A Consistent Hybrid Variational-Smoothing Data Assimilation Method: Application to a Simple Shallow-Water Model of the Turbulent Midlatitude Ocean

MONIKA KRYSTA
Laboratoire des Ecoulements Géophysiques et Industriels, and Laboratoire Jean Kuntzmann, Grenoble, France

ERIC BLAYO
Laboratoire Jean Kuntzmann, Grenoble, France

EMMANUEL COSME AND JACQUES VERRON
Laboratoire des Ecoulements Géophysiques et Industriels, Grenoble, France

(Manuscript received 16 July 2009, in final form 13 April 2011)

ABSTRACT

In the standard four-dimensional variational data assimilation (4D-Var) algorithm the background error covariance matrix $B$ remains static over time. It may therefore be unable to correctly take into account the information accumulated by a system into which data are gradually being assimilated.

A possible method for remedying this flaw is presented and tested in this paper. A hybrid variational-smoothing algorithm is based on a reduced-rank incremental 4D-Var. Its consistent coupling to a singular evolutive extended Kalman (SEEK) smoother ensures the evolution of the $B$ matrix. In the analysis step, a low-dimensional error covariance matrix is updated so as to take into account the increased confidence level in the state vector it describes, once the observations have been introduced into the system. In the forecast step, the basis spanning the corresponding control subspace is propagated via the tangent linear model.

The hybrid method is implemented and tested in twin experiments employing a shallow-water model. The background error covariance matrix is initialized using an EOF decomposition of a sample of model states. The quality of the analyses and the information content in the bases spanning control subspaces are also assessed. Several numerical experiments are conducted that differ with regard to the initialization of the $B$ matrix. The feasibility of the method is illustrated. Since improvement due to the hybrid method is not universal, configurations that benefit from employing it instead of the standard 4D-Var are described and an explanation of the possible reasons for this is proposed.

1. Introduction

The successful integration of data and model information has become an increasing challenge for atmosphere and ocean sciences with the emergence of efficient numerical models and the rapidly increasing availability of remote-sensed and in situ measurements. The data assimilation methodologies offer a mathematical and, in principle, optimal framework for responding to this challenge.

Two main approaches, variational and sequential data assimilation, coexist in geophysical applications, and are encountered in research developments as well as in operational applications such as numerical weather prediction (NWP) and operational oceanography. Today’s meteorological applications are for the most part based on variational assimilation, whereas a number of oceanographic applications have favored sequential methods.

The variational approach is defined within the framework of optimal control theory (Lions 1971; Le Dimet and Talagrand 1986). A cost function is built measuring a distance between a model of a physical phenomenon and observations. Four-dimensional variational data assimilation
(4D-Var) is a variant of the variational approach that employs a cost function embracing all the observations accumulated within a given time interval and compared to the model at the appropriate times. A minimum of the cost function is sought in order to define an improved initial condition and thereby the system’s trajectory for the period under consideration. Most often the issue of finding the minimum takes the form of an ill-posed inverse problem requiring an additional regularization term. Some external knowledge on the system is then introduced, namely an initial guess (or background) $\hat{x}$ (known as $x'$ in the sequential approach), and the corresponding background error covariance matrix $B$ whose sequential counterpart is the forecast error covariance matrix $P_f$. NWP favors purely statistical approaches to specify $B$, as can be found for example in Parrish and Derber (1992), Rabier et al. (1998), Derber and Bouttier (1999), Wu et al. (2002), Buehner (2005), or in an exhaustive review by Bannister (2008a,b), but parameterizations based on physical or mixed physical and statistical or empirical considerations can also be employed (e.g., Weaver and Ricci 2004; Weaver et al. 2005). Nevertheless, in the standard 4D-Var, once $B$ has been specified it remains static over the entire duration of a data assimilation experiment.

Sequential assimilation derives from the Kalman filter (KF) theory (Kalman 1960). Unlike variational assimilation, it provides estimates of the covariance matrices of the forecast and analysis error at each step of the assimilation process, whenever a new set of observations becomes available. Kalman smoothers (KS; e.g., Cohn et al. 1994; Evensen and van Leeuwen 2000; Cosme et al. 2010) generalize the Kalman filter analysis by using data within the entire time span under consideration. Like the 4D-Var, the KF–KS approach is only truly optimal in the linear cases. In practice, the KF–KS and their variants are unable to handle the large size system state vectors that are encountered in realistic geophysical problems. Various forms of adaptation have been developed aimed at reducing the computational burden, among which can be cited the ensemble Kalman filter (EnKF; Evensen 1994) and the singular evolutive extended Kalman filter (SEEK; Pham et al. 1998; Brasseur and Verron 2006).

Numerous efforts have been made to better understand the connections between the variational and the sequential methods and to open new research avenues based on their complementarities (e.g., Lorenc 2003). It must be remembered that assuming a hypothetical linear and perfect model and a linear observation operator, the Kalman filter and the 4D-Var approaches would both provide equivalent estimates at the end of the assimilation window. Weak-constraint 4D-Var and the fixed-interval KS are equivalent in the case of a linear model (Rauch et al. 1965; Ménard and Daley 1996; Li and Navon 2001). The work by Fisher et al. (2005) in particular explores how this can be extended to nonlinear situations. Note also that some equivalence can be found between an incremental approach to variational assimilation (Courtier et al. 1994) and the extended Kalman filter (EKF) approximation proposed by Jazwinski (1970) for sequential estimation.

An analysis of equivalences and differences between 4D-Var and KF suggests the value of combining the two approaches and producing a hybrid version in order to enjoy the practical benefits of each and to provide more powerful algorithms in terms of accuracy and efficiency. Clearly a first suggestion would be to take advantage of the natural complementarity of 4D-Var and KF–KS with regard to $B$, or, in other words, to take full advantage of 4D-Var for updating the state and to rely on the KF–KS to propagate and update the covariance matrix. Adopting such a strategy in a realistic atmospheric or oceanic context involves dealing with some form of order reduction for the KF–KS, and therefore, also dealing with order reduction for the 4D-Var. Reduced-order 4D-Var has already been explored with some success, in particular by Blayo et al. (1999) and Robert et al. (2005). It is therefore tempting to explore a fully consistent scheme for hybridization based on 4D-Var and a reduced-order KF–KS approach.

It is worth mentioning that hybridization involving the 4D-Var and the SEEK filter has already been investigated in an oceanic context by Robert et al. (2006a). For atmospheric applications, Fisher and Andersson (2001) and Beck and Ehrendorfer (2005) examined the possibility of updating the 4D-Var background error covariance matrix with a reduced-rank Kalman filter (RRKF), whereas Liu et al. (2008, 2009) explored coupling between 4D-Var and the EnKF. In a similar framework, Zhang et al. (2009), like Wang et al. (2008) before them (in the context of 3D-Var and ensemble transform Kalman filter), opted for a linear combination of a static and dynamically estimated error covariance matrix, as put forward by Hamill and Snyder (2000).

This work is conducted in the context of an existing incremental 4D-Var system [such as the one used for the Nucleus for the European Modeling of the Ocean (NEMO) system]. The aim of this work is to investigate a possible way of enriching the incremental 4D-Var by making the $B$ matrix evolve and to determine under what circumstances this can improve results with respect to a static case. The means of evolving $B$ are obtained through a consistent hybridization of an incremental 4D-Var and a smoother, obtained from the SEEK filter and referred to as the SEEK smoother, under a perfect-model hypothesis. Note that the use of the SEEK filter–smoother offers a potentially powerful way of evolving a reduced-error
subspace (Brasseur and Verron 2006). The framework of hybrid formalism is thus provided by the incremental 4D-Var, while the error covariance matrix update is made according to the recipe prescribed by the SEEK smoother, whenever a transition between two adjacent assimilation windows occurs.

The present study thus extends the preliminary exploration by (Robert et al. 2006a). It is in line with the full theoretical consistency of the hybridization scheme developed by Veersé et al. (2000), but makes significant advances with respect to demonstrating its applications. It differs from the work by Liu et al. (2008, 2009) and Zhang et al. (2009) in that the scheme is more consistent owing to the use of the SEEK smoother instead of the ensemble Kalman filter. Another important difference lies in the use context. Our interest here focuses on the midlatitude mesoscale ocean where turbulent dynamics develops through barotropic and baroclinic instabilities. The dynamical situation chosen mimics the oceanic context. The framework of this work, but in this section we only need to consider a single assimilation window \([t_0, t_N]\). The subscripts denote the indices of time within this interval. The time window contains \(N\) instants \(t_1, \ldots, t_N\), where observations \(y_1^0, \ldots, y_N^0\) are available, with error covariance matrices \(R_1, \ldots, R_N\). We use \(y^0 = [y_1^0, \ldots, y_N^0]^T\) to denote the concatenated observation vector (T in exponent stands for transposition) and \(R\) to denote the \(R_i\)-block diagonal observation error covariance matrix. We call \(x_b\) the state vector at \(t_0\) and define a concatenated state vector as \(x = [x_1^T, \ldots, x_N^T]^T\), where \(x_i\) is the solution of the dynamical model integration from \(t_0\) to \(t_i\); \(x_i = M_{0-i}(x_0)\). The corresponding linearized model is called \(M_{0-i}(\cdot)\), where \(\cdot\) stands for the trajectory around which linearization takes place. The concatenated linearized model is \(M(\cdot) = [M_{0-1}(\cdot), \ldots, M_{0-N}(\cdot)]^T\).

Finally, the observation operator is denoted as \(H(x) = [H_1(x_1)^T, \ldots, H_N(x_N)^T]^T\) and the linearized operator as \(H(x_i)\), where \(x_i\) is the trajectory around which linearization takes place. It is defined as a block diagonal matrix formed with the single time linearized operators \(H_i(x_i)\). Other notations are described at appropriate times in the text.

### b. Incremental 4D-Var

#### 1) FORMULATION

In the incremental approach to 4D-Var, the cost function is formulated in terms of an increment \(\delta x_b\) to be added to the initial condition at \(t_0\). Following the presentation of Weaver et al. (2002), the transition from a background state \(x_b^0\) toward an analysis \(x_b^0\) occurs via the sequential computation of \(K\) intermediate increments \(\delta x_b^{(k)}\) and analyses \(x_b^{(k)} = x_b^0 + \delta x_b^{(k)} \ (k = 1, \ldots, K)\). As the assumption of a perfect model is adopted throughout this paper, the increments are obtained through the minimization of a sequence of cost functions specified in a strong formulation:

\[
\mathcal{J}(\delta x_b^{(k+1)}) = \frac{1}{2} \delta x_b^{(k+1)} \mathbf{B}_0^{-1} \delta x_b^{(k+1)} + \frac{1}{2} \mathbf{B}_0^{-1} \left[ H(x_b^{(k)}) \mathbf{M}(x_b^{(k)}) \delta x_b^{(k+1)} - d^{(k)} \right]^T \times R^{-1} \left[ H(x_b^{(k)}) \mathbf{M}(x_b^{(k)}) \delta x_b^{(k+1)} - d^{(k)} \right],
\]

(1)

where \(\mathbf{B}_0\) is the background error covariance matrix. The innovation \(d^{(k)}\) is given by

\[
d^{(k)} = y^0 - H(x_b^{(k)}) + H(x_b^{(k)}) \mathbf{M}(x_b^{(k)}) \delta x_b^{(k)},
\]

and \(k\) indexes the outer loop iterations. The benefit of the formulation in terms of increments lies in the use of linearized operators. Once they have been linearized, they no longer depend on the unknown variables (i.e., on the increments \(\delta x_b\)), and the functions that need to be minimized at each step, Eq. (1), are quadratic in \(\delta x_b\). For a given iteration \(k\), the solution can be computed analytically:

\[
\delta x_b^{(k+1),VAR} = \left[ \mathbf{B}_0^{-1} + M^T R^{-1} H M \right]^{-1} M^T R^{-1} d^{(k)} = B_0^{-1} \left[ M^T H [M^T R^{-1} H M]^{-1} M^T R^{-1} \right]^{-1} d^{(k)}.
\]

(3)

To keep the formulas concise we have skipped any reference to the trajectory around which the operators have been linearized. It must be borne in mind, however, that linearization must be repeated each time we pass from the \(k\)th to the \((k + 1)\)th cost function.
Once the 4D-Var analysis has been completed, the state is propagated from \( t_0 \) to \( t_0 + T \) with the nonlinear model, thus providing a background state for the next 4D-Var batch.

2) RANK REDUCTION

In geophysical applications, the 4D-Var problem is generally poorly conditioned, and the use of a preconditioning technique (Gilbert and Lemaître 1989) is required. Formally, this can be achieved with an eigendecomposition of the background error covariance matrix:

\[
B_0 = L_0 U_0^T L_0^T.
\] (4)

Given the fact that all eigenvalues are positive, the square root of \( U_0^T \) can be defined and easily computed. With the change of control variable \( \delta x_0^{(k+1)} = (L_0 U_0^{-1/2}) \chi^{(k+1)} \), the cost function can be rewritten (Blayo et al. 1998):

\[
J[\chi^{(k+1)}] = \frac{1}{2} \chi^{(k+1)T} L_0 \chi^{(k+1)} + \frac{1}{2} [HML_0 U_0^{1/2} \chi^{(k+1)} - d^{(k)}]^T R^{-1} [HML_0 U_0^{1/2} \chi^{(k+1)} - d^{(k)}].
\] (5)

The decomposition of the background error covariance matrix not only allows preconditioning but also rank reduction (Blayo et al. 1999; Robert et al. 2006b). Rank reduction here consists in reducing the dimension of \( U_0^T \), originally of the same size (squared) as the state vector, to only a small value, by preserving exclusively the largest eigenvalues. \( L_0 \) must also be modified to exclude the eigenvectors corresponding to the discarded eigenvalues.

c. Equivalent fixed-interval smoother

We now present the second ingredient of the hybrid method, namely an (incremental) extended Kalman smoother. Since smoothers can be seen as extensions of the filter, they are generally described with the same formalism. The smoother presented immediately below was built with a view to ensuring equivalence with the incremental 4D-Var of the previous section. Hence, the analysis is computed at the beginning of the assimilation window for both the state and the covariance matrix. The incremental aspect makes it different from other fixed-interval smoothers presented by Ménard and Daley (1996), Evensen and van Leeuwen (2000), and Li and Navon (2001). This presentation is based on Verron et al. (2000).

1) FORMULATION

The data assimilation framework coincides with the one presented in section 2b(1) for the 4D-Var. The smoother is built as an iterative procedure and computes a series of incremental analyses with respect to a background state. Let us suppose that the increment \( \delta x_0^{(k)} \) has already been found. The \((k+1)th\) analysis is taken to be a linear function of the innovation vector \( d^{(k)} \) [Eq. (2)], according to

\[
\delta x_0^{(k+1),EKS} = x_0^{E(k+1)} - x_0^b = K^{(k+1)} d^{(k)}.
\] (6)

The gain matrix \( K^{(k+1)} \) is determined by minimizing the trace of the analysis error covariance matrix \( P_0^{(k+1)} \). The computation, not detailed here, is very similar to that of the Kalman filter gain. The solution is written as

\[
K^{(k+1)} = B_0 M^T H^T (H M B_0 M^T H^T + R)^{-1},
\]

where, again, the reference to the trajectory around which linearization takes place [i.e., \( x_0^{E(k)} \)] has been skipped. One should keep in mind this dependence, however, since it is the only factor that accounts for the differences between the gain matrix at successive iterations \( k \). Once \( K^{(k+1)} \) has been used in Eq. (6), which expresses the increment \( \delta x_0^{(k+1),EKS} \) of the \((k+1)th\) Kalman smoother analysis, it can easily be seen that the increment coincides with that given by Eq. (3) and obtained in the variational approach.

In addition to the state increment at each iteration \( k \), the EKS provides an increment to the covariance matrix:

\[
\delta P_0^{(k+1)} = P_0^{(k+1)} - B_0 = -K^{(k+1)} H M B_0.
\] (7)

However, to obtain the final analysis covariance matrix \( P_0^a \), it is only worth computing the covariance matrix increment after the analyzed state has been obtained. Indeed, intermediate matrices are not used in the iterations.

Once the smoother analysis has been performed, the analyzed state at \( t_0 \) can be propagated to \( t_0 + T \) with the nonlinear model, providing the background state for the next time interval. The analysis error covariance matrix can also be propagated as

\[
B_N = M_{0 \rightarrow N} P_0^{a} M_{0 \rightarrow N}^T
\] (8)

to yield the corresponding covariance matrix.

2) RANK REDUCTION

Rank reduction is common in the Kalman filtering approach on account of the computational burden involved. Different algorithms have been presented in the literature (Evensen 1994; Verlaan and Heemink 1998; Pham et al. 1998). The SEEK filter was first described by Pham et al. (1998), then in different versions by Verron.
et al. (1999), Brankart et al. (2003), Brasseur and Verron (2006), or Rozier et al. (2007). The presentation of the reduced-rank smoother below follows the formalism of Pham et al. (1998).

The background error covariance matrix is factorized following Eq. (4). As with the earlier 4D-Var, rank reduction consists in reducing the dimension of $\mathbf{U}_0^k$ to only a few eigenvalues. From Eq. (7), it can be shown (see Pham et al. 1998) that the analysis error covariance matrix at iteration $k$ reads as

$$
\mathbf{P}^{a(k+1)}_0 = \mathbf{L}_0 \mathbf{U}^{a(k+1)}_0 \mathbf{L}_0^T,
$$

(9)

with

$$
[U^{a(k+1)}_0]^{-1} = U^{-1}_0 + L_0^T M^T H^T R^{-1} H M L_0,
$$

(10)

where, again, the dependence of the last term on the intermediate analysis $x^{a(k)}$ is not written in detail. As with the full rank smoother, computing the intermediate matrices $U^{a(k+1)}_0$ is superfluous and can be avoided. The final analysis matrix computed at the last iteration can be propagated over the assimilation window. If we define

$$
L_N = M_{0\rightarrow N}(x^a_0) L_0
$$

(11)

then the error covariance matrix at the end of the interval reads as

$$
B_N = L_N U^a_0 L_N^T.
$$

(12)

d. Hybridization

A comparison of the analysis increments computed by the incremental 4D-Var (section 2b) and by an appropriately designed fixed-interval smoother (section 2c) indicates they are theoretically identical, provided the same ingredients are used. The smoother also offers the opportunity to update and dynamically propagate the background error covariance matrix from one batch to the next. Combining the incremental 4D-Var cycle (for the state vector) with the smoother cycle (for the covariance matrix) results in a perfectly consistent hybrid 4D-Var–fixed-interval smoother. The eigenvalue decomposition of the background covariance matrix is interesting because of various aspects. With respect to 4D-Var, it is useful for preconditioning. For the smoother, it enables the practical propagation of the (reduced rank) background error covariance matrix, when its rank is strongly reduced. Moreover, this rank reduction does not spoil the equivalence and consistency between 4D-Var and the fixed-interval smoother. To summarize, the hybrid 4D-Var/SEEK smoother operates as follows:

1) A sequence of 4D-Var minimizations [Eq. (5), $k$ running from 1 to $K$] is performed to compute an analysis increment and finally the analysis $x^a_k$. The inputs of the algorithm are the background state $x^b_k$, and a decomposition of a reduced-rank error covariance matrix in terms of $L_0$ and $U_0^k$ [see Eq. (4)].

2) The 4D-Var analyzed state is propagated with the dynamical model: $x^a_{k+1} = M_{0\rightarrow N}(x^a_0)$ to establish the background state for the next assimilation window.

3) SEEK smoother analysis $U^a_0$ is conducted for the reduced-rank error covariance matrix, Eq. (10), where the model is linearized around the analyzed trajectory obtained from $x^a_k$.

4) Theoretically, propagation of the basis spanning the control subspace in Eq. (11), is needed to obtain $L_N$, but in practice $U^a_0$ has to be updated, and this was done at the previous step of the algorithm.

Since the analyses produced by both ingredients of the hybrid method are the same, it is sufficient to produce them within the framework of 4D-Var. Consequently, although $x^a_{k}$ in Eq. (11) stands, in theory, for the analyses produced by the SEEK smoother, in practice a linearization procedure is performed for those obtained in 4D-Var.

3. An example of a numerical implementation

This section presents a first implementation of the hybrid method described above, in the context of a simple ocean model. Its aim is to demonstrate that the hybrid algorithm can significantly improve the performance of assimilation with respect to the 4D-Var approach in which $B$ is kept static. In the context of twin experiments, we choose a test case where the background error statistics are partly, but not perfectly known, which is in fact the case in most realistic applications. The general question of identifying those circumstances determining whether the hybrid method leads to substantial improvement over the 4D-Var is not discussed here, but will be addressed in section 4.

a. Shallow-water model

The hybrid method has been tested with the help of a reduced gravity shallow-water model mimicking wind-driven midlatitude circulation. Although rather simple, this model is able to develop strongly nonlinear dynamics, with a meandering zonal jet and associated eddies. Here, we consider a flat bottom square domain $[0, L] \times [y_0 - L/2, y_0 + L/2]$, where $y_0$ is the value of $y$ at the center of the basin. The governing equations are
\[ \frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} + v \frac{\partial u}{\partial y} - f v + g^* \frac{\partial h}{\partial y} h = \frac{\tau_x}{\rho_0 h} - r u + \nu \Delta u, \]

\[ \frac{\partial v}{\partial t} + u \frac{\partial v}{\partial x} + v \frac{\partial v}{\partial y} + f u + g^* \frac{\partial h}{\partial x} h = \frac{\tau_y}{\rho_0 h} - r v + \nu \Delta v, \]

\[ \frac{\partial h}{\partial t} + \frac{\partial}{\partial x} (hu) + \frac{\partial}{\partial y} (hv) = 0, \]

where \((u, v)\) is the horizontal velocity and \(h\) is the water layer thickness. The system is closed with impermeability and no-slip boundary conditions: \(u = v = 0\). The \(g^*\) is a reduced gravity and \(\rho_0\) is the water density; \(f\) is the Coriolis parameter, given by the \(\beta\)-plane approximation; \(f(y) = f(y_0) + \frac{\beta}{(y - y_0)}\) with \(\beta = (\partial f/\partial y)(y_0)\); \(v\) and \(r\) are the diffusivity and bottom friction coefficients, respectively. The wind stress \(\tau = (\tau_x, \tau_y)\) is chosen constant in time: \(\tau_x = \tau_{0x} \cos[2\pi(y - y_0)/L]\) and \(\tau_y = 0\), which leads to a classical double gyre circulation (see Fig. 1). Numerical values of the parameters are chosen as follows: \(L = 2000\) km, \(f(y_0) = 7 \times 10^{-5}\) s\(^{-1}\), \(\beta = 2 \times 10^{-11}\) m\(^{-1}\) s\(^{-1}\), \(\rho_0 = 10^3\) kg m\(^{-3}\), \(g^* = 0.02\) m s\(^{-2}\), \(\tau_0 = 0.015\) N m\(^{-2}\), \(r = 5 \times 10^{-7}\) s\(^{-1}\), and \(\nu = 9\) m\(^2\) s\(^{-1}\). Spatial discretization is performed on an Arakawa C grid with 25-km resolution. The leapfrog numerical scheme stabilized by the Robert-Asselin filter (Robert 1966) is used to perform time integration, with a time step of 30 min.

b. Configuration of the data assimilation experiment

A preliminary 5-yr spinup was allowed starting from rest \((u = v = 0, h = 500\) m\) in order to stabilize the model solution. The end of this spinup period is referred to as the time \(t = 0\) in what follows. Then, a 10-yr simulation obtained from this initial condition was performed. Given the present model dynamics, which is much less complex than in more realistic models, it may be considered that almost the entire range of variability in the model was explored during this 10-yr period. A complete set of 1825 reference EOFs (e.g., Hannachi et al. 2007), denoted \(L^\text{ref}\) (and in order of decreasing magnitude of the corresponding eigenvalues) was constructed from a 2-day sampling of this model trajectory, which is hence regarded as a comprehensive representation of model variability. Note that the first 100 (340) reference EOFs retain 75.7% (95%) of the total variability.

The data assimilation experiment was then set up in the following way. We chose to use the twin-experiment framework and another 10-yr simulation was performed to follow the earlier one. The 15–18-yr period was selected to represent the “true” trajectory of the system, from which measurements were extracted. In a similar manner to the assimilation of altimetric data, observations were chosen to be \(h\) only, assimilated every 10 days. Consequently, we deal with one observed variable, \(h\), and two unobserved ones, \(u\) and \(v\). For the needs of the experiments presented in this paper, \(h\) was observed everywhere in the domain and a white noise representing an error with a 5-cm standard deviation was added to these observations. This noise was chosen artificially small to generate strong and unambiguous corrections, thus emphasizing the subspace directions, described by the covariance matrix \(B\), in which they occur.

The length of the assimilation windows is 30 days so that 3 sets of observations are available for any of them at times \(t_{10}, t_{20}, t_{30}\), where the subscript stands for the corresponding day. The background value for the initial condition of the first assimilation window was chosen as the beginning of the 11th year of simulation (i.e., 4 years before the true state); this background field is fully decorrelated from the true one. The background error covariance matrix, \(B_0\), used in the first assimilation window was defined consistently with this initial condition. It was built using 100 EOFs corresponding to a 2-day sampling of the 200 days preceding this initial condition (i.e., the 200 last days of the 10th year of simulation). This corresponds to a context in which one has only partial knowledge of system variability, which is the case in real-life applications. We shall refer to this configuration as \(\text{EXP\_100\_SHORT}\) since the EOFs used to construct \(B_0\) arise from sampling the reference trajectory over a relatively short period of time.

c. Cost function based on a reduced-rank background error matrix \(B\)

With a view to providing a feasible evolution of the error covariance matrix as proposed by the hybrid method (section 2), the variational ingredient of the algorithm is based on a reduced-rank approach. It has already been shown (Robert et al. 2005), in the context of an ocean circulation model significantly more complex than the shallow-water model employed in this study, that such a reduced-rank 4D-Var is perfectly capable of fulfilling the role of a standard 4D-Var, at least for twin experiments. It may even have an advantage over the full-rank 4D-Var with respect to the minimization cost and speed.

A decomposition of the background error covariance matrix given by Eq. (4) defines the preconditioning for the cost function of the reduced-rank 4D-Var, but also of the hybrid method. This particular decomposition is used for all the inner loops of variational minimization throughout the paper. Although such preconditioning is sufficient for the twin experiments, real applications generally require more sophisticated preconditioning in order to prevent poor convergence of the minimization.

For the experiments presented in this paper, there are five outer loops, each with five inner iterations. An
example of a zoom on the lowest values of $J$ changing with the number of iterations of the minimization procedure for a single assimilation window is presented in Fig. 2. The most significant cost function reduction occurs for the first iteration of each of the outer loops. The first outer loop is normally sufficient for convergence of the whole algorithm. Nevertheless, in some cases, due to a fixed number of iterations in the inner loop, convergence might be achieved later, as in the example shown in Fig. 2. Other choices of the respective numbers of inner and outer iterations could have been made but optimization of the efficiency of minimization is not the main issue here.

EVALUATION OF THE EXPERIMENTAL RESULTS

Within the framework of the EXP_100_SHORT configuration, two 4-yr data assimilation experiments were conducted. In the first, standard incremental 4D-Var keeps the background error covariance matrix constant throughout its duration. In the second, employing the hybrid method, the background error covariance matrix is updated between successive assimilation windows. We compared their results to a so-called free run, which is a simulation without assimilation starting from the background value of the initial condition (i.e., model trajectory corresponding to the 11–14 yr of simulation). Method performance was assessed by means of a relative error, defined as the ratio between the Euclidian norms of the analysis error and the error of the free run: $\|x^a - \hat{x}\| / \|x^\text{free} - \hat{x}\|$. The norm is computed separately for each of the three variables constituting the state vector ($u$, $v$, $h$) and, for a given time instant, is defined as $\|x^a - \hat{x}\| = \sqrt{1/N \sum_{i=1}^{N}(x^a_i - \hat{x}_i)^2}$, where the sum covers all spatial grid cells. A relative error significantly smaller than 1 will then reflect improvement because of data assimilation. In parallel, nonnormalized analysis errors (absolute errors) are also presented.

This relative error is displayed in Fig. 3 and clearly indicates a significantly better performance for the hybrid method than the 4D-Var. The relative error is decreased by roughly 50% for the observed variable, and 25% for the unobserved variables. The reason might be that the EOFs spanning the initial control subspace are obtained from a sample that is not over a long enough period to define a background error covariance matrix of sufficient quality. Therefore, keeping this matrix unchanged does not allow for a fully accurate correction of the model state by the 4D-Var, while making this initial basis evolve ensures significant improvement of data assimilation results. It is to be noted that the absolute errors of $h$ are larger than the expected 5 cm introduced by the observations. This is the result of a suboptimal experimental design, namely, lack of sophistication in the error basis update such as localization, so that the effect of the error subspace evolution could be clearly exhibited.

These results can be interpreted in terms of the quantity of information contained in the bases spanning control subspaces, or, in other words, the degree to which the control subspaces represent the true error directions. The 1825 reference EOFs (see above), which are assumed to contain almost full information on the variability of the model, were projected on the control subspaces employed
in our hybrid experiments, both in their initial and final forms (Fig. 4a). While the basis \( L_1 \) spanning the initial control subspace is distributed over the first 400–500 reference EOFs, an important redistribution of information content, owing to the evolution of the vectors spanning the control subspace, is easily noticeable. The basis \( L_{48} \) at the end of the 4-yr assimilation experiment (i.e., 48 one-month assimilation windows) accumulated a significant quantity of information and is much more focused than the initial basis on the first reference EOFs, which are those containing relevant information.

To be more specific, the quantity of information contained in a given basis \( L \) can be defined as \( Q(L) = \sum_{i=1}^{1825} \lambda^\text{ref}_i \left( \left| \sum_{i=1}^{1825} \lambda^\text{ref}_i \right| \right)^2 \), where \( \lambda^\text{ref}_i \) is the eigenvalue corresponding to the \( i \)th reference EOF, and \( p^i_b \) is the projection of this EOF on \( L \) (i.e., the value which is displayed along the \( y \) axis for \( L_1 \) and \( L_{48} \) in Fig. 4a). The ideal value for \( Q \) is 1, and is reached for the reference EOF basis \( L^\text{ref} \). When dealing with a basis of 100 vectors, this optimum value drops to \( Q(L^\text{ref}_1, \ldots, L^\text{ref}_{100}) = 0.757 \). In the present experiment, the information content corresponding to the initial basis defining the background error covariance matrix is \( Q(L) = 0.24 \), and increases to \( Q(L_{48}) = 0.46 \) at the end. This increase illustrates the positive effect of the evolution of the background error covariance matrix as ensured by the hybrid method.

This result is confirmed by looking at model variability in Fourier space. Thus, a Fourier analysis of the 10-yr reference model trajectory was performed, and a family of reference Fourier coefficients representing almost true model variability was constructed. Each Fourier coefficient was a 2D field defined on the model domain. Here again the first 100 coefficients were projected onto the control subspaces (Fig. 4b). It appears clear that the projection scores are higher for \( L_{48} \) than for \( L_1 \), which illustrates once more that the evolution of the basis improves the representativeness of the background error covariance matrix \( B \) with regard to variability in the true system.

4. Discussion

It has been shown in the preceding section that the hybrid method presented above can lead to significantly improved results in relation to 4D-Var. In this section, we try to investigate the reasons for this. With this purpose in mind, a schematic representation of the data assimilation process is useful. Ideally, data assimilation shifts a modeled system from some background state \( x^b \) to its true state \( x \) according to \( x = x^b + (x - x^b); x - x^b \) is the actual state error. In practice, an analysis of the form \( x = x^b + \sum \alpha_j \mathbf{L}_j \) is produced instead, which means that the state error is approximated in the subspace spanned by \( L = \{L_1, \ldots, L_m\} \), which is at the same time the control subspace for a given experiment. A vital feature of the hybrid method described in this paper is its reduced-rank formulation. Reduced-rank assimilation approaches consist in limiting \( r \), the number of columns of \( L \), to make the numerical problem tractable. Obviously, a major risk is to discover the actual error mostly in the subspace complementary to span(\( L \)). This results in a very limited correction of the background state and hence highlights the necessity of correctly specifying the basis \( L \).

A large- and mesoscale ocean circulation system is commonly described as a dynamical system with an attractor. The system trajectory always lies in the vicinity of the attractor, and cannot significantly depart from it. Within a given time window, only a portion of this attractor will be approached. This portion is (part of) a manifold, and can therefore be included in a reduced-dimension affine subspace. Let \( A \) be a basis of the corresponding vector subspace. The span(\( A \)) is of low dimension and describes most of the variability in the system. Thus, it stands to reason to consider \( A \) as an almost ideal error subspace basis. A challenge for a reduced-order assimilation method is to make the control subspace span(\( L \)) contain the largest possible fraction of the variability subspace span(\( A \)) at each assimilation time window. We may identify two important factors in this regard: the initialization of \( x^b \) and \( L \), and the ability of the system to make \( L \) evolve correctly so that span(\( L \)) significantly intersects span(\( A \)) at each assimilation window. We consider three different configurations for initialization as presented below. It should be noted that the configuration where the initial basis \( L \) does not provide any relevant information about true system variability is discarded here. It was shown in earlier studies, in the context of the SEEK filter, that the dynamics hardly
correct such misspecification (Ballabrera-Poy et al. 2001). Note also that if the system is subject to abrupt changes from one branch of the attractor to the other, the size of $L$ must be large enough to be able to correctly describe such transitions.

a. Almost perfect initialization of $L$

The results of the first experiment discussed in this section (configuration EXP_100_LONG) are shown in Fig. 5. The most remarkable finding is that 4D-Var performs strikingly well over the entire assimilation period. In the case of the observed variable, it roughly needs three assimilation cycles to stabilize its performance, after which it remains stable for the entire duration of the assimilation experiment. The unobserved variables require longer periods in order to settle, and their residual curves show a greater tendency to oscillate. These oscillations, however, are limited for the entire duration of the experiment (4 yr) and even at crests do not exceed the value of the error residual for the first assimilation window.

The 4D-Var performance is linked to the way $L$ has been initialized in this case. The initial basis contains the first 100 reference EOFs (associated with highest eigenvalues). It should be remembered that this reference set of EOFs introduced in section 3 is based on a 10-yr model run and contains 75% of the system variability spectrum, and for this reason the configuration is referred to as EXP_100_LONG. It is almost a perfect initialization in the sense that $\text{span}(A) \subset \text{span}(L)$. As a result, $L$ enables corrections to the background state in all the possible directions of variability in the true system. Because of the way in which the basis is constructed, this property holds true for the whole assimilation period.

As $\text{span}(A)$ accounts for most of the variability in the system, the $L$ basis that has been propagated by the full model would have the same properties as the initial basis. Therefore, one should not expect to find differences in performance between 4D-Var and the hybrid approach. It is indeed the case in Fig. 5 for about 6 initial assimilation cycles (200 days) and all 3 model variables. Nevertheless, the hybrid approach ensures propagation of $L$ with the help of the tangent linear approximation to the true nonlinear system. Since the variability of the former is more limited, basis evolution ensured by the hybrid method leads to a gradual deterioration of the results with respect to the static case (Fig. 5).
b. Short-term initialization of $L$ with an appropriate dimension

Next, the almost perfect initialization becomes partly deteriorated. We consider a case where the initial basis $L$ correctly represents short-term variability in the system, but is defective as far as representation of full system characteristics is concerned. In this case $\text{span}(A)$ and $\text{span}(L)$ partly overlap. Initially, the dimension of the error subspace basis is preserved at a high enough level to represent a large part of full-system variability. We could have realized, in section 4a, that 100 is an appropriate dimension in the case of our present shallow-water model. Such an initial configuration is exactly that illustrated by the test case described in section 3 (EXP_100_SHORT).

With the hybrid method, however, the (linearized) model proves capable of making $L$ evolve in a way that produces improved results compared with those of 4D-Var. It should also be noted that during this procedure the evolving basis of the error subspace accumulates a significant quantity of information on system variability (an increase from 0.24 at the beginning to 0.46 at the end of the experiment, as reported in section 3).

c. Short-term initialization of $L$ with a small dimension

Here we discuss even further deterioration of the definition of $L$. The initial basis is again representative of short-term variability in the system. It is constituted, however, by a number of vectors that is insufficient to represent a significant part of variability in the full system. Figure 6 shows the results for an example of such a configuration, with the initial basis formed by the first 30 EOFs (instead of 100 as it is the case in section 3) and referred to as EXP_30_SHORT. The hybrid method performs better than 4D-Var, although not as clearly as with 100 EOFs. Actually the difference between the two methods lies mostly in the 4D-Var results. In the case of 30 EOFs 4D-Var produces smaller error residuals than with 100 EOFs (Fig. 3), possibly due to a nonoptimal finding of the minimizer in the latter case. Another possible explanation is the higher sensitivity of the last EOFs to sampling errors. Results of the hybrid method, however, are rather similar, and slightly better in the case of a control subspace dimension equal to 100. This is in full agreement with the findings of Zhang et al. (2009), that is, that the results of the EnKF–4D-Var are less sensitive to the size of the ensemble than the 4D-Var alone. The information content of the 30-vector basis, as defined in the previous section, are $Q(L_1) = 0.18$ and $Q(L_{48}) = 0.21$ (cf. 0.24 and 0.46 for 100 vectors). In this case the information accumulated during evolution is very limited. The data assimilation system seems unable to incorporate a significant amount of it into a basis that is too small. But even this limited amount of information turns out to be sufficient to provide a relevant basis for each assimilation window.

d. Structure function evaluation

One way of illustrating the impact of the hybrid method is to analyze the evolution of the covariances of the background error (Thépaut et al. 1996). These structure functions can be assessed with their representatives, constituted by analysis increments resulting from single observation tests. Consequently, single observation tests were performed for an observation of $\delta h = 1$ m placed in the middle of the computational domain. They were obtained for an initial and evolved
(1 year later) form of the background error. In Fig. 7 we compare the resulting increments for the three experiments discussed in this paper. As a result of employing a limited sampling period in constructing the background error, long-distance, nonphysical correlations are present in the structure of some of the representers. They can be seen in the first column, rows 1 and 3 of Fig. 7 for short-term initializations of $L$, EXP_100_SHORT and EXP_30_SHORT, respectively. The representers of the evolved error covariance exhibit different characteristics. One year later, long-distance correlations have almost disappeared. At the same time, the representers have acquired a general structure like that obtained for an almost perfect initialization of background error. As can be seen in the second line of Fig. 7, such a representer does not possess long-distance correlations and these features are preserved by the evolution ensured in the hybrid method.

e. Synthesis

The results of the numerical experiments described in this section clearly indicate that the hybrid method has the potential for improving 4D-Var. Reasonably enough, the perfect $L$ initialization case (EXP_100_LONG) does not benefit from its evolution. Such a data assimilation experimental configuration seems to contain all the available information on the system and hence there is no room for improvement. Gradual deterioration of the hybrid results observed in this case is most probably linked to an approximate (i.e., linearized) evolution of $L$. In the cases where the initial $L$ represents only partial knowledge on the system, EXP_100_SHORT, however, there is plenty of room for improvement. Indeed, the evolving $L$ partly accumulates the missing knowledge. As a result, the hybrid method produces smaller error residuals than the corresponding 4D-Var for almost the entire duration of a 4-yr assimilation experiment. This holds true regardless of whether the dimension of the error space is sufficient to fully acquire the information unavailable at the initial stage (EXP_30_SHORT).

As far as computation costs are concerned, the hybrid method is obviously much more costly than the classic 4D-Var. Both methods have an equal number of iterations in their inner and outer loops and it is the update of the background error covariance matrix that makes the difference. This is not the case for the analysis step, which actually deals with inversion of matrices of a low dimension (not larger than 100 for the applications presented in the manuscript). In contrast, the propagation of the control subspace is costly, requiring as many executions of the tangent linear model as the dimension of the
conclusion. Note that this step could be performed in a parallel mode, which would substantially reduce computation time.

Finally, it is worth mentioning that the improvement observed due to hybridization cannot be achieved by tuning the variances of the background error covariance matrix, at least when the latter is built on the basis of the EOFs, as is the case for the experiments presented in this paper.

5. Conclusions

The fully consistent incremental 4D-Var–SEEK smoother presented in this paper is based on a hybridization of a reduced-rank incremental 4D-Var and an incremental, fixed-interval smoother algorithm derived from the SEEK filter. Consistency results from the theoretical equivalence of these two approaches that produce identical analyses of the system state at the beginning of an assimilation window, provided that input information is the same for both. Owing to the incremental formulation, this equivalence holds even for nonlinear dynamics. In addition, the smoother provides an analysis of the reduced-rank error covariance matrix, also at the beginning of an assimilation window. This last property, associated with the 4D-Var state analysis, forms the hybrid algorithm.

The 4D-Var–SEEK smoother hybrid has been implemented with a nonlinear shallow-water model whose dynamics mimic a simplified midlatitude ocean circulation. Control subspace has been initialized using different sets of the EOFs extracted from free-model simulations. In the case of imperfect initializations of \( B \), resulting from a limited rank of the error covariance matrix and a system that is a priori partly unknown, the hybrid method improves 4D-Var in terms of absolute error. One of the complementary experiments reveals, moreover, that even if the rank of the covariance matrix or the dimension of the control subspace is too low to represent or accumulate a significant amount of information on system variability, the hybrid is still able to provide relevant information at the appropriate times and performs better than the corresponding 4D-Var. An experimental configuration exhibiting a case where 4D-Var consistently produces better results than the hybrid has also been reported. This is the case when the initial basis, and hence background error covariance matrix, already provides a large part of the information on system variability.
Consequently, it may be concluded that the configurations owing considerable improvement to the hybrid algorithm are those in which the EOFs account exclusively for the short-term characteristics of a physical phenomenon. If the long-term characteristics are present, then basis evolution has no positive impact on data assimilation results and can even cause them to deteriorate. An optimal configuration for reduced-rank
data assimilation employing the EOFs to define the control subspace may therefore be described in the following manner. The vectors initialized by the EOFs describing long-term system behavior may be kept static, while those vectors initialized by the EOFs representing short-term system behavior should evolve. A clear-cut distinction between the two families of the EOFs depends on the problem and the configuration and opens perspectives for further investigation. These conclusions echo those of studies undertaken with the background error covariance matrix built as a linear combination of a static term reflecting behavior that is long term and dynamically updated to account for short-term variability (Hamill and Snyder 2000).

The experiments presented in this paper have all been conducted while neglecting the model error term in the propagation equation of the covariance matrix. This choice enhances the chances of investigating the very effect of the dynamical evolution of the covariance matrix, but the choice should be reconsidered in the future. It may be a complex task with the SEEK smoother, but adaptive schemes for tuning the covariance matrix could prove to be appropriate. One such a scheme has been explored by Zhang et al. (2009), and experiments performed by the authors using the Lorenz model (Lorenz 1996) have yielded encouraging results for this method. Adaptive strategies have also been explored in the SEEK filter framework (Testut et al. 2003; Brankart et al. 2009). Incorporating such schemes in the hybrid algorithm will be addressed in forthcoming research.

Acknowledgments. Monika Krysta would like to acknowledge support from the CNRS in the form of a postdoctoral grant. The contributions of Jacques Verron and Emmanuel Cosme were supported by the European Commission under Grant Agreement FP7-SPACE-2007-1-CT-218812-MYOCEAN. This research was also supported by the LEFE program from the CNRS. The authors would also like to express their gratitude to Céline Robert, Jean-Michel Brankart, and Arthur Vidard for having shared their time, knowledge, and experience during preparation of this paper. Finally, the authors thank two anonymous reviewers for careful revisions and for their suggestions on possible improvements of the manuscript.

REFERENCES


