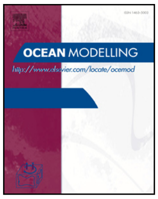
Ocean Modelling 117 (2017) 41–51

Contents lists available at ScienceDirect

Ocean Modelling

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



A hybrid approach to generating search subspaces in dynamically constrained 4-dimensional data assimilation

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

[a] Naval Research Laboratory at Stennis Space Center, USA
[b] Department of Mathematics, Virginia Tech, USA

ARTICLE INFO

Article history:
Received 16 November 2016
Revised 26 July 2017
Accepted 2 August 2017
Available online 3 August 2017

Keywords:
Data assimilation
Adjoint analysis
Regional modeling

ABSTRACT

Development and maintenance of the linearized and adjoint code for advanced circulation models is a challenging issue, requiring a significant proportion of total effort in operational data assimilation (DA). The ensemble-based DA techniques provide a derivative-free alternative, which appears to be competitive with variational methods in many practical applications. This article proposes a hybrid scheme for generating the search subspaces in the adjoint-free 4-dimensional DA method (a4dVar) that does not use a predefined ensemble. The method resembles 4dVar in that the optimal solution is strongly constrained by model dynamics and search directions are supplied iteratively using information from the current and previous model trajectories generated in the process of optimization. In contrast to 4dVar, which produces a single search direction from exact gradient information, a4dVar employs an ensemble of directions to form a subspace in order to proceed. In the earlier versions of a4dVar, search subspaces were built using the leading EOFs of either the model trajectory or the projections of the model-data misfits onto the range of the background error covariance (BEC) matrix at the current iteration. In the present study, we blend both approaches and explore a hybrid scheme of ensemble generation in order to improve the performance and flexibility of the algorithm. In addition, we introduce balance constraints into the BEC structure and periodically augment the search ensemble with BEC eigenvectors to avoid repeating minimization over already explored subspaces. Performance of the proposed hybrid a4dVar (ha4dVar) method is compared with that of standard 4dVar in a realistic regional configuration assimilating real data into the Navy Coastal Ocean Model (NCOM). It is shown that the ha4dVar converges faster than a4dVar and can be potentially competitive with 4dvar both in terms of the required computational time and the forecast skill.

1. Introduction

The ongoing trend toward massive parallelization in computer technologies facilitates the use of ensemble techniques in geophysical data assimilation. The ensemble approach becomes attractive not only because of its favorable parallelization properties (Isaksen, 2011; Desroziers and Berre, 2012). It also brings in more flexibility and realism in representing the background error covariances (e.g., Romine et al., 2014; Ménétrier et al., 2014; Descombes et al., 2015) and appears to be less vulnerable to instabilities associated with model linearization employed by the standard 4dVar technique. Besides, the ensemble approach allows to avoid costly development and maintenance of the linearized models and their adjoints which beyond being costly may impose certain limits on versatility in applying dynamical constraints within a particular adjoint-based assimilation system.

In the last decade, the use of ensembles in DA has been under extensive development in several directions. Apart from improvements in the BEC modeling, major efforts have been made to combine the benefits of the 4dVar and the ensemble methods. In particular, Buehner et al. (2010) have shown that the 4dVar system with the ensemble-generated BEC outperforms the standard 4dVar in the global forecast model. Similar results were obtained by Kuhl et al. (2013) who investigated the performance of the atmospheric DA system (Rosmond and Xu, 2006) with the hybrid BEC formulation. Coupling the regional 4dVar and ensemble KF systems (Zhang and Zhang, 2012; Barker et al., 2012) resulted in a significant reduction of errors for the forecast lead times up to 2.5 days. All these observations underscore the decisive role played by the flow-dependent BECs delivered by ensembles in improving the forecast skill.

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

http://dx.doi.org/10.1016/j.ocemod.2017.08.003
1463-5003/Published by Elsevier Ltd.

Another extensive field of development is related to the so-called 4dEnVar algorithms (Liu et al., 2008; 2009; Fairbairn et al., 2014) which introduce ensembles into the very fabric of the 4dVar optimization. In contrast to 4dVar which implicitly propagates the BEC, these ensemble methods leverage the power of massively parallel computers and explicitly approximate BEC evolution on the model grid. The major issue with this approach is a computationally efficient localization of the raw ensemble-generated BECs which generally suffer from sampling errors caused by the limited number of ensemble members. In their recent studies, Desroziers et al. (2014) and Liu and Xue (2016) established useful relationships between the 4dVar and 4dEnVar variants with different preconditioners and covariance localization schemes. As a result of these developments, hybrid 4dVar and 4dEnVar methods were implemented operationally in the European (Clayton et al., 2013) and Canadian (Buehner et al., 2013; 2015) weather prediction facilities.

In practice, the 4dEnVar technique is formulated as a search for optimal corrections to the control variables which is performed in the range of the background error covariance $\mathbf{B}$. For that reason preconditioning is often made by the square root of $\mathbf{B}$ and the variational optimization problem is considered in the dual (observation space) formulation which usually has much smaller dimension than the original control space formulation and therefore will be more efficient computationally. In particular, this approach has been adopted in the NAVDAS-AR atmospheric DA system (Rosmond and Xu, 2006).

In the ocean, observations are less abundant than in the atmosphere and the ensemble-based BEC estimates which constitute the backbone of 4dEnVar technique tend to be much less accurate. For that reason, one has to rely on heuristic BEC approximations (e.g. Yaremchuk et al., 2013; Weaver et al., 2015). Development of an efficient a4dVar DA method also becomes more problematic as one has to select a few reliable ensemble perturbations with more care. Early predecessors of practical a4dvar algorithms limited optimization to predetermined low-dimensional subspaces spanned either by the reduced-order approximations of the model Green functions (Stammer and Wunsch, 1996; Menemenlis and Wunsch, 1997), or by the dominant principal orthogonal vectors (EOFs) associated with the model statistics (e.g., Robert et al., 2005; Qui et al., 2007; Hoteit, 2008). The 4dEnVar technique proposed by Liu et al. (2008); 2009), generalizes this approach by representing the search subspace by the Schur products of the ensemble members with the eigenvectors of the reduced-order representation of the localization matrix.

In the present paper, we further develop an iterative ensemble-based 4dVar technique (Yaremchuk et al., 2009) which appears to be competitive with 4dVar in oceanographic applications (Panteleev et al., 2015, Yaremchuk et al., 2016a, hereinafter Y16). A distinctive feature of the technique is its self-sufficiency: in contrast to many ensemble estimation methods which employ a given well-trained ensemble to optimize the control variables within a given time window, the a4dvar sequentially generates search subspaces (bundles of search directions) entirely from the statistics of the model trajectories and/or the respective model-data misfits obtained in the course of optimization. In that respect, the a4dVar technique resembles the 4dVar, which uses the adjoint code to generate a new search direction, whereas in a4dVar that direction is replaced by a search subspace spanned by the ensemble of search directions.

In the previously considered versions of the method search subspaces were built using the leading EOFs of either the model trajectory or the projections of the model-data misfits onto the range of $\mathbf{B}$ at the current iteration inheriting information from either dynamical constraints or modeling errors respectively. The present study blends both approaches in an attempt to improve a4dVar performance and flexibility. In addition, search subspaces are explicitly confined to the range of $\mathbf{B}$, whose structure is constrained by the balance operator, which facilitates searches in hydrostatically and geostrophically balanced directions. To avoid searches in the directions nearly parallel to the ones already explored on the previous iterations, the descent process is restarted by augmenting the search subspaces with the eigenvectors of the background error covariance. It is shown that all these modifications result in a significant improvement in the performance of the algorithm.

The paper is organized as follows. In the next section we briefly describe the basics of 4dvar methodology and its ensemble-based (4dEnVar) variants, outline the a4dvar method, and describe considerations in support of the proposed hybrid methodology of selecting the search subspaces. In Section 3, performance of the a4dVar technique is analyzed using NCOM configuration in the Adriatic sea with a particular focus on the impact of balance constraints on the forecast skill and of the new restart procedure on the convergence rate. Summary and discussion of the results conclude the paper.

2. Variational optimization methodologies

We follow the terminological convention proposed by Lorenc (2013), and refer to “4dEnVar” for the adjoint-free optimization algorithms that recover the gradient information from predetermined ensembles which are intended to capture the dominant features of the BEC structure. The a4dVar algorithm being tested here is designed to perform without a given ensemble: Instead of the BEC model derived from the ensemble, we use a heuristic BEC model, which we believe contributes to a more robust strategy in the face of sparse data. We then iteratively retrieve ensemble members (search directions) either from a model trajectory on current iteration, or from dominant spectral modes of the BEC matrix computed off-line.

2.1. 4dVar

In order to better illuminate connections between the 4dVar framework and what follows, the 4dVar approach in this section is formulated as a linear discrete least-squares problem constrained by model dynamics in a small vicinity of the model's background trajectory $\mathbf{x}_b^n$:

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

where $\mathbf{x}^n$ are the deviations of the model state from $\mathbf{x}_b^n$ at time t_n, n enumerates observation times, $\mathbf{B}$ is the BEC matrix of $\mathbf{x}_b^n$ which describes the (Gaussian) statistics of the model state at $n = 0$, $\mathbf{H}_n$ are the model-data projection operators, $\mathbf{d}^n$ are the discrepancies $\mathbf{d}_n^* - \mathbf{H}_n\mathbf{x}_b^n$ between observations $\mathbf{d}_n^*$ and the corresponding background model values, $\mathbf{R}_n$ are the observation error covariances, and $^{\mathrm{T}}$ denotes transposition. If $\mathbf{B}$ is rank-deficient, $\mathbf{B}^{-1}$ is to be understood as a Moore–Penrose pseudoinverse. We will denote the dimension of the discretized model state vector $\mathbf{x}$ by M and the number of observations available at time t_n by L_n.

The correction vectors, $\mathbf{x}^n$, are governed by the recursive relationship

$$\mathbf{x}^n = \mathbf{M}_n\mathbf{x}^{n-1}, \qquad (2)$$

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

$$\mathbf{x}^n = \mathbf{M}_n\mathbf{M}_{n-1}\ldots\mathbf{M}_2\mathbf{M}_1\mathbf{x}^0. \qquad (3)$$

Introduce the preconditioned variable $\mathbf{c} = \mathbf{B}^{-1/2}\mathbf{x}^0$ for the control vector, where $\mathbf{B}^{-1/2}$ is the square root of $\mathbf{B}^{-1}$, and denote the aggregated n-step propagator as $\mathbf{M}^n \equiv \mathbf{M}_n\ldots\mathbf{M}_2\mathbf{M}_1$. Define (briefly)

$\overline{\mathbf{H}}_n = \mathbf{R}_n^{-1/2}\mathbf{H}_n$, $\overline{\mathbf{d}}^n = \mathbf{R}_n^{-1/2}\mathbf{d}^n$, and then simplify the notation by dropping overbars. Taking (3) into account, the minimization problem (1) can be rewritten in terms of the correction $\mathbf{c}$ to the initial state:

$$J = \frac{1}{2}\left(\mathbf{c}^\top\mathbf{c} + \sum_{n=1}^{N} |\mathbf{Q}_n\mathbf{c} - \mathbf{d}^n|^2\right) \to \min_{\mathbf{c}}. \tag{4}$$

where $\mathbf{Q}_n := \mathbf{H}_n\mathbf{M}^n\mathbf{B}^{1/2}$. The 4dVar DA method finds the minimum of J by solving the associated normal equation:

$$\nabla_{\mathbf{c}} J = \mathbf{c} + \sum_n \mathbf{Q}_n^\top(\mathbf{Q}_n\mathbf{c} - \mathbf{d}^n) = 0, \tag{5}$$

The 4dVar method uses a descent algorithm employing gradient information (5) which requires computation of the action of $\mathbf{Q}_n^\top$, and, as a consequence, requires knowledge of the action of the adjoint model $\mathbf{M}^{n\top}$ on the vector of control variables. To keep the discussion that follows concise, denote the identity matrix by $\mathbf{I}$ and introduce the Hessian matrix $\hat{\boldsymbol{H}}$ and the right-hand side $\mathbf{r}$ in Eq. (5),

$$\hat{\boldsymbol{H}} = \mathbf{I} + \sum_n \mathbf{Q}_n^\top\mathbf{Q}_n; \quad \mathbf{r} = \sum_n \mathbf{Q}_n^\top\mathbf{d}^n. \tag{6}$$

Then (5) may be rewritten as $\hat{\boldsymbol{H}}\mathbf{c} - \mathbf{r} = \mathbf{0}$.

In the non-linear case, the background solution $\boldsymbol{x}_b$ is recomputed to initiate the next (outer) loop of the 4dVar minimization process after iteratively solving Eq. (5) (e.g., Courtier et al., 1994). Alternatively, the gradient and the non-linear cost function can be utilized in the iterative loop of a single non-linear minimization algorithm.

Developments in ensemble methods have shown that the use of tangent linear/adjoint models for the gradient computation could be avoided by employing finite-difference approximations using the ensemble members in conjunction with localization. The efficiency of this approach and its superior parallelization and scalability features gave rise to rapid development of the 4dEnVar methods (e.g., Liu and Xue, 2016).

2.2. 4dEnVar

The 4dEnVar approach is based on combining the preconditioning of $\mathbf{x}^0$ by either $\mathbf{B}^{1/2}$ or $\mathbf{B}$ with an algorithm for explicit multiplication of the model state by the preconditioner. In a simple case of representing $\mathbf{B} = \mathbf{B}^{1/2}\mathbf{B}^{\top/2}$ by the raw ensemble average (in which case $\mathbf{B}^{1/2}$ is the $M \times m$ matrix listing m ensemble members columnwise), the matrices $\mathbf{Q}_n$ have a relatively small size ($L_n \times m$) and can be transposed explicitly to compute a gradient for J. Since the raw ensemble covariance estimates are susceptible to sampling errors, localization could be performed by taking the Schur product of $\mathbf{B}$ and the localization matrix $\mathbf{C}$. In the case of $\mathbf{B}^{1/2}$-preconditioning this operation requires an explicit low-rank (l) representation of $\mathbf{C}$ which increases the size of $\mathbf{Q}_n$ to $L_n \times ml$. The low-rank restriction on $\mathbf{C}$ could be avoided because the action of the localized ensemble covariance on a state vector can be computed explicitly at a relatively low cost using various versions of the domain localization technique (e.g., Janjic et al., 2011).

Recently, the 4dEnVar technique was extended by reformulating the problem in the space of model trajectories and using the respective ensemble *space-time* covariances. This generalization relaxes the strong dynamical constraint (2) and allows smooth transition to weak 4dEnVar formulation in both primal and dual forms (see Desroziers et al., 2014 and references therein).

In general, the 4dEnVar approach provides a powerful optimization tool with a number of useful properties: (a) It is flexible in formulating the optimization problem using various dynamical constraints; (b) the BEC model representation is versatile; (c) The potential exists for much better performance on massively parallel computer architectures; and (d) One avoids the necessity of development and maintenance of the tangent linear and adjoint models. For atmospheric DA applications, the 4dEnVar has already demonstrated its competitiveness, and in some cases, superiority, in comparison with standard 4dVar, and this is mostly due to the use of flow-dependent covariances derived from the underlying ensemble (Fairbairn et al., 2014; Lorenc et al., 2015).

2.3. a4dVar

In regional oceanographic practice, observations are sporadic and relatively sparse compared to those in atmosphere. As a consequence, skillful ensembles are rarely available, so the efficiency of optimization largely depends on the structure of $\mathbf{B}$ used for regularization of the problem. In that respect it is worth considering minimization algorithms which do not require availability of the gradient and are able to accumulate information on the Hessian structure in the course of optimization (e.g., Gratton et al., 2014; Ruiz and Sandu, 2016). These "derivative-free" techniques draw search directions from the (Gaussian) pdf specified by $\mathbf{B}$ and show reasonably good performance, especially in the case of significant non-linearity in the dynamics.

A somewhat more heuristic a4dVar approach has been explored by Yaremchuk et al. (2009); 2016a), who proposed an iterative procedure of updating search directions (ensemble members) in the form of model-data misfits smoothed by the background error covariance. A more traditional method of updating the ensemble by the leading EOFs of the model trajectory at the current iteration was tested by Panteleev et al. (2015) in application to the surface wave model (WAM) and demonstrated a reasonably good performance in a set of twin-data assimilation experiments.

The basic idea of a4dVar approach relies on the possibility of low-cost minimization of (4) in a given subspace spanned by the ensemble members using the technique of Zupanski (2005), and low-cost $\hat{\boldsymbol{H}}$-orthogonalization of the current subspace to the previous ones which can be implemented by storing perturbations of the control variables together with respective model-data misfits from the ensemble runs (Appendix B in Y16). Efficiency of the a4dVar method is based on parallelism in minimizing the cost function in the subspace spanned by the ensemble members: In contrast to 4dVar, where a new search direction is found after *sequential* runs of the direct and adjoint codes, the a4dVar method explores multiple search directions, generated by *parallel* runs of the direct code.

Adopting the notation $\mathbf{d} = [\mathbf{0} \;\; \mathbf{d}^1 \ldots \mathbf{d}^N]^\top$, and $\mathbf{A} = [\mathbf{I} \;\; \mathbf{Q}_1^\top \ldots \mathbf{Q}_N^\top]^\top$ for the Hessian square root, the (linearized) cost function (4) is rewritten as

$$J = \frac{1}{2}(\mathbf{A}\mathbf{c} - \mathbf{d})^\top(\mathbf{A}\mathbf{c} - \mathbf{d}) \tag{7}$$

producing a minimizer that leads to the solution of (1):

$$\mathbf{c}_\star = (\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top\mathbf{d} \equiv (\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{r} \tag{8}$$

Solution (8) can be obtained by the iterative process of the form

$$\begin{aligned}\mathbf{c}^{i+1} &= \mathbf{c}^i + \mathbf{P}^i[(\mathbf{A}\mathbf{P}^i)^\top\mathbf{A}\mathbf{P}^i]^{-1}(\mathbf{A}\mathbf{P}^i)^\top\mathbf{d} = \\ &= \mathbf{c}^i + \mathbf{P}^i[\mathbf{P}^{i\top}\hat{\boldsymbol{H}}\,\mathbf{P}^i]^{-1}\mathbf{P}^{i\top}\mathbf{r}\end{aligned} \tag{9}$$

where i is the iteration number, and $\mathbf{P}^i$ is an $M \times m$ matrix of search directions listed columnwise and spanning the m-dimensional minimization subspace. The process (9) converges to the 4dVar solution (8) in at most rank$(\mathbf{B})/m$ steps if all $\mathbf{P}^i$ are kept mutually $\hat{\boldsymbol{H}}$-orthogonal (Appendix A).

For truly non-linear applications as considered below, non-linear effects make it difficult to retain $\hat{\boldsymbol{H}}$-orthogonality with sufficient accuracy throughout the iterative process. Even for linear problems, retaining global orthogonality for Krylov methods (e.g., GMRES) is costly; yet alternative strategies that enforce only local orthogonality (e.g., Lanczos methods) can exhibit a slow deterioration of global orthogonality with a consequent slowing of convergence as a result. Non-linearity tends to contribute a quick loss of orthogonality in most practical (non-linear) applications, so that the iterative process has to be restarted, and the overall efficiency depends on the selection of $\mathbf{P}^i$.

Since the practical number of iterations rarely exceeds one hundred, the key requirement for the search directions is to have sizable projections on the direction towards the minimum. The a4dVar method does not require neither computations nor approximations of the cost function gradient and employs heuristic approaches to generating $\mathbf{P}^i$. These approaches were shown to be competitive with 4dVar in idealized settings (Yaremchuk et al., 2009), and in application to assimilation of the real data in the Adriatic Sea (Y16).

2.4. Selection of search subspaces in a4dVar

In the previous versions, the a4dVar was tested in non-linear regimes with two configurations, where the columns of $\mathbf{P}^i$ are composed of either the leading EOFs of the model trajectory (Yaremchuk et al., 2009; Panteleev et al., 2015), or of the model-data misfits $\mathbf{s}^n$ smoothed by $\mathbf{B}$ (Yaremchuk et al., 2009; 2016a). The leading EOFs of the trajectory $\tilde{\mathbf{x}}^n = \mathbf{x}_b^n + \mathbf{x}^n$ tend to have sizable projections on the most persistent (time-correlated) components of the error fields, whereas search directions specified by $\mathbf{s}^n = \mathbf{B}\mathbf{H}^\mathrm{T}(\mathbf{H}\mathbf{x}^n - \mathbf{d}^n)$ account for the spatial distribution of the observations and bring in information on the background error covariance. When combined, these vectors appear to form search subspaces generally with larger projections on the leading modes of inverse Hessian. In the present study we explore such a hybrid strategy of generating $\mathbf{P}^i$ and assess performance of the hybrid a4dVar (ha4dVar) method formulated as follows:

(a) $\mathbf{P}^i$ are built by extracting the leading modes from the hybrid sequences

$$\mathbf{Z}_i = \{\tilde{\mathbf{x}}_i^1, \mathbf{s}_i^1, \dots, \tilde{\mathbf{x}}_i^N, \mathbf{s}_i^N\}. \tag{10}$$

The modes are computed with respect to the norm induced by the background error covariance:

$$\mathbf{Z}_i\mathbf{Z}_i^\mathrm{T}\hat{\mathbf{P}}_i = \mathbf{B}\hat{\mathbf{P}}_i\boldsymbol{\Lambda}_i; \quad \mathbf{P}_i = \mathbf{B}\hat{\mathbf{P}}_i \tag{11}$$

where $\boldsymbol{\Lambda}_i$ is the $m \times m$ diagonal matrix of the respective eigenvalues and the horizontal means are removed from the *TS* constituents of $\tilde{\mathbf{x}}_i^n$ prior to the analysis (11). The columns of $\mathbf{P}_i$ are $\hat{\boldsymbol{H}}$-orthogonalized to the previous search directions until the accumulated Hessian spectrum degenerates, causing stagnation of the minimization process.

(b) when the minimization process slows down, it is restarted by drawing a basis of the new search subspace from precomputed eigenvectors of $\mathbf{B}$ (see Section 3.3 for the definition of the restart criterion).

(c) the BEC matrix $\mathbf{B}$ is generalized to include balance constraints (see Appendix B)

Note that in accord with the $\mathbf{B}$-preconditioning principle, the optimization process is automatically restricted to the range of $\mathbf{B}$ due to the strategy of selecting the search directions in *(a)* and *(b)*. This makes a4dVar more consistent with the data-space 4dVar formulation which seeks for optimal corrections of the control variables in the range of $\mathbf{B}$ and is used here for comparison purposes.

The updates (*a-b*) of the a4dVar method remain heuristic in nature. However, numerical experimentation shows that they provide better approximations to the directions toward minimum, resulting in a better convergence rate for practical assimilation windows (typically ~ 2–4 days). In that respect it is worthwhile to note that search directions specified by $\mathbf{s}^n$ asymptotically approximate components the cost function gradient when $\mathbf{M}^{n\mathrm{T}} \to \mathbf{I}$). Second, the restarting procedure (*b*) draws search directions from the leading eigenvectors of another limiting case of the inverse Hessian (when observation errors are large and $\hat{\boldsymbol{H}}^{-1} \to \mathbf{B}$). In oceanographic applications, this restarting approach provides a better alternative to the randomized restarts employed, for example, by the breakdown-free GMRES algorithms (e.g., Reichel and Ye, 2005).

The third modification (*c*) of the algorithm has been made to introduce the ability of (partly) constraining the optimization process to the balanced manifold with an option to control the degree of that constraint by tuning the magnitude of unbalanced components.

Overall, the ha4dVar method can be summarized as follows:

0. Specify the dimension m of the search subspaces, the maximum number of iterations I, the perturbation magnitude ε, the restart parameter γ_c, and the maximum number of restarts n_r. Compute the $N \times mn_r$ matrix $\mathbf{B}_n$ of the first mn_r eigenvectors of $\mathbf{B}$ to be used for restarts (Appendix B). Set $\tilde{\mathbf{c}}_0 = \mathbf{x}_b^0$, the iteration number $i = 0$ and the restart parameter $\gamma = 1$.
1. Compute (suboptimal) model trajectory $\tilde{\mathbf{x}}_i^n$, auxiliary vectors $\mathbf{Y}_i = \hat{\boldsymbol{H}}^{1/2}\tilde{\mathbf{c}}_i$, $\mathbf{Z}_i$ and the matrix of search directions $\mathbf{P}_i$ (Eq. (11)).
2. Perturb the initial conditions $\tilde{\mathbf{c}}_i \to \tilde{\mathbf{c}}_i + \varepsilon\mathbf{p}_i^j$ by the jth column of $\mathbf{P}_i$ and run (in parallel) the ensemble of m perturbed models, computing the perturbed values δJ_i^j and $\delta\mathbf{Y}_i^j$, $j = 1, \dots, m$ required for $\hat{\boldsymbol{H}}$-orthogonalization (Y16).
3. $\hat{\boldsymbol{H}}$-orthogonalize the search basis $\{\mathbf{p}_i^j\}$ with respect to the basis vectors obtained on the previous iterations and compute optimal corrections $\delta\mathbf{c}_i$ by solving the normal equation in the search subspace.
4. Set $\tilde{\mathbf{c}}_{i+1} = \tilde{\mathbf{c}}_i + \delta\mathbf{c}_i$.
5. Compute relative contribution $\gamma = 1 - \mathrm{Tr}\hat{\boldsymbol{H}}_{i-1}/\mathrm{Tr}\hat{\boldsymbol{H}}_i$ of the ith search subspace to the Hessian spectrum accumulated since the last restart (see Section 3.3 for details). If the value of γ is less than γ_c, restart the optimization process by populating $\mathbf{P}_{i+1}$ with the next dominant set of unexplored eigenmodes from $\mathbf{B}_n$.
6. If $i = I$ exit. Otherwise set $i \leftarrow i + 1$, then go to 2.

In the linear case, the iterative process outlined above would be equivalent to (9) in the absence of restarts that are caused by gradual degeneration of the search subspaces. In application to the non-linear problem considered, degeneration is due to both non-linearity and a possibility for the new search directions to be spanned by the previous ones. In the linear 4dVar this possibility is (formally) absent since the search directions are built on Hessian polynomials acting on the initial gradient (residual of the normal system). However, in the applications involving non-linearity and finite computer precision, restarting is an obligatory feature of 4dVar as well.

Although the a4dVar method (9)–(11) remains essentially heuristic in nature, it is based on the possibility of the low-cost computation of search directions and their $\hat{\boldsymbol{H}}$-orthogonalization, which in turn provides a background for the algorithm's convergence in a relatively small number of iterations. To assess ha4dVar performance, we conducted a series of data assimilation experiments with NCOM model which had identical configuration to the one used in Y16. Results of the experiments are compared with optimizations performed by the NCOM 4dVar and a4dVar algorithms reported in Y16.

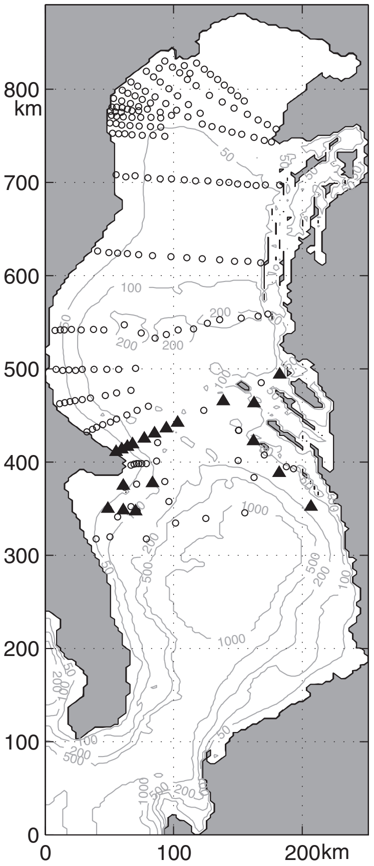


Fig. 1. Model domain with CTD stations (circles) and moorings (triangles) of the DART experiment. Gray contours (m) show bottom topography.

3. Experimental setting

3.1. Data and model

The observations used in this study were conducted in August 2006 (see Burrage et al. (2009) and references therein) and spanned the period August 14–29. The assimilated dataset consisted of 5648 temperature and salinity (*TS*) observations (133 vertical profiles) and 3548 horizontal velocity components (*uv*) acquired at 19 ADCP moorings during the first four days (August 14–17). Velocity observations were available in the depth range 15–150 m every 12 h. The remainder of the data (4002 *TS* observations from 86 profiles and 9958 *uv* values respectively) observed between August 18 and 29 were used to assess the forecast skill improvement delivered by the data assimilation. The results were also compared with 4dVar data assimilation applied to the same data using the same numerical model configuration.

The considered assimilation methods are strongly constrained by the NCOM, 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 the 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 model was configured at δx =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 from the global NCOM (Barron et al., 2004) solution for the region. The model was forced by the river runoff and atmospheric fields derived from the regional atmospheric model (Ivatek-Sahdan and Tudor, 2004). The model setting was identical to the one used in Y16 for the 4dVar/a4dVar comparison.The mean distance of the model fields from their initial state on August 14, 2006 ($n = 0$) averaged over the 4-day assimilation window was 0.58 when normalized by the rms variability of the fields at $n = 0$.

In the described assimilation experiments, initial conditions were used as control variables, i.e., the vector $\mathbf{c}$ 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.

3.2. The background error covariance

The cost function of the ha4dvar assimilation system was slightly different from the a4dVar cost function used in Y16. The difference is in the background error correlation model, which is now identical to the one used by 4dVar. Better compatibility with 4dVar formulation was achieved by $\mathbf{B}^{1/2}$-preconditioning of the problem which allowed us to employ the Gaussian correlation model instead of its low-order approximation used in Y16. In particular, the BEC matrix was defined by the product $\mathbf{VCV}$, where $\mathbf{V}$ is the diagonal matrix of the background error rms variances and $\mathbf{C}$ is the respective correlation matrix represented implicitly by the kernel of the heat transfer equation

$$\mathbf{C} = \frac{\delta x^2}{2\pi r^2} \exp\left(\frac{1}{2} r^2 \Delta\right). \tag{12}$$

Here r is the decorrelation length scale and Δ is the discretized 2d Laplacian operator. Numerically, the action of $\mathbf{C}$ on a state vector is computed by integrating the heat transfer equation (e.g., Weaver and Courtier, 2001). The diagonal elements of $\mathbf{V}$ and $\mathbf{R}$ were computed from the statistics of the first guess model run and observations as in Y16. Similarly, the value of r was chosen to be 9 km to be consistent with the estimates (e.g., Cushman-Roisin and Korotenko, 2007) of the Rossby deformation radius in the Adriatic.

The a4dVar system was also upgraded with an option of incorporating balance constraints into the structure of the background error covariance (Appendix B). The spatial correlations within the unbalanced velocity and SSH fields were described by (12) with a smaller decorrelation scale r_a used as a tunable parameter in the assimilation experiments. The respective error variances were also tuned and found to be close to the mean ratio $\gamma = 0.15$ of the squared magnitudes of the unbalanced (divergent)) and balanced (geostrophic) velocity components in the background solution (Yaremchuk and Martin, 2016b)

3.3. a4dVar parameters

In the reported assimilation experiments we used the data acquired in the 4-day period between 6 UTC August 14, and 6 UTC August 18, 2006. These data were projected on $N = 8$ time layers centered at 0 and 12 UTC. At every iteration the hybrid sequence $\{\mathbf{x}^0, \mathbf{x}^1, \mathbf{s}^1, \ldots, \mathbf{x}^N, \mathbf{s}^N\}$ contained 17 members, including the initial conditions $\mathbf{x}^0$. The ensemble size $m = 16$ was chosen to maximize parallelization efficiency of the code which was run on IBM iDataPlex supercomputer facility.

The a4dVar search directions were computed as the leading modes of the hybrid ensemble which blends two types of state-space vectors: the snapshots of model trajectory $\tilde{\mathbf{x}}^n$ and the $\mathbf{B}$-smoothed model-data misfits $\mathbf{s}^n = \mathbf{BH}_n^{\mathrm{T}}(\mathbf{H}_n\mathbf{x}^n - \mathbf{d}^n)$. To make the contributions of $\tilde{\mathbf{x}}^n$ and $\mathbf{s}^n$ to the ensemble statistics compatible, the vectors $\mathbf{x}^n$ were multiplied by the square root of the respective

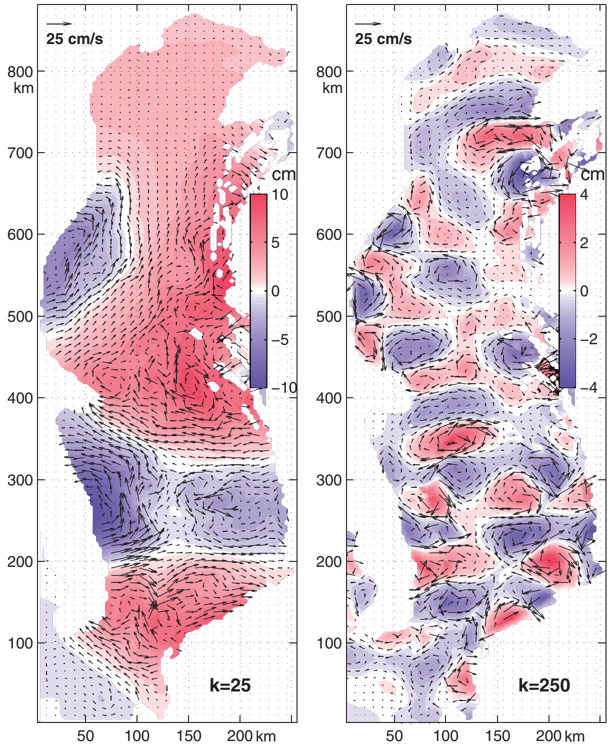


Fig. 2. Surface velocity and SSH constituents of the search directions associated with spectral decomposition of the balanced background error covariance matrix. Index k enumerates the respective eigenvalues in the descending order.

magnitude ratios:

$$\alpha^n=\sqrt{\frac{\mathbf{s}^{n\mathrm{T}}\mathbf{V}^{-2}\mathbf{s}^n}{\mathbf{x}^{n\mathrm{T}}\mathbf{V}^{-2}\mathbf{x}^n}}\equiv\sqrt{\frac{\langle\mathbf{s}^n,\mathbf{s}^n\rangle}{\langle\mathbf{x}^n,\mathbf{x}^n\rangle}} \qquad (13)$$

In the course of experiments it was found that variations of α with time n have little impact on the a4dVar convergence, so we have used a constant value of $\alpha = 0.18$ after the experimental adjustment of its magnitude (see Section 4.3).

A partial spectral decomposition of $\mathbf{B}$, required for restarts, was performed off-line by means of the LAPACK software for finding the largest eigenvalues of $\mathbf{C}^{1/2}$. To obtain the search directions for restarts (Fig. 2), the respective eigenvectors were multiplied by the matrix factorizing $\mathbf{B}$ to block-diagonal form (Eq. B5, Appendix B). Given the relatively small dimension of the search subspace ($m = 16$), we limited ourselves to computing the first 480 eigenvectors, which provided search directions for 480/m=30 restarts.

The search subspace generation was restarted when the spectrum of the Hessian projection reached a prescribed degree of degeneracy γ. Specifically, if the cost function has been minimized over $m(l-1)$ $\hat{\boldsymbol{H}}$-orthogonal directions ($l-1$ subspaces), minimization in the next (lth) subspace was performed if

$$1-\frac{1}{\mathrm{Tr}\hat{\boldsymbol{H}}_{lm}}\sum_{j=1}^{ml-m}\xi_j=\sum_{j=ml-m+1}^{ml}\xi_j\left[\sum_{j=1}^{ml}\xi_j\right]^{-1}>\gamma \qquad (14)$$

where ξ_j are the descending-order eigenvalues of the Hessian projection $\hat{\boldsymbol{H}}_{lm}$ on the new lm-dimensional subspace. If the criterion (14) was not met, l was reset to 1, search directions were initialized as the next dominant modes extracted from the BEC spectrum, and the background state was reset to the current suboptimal state. In handling the effects of non-linearity, the a4dVar restart procedure is analogous to 4dVar outer loop with the only difference that the adjoint-based 4dVar search direction is replaced by the search subspace extracted from the BEC spectrum.

4. Results

To assess the ha4dVar performance, we conducted a series of assimilation experiments and compared the results with 4dVar and a4dVar assimilations reported in Y16. The comparison criteria were the forecast skill and computational cost.

4.1. Computational cost

Fig. 3 compares the total cpu time τ and wall time τ_w required by the NCOM 4dVar and a4dvar algorithms. The vertical axis shows reduction of the model-data misfit given by the second term in the non-linear version of Eq. (1), normalized by its initial value. In both cases computer time is normalized by the time of one a4dVar iteration (i.e., basically one ensemble run) which was executed on 9 cpus. For this reason the ha4dVar method parallelized on 16 cpus appears to be less efficient on the first iteration since both methods deliver similar reduction of J' (compare black and red curves in the left panel). However, after the second iteration the ha4dVar cost function values keep being below the red curve throughout the entire minimization process.

In terms of the wall time (right panel in Fig. 3), the a4dVar and ha4dVar values of J' appear to be nearly identical on the first iteration, which is explained by the dominant role of the eight model-data misfit modes in solving the normal system. In that respect there might be some room for increasing the ha4dVar computational efficiency by implementing an algorithm for automatic selection of the number m of search directions (e.g., Uzunoglu et al., 2007).

Comparison with 4dVar (thick blue curves in Fig. 3) shows certain advantage of a4dVar in terms of τ during the first 10–20 iterations (left panel in Fig. 3) which becomes more prominent if the wall times are considered (right panel). With seven inner iterations per outer loop (marked by squares in Fig. 3) the 4dVar total cpu time per outer loop is approximately equivalent to 6.5 a4dVar and 3.9 ha4dvar iterations, that is roughly 70 direct model runs τ_m in either case.

The (h)a4dVar advantage can be explained by at least two factors: a) compared to the native non-linear NCOM code, execution of the linearized model codes within an outer loop is several times more expensive due to the necessity to extract from memory (or recompute) certain features of the background model trajectory; b) the NCOM 4dVar code has been designed for running operationally in the weak constrained mode and, therefore has more complex structure than the a4dVar code tailored specifically for the NCOM configuration described in Section 3.1. The second factor can be roughly accounted for by assuming that a 4dVar inner iteration requires $5\tau_m$ (i.e. $2.5\tau_m$ for the linearized and adjoint model runs). The respective convergence curves are shown by thin blue lines and labeled by 4dVar*) in Fig. 3. In this case 4dVar outperforms ha4dVar in terms of τ (left panel), but still lags far behind in terms of τ_w due to the necessity to sequentially perform the linearized model runs within the outer loops.

It is also noteworthy that in terms of τ the a4dVar (ha4dVar) schemes deliver similar reductions to J' after 3 (2) iterations as the 4dVar method after the first outer loop (thin blue line in the left panel of Fig. 3). The ha4dVar scheme remains τ-competitive to 4dVar within 2–3 outer loops that are usually executed in practical applications (e.g., Bonavita et al., 2017).

The 4dVar computational efficiency could also be improved by making fewer (than 7) iterations within an outer loop and making less computationally expensive approximations to the linearized

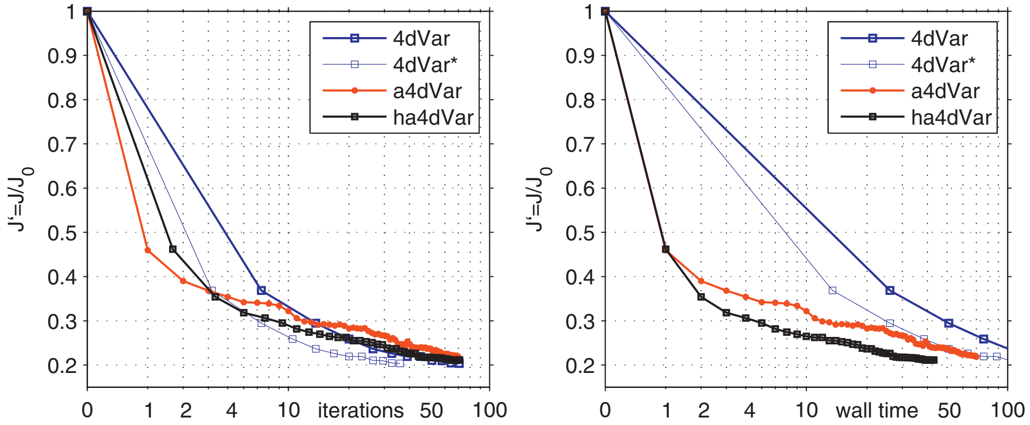


Fig. 3. The normalized cost function value J' against the total cpu time (left) and wall time (right) required by the 4dVar method and two versions of a4dVar algorithm. The x-axis in the left panel is normalized by the total CPU time required by one a4dVar iteration described in Y16. Similar normalization is used in the right panel, but with respect to the wall time.

code (e.g., for example, removing linearizations (Ngodock and Carrier, 2014) with respect to time variations of σ-coordinates in the uppermost layer). Also, we did not take into account the computational expense of the partial spectral decomposition of $\mathbf{B}$ required by ha4dVar (black lines in Fig. 3). Nevertheless, it may seem evident that a4dvar is likely to maintain its wall time advantage on the massively parallel computers because of the sequential nature of the 4dVar method.

4.2. Forecast skill

In assessing the forecast skill F we followed the general approach of Y16 while using the 4dVar skill as a benchmark for normalization. The procedure is outlined as follows. First, an optimal solution is obtained by minimizing the model-data misfits in the first 4 days (August 14–17). Then, the optimized state at 0 UTC August 18 is used for model integration to 0 UTC on August 29, to obtain 20 model snapshots $\mathbf{x}^9, \ldots, \mathbf{x}^{28}$ at 12 h discretization. After that the cost function model-data misfit term J_f^a is daily averaged between August 18 and 28,

$$\tilde{J}_f^k = \sum_{n=7+2k}^{8+2k} (\mathbf{H}_n\mathbf{x}^n - \mathbf{d}^n)^{\mathrm{T}}(\mathbf{H}_n\mathbf{x}^n - \mathbf{d}^n), \quad k = 1, \ldots, 10 \qquad (15)$$

and then normalized by the corresponding 4dVar values J_f^k obtained after 4-day optimization in Y16. The forecast skill is defined as the square root of the respective ratio $F = [\Sigma_k \tilde{J}_f^k / J_f^k]^{1/2}$ To distinguish between the forecast skills for different types of observations $(T, S, \boldsymbol{u})$ we also estimated the values of $F_{T,S,\boldsymbol{u}}$ by separate computations of the respective cost function terms in (15).

The procedure outlined above is limited in scope and does not constitute a comprehensive assessment of a4dvar performance. Such an assessment would require analyses of multiple data sets from various regions of the World Ocean and lies beyond the scope of this study, which has the more modest objective of showing improvements that can be delivered by the ha4dVar method.

Fig. 4 shows the daily-averaged values of the model-data misfit $(J_f^k/n_d)^{1/2}$ in (15) for various assimilation methods normalized by the respective numbers of observations n_d whose relative values are given by shaded rectangles. The ha4dVar model-data misfits (solid red) lie pretty close to the a4dVar (red dashes) and 4dVar (blue) results of Y16. Compared to a4dVar, minor improvements are observed on August 15, 19 and 22–26, while the value of model-data discrepancy is somewhat higher on August 14. This could be explained by the fact that ha4dVar control variables are additionally constrained by the balance relationships. In terms of the overall forecast skill F the ha4dVar demonstrates a modest (3%) improvement compared to a4dVar.

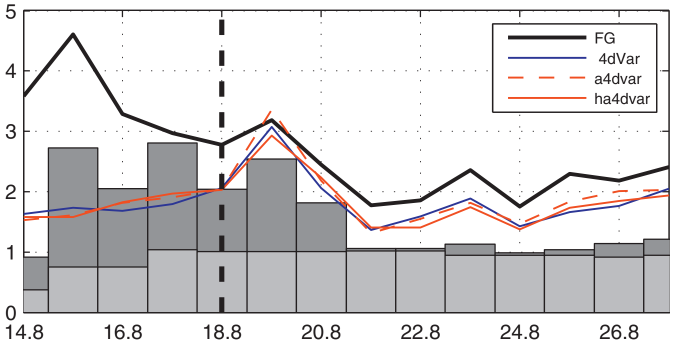


Fig. 4. Forecast skills of the optimized solutions. The relative number of the respective data points for each day is shown by gray shaded rectangles with lighter rectangles corresponding to velocity data. Vertical dashed line shows the time interval of data assimilation. The black line corresponds to the model-data misfits of the first guess (FG) model run.

Table 1
The forecast skill of the a4dVar and ha4dVar techniques relative to 4dVar for the balanced (third line) and unbalanced (second line) BEC models. Values less than 1 correspond to skill improvement as compared to 4dVar.

	F_T	F_S	$F_{\boldsymbol{u}}$	F
a4dVar	1.103	1.021	0.994	1.028
ha4dVar*	1.098	1.025	0.973	1.020
ha4dVar	1.046	1.034	0.924	0.991

Table 1 summarizes the forecast skills F for various fields in three optimization experiments: using the a4dvar method reported in Y16 (first line), and using the hybrid scheme with or without the balanced background error covariance. Compared to 4dVar, the a4dvar method produces slightly (2–10%) less accurate forecast of temperature and salinity, but appears to be 1–8% better in forecasting the velocity field (third column in Table 1). This advantage becomes more significant with the balanced background error covariance version of ha4dVar (last line in Table 1). This experiment tends to perform searches in the directions which generate more persistent (geostrophically balanced) velocity and SSH fields such as those exposed in Fig. 1. Balance constraints also appear to have

Table 2
Sensitivity of the ha4dVar forecast skill with the optimally balanced ($r_a = 4.3$ km, β =0.28) background error covariance to variations of α and γ.

	$\alpha = 0.05$				$\alpha = 0.14$				$\alpha = 0.50$			
γ	0.0	0.012	0.1	0.25	0.0	0.012	0.1	0.25	0.0	0.012	0.1	0.25
F_T	1.184	1.104	1.093	1.198	1.127	*1.046*	1.044	1.183	1.208	1.094	1.089	1.236
F_S	1.066	0.983	0.969	1.095	1.061	*1.034*	1.033	1.089	1.078	1.014	1.005	1.114
$F_{\boldsymbol{u}}$	1.085	0.974	0.971	1.118	1.074	*0.924*	0.931	1.102	1.103	0.994	0.983	1.123
F	1.101	1.008	0.999	1.128	1.081	*0.991*	0.993	1.119	1.119	1.024	1.015	1.146

a positive effect on the temperature field, improving its forecast skill by 5% (first column in Table 1).

On the opposite, salinity forecast skill becomes slightly (1%) worse. This can be explained by the fact that contribution of salinity variations into the density anomalies become significant only in the shallow northern part of the Adriatic Sea strongly affected by the outflow of the Po river. The depths in this region are well below 50 m (Fig. 1), where balance constraints become less applicable.

It should also be noted that the reference 4dvar optimization was performed in Y16 without introducing balance constraints into the background error covariance. This 4dVar option is currently under development and could possibly affect the numbers in Table 1 in favor of 4dVar.

4.3. Sensitivity to a4dVar parameters

A large series of experiments were performed to assess the ha4dVar sensitivity to its free parameters. Specifically, we varied *a)* the decorrelation scale r_a of the unbalanced (ageostrophic) SSH/velocity components (Eq. (17)); *b)* the relative magnitude β of the error variance of unbalanced components $\mathbf{V}_2 = \beta\tilde{\mathbf{V}}$, where $\tilde{\mathbf{V}}$ is the diagonal matrix of rms variations of the velocity and SSH fields in the background solution; *c)* the relative weight α of the model snapshots $\mathbf{x}^n$ in the hybrid ensemble (Eq. (13)); and *d)* the parameter γ (Eq. (14)) triggering the restarts.

The first guess value of β was established through the statistical analysis of the background horizontal velocity field $\boldsymbol{u}_b$. It was defined by

$$\beta = r\frac{\overline{|\mathrm{div}\boldsymbol{u}_b|}}{\overline{|\boldsymbol{u}_b|}} \qquad (16)$$

where $r = 9$km is the local Rossby deformation radius and overbar denotes averaging over the upper 290 m and time of the model integration (August 14,–August 28, 2006). The resulting value was found to be 0.34, indicating the presence of a significant ageostrophic component, primarily associated with inertial oscillations and upper-layer Ekman dynamics. In a subsequent series of assimilation experiments this value was fine-tuned ($\beta = 0.28$) to maximize the forecast skill.

A similar series of experiments was performed to tune the ageostrophic decorrelation scale r_a in the correlation model for the respective background error covariance $\mathbf{B}_2$:

$$\mathbf{B}_2 = \frac{\delta x^2 \beta^2}{2\pi r_a^2}\mathbf{V}\exp\left(\frac{1}{2}r_a^2\Delta\right)\mathbf{V} \qquad (17)$$

The optimized parameters $r_a = 4.3$ km, $\beta = 0.28$ were then used to study sensitivity of the ha4dVar solutions to the variations of α and γ.

Table 2 shows the basic results of these experiments. The best forecast skill achieved with α=0.14 and γ=0.012 is italicized in the middle column (also shown in the last line of Table 1). In the first series of experiments we adjusted the ensemble weighting parameter α using its time-mean background value 0.34 (Eq. (13)) as a first guess. The value of α was varied in the range between 0.05 and 0.5. Although the overall forecast skill F was weakly sensitive to the choice of α (last line in Table 2), it showed a distinct minimum at $\alpha = 0.14$, and, more importantly, had a more significant effect on the convergence rate. All assimilation experiments were terminated when the total cpu time τ reached the benchmark 4dVar value τ_{4dVar} approximately equivalent to 70 a4dVar or 42 ha4dVar iterations (left panel in Fig. 3). However, for the values of α between 0.1 and 0.3 the 1% difference from the limiting value of the cost function was achieved 1.3–1.5 times faster than for $\alpha = 0.05$, 0.5.

Convergence rate was more strongly affected by the parameter γ initiating the restarts. First of all, because of the much slower convergence, exception was made for the τ-termination criterion in the experiments without restarts ($\gamma = 0$). For these cases the forecast skills displayed in Table 2 were achieved after 200 ha4dVar iterations. In general, in the interval $0.01 \leq \gamma_c \leq 0.1$ the first restart usually emerged after 5–8 iterations (searching over 80–128 directions). After that the restart frequency gradually increased, occurring every second iteration by the end of minimization process. For these values of γ_c, the optimization required between 10 and 20 restarts. It is worthwhile to note that we also tried an alternative restarting criterion based on the entropy estimation of the Hessian spectrum (e.g., Uzunoglu et al., 2007), but a simpler method (14) appeared to be more efficient.

Another extreme case ($\gamma_c = 0.25$) shown in the Table triggered a restart at almost every iteration, so that the overall descent process was done mostly within the subspace spanned by the leading eigenvectors of $\mathbf{B}$, requiring more than 600 precomputed modes. In terms of the forecast skill improvement, this type of descent appears to be much less effective than optimization with γ_c varying in the range between 0.01 and 0.1. As a consequence, there exists an optimal range of γ_c determining the number of restarts. In one extreme case (no restarts, $\gamma_c = 0$) the descent process stagnates due to eventual degeneration of the search directions defined by Eq. (11). If restarts are performed on every iteration (large γ_c), the descent loses efficiency, as it is performed along the smoothest eigenvectors of $\mathbf{B}$, which do not contain any information on the model-data misfits and model dynamics present in the ensembles generated by Eq. (11).

The benefit of the new ha4dVar restarting procedure can be seen by comparing the fourth column of Table 2 ($\alpha = 0.05$, $\gamma = 0.1$ – more frequent restarts), with Y16 a4dVar result (Table 1, first line) when restarts were effectively performed by alternating the state space metric in the EOF analysis of the ensemble. Performing restarts with the eigenvectors of the balanced background error covariance improves the forecast skills of all the state vector components with the overall improvement of F from 1.028 to 0.999. What is more important, the eigenvectors of the balanced background error covariance give restart directions a more efficient “kick”, since they tend to persist longer in time. Besides, smoothing of the model-data misfits with the balanced $\mathbf{B}$ effectively distributes observed information between all the state vector components in a manner consistent with geostrophic, hydrostatic and continuity constraints.

Additional assimilation experiments were performed to assess the impact of balance constraints on the convergence rate and the forecast skill. In terms of the forecast skill the best result with unbalanced BEC was obtained with $\alpha = 0.12,\ \gamma = 0.017$ (second line in Table 1). The convergence rate was noticeably slower than for the balanced case (black lines in Fig. 3), although the use of hybrid ensemble delivered somewhat faster reduction of the cost function than that of a4dVar (shown by red lines in Fig. 3). Overall, the impact of balance constraints appears to be beneficial to improving the forecast skill, although the average value of F in Table 2 remains larger than 1 by approximately 2% (excluding the cases with $\gamma = 0$). A few attempts to improve the skill were made by letting β in Eq. (16) to vary in space. Indeed, the values β derived from the background solution have shown a tendency to increase 1.5–2 times in the northern part of the sea. These experiments, however, did not have any positive impact on the forecast skill.

Since the ha4dVar forecast skill is close to that of 4dVar, it is instructive to quantify the ability of the ha4dVar ensemble $\{\delta\tilde{\mathbf{c}}^j\},\ j = 1, \ldots, mI$ to approximate the 4dVar increment $\delta\tilde{\mathbf{c}}_*$. To perform the comparison, the 4dvar increment was restricted to the range of $\mathbf{B}$: $\delta\tilde{\mathbf{c}}_* \leftarrow \mathbf{B}^{1/2}\mathbf{V}^{-1}\delta\tilde{\mathbf{c}}_*$ and then projected onto the set of basis vectors $\{\delta\tilde{\mathbf{c}}^j\}$ generated by the ha4dVar procedure. The resulting vector $\delta\tilde{\mathbf{c}}$ was found to differ by 13% from the ha4dVar increment $\delta\tilde{\mathbf{c}}_h$:

$$\left(\frac{\langle \delta\tilde{\mathbf{c}} - \delta\tilde{\mathbf{c}}_h, \delta\tilde{\mathbf{c}} - \delta\tilde{\mathbf{c}}_h, \rangle}{\langle \delta\tilde{\mathbf{c}}_h, \delta\tilde{\mathbf{c}}_h \rangle}\right)^{1/2} = 0.13 \qquad (18)$$

This value appears to be in a reasonable agreement with the earlier results of Yaremchuk and Martin (2014), and Yaremchuk et al. (2016a) who compared 4dVar/a4dVar increments and sensitivities in idealized linear and non-linear settings.

5. Summary and discussion

The purpose of this work is to describe and assess performance of an adjoint-free assimilation method based on the ensemble approach to generating search directions for minimizing the cost function. Ensemble-based variational optimization methods are rapidly gaining popularity in the meteorological community (e.g., Buehner et al., 2015) due to their relative simplicity and enhanced parallelization capabilities. In oceanography, where observations are less abundant, generation of skillful ensembles (search directions) is more problematic, especially in regional studies that often suffer from poor background error statistics and sparsity of observations.

A distinctive feature of the presented ha4dVar method is the absence of necessity to have neither an ensemble well approximating the background error statistics, nor the adjoint model for efficient gradient estimation. The presented ha4dVar method retrieves search directions from the joint statistics of model trajectory and model-data misfits at the current iteration and employs the leading BEC modes to restart the minimization process. In the present study, the BEC model is generalized to include balance constraints, which tend to guide the new ensemble members towards the directions on the slow manifold.

Performance of the ha4dVar method has been assessed against the traditional dual space 4dVar by comparing the forecast skill and computation time. The comparison has shown faster convergence (Fig. 3) and dummyTXdummy-(improved computational efficiency compared to 4dVar and the previous a4dVar version (Y16). The ha4dVar forecast skill appears to be compatible to 4dVar, although, on average, it seemed to be somewhat worse (Table 2) requiring some additional tuning of ha4dVar parameters whose impact on the efficiency of the algorithm was then studied in more detail. In particular, we have analyzed sensitivity of the ha4dVar scheme to changing the relative contribution of the model states in the ensemble, the effective weight of balance constraints in the background error covariance, and the frequency of the restarts. Numerical experiments have shown that the best forecast skill was achieved when a) contributions of the model states and model-data misfits to the ensemble measured in terms of the respective variances of the background fields were approximately equal; b) the error variance of the unbalanced velocity field was approximately equal to the kinetic energy of ageostrophic motions estimated from the background model trajectory; and c) restarts were performed when new search directions contributed less than 1.2% to the trace of the updated Hessian projection. The first two results indicate that a typical background state of a regional model may still provide reliable estimates of the integral statistical parameters such as time-averaged magnitudes of the oceanographic fields and the relative magnitude of ageostrophic motions in the area.

The analysis presented here is narrowly focused and certainly not comprehensive as it involves a single regional application. In particular, the space-time distribution of observations could bias the forecast skill estimate in favor of ha4dVar, which tends to better retrieve the velocity field from observations, which in this application were three times more abundant than *TS* data during the forecast period of August 19–28 (Fig. 4). More comprehensive testing with a larger variety of domains/datasets and using a comparably tuned 4dVar system are required for a more rigorous comparison. In particular, the 4dVar system should be constrained by the balanced background error covariance because our experiments indicate that constraining search directions by the balance relationships improves both the ha4dVar forecast skill and the convergence rate.

The major advantages of ha4dVar with respect to 4dVar are better parallelization efficiency and simplicity of implementation: the method treats a numerical model as a black box optimizing the background solution to observations by analyzing (the ensemble of) multiple model trajectories. In that respect ha4dVar bears similarity to the DART system (Anderson et al., 2009) which, nevertheless, relies on a user-defined ensemble. Since a4dVar relies on the direct estimation of the cost function derivatives, our major effort in the current research has been focused on elaborating strategies for selecting “best” directions to assess cost function sensitivity, i.e., the best directions of differentiation. In particular, we tried updating the ensembles using basis functions generated by coarse granulation of the initial fields (as in the Green’s function approach of Stammer and Wunsch (1996), and by the breeding technique of Toth and Kalnay (1993). Both methods appeared to be less efficient in minimizing the cost function, but for different reasons. The breeding technique had a tendency to generate search directions localized near the western and southern boundaries of the domain, often far away from the observational arrays, while the Green’s function method was lacking a sufficiently stable re-granulation strategy with iterations. It is noteworthy, however, that the ha4dvar method presented here can certainly be represented as a version of the Green’s function approach with a particular strategy of updating the perturbations. Although this strategy appears somewhat heuristic and dependent on “semi-random” choices of search directions in contrast to the actual the cost function gradients, its success may have connections to recent advances in randomization approaches for extra-large dimensional optimization problems which arise in machine learning (e.g., Bottou et al., 2017).

In the context of oceanographic applications, the ha4dVar algorithm should be considered as an attempt to develop another approach to a large variety of the existing 4d optimization methods based on the ensemble technique which basically employs finite differentiation to retrieve the gradient information. Although this “brute force” technique well fits the current parallelization trend in computer technologies, it still requires an efficient strategy in choosing the optimization subspace which in most applications relies on the accuracy of a given background error covariance. In that respect ha4dVar is self-sufficient because it retrieves ensemble members by blending a heuristic background error covariance model with the statistics of the dynamical model and model-data misfits gained during the search process. This feature makes ha4dVar suitable for oceanographic applications characterized by relatively sparse data and inaccurate background states. We believe that further development to a4dVar and similar self-sufficient/parallelization-friendly techniques has good prospects from both theoretical and numerical points of view.

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. P. Sakaulakis is greatfully acknowledged for help with parallelization issues.

Appendix A. Equivalence of the a4dVar and 4dVar solutions

For simplicity, assume that $\mathbf{c}^0 = 0$ and $\mathrm{rank}(\mathbf{B}) = \rho \le M$ with $\rho/m = L$ an integer. (In particular, we allow for the case that $\mathbf{B}$ is rank deficient.) If the sequence of search subspaces $\mathbf{P}^1, \mathbf{P}^2 \ldots$ are chosen according to criteria given in Section 2.4, then the iteration process (9) will terminate with a solution $\mathbf{c}_\star$ to (4) in no more than L steps, and this leads immediately to the solution that 4dVar provides for (1), $\mathbf{x}_\star^0 = \mathbf{B}^{1/2}\mathbf{c}_\star$.

To see this, note first that if $\mathbf{P}^i$ are generated in accord with conditions (a) and (b) in Section 2.4, then each search subspace is contained in the range of $\mathbf{B}$ and will have full rank m. Since the columns of $\mathbf{P}^i$ are $\hat{\boldsymbol{H}}$-orthogonal, one may introduce a composite matrix $\mathbf{P} = \left[\mathbf{P}^1, \mathbf{P}^2, \ldots, \mathbf{P}^L\right]$ and observe that its columns span the range of $\mathbf{B}$. Thus, at step L of the iteration process (9),

$$\mathbf{c}^L = \sum_{k=1}^{L} \mathbf{P}^k[\mathbf{P}^{k\mathsf{T}}\hat{\boldsymbol{H}}\mathbf{P}^k]^{-1}\mathbf{P}^{k\mathsf{T}}\mathbf{r} = \mathbf{P}[(\mathbf{AP})^\mathsf{T}\mathbf{AP}]^{-1}(\mathbf{AP})^\mathsf{T}\mathbf{d} \tag{A.1}$$

and so, $\mathbf{c}^L = \mathbf{P}\mathbf{a}_\star$ where $\mathbf{a}_\star$ solves

$$\|\mathbf{APa} - \mathbf{d}\| \longrightarrow \min_{\mathbf{a}} \tag{A.2}$$

However, $\mathbf{c}_\star$ that solves (8) must also lie in the range of $\mathbf{B}$ which by our discussion above coincides with the range of $\mathbf{P}$. But this means,

$$\min_{\mathbf{a}} \|\mathbf{APa} - \mathbf{d}\| = \min_{\mathbf{c}} \|\mathbf{Ac} - \mathbf{d}\|. \tag{A.3}$$

and so, $\mathbf{c}_\star = \mathbf{P}\mathbf{a}_\star = \mathbf{c}^L$ and the a4dVar solution, $\mathbf{B}^{1/2}\mathbf{c}^L$, coincides with the exact solution of the 4dVar problem, $\mathbf{x}_\star^0 = \mathbf{B}^{1/2}\mathbf{c}_\star$.

Appendix B. Balanced BEC model

Following the definition of the balance operator (e.g., Weaver et al., 2005), we partition the state vector $\mathbf{x} := \{T, S, \boldsymbol{u}, \zeta\}$ into two components $\mathbf{x}_1 = \{T, S\}$ and $\mathbf{x}_2 = \{\boldsymbol{u}, \zeta\}$, where $T, S, \boldsymbol{u}$ are the 3d fields of temperature, salinity and horizontal velocity, and ζ is the 2d sea surface height. We further split $\mathbf{x}_2$ into the balanced $\bar{\mathbf{x}}_2$ and unbalanced $\tilde{\mathbf{x}}_2$ components, assuming that $\bar{\mathbf{x}}_2$ (the balanced component) linearly depends on $\mathbf{x}_1$: $\bar{\mathbf{x}}_2 = \mathbf{L}\mathbf{x}_1$, where $\mathbf{L}$ is the finite-difference discretization of the following balance operator:

$$\rho = \rho_0 + \alpha(\boldsymbol{x}, z)T + \beta(\boldsymbol{x}, z)S, \tag{B.1}$$

$$\nabla h(\boldsymbol{x})\nabla\zeta = \mathrm{div}\int_{h(\boldsymbol{x})}^{0}\int_{z}^{0} \nabla\rho(\boldsymbol{x}, z')dz'dz, \tag{B.2}$$

$$\boldsymbol{u} = \frac{g}{f}\boldsymbol{k} \times \nabla\left[\zeta + \int_z^0 \frac{\rho(\boldsymbol{x}, z')}{\rho_0}dz'\right] \tag{B.3}$$

where ρ_0 is the background density of seawater, ρ stands for the deviations from the background associated with variations of the temperature T and salinity S fields, $h(\boldsymbol{x})$ is the bottom topography, g is acceleration due to gravity, f is the Coriolis parameter, and $\boldsymbol{k}$ is the vertical unit vector.

Dynamically, these equations constrain $\bar{\mathbf{x}}_2$ to be in hydrostatic and geostrophic balance (B.3), satisfy the vertically integrated continuity constraint (B.2), and the linearized equation of state of seawater (B.1).

The specified structure of the state vector $\mathbf{x}$, implies the following form of the BEC matrix

$$\mathbf{B} = \langle\mathbf{x}\mathbf{x}^\mathsf{T}\rangle = \begin{bmatrix} \mathbf{B}_1 & \mathbf{B}_1\mathbf{L}^\mathsf{T} \\ \mathbf{L}\mathbf{B}_1 & \mathbf{L}\mathbf{B}_1\mathbf{L}^\mathsf{T} + \mathbf{B}_2 \end{bmatrix} = \tag{B.4}$$
$$= \begin{bmatrix} \mathbf{I}_1 & \mathbf{0} \\ \mathbf{L} & \mathbf{I}_2 \end{bmatrix}\begin{bmatrix} \mathbf{B}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_2 \end{bmatrix}\begin{bmatrix} \mathbf{I}_1 & \mathbf{L}^\mathsf{T} \\ \mathbf{0} & \mathbf{I}_2 \end{bmatrix}$$

where $\mathbf{B}_{1,2}$ are the BEC matrices of $\mathbf{x}_1$ and $\tilde{\mathbf{x}}_2$ and $\mathbf{I}_{1,2}$ are the identity matrices of the respective size.

The background error covariances $\mathbf{B}_{1,2}$ are factorized using the Gaussian correlation model (12) with the decorrelation scale r_a of the unbalanced components $\mathbf{B}_2$ being set to 4.3 km. The respective diagonal values of $\mathbf{V}_2$ depend on spatial coordinates and were estimated from the statistics of the divergent component of the background solution (Yaremchuk and Martin, 2016b).

Relationships (B.4) and (12) allow to obtain explicit factorization $\mathbf{B} = \mathbf{B}^{1/2}\mathbf{B}^{\mathsf{T}/2}$ with

$$\mathbf{B}^{1/2} = \begin{bmatrix} \mathbf{V}_1 & \mathbf{0} \\ \mathbf{L}\mathbf{V}_1 & \mathbf{V}_2 \end{bmatrix}\begin{bmatrix} \mathbf{C}_1^{1/2} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_2^{1/2} \end{bmatrix} \tag{B.5}$$

where

$$\mathbf{C}^{1/2} = \frac{\delta x^2}{\pi r^2}\exp\left[\frac{r^2}{4}\Delta\right] \tag{B.6}$$

Factorization (B.4) provides a shortcut for computing the leading search directions $\mathbf{p}_m$ defined through the spectral decomposition of $\mathbf{B}$

$$\mathbf{B} = \sum_{m=1}^{M} \lambda_m^2\mathbf{p}_m\mathbf{p}_m^\mathsf{T} = (\mathbf{P}\boldsymbol{\Lambda})(\mathbf{P}\boldsymbol{\Lambda})^\mathsf{T}. \tag{B.7}$$

where λ_m^2 are the eigenvalues of the correlation matrix in the descending order, $\boldsymbol{\Lambda} = \mathrm{diag}\{\lambda_m\}$ and $\mathbf{P}$ is the matrix of search directions $\mathbf{p}_m$ listed columnwise. Substitution of (B.5) for $\mathbf{P}\boldsymbol{\Lambda}$ in the rhs of (B.7), shows that the leading search directions can be defined by

$$\mathbf{p}_m = \{\mathbf{V}_1\mathbf{e}_m^1; \quad \mathbf{L}\mathbf{V}_1\mathbf{e}_m^1 + \beta\mathbf{V}_2\mathbf{e}_m^2\}^\mathsf{T} \tag{B.8}$$

where $\mathbf{e}_m^{1,2}$ are the leading eigenvectors of $\mathbf{C}_{1,2}^{1/2}$ and β is a parameter controlling projection of a search direction on the unbalanced subspace. In the reported experiments the value of β was set to 0.2 and $\mathbf{e}^{1,2}$ were computed as eigenvectors corresponding to the smallest eigenvalues of the operators $\mathbf{I}_{1,2} + r_{1,2}^2/4\Delta_{1,2}$.

References

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, 12831296.

Barker, D.M., Huang, X.Y., Liu, Z., AulignÃ, T., Zhang, X., Rugg, S., Ajjaji, R., Bourgeois, A., Bray, J., Chen, Y., Demirtas, M., Guo, Y.R., Henderson, T., Huang, W., Lin, C., Michalakes, J., Rizvi, S., Zhang, X., 2012. The weather research and forecasting models community variational/ensemble data assimilation system: WRFDA. Bull. Am. Meteorol. Soc. 93, 831843. doi:10.1175/BAS-D-11-00167.1.

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), 18761894.

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, 347375. doi:10.1016/j.ocemod.2005.01.004.

Bonavita, M., Tremolet, Y., Holm, E., Lang, S., Chrust, M., Janiskova, M., Lopez, P., Laloyaux, P., Rosnay, P., Fisher, M., Harmud, M., English, S., 2017. A strategy for data assimilation. ECMWF Tech. Memo. 800, 42.

Bottou, L., Curtis, F.E., Nocedal, J., 2017. Optimization methods for large-scale machine learning. Mach. Learn. 108. (in press). Available at https://arxiv.org/pdf/1606.04838.pdf.

Buehner, M., Houtekamer, P.L., Charette, C., Mitchell, H.L., He, B., 2010. Intercomparison of variational data assimilation and ensemble kalman filter for global deterministic NWP. Mon. Weather Rev. 138, 1550–1586.

Buehner, M., McTaggart-Cowan, R., Beaulne, A., Charette, C., Garand, L., Heilliette, S., Lapalme, E., Laroche, S., Macpherson, S.P., Morneau, J., Zadra, A., 2015. Implementation of deterministic weather forecasting systems based on ensemble–variational data assimilation in environment canada. part i: the global system. Mon. Wea. Rev. 143, 2532–2559.

Buehner, M., Morneau, J., Charette, C., 2013. Four-dimensional ensemble-variational data assimilation for global deterministic weather prediciton. Nonlinear Processes Geophys. 20, 669–682. doi:10.5194/npg-20-669-2013.

Burrage, D.M., Book, J.W., Martin, P.J., 2009. Eddies and filaments of the Western Adriatic Current: Analysis and prediction. J. Mar. Sys. 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. Meteor. Soc. doi:10.1002/qj.2054.

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

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.

Descombes, G., Aulign, T., Vandenberghe, F., Barker, D.M., Barr, J., 2015. Generalized background error covariance matrix model. Geosci. Model. Dev. 8, 669–696. doi:10.5194/gmd-8-669-2015.

Desroziers, G., Berre, L., 2012. Accelerating and parallelizing minimizations in ensemble and deterministic variational assimilations. Q. J. R. Meteorol. Soc. 138, 1599–1610.

Desroziers, G., Camino, J.-T., Berre, L., 2014. 4DEnvar: link with 4d state formulation of variational assimilation and different possible implementations. Q. J. R. Meteorol. 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. Meteorol. Soc. 140, 281–294.

Gratton, S., Laloyaux, P., Sartenaer, A., 2014. Derivative-free optimization for large-scale nonlinear data assimilation problems. Q. J. R. Meteorol. Soc. 140, 164–179.

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, 11811199.

Isaksen, L., 2011. Data assimilation on future computer architectures. In: Proc. Seminar on Data Assimilation for Atmosphere and Ocean. ECMWF, Reading, UK, pp. 301–322.

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

Janjic, T., Nerger, L., Schr'oter, J., Skachko, S., 2011. On domain localization in ensemble-based kalman filter algorithms. Mon. Wea. Rev. 139, 2046–2060.

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. Wea. Rev. 141, 2740–2758.

Liu, C., Xiao, Q., Wang, B., 2008. An ensemble-based four-dimensional variational data assimilation scheme. part i: technical formulation and preliminary test. Mon. Wea. Rev. 136, 33633373.

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.Wea. Rev. 137, 1687–1704.

Liu, C., Xue, M., 2016. Relationships among four-dimensional hybrid ensemble-variational data assimilation algorithms with full and approximate ensemble covariance localization. Mon. Wea. Rev. 144, 591–606.

Lorenc, A., 2013. Recommended nomenclature for envar data assimilation methods. WHNE. Blue Book. [Available online at http://www.wcrp-climate.org/WGNE/BlueBook/2013/individual-articles/01_Lorenc_Andrew_EnVar_nomenclature.pdf].

Lorenc, A.C., Bowler, N.E., Clayton, A.M., Pring, S.R., 2015. Comparison of hybrid-4DEnvar and hybrid-4DVar data assimilation methods for global NWP. Mon. Wea. Rev. 143, 212–229.

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

Menemenlis, D., Wunsch, C., 1997. Linearization of an oceanic general circulation model for data assimilation and climate studies. J. Atmos. Oceanic Technol. 14, 14201443.

Ménétrier, B., Montmerle, T., Berre, L., Michel, Y., 2014. Estimation and diagnosis of heterogeneous flow-dependent background- error covariances at the convective scale using either large or small ensembles. Q. J. R. Meteorol. Soc. 140, 20502061. doi:10.1002/qj.2267.

Morey, S.L., Martin, P.J., O'Brien, 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 monterrey bay. Mon. Wea. Rev. 142, 2085–2107.

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.

Romine, G.S., Schwartz, C.S., Berner, J., Fossell, R.K., Snyder, C., Anderson, J., Weisman, M.L., 2014. Representing forecast error in a convection-permitting ensemble system. Mon. Weather Rev. 142, 45194541. doi:10.1175/MWR-D-14-00100.1.

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

Ruiz, E.N., Sandu, A., 2016. A derivative-free trust region framework for variational data assimilation. J. Comp. Appl. Math. 293, 164–179.

Stammer, D., Wunsch, C., 1996. The determination of the large-scale circulation of the pacific ocean from satellite altimetry using model greenâs functions. J. Geophys. Res. 101, 1840918432.

Toth, Z., Kalnay, E., 1993. Ensemble forecasting at NMC: the generation of perturbations. Bul. Am. Meteorol. Soc. 74, 2317–2330.

Uzunoglu, B., Fletcher, C.J., Zupanski, M., Navon, I.M., 2007. Adaptive ensemble reduction and infaltion. Q. J. R. Meteorol. Soc. 133, 1281–1294.

Weaver, A., Deltel, C., Machu, E., RIcci, S., Daget, N., 2005. A multivariate balance operator for variaional ocean data assimilation. Q. J. R. Meteorol. Soc. 131, 3605–3625.

Weaver, A.T., Courtier, P., 2001. Correlation modelling on a sphere using a generalized diffusion equation. Q. J. R. Meteorol. Soc. 127, 18151846.

Weaver, A.T., Tshimanga, J., Piacentini, A., 2015. Correlation operators based on an implicitly formulated diffusion equation solved with chebyshev iteration. Q. J. R. Meteorol. Soc. 142, 455–471. doi:10.1002/qj.2664.

Yaremchuk, M., Carrier, M., Smith, S., Jacobs, G., 2013. Background error correlation modeling with diffusion operators. In: Park, S.K., Xu, L. (Eds.), Data Assimilation for Atmospheric, Oceanic and Hydrological Applications, vol. 2. Springer, p. 177203.

Yaremchuk, M., Martin, P., 2014. On sensitivity analysis in the 4dvar framework. Mon. Wea. Rev. 142, 774787.

Yaremchuk, M., Martin, P., 2016b. Implementation of the Balance Operator in NCOM. NRL Report 7320-16-9649. Stennis Space Center, MS, USA 13pp.

Yaremchuk, M., Martin, P., Koch, A., Beattie, C., 2016a. Comparison of the adjoint and adjoint-free 4dvar assimilation of the hydrographic and velocity observations in the adriatic sea. Ocean Modell. 97, 129–140.

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

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. Wea. Rev. 140, 587600.

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