $\mathbf{z}$, the solution of the problem,

$$[\mathbf{H}\mathbf{P}^{\mathrm{b}}\mathbf{H}^{\mathrm{T}} + \mathbf{R}]\mathbf{z} = \mathbf{y} - \mathbf{H}\mathbf{x}^{\mathrm{b}}, \tag{22}$$

where $[\mathbf{y} - \mathbf{H}\mathbf{x}^{\mathrm{b}}]$ is the innovation vector. Equations such as (22) are usually solved iteratively using descent methods such as the conjugate gradient algorithm (see Daley and Barker, 2000, section 3). In all of these descent methods it is necessary to perform the matrix vector multiplication

$$\mathbf{q} = [\mathbf{H}\mathbf{P}^{\mathrm{b}}\mathbf{H}^{\mathrm{T}} + \mathbf{R}]\mathbf{p}, \tag{23}$$

at each iteration of the descent. Here, $\mathbf{p}$ is a known vector of length K and $\mathbf{q}$ is a vector of length K, which is the result of the matrix/vector multiplication.

Experience has shown that to reduce the condition number of the solver matrix and improve the convergence properties of conjugate gradient descent algorithms, it is useful to non-dimensionalize (23) with the observation errors. If we assume that $\mathbf{R}$ is a pre-specified diagonal matrix of observation error variance and multiply both sides of (23) by $\sqrt{\mathbf{R}^{-1}}$, then

$$\sqrt{\mathbf{R}^{-1}}\mathbf{q} = \sqrt{\mathbf{R}^{-1}}\mathbf{H}\mathbf{P}^{\mathrm{b}}\mathbf{H}^{\mathrm{T}}\mathbf{p} + \sqrt{\mathbf{R}}\mathbf{p}, \tag{24}$$

or,

$$\hat{\mathbf{q}} = \mathbf{q}^{*} + \mathbf{q}^{+}, \tag{25}$$

where $\hat{\mathbf{q}} = \sqrt{\mathbf{R}^{-1}}\mathbf{q}$, $\mathbf{q}^{*} = \sqrt{\mathbf{R}^{-1}}\mathbf{H}\mathbf{P}^{\mathrm{b}}\mathbf{H}^{\mathrm{T}}\mathbf{p}$ and $\mathbf{q}^{+} = \sqrt{\mathbf{R}}\mathbf{p}$.

Equation (25) consists of two separate operations, namely $\mathbf{q}*$ and $\boldsymbol{q}^{+}$. The second operation $\mathbf{p}^{+} = \sqrt{\mathbf{R}}\mathbf{p}$ is trivial. The first operation $\mathbf{q}^{*} = \sqrt{\mathbf{R}^{-1}}\mathbf{H}\mathbf{P}^{\mathrm{b}}\mathbf{H}^{\mathrm{T}}\mathbf{p}$ is much more complex, but can be easily understood by analogy with the post-multiplier eqs (18)–(19). Define, $\mathbf{p}^{\mathrm{T}} = [\mathbf{p}_1^{\mathrm{T}} \dots \mathbf{p}_n^{\mathrm{T}} \dots \mathbf{p}_N^{\mathrm{T}}]$ and $[\mathbf{q}^{*}]^{\mathrm{T}} = [[\mathbf{q}_1^{*}]^{\mathrm{T}} \dots [\mathbf{q}_n^{*}]^{\mathrm{T}} \dots [\mathbf{q}_N^{*}]^{\mathrm{T}}]$, where $\mathbf{p}_n^{\mathrm{T}}$ and $[\mathbf{q}_n^{*}]^{\mathrm{T}}$ are vectors of length k_n, with $\mathbf{p}_n^{\mathrm{T}}$ assumed to be known for $1 \le n \le N$. We also assume that we have working vectors $\mathbf{f}$ and $\mathbf{g}$ available as in Section 2.2.1 Following eq. (18), we then have

$$\mathbf{f}_n = \mathbf{M}_n^{\mathrm{T}}\mathbf{f}_{n+1} + \mathbf{H}_n^{\mathrm{T}}\mathbf{p}_n, \quad \text{for } 1 \le n \le N-1, \quad \text{subject to } \mathbf{f}_N = \mathbf{H}_N^{\mathrm{T}}\mathbf{p}_N. \tag{26}$$

This is the backward sweep. Now, follow this with the forward sweep using the forward model starting at time t_0,

$$\mathbf{g}_n = \mathbf{M}_{n-1}\mathbf{g}_{n-1} + \mathbf{Q}_n\mathbf{f}_n, \quad \text{for } 1 \le n \le N, \quad \text{subject to } \mathbf{g}_0 = \mathbf{P}_0^{\mathrm{b}}\mathbf{f}_0. \tag{27}$$

Finally, we can write the first operation in (25) as

$$\mathbf{q}_n^{*} = \sqrt{\mathbf{R}^{-1}}\mathbf{H}_n\mathbf{g}_n \quad \text{for } 1 \le n \le N. \tag{28}$$

Equations (26)–(27) are essentially the same as eqs (18)–(19). In the event that the model is perfect, eq. (27) is again reduced to

$$\mathbf{g}_n = \mathbf{M}_{n-1}\mathbf{g}_{n-1}, \quad \text{for } 1 \le n \le N, \quad \text{subject to } \mathbf{g}_0 = \mathbf{P}_0^{\mathrm{b}}\mathbf{f}_0. \tag{29}$$

The operation (26)–(28) is performed once per iteration of the solver (23).

3. Implementation of NAVDAS-AR

The accelerated representer algorithm (Xu and Daley, 2002) as highlighted in Section 2 is designed for both global and mesoscale atmospheric data assimilation applications due to the fact it is an observation-space algorithm. It appears that one can simply change the forward and adjoint models in the "solver" and the "post-multiplier" described in Section 2. However, there are at least two main difficulties in the mesoscale implementation of NAVDAS-AR. First, forward and adjoint models of either mesoscale or global NWP applications require major development efforts in their own right, and the problem of lateral boundary conditions presents special problems in the mesoscale case. Second, the pre-specified initial background and model error covariance, $\mathbf{P}_0^{\mathrm{b}}$ and $\mathbf{Q}$, are quite different for the corresponding global and mesoscale applications. For example, we speculate that the errors in the lateral boundary conditions in the mesoscale model may pose a significant difficulty in the commonly used perfect model assumption. In this paper, we focus on the global (NOGAPS) application of NAVDAS-AR. In the following subsections, we give a brief description of the major components of the inner loop of NAVDAS-AR for NOGAPS.

3.1. The innovation vector: $(\mathbf{y} - \mathbf{H}\mathbf{x}^{\mathrm{b}})$

One of the key elements in any given data assimilation system is trying to optimally combine the sampled atmosphere through various measurements, the observations, with the modelled atmospheric, the background. The innovation vector $[\mathbf{y} - \mathbf{H}\mathbf{x}^{\mathrm{b}}]$ is essentially the differences between the sampled and the modelled atmospheric states at the observation locations. The observation operator $\mathbf{H}$ is used to transform the forecasted variables (background) in model space to the observed variables in observation space. Any observation operator, such as temperature to radiance, wind components to wind speed, specific humidity to total precipitable water, etc. must be included here. There are two major steps involved in obtaining the innovation vector, namely the background $\mathbf{x}^{\mathrm{b}}$ and the modelled observations $\mathbf{H}\mathbf{x}^{\mathrm{b}}$.

3.1.1. The background trajectories $\mathbf{x}^{\mathrm{b}}$. The background $\mathbf{x}^{\mathrm{b}}$ is obtained using eq. (16) in the linear coupled Euler–Lagrange system case. Two 6 h background trajectories of $\mathbf{x}^{\mathrm{b}}$ are generated from the integration of NOGAPS: a high-resolution (e.g. T239L30) time-series of model state variables that are used for computing observation innovations, and a low-resolution (e.g. T79L30) time-series of these variables that are used as the linearization trajectories in $\mathbf{M}$ and $\mathbf{M}^{\mathrm{T}}$. The high-resolution case is the so-called outer loop resolution, which will become more relevant when we introduce non-linearity into NAVDAS-AR in a subsequent paper. The low-resolution case is the inner loop resolution, at which both the solver and post-multiplication computations are performed.

3.1.2. The observation operator $\mathbf{H}$. As we discussed above, the observation operator $\mathbf{H}$ is essentially a one-way bridge from model grid space to observation space; we use the following procedures to make the operation $\mathbf{H}\mathbf{x}^{\mathrm{b}}$ very efficient.

Time interpolation: All observations are sorted into 30 min time bins. The forward observation operator is applied to the subset of observations that fall within each time bin, neglecting the time variation over the 30 min period. For a 6 h cycle, there would be 13 such time bins, i.e. in eqs (1c) and (1d), $\mathrm{N} = 12$.

Horizontal and vertical interpolation: All observations (single level, profiles or soundings) have a horizontal location that is defined by its latitude and longitude. In the observation operator $\mathbf{H}$, we interpolate horizontally from the locations of the model grid points to the observation locations. Knowing the latitude and longitude of each model grid point, we use a four-point bilinear interpolation from the four nearest grid points to the observation locations. The weights for a Lagrange-type interpolation are based on the great circle distance between the observation location and each of the four surrounding grid-point locations.

Also in the observation operator $\mathbf{H}$ we vertically interpolate to observation locations linearly in log pressure. The pressure coordinates are defined from the model's background terrain pressure and vertical coordinate (hybrid sigma). The vertical interpolation therefore requires only the two model levels above and below the observation pressure, so a total of eight model points are required for each 3-D interpolation to an observation location. This is a particularly cheap interpolation procedure. All observation types except geopotential and SSM/I precipitable water observations are interpolated this way. These two data types are a special case and are described below.

SSM/I total precipitable water: The vertical operators relating normalized total precipitable water to specific humidity profiles are described in Section 5.6 of Daley and Barker (2000). The SSM/I $\mathbf{H}$ operator, in addition to doing horizontal interpolation as described above, must also project vertically integrated water mass innovations onto profiles of relative humidity innovations. This requires a Jacobian operator that is a function of the forecast model vertical pressure levels and the saturation specific humidity as a function of background temperature at each SSM/I observation location. The background temperature obviously varies a great deal as a function of geographical location. However, because SSM/I observations are all taken over the ocean, we can safely ignore the horizontal variation of model coordinate pressures, a significant simplification. For more details see Section 5.6 of Daley and Barker (2000).

Height observations: NAVDAS-AR does not use height observations except to analyse the surface (terrain) pressure. Innovations are computed by subtracting the height at the observation pressure from background heights at that pressure. These height innovations are converted to terrain pressure innovations by hydrostatic integration between the background terrain pressure and the observation pressure using the background temperature at the surface. However, this procedure is subject to considerable error if the pressure difference is too great. If the observation pressure is less than the background terrain pressure, i.e. in the free atmosphere, the maximum pressure difference allowed is 50 hPa, and if the observation pressure is greater than the terrain pressure, i.e. subterranean, the difference is only 12 hPa. If either threshold is exceeded, the observation is rejected. This is essentially a variation on the well-known reduction to sea level problem necessary for sea level pressure analysis, and we have made a design decision to avoid the problem in NAVDAS-AR. The operational NAVDAS uses a similar design strategy, incorporating a separate 2DVAR univariate analysis for sea level pressure over land areas where the observations are often subterranean. We plan in implement a similar strategy for NAVDAS-AR.

3.2. Adjoint and forward models of NOGAPS

Two key elements in the backward and forward sweeps of NAVDAS-AR are the forward and adjoint models, $\mathbf{M}$ and $\mathbf{M}^{\mathrm{T}}$, of NOGAPS. They were developed at NRL by Rosmond (1997) and have been extensively used in targeted observation studies and predictability research. Currently they contain linearizations of the NOGAPS dynamical core, a simplified surface drag and

vertical mixing parametrization, and no moist physics. Detailed technical descriptions of the NOGAPS forward and adjoint modelling system can be found in Rosmond (1997).

3.3. Convolution of the initial background error covariance

One of the important components of NAVDAS-AR is the convolution of adjoint fields at the initial time with the initial background error covariance, $\mathbf{g}_0 = \mathbf{P}_0^{\mathrm{b}}\mathbf{f}_0$, which occurs in both the "solver" and the "post-multiplier". It is essentially a multiplication of an $I \times I$ matrix by a vector of length I in analysis/model grid space. The convolution can be also seen as a diffusion or smoothing process where observation-space forcing is properly distributed to the analysis/model grid space according to the initial background error covariance matrix $\mathbf{P}_0^{\mathrm{b}}$. It produces an initial condition, $\mathbf{g}_0$ in the model grid space, to be used to initialize the forward sweep. We are currently using a similar algorithm as used in NAVDAS to calculate the convolution in NAVDAS-AR.

In NAVDAS we must compute $\mathbf{a} = \mathbf{P}^{\mathrm{b}}\mathbf{H}^{\mathrm{T}}\mathbf{b}$, which is a matrix/vector multiplication with an incoming observation-space vector $\mathbf{b}$ of length K and an outgoing analysis/model grid-space vector $\mathbf{a}$ of length I. The $\mathbf{H}^{\mathrm{T}}$ operation is an integral part of the covariance calculations, requiring that correlations are computed between actual observation locations. In NAVDAS-AR, however, the observation-space to model-space calculations are integrated into the backward sweep as described in Sections 2.2.1 and 2.2.2, so that at $t = t_0$ the input to $\mathbf{P}_0^{\mathrm{b}}$ is already in model space. This has profound implications for the potential computational cost of NAVDAS-AR *vis-à-vis* NAVDAS that we will discuss in Section 5. In our current NAVDAS-AR implementation of the background error convolution we have adapted an efficient, scalable matrix/vector multiplication algorithm used in NAVDAS by mimicking observation locations at the grid-point locations of the analysis/model grid. This yields much more straightforward operations than the highly irregular and constantly changing observation-space operations performed in NAVDAS.

The variables of NOGAPS are temperature, wind components, specific humidity and surface (terrain) pressure. They are defined on a Gaussian grid with a hybrid vertical coordinate. In the following, we consider first the covariances between temperatures, wind components and moisture, which are similar to the NAVDAS covariances. We then consider surface pressure covariances, which do not occur in NAVDAS.

3.3.1. Covariances between temperatures, wind components and moisture. Since we can define the pressure at any point in the model domain, and from the background temperature the potential temperature at any model point, we have three choices for the vertical coordinate for defining the level surfaces of the error covariance. As noted in Section 4.5 of Daley and Barker (2000), it is only necessary to find the location of two given points in some vertical metric in order to define a covariance. This can be done in pressure, sigma or isentropic coordinates.

We have decided (for the present) to define the level surfaces for the horizontal correlations as pressure surfaces, even though the model surfaces are hybrid sigma surfaces. The relatively large correlation length scales we use produce correlation structures that are more appropriate for nearly level pressure surfaces than for the highly irregular model coordinate surfaces.

The procedure we have devised to compute $\mathbf{g}_0 = \mathbf{P}_0^{\mathrm{b}}\mathbf{f}_0$ is designed to use as much existing NAVDAS code as possible. It is as follows:

(1) Calculate the model coordinate vertical pressure profiles as a function of model terrain pressure at horizontal grid points.

(2) Normalize the elements of $\mathbf{f}_0$ by the appropriate background error variance on the model pressure surfaces (see Daley and Barker, 2000, Section 4.1).

(3) Vertically decompose each vertical column of the normalized $\mathbf{f}_0$ using the vertical eigenvector decomposition of Daley and Barker (2000), Section 4.4.

(4) Perform the horizontal correlations (see Daley and Barker, 2000, Section 4.6) on the eigenvectors from step (3). These are matrix/vector multiples that spread observation influence to adjacent horizontal grid points.

(5) At each horizontal grid point vertically recompose the horizontally correlated eigenvectors back to model coordinate pressure profiles (see Daley and Barker, 2000, Section 4.4). This yields normalized $\mathbf{g}_0$.

(6) Denormalize the results of step (5) using the appropriate background error variance to yield $\mathbf{g}_0$ (see Daley and Barker, 2000, Section 4.1).

3.3.2. Surface pressure covariances. The surface (terrain) pressure calculation is unique to NAVDAS-AR. Since the surface pressure field is a 2-D it is only necessary to use 2-D (horizontal) covariances.

The most obvious thing to do is to define univariate (horizontal) covariances between the incoming surface pressures and the outgoing surface pressures. However, this ignores one important feature of NAVDAS and most MVOI schemes, that is, the geostrophic covariance (in the extratropics) between the surface pressure (or height) and the surface wind fields. This can be quite important in extracting the maximum amount of information from the observations. Consequently, we have decided to include cross-correlations between the surface pressure and the lowest model level winds.

At the Earth's surface the geostrophic relation can be written as

$$f\mathbf{k} \times \mathbf{v}_{\mathrm{s}} = -\nabla\Phi_{\mathrm{s}} - RT_{\mathrm{s}}p_{\mathrm{s}}^{-1}\nabla p_{\mathrm{s}} \tag{30}$$

where f is the Coriolis parameter, $\mathbf{v}_{\mathrm{s}}$ is the horizontal vector wind, $\mathbf{k}$ is the vertically pointing unit vector, Φ_{s} is the specified terrain geopotential, T_{s} is the temperature, R is the gas constant, p_{s} is the surface (terrain) pressure and ∇ is the horizontal gradient

following the lowest model surface. Consider a perturbation $\Delta \mathbf{v}_s$. Then, from (30), we can write,

$$f\mathbf{k} \times \Delta \mathbf{v}_s \approx -RT_s^b \left(p_s^b\right)^{-1} \nabla \Delta p_s, \tag{31}$$

where p_s^b and T_s^b are background values of the surface pressure and temperature respectively and Δp_s is a surface pressure perturbation. We use eq. (31) to define covariances and correlations between the surface pressure and the lowest model level wind components (in the extratropics). Thus, incoming surface pressure innovations can influence outgoing surface wind increments and incoming surface wind innovations can influence outgoing surface pressure increments.

In practice, we create pseudo-heights from the surface pressure innovations in order to make use of the scaling already used in the NAVDAS software.

3.4. The descent algorithm

The conjugate gradient algorithm (Golub and Van Loan, 1996) is used to find an iterative solution of the "solver" eq. (25) instead of eq. (23). One can view (25) as a non-dimensionalized version of (23). Based on our experience, the condition number of eq. (25) is apparently smaller than that of (23). Consequently, fewer iterations are needed to find a solution with the same accuracy using (25) than using (23). A similar pre-conditioning practice has been used in other analysis/model grid-space algorithms, such as the incremental 4DVAR of Courtier et al. (1994).

4. Initial data assimilation tests of NAVDAS-AR

Various data assimilation experiments were conducted to examine the ability of NAVDAS-AR to assimilate observations. In an idealized experiment, we generated seven surface pressure observations by adding 4 hPa to the background surface pressure at collocated model grid points and arbitrary times in the 6 h time window. We assumed the observations were perfect, i.e. zero error, and NAVDAS-AR produced the expected result of drawing exactly to the data. We also tested NAVDAS-AR with single observations as well as a single radiosonde sounding. Results from these experiments produced the expected results that closely matched those from similar idealized experiments conducted by Rabier et al. (1997) and are not presented here. We will present instead NAVDAS-AR data assimilation results using the same observations typically used in a 6 h window of the operational NAVDAS.

4.1. Experiment design

For the experiment, NAVDAS and NAVDAS-AR each use an identical background forecast and observation data set. The background forecast is from the NOGAPS spectral model with T239L30 resolution. The total observation count is about 420 000, which included a comprehensive mix of conventional observations and data from polar orbiting and geostationary satellites. Some of the satellite observations are "superobs". However, although the two systems use the same background forecast, there is a significant difference in how the (observation – background) differences, i.e. the innovations, are computed. For NAVDAS the background fields of temperature, heights, winds and moisture are produced on 30 constant pressure surfaces and a 0.5° latitude–longitude grid, interpolated from a native Gaussian grid/sigma coordinate 6 h forecast history of NOGAPS. Then another interpolation of background values to the observation locations is done, and the innovations are computed. The time dimension is crudely treated by computing innovations using background fields that are valid within 1 h of the actual observation time. However, as in any 3DVAR system such as NAVDAS the resulting innovations are assumed to be valid at the specified analysis time.

NAVDAS-AR improves on the NAVDAS procedures in two ways. First, the time dimension is treated explicitly, with innovations computed and assimilated into the backward sweep of eq. (18) with 30 min resolution. Second, NAVDAS-AR bypasses the first spatial interpolation, interpolating directly from the native NOGAPS grid to the observation locations for innovation calculation. We feel there is a clear benefit from the NAVDAS-AR method, although care must be taken with surface data over land, where the NOGAPS terrain height can depart significantly from the actual Earth's surface terrain. In practice, surface data from areas where there is a large mismatch between the real and model terrain is rejected from NAVDAS-AR to avoid non-representativeness errors.

The innovations computed for each system are subject to identical quality control checks. Any innovation larger in absolute value than three times the expected observation error is rejected. This is based on our experience (and at other NWP centres as well) that innovations that exceed this threshold (or something similar) are often erroneous, and that accepting bad observations in an analysis often does more damage than rejecting potentially good observations. Typically 5–6% of the observations fail the innovation check in each system. In both systems there is also a provision for a "buddy check" quality control test after a specified number of iterations of the "solver" descent algorithm. The buddy check uses the same rejection criteria as the innovation check, so if an observation fails the buddy check, it means the solver is having trouble "fitting" the data, presumably because of inconsistency of that observation with others in close proximity. This often slows the rate of convergence of the descent algorithm, so the observation is rejected and the descent algorithm is restarted with the revised observation set. In mathematical terms rejecting these observations reduces the condition number of $\mathbf{P}^b$ and improves the convergence properties of the solver. In NAVDAS the buddy check is used routinely and typically rejects an additional ≈0.5% of the observations. In NAVDAS-AR we choose not to invoke the buddy check because we find that it

has little impact on the rate of descent algorithm convergence, and therefore see no need to arbitrarily discard data which have already passed the innovation quality control check. We also see negligible differences in NAVDAS-AR results with or without the buddy check.

4.2. Background and trajectories

We employ the full-resolution (T239L30) operational NOGAPS to run a 9 h forecast from 18Z11MAR2004 to 03Z12MAR2004, using the last 6 h to produce the background fields products used by NAVDAS and the background trajectories used by NAVDAS-AR. As described above, the NAVDAS background fields are not on the native NOGAPS grid, necessitating an extra interpolation step to compute innovations. The NAVDAS-AR background trajectories are saved every 30 min on the native grid, and, because NOGAPS is a spectral model, are stored as a file of spherical harmonic coefficients, greatly reducing storage requirements in exchange for the modest computational cost of spectral transforming back to the native grid. The full resolution 3-D native grid is a 720 $\times$ 360 Gaussian grid and 30 hybrid sigma levels from the surface to 1 hPa.

A further advantage of defining the NAVDAS-AR background trajectories in spectral form is that it facilitates implementing a lower resolution for the inner loop than that of the background forecast. The inner loop requires basic state trajectories for the forward $\mathbf{M}$ and adjoint $\mathbf{M}^{\mathrm{T}}$ models. These are easily produced by truncating the full-resolution spectral trajectories to the inner loop spectral resolution. In the current experiment the inner loop resolution is T79L30. During the background forecasts two spectral coefficient files are produced, a T239L30 trajectory and a T79L30 trajectory. The native grid corresponding to the T79L30 inner loop is a 240 $\times$ 120 Gaussian grid with the same vertical levels as the T239L30 trajectory.

4.3. Specification of error covariances

For the experiment we needed to specify various prior error covariances. We assume that all observation errors are uncorrelated spatially, leading to a diagonal observation error covariance matrix, i.e. error variances. These observation error variances are the same as those specified for the operational NAVDAS. The same error correlation functions and background error variances are used to produce the background error covariance matrix at 00Z12MAR2004 for NAVDAS and the initial background error covariance at 21Z11MAR2004 for NAVDAS-AR. Since no dynamic model term is explicitly present in NAVDAS, there is no model error to be specified in the case of NAVDAS. Although NAVDAS-AR is capable of dealing with model error, we use the perfect model assumption in this study, i.e. the model error covariance is zero.

5. Results

In this section, the temperature analysis increments from NAVDAS and NAVDAS-AR are compared at a common analysis time, 00Z12MAR2004. The purpose is not to examine the impact of each data assimilation system on the entire forecast system, but rather to examine the ability of NAVDAS-AR to generate comparable analysis increments.

5.1. Comparison of analysis increments

The horizontal structure of the temperature analysis increments at 500 hPa produced by NAVDAS-AR and NAVDAS at 00Z12MAR2004 are shown in Figs 1a and b, respectively. Overall there is considerable similarity between the two increment fields, which is the desired outcome for an experiment where we tried to make NAVDAS-AR mimic NAVDAS as closely as possible, except for the buddy check quality control test. There are some noticeable local differences, however. In NAVDAS-AR a 1–2° increment occurs over northern Mexico which is absent in NAVDAS. On close examination of this difference we determined it is due to several observations being rejected by the NAVDAS buddy check. The differences over the North Atlantic are also due to buddy check observation rejection by NAVDAS. Only future forecast error sensitivity experiments will tell us which solution is better, but because NAVDAS-AR was able to converge to a seemingly acceptable solution, we are not convinced of the need for a buddy check quality control test yet.

Vertical cross sections of temperature analysis increments from NAVDAS-AR and NAVDAS along 45°N at 00Z12MAR2004 are shown in Figs 2a and b, respectively. We again see similar overall temperature analysis increment patterns from the two systems. The NAVDAS-AR results are somewhat noisier, suggesting a somewhat less dissipative system. Whether or not this extra detail is desirable will also have to wait for future forecast sensitivity experiments. However, we can easily introduce time and space filters into NAVDAS-AR during the time integrations of $\mathbf{M}$ and $\mathbf{M}^{\mathrm{T}}$, making it relatively easy to control undesirable noise in the increment fields.

Noticeable differences between the two systems are present around the two major mountain ranges masked out in the cross sections. This is not surprising considering the different way near-surface observations are treated in the two systems, particularly around mountains. In particular a much stronger temperature signal appears in the NAVDAS-AR increments at low levels just east of the Rocky Mountains.

Figures 3a, b and c show the NAVDAS-AR 500 hPa temperature increment fields at 21Z11MAR2004, 00Z12MAR2004 and 03Z12MAR2004, respectively. Close examination clearly shows that evolution of the increments during the 6 h time window, demonstrating a powerful feature of any 4DVAR data

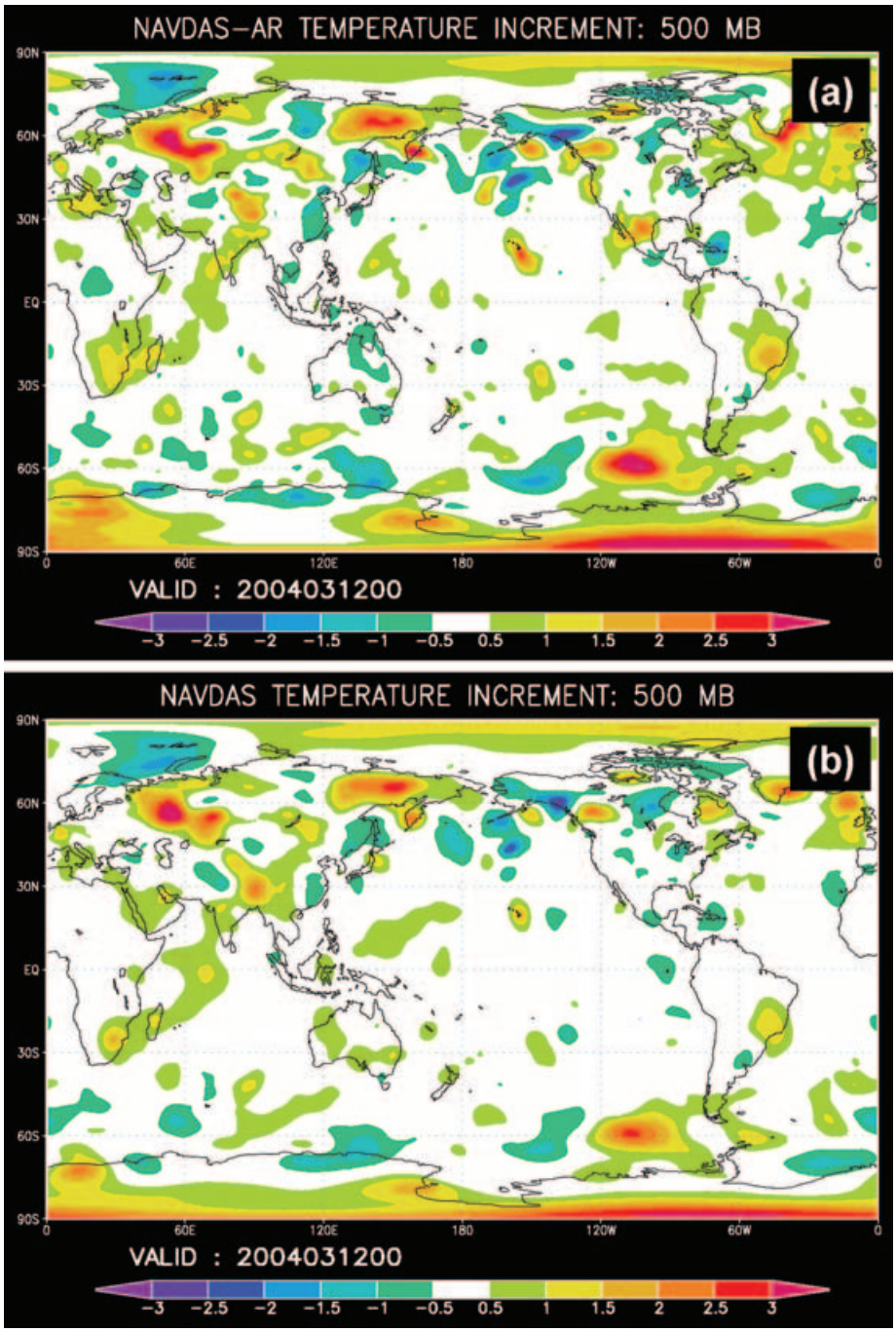

Fig. 1. (a) NAVDAS-AR 500 hPa temperature increments, 00Z12MAR2004. (b) NAVDAS 500 hPa temperature increments, 00Z12MAR2004.

assimilation system; the ability to propagate flow-dependent information. In principle, each of these realizations can be legitimately called an "analysis", blurring the definition of an analysis time, as has been pointed out by Courtier et al. (1994). Furthermore, we can add any of the realizations to its time corresponding background to produce initial conditions for a forecast model, a feature that will be explored in our future testing of possible operational configurations of NAVDAS-AR.

5.2. *Computational details*

Four-dimensional data assimilation is a computationally expensive undertaking no matter what simplifying assumptions and

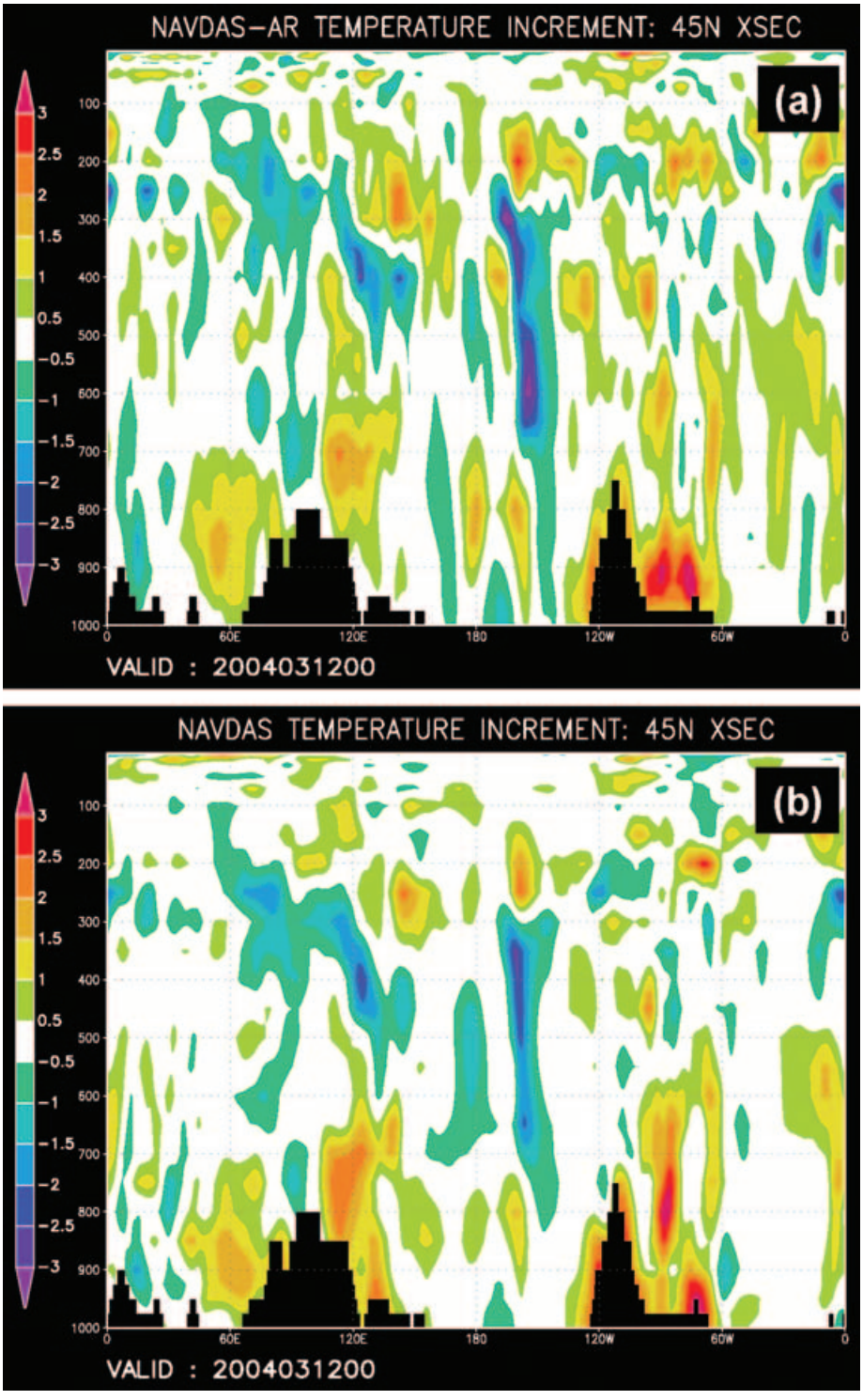

Fig. 2. (a) NAVDAS-AR longitude/pressure 45N X-sec, temperature increments, 00Z12MAR2004. (b) NAVDAS longitude/pressure 45N X-sec, temperature increments, 00Z12MAR2004.

approximations are used. The results show here were produced with an outer loop (i.e. background forecast) resolution of T239L30 and a conjugate gradient inner loop (i.e. analysis increment) resolution of T79L30. These were chosen because we felt they were representative of what was operationally feasible on the computational resources likely to be available for Navy operational NWP applications in the near future. The computational processes can be broken down as follows:

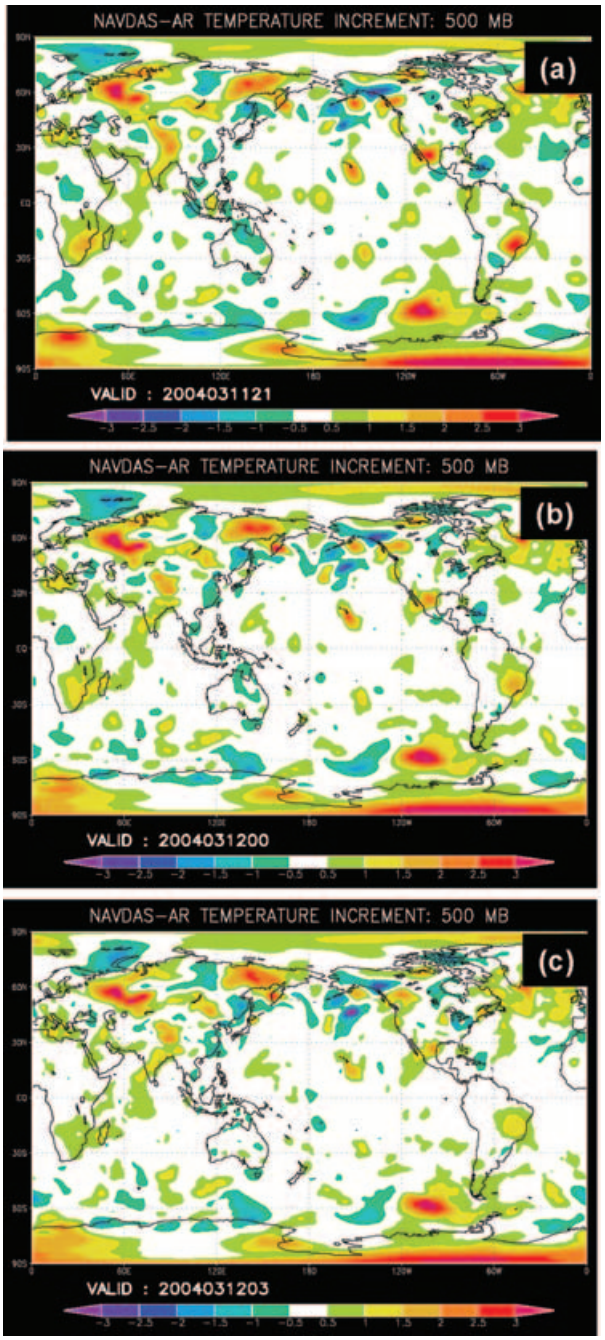

Fig. 3. (a) NAVDAS-AR 500 hPa temperature increments, 21Z11MAR2004. (b) NAVDAS-AR 500 hPa temperature increments, 00Z12MAR2004. (c) NAVDAS-AR 500 hPa temperature increments, 03Z12MAR2004.

(1) Background forecast, typically run at high resolution, e.g. T239L30, which nevertheless is not a major part of the computational effort because it is only run once while the lower-resolution inner loop must be iterated many times.

(2) $\mathbf{H}$ and $\mathbf{H}^{\mathrm{T}}$ operations, which have costs that are a linear function of the number of observations.

(3) $\mathbf{M}$ and $\mathbf{M}^{\mathrm{T}}$ integrations; two low-resolution (T79L30) 6 h forecasts for each inner loop iteration.

(4) Background error covariance calculations with $\mathbf{P}_0^{\mathrm{b}}$, which have costs that are a quadratic function of the inner loop resolution.

NAVDAS-AR costs are dominated by items (3) and (4), because they are the heart of the inner loop and the conjugate gradient algorithm that must be iterated many times to reduce the cost function gradient to acceptably small values. The $\mathbf{H}$ operations (item 2) are also part of the inner loop and its cost is relatively insensitive to observation numbers. This is in sharp contrast to the cost of NAVDAS, which is dominated by a quadratic dependence on observation count. The volume of satellite data is projected to increase by orders of magnitude in the near future, which will necessitate extensive superobbing and/or data thinning to keep NAVDAS computational costs acceptable. NAVDAS-AR, which for the data volumes used in the experiments described here is five to six times as expensive as NAVDAS, will be almost unaffected by these increased data volumes. NAVDAS-AR computational cost is determined by the choice of inner loop resolution. Once this choice is made, the cost is essentially fixed, a significant benefit to operational scheduling.

NAVDAS-AR is a message-passing interface (MPI) application. The experiments described here were run on a SGI ORIGIN 3000 with 40 processors. A particular challenge with an MPI-based data-assimilation system is the need to transform between observation space and analysis space efficiently. The design of the $\mathbf{H}$ and $\mathbf{H}^{\mathrm{T}}$ operators and the ordering of observations in the innovation vector is critical to this efficiency. The minimal cost of the item (2) operations described above is a result of our success in this effort. Likewise the $\mathbf{M}$ and $\mathbf{M}^{\mathrm{T}}$ integrations are very efficient on 40 processors. This is not surprising since they are essentially spectral model dynamical cores which are ideally suited for scalable architectures. On the other hand, the background error covariance calculations do not scale well on 40 processors, and in fact account for almost 75% of NAVDAS-AR wall time. As mentioned in Section 3 we adapted the observation space algorithm from NAVDAS, and it does not exploit the fact that NAVDAS-AR is actually doing the covariance calculations in model grid space and on constant pressure surfaces. Several methods to potentially improve the computational efficiency of the convolution of the initial background error covariance are currently under investigation. Details of the methods will be presented in a future report.

In addition to needing to improve the efficiency of the background error covariance operator, the convergence properties of the conjugate gradient algorithm are far from optimum. Figure 4 shows the cost function gradient as a function of inner loop iteration. The y-axis is actually the ratio of the gradient to the initial norm of the innovation vector, which is the starting value of this gradient in the conjugate gradient algorithm. For the first $\approx$20 iterations this ratio is greater than 1,

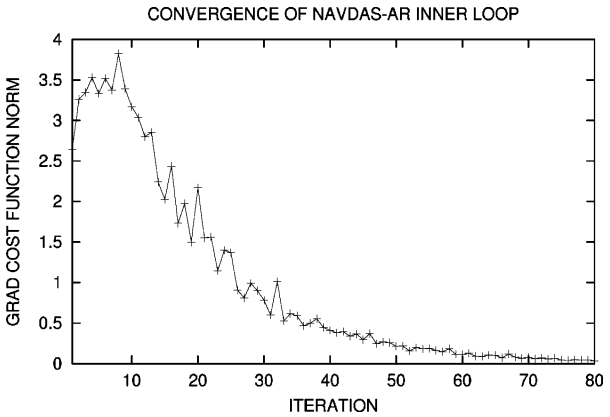

Fig. 4. Convergence of $[\mathbf{HP}^{\mathrm{b}}\mathbf{H}^{\mathrm{T}} + \mathbf{R}]^{-1}[\mathbf{y} - \mathbf{Hx}^{\mathrm{b}}]$, the NAVDAS-AR inner loop.

meaning that the estimated analysis residual is a poorer solution than the original innovations. Daley and Barker (2001) show that a pre-conditioning matrix, which is an approximate inverse to the reproducing kernel (15), is likely to correct this problem of poor initial convergence. NAVDAS uses such a pre-conditioner with great advantage. Unfortunately, the presence of $\mathbf{M}$ and $\mathbf{M}^{\mathrm{T}}$ complicates our problem, and we are still exploring alternative ideas for a useful pre-conditioner for NAVDAS-AR.

6. Summary and conclusions

A natural 4-D variational (4DVAR) extension of NAVDAS, NAVDAS-AR, has been developed by the Naval Research Laboratory in Monterey. NAVDAS-AR is essentially a weak-constraint 4-D variational data assimilation system and is designed for use in both regional mesoscale and global atmospheric 4DVAR applications. It parallels the 4DVAR algorithm implemented at ECMWF except that it is cast in observation space and is capable of including model error during the minimization processes. NAVDAS-AR uses the existing infrastructure of NAVDAS and NOGAPS to solve a series of weighted linear least-square problems. The non-linearity in NAVDAS-AR is treated through so-called outer loops, where NOGAPS and the observation operators are linearized around the analysis in the previous iterations, respectively. The accelerated representer algorithm (Xu and Daley, 2002) is used to decouple and solve the coupled linear Euler–Lagrange system.

In this study, we emphasize the linear global aspect of NAVDAS-AR, i.e. the inner loop using NOGAPS. The solution to the linear problem is formally like the one in Daley and Barker (2001) for NAVDAS except that NAVDAS-AR incorporates the time dimension explicitly. There are two major components to the inner loop of NAVDAS-AR. The first is the solution (in observation space) to the "solver". The second is the solution (in model grid space) of the "post-multiplier". A standard conjugate gradient method is used to obtain an iterative solution to the "solver". The input to the "solver" is the innovation vector, $[\mathbf{y} - \mathbf{Hx}^{\mathrm{b}}]$, while the output from the "solver" is the normalized analysis residual, $[\mathbf{HP}^{\mathrm{b}}\mathbf{H}^{\mathrm{T}} + \mathbf{R}]^{-1}[\mathbf{y} - \mathbf{Hx}^{\mathrm{b}}]$, in observation space. The input to the "post-multiplier" is the analysis residual and the output is the analysis in model grid space.

Our preliminary experimental results show that NAVDAS-AR produces very similar analysis increment fields to NAVDAS, given the same input data and background forecast, although details in implementation introduce subtle differences. We are in the processes of examining these differences to determine if they demonstrate advantages for NAVDAS-AR over NAVDAS.

Future development efforts will be in three areas: (1) examination of the impact of non-linearity with the implementation of outer loops, (2) investigation of alternative representations of the background error covariance calculations to improve computational efficiency, and (3) starting cycling experiments to compare impact of NAVDAS and NAVDAS-AR initial conditions on NOGAPS forecast skill.

7. Acknowledgments

We are deeply indebted to our colleague and co-author, the late Roger Daley, who initiated and set the foundation of NAVDAS-AR. We also thank Andrew Bennett and Boon Chua of Oregon State University and Ed Barker of NRL in Monterey for their encouragement and help during the development. Support of the sponsor, the Naval Research Laboratory, under base programme elements 0601153N and 0602436N, is gratefully acknowledged. We also thank the anonymous reviewers who gave us numerous helpful comments and corrections that greatly improved the manuscript in ways that should help other readers.

References

Amodei, L. 1995. Solution approchee pour un probleme d'assimilation de donnees meteorologiques avec prise en compte de l'erreur de modele. *C. R. Acad. Sci. Ser. IIa* **321**, 1087–1094.

Bennett, A. F. 1992. *Inverse Methods in Physical Oceanography*, Monographs on Mechanics and Applied Mathematics. Cambridge University Press, Cambridge.

Bennett, A. F. 2002. *Inverse Modeling of the Ocean and Atmosphere*. Cambridge University Press, Cambridge.

Bennett, A. F. and McIntosh, P. C. 1982. Open ocean modeling as an inverse problem: tidal theory. *J. Phys. Oceanogr.* **12**, 1004–1018.

Bennett, A. F. and Thorburn, M. A. 1992. The generalized inverse of a nonlinear quasigeostrophic ocean circulation model. *J. Phys. Oceanogr.* **22**, 213–230.

Chua, B. S. and Bennett, A. F. 2001. An inverse ocean modeling system. *Ocean Modeling* **3**, 137–165.

Courtier, P. 1998. Dual formulation of four-dimensional data assimilation. *Q. J. R. Meteorol. Soc.* **123**, 2449–2461.

Courtier, P., Thepaut, J. N. and Hollingsworth, A. 1994. A strategy for operational implementation of 4D-VAR, using an incremental approach. *Q. J. R. Meteorol. Soc.* **120**, 1367–1387.

Daley, R. and Barker, E. 2000. *The NAVDAS Source Book*, Naval Research Laboratory Publication NRL/PJ/7530-01-441. NRL, Monterey, CA.

Daley, R. and Barker, E. 2001. NAVDAS—formulation and diagnostics. *Mon. Weather Rev.* **129**, 869–883.

Egbert, G., Bennett, A. and Foreman, M. 1994. TOPEX/POSEIDON tides estimated using a global inverse method. *J. Geophys. Res.* **99**, 24 821–24 852.

Golub, G. H. and Van Loan, C. F. 1996. *Matrix Computations* 3rd Edition. The Johns Hopkins University Press, Baltimore, MD.

Hodur, R. M. 1997. The Naval Research Laboratory's coupled ocean/atmosphere mesoscale prediction system (COAMPS). *Mon. Weather Rev.* **125**, 1414–1430.

Hogan, T. and Rosmond, T. 1991. The description of the Navy Operational Global Atmospheric Prediction System's spectral forecast model. *Mon. Weather Rev.* **119**, 1786–1815.

Rabier, F., Mahfouf, J.-F., Fisher, M., Jarvinen, H., Simmons, A. and co-authors 1997. *Recent Experimentation on 4D-VAR and First Results from a Simplified Kalman Filter*, ECMWF Research Department Technical Memorandum No. 240. ECMWF, Reading.

Rosmond, T. E. 1997. *A Technical Description of the NRL Adjoint Modeling System*, Naval Research Laboratory Publication NRL/MR/7532/97/7230. NRL, Monterey, CA.

Thepaut, J., Courtier, P., Beland, G. and Leamitre, G. 1996. Dynamical structure functions in a four-dimensional variational assimilation: a case study. *Q. J. R. Meteorol. Soc.* **122**, 535–561.

Xu, L. 1995. *The study of mesoscale land-air-sea interaction processes using a nonhydrostatic model.* Ph.D Dissertation, North Carolina State University, Raleigh, NC.

Xu, L. and Daley, R. 2000. Towards a true 4-dimensional data assimilation algorithm: application of a cycling representer algorithm to a simple transport problem. *Tellus* **52A**, 109–128.

Xu, L. and Daley, R. 2002. Data assimilation with a barotropically unstable shallow water system using representer algorithms. *Tellus* **54A**, 125–137.

Xu, L. and Rosmond, T. 2004. *Formulation of the NRL Atmospheric Variational Data Assimilation System-Accelerated Representer (NAVDAS-AR)*, Naval Research Laboratory Publication NRL/MR/7532-04-36. NRL, Monterey, CA.