Ocean Modelling 97 (2016) 129–140

Contents lists available at ScienceDirect

Ocean Modelling

journal homepage: www.elsevier.com/locate/ocemod

Comparison of the adjoint and adjoint-free 4dVar assimilation of the hydrographic and velocity observations in the Adriatic Sea

Max Yaremchuk[a,*], Paul Martin[a], Andrey Koch[b], Christopher Beattie[c]

[a] *Naval Research Laboratory at Stennis Space Center, USA*
[b] *Department of Marine Science, University of Southern Mississippi, USA*
[c] *Department of Mathematics, Virginia Tech, USA*

ARTICLE INFO

Article history:
Received 12 March 2015
Revised 25 September 2015
Accepted 12 October 2015
Available online 10 November 2015

Keywords:
Data assimilation
Ajoint analysis
Regional modeling

ABSTRACT

Performance of the adjoint and adjoint-free 4-dimensional variational (4dVar) data assimilation techniques is compared in application to the hydrographic surveys and velocity observations collected in the Adriatic Sea in 2006. Assimilating the data into the Navy Coastal Ocean Model (NCOM) has shown that both methods deliver similar reduction of the cost function and demonstrate comparable forecast skill at approximately the same computational expense. The obtained optimal states were, however, significantly different in terms of distance from the background state: application of the adjoint method resulted in a 30–40% larger departure, mostly due to the excessive level of ageostrophic motions in the southern basin of the Sea that was not covered by observations.

1. Introduction

In recent years the ensemble approach to data assimilation has developed rapidly due the growth of computer power made available through massive parallelization. An attractive feature of the ensemble technique is its ability to probe the structure of a dynamical system and assess the sensitivity of its outputs (e.g., measured quantities) to variations in poorly known inputs (e.g., initial conditions) without using the adjoint code. In particular, feasibility of the ensemble method to the estimation of sensitivities was demonstrated in meteorological (Ancell and Hakim, 2007; Torn and Hakim, 2008) and oceanographic (Yaremchuk and Martin, 2014) applications.

Another important advantage of the ensemble approach is the opportunity it offers to treat the numerical model as a black box, thus avoiding the burden of the development and maintenance of the tangent linear and adjoint codes required by the variational methods (e.g., Le Dimet and Tlagrand, 1986). Employing this property, Anderson et al. (2009) and Hoteit et al. (2013) developed the data assimilation research testbed (DART) system on the basis of the widely used ensemble Kalman filter (EnKF).

Recently, significant progress also has been made in extending the EnKF technique into the particle filtering framework (e.g., Hoteit et al., 2012) and in coupling EnKF techniques with 3d and 4d variational methods (e.g., Zupanski, 2005; Liu et al., 2008; Zhang et al., 2009). Of particular interest in the present context has been the development of the Maximum Likelihood Ensemble Filter (Zupanski, 2005) based on the explicit computation of the square root of the Hessian matrix in the subspace spanned by the ensemble members.

Merging ensemble approaches with variational techniques has developed along two lines: 1) improvement of the background error covariances (BECs) through introduction of the ensemble-based estimates and/or their hybrid generalizations (Clayton et al., 2013; Kuhl et al., 2013) and 2) searching for the optimal solution within the subspaces spanned by the leading error modes of the BECs, a technique pursued by many authors in the last decade (e.g., Liu et al., 2008; Zhang et al., 2009; Zhang and Zhang, 2012; Trevisan et al., 2010). This assumption implicitly assumes that the BEC structure is well described by these (possibly localized) modes. More recently, performance the adjoint-free family of methods (4dEnVar) based on the formulation by Liu et al. (2008) has been compared with the 4dVar technique in the framework of idealized experiments with the Lorenz-05 model (Fairbairn et al., 2014). The results show a significantly better performance of the 4dEnvar for moderate-length assimilation windows with low-density observations. Desroziers et al. (2014) demonstrated a close relationship between the 4dEnVar and 4dVar state space formulations and compared various implementations of 4dEnVar with 4dVar in an idealized setting.

The above cited developments mostly deal with meteorological applications, where the ensembles are supported by significantly denser data than are available in the ocean. High data density allows one to obtain reasonably good estimates of the BECs from the ensemble using truncated representation of the localization matrices and to efficiently compute the cost function gradient on the model

[*] Corresponding author. Tel.: +1 2286885259.
E-mail address: max.yaremchuk@nrlssc.navy.mil (M. Yaremchuk).

http://dx.doi.org/10.1016/j.ocemod.2015.10.010
1463-5003/Published by Elsevier Ltd.

grid directly from ensemble perturbations (Liu et al., 2009; Tian and Xie, 2012). A significant advantage of such an approach is the absence of the necessity to develop and maintain tangent linear and adjoint codes and its flexibility in adaptation to various dynamical constraints.

In the ocean, the ensemble-based BEC estimates tend to be less accurate, and one has to rely on ad hoc BEC representations (e.g., Mirouze and Weaver, 2010; Yaremchuk and Sentchev, 2012). Development of an efficient adjoint-free assimilation method also becomes more problematic as one has to select a few reliable perturbations with more care. Early attempts to develop practical a4dVar algorithms in oceanography were limited to predetermined low-dimensional subspaces spanned either by the reduced-order approximations of the model Green's functions (Stammer and Wunsch, 1996; Menemenlis and Wunsch, 1997), or by the dominant principal component vectors (EOFs) associated with the model statistics (e.g., Robert et al., 2005; Qui et al., 2007; Hoteit, 2008). In fact, the 4dEnVar technique pursues a similar, but more general approach, parameterizing the search subspace by Schur products of the ensemble members with the eigenvectors of the reduced-order representation of the localization matrix.

In the present paper, the a4dVar approach of Yaremchuk et al. (2009) is tested against the observation space 4dVar. Both algorithms are dynamically constrained by the Navy Coastal Ocean Model (NCOM) and applied to the real observations collected in the Adriatic Sea in August, 2006. A specific feature of the a4dVar method tested here is that it employs an iterative search over a sequence of low-dimensional subspaces to find the cost function minimum.

The paper is organized as follows. In the next section we briefly describe the 4dVar methodology, outline the a4dVar method to be tested, and provide a detailed description of the experimental setting. In Section 3 performance of the 4dVar and a4dVar methods is compared in terms of the convergence rate, forecast skill, computational expense, and particular properties of the optimized solutions. Conclusions and discussion of the results are presented in Section 4.

2. Methodology

2.1. 4dVar assimilation

For the sake of clarity, consider the 4dVar method as the following linear discrete least-squares problem constrained by model dynamics in a small vicinity of the model's background trajectory $\mathbf{x}_b^n$:

$$J = \frac{1}{2}\left[\mathbf{x}^{0\mathsf{T}}\mathbf{B}^{-1}\mathbf{x}^0 + \sum_{n=0}^{N}(\mathbf{H}_n\mathbf{x}^n - \mathbf{d}^n)^{\mathsf{T}}\mathbf{R}_n^{-1}(\mathbf{H}_n\mathbf{x}^n - \mathbf{d}^n)\right] \to \min_{\mathbf{x}^0}. \quad (1)$$

where n enumerates observation times, $\mathbf{B}$ is the background error covariance matrix of $\mathbf{x}_b^0$ which describes the (Gaussian) error statistics of the model state at $n = 0$, $\mathbf{H}_n$ is the model-data projection operator, $\mathbf{d}^n$ is the misfit $\mathbf{y}_n^o - \mathbf{H}_n\mathbf{x}_b^n$ between observations $\mathbf{y}_n^o$ and the corresponding background model values, $\mathbf{R}_n$ is the observation error covariance, and $^{\mathsf{T}}$ denotes transposition. Further below we denote the dimension of the discretized model state vector $\mathbf{x}$ by M and the total number of observations by K.

The optimal correction vector $\mathbf{x}^n$ is governed by the recursive relationship $\mathbf{x}^n = \mathbf{M}_n\mathbf{x}^{n-1}$, where $\mathbf{M}_n$ is the dynamical operator of the model linearized in the vicinity of the background trajectory $\mathbf{x}_b^n$ at the time interval (t_{n-1}, t_n), so that

$$\mathbf{x}^n = \mathbf{M}_n\mathbf{M}_{n-1}\ldots\mathbf{M}_2\mathbf{M}_1\mathbf{x}^0. \quad (2)$$

Introducing new notation $\mathbf{c} = \mathbf{x}^0$ for the control vector, $\mathbf{M}^n \equiv \mathbf{M}_n\ldots\mathbf{M}_2\mathbf{M}_1$ for the aggregated n-step propagator, $\overline{\mathbf{H}}_n = \mathbf{R}_n^{-1/2}\mathbf{H}_n$, $\overline{\mathbf{d}}^n = \mathbf{R}_n^{-1/2}\mathbf{d}^n$, omitting over-bars, and taking (2) into account, the minimization problem (1) can be rewritten in terms of the optimal correction $\mathbf{c}$ to the initial state:

$$J = \frac{1}{2}\left[\mathbf{c}^{\mathsf{T}}\mathbf{B}^{-1}\mathbf{c} + \sum_{n=0}^{N}(\mathbf{H}_n\mathbf{M}^n\mathbf{c} - \mathbf{d}^n)^{\mathsf{T}}(\mathbf{H}_n\mathbf{M}^n\mathbf{c} - \mathbf{d}^n)\right] \to \min_{\mathbf{c}}. \quad (3)$$

A 4dVar data assimilation method finds the minimum of J by solving the normal equation:

$$\nabla_{\mathbf{c}}J = \mathbf{B}^{-1}\mathbf{c} + \sum_n \mathbf{M}^{n\mathsf{T}}\mathbf{H}_n^{\mathsf{T}}(\mathbf{H}_n\mathbf{M}^n\mathbf{c} - \mathbf{d}^n) = 0, \quad (4)$$

To simplify further treatment, introduce the following notation for the Hessian matrix $\tilde{\mathbf{H}}$ and the right-hand side $\mathbf{b}$,

$$\tilde{\mathbf{H}} = \mathbf{B}^{-1} + \sum_n \mathbf{M}^{n\mathsf{T}}\mathbf{H}_n^{\mathsf{T}}\mathbf{H}_n\mathbf{M}^n; \quad \mathbf{b} = \sum_n \mathbf{M}^{n\mathsf{T}}\mathbf{H}_n^{\mathsf{T}}\mathbf{d}^n, \quad (5)$$

which define the solution of the normal equation $\tilde{\mathbf{H}}\mathbf{c} - \mathbf{b} = 0$.

2.1.1. Adjoint techniques

There are two major approaches to 4dVar assimilation. The first one (the so-called "state space approach") typically solves (4) iteratively through a conjugate gradient descent or related algorithm, which requires on every iteration the formation of a matrix-vector product using the Hessian, $\tilde{\mathbf{H}}$, or an equivalent process, which in either case involves application of the non-linear model operator, $\mathbf{M}^{\mathsf{T}}$ (the "adjoint model") and (in many cases) the linearized model $\mathbf{M}$. The method is widely used in a number of community OGCMs (MIT, ROMS), and in operational meteorology (ECMWF).

Such iterative solution approaches generate a sequence of residuals $\mathbf{r}_i = \tilde{\mathbf{H}}\mathbf{c}_i - \mathbf{b}$, which is the cost function gradient (4) evaluated at the current solution iterate, $\mathbf{c}_i$. However, knowledge of the steepest descent direction for the cost function requires access to the adjoint model, $\mathbf{M}^{\mathsf{T}}$; note that it enters the expressions (5) for both the Hessian and the right-hand side of the normal system.

Numerically, the procedure of calculating $\mathbf{r}_i$ involves two major steps:

(1) Sequential calculation of $\mathbf{x}_i^n = \mathbf{M}^n\mathbf{c}_i$ (forward run of the tangent linear model) supplemented by additional calculation of the model-data misfits

$$\mathbf{q}_i^n = \mathbf{H}_n^{\mathsf{T}}(\mathbf{H}_n\mathbf{x}_i^n - \mathbf{d}^n). \quad (6)$$

(2) Summation of the products $\mathbf{M}^{n\mathsf{T}}\mathbf{q}_i^n$ conveniently performed in reverse-time order, because $\mathbf{M}^{n\mathsf{T}} = (\mathbf{M}_n\ldots\mathbf{M}_1)^{\mathsf{T}} = \mathbf{M}_1^{\mathsf{T}}\ldots\mathbf{M}_n^{\mathsf{T}}$. This corresponds to backward-in-time integration of the adjoint model forced by $\mathbf{q}_i^n$.

The second approach to 4dVar ("observation space method") appears to be more rigorous in that it can include more easily explicit treatment of the model errors $\mathbf{e}^n = \mathbf{x}^n - \mathbf{M}_{n-1}\mathbf{x}^{n-1}$. In this formulation, the cost function (1) is augmented with an additional term $\sum_n \mathbf{e}^{n\mathsf{T}}\mathbf{B}_n^{-1}\mathbf{e}^n$ involving model errors. In other words, the background error is explicitly separated into the components associated with the uncertainty $\mathbf{B}_0$ of the initial state, and the uncertainty $\mathbf{B}_n$ of the model equations/forcing. The normal equation in this case is more complicated than (4), and can be solved numerically using the representer method (e.g., Bennett, 2002; Rosmond and Xu, 2006). The latter is closely related to the optimal interpolation technique as it seeks a solution of the normal equation in the form of a linear mapping of the model-data misfits, $\mathbf{d}^n$ on the control space. Mathematically, the approach employs the Sherman-Morrison-Woodbury formula to transform the Hessian inverse from the state space to the observational space, which has a smaller dimension in oceanographic applications. On the other hand, minimization of the (non-linear) cost function in the observation space contains two embedded loops, and involves multiple convolutions with the non-sparse matrix $\mathbf{B}$, making the method sometimes more computationally expensive. It has been implemented in ROMS (Moore et al., 2011) as an optional feature, and in the Naval Research Laboratory for both atmospheric (Xu and Rosmond, 2004; Xu et al., 2005) and oceanic (Ngodock and Carrier, 2014) data assimilation systems.

2.1.2. *Adjoint-free methods*

With the ongoing trend toward parallelization in computer technologies, directly perturbing a large number of control variables becomes computationally feasible, allowing the replacement of tangent linear and adjoint codes by the finite differences of perturbed numerical models (”ensemble members”) that may be run in parallel. For example, the fast-developing 4dEnVar method (e.g., Liu et al., 2009; Buehner et al, 2010; Desroziers et al., 2014) restricts the search for an optimal increment to a pre-determined subspace spanned by the eigenvectors of a localized ensemble covariance matrix.

When skillful ensembles are unavailable, a similar search could be performed using the leading eigenvectors of an ad hoc model of $\mathbf{B}$, provided that search directions (SDs) are kept orthogonal with respect to the inner product associated with the Hessian matrix (Appendix B). The adjoint-free (a4dVar) method considered here follows this approach through iterative minimization of the cost function in a sequence of subspaces spanned by the SDs

$$\mathbf{s}^n := \left[\mathbf{B}^{-1} + \mathbf{H}_n^T\mathbf{H}_n\right]^{-1}\mathbf{q}^n = \mathbf{B}\left[\mathbf{I} + \mathbf{H}_n^T\mathbf{H}_n\mathbf{B}\right]^{-1}\mathbf{q}^n; \quad n = 0, .., N, \quad (7)$$

where $\mathbf{I}$ is the identity operator in state space.

Compared to the search directions defined by the leading eigenvectors of $\mathbf{B}$, the a4dVar subspaces spanned by $\{\mathbf{s}^n\}$ contain extra information on the structure of the Hessian (through the observation operators $\mathbf{H}_n$) and on the magnitudes of model-data misfits $\mathbf{q}^n$ at the current iteration. The overall a4dVar strategy is to replace the *single* SD obtained by projecting model-data misfits $\mathbf{q}^n$ on the initial condition with the adjoint model by parallel searches in *multiple* directions (7). Although these directions are unlikely to include the direction of *local* steepest descent, they implicitly accumulate information on the Hessian structure through the $\tilde{\mathbf{H}}$-orthogonalization process and may therefore be competitive in efficiency with the adjoint method in a typical oceanographic application.

Feasibility of the presented a4dVar approach is also motivated by the following considerations:

(1) the search subspaces can be easily $\tilde{\mathbf{H}}$-orthogonalized to each other to avoid redundant “nearly parallel” searches in the directions that have been largely explored, and, therefore, bring only minor reduction of the cost function;
(2) search subspaces spanned by $\{\mathbf{s}^n\}$ are unlikely to be strictly $\tilde{\mathbf{H}}$-orthogonal to the cost function gradient and may provide a larger overall projection on the direction towards the cost function minimum than the direction of steepest descent;
(3) adjoint and tangent linear codes for the state-of-the art GCMs are never exact and require several times more computational resources compared to direct model runs, providing an incentive to avoid full-adjoint model evaluation and to take advantage of the parallelism afforded by the multiple direct model runs used by a4dVar.

Note that the number of SDs is not restricted to the number of observation times. In principle, a simultaneous search can be performed in all the directions corresponding to every single observation within the assimilation window. However, care should be taken to exclude ”nearly parallel” SDs from consideration. This can be accomplished by extracting the leading eigenvectors from the given set of search directions. In the reported experiments, we assigned a SD to all observations collected at an observation time.

Plausible forms of SD generation are not restricted to (7). For instance, search subspaces could be built using the leading modes of the model trajectory at subsequent iterations provided the SDs are $\tilde{\mathbf{H}}$-orthogonal and have sizable projections on the local gradient. This strategy has been widely used in the early versions of ensemble-4dVar techniques (Qui et al., 2007; Hoteit, 2008) and proved to be rather effective in a number of applications including the tested a4dVar method (Yaremchuk et al., 2009; Panteleev et al., 2015).

As an alternative option, one can simply use eigenvectors of $\mathbf{B}$ in the descending order of their eigenvalues to build the search subspaces. In the latter case, the convergence rate for the respective a4dVar algorithm can be estimated (more details on the subject can be found in Appendix A). However, this approach demonstrated a significantly slower convergence in the realistic application considered below. We partly attribute this behavior to the above mentioned absence of information on the (unevenly distributed) observational locations in the subspaces spanned by the eigenvectors of $\mathbf{B}$.

The primary purpose of the present study is to test the performance of the a4dVar algorithm in application to a realistic oceanographic data set constrained by a state-of-the-art numerical model and compare its performance with a standard 4dVar approach. The method is based on (7) and outlined as follows:

0. Specify the dimension m_s of the search subspaces, their number k to be kept in memory for $\tilde{\mathbf{H}}$-orthogonalization, the maximum number of iterations I, the perturbation magnitude ε and the background model trajectory $\mathbf{x}_b^n$. Set the iteration number i to zero, $\mathbf{c}_0 = 0$, and compute $\overline{\mathbf{d}}^n$.
1. Compute $\mathbf{x}_i^n, J_i, \mathbf{Y}_i = \tilde{\mathbf{H}}^{1/2}\mathbf{c}_i$ and the search directions $\mathbf{s}_i^n$ Eq. (7).
2. Extract the m_s leading EOFs $\mathbf{p}_i^m, m = 1, \ldots, m_s$ of the search directions to form the basis in the search subspace.
3. Perturb the initial conditions $\mathbf{c}_i \to \mathbf{c}_i + \varepsilon\mathbf{p}_i^m$ and run (in parallel) the ensemble of m_s perturbed models, computing the respective perturbed values of J_i^m and $\mathbf{Y}_i^m$.
3. $\tilde{\mathbf{H}}$-orthogonalize the search basis $\{\mathbf{p}_i^m\}$ with respect to at most k basis vectors obtained on the previous iterations and compute optimal corrections $\delta\mathbf{c}_i$ (see Appendix B).
4. Set $\mathbf{c}_{i+1} = \mathbf{c}_i + \delta\mathbf{c}_i$.
5. If $i = I$ exit. Otherwise set $i \leftarrow i + 1$, then go to 1.

The above listed algorithm was configured for the purpose of comparison with 4dVar both in terms of accuracy and computational expense. For this reason the value of I was set by the total CPU time required for convergence of the 4dVar algorithm. Parameters m_s, k, and ε were selected by experimentation. The $\tilde{\mathbf{H}}$-orthogonalization procedure and computation of $\delta\mathbf{c}_i$ are based on the technique proposed by Zupanski (2005), with more details given in Appendix B.

Comparison of 4dVar and a4dVar is done by constraining NCOM with observations collected in the Adriatic Sea in August, 2006. Description of the model, data, and parameters of the assimilation algorithms are given below.

2.2. *Model and data*

2.2.1. *The model*

The NCOM is a free-surface primitive-equation hydrostatic ocean model with σ coordinates in the upper layers and, optionally, fixed depths below a user-specified distance from the surface. Algorithms that comprise a NCOM computational kernel are described in detail by Martin (2000) and with some improvements by Morey et al. (2003) and Barron et al. (2006). The vertical mixing model utilized is the Mellor-Yamada level 2 closure scheme (Mellor and Yamada, 1974) and the equation of state of Mellor (1991) is used. Biharmonic horizontal diffusion is prescribed implicitly via third-order upwind advection (Holland et al., 1998).

The model was configured at 3 km resolution on an 85 × 294 horizontal grid (Fig. 1) with 32 levels in the vertical. The top 22 σ levels follow the bathymetry, stretching from the surface to a fixed depth of 291 m, and 10 fixed-depth levels are used below 291 m.

Initial and open boundary conditions for the sea surface height ζ, temperature T, salinity S, and horizontal velocities u, v were provided

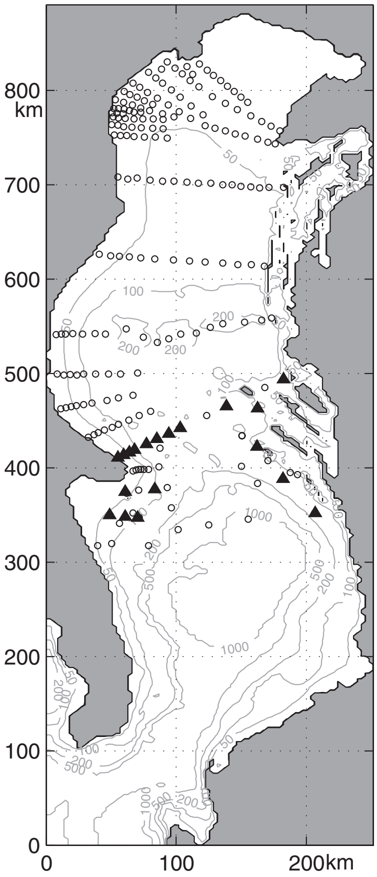

Fig. 1. Model domain with CTD stations (circles) and moorings (triangles) of the DART experiment. Gray contours (m) show bottom topography.

from the global NCOM (Barron et al., 2004) solution for the region. Tidal forcing is not used. The model was forced by the river runoff and atmospheric fields derived from the regional ALADIN atmospheric model with 8 km horizontal resolution (Ivatek-Sahdan and Tudor, 2004).

In the described assimilation experiments, initial conditions were used as control variables, i.e., the vector **c** comprised all the grid point values of ζ, T, S, u, v at $n = 0$. With the given 3-dimensional grid and bathymetry, the inverse problem has M=1,493,570 unknowns.

The first guess (background) values of **c** were taken from the NCOM simulation described by Martin et al. (2009) and then adjusted to suppress temperature and salinity biases during the assimilation time interval (0.00 UTC on 08/14 to 0.00 UTC on 08/29/2006). After the adjustment, the horizontal-and-time average misfits between the background solutions and *TS* observations did not exceed 0.02 °C and 0.005 psu, respectively.

Vertical profiles of the NCOM standard deviations (Fig. 2) demonstrate quantitative similarity between the variability of the first guess model solution and the mean variability at the observation points. However, the root-mean square (rms) model-data misfits were found to be of the same order of magnitude as the observed variability, indicating that the model has limited simulation skill of the smaller-scale features without further adjustment of the poorly known parameters (e.g., initial conditions) provided by data assimilation.

It is also noteworthy, that the obtained background solution provides a rather challenging environment for variational assimilation, as it contains a large number of transient mesoscale features induced by in wind forcing events and instabilities of the coastal boundary currents (see, for example, Figs. 7 and 8 in Burrage et al., 2009). In particular, in the upper 150 m (sampled by 90% of observations) the ratio of the time-averaged magnitudes of non-linear to linear terms in the momentum equation was close to 0.3, indicating a considerable degree of dependence of the tangent linear and adjoint models on the background solution.

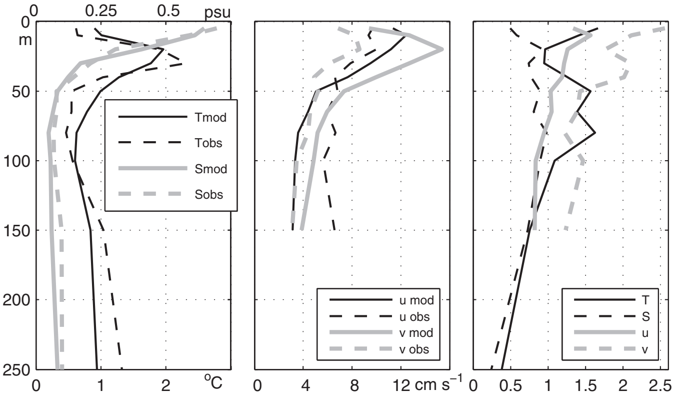

Fig. 2. Vertical profiles of the horizontal and time averaged rms variations of temperature and salinity (left) and velocity components (middle) derived from the background model run and observations. Right panel shows vertical profiles of the rms model-data differences normalized by the respective rms variabilities of observations.

2.2.2. DART06 experiment

The data used in this research were acquired in the course of a collaborative field experiment in the central Adriatic, the Dynamics of the Adriatic in Real Time (DART) (Martin et al., 2009; Burrage et al., 2009). In the present study, ADCP and CTD observations from August 14 to August 29, 2006 are used (Fig. 1).

Current velocities u, v were measured by 19 moored ADCPs at locations shown by triangles in Fig. 1. Due to the inaccuracy of ADCP measurements in proximity to the surface and bottom, the data spanned the depth range from 15 to 150 m. All the velocity data were detided and averaged over 29, 12 h intervals centered at the assimilation times t_n of 0 and 12 UTC. The average duration of an ADCP time series employed in the data assimilation experiments was 12.7 days.

Temperature T and salinity S were measured at 219 CTD stations occupied in the northern and central parts of the basin. As it can be seen from Fig. 1, most of the CTD soundings (216) were shallower than 280 m, with only a three casts taken at deeper locations. The total number of TS observations used in the assimilation is 9650. With the total number of the observed velocities 13,856 the dimension of the observation space was $K = 23{,}506$.

2.3. Assimilation parameters

In the course of the experiments described in the next section, we tried to keep the parameters of the tested 4dVar and a4dVar systems as close as possible to each other. However, due to the different formulations (observation space for 4dVar and state space for a4dVar), certain discrepancies remained in the shape of the background error covariance **B**. In both algorithms **B** is represented by the product **VCV** where **V** is the diagonal matrix of the background error rms variances and **C** is the respective correlation matrix.

In the 4dVar algorithm, **C** is represented implicitly by the kernel of the heat transfer equation

$$\mathbf{C} \simeq \exp\left(\frac{1}{2}a^2\Delta\right), \tag{8}$$

where a is the decorrelation length scale and Δ is the discretized 2d Laplacian operator. Numerically, the action of **C** on a state vector is computed by integrating the heat transfer equation (e.g., Weaver and Courtier, 2001). In the vertical, the decorrelation scale was set to zero. The correlation matrix (8) is rank-deficient, so the 4dVar solution is obtained in the range of **B**.

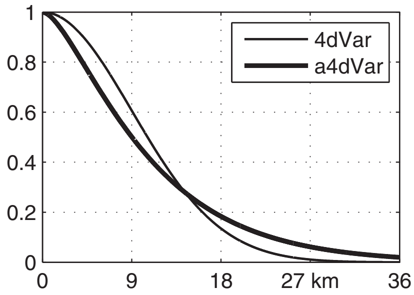

Fig. 3. The background error correlation functions.

The a4dVar algorithm is formulated in the state space and requires a definition of $\mathbf{C}^{-1}$ which was explicitly specified as the inverse of the second-order approximation of the exponent in (8)

$$\mathbf{C}_a^{-1} = \left[\mathbf{I} - \frac{1}{2}b^2\Delta\right]^2. \tag{9}$$

where $\mathbf{I}$ is the identity operator in state space, and $b = \sqrt{8/\pi}\,a$ to preserve the value of the integral decorrelation scale specified in 4dVar (e.g., Yaremchuk and Smith, 2011).

Although the respective correlation functions are somewhat different in shape (Fig. 3), we assume that this difference has minor effects on the overall results of the assimilation compared to the effect of non-linearity.

The rest of the assimilation parameters were identical for both the 4dVar and a4dVar assimilation systems. The value of a was chosen to be 9 km, consistent with typical estimates of the baroclinic deformation radius in the region (e.g., Cushman-Roisin and Korotenko, 2007). The background error rms variances (diagonal elements of $\mathbf{V}$) were assumed to be proportional to the rms variability of the respective NCOM fields from the first guess solution with the typical errors of 1 °C, 0.1 psu, and 10 cm/s near the surface. Observation errors were assumed to be spatially uncorrelated with the rms variances (diagonal elements of $\mathbf{R}^{1/2}$) dependent only on depth in the manner shown by the solid profiles in Fig. 2). Actual vertical distributions of the observational rms error variances were specified by multiplying these curves by 0.33 for temperature and salinity, and by 0.5 for velocity.

The a4dVar EOF analysis was performed with respect to the diagonal metric specified by the inverse background error variances $\mathbf{V}^{-2}$.

The stopping criteria for the iterative processes were selected as follows: for the 4dVar system the solution of the system for the representer coefficients was terminated after n_t=7 iterations, when the accuracy of the conjugate gradient (CG) solver was, as a rule, below 10^{-3}. Experiments with the larger number of CG iterations (inner loops) have shown only minor effects on the final optimal solution, whereas reduction of n_t resulted in amplification of the effects related to the instabilities of the tangent linear model and/or its adjoint, causing a sharp increase of the cost function after 3–4 outer loops. With the value of n_t=7, 8–10 outer loops were executed before the values of J started to increase.

For the a4dVar system, the minimization was terminated when the total CPU time reached the value used by the respective 4dVar experiment. The number of ensemble members was kept constant at $m_s = 9$ through all the experiments.

An important technical issue in a4dVar was the choice of the perturbation magnitude ε. Ideally, this value should be as small as possible to keep the perturbed system linear. In practice, values of ε significantly below 10^{-2} were ineffective due to the loss of accuracy in calculating the perturbations of $\mathbf{Y}_i$, especially in the course of 14 days of model integration. For that reason, on each iteration, the value of ε was selected in a way that only *one field* among all the m_s perturbations $\mathbf{p}_i$ could reach its critical magnitude at one point of the domain. The respective critical values of the temperature, salinity, and velocity perturbations were set to 4 °C, 2 psu, and 0.5 m/s, respectively. This strategy allowed fast convergence while avoiding development of instabilities within the perturbed model runs.

Preliminary experiments were also performed to tune the number of search subspaces k to be kept in memory along with the vectors $\mathbf{Y}_i$ needed for $\tilde{\mathbf{H}}$-orthogonalization. The large number of elements in $\mathbf{Y}$ made the orthogonalization process rather time-consuming for $k > 10$. Besides, $\tilde{\mathbf{H}}$-orthogonality was quickly lost with iterations due to strong non-linearity of the model and the above mentioned inaccuracy in estimating the perturbations of $\mathbf{Y}$ due to the finite value of ε. After some experimentation, it was found that $k = 3$ with the above mentioned strategy of selecting ε provided the fastest convergence for the a4dVar algorithm.

The EOF reduction of the search subspace (step 2 in the layout of Section 2.1.2) may seem to be redundant for a linear system, but appears to be important in the considered a4dVar application: First, performing the search along the few principal modes extracted from the time sequence $\{\mathbf{r}^n\}$ tends to keep the minimization process within the most persistent (geostrophically and hydrostatically balanced) manifold, thus avoiding searches over initial states that tend to generate excessive ageostrophic (i.e.,smaller-scale) motions. Second, rescaling the EOF metric $\mathbf{V}$ proved to be useful in restarting the $\tilde{\mathbf{H}}$-orthogonalization: rescaling was done every time when the relative reduction $\gamma = \delta J/J$ of the cost function at the start of the new orthogonalization cycle was ten times smaller than the mean value of $\overline{\gamma}$ on the previous cycle. In the event $\gamma < 0.1\overline{\gamma}$, salinity entries of $\mathbf{V}$ were inflated by the factor of 5 and then restored to their original values on the next occurrence of the event.

3. Results

In the reported experiments we varied the length of the assimilation window from short (4 days, $N = 9$) to moderate (8 days, $N = 17$) and long (14 days, N=29) duration. Performance of the assimilation algorithms was evaluated in three categories: the forecast skill at the end of the assimilation window (for $N = 917$), the rate of convergence, and by qualitative inspection of the optimal model trajectories.

3.1. Convergence rates and computational expense

To assess the rates of convergence, one has to have an ability to compare the reduction of the cost function with iterations, which is not straightforward for two reasons.

First, in the 4dVar algorithm considered here, the regularization term of the cost function can be evaluated only within the range of the correlation matrix defined by (9). To avoid the burden of restricting $\mathbf{C}_a$ to the range of $\mathbf{C}$, we compared only the observational parts of the 4dVar and a4dVar cost functions (second term in Eq. (3)).

Second, the number of iterations required for convergence cannot be considered as an objective criterion because 4dVar and a4dVar iterations are different in nature. Due to the non-linearity of the problem, an iteration (either 4dVar or a4dVar) performs minimization in the vicinity of the current (suboptimal) state, but 4dVar does that in the range of $\mathbf{B}$, whereas a4dVar minimizes in the subspace of a much smaller dimension spanned by $\mathbf{p}^m$. For that reason, iterations require quite different computational resources and should be compared in terms of CPU time.

Fig. 4 shows such a comparison by rescaling the horizontal axis with the total CPU time τ_a required by one a4dVar iteration. The value of τ_a was 11 times larger than the CPU time τ_m of a direct NCOM model run for a given experiment, i.e. $\tau_a \simeq 11\tau_m$. The major contribution to τ_a is given by the ensemble run ($9\tau_m$, p.3 in the layout of Section 2.1.2), while the master NCOM run (p.1) and operations listed in pp.2 and 4 require τ_m and $0.8\tau_m$, respectively. Overall, convergence was achieved at an expense of 60–70 iterations (650–800 NCOM runs).

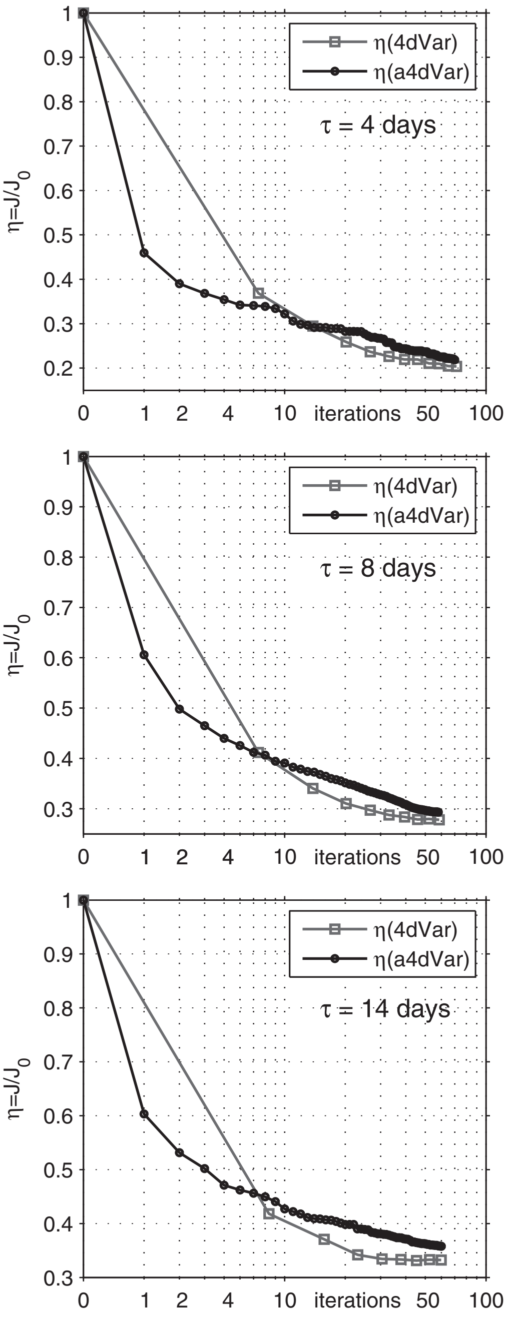

Fig. 4. Relative reduction η of the cost function with iterations (marked by circles) for different assimilation periods. The horizontal axis is scaled by the CPU time required for the a4dVar iteration. Squares label the 4dVar outer loops.

As may be seen in Fig. 4, a single 4dVar iteration was approximately equivalent to 6–7 a4dVar iterations, or 70–80 direct model runs. This computational expense arises because sequential execution of the adjoint and tangent linear codes (inner loops of the CG solver) required around $11\tau_m$, whereas one 4dVar outer loop included seven inner loops to solve the system of linear equations for the representer coefficients. In a series of experiments, it was found that executing seven inner loops provided a 10^3-fold reduction of the system's residual and was optimal with regard to the total CPU time of the 4dVar algorithm.

Fig. 4 shows that, in general, the tested a4dVar method is computationally comparable to the observation space 4dVar. Although the total CPU time required for reduction of J by the factor of 0.4 (attained after the first outer loop of the 4dVar) appears to be similar for the 4dVar and a4dVar methods, the a4dVar minimization noticeably slows down at subsequent iterations, especially for longer assimilation windows (τ=8,14 days). This could be partly explained by the fact that, with longer windows, operators $\mathbf{M}^n$ tend to depart farther away from the identity and it becomes increasingly more difficult for the a4dVar algorithm to find the minimum without the additional information on the structure of $\tilde{\mathbf{H}}$ provided by the adjoint code in 4dVar.

Fig. 5 demonstrates the time evolution of the quantities

$$f_q^n = \left\langle \left[(\mathbf{H}_n \mathbf{x}_q^n - \mathbf{d}_q^n)^\top (\mathbf{H}_n \mathbf{x}_q^n - \mathbf{d}_q^n)/n_q \right]^{1/2} \right\rangle \qquad (10)$$

characterizing the daily averaged $\langle\rangle$ model-data misfits of the various state vector components before (black lines) and after (gray lines) optimization with a 14-day assimilation window (i.e., using all the available data). The subscript q takes the values of the labels in the mid-bottom parts of Fig. 5 which indicate the observed variables (temperature, salinity and velocity vector) for which the statistics f_q were computed, whereas n_q stands for the total number of respective observations taken at a given day.

Comparison of the model-data differences for the background forecast (thick black lines in Fig. 5) with those computed for persistence (thin lines) shows their approximate similarity, especially for f_T and f_S. This similarity can be partly explained by small biases in the temperature and salinity fields of the background solution and large discrepancies in representation of the mesoscale structures by the background solution. As a consequence, persistence assumption appears to be much less valid for the velocity field (Fig. 5, lower panel) which provides the major contribution to the combined behavior of f_q shown in the upper panel.

The upper panel in Fig. 5 also shows a remarkable similarity in the time evolution of the combined model-data misfit for the 4dVar- and a4dVar-optimized NCOM states. The a4dVar algorithm has, however, a noticeable tendency to provide a better fit at the beginning of the assimilation window, clearly visible in the lower panels for f_S and f_v. This can be explained by the above mentioned property of a4dVar to better retrieve optimal states at shorter integration times.

When separated into different components, behavior of f_T^n, f_S^n, and f_v^n reveals more differences. In particular, the 4dVar method provides a much better fit to the temperature data after August 20 (in the second half of the assimilation window), but appears to be 10–13% worse than a4dVar with respect to salinity and velocity data.

A large contribution to a better salinity fit is given by the first two days of the a4dVar model trajectory (third panel in Fig. 5). However, certain gains relative to 4dVar are also observed at the end of the assimilation, which is quite opposite to the difference in the values of f_T. We attribute the better salinity fit to the variable EOF metric used during generation of the a4dVar search directions (Section 2.3).

Compared to f_T and f_S, the overall improvement of the model-data misfit is the smallest for velocity (bottom panel in Fig. 5), which was characterized by the observation errors of $\mathbf{R}^{1/2} \sim$ 7–10 cm/s in the cost function. Several assimilation runs were made with significantly smaller (3–5 cm/s) errors, but they were found to be inconsistent with *a posteriori* statistics of the model-data misfits as the optimal cost function values in these cases were much larger than those obtained in the reported experiments. The a4dVar-optimized value of f_v is persistently smaller during the entire assimilation period providing the 13% better value (as compared to 4dVar) in the 14-day average. This advantage could be partly attributed to the fact that the a4dVar search directions are derived from the most persistent patterns of the model-data misfits and therefore tend to be closer to the slowly evolving (geostrophically and hydrostatically balanced) modes of the flow. This property was also reflected in the better velocity forecast skill of the a4dVar solutions.

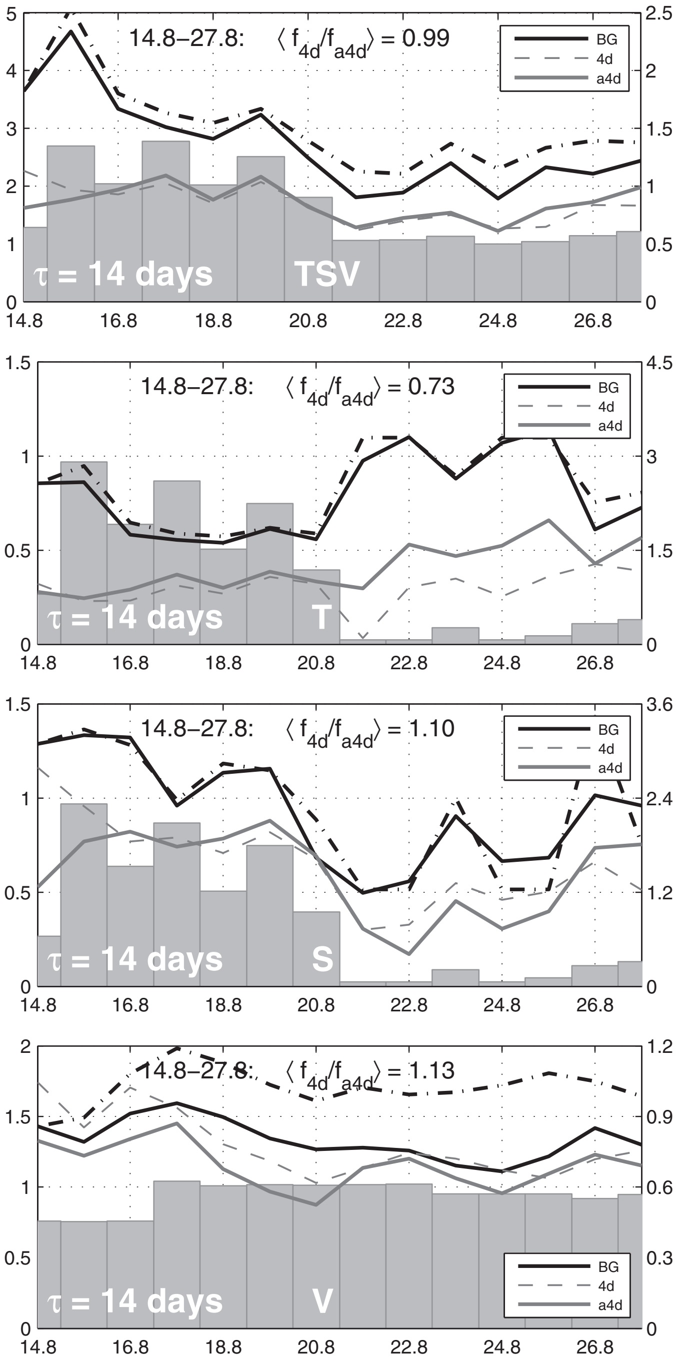

Fig. 5. Evolution of the root-mean-square model-data misfits f_q characterizing the background (BG, thick black lines), 4dVar-optimized (4d, thin dashed gray lines) and a4dVar-optimized (a4d) solutions. Thick dash-dotted line shows the misfit with the background fields at $t = 0$ (persistence). The values of f_q are shown on the right axis of each panel. The left axis quantifies the number of the data points for each day in thousands (shown by gray shaded rectangles). The ratio of the mean values of f_q averaged over the assimilation window for the 4dVar and a4dVar methods is given.

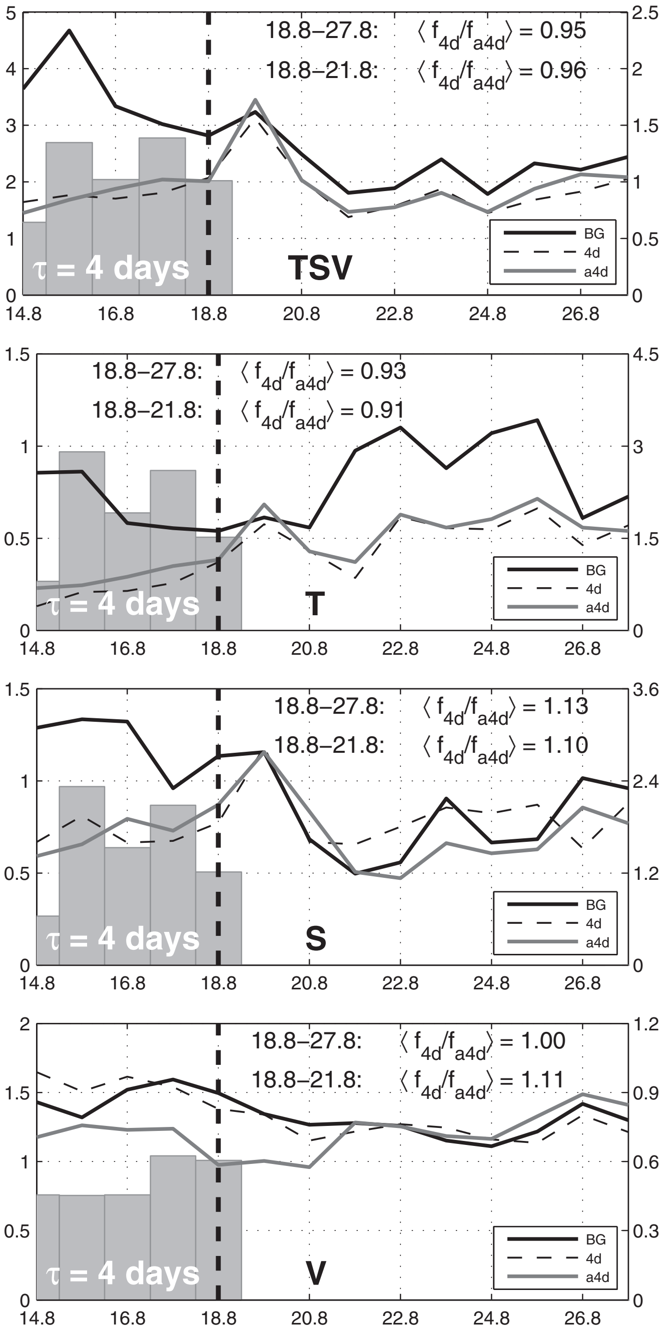

Fig. 6. Forecast skills f_s, f_v and f_t of the 4dVar and a4dVar-optimized solutions for the 4-day assimilation window. The relative number of the respective data points for each day is shown by gray shaded rectangles. Vertical dashed line show the time interval of data assimilation. Ratios of the mean f values averaged over the 3- and 9-days intervals are shown.

3.2. Forecast skill

The quality of the assimilated solutions was assessed for 4- and 8-day experiments using comparison with observations outside the respective assimilation windows. Evolution of the quantities f^n for the background and optimized solutions is shown in Fig. 6 for the 4-day assimilation experiment.

The general behavior of f is consistent with the one obtained in the 14-day experiment, showing persistently better 4dVar forecasts in temperature and the advantage of a4dVar in the salinity and velocity forecasts. The upper panel in Fig. 6 summarizes the forecast skill and indicates that 4dVar slightly (4–5%) outperforms a4dVar, mostly because of the better temperature forecasts (second panel from above). On the other hand, the 4dVar-optimized salinity is characterized by very low forecast skill, especially during August 21–25, when it was even farther away from the observations than the background forecast.

The 4dVar-optimized velocities show only small improvements compared to the background solution (bottom panel in Fig. 6). In contrast, the a4dVar-optimized velocities demonstrate 10–30% reduction of the model-data misfit within the assimilation window, which persists for up to three days (August 18–21) of the free model run. After August 21, the velocity mismatch of the background, a4dVar and 4dVar-optimized solutions are nearly identical. Qualitatively similar behavior of the forecast skill and its distribution among the state vector components was observed in the results of the 8-day assimilation experiment.

In interpreting the forecast skill assessment, it is necessary to take into account that the temperature and salinity observations after August 21 are much less numerous than those taken between the 18 and 21 of August (cf. gray rectangles of the second and third panels in Fig. 5), whereas the velocity data cover a rather limited area shown by the triangles in Fig. 1.

In general, the overall forecast skill provided by the a4dVar method appears to be comparable with that of the 4dVar (upper panel in Fig. 6), and in some aspects (such as short-term velocity forecast), the a4dVar technique provides noticeably better results. It should be noted that available observations could effectively constrain only a small part K/M =23,506/1,493,570 ∼ 1.5% of the model's degrees of freedom, so one should expect substantial differences in the small-scale structure of the optimal solutions obtained by two different methods.

3.3. Comparison of the optimal solutions

Temperature and velocity increments for the optimal states of the 14-day assimilation experiment are shown in Fig. 7. A certain coherence between the larger scale corrections to the background temperature field are clearly seen in the northern part of the model domain that is well covered by observations (cf. Fig. 1). The time-mean correlation coefficients ρ between the low-pass filtered temperature and salinity increments of the 4dVar and a4dVar solutions are 0.61 and 0.45, respectively if averaging is performed in the upper 200 m over the northern part of the domain. In the data-free region south of the 340 km mark, the correlations are substantially lower (respectively, 0.26 and 0.32) and lie below the 95% confidence level of nonzero correlation (0.36). Similar values of ρ (0.59 and 0.32 in the northern and southern subregions, respectively) were obtained for the sea surface height field.

Velocity increments appear to have the lowest correlations among the model fields with time-averaged values of ρ_v =0.36, 0.27 for the northern and southern subregions, respectively. The lowest correlations ($\rho_v = 0.09$, $\rho_T = 0.21$, and $\rho_S = 0.12$) were observed in the data-free southern subregion during the first 4 days (8/14–8/18) of the assimilation. Such incoherence between the increments is caused by excessive ageostrophic activity (left panel in Fig. 7) of the 4dVar solution at the beginning of the assimilation window. Ageostrophic nature of this disturbances is quantified by 4 times larger magnitude of the divergence field as compared to the rest of the domain. The ageostrophic mode disappears at the later times and does not affect the cost function because the southern subregion is virtually data-free, whereas smoothness constraints are imposed on the model fields only at the initial time.

The problem could be solved, apparently, by introducing balance constraints (e.g., Weaver et al., 2005) into the definition of the background error covariance at $n = 0$, which may not be necessary if the NCOM 4dVar were run in the weakly constrained mode, i.e., if background errors were prescribed throughout the entire assimilation window. For comparison purposes we ran the 4dVar system in the strongly constrained mode and the effect became visible after several outer loops. It is remarkable that the a4dVar algorithm appears to be much less susceptible to excitation of the ageostrophic modes (right panel in Fig. 7), possibly because the EOF-derived descent directions span subspaces characterized by slower time variation of the model trajectory and, therefore, tend to be closer to geostrophic and hydrostatic balance. The null-space nature of the ageostrophic features at the beginning of the 4dVar-optimized model trajectory can be also traced by looking at the evolution with iterations of the distances $\overline{\delta\mathbf{x}}$ between the background and (sub)optimal model trajectories (Fig. 8):

$$\overline{|\delta\mathbf{x}|} = \frac{1}{N+1}\sum_{n=0}^{N}\mathbf{x}^{nT}\mathbf{V}^{-1}\mathbf{x}^{n}. \tag{11}$$

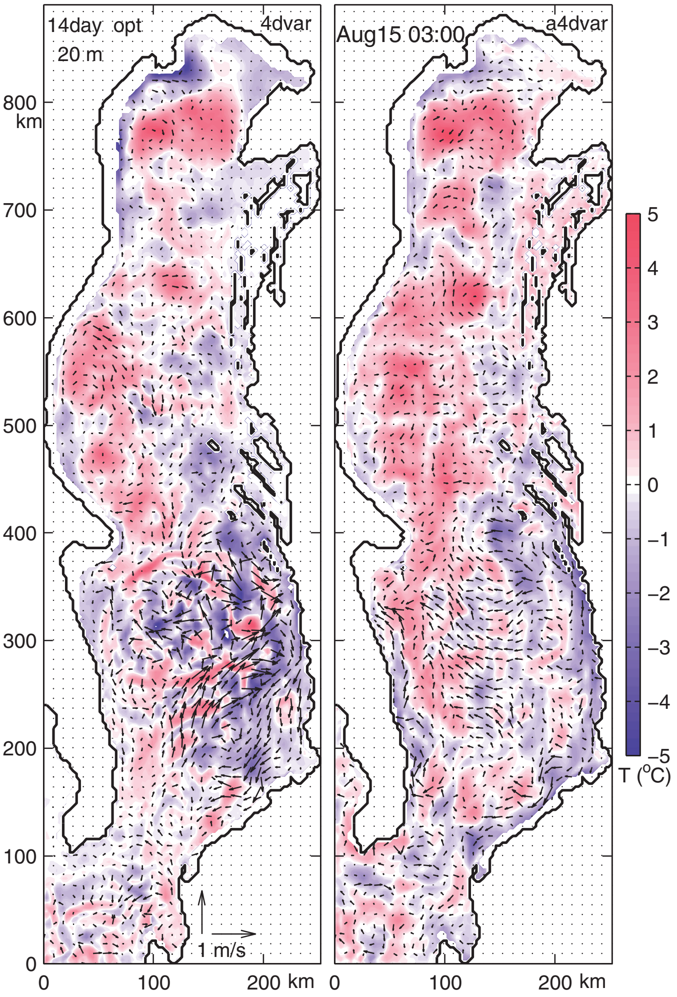

Fig. 7. Temperature and velocity differences between the background and optimized NCOM states at 20 m on August 15 03 UTC. Results of 4dVar and a4dVar optimizations are shown in the left and right panels, respectively. Velocity scale is shown at the bottom of the left panel.

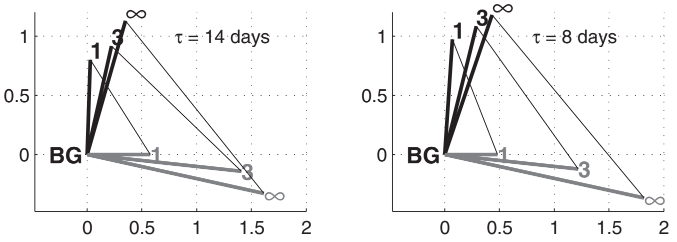

Fig. 8. Distances between the background (BG)-a4dVar (thick black lines), BG-4dVar (gray lines), and a4dVar-4dVar (thin black lines) solutions after the first 4dVar outer loop (1), third loop (3) and upon convergence (∞). Respective 4dVar outer loops are labeled by squares in the middle and lower panels of Fig. 4. The a4dVar distances are given on the iterations (circles in Fig. 4) corresponding to the equivalent CPU times of the respective 4dVar outer loops (squares in Fig. 4). All the distances are normalized by the magnitude of the background trajectory. Angles are assessed using the scalar product, associated with Eq. (12).

As can be seen, the largest (60–70%) a4dVar correction of the background trajectory occurs at the first few iterations, whereas 4dVar at the first iteration makes a significantly smaller correction and subsequently produces larger updates that are characterized by a relatively small reduction of the cost function (cf. Fig. 4). Such a behavior is typical for a null space search which appears to be inhibited in the a4dVar case. Introduction of the balance constraints into $\mathbf{B}$ will certainly improve the performance of both algorithms with a potentially larger benefit for the 4dVar case.

4. Discussion and conclusions

The major goal of the present study was to show the feasibility of the a4dVar technique (Yaremchuk et al., 2009) in a realistic application and compare its performance with the 4dVar method. We have found that the adjoint-free approach is capable of producing optimized solutions of similar quality to 4dVar with comparable computational expense. It was also found that the a4dVar technique is less susceptible to excitation of ageostrophic modes in the data-free regions if balance constraints are not imposed on the background error covariances.

A distinctive feature of the a4dVar technique presented here is the iterative approach to minimization of the cost function. The adjoint-free methods of Qui et al. (2007) and Liu et al. (2008) employ minimization in a predetermined subspace derived from the background error statistics (with or without localization of the background error covariance). The tested a4dVar method follows the same principle as the model-data misfits are projected on the range of $\mathbf{B}$ (Eq. (7)), nonetheless convergence to $\mathbf{x}_{opt}$ cannot generally be guaranteed since the process may be subject to breakdown (situations when new search directions are linearly dependent on previous ones). In future applications, the issue could possibly be resolved using the methods already developed for the GMRES-type algorithms (e.g., Reichel and Ye, 2005), which are not immune to breakdowns, similar to the a4dVar minimization technique. Experiments reported here do show, however, that the observed rate of convergence of the a4dVar minimization process may slow down relative to what is seen with 4dVar and this effect may be exacerbated with increasing size of assimilation window (Fig. 4).Nonetheless, a4dVar tends to produce better results at the earlier stages of assimilation than 4dVar and in general its performance could still be viewed as satisfactory.

Taking advantage of the trend toward massive parallelization in computer technologies, the adjoint-free variational methods estimate the cost function gradient with a “brute force” approach that employs finite difference approximations along predetermined directions in state space. Selection of the most effective directions (search subspaces $\mathbb{S}_i$) based on this sampling becomes an issue of primary importance. Experience shows that there exists a considerable freedom in generating search directions (Yaremchuk et al., 2009; Panteleev et al., 2015) as long as they are kept being spatially smooth and $\tilde{\mathbf{H}}$-orthogonal. In particular, building $\mathbb{S}_i$ on the eigenvectors of $\mathbf{B}$ in the descending order of their eigenvalues may also work reasonably well, as shown in Appendix A. In the reported experiments, we investigated a number of methodologies in building $\mathbb{S}_i$ and found Eq. (7) to be the most effective computationally. This methodology can be developed further by introducing balance constraints into $\mathbf{B}^{-1}$. In this case the inverse background error covariance should be replaced by the composite matrix

$$\mathbf{B}_{bal}^{-1} = \begin{bmatrix} \mathbf{B}_1^{-1} + \mathbf{L}^{\mathsf{T}}\mathbf{B}_2^{-1}\mathbf{L} & -\mathbf{L}^{\mathsf{T}}\mathbf{B}_2^{-1} \\ -\mathbf{B}_2^{-1}\mathbf{L} & \mathbf{B}_2^{-1} \end{bmatrix}, \tag{12}$$

where $\mathbf{B}_1^{-1}$ and $\mathbf{B}_2^{-1}$ are the inverse covariances of the unbalanced components of the state vector (e.g., defined by (9)) and $\mathbf{L}$ is the balance operator. Further improvements can be made by replacing the diffusion operator in Eq. (9) with a more general expression (e.g., Weaver and Mirouze, 2012; Yaremchuk and Nechaev, 2013).

An important issue with the a4dVar technique is its extension to optimization of other sets of variables that may control the model trajectory, such as surface forcing fields. One of the possible solutions in this case augments the search subspaces (ocean model states) by the leading EOFs of the surface forcing error fields. This will require a better knowledge of error statistics of the atmospheric model used to force the ocean in a particular application. In view of recent rapid development of the observational systems and data acquizition techniques in the atmosphere, the issue of accessibility to the above mentioned statistics seems likely to be resolvable in the near term. Moreover, the a4dVar technique appears to be even more suitable for coupled ocean-atmosphere systems, where external forcing errors tend to play a lesser role at the time scale of a typical assimilation window.

In terms of the computational expense, the a4dVar technique appears roughly comparable to 4dVar, mostly because of the excessive computational cost of tangent linear and adjoint codes that were, on average, several times more expensive than a direct run of the nonlinear NCOM model (a typical situation with state-of-the-art OGCMs, e.g., Oldenborgh et al., 1999 and Heimbach et al., 2005). On massively parallel machines, the advantage of a4dVar will be more noticeable due to the limited parallel scalability of an OGCM code, be it original, adjoint, or tangent linear.

A much larger computational advantage is evident when considering the wall time in a massively parallel environment, which formally allows a4dVar to search over multiple directions at a fraction of the wall time used by 4dVar to generate a steepest descent direction. In fact, in the reported experiments with NCOM model, one a4dVar run was executed almost five times faster if all the ensemble members were run on separate nodes. This property of the a4dVar approach gives good prospects for its further development in sync with and other types of ensemble data assimilation techniques that are based on relaxed communication requirements between processors.

Acknowledgments

The authors were supported by the Office of Naval Research under program element 61153N “Adjoint-free 4dVar for Ocean Models”. The work of C. Beattie was additionally supported in part by the Einstein Foundation of Berlin and the NSF through Grant DMS-1217156. We gratefully acknowledge P. Sakaulakis for help with parallelization issues and M. Carrier and S. Smith for their kind assistance in configuring the NCOM 4dVar data assimilation system. Helpful discussions with I. Shulman are also acknowledged.

Appendix A. 4dVar/a4dVar comparison in the linear case

To compare the performance of 4dVar and a4dVar techniques in a simple environment, consider the problem of retrieving the initial field of tracer concentration $\eta(\boldsymbol{x}, 0)$ from observations at some distant time T. The tracer evolution is governed by

$$\partial_t \eta + \boldsymbol{u}\nabla\eta - \mu\Delta\eta = f(\boldsymbol{x}, t) \tag{A.1}$$

in a closed rectangular 49 × 91 domain Ω with the boundary condition $\eta(\partial\Omega, t) = 0$. Eq. (A.1) is discretized on a regular grid using first-order RK time-stepping, upwind advection, and a standard 5-point stencil for the Laplacian with unit steps in temporal and spatial directions. Velocity $\boldsymbol{u} = (u, v)$ at any space-time location was defined by $u = -0.2 + 0.01\nu$; $v = -0.1 + 0.01\nu$, where ν is the white noise on unit interval. The forcing f was generated by setting $f(\boldsymbol{x}, t) = .001\nu$ in every point of the space-time grid. The coefficient μ was set to 10^{-5}, so that diffusion was largely determined by the numerics.

The simulated data comparison experiment was set as follows. Given the initial tracer distribution $\hat{\eta} = \eta(\boldsymbol{x}, 0) = \exp[-(\boldsymbol{x} - \boldsymbol{x}_0)^2/9]$ with $\boldsymbol{x}_0 = (70, 35)$ (bell-shaped disturbance in Fig. A1a), the model

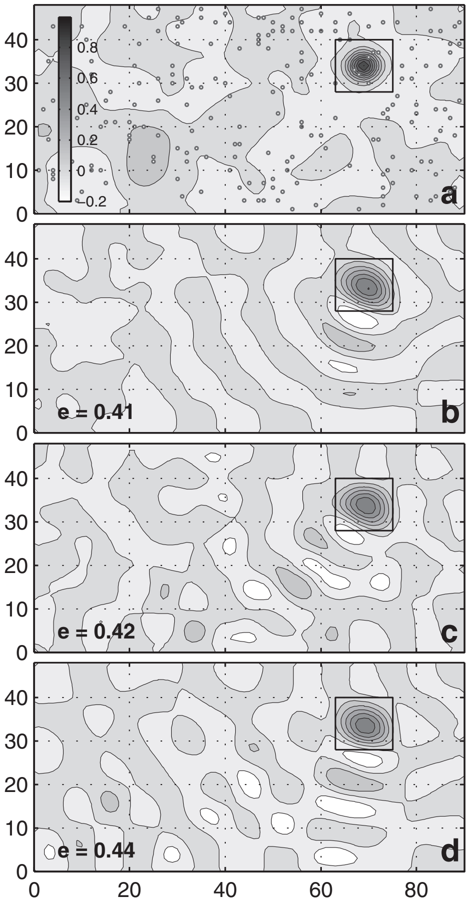

Fig. A1. Reconstruction of the initial condition of the tracer field by 4dVar (b) and a4dVar (c,d) techniques. Composite map of the reconstructed tracer field evolution is shown in the upper panel. (a): Initial position of the reconstructed feature (Gaussian eddy at $x = 70$, $y = 35$ km) is superimposed on the tracer field (contours) at the observation time $t = 200$ when the eddy diplaced to $x = 25$, $y = 15$ km. Circles denote observation points. The errors in approximation of the true field at $t = 0$ are shown in the left corners of the panels b–d, showing the initial field, reconstructed by various methods.

was integrated for $T = 200$ time steps to obtain the final distribution $\eta(\boldsymbol{x}, T)$ shown by contours in the same panel. The initial disturbance almost completely dispersed and migrated to $\boldsymbol{x}_T \sim (25, 15)$ (see contours in the same panel). After that, $\eta(\boldsymbol{x}, T)$ was sampled at 200 points shown in Fig. A1a, and obtained numbers were used to reconstruct $\hat{\eta}$ by minimizing the cost function (3) under the dynamical constraint (A.1) with the background error covariance defined by (9).

Optimal approximations $\tilde{\eta}$–$\hat{\eta}$ obtained by 4dVar and a4dVar techniques are shown in Fig. A1b and Fig. A1c, respectively). To specify search directions in the a4dVar method, 200 observations were split into $m_s = 10$ equal groups so an observation operator for each search direction in (7) had 20 observation locations.

The quality of reconstruction was assessed by the parameter

$$e = \sqrt{\langle(\tilde{\eta} - \hat{\eta})^2\rangle / \langle\hat{\eta}^2\rangle} \tag{A.2}$$

where angular brackets denotes averaging over small rectangles in Fig. A1. Comparison of Fig. A1b and c suggests that the a4dVar method is capable of providing a solution of the same quality with 4dVar.

Fig. A1d illustrates another a4dVar solution, using search subspaces specified by the eigenvectors of $\mathbf{B}$ in the decreasing order of their eigenvalues (in this case, sines with decreasing wavelengths in both directions). The result obtained is of similar quality, suggesting that the general a4dVar strategy of minimizing J over a sequence of smooth $\tilde{\mathbf{H}}$-orthogonal SDs may work well with various methods of generating search directions.

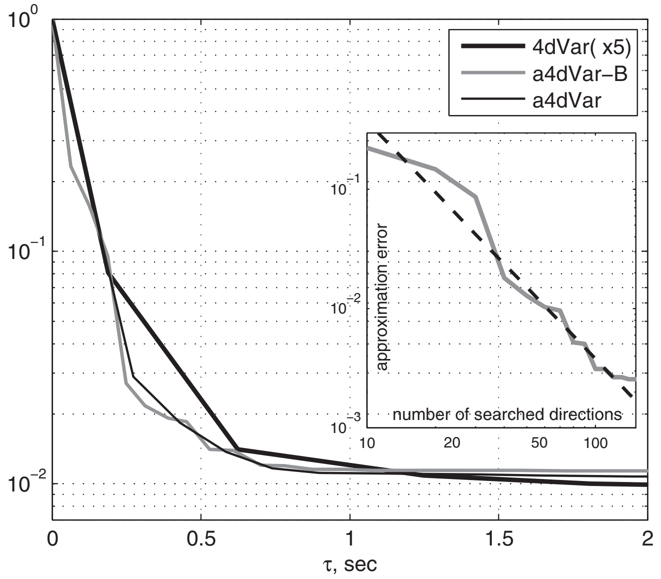

Fig. A2. Reduction of the cost function against CPU time for 4dVar and a4dVar techniques. The 4dVar CPU time is multiplied by five to mimic larger CPU requirements of the state-of-the-art adjoint models. Inset: Convergence of the a4dVar-B solution (Fig. A1d) to the exact solution. Dashed line shows the convergence rate given by (A.5).

In terms of computational expense, the 4dVar method provided approximately five times faster reduction of the cost function (Fig. A2) due to high efficiency of the adjoint model. In this simple case, an adjoint model run required the same amount of time as the direct model run. In real applications, the tangent linear and adjoint codes are several times more expensive to run and the a4dVar techniques may prove to be more competitive, as shown in Section 3 of the present study.

Finally, the known spectrum of $\mathbf{B}$ provides an opportunity to assess the convergence rate of the ad4Var solution exposed in Fig. A1d. Assume that after k a4dVar iterations $m = km_s$ $\tilde{\mathbf{H}}$-orthogonal directions have been already searched and the kth approximation $\hat{\eta}_k$ to the optimal solution $\hat{\eta} = \tilde{\mathbf{H}}^{-1}\mathbf{b}$ have been found. Without loss of generality, the eigenvectors ϕ_i of $\mathbf{B}$ could be normalized to satisfy $\phi_i \mathbf{B}^{-1}\phi_i = 1$, so that their (Euclidean) norm is equal to the associated eigenvalue σ_i. The magnitude e_m of the approximation error $\mathbf{e}_m = \hat{\eta} - \hat{\eta}_m$ with respect to the norm induced by the inverse covariance can be assessed by projecting $\hat{\eta}$ on the *unexplored* directions:

$$e_m = \mathbf{e}_m^{\mathrm{T}}\mathbf{B}^{-1}\mathbf{e}_m \le \sum_{k=m+1}^{\infty} |\hat{\eta}^{\mathrm{T}}\mathbf{B}^{-1}\phi_k|^2 \tag{A.3}$$

Furthermore, since the optimal solution $\hat{\eta} = \tilde{\mathbf{H}}^{-1}\mathbf{b}$ allows representation in the (dual) form $\hat{\eta} = \mathbf{B}\xi$ (ξ is the optimal linear combination of the representers), the upper bound of the terms under summation in (A.3) can be assessed by

$$|\hat{\eta}^{\mathrm{T}}\mathbf{B}^{-1}\phi_k| = |\xi^{\mathrm{T}}\phi_k| \le \sigma_k(\xi^{\mathrm{T}}\mathbf{B}^{-1}\xi)^{1/2} \tag{A.4}$$

Substituting (A.4) into (A.3) yields the following upper bound on the error magnitude:

$$e_m \le \xi^{\mathrm{T}}\mathbf{B}^{-1}\xi \sum_{k>m} \sigma_k \sim O(m^{-2}) \tag{A.5}$$

This estimate remains intact if we assess e_m with respect to the norm induced by the Hessian matrix. In the latter case, the right-hand side of (A.5) will be multiplied by a scaling factor $||\,\tilde{\mathbf{H}}\,||/||\mathbf{B}^{-1}|| > 1$.

Dependence of the distance between the 4dVar solution (Fig. A1b) and the consecutive approximations to the a4dVar solution (Fig. A1d) shown in the inset to Fig. A2, confirms the above estimate.

Appendix B. $\tilde{\mathbf{H}}$-orthogonalization and related issues

The a4dVar method utilizes the technique employed by Zupanski (2005) in the Maximum Likelihood Ensemble Filter, which is based on the explicit inversion of the Hessian matrix in the subspace spanned by the model perturbations. In view of the definition (9), $\mathbf{B}^{-1/2}$ can be explicitly represented using the expression for the square root of the inverse error covariance:

$$\mathbf{B}^{-1/2} = \mathbf{V}^{-1}\left(\mathbf{I} - \frac{b^2}{2}\Delta\right) \tag{B.1}$$

which allows a symmetric Hessian factorization

$$\tilde{\mathbf{H}} = \tilde{\mathbf{H}}^{\mathsf{T}/2}\tilde{\mathbf{H}}^{1/2}, \tag{B.2}$$

where

$$\tilde{\mathbf{H}}^{1/2} = \begin{bmatrix} \mathbf{B}^{-1/2} \\ \mathbf{H}_0 \\ \mathbf{H}_1\mathbf{M}^1 \\ \vdots \\ \mathbf{H}_N\mathbf{M}^N \end{bmatrix} \tag{B.3}$$

is the Hessian square root.

For sufficiently small perturbations $\delta\mathbf{c}_m = \varepsilon\mathbf{p}_m$, perturbations of the auxiliary vector

$$\delta\mathbf{Y}_m = \tilde{\mathbf{H}}^{1/2}\delta\mathbf{c}_m \tag{B.4}$$

are linear in $\delta\mathbf{c}_m$, so that computation of the dot products between the vectors $\delta\mathbf{Y}_m$ provides the inner product in the control space associated with the Hessian matrix

$$\delta\mathbf{Y}_1^{\mathsf{T}}\delta\mathbf{Y}_2 = \delta\mathbf{c}_1^{\mathsf{T}}\tilde{\mathbf{H}}\delta\mathbf{c}_2 = \langle\delta\mathbf{c}_1, \delta\mathbf{c}_2\rangle_{\tilde{\mathbf{H}}}, \tag{B.5}$$

which can be used for $\tilde{\mathbf{H}}$-orthogonalization of the search subspaces of the a4dVar algorithm.

We seek the optimal correction of the control variable $\mathbf{c}$ in the search subspace $\mathbb{S}$ spanned by $\mathbf{p}_m$:

$$\mathbf{c} \leftarrow \mathbf{c} + \sum_{l=1}^{m_s} s_l\mathbf{p}_l,$$

where the coefficients s_l satisfy for $m = 1, 2, \ldots, m_s$,

$$\mathbf{p}_m^{\mathsf{T}}\left(\tilde{\mathbf{H}}\left(\mathbf{c} + \sum_{l=1}^{m_s} s_l\mathbf{p}_l\right) - \mathbf{b}\right) = 0. \tag{B.6}$$

This constitutes a Ritz–Galerkin projection of the normal system (4) to the search subspace, $\mathbb{S}$. Rearranging, we obtain the linear system of m_s equations in the m_s unknowns $s_1, s_2, \ldots, s_{m_s}$:

$$\sum_{l=1}^{m_s}\mathbf{p}_m^{\mathsf{T}}\tilde{\mathbf{H}}\,\mathbf{p}_l s_l = \mathbf{p}_m^{\mathsf{T}}(\mathbf{b} - \tilde{\mathbf{H}}\mathbf{c}). \tag{B.7}$$

Substituting $\mathbf{p}_m = \delta\mathbf{c}_m/\varepsilon$ into (B.7), multiplying by ε^2, and using (B.2) and (B.4) yields

$$\sum_{l=1}^{m_s}\delta\mathbf{Y}_m^{\mathsf{T}}\delta\mathbf{Y}_l s_l = \varepsilon\delta\mathbf{c}_m^{\mathsf{T}}(\mathbf{b} - \tilde{\mathbf{H}}\mathbf{c}). \tag{B.8}$$

The right-hand side of (B.8) cannot be computed directly because evaluation of $\mathbf{b} - \tilde{\mathbf{H}}\mathbf{c}$ requires the adjoint code (Eq. (5)). Nonetheless, for each m, $\delta\mathbf{c}_m^{\mathsf{T}}(\mathbf{b} - \tilde{\mathbf{H}}\mathbf{c})$ can be calculated directly from the variations of the cost function $\delta J_m = J(\mathbf{c} + \delta\mathbf{c}_m) - J(\mathbf{c})$ induced by $\delta\mathbf{c}_m$:

$$\begin{aligned}\delta J_m &= \frac{1}{2}\delta\mathbf{c}_m^{\mathsf{T}}\tilde{\mathbf{H}}\delta\mathbf{c}_m + \delta\mathbf{c}_m^{\mathsf{T}}\left(\tilde{\mathbf{H}}\mathbf{c} - \mathbf{b}\right)\\ &= \frac{1}{2}\delta\mathbf{Y}_m^{\mathsf{T}}\delta\mathbf{Y}_m - \delta\mathbf{c}_m^{\mathsf{T}}\left(\mathbf{b} - \tilde{\mathbf{H}}\mathbf{c}\right).\end{aligned} \tag{B.9}$$

Thus, the coefficients for the optimal correction of the control variable $\mathbf{c}$ within the search subspace $\mathbb{S}$ are given as the solution to a linear system posed in terms of the quantities δJ_m and $\delta\mathbf{Y}_m$ computed by the a4dVar algorithm:

$$\sum_{l=1}^{m_s}\delta\mathbf{Y}_m^{\mathsf{T}}\delta\mathbf{Y}_l s_l = \varepsilon\left(\frac{1}{2}\delta\mathbf{Y}_m^{\mathsf{T}}\delta\mathbf{Y}_m - \delta J_m\right). \tag{B.10}$$

In the $\tilde{\mathbf{H}}$-orthonormal coordinate system $\delta\mathbf{Y}_m^{\mathsf{T}}\delta\mathbf{Y}_m = \varepsilon^2$, and equations (B.10) are simplified to

$$s_l = \sum_m \alpha_{lm}\left(\frac{\varepsilon}{2} - \frac{\delta J_m}{\varepsilon}\right), \tag{B.11}$$

where α_{lm} are the matrix elements of the linear transformation of the original basis $\delta\mathbf{c}_m$ that are obtained in the orthogonalization process.

For a differentiable numerical model and sufficiently small ε, the quadratic term in the right hand side of (B.10) is negligible. In the reported experiments we kept it in place since the value of ε was close to 0.01 and could not be reduced any further without affecting the rate of convergence. The relatively large limit on the value of ε was caused by a number of factors deteriorating the linear dependence between the magnitude of the model perturbations and ε. These factors include rounding errors (especially for temperature and salinity in the upper layers), non-differentiable operators in the model code, particularly at the open boundary, and small-scale instabilities of the flow developing at the mouth of the Po river and in the boundary currents, especially prominent in the experiments with the 14-day assimilation window.

The finite value of ε also affected the orthogonalization process, resulting in non-zero (of the order of 3–10%) off-diagonal elements observed in the Hessian projection after orthogonalization of the perturbations. For that reason, the optimal coefficients s_l were computed through the direct solution of Eq. (B.10), which was not expensive due to the low dimension of the system.

References

Ancell, B., Hakim, G.J., 2007. Comparing adjoint- and ensemble-based sensitivity analysis with applications to observation targeting. Mon. Weather Rev. 135, 4117–4134.

Anderson, J.L., Hoar, T., Raeder, K., Liu, H., Collins, N., Torn, R., Arellano, A., 2009. The data assimilation research testbed: a community facility . Bull. Am. Meteorol. Soc. 90, 1283–1296.

Barron, C.N., Kara, A.B., Hurlburt, H.E., Rowley, C., Smedstad, L.F., 2004. Sea surface height predictions from the global Navy Coastal Ocean Model (NCOM) during 1998–2001. J. Atmos. Oceanic Technol. 21 (12), 1876–1894.

Barron, C.N., Kara, A.B., Martin, P.J., Rhodes, R.C., Smedstad, L.F., 2006. Formulation, implementation and examination of vertical coordinate choices in the global Navy Coastal Ocean Model (NCOM). Ocean Modell. 11, 347–375. doi:10.1016/j.ocemod.2005.01.004.

Bennett, A.F., 2002. Inverse Modeling of the Ocean and Atmosphere. Cambridge University Press, p. 234. ISBN 0-521-81373-5.

Buehner, M., Houtekamer, P.L., Charette, C., Mitchell, H.L., He, B., 2010. Intercomparison of variational data assimilation and the ensemble Kalman filter for a global deterministic NWP. Mon. Wea. Rev. 138, 1550–1566.

Burrage, D.M., Book, J.W., Martin, P.J., 2009. Eddies and filaments of the Western Adriatic cCurrent: analysis and prediction. J. Mar. Syst. 78, S205–S226.

Clayton, A.M., Lorenc, A.C., Barker, D.M., 2013. Operational implementation of a hybrid ensemble/4D-Var global data assimilation system at the met office. Q. J. R. Meteorol. Soc. doi:10.1002/qj.2054.

Cushman-Roisin, B., Korotenko, K.A., 2007. Mesoscale-resolving simulations of summer and winter bora events in the Adriatic Sea. J. Geophys. Res. 112, C11S91. doi:10.1029/2006JC003516.

Desroziers, G., Camino, J.-T., Berre, L., 2014. 4DEnVar: link with 4D state formulation of variational assimilation and different possible implementations. Q. J. R. Meterol. Soc. 140, 2097–2110.

Fairbairn, D., Pring, S.R., Lorenc, A.C., Roulstone, I., 2014. A comparison of 4dVar with ensemble data assimilation methods. Q. J. R. Meterol. Soc. 140, 281–294.

Heimbach, P., Hill, C., Giering, R., 2005. An efficient adjoint of the parallel MIT GCM generated via automatic differentiation. Future Gener. Comput. Syst. 21, 1356–1371.

Holland, W.R., Chow, J.C., Bryan, F.O., 1998. Application of a third-order-upwind scheme in the NCAR Ocean Model. J. Clim. 11, 1487–1493.

Hoteit, I., 2008. A reduced-order simulated annealing approach for four-dimensional variational data assimilation in meteorology and oceanography. Int. J. Numer. Methods Fluids 58, 1181–1199.

Hoteit, I., Hoar, T., Gopalakrishnan, G., Anderson, J., Collins, N., Cornuelle, B., Kohl, A., Heimbach, P., 2013. A MITgcm/DART ocean prediction and analysis system with application to the Gulf of Mexico. Dyn. Atmos. Oceans 63, 1–23.

Hoteit, I., Luo, X., Pham, D.T., 2012. Particle Kalman filtering: a nonlinear Bayesian framework for ensemble Kalman filters. Mon. Weather Rev. 140, 528–542.

Ivatek-Sahdan, S., Tudor, M., 2004. Use of high-resolution dynamical adaptation in operational suite and research studies. Meteorol. Z. 13, 1–10.

Kuhl, D.D., Rosmond, T.E., Bishop, C.H., McLay, J., Baker, N., 2013. Comparison of hybrid ensemble/4DVar and 4dVar within the NAVDAS-AR data assimilation framework. Mon. Weather Rev. 141, 2740–2758.

Le Dimet, F.-X., Tlagrand, O., 1986. Variational algorithms for assimilation and analysis of meteorological observtaions: theoretical aspects. Tellus A 38A (2), 97–110.

Liu, C., Xiao, Q., Wang, B., 2008. An ensemble-based four-dimensional variational data assimilation scheme. Part I: technical formulation and preliminary test. Mon. Weather Rev. 136, 3363–3373.

Liu, C., Xiao, Q., Wang, B., 2009. An ensemble-based four-dimensional variational data assimilation scheme. Part II: observing system simulation experiments with advanced research WRF (ARW). Mon.Weather Rev. 137, 1687–1704.

Martin, P.J., 2000. A Description of the Navy Coastal Ocean Model Version 1.0. NRL Rep. NRL/FR/7322 00-9962, 42 pp. Naval Research Laboratory, Stennis Space Center, MS.

Martin, P.J., Book, J.W., Burrage, D.M., Rowley, C.D., Tudor, M., 2009. Comparison of model-simulated and observed currents in the central adriatic during DART. J. Geophys. Res. 114, C01S05. doi:10.1029/2008JC004842.

Mellor, G.L., 1991. An equation of state for numerical models of oceans and estuaries. J. Atmos. Oceanic Technol. 8, 609–611.

Mellor, G.L., Yamada, T., 1974. A hierarchy of turbulence closure models for planetary boundary layers. J. Atmos. Sci. 31, 1791–1806.

Menemenlis, D., Wunsch, C., 1997. Linearization of an oceanic general circulation model for data assimilation and climate studies. J. Atmos. Oceanic Technol. 14, 1420–1443.

Mirouze, I., Weaver, A., 2010. Representation of correlation functions in variational data assimilation using an implicit diffusion operator. Q. J. R. Meteorol. Soc. 136, 1421–1443.

Moore, A.M., Arango, H.G., Broquet, G., Powell, B.S., Weaver, A.T., Zavala-Garay, J., 2011. The Regional Ocean Modeling System (ROMS) 4-dimensional variational data assimilation systems: Part I System overview and formulation. Prog. Oceanogr. 91 (1), 34–49.

Morey, S.L., Martin, P.J., OBrien, J.J., Wallcraft, A.A., Zavala-Hidalgo, J., 2003. Export pathways for river discharged fresh water in the northern Gulf of Mexico. J. Geophys. Res. 108 (C10), 3303. doi:10.1029/2002JC001674.

Ngodock, H., Carrier, M., 2014. A 4dVar system for the navy coastal ocean model. Part I: system description and assimilation of synthetic observations in Monterey Bay. Mon. Weather Rev. 142 (6), 2085–2107.

Oldenborgh, G.J., Burgers, G., Venzke, S., Eckart, C., Giering, R., 1999. Tracking down the ENSO delayed oscillator with an adjoint OGCM. Mon. Weather Rev. 127, 1477–1495.

Panteleev, G., Yaremchuk, M., Rogers, E., 2015. Adjoint-free variational data assimilation into a regional wave model. J. Atmos. Oceanic Technol. 32, 1386–1399.

Qui, C., Shao, A., Wei, L., 2007. Fitting model fields to observations by using singular value decomposition: an ensemble-based 4dVar approach. J. Geophys. Res. 112, D11105. doi:10.1029/2006JD007994.

Reichel, L., Ye, Q., 2005. Breakdown-free GMRES for singular systems. SIAM J. Matrix Anal. Appl. 26 (4), 1001–1021.

Robert, C., Durbiano, S., Blayo, E., Verron, J., Blum, J., Dimet, F.-X. L., 2005. A reduced-order strategy for 4dVar data assimilation. J. Mar. Sys. 57 (1–2), 70–82.

Rosmond, T., Xu, L., 2006. Development of the NAVDAS-AR: non-linear formulation and outer loop tests. Tellus 58A, 45–58.

Stammer, D., Wunsch, C., 1996. The determination of the large-scale circulation of the Pacific Ocean from satellite altimetry using model Green's functions. J. Geophys. Res. 101, 18409–18432.

Tian, X., Xie, Z., 2012. Implementations of a square-root ensemble analysis and a hybrid localization into the POD-based ensemble 4DVar. Tellus A 64, 1–10. doi:10.3402/tellusa.v64i0.18375.

Torn, R.D., Hakim, G.J., 2008. Ensemble-based sensitivity analysis. Mon. Weather Rev. 136, 663–677.

Trevisan, A., DIsidoro, M., Talagrand, O., 2010. Four-dimensional variational assimilation in the unstable subspace and the optimal subspace dimension. Q. J. R. Meteorol. Soc. 136, 487–496.

Weaver, A.T., Courtier, P., 2001. Correlation modelling on a sphere using a generalized diffusion equation. Q. J. R. Meteorol. Soc. 127, 1815–1846.

Weaver, A.T., Deltel, C., Machu, E., Ricci, S., Daget, N., 2005. A multi-variate balance operator for variational data assimilation. Q. J. R. Meteorol. Soc. 131, 3605–3625.

Weaver, A.T., Mirouze, I., 2012. On the diffusion equation and its application to isotropic and anisotropic correlation modeling in variational assimilation. Q. J. R. Meteorol. Soc. 138. doi:10.1002/qj.1953.

Xu, L., Rosmond, T., 2004. Formulation of the NRL Atmospheric Variational Data Assimilation System – Accelerated Representer (NAVDAS-AR). Naval research Laboratory, p. 28. NRL/MR/7532-36.

Xu, L., T. Rosmond, T., Daley, R., 2005. Development of the NAVDAS-AR: formulation and initial tests of the linear problem. Tellus 57A, 546–559.

Yaremchuk, M., Martin, P., 2014. On sensitivity analysis in the 4dVar framework. Mon. Weather Rev. 142, 774–787.

Yaremchuk, M., Nechaev, D., 2013. Covariance localization with the diffusion-based correlation models. Mon. Weather Rev. 141, 848–860.

Yaremchuk, M., Nechaev, D., Panteleev, G., 2009. A method of successive corrections of the control subspace in the reduced-order variational data assimilation. Mon. Weather Rev. 137, 2966–2978.

Yaremchuk, M., Sentchev, A., 2012. Multi-scale correlation functions associated with polynomials of the diffusion operator. Q. J. R. Meterol. Soc. 138, 1948–1953.

Yaremchuk, M., Smith, S., 2011. On the correlation functions associated with polynomials of the diffusion operator. Q. J. R. Meterol. Soc. 137, 1927–1932.

Zhang, F., Zhang, M., Hansen, J.A., 2009. Coupling ensemble Kalman filter with four-dimensional variational data assimilation. Adv. Atmos. Sci. 26, 19.

Zhang, M., Zhang, F., 2012. E4DVar: Coupling an ensemble Kalman filter with four-dimensional variational data assimilation in a limited-area weather prediction model. Mon. Weather Rev. 140, 587600.

Zupanski, M., 2005. Maximum likelihood ensemble filter: theoretical aspects. Mon. Weather Rev. 133, 1710–1726.