A reduced and limited memory preconditioned approach for the 4D-Var data assimilation problem

S. Gratton*, P. Laloyaux†, A. Sartenaer‡, J. Tshimanga§

December 2008

Report number : TR/PA/09/98

Abstract

Data assimilation is a concept involving any method which estimates the initial state of a dynamical system by combining both the information from a numerical model and from observations. The computed estimated initial state of the system can then be integrated in time to obtain a forecast. There are two main ways to solve data assimilation problems. The *sequential methods* are based on statistical estimation theory and regroup the different Kalman filtering approaches. The *variational methods* are based on optimal control theory and state nonlinear weighted least-squares problems as the four-dimensional variational (4D-Var) formulation. Approximations of these methods have been defined to make them suitable for solving large-scale data assimilation problems. In the first part of this paper, we present a theoretical work on the equivalence between the Kalman filter and the 4D-Var, that we then generalize to the approximate case, for the SEEK filter and the reduced 4D-Var. We next concentrate on the solution of the 4D-Var which is usually computed with a Gauss-Newton algorithm using a preconditioned conjugate-gradient-like (CG) method. Motivated by the equivalences shown in the first part, we explore in a second part the techniques used in the SEEK filter, which are based on relevant information contained in the empirical orthogonal functions (EOFs), as an attempt to further accelerate the Gauss-Newton method. This leads to the development of an appropriate starting point for the CG method together with that of a powerful limited memory preconditioner (LMP), as shown by preliminary numerical experiments performed on a shallow water model.

*CERFACS / SUC URA 1875, Toulouse, France
†FUNDP, Namur, Belgium
‡FUNDP, Namur, Belgium
§CERFACS / SUC URA 1875, Toulouse, France

1 Introduction

To compute forecasts of the state of the ocean and of the atmosphere, dynamical systems are integrated on high performance computers within a given time window starting from an initial condition. This time window is related to the relevant time scales of the considered physical situation. The ability to make an accurate forecast requires both that the dynamical systems be a good representation of the reality and that the initial conditions be known precisely. Numerical weather forecasts are performed since the 1950s and have witnessed an explosion of activity since the 1980s. Over the years, the quality of the models and methods for using observation has improved continuously, resulting in major forecast advancements. These progresses are also due to the increased amount of observational data and the increased power of supercomputer that is available. The problem of determining the initial condition for a forecast is very complex and has become an active area of research on its own. *Data assimilation methods* aim at specifying an initial condition of a dynamical system by combining both the information from a numerical model and from observations (see, Kalnay, 2003; Rabier, 2005, for a review).

There are two main categories of data assimilation algorithms : *sequential methods* based on statistical estimation theory and *variational methods* based on optimal control theory. Sequential methods, originally intended for linear dynamical systems and linear observational operators, have been introduced by Kalman (1960) with the *Kalman filter*, and have been successfully applied in a wide range of engineering applications. Adaptations of this filter have been developed to overcome the nonlinearities and the computational cost inherent to large-scale observational systems (Todling and Cohn, 1994; Rozier et al., 2007). It has led to the development of some suboptimal Kalman filters, among which the well known *singular evolutive extended Kalman* (SEEK) filter (Pham et al., 1998; Hoteit and Pham, 2003) using empirical orthogonal functions (EOFs) . Variational data assimilation methods have undergone a rapid development since their introduction (Le Dimet and Talagrand, 1986). These are based on the minimization of a sum of squares of misfits, weighted to take into acount the statistical property of the noise in the problem. The conventional formulation is the *four-dimensional variational problem* (4D-Var) (Courtier et al., 1994). As for the Kalman filter, adaptations have been made for large nonlinear systems in numerical weather prediction. The focus of the research has recently shifted towards the developement of reduced control space methods. A *reduced 4D-Var* formulation has been proposed by Robert et al. (2005), which solve the 4D-Var problem in a subspace spanned by empirical orthogonal functions. Yaremchuk et al. (2009); Wang and Li (2009) consider other strategies using also the empirical orthogonal functions. Another method studied by (Daescu and Navon, 2007; Lawless et al., 2006) is based on the reduction of

the model itself.

To our knowledge, while equivalence between Kalman filter and 4D-Var has been established for linear model and observation operators (Strang and Borre, 1997; Li and Navon, 2001), such an equivalence does not exist for the SEEK filter and the reduced 4D-Var. It is the purpose of the first part of this paper to prove this equivalence. The second part of the paper concentrates on the solution of the 4D-Var which amounts to the solution of a nonlinear weighted least-squares problem. This problem is generally solved by a Gauss-Newton method (Gratton et al., 2007) using a preconditioned conjugate gradient algorithm. In Tshimanga et al. (2008), a class of limited memory preconditioners (LMP) is proposed and studied to accelerate the convergence of the Gauss-Newton process. Motivated by the equivalence proved in the first part, our goal here is to go one step further and to explore the techniques used in the SEEK filter, which are based on relevant information contained in the so-called empirical orthogonal functions (EOFs), as an attempt to further accelerate the Gauss-Newton method. We will show that this leads leads to the development of an appropriate starting point for the CG together with that of a powerful preconditioner belonging to the LMP class proposed in (Tshimanga et al., 2008) .

This paper describes, in Section 2, the sequential and the variational approaches in data assimilation and investigates their theoretical connections in Section 3. The major result is the equivalence between the SEEK filter and the reduced 4D-Var. In Section 4, we describe the Gauss-Newton method to solve the 4D-Var problem, introduce the Ritz-Galerkin starting point for the CG and present some limited memory preconditioning techniques. Section 5 proposes a new approach to solve the 4D-Var problem which improves the Gauss-Newton algorithm by combining the use of a Ritz-Galerkin starting point and of a limited memory preconditioner based on the first empirical orthogonal functions. In Section 6, we interpret the numerical results obtained for a data assimilation problem on a shallow water model. Finally, we conclude and examine perspectives for future research in Section 7.

2 Sequential and variational approaches

For consistency with the literature on the subject, the notation proposed by Ide et al. (1997) will be adopted here as far as possible. A discrete model for the evolution of an atmospheric, oceanic or coupled system is governed by the equation

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

where $\mathbf{x}_i^t \in \mathbb{R}^n$ is a vector representing the true state at time t_i and $\mathcal{M}_{i+1,i}$ is a model operator. The exponent t is introduced to denote the *true* state of the system and $\mathbb{R}^n$ is the state space. The model operator $\mathcal{M}_{i+1,i}$ describes

the system transition from t_i to t_{i+1} and represents the integration in time of some partial differential equations governing the evolution of the system. Calligraphic fonts are used whenever the considered operators are assumed to be nonlinear. Notice that the model operator takes $\mathbf{x}_i^t$ as argument and is perfect since no noise has been added in this system. At each time t_i, the state vector is observed according to

$$\mathbf{y}_i = \mathcal{H}_i \mathbf{x}_i^t + \epsilon_i, \tag{2}$$

where $\mathbf{y}_i \in \mathbb{R}^{p_i}$ is the observation vector with $\mathbb{R}^{p_i}$ the observation space at time t_i, $\mathcal{H}_i$ is an observation operator and where $\epsilon_i \in \mathbb{R}^{p_i}$ is the observational noise due to instrumental and representativeness errors. These noises are assumed to be uncorrelated random vectors with zero means and symmetric positive definite covariance matrices $\mathbf{R}_i \in \mathbb{R}^{p_i \times p_i}$. When the model and observation operators are linear, the dynamical system is described by

$$\mathbf{x}_{i+1}^t = \mathbf{M}_{i+1,i} \mathbf{x}_i^t, \tag{3}$$

where $\mathbf{M}_{i+1,i} \in \mathbb{R}^{n \times n}$ is an n by n matrix, and the observations are obtained by

$$\mathbf{y}_i = \mathbf{H}_i \mathbf{x}_i^t + \epsilon_i, \tag{4}$$

where $\mathbf{H}_i \in \mathbb{R}^{p_i \times n}$ is an p_i by n matrix. The distribution of the observations in space and time could be very nonuniform with regions which are relatively data rich, while others are much more poorly observed. Moreover, the size of the observation vectors $\mathbf{y}_i$ is very much smaller than the size of the state space ($p_i << n$). The *background*, denoted by $\mathbf{x}^b \in \mathbb{R}^n$, is a prior estimation of the state vector. This additional information is available everywhere in the space at the initial time t_0. We suppose that the background is unbiased and uncorrelated with the observations. The background covariance matrix is symmetric positive definite and denoted by $\mathbf{B} \in \mathbb{R}^{n \times n}$.

We are now able to describe the two main approaches to solve data assimilation problems. We suppose that a background $\mathbf{x}^b$ is available at the initial time t_0 and that we can perform N observations from time t_1 to time t_N. All information use to estimate the state of the system comes from the time interval $[t_0, t_N]$ which is called the assimilation window. In sequential data assimilation, the observations $\mathbf{y}_1, \ldots, \mathbf{y}_N$ are available sequentially at times $t_1, \ldots, t_N$. At each observation time t_i, the method constructs an *analysis state* $\mathbf{x}_i^a$ which is an estimate of the true model state $\mathbf{x}_i^t$ using the available observations up to this time, $\mathbf{y}_1, \ldots, \mathbf{y}_i$, and the background $\mathbf{x}^b$. The Kalman filter is originally intended for linear dynamical systems (3) and linear observational operators (4). It computes the analysis state $\mathbf{x}_N^a$ which is the best linear unbiased estimate of the model state at the end of the assimilation window. It means that the estimate is an unbiased linear combination of the the background and the observations with minimal variance.

The n by n error covariance matrix of $\mathbf{x}_N^a$, denoted $\mathbf{P}_N^a$, is also computed by the filter. We give the equations of the Kalman filter in Algorithm 1, that mainly involves two steps. In the *forecast step*, we integrate the last analysis state to obtain a forecast $\mathbf{x}_i^f$ and accordingly update the covariance matrix. In the *analysis step*, we correct the forecast using the observation $\mathbf{y}_i$ and the Kalman gain $\mathbf{K}_i$ to obtain the analysis at time t_i. We also update the analysis covariance matrix denoted by $\mathbf{P}_i^a$.

Algorithm 1 Kalman filter with perfect model operators

1: $\mathbf{x}_0^a = \mathbf{x}^b$
2: $\mathbf{P}_0^a = \mathbf{B}$
3: **for** $i = 1$ to N **do**
4: (* **Forecast step** *)
5: $\mathbf{x}_i^f = \mathbf{M}_{i,i-1}\mathbf{x}_{i-1}^a$
6: $\mathbf{P}_i^f = \mathbf{M}_{i,i-1}\mathbf{P}_{i-1}^a\mathbf{M}_{i,i-1}^{\mathrm{T}}$
7: (* **Analysis step** *)
8: $\mathbf{K}_i = \mathbf{P}_i^f\mathbf{H}_i^{\mathrm{T}}(\mathbf{H}_i\mathbf{P}_i^f\mathbf{H}_i^{\mathrm{T}} + \mathbf{R}_i)^{-1}$
9: $\mathbf{x}_i^a = \mathbf{x}_i^f + \mathbf{K}_i(\mathbf{y}_i - \mathbf{H}_i\mathbf{x}_i^f)$
10: $\mathbf{P}_i^a = \mathbf{P}_i^f - \mathbf{K}_i\mathbf{H}_i\mathbf{P}_i^f$
11: **end for**

The Kalman filter has a simple expression, but its application into realistic ocean and atmospheric models encounters three major difficulties : non-linearity, computational cost and storage for large-scale problems. Non-linearity can be partially solved by linearization of the model and of the observation operators around the state estimate, and by performing iterations. It leads to the so-called *extended Kalman filter* and *iterated Kalman filter* (Jazwinski, 1970; Gelb, 1974; Bell and Cathey, 1993). The dominant significant computational cost of the Kalman filter often comes from the propagation of the error covariance matrix $\mathbf{P}_{i-1}^a$ at line 6 of Algorithm 1 since it implies a number of model integrations equals two matrix products where their rank is equal to the size of the state vector which is larger than 10^6 in many realistic applications. It comes also from the computation of $\mathbf{H}_i\mathbf{P}_i^f\mathbf{H}_i^{\mathrm{T}}$ and from the inversion at line 8. Moreover, storing explicitely the error covariance matrices $\mathbf{P}_i^a$ may be unrealistic for large-scale problems. To overcome these difficulties, a set of suboptimal Kalman filters has been developed (see, Rozier, Birol, Cosme, Brasseur, Brankart, and Verron, 2007, for a review). The Singular Evolutive Extended Kalman filter (SEEK filter) is one of them, and has been proposed in Pham, Verron, and Roubaud (1998). It consists of approximating the error covariance matrices by singular low-rank matrices. Finding a realistic background error covariance matrix in the operational models is also a serious problem. The SEEK filter determines it from an ensemble of state vectors $\mathbf{x}_1, \ldots, \mathbf{x}_l \in \mathbb{R}^n (l \geq 2)$

supposedly representative of the variability of the system and computed on a time interval before the assimilation window. Regarding the state vector as a random vector and having a sample of realizations, we calculate the *sample covariance matrix* using the formula

$$\mathbf{S} = \frac{1}{l-1} \sum_{i=1}^{l} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^{\mathrm{T}},$$

where $\bar{\mathbf{x}}$ is the sample mean. Formaly, one considers a spectral decomposition on this covariance matrix

$$\mathbf{S} = \mathbf{L}_0 \mathbf{U}_0 \mathbf{L}_0^{\mathrm{T}} + \hat{\mathbf{L}}_0 \hat{\mathbf{U}}_0 \hat{\mathbf{L}}_0^{\mathrm{T}}, \tag{5}$$

where the diagonal matrix $\mathbf{U}_0 \in \mathbb{R}^{r \times r}$ contains the r largest eigenvalues of $\mathbf{S}$ and the diagonal matrix $\hat{\mathbf{U}}_0 \in \mathbb{R}^{(n-r)\times(n-r)}$ contains the rest of the eigenvalues, while the matrices $\mathbf{L}_0 \in \mathbb{R}^{n \times r}$ and $\hat{\mathbf{L}}_0 \in \mathbb{R}^{n \times (n-r)}$ are formed with the corresponding eigenvectors, called *empirical orthogonal functions* (EOFs). Note that in most practical situations, only $\mathbf{L}_0$ and $\mathbf{U}_0$ are really needed. In particular, the initial error covariance matrix in the SEEK filter is then defined by

$$\mathbf{P}_0^a = \mathbf{L}_0 \mathbf{U}_0 \mathbf{L}_0^{\mathrm{T}},$$

where only $\mathbf{L}_0$ and $\mathbf{U}_0$ have to be computed. This low-rank matrix plays the role of the background error covariance matrix $\mathbf{B}$ and keeps most of the information contained in $\mathbf{S}$ since $\mathbf{U}_0$ contains the largest eigenvalues. The columns of the partition matrix $[\mathbf{L}_0 \ \hat{\mathbf{L}}_0]$ form an orthonormal basis of $\mathbb{R}^n$ called EOFs basis The equations of the SEEK filter, given in Algorithm 2, follow from the equations of the Kalman filter and from this low-rank approximation (see, Pham et al., 1998). Notice that the error covariance matrices at line 7 and 12 are low-rank matrices and are computed only for diagnostic reasons. The SEEK filter performs corrections only in the directions spanned by the columns of $\mathbf{L}_i$. These columns evolve since they are integrated in time at each iteration.

There are physical considerations with oceanographic or atmospheric models which support the above approximation. The reason is that most of the variability of the oceanic and atmospheric systems can be described by a low dimensional subspace. Indeed, the ocean and atmosphere are basically forced and dissipative dynamical systems that exhibit an attractor, meaning that asymptotically the trajectories of the state vector belong only to a small part of the phase space. In the vicinity of this attractor, orthogonal perturbations will be naturally damped while tangent perturbations will not. This is why the SEEK filter performs corrections only in the directions spanned by the columns of $\mathbf{L}_i$ which are the main directions of variability tangent to the attractor. The choice of the subspace defining the directions of correction is not unique. It is possible to use singular vectors, Liapunov

Algorithm 2 SEEK filter with perfect model operators

$\mathbf{x}_0^a = \mathbf{x}^b$
$\mathbf{P}_0^a = \mathbf{L}_0\mathbf{U}_0\mathbf{L}_0^{\mathrm{T}}$
for $i = 1$ to N **do**
  (* **Forecast step** *)
  $\mathbf{x}_i^f = \mathbf{M}_{i,i-1}\mathbf{x}_{i-1}^a$
  $\mathbf{L}_i = \mathbf{M}_{i,i-1}\mathbf{L}_{i-1}$
  $\mathbf{P}_i^f = \mathbf{L}_i\mathbf{U}_{i-1}\mathbf{L}_i^{\mathrm{T}}$
  (* **Analysis step** *)
  $\mathbf{U}_i = (\mathbf{U}_{i-1}^{-1} + \mathbf{L}_i^{\mathrm{T}}\mathbf{H}_i^{\mathrm{T}}\mathbf{R}_i^{-1}\mathbf{H}_i\mathbf{L}_i)^{-1}$
  $\mathbf{K}_i = \mathbf{L}_i\mathbf{U}_i\mathbf{L}_i^{\mathrm{T}}\mathbf{H}_i^{\mathrm{T}}\mathbf{R}_i^{-1}$
  $\mathbf{x}_i^a = \mathbf{x}_i^f + \mathbf{K}_i(\mathbf{y}_i - \mathbf{H}_i\mathbf{x}_i^f)$
  $\mathbf{P}_i^a = \mathbf{L}_i\mathbf{U}_i\mathbf{L}_i^{\mathrm{T}}$
end for

vectors or bred grown vectors. They are computed using the tangent linear model or the nonlinear model and contain, in a sense, the main directions of variability of the system. Durbiano (2001) performed a study of these families of vectors. When the data assimilation is performed on a shallow water model, she concluded to the clear superiority of the EOFs basis with regards to the other subspaces.

The variational data assimilation is the second approach for data assimilation. The conventional formulation of this approach is the so-called four-dimensional variational problem (4D-Var). The 4D-Var problem can be stated as that of determining the system state $\mathbf{x}_0$ at the initial time t_0 that produces a trajectory that best fits, in a nonlinear weighted least-squares sense, both the background $\mathbf{x}^b$ and the observations $\mathbf{y}_1, \ldots, \mathbf{y}_N$. Mathematically, we formulate the minimization problem as a nonlinear least-squares problem

$$\min_{\mathbf{x}\in\mathbb{R}^n} J(\mathbf{x}) = \frac{1}{2}(\mathbf{x}-\mathbf{x}^b)^{\mathrm{T}}\mathbf{B}^{-1}(\mathbf{x}-\mathbf{x}^b) + \frac{1}{2}\sum_{i=1}^{N}(\mathbf{y}_i - \mathcal{H}_i\mathcal{M}_{i,0}\mathbf{x})^{\mathrm{T}}\mathbf{R}_i^{-1}(\mathbf{y}_i - \mathcal{H}_i\mathcal{M}_{i,0}\mathbf{x}), \quad (6)$$

where $\mathcal{M}_{i,0} = \mathcal{M}_{i,i-1}\ldots\mathcal{M}_{1,0}$ is an operator which describes the system transition from t_0 to t_i . This method uses all available observations at the same time to define the objective function. If the model and observation operators are linear, we replace the nonlinear operators by $\mathbf{M}_{i,0} = \mathbf{M}_{i,i-1}\ldots\mathbf{M}_{1,0}$ and $\mathbf{H}_i$ and the 4D-Var becomes a linear least-squares problem. In every instance, the 4D-Var formulation can be written in the more

compact form

$$\min_{\mathbf{x}\in\mathbb{R}^n} J(\mathbf{x}) = \frac{1}{2}||\mathbf{x}-\mathbf{x}^b||^2_{\mathbf{B}^{-1}} + \frac{1}{2}\sum_{i=1}^{N}||\mathbf{y}_i - \mathcal{H}_i\mathcal{M}_{i,0}\mathbf{x}||^2_{\mathbf{R}_i^{-1}}, \quad (7)$$

where weighted norms, defined by $||\mathbf{z}||^2_{\mathbf{B}^{-1}} = \mathbf{z}^{\mathrm{T}}\mathbf{B}^{-1}\mathbf{z}$ and $||\mathbf{z}||^2_{\mathbf{R}_i^{-1}} = \mathbf{z}^{\mathrm{T}}\mathbf{R}_i^{-1}\mathbf{z}$, are used.

As for the Kalman filter, it is possible to reduce the computational cost of the method. The 4D-Var problem (7) can be defined in a subspace of low dimension. One can uniquely express the variable $\mathbf{x}$ in the EOFs basis as

$$\mathbf{x} = \mathbf{L}_0\mathbf{x}_1 + \hat{\mathbf{L}}_0\mathbf{x}_2.$$

The 4D-Var problem using this decomposition can then be stated as

$$\min_{\mathbf{x}\in\mathbb{R}^n} J(\mathbf{x}) = \frac{1}{2}||\mathbf{L}_0\mathbf{x}_1 + \hat{\mathbf{L}}_0\mathbf{x}_2 - \mathbf{x}^b||^2_{\mathbf{B}^{-1}} + \frac{1}{2}\sum_{i=1}^{N}||\mathbf{y}_i - \mathcal{H}_i\mathcal{M}_{i,0}(\mathbf{L}_0\mathbf{x}_1 + \hat{\mathbf{L}}_0\mathbf{x}_2)||^2_{\mathbf{R}_i^{-1}}.$$

To perform corrections in the same directions as those generated by the SEEK filter, the components of $\mathbf{x}$ in the range of $\hat{\mathbf{L}}_0$ are neglected. A new optimization problem depending only on $\underline{\mathbf{x}} = \mathbf{x}_1 \in \mathbb{R}^r$ can thus be defined, which approximates the previous one

$$\min_{\underline{\mathbf{x}}\in\mathbb{R}^r} \underline{J}(\underline{\mathbf{x}}) = \frac{1}{2}||\mathbf{L}_0\underline{\mathbf{x}} - \mathbf{x}^b||^2_{\mathbf{B}^{-1}} + \frac{1}{2}\sum_{i=1}^{N}||\mathbf{y}_i - \mathcal{H}_i\mathcal{M}_{i,0}\mathbf{L}_0\underline{\mathbf{x}}||^2_{\mathbf{R}_i^{-1}}. \quad (8)$$

In many oceanographic and atmospheric problems (Hoteit and Pham, 2003), one can choose the dimension of the subspace $\mathbf{L}_0$ very small compared to the dimension of the state space, since only a few EOFs explains most of the variability of the system. As a consequence, the computational cost to solve the approximated problem (8) is very much lower than the solution of the initial 4D-Var problem (7).

3 Connections between sequential and variational approaches

The connections between sequential and variational methods have been known for long. For linear and noisy model and observation operators, both

the Kalman filter and the 4D-Var yield the same values for the state vector at each observation time in the assimilation window (Strang and Borre, 1997; Li and Navon, 2001). Their proofs are based respectively on a block matrix and on a statistical view. In this section, we prove the equivalence between the Kalman filter and the 4D-Var in the specific case of a perfect linear model operator and of a noisy linear observation operator. We propose an alternative proof, which is based on quadratic optimization techniques, and is both short and elementary. Moreover, we use some results from this proof to establish a useful connection between the SEEK filter and the reduced 4D-Var problem presented in Section 1, that we believe to be new. In order to be as concise as possible in what follows, we set $\mathbf{M}_i = \mathbf{M}_{i,0}$, $\mathbf{H} = \mathbf{H}_i$ and $\mathbf{R} = \mathbf{R}_i$. Generalizing the proof would just result in heavier notation.

3.1 Connection between 4D-Var / Kalman filter

Theorem 1 *Suppose that the model operator is perfect and linear, that the observation operator is linear and that the same background* $\mathbf{x}^b$*, observations* $\mathbf{y}_i$*, background covariance matrix* $\mathbf{B}$*, and observation covariance matrices* $\mathbf{R}_i$*, are given. Then the analysis state* $\mathbf{x}_i^a$ *computed at time* t_i *by the Kalman filter and the solution produced by the 4D-Var, using the first* i *observations and integrated up to time* t_i*, are identical and have same covariance matrices. Both methods hence produce the same results at the end of the assimilation window.*

Proof. We first introduce the notation $\mathbf{x}_l^k$ to denote the solution of the 4D-Var (6) using the first l observations,

$$\min_{\mathbf{x}\in\mathbb{R}^n} J_l(\mathbf{x}) = \frac{1}{2}||\mathbf{x}-\mathbf{x}_b||^2_{\mathbf{B}^{-1}} + \frac{1}{2}\sum_{i=1}^{l}||\mathbf{y}_i - \mathbf{H}\mathbf{M}_i\mathbf{x}||^2_{\mathbf{R}^{-1}},$$

and integrated up to time t_k, and $\mathbf{P}_l^k$ to denote the covariance matrix of $\mathbf{x}_l^k$. We thus want to prove, for each observation time $l = 1, \ldots, N$, that the solution of the 4D-Var using the first l observations and integrated up to time t_l is equal to the analysis state produced by the Kalman filter after l iterations, i.e., $\mathbf{x}_l^l = \mathbf{x}_l^a$, and that both state vectors have the same covariance matrices, i.e., $\mathbf{P}_l^l = \mathbf{P}_l^a$. The proof is by induction.
To prove that $\mathbf{x}_1^1 = \mathbf{x}_1^a$ and $\mathbf{P}_1^1 = \mathbf{P}_1^a$, we first notice that $\mathbf{x}_0^0 = \mathbf{x}^b = \mathbf{x}_0^a$ since $\mathbf{x}^b$ is the solution of the 4D-Var problem using no observations. Moreover, Rabier and Courtier (1992) proved that the covariance matrix of a 4D-Var solution is equal to the Hessian inverse. One can thus conclude that the covariance matrix of $\mathbf{x}_0^0$ is given by $\mathbf{P}_0^0 = \mathbf{B} = \mathbf{P}_0^a$ since $\mathbf{B}$ is the Hessian inverse of the 4D-Var problem using no observations. We next define the 4D-Var problem using only the first observation $\mathbf{y}_1$ at time t_1 as

$$\min_{\mathbf{x}\in\mathbb{R}^n} J_1(\mathbf{x}) = \frac{1}{2}||\mathbf{x}-\mathbf{x}_0^0||^2_{(\mathbf{P}_0^0)^{-1}} + \frac{1}{2}||\mathbf{y}_1 - \mathbf{H}\mathbf{M}_1\mathbf{x}||^2_{\mathbf{R}^{-1}}. \tag{9}$$

The solution $\mathbf{x}_1^0$ of this problem is computed by nullifying its gradient

$$(\mathbf{P}_0^0)^{-1}(\mathbf{x}_1^0 - \mathbf{x}_0^0) - \mathbf{M}_1^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{y}_1 - \mathbf{H}\mathbf{M}_1\mathbf{x}_1^0) = 0. \tag{10}$$

Defining

$$\mathbf{W} \equiv (\mathbf{P}_0^0)^{-1} + \mathbf{M}_1^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}\mathbf{M}_1, \tag{11}$$

we obtain, by simple extraction of the vector $\mathbf{x}_1^0$ from (10)

$$\mathbf{x}_1^0 = \mathbf{W}^{-1}\left(\mathbf{M}_1^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{y}_1 + (\mathbf{P}_0^0)^{-1}\mathbf{x}_0^0\right). \tag{12}$$

Note that the matrix $\mathbf{W}$ is invertible since it is the sum of a positive definite and a positive semidefinite positive matrix which gives a positive definite matrix. From the Sherman-Morisson-Woodbury formula (see, for example, Horn and Johnson, 2006, p.18), we have that

$$\mathbf{W}^{-1} = \mathbf{P}_0^0 - \mathbf{P}_0^0\mathbf{M}_1^{\mathrm{T}}\mathbf{H}^{\mathrm{T}} (\mathbf{R} + \mathbf{H}\mathbf{M}_1\mathbf{P}_0^0\mathbf{M}_1^{\mathrm{T}}\mathbf{H}^{\mathrm{T}})^{-1}\mathbf{H}\mathbf{M}_1\mathbf{P}_0^0. \tag{13}$$

Using this formula in equation (12), developing the product and reordering the terms, we obtain

$$\begin{aligned}\mathbf{x}_1^0 =& \mathbf{x}_0^0 + \mathbf{P}_0^0\mathbf{M}_1^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{y}_1 \\ &- \mathbf{P}_0^0\mathbf{M}_1^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}(\mathbf{R} + \mathbf{H}\mathbf{M}_1\mathbf{P}_0^0\mathbf{M}_1^{\mathrm{T}}\mathbf{H}^{\mathrm{T}})^{-1} \\ &\mathbf{H}\mathbf{M}_1\mathbf{P}_0^0\left(\mathbf{M}_1^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{y}_1 + (\mathbf{P}_0^0)^{-1}\mathbf{x}_0^0\right).\end{aligned} \tag{14}$$

We observe that the last line of this equation can be written as

$$\begin{aligned}&\mathbf{H}\mathbf{M}_1\mathbf{P}_0^0\left(\mathbf{M}_1^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{y}_1 + (\mathbf{P}_0^0)^{-1}\mathbf{x}_0^0\right) \\ =&\mathbf{H}\mathbf{M}_1\mathbf{P}_0^0\mathbf{M}_1^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{y}_1 + \mathbf{H}\mathbf{M}_1\mathbf{x}_0^0 \\ =&(\mathbf{H}\mathbf{M}_1\mathbf{P}_0^0\mathbf{M}_1^{\mathrm{T}}\mathbf{H}^{\mathrm{T}} + \mathbf{R} - \mathbf{R})\mathbf{R}^{-1}\mathbf{y}_1 + \mathbf{H}\mathbf{M}_1\mathbf{x}_0^0 \\ =&(\mathbf{H}\mathbf{M}_1\mathbf{P}_0^0\mathbf{M}_1^{\mathrm{T}}\mathbf{H}^{\mathrm{T}} + \mathbf{R})\mathbf{R}^{-1}\mathbf{y}_1 - \mathbf{y}_1 + \mathbf{H}\mathbf{M}_1\mathbf{x}_0^0.\end{aligned} \tag{15}$$

Substituting now equation (15) in (14), we obtain after appropriate simplifications

$$\mathbf{x}_1^0 = \mathbf{x}_0^0 + \mathbf{P}_0^0\mathbf{M}_1^{\mathrm{T}}\mathbf{H}^{\mathrm{T}} (\mathbf{R} + \mathbf{H}\mathbf{M}_1\mathbf{P}_0^0\mathbf{M}_1^{\mathrm{T}}\mathbf{H}^{\mathrm{T}})^{-1}(\mathbf{y}_1 - \mathbf{H}\mathbf{M}_1\mathbf{x}_0^0) \tag{16}$$

as solution of the 4D-Var (9). If we next integrate this solution $\mathbf{x}_1^0$ up to the first time-step, we find

$$\mathbf{x}_1^1 = \mathbf{M}_1\mathbf{x}_0^0 + \mathbf{M}_1\mathbf{P}_0^0\mathbf{M}_1^{\mathrm{T}}\mathbf{H}^{\mathrm{T}} (\mathbf{R} + \mathbf{H}\mathbf{M}_1\mathbf{P}_0^0\mathbf{M}_1^{\mathrm{T}}\mathbf{H}^{\mathrm{T}})^{-1}(\mathbf{y}_1 - \mathbf{H}\mathbf{M}_1\mathbf{x}_0^0).$$

We can thus conclude from Algorithm 1 that $\mathbf{x}_1^1 = \mathbf{x}_1^a$ since $\mathbf{M}_1\mathbf{x}_0^0 = \mathbf{M}_1\mathbf{x}_0^a = \mathbf{x}_1^f$, $\mathbf{M}_1\mathbf{P}_0^0\mathbf{M}_1^{\mathrm{T}} = \mathbf{M}_1\mathbf{P}_0^a\mathbf{M}_1^{\mathrm{T}} = \mathbf{P}_1^f$ and since we use the same observation $\mathbf{y}_1$. Moreover, observe from equation (10) that $\mathbf{W}$, as defined in (11), is the Hessian matrix of J_1. We can thus conclude that the covariance matrix of $\mathbf{x}_1^0$ is given by

$$\mathbf{P}_1^0 = \mathbf{W}^{-1}. \tag{17}$$

We have that the covariance matrix of $\mathbf{x}_1^1$ is given by

$$\begin{aligned}\mathbf{P}_1^1 &= \mathrm{Cov}(\mathbf{x}_1^1, \mathbf{x}_1^1) = \mathrm{Cov}(\mathbf{M}_1\mathbf{x}_1^0, \mathbf{M}_1\mathbf{x}_1^0) \\ &= \mathbf{M}_1\mathrm{Cov}(\mathbf{x}_1^0, \mathbf{x}_1^0)\mathbf{M}_1^{\mathrm{T}} = \mathbf{M}_1\mathbf{P}_1^0\mathbf{M}_1^{\mathrm{T}}.\end{aligned}$$

By (13) and (17), we deduce from this last expression that

$$\begin{aligned}\mathbf{P}_1^1 = \mathbf{M}_1\mathbf{P}_0^0\mathbf{M}_1^{\mathrm{T}} - \mathbf{M}_1\mathbf{P}_0^0\mathbf{M}_1^{\mathrm{T}}\mathbf{H}^{\mathrm{T}} \\ (\mathbf{R} + \mathbf{H}\mathbf{M}_1\mathbf{P}_0^0\mathbf{M}_1^{\mathrm{T}}\mathbf{H}^{\mathrm{T}})^{-1}\mathbf{H}\mathbf{M}_1\mathbf{P}_0^0\mathbf{M}_1^{\mathrm{T}},\end{aligned}$$

which corresponds to the analysis error covariance matrix $\mathbf{P}_1^a$ produced by the Kalman filter (see Algorithm 1), since $\mathbf{M}_1\mathbf{P}_0^0\mathbf{M}_1^{\mathrm{T}} = \mathbf{P}_1^f$.
Now, we suppose that when the j first observations vectors are available, the solution $\mathbf{x}_j^0$ of the corresponding 4D-Var problem

$$\min_{\mathbf{x}\in\mathbb{R}^n} J_j(\mathbf{x}) = \frac{1}{2}||\mathbf{x} - \mathbf{x}_0^0||^2_{(\mathbf{P}_0^0)^{-1}} + \frac{1}{2}\sum_{i=1}^{j}||\mathbf{y}_i - \mathbf{H}\mathbf{M}_i\mathbf{x}||^2_{\mathbf{R}^{-1}}$$

satisfies $\mathbf{M}_j\mathbf{x}_j^0 = \mathbf{x}_j^j = \mathbf{x}_j^a$ and $\mathbf{P}_j^j = \mathbf{P}_j^a$. We also assume that the Hessian matrix of J_j is $(\mathbf{P}_j^0)^{-1}$, allowing us to express the quadratic function J_j using its Taylor expansion around the solution $\mathbf{x}_j^0$ as

$$J_j(\mathbf{x}) = J_j(\mathbf{x}_j^0) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_j^0)^{\mathrm{T}}(\mathbf{P}_j^0)^{-1}(\mathbf{x} - \mathbf{x}_j^0).$$

We add the next available observation vector $\mathbf{y}_{j+1}$ to this expression and obtain the expression of the 4D-Var problem using the first $j+1$ observation vectors

$$\begin{aligned}\min_{\mathbf{x}\in\mathbb{R}^n} J_{j+1}(\mathbf{x}) = J_j(\mathbf{x}_j^0) + \frac{1}{2}||\mathbf{x} - \mathbf{x}_j^0||^2_{(\mathbf{P}_j^0)^{-1}} \\ + \frac{1}{2}||\mathbf{y}_{j+1} - \mathbf{H}\mathbf{M}_{j+1}\mathbf{x}||^2_{\mathbf{R}^{-1}}.\end{aligned} \tag{18}$$

It remains to prove that $\mathbf{x}_{j+1}^{j+1} = \mathbf{x}_{j+1}^a$ and that $\mathbf{P}_{j+1}^{j+1} = \mathbf{P}_{j+1}^a$. Observing that the first term of (18) is constant, we can apply the same reasoning as

the one used to deduce (16) and (17) from (9) to write the solution of (18) as

$$\mathbf{x}_{j+1}^0 = \mathbf{x}_j^0 + \mathbf{P}_j^0 \mathbf{M}_{j+1}^{\mathrm{T}} \mathbf{H}^{\mathrm{T}} (\mathbf{R} + \mathbf{H}\mathbf{M}_{j+1}\mathbf{P}_j^0\mathbf{M}_{j+1}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}})^{-1}(\mathbf{y}_{j+1} - \mathbf{H}\mathbf{M}_{j+1}\mathbf{x}_j^0) \quad (19)$$

with the covariance matrix

$$\mathbf{P}_{j+1}^0 = \left((\mathbf{P}_j^0)^{-1} + \mathbf{M}_{j+1}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}\mathbf{M}_{j+1}\right)^{-1}. \quad (20)$$

If we integrate the solution $\mathbf{x}_{j+1}^0$ up to the time-step t_{j+1}, we have

$$\mathbf{x}_{j+1}^{j+1} = \mathbf{M}_{j+1}\mathbf{x}_j^0 + \mathbf{M}_{j+1}\mathbf{P}_j^0\mathbf{M}_{j+1}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}} (\mathbf{R} + \mathbf{H}\mathbf{M}_{j+1}\mathbf{P}_j^0\mathbf{M}_{j+1}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}})^{-1}(\mathbf{y}_{j+1} - \mathbf{H}\mathbf{M}_{j+1}\mathbf{x}_j^0).$$

We can thus conclude, from Algorithm 1 and the recurrence assumptions, that $\mathbf{x}_{j+1}^{j+1} = \mathbf{x}_{j+1}^a$, since we have

$$\begin{aligned}\mathbf{M}_{j+1}\mathbf{x}_j^0 &= \mathbf{M}_{j+1,j}\mathbf{M}_j\mathbf{x}_j^0 = \mathbf{M}_{j+1,j}\mathbf{x}_j^j \\ &= \mathbf{M}_{j+1,j}\mathbf{x}_j^a = \mathbf{x}_{j+1}^f,\end{aligned}$$

and

$$\begin{aligned}\mathbf{M}_{j+1}\mathbf{P}_j^0\mathbf{M}_{j+1}^{\mathrm{T}} &= \mathbf{M}_{j+1,j}\mathbf{M}_j\mathbf{P}_j^0\mathbf{M}_j^{\mathrm{T}}\mathbf{M}_{j+1,j}^{\mathrm{T}} \\ &= \mathbf{M}_{j+1,j}\mathbf{P}_j^j\mathbf{M}_{j+1,j}^{\mathrm{T}} \\ &= \mathbf{M}_{j+1,j}\mathbf{P}_j^a\mathbf{M}_{j+1,j}^{\mathrm{T}} \\ &= \mathbf{P}_{j+1}^f, \end{aligned} \quad (21)$$

and the same observation $\mathbf{y}_{j+1}$. The covariance matrix of $\mathbf{x}_{j+1}^{j+1}$ is given by

$$\begin{aligned}\mathbf{P}_{j+1}^{j+1} &= \mathrm{Cov}(\mathbf{M}_{j+1}\mathbf{x}_{j+1}^0, \mathbf{M}_{j+1}\mathbf{x}_{j+1}^0) \\ &= \mathbf{M}_{j+1}\mathrm{Cov}(\mathbf{x}_{j+1}^0, \mathbf{x}_{j+1}^0)\mathbf{M}_{j+1}^{\mathrm{T}} \\ &= \mathbf{M}_{j+1}\mathbf{P}_{j+1}^0\mathbf{M}_{j+1}^{\mathrm{T}}.\end{aligned}$$

Replacing $\mathbf{P}_{j+1}^0$ by its expression (20) and using (13) (with subscripts j and $j+1$ instead of 0 and 1, respectively), we obtain

$$\mathbf{P}_{j+1}^{j+1} = \mathbf{M}_{j+1}\mathbf{P}_j^0\mathbf{M}_{j+1}^{\mathrm{T}} - \mathbf{M}_{j+1}\mathbf{P}_j^0\mathbf{M}_{j+1}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}} (\mathbf{R} + \mathbf{H}\mathbf{M}_{j+1}\mathbf{P}_j^0\mathbf{M}_{j+1}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}})^{-1}\mathbf{H}\mathbf{M}_{j+1}\mathbf{P}_j^0\mathbf{M}_{j+1}. \quad (22)$$

Using (21) in (22), we obtain $\mathbf{P}_{j+1}^{j+1} = \mathbf{P}_{j+1}^a$ which ends the proof. □

Note that the proof of this theorem shows (see equations (19) and (20)) how to update the solution of the 4D-Var and its covariance matrix incrementally when observations are obtained sequentially, without processing all the information from the beginning.

3.2 Connection between reduced 4D-Var / SEEK filter

Connection between reduced 4D-Var and the SEEK filter can be proved when the background covariance matrix $\mathbf{B}$ is defined using the sample covariance matrix $\mathbf{S}$ computed from a set of state vectors. In this specific case, we write the reduced 4D-Var problem as

$$\min_{\underline{\mathbf{x}}\in\mathbb{R}^r} \underline{J}(\underline{\mathbf{x}}) = \frac{1}{2}||\mathbf{L}_0\underline{\mathbf{x}} - \mathbf{x}^b||^2_{\mathbf{S}^{-1}} + \frac{1}{2}\sum_{i=1}^{N}||\mathbf{y}_i - \mathbf{H}_i\mathbf{M}_{i,0}\mathbf{L}_0\underline{\mathbf{x}}||^2_{\mathbf{R}_i^{-1}}. \quad (23)$$

Before proving that this problem is equivalent to the SEEK filter, we have to reformulate it in another way. This is the purpose of the following lemma.

Lemma 1 *The solution of the reduced 4D-Var (23) is equal to the solution of*

$$\min_{\underline{\mathbf{x}}\in\mathbb{R}^r} \underline{J}(\underline{\mathbf{x}}) = \frac{1}{2}||\underline{\mathbf{x}} - \underline{\mathbf{x}}^b||^2_{\mathbf{U}_0^{-1}} + \frac{1}{2}\sum_{i=1}^{N}||\mathbf{y}_i - \mathbf{H}\mathbf{M}_{i,0}\mathbf{L}_0\underline{\mathbf{x}}||^2_{\mathbf{R}_i^{-1}}, \quad (24)$$

where $\underline{\mathbf{x}}^b = \mathbf{L}_0^{\mathrm{T}}\mathbf{x}^b$ *is the reduced background and* $\mathbf{U}_0^{-1}$ *is the reduced background covariance matrix which is a diagonal matrix containing the inverse of the* r *largest eigenvalues of* $\mathbf{S}$*.*

Proof. Since the columns of $[\mathbf{L}_0\ \hat{\mathbf{L}}_0]$ form an orthonormal basis of $\mathbb{R}^n$, we have the equality $[\mathbf{L}_0\ \hat{\mathbf{L}}_0][\mathbf{L}_0\ \hat{\mathbf{L}}_0]^{\mathrm{T}} = \mathbf{I}_n$ (see Meyer, 2000, p.320) which can be rewritten as $\mathbf{L}_0\mathbf{L}_0^{\mathrm{T}} + \hat{\mathbf{L}}_0\hat{\mathbf{L}}_0^{\mathrm{T}} = \mathbf{I}_n$ and used to decompose the background vector $\mathbf{x}^b$ as

$$\mathbf{x}^b = \mathbf{L}_0\mathbf{L}_0^{\mathrm{T}}\mathbf{x}^b + \hat{\mathbf{L}}_0\hat{\mathbf{L}}_0^{\mathrm{T}}\mathbf{x}^b. \quad (25)$$

Substituting (25) in the background term of (23), we obtain

$$||\mathbf{L}_0\underline{\mathbf{x}} - \mathbf{x}^b||^2_{\mathbf{S}^{-1}} = ||\mathbf{L}_0(\underline{\mathbf{x}} - \mathbf{L}_0^{\mathrm{T}}\mathbf{x}^b) + \hat{\mathbf{L}}_0\hat{\mathbf{L}}_0^{\mathrm{T}}\mathbf{x}^b||^2_{\mathbf{S}^{-1}}$$

which can be reformulated as

$$||\mathbf{L}_0(\underline{\mathbf{x}} - \mathbf{L}_0^{\mathrm{T}}\mathbf{x}^b)||^2_{\mathbf{S}^{-1}} + ||\hat{\mathbf{L}}_0\hat{\mathbf{L}}_0^{\mathrm{T}}\mathbf{x}^b||^2_{\mathbf{S}^{-1}} + 2(\mathbf{L}_0(\underline{\mathbf{x}} - \mathbf{L}_0^{\mathrm{T}}\mathbf{x}^b))^{\mathrm{T}}\mathbf{S}^{-1}(\hat{\mathbf{L}}_0\hat{\mathbf{L}}_0^{\mathrm{T}}\mathbf{x}^b).$$

We observe that the third term is equal to zero, using (5) and the decomposition of $\mathbf{S}^{-1}$ in the basis $[\mathbf{L}_0\ \hat{\mathbf{L}}_0]$

$$\mathbf{S}^{-1} = \mathbf{L}_0\mathbf{U}_0^{-1}\mathbf{L}_0^{\mathrm{T}} + \hat{\mathbf{L}}_0\hat{\mathbf{U}}_0^{-1}\hat{\mathbf{L}}_0^{\mathrm{T}}, \quad (26)$$

and the equalities $\hat{\mathbf{L}}_0^{\mathrm{T}}\hat{\mathbf{L}}_0 = \mathbf{I}_{n-r}$ and $\mathbf{L}_0^{\mathrm{T}}\hat{\mathbf{L}}_0 = 0$. The second term is constant and can thus be ignored for the minimization. Finally, we can reformulate the first term as

$$||\underline{\mathbf{x}} - \mathbf{L}_0^{\mathrm{T}}\mathbf{x}^b||^2_{\mathbf{L}_0^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{L}_0}$$

which is equal to

$$||\underline{\mathbf{x}} - \underline{\mathbf{x}}^b||^2_{\mathbf{U}_0^{-1}}, \tag{27}$$

by (26). It concludes the proof since the background term of (23) is equal to (27) plus a constant term. □

We can now prove the main theoretical result of this paper, i.e., the equivalence between the SEEK filter and a specific reduced 4D-Var problem. To continue to be as concise as possible in what follows, we set $\mathbf{M}_i = \mathbf{M}_{i,0}$, $\mathbf{H} = \mathbf{H}_i$ and $\mathbf{R} = \mathbf{R}_i$.

Theorem 2 *Suppose that the model operator are perfect and linear, that the observation operator are linear and that the same background* $\mathbf{x}^b$*, observations* $\mathbf{y}_i$*, background covariance matrix* $\mathbf{S}$*, and observation covariance matrices* $\mathbf{R}_i$ *are given. Assume moreover that the initial analysis of the SEEK filter* $\mathbf{x}_0^a$ *is equal to* $\mathbf{L}_0\mathbf{L}_0^{\mathrm{T}}\mathbf{x}^b$ *and that its covariance matrix* $\mathbf{P}_0^a$ *is equal to* $\mathbf{L}_0\mathbf{U}_0\mathbf{L}_0^{\mathrm{T}}$*. Then the analysis state* $\mathbf{x}_i^a$ *computed at time* t_i *by the SEEK filter and the solution produced by the reduced 4D-Var (23), using the first i observations, prolongated in the full space and integrated up to time* t_i*, are identical and have same covariance matrices. Both methods hence produce the same results at the end of the assimilation period.*

Proof. We have to prove that the reduced 4D-Var (23) is equivalent to a SEEK filter. Using Lemma 1, it amounts to prove that (24) is equivalent to the SEEK filter. At this aim, we use the same scheme as in the proof of Theorem 1. We first introduce the notation $\underline{\mathbf{x}}_l^0$ to denote the solution of the reduced 4D-Var (24) using the first l observations, and $\underline{\mathbf{P}}_l^0$ to denote the covariance matrix of $\underline{\mathbf{x}}_l^0$. Moreover, we introduce the notation $\mathbf{x}_l^k$ to denote the prolongation solution of $\underline{\mathbf{x}}_l^0$ integrated up to time t_k and $\mathbf{P}_l^k$ to denote the covariance matrix of $\mathbf{x}_l^k$. We thus want to prove that the solution of the reduced 4D-Var (24) using the first l observations, prolongated and integrated up to time t_l is equal to the analysis state produced by the SEEK filter after l iterations, i.e., $\mathbf{x}_l^l = \mathbf{x}_l^a$, and that both state vectors have the same covariance matrices, i.e., $\mathbf{P}_l^l = \mathbf{P}_l^a$, for each observation time $l = 1, \ldots, N$. The proof is by induction.
To prove that $\mathbf{x}_1^1 = \mathbf{x}_1^a$ and $\mathbf{P}_1^1 = \mathbf{P}_1^a$, we first notice that $\mathbf{x}_0^0 = \mathbf{L}_0\underline{\mathbf{x}}_0^0 = \mathbf{L}_0\underline{\mathbf{x}}^b = \mathbf{L}_0\mathbf{L}_0^{\mathrm{T}}\mathbf{x}^b = \mathbf{x}_0^a$, since $\underline{\mathbf{x}}^b$ is the solution of the reduced 4D-Var problem using no observations and from the definition of $\underline{\mathbf{x}}^b$ in Lemma 1 and the

assumption on $\mathbf{x}_0^a$. We also note that

$$\begin{aligned}\mathbf{P}_0^0 &= \mathrm{Cov}(\mathbf{x}_0^0, \mathbf{x}_0^0) = \mathrm{Cov}(\mathbf{L}_0\underline{\mathbf{x}}_0^0, \mathbf{L}_0\underline{\mathbf{x}}_0^0)\\ &= \mathbf{L}_0\mathrm{Cov}(\underline{\mathbf{x}}_0^0, \underline{\mathbf{x}}_0^0)\mathbf{L}_0^{\mathrm{T}}\\ &= \mathbf{L}_0\underline{\mathbf{P}}_0^0\mathbf{L}_0^{\mathrm{T}} = \mathbf{L}_0\mathbf{U}_0\mathbf{L}_0^{\mathrm{T}} = \mathbf{P}_0^a,\end{aligned}$$

since $\mathbf{U}_0$ is the Hessian inverse and thus the covariance matrix of the solution $\underline{\mathbf{x}}_0^0$. We next define the reduced 4D-var problem when only one observation vector is available at the first time-step as

$$\min_{\underline{\mathbf{x}}\in\mathbb{R}^r} \underline{J}_1(\underline{\mathbf{x}}) = \frac{1}{2}||\underline{\mathbf{x}} - \underline{\mathbf{x}}_0^0||^2_{(\underline{\mathbf{P}}_0^0)^{-1}} + \frac{1}{2}||\mathbf{y}_1 - \mathbf{H}\mathbf{M}_1\mathbf{L}_0\underline{\mathbf{x}}||^2_{\mathbf{R}^{-1}}.$$

Using the equivalent of equations (11), (16) and (17) from Theorem 1 adapted to the reduced space, we obtain as solution

$$\underline{\mathbf{x}}_1^0 = \underline{\mathbf{x}}_0^0 + \underline{\mathbf{P}}_1^0\mathbf{L}_0^{\mathrm{T}}\mathbf{M}_1^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{y}_1 - \mathbf{H}\mathbf{M}_1\mathbf{L}_0\underline{\mathbf{x}}_0^0) \quad (28)$$

where

$$\underline{\mathbf{P}}_1^0 = \left((\underline{\mathbf{P}}_0^0)^{-1} + \mathbf{L}_0^{\mathrm{T}}\mathbf{M}_1^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}\mathbf{M}_1\mathbf{L}_0\right)^{-1} \quad (29)$$

is its covariance matrix. We observe that the inverse of $\underline{\mathbf{P}}_1^0$ is the Hessian matrix of $\underline{J}_1$. If we formulate this reduced solution in the full space $\mathbb{R}^n$ and integrate it up to the first time-step, we obtain

$$\mathbf{x}_1^1 = \mathbf{M}_1\mathbf{L}_0\underline{\mathbf{x}}_0^1 = \mathbf{M}_1\mathbf{L}_0\underline{\mathbf{x}}_0^0 + \mathbf{M}_1\mathbf{L}_0\underline{\mathbf{P}}_1^0\mathbf{L}_0^{\mathrm{T}}\mathbf{M}_1^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{y}_1 - \mathbf{H}\mathbf{M}_1\mathbf{L}_0\underline{\mathbf{x}}_0^0),$$

and find back the solution of Algorithm 2 after one iteration. Indeed, by the algorithm, we have $\mathbf{M}_1\mathbf{L}_0\underline{\mathbf{x}}_0^0 = \mathbf{M}_1\mathbf{x}_0^0 = \mathbf{M}_1\mathbf{x}_0^a = \mathbf{x}_1^f$ and $\underline{\mathbf{P}}_1^0 = \mathbf{U}_1$, and we use the same observation $\mathbf{y}_1$. Moreover, the covariance matrix of $\mathbf{x}_1^1$ is given by

$$\begin{aligned}\mathbf{P}_1^1 &= \mathrm{Cov}(\mathbf{x}_1^1, \mathbf{x}_1^1) = \mathrm{Cov}(\mathbf{M}_1\mathbf{L}_0\underline{\mathbf{x}}_1^0, \mathbf{M}_1\mathbf{L}_0\underline{\mathbf{x}}_1^0)\\ &= \mathbf{M}_1\mathbf{L}_0\mathrm{Cov}(\underline{\mathbf{x}}_1^0, \underline{\mathbf{x}}_1^0)\mathbf{L}_0^{\mathrm{T}}\mathbf{M}_1^{\mathrm{T}}\\ &= \mathbf{M}_1\mathbf{L}_0\underline{\mathbf{P}}_1^0\mathbf{L}_0^{\mathrm{T}}\mathbf{M}_1^{\mathrm{T}} = \mathbf{L}_1\mathbf{U}_1\mathbf{L}_1^{\mathrm{T}} = \mathbf{P}_1^a,\end{aligned}$$

which is the analysis error covariance matrix given by the SEEK filter after one iteration.

Now, we suppose that when the j first observation vectors are available, the solution $\underline{\mathbf{x}}_j^0$ of the corresponding reduced 4D-Var problem

$$\min_{\underline{\mathbf{x}}\in\mathbb{R}^r} \underline{J}_j(\underline{\mathbf{x}}) = \frac{1}{2}||\underline{\mathbf{x}} - \underline{\mathbf{x}}^b||^2_{(\underline{\mathbf{P}}_0^0)^{-1}} + \frac{1}{2}\sum_{i=1}^{j}||\mathbf{y}_i - \mathbf{H}\mathbf{M}_i\mathbf{L}_0\underline{\mathbf{x}}||^2_{\mathbf{R}^{-1}}$$

satisfies $\mathbf{M}_j\mathbf{L}_0\underline{\mathbf{x}}_j^0 = \mathbf{x}_j^j = \mathbf{x}_j^a$ and $\mathbf{P}_j^j = \mathbf{P}_j^a$. We also assume that $\underline{\mathbf{P}}_j^0 = \mathbf{U}_j$, whose inverse equals the Hessian matrix of $\underline{J}_j$. It allows us to express the quadratic function $\underline{J}_j$ using its Taylor expansion around the solution $\underline{\mathbf{x}}_j^0$ as

$$\underline{J}_j(\underline{\mathbf{x}}) = \underline{J}_j(\underline{\mathbf{x}}_j^0) + \frac{1}{2}(\underline{\mathbf{x}} - \underline{\mathbf{x}}_j^0)^{\mathrm{T}}(\underline{\mathbf{P}}_j^0)^{-1}(\underline{\mathbf{x}} - \underline{\mathbf{x}}_j^0).$$

We add the next available observation vector $\mathbf{y}_{j+1}$ to this expression and obtain the expression of the reduced 4D-Var problem using the first $j+1$ observation vectors

$$\min_{\underline{\mathbf{x}}\in\mathbb{R}^r} \underline{J}_{j+1}(\underline{\mathbf{x}}) = \underline{J}_j(\underline{\mathbf{x}}_j^0) + \frac{1}{2}||\underline{\mathbf{x}} - \underline{\mathbf{x}}_j^0||^2_{(\underline{\mathbf{P}}_j^0)^{-1}} + \frac{1}{2}||\mathbf{y}_{j+1} - \mathbf{H}\mathbf{M}_{j+1}\mathbf{L}_0\underline{\mathbf{x}}||^2_{\mathbf{R}^{-1}}.$$

The solution of this problem, based on the same reasoning used to derive (28) and (29), writes

$$\underline{\mathbf{x}}_{j+1}^0 = \underline{\mathbf{x}}_j^0 + \underline{\mathbf{P}}_{j+1}^0\mathbf{L}_0^{\mathrm{T}}\mathbf{M}_{j+1}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{y}_{j+1} - \mathbf{H}\mathbf{M}_{j+1}\mathbf{L}_0\underline{\mathbf{x}}_j^0)$$

with the covariance matrix

$$\underline{\mathbf{P}}_{j+1}^0 = \left((\underline{\mathbf{P}}_j^0)^{-1} + \mathbf{L}_0^{\mathrm{T}}\mathbf{M}_{j+1}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}\mathbf{M}_{j+1}\mathbf{L}_0\right)^{-1}.$$

By Algorithm 2, we remark that $\underline{\mathbf{P}}_{j+1}^0 = \mathbf{U}_{j+1}$ since $\underline{\mathbf{P}}_j^0 = \mathbf{U}_j$. If we formulate the solution $\underline{\mathbf{x}}_{j+1}^0$ in the full space and integrate it up to the time-step $j+1$, we have

$$\mathbf{x}_{j+1}^{j+1} = \mathbf{M}_{j+1}\mathbf{L}_0\underline{\mathbf{x}}_{j+1}^0 = \mathbf{M}_{j+1}\mathbf{L}_0\underline{\mathbf{x}}_j^0 + \mathbf{M}_{j+1}\mathbf{L}_0\underline{\mathbf{P}}_{j+1}^0\mathbf{L}_0^{\mathrm{T}}\mathbf{M}_{j+1}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{y}_{j+1} - \mathbf{H}\mathbf{M}_{j+1}\mathbf{L}_0\underline{\mathbf{x}}_j^0).$$

Using the recurrence assumptions, we find back the Kalman filter solution after $j+1$ iterations since

$$\begin{aligned}\mathbf{M}_{j+1}\mathbf{L}_0\underline{\mathbf{x}}_j^0 &= \mathbf{M}_{j+1,j}\mathbf{M}_j\mathbf{L}_0\underline{\mathbf{x}}_j^0 = \mathbf{M}_{j+1,j}\mathbf{x}_j^j \\ &= \mathbf{M}_{j+1,j}\mathbf{x}_j^a = \mathbf{x}_{j+1}^f,\end{aligned}$$

by Algorithm 2, and since

$$\begin{aligned}\mathbf{P}_{j+1}^{j+1} &= \mathrm{Cov}(\mathbf{x}_{j+1}^{j+1}, \mathbf{x}_{j+1}^{j+1}) \\ &= \mathrm{Cov}(\mathbf{M}_{j+1}\mathbf{L}_0\underline{\mathbf{x}}_{j+1}^0, \mathbf{M}_{j+1}\mathbf{L}_0\underline{\mathbf{x}}_{j+1}^0) \\ &= \mathbf{M}_{j+1}\mathbf{L}_0\mathrm{Cov}(\underline{\mathbf{x}}_{j+1}^0, \underline{\mathbf{x}}_{j+1}^0)\mathbf{L}_0^{\mathrm{T}}\mathbf{M}_{j+1}^{\mathrm{T}} \\ &= \mathbf{M}_{j+1}\mathbf{L}_0\underline{\mathbf{P}}_{j+1}^0\mathbf{L}_0^{\mathrm{T}}\mathbf{M}_{j+1}^{\mathrm{T}} \\ &= \mathbf{L}_{j+1}\mathbf{U}_{j+1}\mathbf{L}_{j+1}^{\mathrm{T}} \\ &= \mathbf{P}_{j+1}^a,\end{aligned}$$

again by Algorithm 2 and by $\underline{\mathbf{P}}^0_{j+1} = \mathbf{U}_{j+1}$. □

4 Enhanced Gauss-Newton method for the 4D-Var

In atmospheric and oceanographic applications the model and observation operators are nonlinear and thus the 4D-Var (6) is a nonlinear least-squares problem. A common method for solving this problem is the Gauss-Newton algorithm (see, Bjorck (1996) and Nocedal and Wright (2006)). To present it, we reformulate the 4D-Var problem as

$$\min_{\mathbf{x}\in\mathbb{R}^n} J(\mathbf{x}) = \frac{1}{2}[\mathbf{x}-\mathbf{x}^b]^{\mathrm{T}}\mathbf{B}^{-1}[\mathbf{x}-\mathbf{x}^b] + \frac{1}{2}[\mathcal{G}(\mathbf{x})-\mathbf{y}]^{\mathrm{T}}\mathbf{R}^{-1}[\mathcal{G}(\mathbf{x})-\mathbf{y}]. \quad (30)$$

using the following observation vector $\mathbf{y} = ((\mathbf{y}_1)^{\mathrm{T}}, \ldots, (\mathbf{y}_N)^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^m$ where $\mathbf{y}_i \in \mathbb{R}^{m_i}$ is the observation vector at time i, and $\sum_{i=1}^N m_i = m$. The covariance matrix $\mathbf{R} \in \mathbb{R}^{m\times m}$ is formed with each $\mathbf{R}_i$ for $i = 1, \ldots, N$. The operator $\mathcal{G} : \mathbb{R}^n \to \mathbb{R}^m$ is mapping the initial condition into the space of the observation vector and is given by

$$\mathcal{G}(\mathbf{x}) = \begin{pmatrix} \mathcal{H}_1\mathcal{M}_{1,0}\mathbf{x} \\ \vdots \\ \mathcal{H}_i\mathcal{M}_{i,0}\mathbf{x} \\ \vdots \\ \mathcal{H}_N\mathcal{M}_{N,0}\mathbf{x} \end{pmatrix}.$$

where the observation and model operator are assumed to be nonlinear. We can rewrite the problem (30) in a more compact form

$$\min_{\mathbf{x}\in\mathbb{R}^n} J(\mathbf{x}) = \frac{1}{2}||\mathbf{f}(\mathbf{x})||_2^2 \quad (31)$$

where

$$\mathbf{f}(\mathbf{x}) = \begin{pmatrix} \mathbf{B}^{-1/2}(\mathbf{x}-\mathbf{x}^b) \\ \mathbf{R}^{-1/2}(\mathcal{G}(\mathbf{x})-\mathbf{y}) \end{pmatrix}.$$

The matrices $\mathbf{B}^{-1/2}$ and $\mathbf{R}^{-1/2}$ are the inverse of the square-root factors of $\mathbf{B} = \mathbf{B}^{1/2}(\mathbf{B}^{1/2})^{\mathrm{T}}$ and $\mathbf{R} = \mathbf{R}^{1/2}(\mathbf{R}^{1/2})^{\mathrm{T}}$.

In the Gauss-Newton problem, the solution of the nonlinear least-squares problem (31) is computed through solving a sequence of linear least-squares problems

$$\min_{\delta\mathbf{x}_k\in\mathbb{R}^n} \frac{1}{2}||\mathbf{F}_k\delta\mathbf{x}_k + \mathbf{f}(\mathbf{x}_k)||_2^2, \quad (32)$$

where $\mathbf{F}_k$ is the Jacobian of f at $\mathbf{x}_k$ and is given by

$$\mathbf{F}_k = \begin{pmatrix} \mathbf{B}^{-1/2} \\ \mathbf{R}^{-1/2}\mathbf{G}_k \end{pmatrix},$$

with $\mathbf{G}_k$ the Jacobian matrix of $\mathcal{G}$ at $\mathbf{x}_k$. The minimum of the linear lest-squares problem can be computed by nullifying its gradient. Thus, we have to solve a sequence of positive definite linear system

$$\mathbf{F}_k^{\mathrm{T}}\mathbf{F}_k\delta\mathbf{x}_k = -\mathbf{F}_k^{\mathrm{T}}\mathbf{f}(\mathbf{x}_k)$$

known as the normal equation. We rewrite the system in a more compact form

$$\mathbf{A}_k\delta\mathbf{x}_k = \mathbf{b}_k \tag{33}$$

where $\mathbf{A}_k = \mathbf{B}^{-1}+\mathbf{G}_k^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{G}_k$ is a symmetric positive definite approximation of the hessian matrix of J without considering the second-order terms and where $\mathbf{b}_k = \mathbf{B}^{-1}(\mathbf{x}^b - \mathbf{x}_k) + \mathbf{G}_k^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{y} - \mathcal{G}(\mathbf{x}_k))$ is the gradient of J. A new approximation of the solution of (31) is computed using the update formula $\mathbf{x}_{k+1} = \mathbf{x}_k + \delta\mathbf{x}_k$. A nonlinear model integration is then performed from this new approximation to update the reference trajectory and compute the new value of the cost function $\mathbf{f}(\mathbf{x}_{k+1})$. We resume the Gauss-Newton method in Algorithm 3. The linear least-squares problems can be viewed as a sequence

Algorithm 3 Gauss-Newton

Set $\mathbf{x}_0 = \mathbf{x}^b$
for k=0,1,2,... **do**
  Solve $\min_{\delta\mathbf{x}_k\in\mathbb{R}^n} \frac{1}{2}||\mathbf{F}_k\delta\mathbf{x}_k + \mathbf{f}(\mathbf{x}_k)||_2^2$
  Set $\mathbf{x}_{k+1} = \mathbf{x}_k + \delta\mathbf{x}_k$
end for

of quadratic problems which approximate the nonlinear function. They can be rewritten into the well-known incremental 4D-Var formulation

$$\min_{\delta\mathbf{x}_k\in\mathbb{R}^n} \frac{1}{2}(\delta\mathbf{x}_k - \mathbf{d}^{bk})^{\mathrm{T}}\mathbf{B}^{-1}(\delta\mathbf{x}_k - \mathbf{d}^{bk}) + \frac{1}{2}(\mathbf{G}_k\delta\mathbf{x}_k - \mathbf{d}^{ok})^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{G}_k\delta\mathbf{x}_k - \mathbf{d}^{ok}) \tag{34}$$

where $\mathbf{d}^{bk} = \mathbf{x}^b - \mathbf{x}_k$ and $\mathbf{d}^{ok} = \mathbf{y} - \mathcal{G}(\mathbf{x}_k)$ are departure vectors.

There are direct methods to solve (33) such as the Cholesky factorization of $\mathbf{A}_k$ or the QR factorization of $\mathbf{F}_k$. We opt for a conjugate-gradient-like (CG) method because it is better suited for large operational system (Weaver, Viliard, and Anderson, 2003). Moreover, we can not store the Hessians $\mathbf{A}_k$ and can only compute the product between them and any vector.

For computational cost considerations, we do not compute the exact Jacobian of $\mathbf{G}$ in the incremental 4D-Var (34) and perform only a limited number of outer Gauss-Newton iterations and a limited number of inner conjugate gradient iterations to reduce the computational cost. As pointed by Lawless, Gratton, and Nichols (2004) and Gratton, Lawless, and Nichols (2007), we find back the *incremental 4D-Var* introduced by Courtier, Thépaut, and Hollingsworth (1994). The convergence rate of the conjugate-gradient-like method is influenced by the starting point, the right-hand side, and the actual distribution of the eigenvalues (e.g. presence of eigenvalue clusters). Therefore, the development of good starting points and preconditioners is essential to obtain a better convergence rate.

To define an appropriate starting point for the conjugate-gradient-like method, we will derive a Gauss-Newton algorithm for the the reduced 4D-Var (8). This reduced Gauss-Newton method define a sequence of reduced space linear-least squares problem given by

$$\min_{\delta\underline{\mathbf{x}}_k \in \mathbb{R}^r} \frac{1}{2}||\mathbf{F}_k\mathbf{L}_0\delta\underline{\mathbf{x}}_k + \mathbf{f}(\mathbf{x}_k)||_2^2. \tag{35}$$

It leads to the sequence of positive definite linear system

$$\mathbf{L}_0^{\mathrm{T}}\mathbf{F}_k^{\mathrm{T}}\mathbf{F}_k\mathbf{L}_0\delta\underline{\mathbf{x}}_k = -\mathbf{L}_0^{\mathrm{T}}\mathbf{F}_k^{\mathrm{T}}\mathbf{f}(\mathbf{x}_k)$$

which can be written

$$\mathbf{L}_0^{\mathrm{T}}\mathbf{A}_k\mathbf{L}_0\delta\underline{\mathbf{x}}_k = \mathbf{L}_0^{\mathrm{T}}\mathbf{b}_k. \tag{36}$$

A new approximation of the solution is computed using the update formula $\underline{\mathbf{x}}_{k+1} = \underline{\mathbf{x}}_k + \delta\underline{\mathbf{x}}_k$. The corresponding approximation in the full space is given by $\mathbf{L}_0\underline{\mathbf{x}}_{k+1}$. A nonlinear model integration is then performed from this new approximation to update the reference trajectory and to compute $\mathbf{f}(\mathbf{x}_{k+1})$. This reduced version is summarized in Algorithm 4. As previously,

Algorithm 4 Reduced Gauss-Newton

Set $\mathbf{x}_0 = \mathbf{x}^b$
for k=0,1,2,... **do**
  Solve $\min_{\delta\underline{\mathbf{x}}_k \in \mathbb{R}^r} \frac{1}{2}||\mathbf{F}_k\mathbf{L}_0\delta\underline{\mathbf{x}}_k + \mathbf{f}(\mathbf{x}_k)||_2^2$
  Set $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{L}_0\delta\underline{\mathbf{x}}_k$
end for

the reduced linear-least squares problems (35) can be rewritten into the incremental formulation

$$\begin{aligned}\min_{\delta\underline{\mathbf{x}}_k \in \mathbb{R}^r} = \frac{1}{2}(\mathbf{L}_0\delta\underline{\mathbf{x}}_k - \mathbf{d}^{bk})^{\mathrm{T}}\mathbf{B}^{-1}(\mathbf{L}_0\delta\underline{\mathbf{x}}_k - \mathbf{d}^{bk}) \\ + \frac{1}{2}(\mathbf{G}_k\mathbf{L}_0\delta\underline{\mathbf{x}}_k - \mathbf{d}^{ok})^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{G}_k\mathbf{L}_0\delta\underline{\mathbf{x}}_k - \mathbf{d}^{ok}).\end{aligned} \tag{37}$$

and called reduced incremental 4D-Var problems. The solution of the reduced linear-least squares problems (35) in the full space is given by

$$\begin{aligned} \mathbf{L}_0 \delta \underline{\mathbf{x}}_k &= \mathbf{L}_0 (\mathbf{L}_0^{\mathrm{T}} \mathbf{A}_k \mathbf{L}_0)^{-1} \mathbf{L}_0^{\mathrm{T}} \mathbf{b}_k \\ &= \mathbf{L}_0 (\mathbf{L}_0^{\mathrm{T}} \mathbf{A}_k \mathbf{L}_0)^{-1} \mathbf{L}_0^{\mathrm{T}} \mathbf{A}_k \delta \mathbf{x}_k \end{aligned} \tag{38}$$

and is equivalent to the solution of the full linear-least squares problems (32) projected on $\mathbf{L}_0$ along the orthogonal of $\mathbf{A}_k \mathbf{L}_0$ (see Saad, 2008, for a review on projection). Hence, we can choose (38) as a starting point to solve the full linear-least squares problems. This staring point will be called the Ritz-Galerkin starting point with respect to the subspace $\mathbf{L}_0$ (van der Vorst, 2003).

As already mentioned, the convergence rate of the conjugate-gradient-like method is influenced by the distribution of the eigenvalues. Particularly, it is roughly bounded as a function of the condition number of $\mathbf{A}_k$ (see, for details, Nocedal and Wright (2006)). When this condition number is large, it is advisable to use a *preconditioned* version of the conjugate gradient method. It is widely recognized that there is no universal way to design a preconditioner for all type of problems (Benzi, 2002) and that designing preconditioners is a very broad and active area of research (van der Vorst, 2003). Ideally, the preconditioner must be a symmetric positive definite matrix which is cheap to compute easy to apply. It has to reduce the condition number or to produce a more clustered spectrum. The *limited memory preconditioner* (LMP) is a class of preconditioner defined by

$$\begin{aligned} \mathbf{H}_k = [\mathbf{I}_n - \mathbf{Z}(\mathbf{Z}^{\mathrm{T}} \mathbf{A}_k \mathbf{Z})^{-1} \mathbf{Z}^{\mathrm{T}} \mathbf{A}] \mathbf{M} \\ [\mathbf{I}_n - \mathbf{A}_k \mathbf{Z}(\mathbf{Z}^{\mathrm{T}} \mathbf{A}_k \mathbf{Z})^{-1} \mathbf{Z}^{\mathrm{T}}] + \mathbf{Z}(\mathbf{Z}^{\mathrm{T}} \mathbf{A}_k \mathbf{Z})^{-1} \mathbf{Z}^{\mathrm{T}} \end{aligned}$$

where the column of $\mathbf{Z} \in \mathbb{R}^{n \times r}$ are linearly independent and $\mathbf{M}$ plays the role of the first-level preconditioner (see, Gratton, Sartenaer, and Ilunga (2008)). The first-level preconditioner usually depends on the physics of the application. In variational data assimilation problems, it is equal to the background error covariance matrix B and is able to cluster most eigenvalues at 1. In order to improve the efficiency of this first-level preconditioner, the limited memory preconditioner use directions in a low dimensional subspace Z that are left out by the first-level preconditioner and slowing down the convergence of the conjugate-gradient-like method. Recently, Tshimanga et al. (2008) described a limited memory preconditioner techniques in a 4D-Var context. Their idea was to exploit information gained when solving one system to build the limited memory preconditioner for the next system of the sequence. The information could be, for example, Ritz pairs or descent directions. The drawback of this method is the impossibility of constructing a limited memory preconditioner for the first system because no information

is available.

5 New approach

Robert, Blayo, and Verron (2006) utilized the empirical orthogonal functions for the definition of the reduced control space and parametrization of the background error covariance matrix. They have developed a method to combine the full 4D-Var (6) and the reduced version of the 4D-Var (8). In this approach, an optimal correction in the full space is sought, but the problem of the computational cost is addressed by using a reduced 4D-Var, to provide a relevant guess for the full space minimization. Their approach consists in performing a few iterations to solve approximately the first reduced incremental 4D-Var (37) where the background error covariance matrix B is equal to the sample covariance matrix S

$$\min_{\delta\underline{\mathbf{x}}_0\in\mathbb{R}^r} \frac{1}{2}||\mathbf{L}_0\delta\underline{\mathbf{x}}_0 - \mathbf{d}^{b0}||^2_{\mathbf{S}^{-1}} + \frac{1}{2}||\mathbf{G}_0\mathbf{L}_0\delta\underline{\mathbf{x}}_k - \mathbf{d}^{o0}||^2_{\mathbf{R}^{-1}}. \quad (39)$$

Indeed, we can rewrite this problem to find back the same expression as in Robert et al. (2006)

$$\min_{\delta\underline{\mathbf{x}}_0\in\mathbb{R}^r} \frac{1}{2}||\mathbf{L}_0\delta\underline{\mathbf{x}}_0||^2_{\mathbf{L}_0\mathbf{U}_0^{-1}\mathbf{L}_0^T} + \frac{1}{2}||\mathbf{G}_0\mathbf{L}_0\delta\underline{\mathbf{x}}_0 - \mathbf{d}^{o0}||^2_{\mathbf{R}^{-1}}$$

using the equalities $\mathbf{d}^{b0} = 0$, $\mathbf{L}_0^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{L}_0 = \mathbf{U}_0^{-1}$ and $\mathbf{L}_0^T\mathbf{L}_0 = \mathbf{I}_r$. The prolongation of the approximate solution is then used as starting point of the Gauss-Newton method to solve approximately the full 4D-Var (6) where the background error covariance matrix B is used as preconditioner. They apply this reduced-order approach in the OPA model with its TDH configuration and the variational data assimilation package OPAVAR (Weaver et al., 2003). Using theorem 2, we can prove that the solution of (39) is equivalent to the solution of a SEEK filter where x_0^a is a zero vector, where P_0^a is given by $L_0U_0L_0^T$, where the i-th observation is given by the i-th component of $\mathbf{d}^{o0}$ and where the observation and model operators are the Jacobian of the nonlinear one at $\mathbf{x}_0$. This equivalence helps justify the introduction of this reduced incremental 4D-Var and the basis of their approach.

In this work, we present a new approach to solve data assimilation problem which joins up some previous presented ideas. We compute a solution

of the 4D-Var problem (6) using the Gauss-Newton algorithm with a preconditioned conjugate-gradient-like method. For the first outer iteration of the Gauss-Newton algorithm, we choose as starting point of the CG method a nonzero vector which is the Ritz-Galerkin starting point (38) with respect to the subspace $\mathbf{L}_0$ given by

$$\delta\mathbf{x}_0^0 = \mathbf{L}_0(\mathbf{L}_0^{\mathrm{T}}\mathbf{A}_0\mathbf{L}_0)^{-1}\mathbf{L}_0^{\mathrm{T}}\mathbf{b}_0. \tag{40}$$

A direct method compute the matrix inverse since its dimension is equal to r and is very small in apllications. This is the solution of the first reduced incremental 4D-Var (37) with the background error covariance matrix B prolongated into the full space. It would be a relevant starting point for the CG method since the subspace $\mathbf{L}_0$ contains the main directions of variability and since $\mathbf{L}_0$ shows its worth in the SEEK filter. We choose as preconditioner the limited memory preconditioner

$$\begin{aligned}\mathbf{H}_0 = [\mathbf{I}_n - \mathbf{L}_0(\mathbf{L}_0^{\mathrm{T}}\mathbf{A}_0\mathbf{L}_0)^{-1}\mathbf{L}_0^{\mathrm{T}}\mathbf{A}_0]\mathbf{B}& \\ [\mathbf{I}_n - \mathbf{A}_0\mathbf{L}_0(\mathbf{L}_0^{\mathrm{T}}\mathbf{A}_0\mathbf{L}_0)^{-1}\mathbf{L}_0^{\mathrm{T}}] + \mathbf{L}_0(\mathbf{L}_0^{\mathrm{T}}\mathbf{A}_0\mathbf{L}_0)^{-1}\mathbf{L}_0^{\mathrm{T}}.&\end{aligned} \tag{41}$$

This preconditioner is available from the first to the last system of the sequence. The cost to build this preconditioner is dominated by the product $\mathbf{A}_0\mathbf{L}_0$. Therefore, it is available without any extra computational cost since this product as been computed for the starting point (40). We assume that the Hessian matrix does not change significantly from one outer iteration to the next. Thus, we can keep unchanged this preconditioner for each system.

6 Numerical experiments

The model we wish to consider is the one-dimensional shallow-water system describing the flow of a fluid over an obstacle (see, Lawless (2001)). The governing equation can be written

$$\begin{aligned}\frac{Du}{Dt} + \frac{\partial\phi}{\partial z} &= -g\frac{\partial\bar{h}}{\partial z},\\ \frac{D\phi}{Dt} + \phi\frac{\partial u}{\partial z} &= 0,\end{aligned}$$

where $\frac{D}{Dt} = \frac{\partial}{\partial t} + u\frac{\partial}{\partial u}$ is the material derivative. In these equations $\bar{h} = \bar{h}(z)$ is the height of bottom orography, $u = u(z,t)$ is the velocity of the fluid and $\phi = gh(z,t)$ is the geopotential, where g is the gravitational constant and $h > 0$, the depth of the fluid above the orography. The problem is defined on the domain $z \in [0,l]$ with periodic boundary condition such that $z(0) = z(l)$ and we let $t \in [0,T]$. The values of u and ϕ are specified everywhere at the initial time, such that

$$\begin{aligned}u(z,0) &= u_0(z)\\ \phi(z,0) &= \phi_0(z).\end{aligned}$$

For time $t < 0$ the fluid is at rest and the geopotential ϕ is equal to $g(h_0 - \bar{h}(z))$, with h_0 constant. At $t = 0$, the fluid is impulsively set in motion with a constant velocity u_0 for all z. From this impulse a wave motion develops and moves away from the obstacle in both directions. In our example, the domain is defined to be periodic over 250 grid points, with a distance $\Delta z = 0.01m$ between them, so that $z \in [0m, 2.5m]$. The height of the

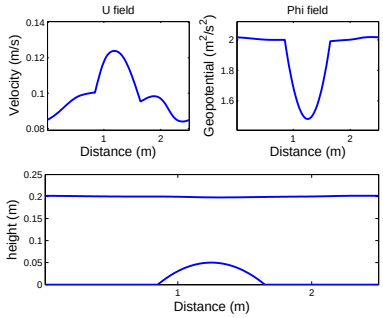

Figure 1: Velocity, geopotential and water level in an instant for the shallow water model

obstacle is given by

$$\bar{h}(z) = \bar{h}_c \left(1 - \frac{z^2}{a^2}\right) \quad \text{for} \quad 0 \leq |z| \leq a,$$

and $\bar{h}(z) = 0$ otherwise. $\bar{h}_c$ is the maximum height of the obstacle and a is half the length over which the base of the obstacle extends. The value of a is taken to be $0.4m$ and the height of the obstacle $\bar{h}_c = 0.05m$. As initial fields, we choose $u_0(z) = 0.1m/s$ and $h_0(z) = 0.2m$ for all z. For simplicity, we set the gravitational constant g at $10m/s^2$. The time step Δt for model integration is $4.6 \times 10^{-3}s$. The state vector of this system at an instant t is the combination of the velocity and the geopotential at each grid point and is noted by $\mathbf{x}(t)$. Its dimension is 500 since there are 250 grid points in the discretized space.

The framework for the numerical tests is the classical twin experiment. To create the reference states, we integrate the model from an arbitrary initial condition during 400 time steps which is approximately the system period. These reference states will be considered to be our truth and could be compared with the states produced by the assimilation method. The experiments are performed using pseudo-measurement which are extracted from the reference states. More precisely, the observations used during the assimilation are created by adding Gaussian noises to the reference fields. We observe only two components of the state vector at each time step. The assimilation problem is divided into 10 assimilation windows. Each of

them is composed of 40 time steps. The background $\mathbf{x}^b$ is the true state perturbed by a Gaussian noise. The background covariance matrix $\mathbf{B}$ is computed using a Laplacian-based correlation model.

To build the sampling covariance matrix $\mathbf{S}$ needed to compute the subspace $\mathbf{L}_0$, we create a set of state vectors. We integrate the model during 400 time steps but from an arbitrary initial condition which occurs before the assimilation windows. We retain the state vector with a 8-time step periodicity since successive states are similar and do not bring additional information.We apply the spectral decomposition (5) to compute the EOF basis. The percentage of variation accounted for by the first r EOFs is given by

$$100\frac{\sum_{i=1}^{r}\lambda_i}{\sum_{i=1}^{n}\lambda_i}$$

where λ_i are the eigenvalues of $\mathbf{S}$ ordered in decreasing order. We plot this percentage in figure 2 and see that retaining 5 EOFs enables to explain 80% of the system's variability. This is a good compromise since we account

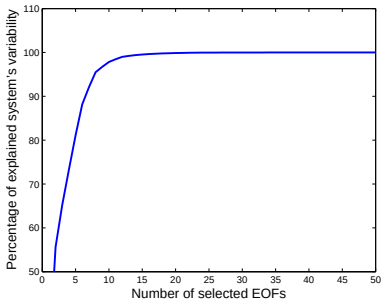

Figure 2: Percentage of explained system's variability versus the number of selected EOFs

for most of the variation without increasing to much the dimension of the subspace spanned by the EOFs. All of these parameters are chosen according to the choice made for operational data assimilation problem and the dimension of the model state vector.

The set of experiments is designed to illustrate the impact of the Ritz-Galerkin starting point and of the LMP when the subspace is defined by the first EOFs. We allow three outer integrations in the Gauss-Newton algorithm with 5 inner CG iterations to solve each linear system. These are possible heuristics to solve the 4D-Var problem. In Figure 3, we plot the history of the quadratic cost function and the non- linear function (the quadratic with lines and the nonlinear with circles) of the 3 systems. The curves are placed one after the other in sequence and the inner iterations are cumulated. Note that the nonlinear function value is available only at outer iter-

ations when linearization is performed in the Gauss-Newton algorithm. The

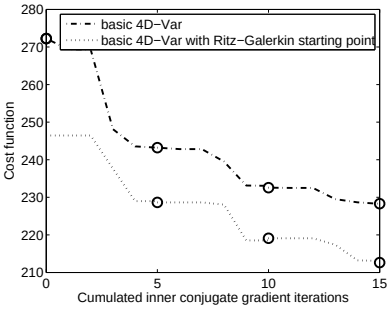

Figure 3: Convergence of the basic 4D-var and of the basic 4D-Var with the Ritz-Galerkin starting point

dash-dot line is obtained using a basic Gauss-Newton algorithm using the background covariance matrix $\mathbf{B}$ as preconditioner in the CG method. The dotted line is obtained using the same approach but with the Ritz-Galerkin starting point (40). It means that the starting point for the minimization is the solution of a reduced incremental 4D-Var problem prolongated into the full space. We see that the Ritz-Galerkin starting point provides a good reduction for the first quadratic problem and improves the convergence rate of the Gauss-Newton method. Nevertheless, both approaches do not have the same cost for the whole minimization since it is dominated by the number of Hessian-vector products. Indeed, the calculation of the Ritz-Galerkin starting point needs five extra Hessian-vector products to evaluate $\mathbf{A}_0\mathbf{L}_0$ while each CG iteration needs one Hessian-vector product. The nonlinear function is equal to 228.3 with the basic 4D-Var after 15 cumulated inner iterations and is equal to 219.1 after 10 cumulated inner iterations with the Ritz-Galerkin. Thus, for the same computational effort, the Ritz-Galerkin starting point gives a better reduction in the nonlinear function.

The choice of the preconditioner has also an influence on the convergence rate. In figure 4, we take back the dotted line which is obtained with the background covariance matrix $\mathbf{B}$ as preconditioner and with the Ritz-Galerkin starting point. The use of $\mathbf{B}$ as preconditioner is a simple choice often used for data assimilation problem. The dashed line shows the convergence curve when we use the limited memory preconditioner based on the first EOFs (41) with the Ritz-Galerkin starting point. As seen in Section 4, this LMP is available without any extra computational cost since the product $\mathbf{A}_0\mathbf{L}_0$ has already been computed for the Ritz-Galerkin starting point. With this approach, the nonlinear cost function is equal to 193.3 after 10 cumulated inner iterations. It is a significant asset for the use of the LMP since we have obtain, for a same computational effort, a better

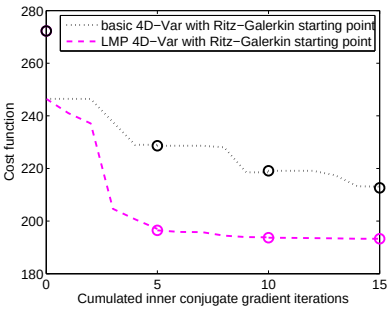

Figure 4: Convergence of the basic and LMP 4D-Var with Ritz-Galerkin starting point

reduction in the nonlinear function than when using $\mathbf{B}$ as preconditioner with the Ritz-Galerkin starting point.

We can partly explain this good result from the spectrum of the first matrix system since the rate of convergence of the conjugate gradient method dpends, to a large extend, on the condition number and on the clustering of the eigenvalues. In figure 5, we plot the spectrum of the first matrix system $\mathbf{A}_0$ and its preconditioned versions $\mathbf{BA}_0$ and $\mathbf{H}_0\mathbf{A}_0$. The matrix $\mathbf{A}_0$ has a continuous spectrum and its condition number is equal to 9.58×10^7. If we use $\mathbf{B}$ as preconditioner, the preconditioned matrix $\mathbf{BA}_0$ is given by

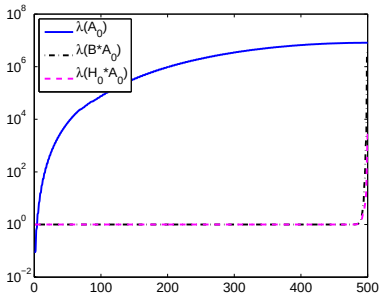

Figure 5: Spectrum of the first preconditioned matrix system

$\mathbf{I}_n + (\mathbf{G}_0)^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{G}_0$. Thus, the spectrum becomes bounded below 1 and has a cluster at 1 of size at least $\max\{0, n-m\}$ where m is the dimension of the observation vector $\mathbf{y}$ (Tshimanga et al., 2008). In our case the size of the cluster is at least 460 and the condition number is 3.66×10^6. With the LMP preconditioner, the condition number falls to 3.31×10^3 while the cluster do not decrease. In figure 6, we zoom in the largest eigenvalue to make easier the comparison between the spectrum of the preconditioned matrices. This spectral analysis shed some light on the good performance of the LMP on the first linear system.

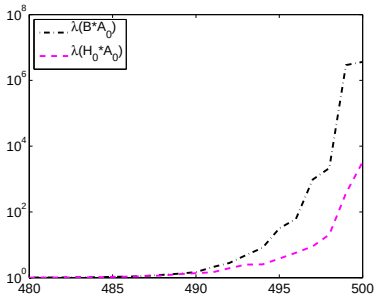

Figure 6: End of spectrum of the first preconditioned matrix system

7 Conclusions and perspectives

In the data assimilation community, the variants of Kalman filter and the 4D-Var formulation are the two major approaches to solve data assimilation problems. We have presented a theoretical work about the correspondences between these approaches and generalized them to the reduced case. These results have been used as a baseline for combining both methods.

We have enhanced the Gauss-Newton method using a Ritz-Galerkin starting point with respect to the first empirical orthogonal functions. We have shown that it is equivalent to the solution of a reduced incremental 4D-Var problem. Moreover, we have developed a limited memory preconditioner based on the same information. The combination of these two elements improves the convergence rate of the Gauss-Newton method for a data assimilation problem on a shallow water model.

For further research, we are concerned by the application of our algorithm in an operational oceanographic or atmospheric application to show that the same conclusion can be drawn for more realistic assimilation problems. We also want to use information as Ritz pairs or descent directions, gained when solving one system, to update the limited memory preconditioner based on the EOFs for the next system of the sequence.

References

B. Bell and F. Cathey. The iterated kalman filter update as a gauss-newton method. *IEEE Transactions on Automatic Control*, 38:294–297, 1993.

M. Benzi. Preconditioning techniques for large linear systems: a survey. *J. Comput. Phys.*, 182(2):418–477, 2002.

A. Bjorck. *Numerical Methods for Least Squares Problems*. SIAM, 1996.

P. Courtier, J. Thépaut, and A. Hollingsworth. A strategy for operational

implementation of 4d-var, using an incremental approach. *Quarterly Journal of the Royal Meteorological Society*, 120:1367–1388, 1994.

D. Daescu and I. Navon. Efficiency of a pod-based reduced second-order adjoint model in 4d-var data assimilation. *International journal for numerical methods in fluids*, 53:985–1004, 2007.

S. Durbiano. *Vecteurs caractéristiques de modèles océaniques pour la réduction d'ordre en assimilation de données*. PhD thesis, Universit Joseph Fourier, 2001.

A. Gelb. *Applied optimal estimation*. MIT Press, 1974.

S. Gratton, A. Lawless, and N. Nichols. Approximate gauss–newton methods for nonlinear least squares problems. *SIAM Journal on Optimization*, 18 (1):106–132, 2007.

S. Gratton, A. Sartenaer, and J. Tshimanga Ilunga. On a class of limited memory preconditioners for large scale linear systems with multiple right-hand sides. *Numerical analysis report, The university of Namur*, 2008.

R. Horn and C. Johnson. *Matrix Analysis*. Cambridge University Press, 2006.

I. Hoteit and D. T. Pham. Evolutivity of the reduced state space and data assimilation schemes based on the kalman filter. *Journal of the Meteorological Society of Japan*, 81:21–39, 2003.

K. Ide, P. Courtier, M. Ghil, and A.C. Lorenc. Unified notation for data assimilation: operational, sequential and variational. *J. Meteorol. Soc. Japan*, 75:181–189, 1997.

A. Jazwinski. *Stochastic Processes and Filtering Theory*. Academic Press, 1970.

R. Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(Series D): 35–45, 1960.

E. Kalnay. *Atmosheric modeling, data assimilation and predictability*. Cambridge University Press, 2003.

A. Lawless. *Development of linear models for data assimilation in numerical weather prediction*. PhD thesis, Department of Mathematics, University of Reading, Reading, United Kingdom, 2001.

A. Lawless, S. Gratton, and N. Nichols. Approximate iterative method for variational for data assimilation. *International journal for numerical methods in fluids*, 1:1–6, 2004.

A. Lawless, N. Nichols, C. Boess, and A. Bunse-gerstner. Using model reduction methods within incremental 4d-var. *Numerical analysis report, The university of Reading*, 2006.

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

Z. Li and I. M. Navon. Optimality of 4d-var and its relationship with the kalman filter and kalman smoother. *Quarterly Journal of the Royal Meteorological Society*, 127:661–684, 2001.

C. Meyer. *Matrix analysis and applied linear algebra*. SIAM, 2000.

J. Nocedal and S. Wright. *Numerical optimization*. Springer series in operation research, 2006.

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

F. Rabier. Overview of global data assimilation developments in numerical weather-prediction centres. *Quarterly Journal of the Royal Meteorological Society*, 131:3215–3233, 2005.

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

C. Robert, S. Durbiano, E. Blayo, J. Verron, J. Blum, and F.-X. Le Dimet. A reduced order strategy for 4D-var data assimilation. *J. Mar. Syst.*, 57 (1-2):70–82, 2005.

C. Robert, E. Blayo, and J. Verron. Reduced-order 4d-var: a preconditioner for the incremental 4d-var data assimilation method. *Geophys. Res. Lett.*, 33:1–4, 2006.

D. Rozier, F. Birol, E. Cosme, P. Brasseur, J. M. Brankart, and J. Verron. A reduced-order kalman filter for data assimilation in physical oceanography. *SIAM Review*, 49(3):449–465, 2007.

Y. Saad. *Iterative Methods for Sparse Linear Systems*. SIAM, second edition, 2008.

G. Strang and K. Borre. *Linear Algebra, geodesy, and GPS*. Wellesley-Cambridge Press, 1997.

R. Todling and S. E. Cohn. Suboptimal schemes for atmospheric data assimilation based on the kalman filter. *Mon. Wea. Rev.*, 122:2530–2557, 1994.

J. Tshimanga, S. Gratton, A. Weaver, and A. Sartenaer. Limited-memory preconditioners, with application to incremental four-dimensional variational data assimilation. *Quarterly Journal of the royal Meteorological Society*, 134:751–769, 2008.

H. van der Vorst. *Iterative Krylov Methods for Large Linear Systems*. Cambridge Monographs on Applied and Computational Mathematics, 2003.

J. Wang and J. Li. A four-dimensional scheme based on singular value decomposition (4dsvd) for chaotic-attractor-theory-oriented data assimilation. *Journal of geophysical research*, 114, 2009.

A. Weaver, J. Viliard, and D. Anderson. Three- and four-dimensional variational assimilation with an ocean general circulation model of the tropical pacific ocean : Part i. formulation, internal diagnostics and concistency checks. *Mon. Weather Rev.*, 131:1360–1378, 2003.

M. Yaremchuk, D. Nechaev, and G. Panteleev. A method of successive corrections of the control subspace in the reduced-order variational data assimilation. *Monthly weather review*, 137:2966–2978, 2009.