Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Marine Systems 40–41 (2003) 253–289
JOURNAL OF MARINE SYSTEMS
www.elsevier.com/locate/jmarsys

A demonstration of ensemble-based assimilation methods with a layered OGCM from the perspective of operational ocean forecasting systems

K. Brusdal[a,*], J.M. Brankart[b], G. Halberstadt[c], G. Evensen[a,d], P. Brasseur[b], P.J. van Leeuwen[c], E. Dombrowsky[e], J. Verron[b]

[a] *Nansen Environmental and Remote Sensing Center, Bergen 5059, Norway*
[b] *Université Joseph Fourier, Laboratoire des Ecoulements, Géo-physiques et Industriels, Grenoble, France*
[c] *Institute for Marine and Atmospheric Research, University of Utrecht, Utrecht, Netherlands*
[d] *Department of Mathematics, University in Bergen, Bergen, Norway*
[e] *Collecte Localisation Satellites, Toulouse, France*

Received 15 October 2001; accepted 9 August 2002

Abstract

A demonstration study of three advanced, sequential data assimilation methods, applied with the nonlinear Miami Isopycnic Coordinate Ocean Model (MICOM), has been performed within the European Commission-funded DIADEM project. The data assimilation techniques considered are the Ensemble Kalman Filter (EnKF), the Ensemble Kalman Smoother (EnKS) and the Singular Evolutive Extended Kalman (SEEK) Filter, which all in different ways resemble the original Kalman Filter.

In the EnKF and EnKS an ensemble of model states is integrated forward in time according to the model dynamics, and statistical moments needed at analysis time are calculated from the ensemble of model states. The EnKS, as opposed to the EnKF, update the analysis also backward in time whenever new observations are available, thereby improving the estimated states at the previous analysis times. The SEEK filter reduces the computational burden of the error propagation by representing the errors in a subspace which is initially calculated from a truncated EOF analysis.

A hindcast experiment, where sea-level anomaly and sea-surface temperature data are assimilated, has been conducted in the North Atlantic for the time period July until September 1996. In this paper, we describe the implementation of ensemble-based assimilation methods with a common theoretical framework, we present results from hindcast experiments achieved with the EnKF, EnKS and SEEK filter, and we discuss the relative merits of these methods from the perspective of operational marine monitoring and forecasting systems. We found that the three systems have similar performances, and they can be considered feasible technologically for building preoperational prototypes.

Keywords: North Atlantic; Assimilation methods; Operational ocean forecasting

1. Introduction

An ocean monitoring and prediction system must rely on integrated use of available remotely sensed

* Corresponding author.
E-mail address: kari.brusdal@hydro.com (K. Brusdal).

0924-7963/03/$ - see front matter © 2003 Elsevier Science B.V. All rights reserved.
doi:10.1016/S0924-7963(03)00021-6

and in situ measured observations together with dynamical models to achieve a best possible estimation of the true state of the ocean. Such integrated use of observations and model tools is best done using so-called data assimilation methods which provide a mean for optimal combination of the information about the real world contained in observations and the information about dynamical processes described by the models. The assimilation of observations allows numerical ocean models to simulate realistic meso-scale features of the ocean, i.e., the model evolution can be kept close to the true state of the ocean, and one avoids that the model drifts away from the observed state.

In the EC MAST-III-funded project, Development of Advanced Data Assimilation Systems for Operational Monitoring and Forecasting of the North Atlantic and Nordic Seas (DIADEM), three state-of-the-art advanced data assimilation techniques have, for the first time, been implemented with the Miami Isopycnic Coordinate Ocean Model (MICOM) to build a preoperational marine monitoring and forecasting system. The data assimilation methods considered are the Ensemble Kalman Filter (EnKF) by Evensen (1994), the Ensemble Kalman Smoother (EnKS) by Evensen and van Leeuwen (2000) and the Singular Evolutive Extended Kalman (SEEK) filter by Pham et al. (1998), which all are based on dynamically consistent estimates of the model error statistics.

The EnKF and EnKS schemes are based upon Monte Carlo forecasting or ensemble integration to compute the time evolution of error statistics. Monte Carlo methods avoid problems associated with the traditional Extended Kalman Filter (EKF), see, e.g., Evensen (1992), which neglects contributions from higher-order statistical moments when solving the error covariance equation. The Monte Carlo methods considered in this paper represent an alternative to integrating the error covariance equation (as in EKF) and are equivalent to solving the equation for the time evolution of the probability density function for the model error statistics. All the statistical information (mean state and its error covariance) needed at analysis time can then be computed from the ensemble of model states. The EnKF was first introduced in Evensen (1994), where it was applied with success in a twin experiment and intercompared with the EKF. In Evensen and van Leeuwen (1996), it was used in a realistic application for the Agulhas Current assimilating Geosat altimeter data into a quasi-geostropic model. The EnKS bears a strong resemblance with the EnKF; the difference is that whenever new observations are available during the forward integration, new analyses are calculated for all previous and the current time. Consequently, the first guess for the EnKS equals the EnKF solution, and subsequent smoother estimates are improvements of the first guess solution.

The SEEK filter reduces the computational burden of the error propagation by representing the model error in a subspace of small dimension. The initial error covariance is represented as a truncated series of orthogonal perturbations from an Empirical Orthogonal Function (EOF) analysis. The error subspace spanned by these EOFs evolves during the assimilation according to the model dynamics. The prediction error statistics are approximated from this error subspace, where only the dominant modes of variability are retained from the EOF analysis. As the number of dominant modes to be considered in the subspace is usually smaller than the size of the ensemble, the computational burden of the evolution error covariance with the SEEK can be reduced compared with the EnKF and EnKS. The SEEK filter has been applied in an academic test case in Pham et al. (1998). In Brasseur et al. (1999), they implemented an improved version of the SEEK filter which progressively learns the statistical structure of the estimation error from the residual innovation.

The purpose of this work is to validate the implementation of the three assimilation schemes in a hindcast experiment in the North Atlantic covering the time period July till September 1996. MICOM is a dynamic and thermodynamic isopycnic ocean general circulation model which has certain properties that make it ideal for use in a data assimilation system, e.g., it allows for updating the vertical density profile solely by moving layer interfaces up and down. In the data assimilation experiment presented in this paper, satellite observations of Sea-Level Anomalies (SLA) from radar altimeter data and Sea-surface Temperature (SST) from AVHRR data are assimilated into the ocean model. The capability of EnKF, EnKS and SEEK schemes to track the true evolution of the ocean will be examined, interpreted and, to some extent, intercompared.

In Section 2, the Ocean General Circulation Model, MICOM and the model setup is presented. A theoretical presentation of the three data assimilation techniques and implementation issues are given in Section 3. A discussion and comparison of the EnKF and the SEEK filter is given in Section 3.7. Section 4 describes the processing of the satellite observations used, and Section 5 presents the setup of the assimilation experiment. Finally, the numerical results of the hindcast experiment followed by a discussion and conclusions are given in Sections 6 and 7, respectively.

2. Model description

The Miami Isopycnic Coordinate Ocean Model (MICOM) was developed by Bleck and Boudra (1986), Bleck et al. (1989, 1992) and Smith et al. (1990) at the University of Miami. MICOM is characterized by the use of potential density as the vertical coordinate. This is possible since density is a monotonic function of depth. The motivation for using an isopycnic ocean model is that mixing along neutral surfaces, i.e., approximately surfaces of constant density, is orders of magnitude larger than the mixing across density surfaces. Hence, in MICOM, the ocean is divided into a number of constant density layers where the density increases with depth. The interaction between the different layers is mostly through hydrostatic pressure forces, but also includes explicitly prescribed diapycnal mixing and convection processes. The isopycnic nature of the model allows for high resolution in areas with large density gradients, and artificial mixing only occurs along density surfaces where it is negligible compared to the prescribed eddy mixing.

The upper layer is treated as a bulk mixed layer which allows for horizontal variations in the thermodynamic variables and density. It is based on an implementation of the Kraus–Turner mixed layer formulation (Bleck et al., 1989) and the formulation by Gaspar et al. (1990). This allows the model to use realistic heat and freshwater fluxes and to interact with a dynamic and thermodynamic ice model.

The model solves a number of equations consisting of the momentum equation for the velocity vector, one conservation equation for salt or heat, a continuity equation for mass conservation and, finally, the hydrostatic pressure equation which relates pressure to depth (see Bleck et al., 1992). Temperature is calculated by integrating a transport equation, and the salinity can then be diagnosed from the equation of state since the reference density of the layer is known. The model permits motion in all layers (contrary to reduced gravity models which suppress the barotropic mode), and in order to reduce the numerical cost of carrying the barotropic waves, a split-explicit scheme is used (Higdon and Bennett, 1996).

In MICOM, it is assumed that all model layers exist everywhere in the model domain. The thicknesses of the layers are entirely determined by the time evolving model equations, and special transport algorithms are required to maintain positive thickness for the model layers at all times. The internal layers (below the upper mixed layer) will outcrop to the bathymetry and to the upper mixed layer and they are, therefore, allowed to become massless with zero thickness. There are thermodynamic variables in all layers.

Vertical mixing processes include parameterizations for diapycnal mixing (transfer of salt and potential temperature between layers), convection (when the mixed layer water becomes denser than the isopycnal layers below) and entrainment/detrainment of mixed layer water (due to deepening/retreat of the mixed layer depth).

3. Data assimilation methods

The theoretical foundation for the three assimilation schemes, i.e., the EnKF, the EnKS and the SEEK filter, is now given in order to make it possible for the reader to understand the fundamental differences between them.

Given a vector of μ measurements, $\boldsymbol{d}\in\Re^{\mu}$ with an error covariance matrix $\boldsymbol{W}\in\Re^{\mu\times\mu}$, and a model state vector, $\boldsymbol{\psi}\in\Re^{n}$ with its error covariance matrix $\boldsymbol{P}^{\mathrm{f}}\in\Re^{n\times n}$. A linear variance minimizing analysis then becomes

$$\boldsymbol{\psi}^{\mathrm{a}} = \boldsymbol{\psi}^{\mathrm{f}} + \boldsymbol{K}(\boldsymbol{d} - \boldsymbol{H}\boldsymbol{\psi}^{\mathrm{f}}). \tag{1}$$

The superscripts, ‘a’ and ‘f’, respectively, denote analysis and forecast. The matrix $\boldsymbol{H}\in\Re^{\mu\times n}$ is the observation operator which ‘‘measures’’ the model

variables at the location of the observations, i.e., the "innovation vector", $\boldsymbol{d}-\boldsymbol{H}\boldsymbol{\psi}^{\mathrm{f}}$, computes the misfit between the observation vector and the model prediction. The matrix $\boldsymbol{K}\in\Re^{n\times\mu}$ is the Kalman gain given as

$$\boldsymbol{K}=\boldsymbol{P}^{\mathrm{f}}\boldsymbol{H}^{\mathrm{T}}(\boldsymbol{H}\boldsymbol{P}^{\mathrm{f}}\boldsymbol{H}^{\mathrm{T}}+\boldsymbol{W})^{-1}. \tag{2}$$

Finally, the error covariance matrix for the analysed estimate becomes

$$\boldsymbol{P}^{\mathrm{a}}=(\boldsymbol{I}-\boldsymbol{K}\boldsymbol{H})\boldsymbol{P}^{\mathrm{f}}. \tag{3}$$

As is well known, these equations are optimal, in the sense of variance minimizing, for linear models, and the optimal estimate is the maximum likelihood estimator for Gaussian distributed prior probability densities for the model and measurement errors. However, in our case, the model is nonlinear, so the optimality is not assured.

In between measurement times, the model state and the associated error covariance matrix is evolved in time according to the model dynamics (assuming a linear model),

$$\boldsymbol{\psi}_{k+1}=\boldsymbol{F}\boldsymbol{\psi}_k, \tag{4}$$

where $\boldsymbol{F}$ is the model operator. The error covariance equation is given by

$$\boldsymbol{P}_{k+1}=\boldsymbol{F}\boldsymbol{P}_k\boldsymbol{F}^{\mathrm{T}}+\boldsymbol{Q}, \tag{5}$$

where $\boldsymbol{Q}$ is the model error covariance matrix representing the uncertainties in the model formulation.

With a nonlinear model

$$\boldsymbol{\psi}_{k+1}=\boldsymbol{f}(\boldsymbol{\psi}_k), \tag{6}$$

the error covariance equation is written on the same form as above, but $\boldsymbol{F}$ is now the tangent linear operator or Jacobian of the nonlinear model operator, $\boldsymbol{f}$ evaluated at the current value of the model state vector. Thus, a linearization is applied leading to a linear equation for the evolution of the error covariance statistics. This is exactly the algorithm used in the so-called Extended Kalman Filter (EKF).

3.1. Ensemble Kalman Filter

The EnKF was designed to resolve two major problems related to the use of the EKF with nonlinear dynamics in large state spaces. The EKF applies a closure scheme where third- and higher-order moments in the error covariance equation are discarded. This linearization has been shown to be invalid in a number of applications, e.g., Evensen (1992) and Miller et al. (1994). In fact, the equation is no longer the fundamental equation for the error evolution when the dynamical model is nonlinear. In Evensen (1994), it was shown that a Monte Carlo method can be used to solve an equation for the time evolution of the probability density of the model state, as an alternative to using the approximate error covariance equation in the EKF.

For a nonlinear model where we appreciate that the model is not perfect and contains model errors, we can write it as a stochastic differential equation (on continuous form) as

$$\mathrm{d}\boldsymbol{\psi}=\boldsymbol{f}(\boldsymbol{\psi})\mathrm{d}t+\boldsymbol{g}(\boldsymbol{\psi},\mathrm{d}\boldsymbol{q})\mathrm{d}\boldsymbol{q}. \tag{7}$$

This equation states that an increment in time will yield an increment in $\boldsymbol{\psi}$, which, in addition, is influenced by a random contribution from the stochastic forcing term, $\boldsymbol{g}\mathrm{d}\boldsymbol{q}$, representing the model errors. The $\mathrm{d}\boldsymbol{q}$ describe a vector Brownian motion process with covariance $\boldsymbol{Q}\mathrm{d}t$. Because the model is nonlinear, $\boldsymbol{g}$ is not an explicit function of the random variable $\mathrm{d}\boldsymbol{q}$ so the Ito interpretation of the stochastic differential equation has to be used instead of the Statonovitz interpretation Jazwinski (1970).

When the model errors are additive, i.e., when $\boldsymbol{g}(\boldsymbol{\psi},\mathrm{d}\boldsymbol{q})\mathrm{d}\boldsymbol{q}=g(\boldsymbol{\psi})\mathrm{d}\boldsymbol{q}$, one can derive the Fokker–Planck or Kolmogorov's equation which describes the time evolution of the probability density $\phi(\boldsymbol{\psi})$ of the model state,

$$\frac{\partial\phi}{\partial t}+\sum_i\frac{\partial(f_i\phi)}{\partial\boldsymbol{\psi}_i}=\frac{1}{2}\sum_{i,j}\frac{\partial^2\phi(\boldsymbol{g}\boldsymbol{Q}\boldsymbol{g}^{\mathrm{T}})_{\boldsymbol{ij}}}{\partial\boldsymbol{\psi}_i\partial\boldsymbol{\psi}_j}, \tag{8}$$

where f_i is the component number i of the model operator $\boldsymbol{f}$ and $\boldsymbol{g}\boldsymbol{Q}\boldsymbol{g}^{\mathrm{T}}$ is the covariance matrix for the model errors.

This equation does not apply any important approximations and can be considered as the fundamental equation for the time evolution of error statistics. A detailed derivation is given in Jazwinski (1970). The equation describes the change of proba-

bility density in a local "volume" which is dependent on the divergence term describing a probability flux into the local "volume" (impact of the dynamical equation) and the diffusion term which tends to flatten the probability density due to the effect of stochastic model errors. If Eq. (8) could be solved for the probability density function, it should be possible to calculate statistical moments like the mean state and the error covariance for the model forecast to be used in the analysis scheme.

The EnKF applies a Markov Chain Monte Carlo (MCMC) method to solve Eq. (8). The probability density can be represented using a large ensemble of model states, and by integrating these model states forward in time according to the model dynamics described by the stochastic differential Eq. (7), this ensemble prediction is equivalent to solving the Fokker Planck equation using a MCMC method. This procedure forms the backbone for the EnKF.

The analysis in the EnKF is based on Eqs. (1)–(3), but now assumes that the covariance matrix $\boldsymbol{P}^{\mathrm{f}}$ can be represented by the forecast ensemble of model states. In the EnKF, the update is performed on each ensemble member separately, such that the new ensemble automatically has the correct spreading. This makes the scheme very efficient, avoiding the need for resampling algorithms to generate the new ensemble from the updated covariance each time measurement have been assimilated. Furthermore, the storage of the huge covariance matrix is not required. The implementation of the filter is discussed in more detail below.

By using these widely adapted Eqs. (1)–(3) for computing the analysis, the fundamental assumption is that the probability densities of the model prediction and the measurements are both close to a Gaussian. This is an implicit approximation when using the EKF since higher-order statistical moments are discarded and only the error covariance matrix is evolved in time. However, the EnKF applies a fully nonlinear evolution of error statistics and will predict non-Gaussian error statistics if the dynamical model is nonlinear. The nonlinear filter equations, or analysis equations, becomes more difficult to use in practical applications and are discussed in Evensen and van Leeuwen (2000). One preliminary conclusion is that if sufficient number of measurements with Gaussian distributed errors are available, then the model state will stay close to a Gaussian distribution, too. Unfortunately, no general theorems are available to quantify these statements, but in some papers this assumption has been tested by investigating the magnitude of the third-order moment, or even of higher-order moments (see, e.g., the works by van Leeuwen, 2001; Natvik and Evensen (2003a,b, this issue). The experience from EnKF applications with relatively low-resolution (eddy permitting) ocean models seems to be that the Gaussian assumption adopted in the analysis equations is not so bad after all. Clearly, this needs further investigation.

3.2. Ensemble Kalman smoother

The EnKS is a sequential algorithm which builds on the EnKF. It uses the same assumption as the EnKF about Gaussian statistics at analysis times. Previous work by van Leeuwen (2001) had shown that a nonsequential ensemble smoother (ES) can be formulated, but the assimilation time interval is restricted, at least for nonlinear models. van Leeuwen and Evensen (1996) had earlier found that the EnKF results were superior to the ES results due to the assumption of Gaussianity. In the ES, this assumption is used for the probability density over the whole integration interval, while the EnKF only makes it at analysis times. More importantly, however, is the fact that the prior estimate in the ES (the mean of an ensemble of model runs over the whole time interval), is rather poor and actually resembles the model climatology. An important result was found by Evensen and van Leeuwen (2000), where it was pointed out that any smoother solution can be obtained sequentially as long as measurement errors are uncorrelated in time. A new smoother was derived from basic principles which uses the EnKF solution as a first guess and computes further updates backward in time to obtain the smoother solution. Thus, the EnKS computes an additional contribution to the EnKF analysis from future measurements, which leads to a further improvement in the estimate. The additional smoother updates do not involve backward integrations as in many other smoother algorithms, with all their potential problems. Instead, the assumption of Gaussianity is made again, so that the smoother solution can again be written in terms of covariances using a slightly

modified form of the Kalman update Eqs. (1) and (2). The analysis equation now becomes

$$\psi^{a}(t-\tau) = \psi^{f}(t-\tau) + \boldsymbol{K}(t,\tau)(\boldsymbol{d} - \boldsymbol{H}\psi^{f}(t)), \qquad (9)$$

where t denotes the time for the current data set and τ is the time instant for the smoother update. The Kalman gain $\boldsymbol{K}(t,\tau)$ is now using the covariance of the model state between the times t and τ,

$$\boldsymbol{K}(t,\tau) = \boldsymbol{P}^{f}(t,\tau)\boldsymbol{H}^{T}(\boldsymbol{H}\boldsymbol{P}^{f}(t,t)\boldsymbol{H}^{T} + \boldsymbol{W})^{-1} \qquad (10)$$

Interestingly, no further model integrations have to be carried out to obtain the EnKS from the EnKF. Even the inversion of $\boldsymbol{H}\boldsymbol{P}^{f}\boldsymbol{H}^{T} + \boldsymbol{W}$, when computing the Kalman gain, becomes the same as in the EnKF. One only needs to store the relevant part of the EnKF ensemble at those times where the smoother solution is needed. Obviously, this is computationally a very efficient scheme. The implementation issues are further discussed below, being very close to those of the EnKF.

3.3. Singular evolutive Extended Kalman Filter

A critical issue for the application of Monte Carlo methods is the determination of a minimum number of ocean states needed to properly solve the forecast error equation and to estimate the characteristics of the forecast distribution. Due to the huge dimension of the numerical ocean state vector, the convergence of a Monte Carlo method may be slow, and the required computation may quickly exceed available resources.

A further simplification of the propagation of error statistics may be achieved using the concept of an error subspace. Assume that m random members are necessary to sample the probability distribution (or actually the error covariance matrix). An EOF decomposition of the ensemble may show that only a few dominant directions in the state space, $r \ll m$, are needed to explain the spreading of the ensemble (or actually the error covariance matrix).

The idea of the SEEK filter is to compute these r dominant directions to parameterize the error covariance at the initial time of an assimilation sequence and to solve the Kalman Filter equations in terms of that subspace afterward (Pham et al., 1998). Depending on the norm of the error modes and the importance of nonlinear effects, the propagation of the forecast error can be achieved using the nonlinear model (in a similar way as in the EnKF) or using a tangent linear approximation of it (Verron et al., 1999).

While the EnKF computes the analysis in a space spanned by the m random ensemble members, the SEEK computes the analysis in the space spanned by the r dominant EOFs. Further, while the EnKF updates the error statistics by performing an analysis for each ensemble member, the SEEK filter updates the statistics of the error covariance in the reduced space directly.

In order to compensate for the EOF truncation and more generally for the limitations inherent to the error subspace parameterization, an adaptive mechanism is included in the SEEK filter which performs a consistency check between the error statistics predicted by the filter and the information contained in the innovation vector (Brasseur et al., 1999). In the present application, this mechanism operates through the self-tuning of the model error in order to achieve a balance between the expected and the observed innovation variance (Brankart et al., 2003).

3.4. Practical implementation of the EnKF

The EnKF and EnKS analysis schemes have been implemented in a similar manner, the difference is that the EnKS implementation has been extended such that the observations also update model states backward in time. Therefore, the following description of the analysis implementation is valid for both the EnKF and EnKS.

The model state is denoted by $\psi \in \Re^{n}$, and the ensemble of forecasted model states are stored in a matrix $\boldsymbol{A} \in \Re^{n \times m}$, where n is the number of elements in the model state and m is the number of members in the ensemble.

An ensemble approximation of the forecast error covariance matrix is given by

$$\boldsymbol{P}_{e}^{f} = \frac{1}{m-1}(\boldsymbol{A}^{f} - \overline{\boldsymbol{A}}^{f})(\boldsymbol{A}^{f} - \overline{\boldsymbol{A}}^{f})^{T}, \qquad (11)$$

where each column of the matrix $\overline{\boldsymbol{A}}^{f}$ contains the mean of the ensemble, $\overline{\psi}^{f} = (1/m)\sum_{j=1}^{m}\psi_{j}^{f}$. The ensemble mean is considered to be the best guess estimate (it is the variance-minimizing estimate), and the spreading of the ensemble around the mean gives the error variance in the ensemble.

In order to obtain a variance minimizing analysis scheme, we also have to create an ensemble of observation vectors, $\boldsymbol{d}_j \in \Re^{\mu}$, which is generated by adding vectors of observation noise, $\boldsymbol{\epsilon}_j \in \Re^{\mu}$, to the observations, $\boldsymbol{d}$, i.e.,

$$\boldsymbol{d}_j = \boldsymbol{d} + \boldsymbol{\epsilon}_j, \quad j = 1, \ldots, m, \tag{12}$$

where each $\boldsymbol{\epsilon}_j$ is picked randomly from a Gaussian distribution with zero mean and standard deviation determined by error covariance matrix for the measurements, $\boldsymbol{W}$. The ensemble of measurements can be stored in the columns of a matrix

$$\boldsymbol{D} = (\boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_m) \in \Re^{\mu \times m}. \tag{13}$$

In the analysis scheme, each ensemble member, ψ_j, $j = 1, \ldots, m$ is updated according to

$$\psi_j^{\mathrm{a}} = \psi_j^{\mathrm{f}} + \boldsymbol{K}_e(\boldsymbol{d}_j - \boldsymbol{H}\psi_j^{\mathrm{f}}), \tag{14}$$

or written in matrix form,

$$\boldsymbol{A}^{\mathrm{a}} = \boldsymbol{A}^{\mathrm{f}} + \boldsymbol{K}_e(\boldsymbol{D} - \boldsymbol{H}\boldsymbol{A}^{\mathrm{f}}), \tag{15}$$

where the Kalman gain computed from the ensemble is

$$\boldsymbol{K}_e = \boldsymbol{P}_e^{\mathrm{f}}\boldsymbol{H}^{\mathrm{T}}(\boldsymbol{H}\boldsymbol{P}_e^{\mathrm{f}}\boldsymbol{H}^{\mathrm{T}} + \boldsymbol{W})^{-1}. \tag{16}$$

Define now the "measurements" of the ensemble perturbations

$$\boldsymbol{S}_e = \boldsymbol{H}(\boldsymbol{A}^{\mathrm{f}} - \overline{\boldsymbol{A}}^{\mathrm{f}}), \tag{17}$$

and

$$\boldsymbol{C}_e = (\boldsymbol{H}\boldsymbol{P}_e^{\mathrm{f}}\boldsymbol{H}^{T} + \boldsymbol{W}) = \frac{\boldsymbol{S}_e\boldsymbol{S}_e^{\mathrm{T}}}{m-1} + \boldsymbol{W} = \boldsymbol{U}_e\boldsymbol{\Lambda}_e\boldsymbol{U}_e^{\mathrm{T}}. \tag{18}$$

where $\boldsymbol{U}_e$ contains the eigenvectors of $\boldsymbol{C}_e$ and $\boldsymbol{\Lambda}_e$ the corresponding eigenvalues. The $\boldsymbol{C}_e^{-1}$ matrix is computed from an eigenvalue decomposition of $\boldsymbol{C}_e$, since this makes it possible to handle the situation where $\boldsymbol{C}$ is singular (this may occur if dependent measurements are used and if the number of measurements is larger than the number of ensemble members). In order to avoid this problem, the spectrum of the eigenvalues is truncated to only retain significant nonzero eigenvalues and the pseudo inverse of $\boldsymbol{C}_e$ is approximated by $\boldsymbol{C}_e^{-1} = \boldsymbol{U}_e\boldsymbol{\Lambda}_e^{-1}\boldsymbol{U}_e^{\mathrm{T}}$. The analysis Eq. (15) now becomes

$$\boldsymbol{A}^{\mathrm{a}} = \boldsymbol{A}^{\mathrm{f}} + \frac{1}{m-1}(\boldsymbol{A}^{\mathrm{f}} - \overline{\boldsymbol{A}}^{\mathrm{f}})\boldsymbol{S}_e^{\mathrm{T}}\boldsymbol{U}_e\boldsymbol{\Lambda}_e^{-1}\boldsymbol{U}_e^{\mathrm{T}}(\boldsymbol{D} - \boldsymbol{H}\boldsymbol{A}^{\mathrm{f}}). \tag{19}$$

Following the analysis step, each individual member of the ensemble is integrated forward in time until the next time observations are available using the stochastic Eq. (7). The prediction of error statistics would be exact in the case of an infinite ensemble size. Simulation of model errors is included in a realistic way (if the model error statistics is actually known). Thus, the major approximations used in the EnKF are related to the use of a finite ensemble size and the use of a Gaussian assumption on the predicted error statistics at analysis time.

3.5. Practical implementation of the EnKS

The EnKS analysis can be computed using the same formula (Eq. (19)) for different prior times τ just by writing it as

$$\boldsymbol{A}^{\mathrm{a}}(t-\tau) = \boldsymbol{A}^{\mathrm{f}}(t-\tau) + \frac{1}{m-1}(\boldsymbol{A}^{\mathrm{f}}(t-\tau) - \overline{\boldsymbol{A}}^{\mathrm{f}}(t-\tau)) \times \boldsymbol{S}_e^{\mathrm{T}}\boldsymbol{U}_e\boldsymbol{\Lambda}_e^{-1}\boldsymbol{U}_e^{\mathrm{T}}(\boldsymbol{D} - \boldsymbol{H}\boldsymbol{A}^{\mathrm{f}}). \tag{20}$$

Thus, it is not necessary to recompute the coefficients already used in the EnKF analysis. The analysis error at the prior times is diagnosed from the updated ensemble as

$$\boldsymbol{P}_e^{\mathrm{a}}(t-\tau) = \frac{1}{m-1}(\boldsymbol{A}^{\mathrm{a}}(t-\tau) - \overline{\boldsymbol{A}}^{\mathrm{a}}(t-\tau)) \times (\boldsymbol{A}^{\mathrm{a}}(t-\tau) - \overline{\boldsymbol{A}}^{\mathrm{a}}(t-\tau))^{\mathrm{T}}. \tag{21}$$

3.6. Practical implementation of the SEEK filter

The general equations of the SEEK analysis are quite similar to their EnKF and EnKS counterparts, although several differences exist in the numerical algorithm. The major difference is linked to the fact that only one oceanic state is corrected with the data, while the error statistics are modified in the reduced space.

In practice, the conventional procedure to evaluate the initial error covariance assumes that (i) the covariance of the oceanic variability can be used as a proxy of the initial error covariance; (ii) the model variability is identical to the real ocean variability; (iii) a sample of model snapshots adequately represents the model variability; and (iv) the EOF analysis of the sample is dominated by r significant modes.

With r being the dimension of the error subspace, the forecast error covariance matrix of the SEEK filter can be computed as

$$\boldsymbol{P}_s^{\mathrm{f}} = \frac{1}{r-1}\boldsymbol{N}^{\mathrm{f}}(\boldsymbol{N}^{\mathrm{f}})^{\mathrm{T}}, \tag{22}$$

where $\boldsymbol{N}^{\mathrm{f}} \in \Re^{n \times r}$ initially contains the r dominant vectors of the EOF decomposition. Thus, $\boldsymbol{N}^{\mathrm{f}}$ takes the role of the $\boldsymbol{A}^{\mathrm{f}} - \overline{\boldsymbol{A}}^{\mathrm{f}}$ in the EnKF. At the analysis step, the ocean state vector is updated according to the same equation as Eq. (14) in the EnKF, i.e.,

$$\boldsymbol{\psi}^{\mathrm{a}} = \boldsymbol{\psi}^{\mathrm{f}} + \boldsymbol{K}_s(\boldsymbol{d} - \boldsymbol{H}\boldsymbol{\psi}^{\mathrm{f}}), \tag{23}$$

where

$$\boldsymbol{K}_s = \boldsymbol{P}_s^{\mathrm{f}}\boldsymbol{H}^{\mathrm{T}}(\boldsymbol{H}\boldsymbol{P}_s^{\mathrm{f}}\boldsymbol{H}^{\mathrm{T}} + \boldsymbol{W})^{-1} \tag{24}$$

is the Kalman gain as computed in the SEEK filter. Clearly, the basic analysis equations in the EnKF and the SEEK are similar. The major difference is that while the SEEK filter uses the error subspace $\boldsymbol{N}^{\mathrm{f}}$ to represent the prediction error statistics, the EnKF uses the ensemble perturbations, $\boldsymbol{A}^{\mathrm{f}} - \overline{\boldsymbol{A}}^{\mathrm{f}}$ (see discussion below).

Defining the ''measurements'' of the elements in the error subspace

$$\boldsymbol{S}_s = \boldsymbol{H}\boldsymbol{N}^{\mathrm{f}}, \tag{25}$$

and when using some algebraic transformations, the Kalman gain from the SEEK can be rewritten as

$$\boldsymbol{K}_s = \boldsymbol{N}^{\mathrm{f}}((r-1)\boldsymbol{I} + \boldsymbol{S}_s^{\mathrm{T}}\boldsymbol{W}^{-1}\boldsymbol{S}_s)^{-1}\boldsymbol{S}_s^{\mathrm{T}}\boldsymbol{W}^{-1} \tag{26}$$

This form for the Kalman gain is particularly useful since it allows for a simple computation of the analysed error subspace $\boldsymbol{N}^{\mathrm{a}}$.The expression requires the inverse of an $r \times r$ matrix but rather than computing an explicit inversion, we proceed with an eigenvalue decomposition of the symmetric positive definite matrix

$$\frac{1}{r-1}\boldsymbol{S}_s^{\mathrm{T}}\boldsymbol{W}^{-1}\boldsymbol{S}_s = \boldsymbol{U}_s\boldsymbol{\Lambda}_s\boldsymbol{U}_s^{\mathrm{T}} \tag{27}$$

This allows for a reformulation of the Kalman gain, as

$$\boldsymbol{K}_s = \frac{1}{r-1}\boldsymbol{N}^{\mathrm{f}}\boldsymbol{U}_s(\boldsymbol{I} + \boldsymbol{\Lambda})^{-1}\boldsymbol{U}_s^{\mathrm{T}}\boldsymbol{S}_s^{\mathrm{T}}\boldsymbol{W}^{-1}, \tag{28}$$

and, finally, the analysis update becomes

$$\boldsymbol{\psi}^{\mathrm{a}} = \boldsymbol{\psi}^{\mathrm{f}} + \frac{1}{r-1}\boldsymbol{N}^{\mathrm{f}}\boldsymbol{U}_s(\boldsymbol{I} + \boldsymbol{\Lambda})^{-1}\boldsymbol{U}_s^{\mathrm{T}}\boldsymbol{S}_s^{\mathrm{T}}\boldsymbol{W}^{-1}(\boldsymbol{d} - \boldsymbol{H}\boldsymbol{\psi}^{\mathrm{f}}). \tag{29}$$

The error covariance of the updated state is given as usual in terms of the Kalman gain

$$\boldsymbol{P}_s^{\mathrm{a}} = (\boldsymbol{I} - \boldsymbol{K}_s\boldsymbol{H})\boldsymbol{P}_s^{\mathrm{f}}, \tag{30}$$

or, using Eq. (28),

$$\boldsymbol{P}_s^{\mathrm{a}} = \frac{1}{r-1}\boldsymbol{N}^{\mathrm{f}}\boldsymbol{U}_s(\boldsymbol{I} + \boldsymbol{\Lambda})^{-1}\boldsymbol{U}_s^{\mathrm{T}}(\boldsymbol{N}^{\mathrm{f}})^{\mathrm{T}} = \frac{1}{r-1}\boldsymbol{N}^{\mathrm{a}}(\boldsymbol{N}^{\mathrm{a}})^{\mathrm{T}}, \tag{31}$$

where we define

$$\boldsymbol{N}^{\mathrm{a}} = \boldsymbol{N}^{\mathrm{f}}\boldsymbol{U}_s(\boldsymbol{I} + \boldsymbol{\Lambda})^{-1/2}\boldsymbol{U}_s^{\mathrm{T}}. \tag{32}$$

This last equation describes the transformation of the error subspace associated to the analysis step, and is the counterpart of Eq. (19) in the EnKF.

The subsequent model forecast is then achieved with the best estimate (Eq. (29)) as initial conditions:

$$\boldsymbol{\psi}^{\mathrm{f}} = \boldsymbol{f}(\boldsymbol{\psi}^{\mathrm{a}}), \tag{33}$$

while the dynamical propagation of the associated error covariance is computed as an ensemble integration of the error subspace components:

$$\tilde{\boldsymbol{N}}_j^{\mathrm{f}} = \frac{1}{\alpha}(\boldsymbol{f}(\boldsymbol{\psi}^{\mathrm{a}} + \alpha\boldsymbol{N}_j^{\mathrm{a}}) - \boldsymbol{f}(\boldsymbol{\psi}^{\mathrm{a}})). \tag{34}$$

α is a scalar parameter which determines the size of the error modes to be considered for the nonlinear model integrations. This parameter is typically of order 1, but it can be adjusted according to the actual spread of the perturbations around the central state. If the model is perfectly linear, the forecast error modes will not be affected by the numerical value specified for α.

As in Pham et al. (1998), the model error is taken into account by means of a ''forgetting'' factor $\rho \in [0,1]$, leading to:

$$\boldsymbol{N}_j^{\mathrm{f}} = \frac{1}{\rho}\tilde{\boldsymbol{N}}_j^{\mathrm{f}}. \tag{35}$$

where a value of $\rho < 1$ leads to an amplification of the prediction errors.

Because a model is characterized by several dynamical regimes and the model error probably is inhomogeneous in space, the ρ coefficient is determined as a function of the spatial coordinates to get an approximate balance between the expected and the observed innovation variance Brankart et al. (2003). In practice, we estimate the variance of the innovation vector from the full sequence of its realizations up to the current time giving a larger weight to the most recent events (an exponential decay of about 2 months was used for the experiments presented below).

3.7. Discussion

Some particular issues related to both the EnKF and the SEEK are discussed below.

3.7.1. Subsampling of data

The estimation of the small correlations associated with remote observations is a well-known difficulty of ensemble methods (Houtekamer and Mitchell, 1998). In both the EnKF and the SEEK, a local parameterization is used in the forecast error covariance matrix, enforcing to zero the correlation coefficients between distant state variables. In practice, this is implemented by assuming that distant observations have negligible influence on the analysis. The global system is split into small subsystems where the traditional analysis is computed for each of these. Here, it was chosen to update variables grid-point by grid-point in the horizontal dimensions using a radius of influence parameter in order to limit the number of observations to be used when updating each grid-point. Only data points located within a circle with a specified radius of influence, centered at the particular model grid-point to be updated, will contribute in the update. This is an approximation but it makes sense since only data points located in the ''neighborhood'' of a model grid-point should have a significant impact on the analysis for that grid-point. Further, this algorithm leads to a reduction in the numerical cost (since the inversion of the full $\boldsymbol{W}$ is replaced with many inversions of small subparts of $\boldsymbol{W}$, one for each model grid-point). Further, we have observed that this also improves the analysis since the size of the ensemble in the EnKF, or the error subspace in the SEEK, relative to the number of state variables at a particular grid-point is of the same order and, therefore, span a larger part of the model state space.

3.7.2. Consistency check

In addition to the purely statistical procedure used in the analysis schemes, there are several dynamical constraints that must be satisfied following an analysis step to ensure that one proceeds with a dynamically acceptable model state. Of particular importance in isopycnal models is that the isopycnic layers always must have a positive thickness. Since the Kalman analysis is written in the context of linear estimation theory, it is not easily capable of taking such a constraint into account. Thus, following an analysis step eventually negative layer thicknesses must be reset to zero. This has been observed to happen occasionally, often in connections with already thin layers in the forecast, when strong updates are computed based on large differences between a model forecast and accurate (and highly weighted) observations. A number of additional adjustments are also implemented in order to respect common-sense criteria (the density of the mixed layer must be lower than the density of the layer underneath; the temperature of sea water must always be higher than the freezing point, the salinity must be positive) or to keep the system within a reasonable range (temperature lower than 32 °C and salinity between 10 and 39 psu). These checks were applied on the best guess estimate in the SEEK. In the EnKF, only a consistency check on the layer thicknesses was performed after analysis. Still, no problems with non-physical temperature and salinity values occurred during the assimilation experiment.

3.7.3. Comparison of EnKF and SEEK

Finally, we attempt to provide an explanation of the differences and similarities between the EnKF and the SEEK. The previous discussion has shown that both methods start out with the same basic assumptions for a linear unbiased variance minimizing analysis scheme.

3.7.3.1. Representation of error statistics. A major difference resides in the way error statistics is represented. The EnKF applies a random ensemble of model states which samples the model probability density function. The larger number of model states contained in the ensemble, the better it will represent the probability density function. Thus, a very accurate representation can be obtained by using a large enough ensemble. The SEEK filter computes an orthogonal error subspace which should contain the dominant variability of the system. This subspace could *initially* be determined by using the dominant singular vectors from a singular value decomposition (SVD) of an ensemble of model state perturbations, $\boldsymbol{A}-\overline{\boldsymbol{A}}$, or alternatively using the dominant eigenvectors of an error covariance matrix. In practical applications it is computed from an SVD of a number of model states anomalies subsampled from a representative model simulation. Using the orthogonal error subspace, it is possible to represent the error statistics using a smaller number of EOFs than the number of ensemble members normally needed in the EnKF.

3.7.3.2. One model state vs. ensemble mean as best estimate. The EnKF computes an analysis for each model state in the ensemble. The only reason for doing this is that one then gets a new ensemble which has the correct error covariance statistics for the analysis. Thus, there is no need for an additional resampling to create a new ensemble for the further integration, as is seen in some of the more sophisticated nonlinear filters (Anderson and Anderson, 1999; Evensen and van Leeuwen, 2000; Pham, 2001). Note also that in the EnKF, it is possible to compute the analysis for the ensemble mean directly from

$$\overline{\boldsymbol{\psi}}^{\mathrm{a}} = \overline{\boldsymbol{\psi}}^{\mathrm{f}} + \frac{1}{m-1}(\boldsymbol{A}^{\mathrm{f}} - \overline{\boldsymbol{A}}^{\mathrm{f}})\boldsymbol{S}_e^{\mathrm{T}}\boldsymbol{U}_e\boldsymbol{\Lambda}_e^{-1}\boldsymbol{U}_e^{\mathrm{T}}(\boldsymbol{d} - \boldsymbol{H}\overline{\boldsymbol{\psi}}^{\mathrm{f}}). \tag{36}$$

This equation should be compared with the SEEK analysis (Eq. (23)), which updates a single-model state used as the best estimate (see below).

3.7.3.3. Specification of model errors. Another difference resides in the specification of the model errors. In the EnKF, this error is taken into account by means of stochastic forcing during the ensemble forecast. This allows for the use of realistic random model errors spanning the whole n-dimensional model space. Thus, if the actual model errors are known or can be estimated, it is possible to introduce a realistic simulation of them.

In the SEEK filter, the so-called forgetting factor, ρ, is used. This is a number which is multiplied with the EOFs in the error subspace to increase the error variance. Note that the error subspace is changing with the forgetting factor because ρ is space-dependent. This is an approximate way of introducing the effect of model errors, and it could probably be done more consistently if needed. On the other hand, model error statistics is normally poorly known, justifying the use of a simplistic but efficient modeling of them. This approach also has the merit to allow the use of a simple adaptive mechanism to tune the forecast error according to the statistics of the innovation sequence and in this way to enforce internal consistency of error statistics Brankart et al. (2003).

3.7.3.4. Prediction of error statistics. It is of interest to compare the procedure used for the prediction of error statistics in the EnKF and the SEEK. First, note that the error components used in the SEEK could be defined as $\boldsymbol{N}_j = \boldsymbol{U}_j\sigma_j$, where $\boldsymbol{U}_j$ is the singular vector j and σ_j the corresponding singular value of an SVD of the ensemble perturbations $\boldsymbol{A}^{\mathrm{f}} - \overline{\boldsymbol{A}}^{\mathrm{f}}$. On matrix form, this would just be $\boldsymbol{N} = \boldsymbol{U}\boldsymbol{\Sigma}$. In the EnKF, the time evolution of error statistics is modeled by integration of each individual ensemble member forward in time according to the stochastic form of the model equations. On matrix form, this process can be written as

$$\begin{aligned}\boldsymbol{A}^{\mathrm{f}} &= \boldsymbol{f}(\boldsymbol{A}^{\mathrm{a}}) + \boldsymbol{q} = \boldsymbol{f}(\overline{\boldsymbol{A}}^{\mathrm{a}} + (\boldsymbol{A}^{\mathrm{a}} - \overline{\boldsymbol{A}}^{\mathrm{a}})) + \boldsymbol{q} \\ &= \boldsymbol{f}(\overline{\boldsymbol{A}}^{\mathrm{a}} + \boldsymbol{U}^{\mathrm{a}}\boldsymbol{\Sigma}^{\mathrm{a}}(\boldsymbol{V}^{\mathrm{a}})^{\mathrm{T}}) + \boldsymbol{q}.\end{aligned} \tag{37}$$

Thus,

$$\begin{aligned}\boldsymbol{A}^{\mathrm{f}} - \overline{\boldsymbol{A}}^{\mathrm{f}} &= \boldsymbol{f}(\overline{\boldsymbol{A}}^{\mathrm{a}} + \boldsymbol{U}^{\mathrm{a}}\boldsymbol{\Sigma}^{\mathrm{a}}(\boldsymbol{V}^{\mathrm{a}})^{\mathrm{T}}) + \boldsymbol{q} \\ &\quad - \overline{\boldsymbol{f}(\overline{\boldsymbol{A}}^{\mathrm{a}} + \boldsymbol{U}^{\mathrm{a}}\boldsymbol{\Sigma}^{\mathrm{a}}(\boldsymbol{V}^{\mathrm{a}})^{\mathrm{T}})}.\end{aligned} \tag{38}$$

The EnKF evolves an ensemble where each individual perturbation (ensemble member) consists of all directions in $\boldsymbol{U}$ space using the linear combination defined in $\boldsymbol{V}^{\mathrm{T}}$. Following an ensemble integration, a

new SVD could be formed which would define a set of new orthogonal directions and with a new set of singular values holding information about the error variance change obtained during the integration. Note that only $\boldsymbol{U}$ and $\boldsymbol{\Sigma}$ are needed to compute $\boldsymbol{P}_e$.

Eq. (38) can be compared with the prediction Eq. (34) used for the error components in the SEEK which evolves an ensemble where member j contains a perturbation proportional to and along an independent direction $\boldsymbol{N}_j$. The EnKF integrates ensemble members which all may hold contributions from all directions in $\boldsymbol{U}$. The equations for the EnKF and the SEEK also differ in the use of the ensemble mean and the "central forecast" as the best estimate. Following an integration of the r error subspace components $\boldsymbol{N}_j$, an orthogonalization procedure is used to produce a new set of independent error components.

For a linear model, the ensemble integration Eq. (34) for the EnKF becomes

$$\boldsymbol{A}^{\mathrm{f}} - \overline{\boldsymbol{A}}^{\mathrm{f}} = \boldsymbol{F}\boldsymbol{U}^{\mathrm{a}}\boldsymbol{\Sigma}^{\mathrm{a}}(\boldsymbol{V}^{\mathrm{a}})^{\mathrm{T}} + \boldsymbol{q}, \tag{39}$$

where we have used that $\overline{\boldsymbol{U}^{\mathrm{a}}\boldsymbol{\Sigma}^{\mathrm{a}}(\boldsymbol{V}^{\mathrm{a}})^{\mathrm{T}}} \equiv 0$, while the error prediction Eq. (34) used in the SEEK can be written on matrix form as

$$\tilde{\boldsymbol{N}}^{\mathrm{f}} = \boldsymbol{F}\boldsymbol{N}^{\mathrm{a}} = \boldsymbol{F}\boldsymbol{U}^{\mathrm{a}}\boldsymbol{\Sigma}^{\mathrm{a}}(\boldsymbol{I}^{\mathrm{a}})^{\mathrm{T}}, \tag{40}$$

where $\boldsymbol{I}$ is just the identity matrix. When neglecting the model error implementation in the EnKF and the SEEK (i.e., $\rho = 1$ in Eq. (35)), the time evolution in the SEEK and the EnKF will give identical results (for linear model dynamics), when the same number of orthogonal directions are used. Note that for a linear model, the ensemble mean, $\overline{\psi}$, becomes identical to the central forecast used in the SEEK. Eqs. (39) and (40) show that the only difference is the choice of ensemble used. In the EnKF, any unitary matrix (all columns orthonormal to each other) can be used for $\boldsymbol{V}$, where a different $\boldsymbol{V}$ just represents a different ensemble of model states in the same error subspace. The identity matrix is one such matrix, of course.

For a nonlinear model, the time evolution will clearly be different for the two methods since there are nonlinear interactions between the error modes. However, the same will be true between two different $\boldsymbol{V}$'s in the EnKF. However, both approaches are funded on sound mathematical theory.

3.7.3.5. The analysis equations. The transformations used for the Kalman gain in the SEEK (see Eqs. (26)–(28)) are only needed for the computation of the analysed error components $\boldsymbol{N}_j^{\mathrm{a}}$, and the SEEK analysis (Eq. (29)) can also be solved on the form,

$$\psi^{\mathrm{a}} = \psi^{\mathrm{f}} + \frac{1}{r-1}\boldsymbol{N}^{\mathrm{f}}\boldsymbol{S}_s^{\mathrm{T}}\boldsymbol{U}'_s\boldsymbol{\Lambda}'^{-1}_s\boldsymbol{U}'^{\mathrm{T}}_s(\boldsymbol{d} - \boldsymbol{H}\psi^{\mathrm{f}}). \tag{41}$$

where the following decomposition is used:

$$\boldsymbol{C}_s = (\boldsymbol{H}\boldsymbol{P}_s^{\mathrm{f}}\boldsymbol{H}^{\mathrm{T}} + \boldsymbol{W}) = \frac{\boldsymbol{S}_s\boldsymbol{S}_s^{\mathrm{T}}}{r-1} + \boldsymbol{W} = \boldsymbol{U}'_s\boldsymbol{\Lambda}'_s\boldsymbol{U}'^{\mathrm{T}}_s. \tag{42}$$

Thus, the similarity of the analysis in the SEEK and the EnKF is obvious when Eq. (41) is compared with the EnKF analysis equation for the ensemble mean (Eq. (36)).

4. The remotely sensed data

The observed parameters assimilated in the MICOM model are Sea-Level Anomaly (SLA) and Sea-Surface Temperature (SST). The same data set is assimilated by the EnKF, the EnKS and SEEK. It consists of gridded SLA and SST on a 1/4° grid every 10 days derived from the available satellite data for the assimilation experiment period.

4.1. Sea-level anomalies gridded products

The SLA data set assimilated into MICOM consists of data obtained from NASA/CNES TOPEX POSEIDON (T/P) and ESA ERS satellite altimeter observations after correction and interpolation on the 1/4° horizontal grid every 10 days.

In order to get the most homogeneous data set to be used in the interpolation, we have used state-of-the-art correction algorithms for both satellites. We have used the NASA-JGM3 orbit for T/P and the D-PAF precise orbit with reference to TOPEX ellipsoid for ERS. The wet tropospheric corrections come from the TMR radiometer for T/P and from ATSR-M radiometer for ERS. The ERS radiometer wet tropospheric correction has been extrapolated near the coast in order to avoid the pollution of the radiometer data by the continent near the coastline.

This interpolation allows to retrieve a significant amount of altimeter data points near the shore where the wet tropospheric correction is flagged "bad data". Dry tropospheric correction comes from ECMWF atmospheric fields. The ionospheric correction is computed from the dual frequency altimeter for TOPEX after applying a 300-km Lanczos filter, from DORIS data for POSEIDON, and from the BENT model for ERS. The ocean tide, and the loading effect are corrected using CSR3.0 model (Eanes and Bettadpur, 1996). The electromagnetic bias is corrected using BM4 formula (Gaspar et al., 1994) for T/P and −5.5% of the significant wave height for ERS. The inverse barometer correction is applied. The instrumental noise is reduced applying a Lanczos low pass filter along each track. The anomalies are computed as departures from a 3-year mean pass (1993–1995) for T/P, and a mean pass for the same period mean using T/P data to correct for orbit error and oceanic signal for ERS.

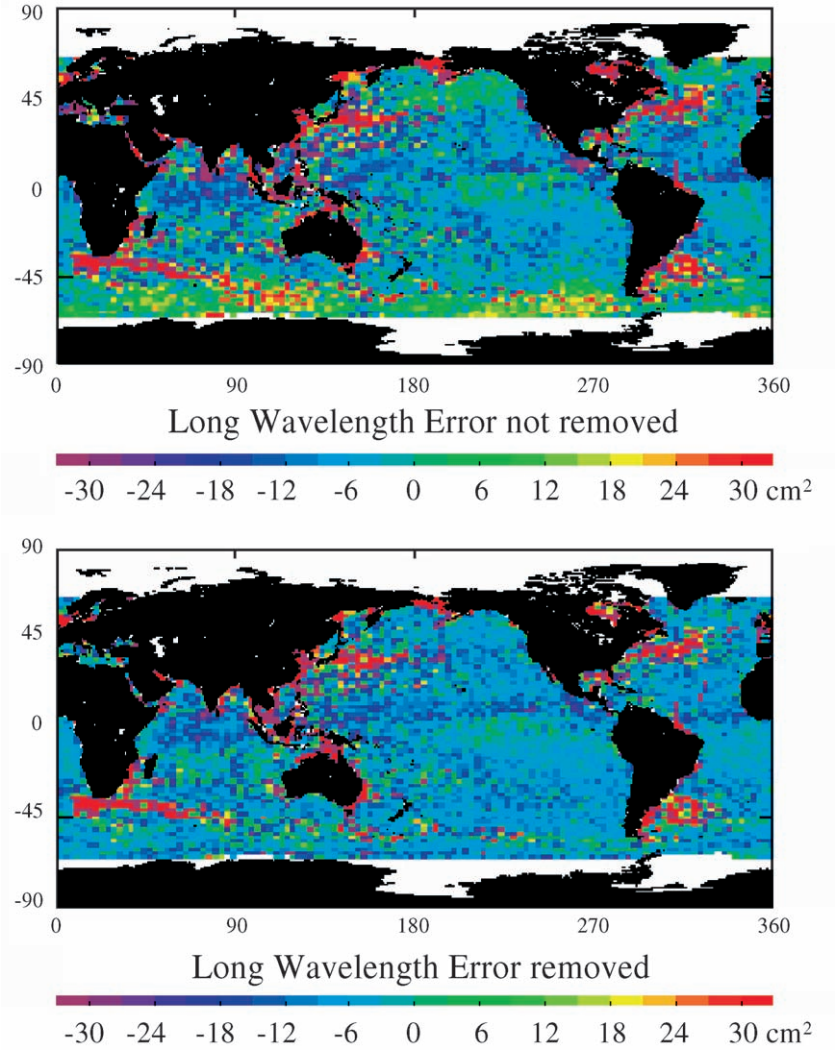

Fig. 1. The difference between the 1-year SLA variance (in cm^2) observed by T/P and ERS. Top panel: before Long Wavelength Errors (LWE) are removed, bottom panel: after LWE removal.

The interpolation of along track data onto the 1/4° grid was done using an optimal interpolation method described by Traon et al. (1998). The originality of this interpolation method is that the long wavelength errors along each track which are mainly radial orbit errors are explicitly taken into account in the inversion by adding extra off-diagonal terms in the observation error covariance matrix. This method improves considerably the results in terms of horizontal coherence (between each ground track) and also in terms of consistency between the different satellite data sets as we can see in Fig. 1, which shows the difference between the SLA variance obtained with ERS and with T/P for a 1-year period. The differences are high in energetic regions because the two satellites do not have the same repeat period and thus the same sampling of the meso-scale. We see that the long wavelength error removal reduces the differences, especially in the regions where the effect of the atmospheric forcing is stronger (near Antarctica), which indicates that the two data sets are more homogeneous after correction.

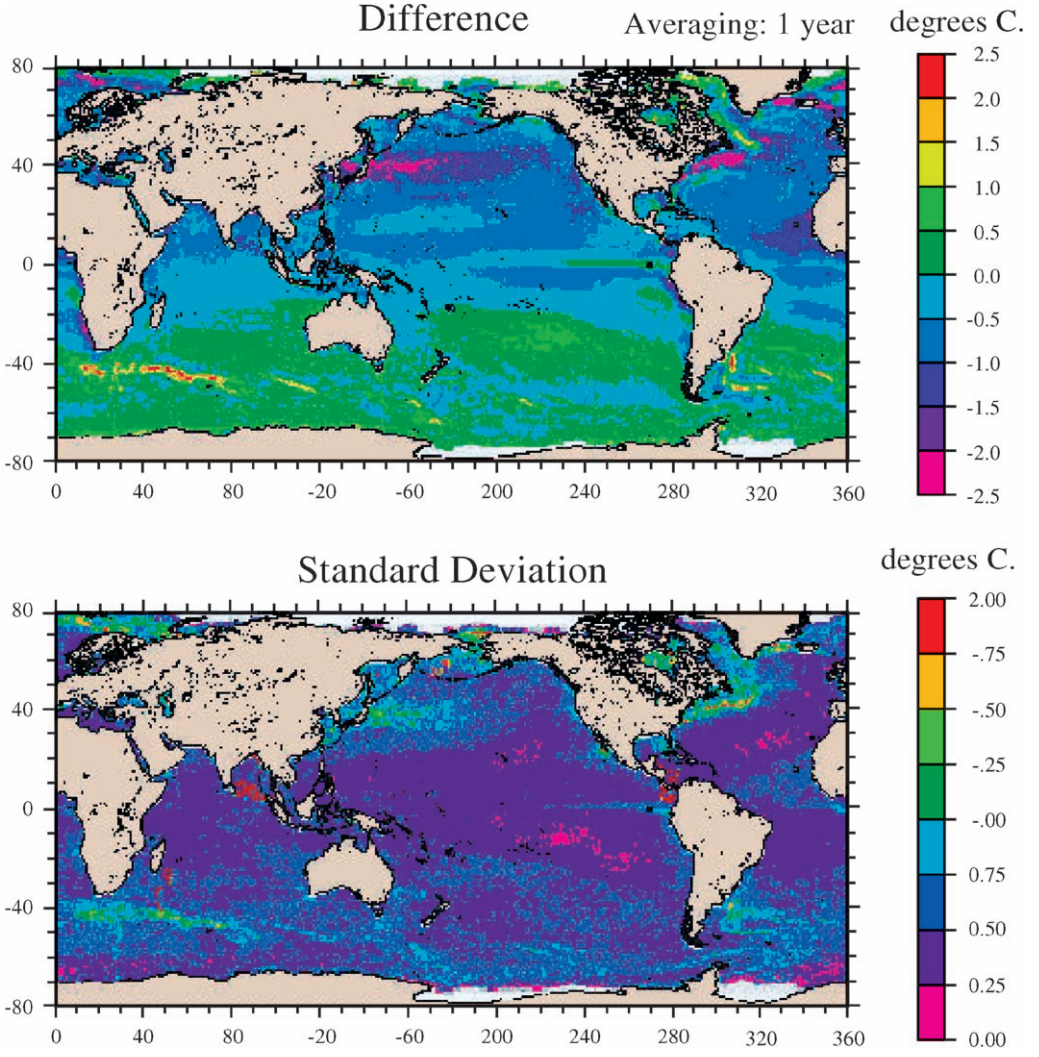

Fig. 2. Top panel: 1-year (1993) average of the differences between our SST product and Reynolds SST. Lower panel: 1-year standard deviation of the differences between our SST product and Reynolds SST.

4.2. Sea-surface temperature

To simplify the assimilation procedure, we have chosen to produce coherent gridded SST and SLA data sets. Thus, daily SST images from the NASA Pathfinder AVHRR with 9-km horizontal resolution has been interpolated to same 1/4° resolution grid as the SLA data set described above, with the same temporal frequency of 10-day period. Thus, the same space timescales are expected to be resolved by both data sets.

Only night images were used in order to avoid the potential problem of skin SST occurring during sunshine at daytime. The data were gridded using a spatial median filter and a temporal mean at 1/4° horizontal resolution every 10 days. All data points from the daily images closer than 18 km (two 9-km pixels in the original images) to the estimation point were selected and the median value was retained. Doing so, we get gridded SST every day. The spatial median filter allows to avoid the effect of spurious data points near the clouds. The temporal mean was then computed so that we get a SST value for a, as large as possible, number of grid-points, especially when the cloud coverage is heavy, while keeping the high frequency signals that can be sampled with a 10-day period. In order to do so, we have computed the average of all the good points available on the 1/4° grid for an 11-day time period around the estimation date. This means that for some grid-points in cloudy regions, the average does not correspond to an 11-day mean value, but to an average obtained with an uneven sampling of the 11-day period. Even so, there are still several places in the world where there is no valid measurement of SST during the 11-day period. When this is the case, a missing value ''flag'' is put in the product instead of a SST value, which implies that the SST is not assimilated at this grid-point.

We have compared our gridded SST with the Reynolds SST over the year 1993. The comparison shows as expected that Reynolds SST is smoother than our gridded SST. Our gridded SST is colder than the Reynolds SST (0.4 °C). This is mainly due to the fact that we use only night images. We can see in Fig. 2 that our gridded SST is generally colder than Reynolds SST in cold areas (North of Kuroshio and Kuroshio extension, upwelling along the African coast, North of the Gulf stream, Irminger current, east Greenland current) and warmer in warm areas (south of Gulf stream, in the convergence zone, on the eastern flank of the Labrador current, North of the ACC) which indicates that our gridded product is better capable to reproduce the temperature gradients in these regions than Reynolds SST. The differences are larger in highly energetic areas. The highest standard deviations occur east of south India, near Panama and southeast of Madagascar. These differences in the equatorial band can be explained by the cloud coverage which is high in these areas, and by the occasional presence of aerosols (Reynolds corrects for the effect of aerosols while we do not). In these area, our gridded SST are generally colder than Reynolds SST.

5. Description of experiment

5.1. Model configuration

The MICOM model, described in Section 2, has been implemented on a grid covering the North Atlantic, the Nordic Seas and the Arctic Ocean. The model grid is an orthogonal curvilinear grid with enhanced resolution in the Nordic Seas, as shown in Fig. 3, with enhanced resolution in the Nordic Seas. The size of the grid cells is approximately 20 km in the North Sea, 40 km in the Gulf Stream region and 80 km in the subtropical gyre. The grid has 140×130 grid-points in the horizontal, while the number of layers in the vertical is 17 (including the mixed layer). This results in 309,400 grid-points which each holds four variables (layer thickness, temperature and two velocity components). With the additional four two-dimensional variables (barotropic pressure, two velocity components and another thermodynamic variable for the mixed layer), the total number of unknowns in the model state becomes $130 \times 140 \times (17 \times 4 + 4) = 1{,}310{,}400$. This is a substantially sized model state compared to previous assimilation experiments using sophisticated assimilation techniques. The use of a variable model resolution and a large model domain allow us to apply ''closed'' boundaries far from the area of interest where a relaxation to climatology is applied.

The model bathymetry was interpolated from the ETOPO-5 data set named DS759.2 from the Terrain-Base project conducted by the National Geophysical

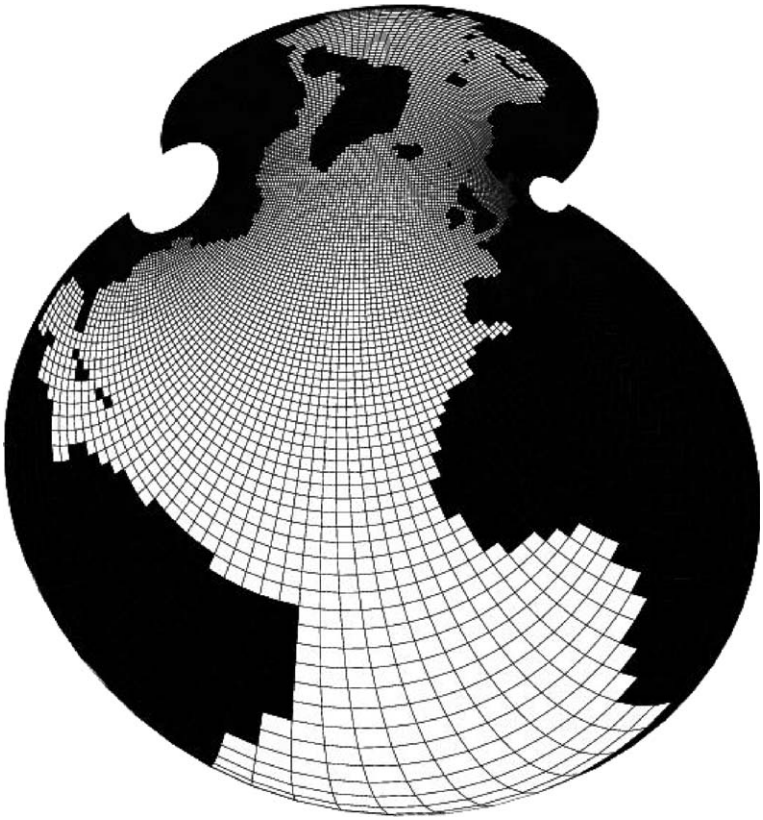

Fig. 3. The diadem1 model grid used in the assimilation experiment.

Data Center and World Data Centers-A for Solid Earth Geophysics and for Marine Geology and Geophysics (NGDC/WDC-A).

5.2. Model forcing

The model is forced by atmospheric data from ECMWF, which is available on an approximately $1 \times 1^\circ$ grid and with six hourly resolution. This includes full thermodynamic forcing with computation of heat and freshwater fluxes in addition to the transfer of wind stress. The atmospheric fluxes are computed from air temperature, relative humidity, dew-point temperature, cloud cover, precipitation and mean sea-level pressure, and wind stresses are derived from the atmospheric wind data.

In the EnKF and EnKS runs, a weak SST relaxation was applied, while this was excluded in the runs with the SEEK filter. All the simulations used a weak relaxation for the fresh water fluxes. The timescale for the SST and SSS relaxation was 60.0 days.

5.3. Model initialization

The model was initialized using outputs from a simulation which started from Levitus temperature and salinity data and with twice the horizontal grid resolution used in the assimilation experiments. This model was first run for 10 years in a spin-up simulation subject to monthly averaged climatological forcing data. Another 5 years of simulation were carried out until June 1996, forcing the model with interannual atmospheric data from the ECMWF.

5.4. Initialization of assimilation experiments

An ensemble of 150 model states was created for June 1, 1996 by perturbing layer interfaces in the

model state resulting from the simulation in Section 5.3. The perturbations were drawn from a sample of smooth pseudo-random fields with mean equal to zero and a prescribed covariance. The algorithm described by Evensen (1994) was used to generate the fields. Evensen (1994) also presented an algorithm for introducing a specified predetermined correlation between a sample of pseudo random fields by choosing each new sample as a specific linear combination of the original fields. Multiplication of each pseudo-random field, having zero mean and variance equal to one, with a constant σ will lead to a new sample with variance equal to σ.

The initial ensemble was, thus, created by adding smooth and vertically correlated pseudo-random fields drawn from a sample with mean equal to zero and for each vertical layer the variance represents 10% of the layer thickness at that location. This results in a sample of model states which all have a slightly different stratification. These differences should be a measure of our confidence in the watermass characteristics of the model initial conditions. From this initial ensemble, a short spin-up simulation is required to ensure that the model state for each individual ensemble member is in dynamic balance.

The ensemble spin-up was done by a 1 month integration until July 3, 1996 (day 184). The mean of the ensemble at day 184 represents the best guess estimate of the true state of the ocean in the EnKF and EnKS, and was used as the initial condition for the SEEK assimilation experiment. The spreading of the ensemble represents a measure of the error covariance of this best guess estimate in the EnKF and EnKS.

In the SEEK filter, the variability of a prior simulation of year 1993 was sampled every 5 days, and the first seven EOFs of the sequence (explaining more than 95% of the total variance) was used to build the initial error subspace.

Indeed, an open question related to the practical implementation of assimilation systems is the specification of the background error, whose impact on the assimilated trajectory can remain during several cycles after initialisation. As there is no unique answer to this question, the idea beyond the use of EOFs is to investigate the relevance of a subspace initialised from the natural variability of a prior run, by comparison with initial ensemble covariances obtained by perturbing layer interfaces using prescribed statistics.

5.5. Random atmospheric forcing

Model errors include both errors of model physics (forcing errors, errors in parameterization, etc.) and errors in the numerical solution methods. In the EnKF, EnKS and in the spin-up of the ensemble, it is assumed that the dominant model errors are connected to misspecifications of atmospheric forcing fields used to compute the surface fluxes. Thus, we have treated the atmospheric forcing as a stochastic process and added spatially and temporally correlated random fields to the atmospheric data. Hence, random atmospheric forcing has been added in order to simulate modeling errors. Pseudo-random fields have been calculated from a Gaussian distribution with zero mean and given standard deviation and were added to the atmospheric forcing data. The random fields are independent for each member and the horizontal decorrelation length is 1500 km. For the different atmospheric fields, we have used the following standard deviations: $\sigma_{\text{air temperature}} = 3.0\ ^{\circ}\text{C}$, $\sigma_{\text{wind speed}} = 5.0\ \text{m s}^{-1}$, $\sigma_{\text{wind stress}} = 0.5\ \text{g cm}^{-1}\ \text{s}^{2}$.

As pointed out by an anonymous reviewer, the wind speed enters the equations through the nonlinear model operator. This means that the noise is not additive, and it is difficult to derive an equation for the full probability density, like Kolmogorov's equation (Eq. (8)). However, this poses no serious problem on the ensemble methods presented here because we still can use Eq. (7) for the model evolution. As mentioned in Section 3.1, this equation is meaningful when the Ito interpretation is used. Thus, eventhough an evolution equation for the probability density is not known, we can approximate its evolution by applying ensemble integrations (and in the limit of an infinite number of ensemble members, an exact evolution is obtained).

5.6. Observations errors

The observations are gridded data sets derived from satellite measurements. Thus, the observation errors will be spatially correlated with a correlation determined by the horizontal scales of the objective analysis scheme used in the gridding process. In order to model this correlation, we have for the EnKF and EnKS assumed a spatial correlation function

$$\text{corr}(i,j) = \exp(-\text{dist}(i,j)^2/\text{dist}0^2), \tag{43}$$

where dist(i,j) is the spherical distance between observation point i and j and dist0 = 10 km is a chosen decorrelation length parameter. Elements $\boldsymbol{W}_{ij}$ in the error covariance matrix are, thus, given by $\sigma(i)\sigma(j)\mathrm{corr}(i,j)$, $j,i = 1, 2, \ldots, \mu$, where μ is the number of observations. In the SEEK filter, a function similar to Eq. (43) is used except that the distances are computed in grid cell units.

Except for in the Nordic Seas, the resolution in the model grid is coarser than the observations grid, hence, the SLA data describes fine-scale variations that the model is not capable of assimilating. In the SLA assimilation context, this fine-scale variation in observed SLA data is considered as noise. The observed SLA data, $\boldsymbol{d}_{\mathrm{SLA}}$, is, therefore, written as

$$\boldsymbol{d}_{\mathrm{SLA}} = \psi^{\mathrm{t}}_{\mathrm{SLA}} + \varepsilon_m + \varepsilon_r, \tag{44}$$

where ε_m is the measurement error and ε_r is a term that is due to the representation error in the model. The true SLA data is given by $\psi^{\mathrm{t}}_{\mathrm{SLA}}$. The error statistics for ε_m is assumed to be Gaussian with zero mean and standard deviation equal to 0.05 m.

In order to approximate statistics for ε_m, we have used 100 SLA data sets (a data set every 10 days starting July 3, 1996) and smoothed these data sets by averaging according to the model resolution. The smoothing of the SLA data is done by averaging the original SLA data within a box centered at each model grid-point. The size of the averaging box decreases with increasing model resolution and, therefore, depends on the typical distance between two neighboring model grid cells. Moreover, the averaging is performed grid-point by grid-point, and the resulting SLA data is constant within the average box. A SLA data remains undefined if the original SLA data is undefined. The smoothed SLA field on measurement day k and observation grid-point (i,j) is denoted $\overline{d}^k_{ij}$, and equals the mean value of the observations located within the averaging box. The SLA variance due to the representation error in the model is then given by

$$v_{ij} = \frac{1}{nrt-1}\sum_{k=1}^{nrt}(d^k - \overline{d}^k)^2_{ij}, \tag{45}$$

where nrt is the number of times a SLA data is defined at gridpoint (i,j) in the SLA data set.

The total variance for the SLA data is given by

$$v_{ij} = v(\varepsilon_m) + v_{ij}(\varepsilon_r). \tag{46}$$

A similar approach is not considered for the observed SST data since the SST field is smoother than the observed SLA field. The standard deviation of the SST measurement errors is set to 0.5° C and the radius of influence, which is used in the subsampling of data, is set to 40 km.

5.7. The assimilation and free-run experiment

The data assimilation experiment was run in hindcast mode, starting on July 3, 1996 and ending 60 days later on September 1, 1996. Satellite data were assimilated every 10 days on a region spanning from 100°W to 40°E and 10°S to 80°N in the EnKF and EnKS assimilation runs, and up to 70°N in the SEEK assimilation run. For the SLA assimilation, we needed a mean Sea-Surface Height (SSH) field in order to generate the model-predicted SLA field. This is because the satellite data are given as SLA data while the model output is SSH data. The model SLA is calculated by subtracting a mean SSH field from the model SSH field. This mean SSH field has been generated as a time mean of SSH fields from a high-resolution model simulation with MICOM. The model has been run for 3 years and SSH fields have been sampled throughout this time period. A time mean SSH field has been calculated from the sampled SSH states. The resulting SSH mean field is given in Fig. 4. The signature of the large-scale features of the North Atlantic surface circulation is imprinted on the mean SSH, with essentially the cyclonic circulation of the subpolar gyre, the anticyclonic circulation of the subtropical gyre and the intensification of the western boundary current. The Gulf Stream and the North Atlantic drift, however, are not properly reproduced and display a number of flaws typically found with medium-resolution models, i.e., spurious anticyclonic eddy at Cape Hatteras and overshooting of the Gulf Stream northward. These flaws are not crippling for this demonstration study, but a better mean SSH will be needed at a later stage for improving the realism of the assimilation products.

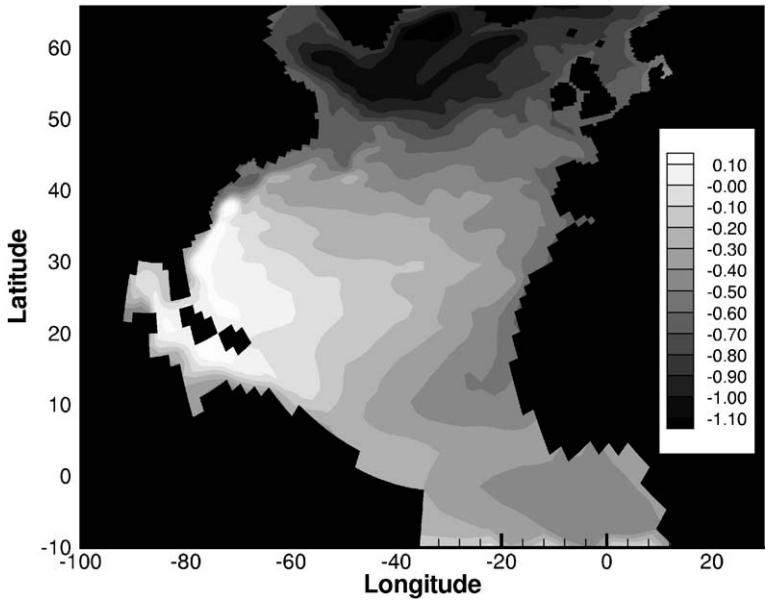

Fig. 4. The annual averaged SSH field used in the assimilation experiments.

In addition, a so-called free-run simulation was performed for the same time period as the assimilation experiment. In this experiment, the model run in free mode, i.e., with no assimilation of SLA and SST data. The results from the assimilation runs and the free-run have been intercompared in Section 6 in order to assess the impact of assimilating SLA and SST data.

6. Discussion of results

The focus of this section is put on the interpretation and examination of the impact of the assimilation for each of the assimilation schemes. In order to do so, various diagnostics have been computed; we considered plots of the observed data, the difference between the forecast and observed data, the difference between the analysis and observed data and the difference between free-run model data and observations. The innovation plots have been studied to examine to which extent the different data assimilation schemes ''force'' the model state towards the observations. In addition, the assimilation schemes have been validated by calculating the Root Mean Square Error (RMSE) values of the innovation fields, i.e., the difference between the modelled data and observed data before and after the analysis. The impact of assimilating SLA and SST over time has also been examined by plotting observed and simulated temperature, salinity, zonal and meridional velocity for the upper 500 m for a selected location in the Gulf Stream. Finally, we compared the assimilation results with independent XBT data from the TOGA/WOCE/CLIVAR data bases, made available by the SYSMER/IFREMER data center in Brest.

6.1. Correlations between MICOM variables

The SEEK filter, EnKF and EnKS are all multivariate assimilation schemes, which means that all the model variables are updated in a systematic manner according to the approximations of the covariances between the variables. The correlations between the variables are calculated using the ensemble of model states in the EnKF and EnKS, while the SEEK filter uses the dominant modes of its error subspace. Some of the model variables are more correlated than others, as can be seen from Fig. 5. Plots of correlations between SSH and SSH, mixed layer density (TH(1)), mixed layer thickness (DP(1)) and SST are given left and similar correlations plots for SST are given right. The calculations are done using the EnKF

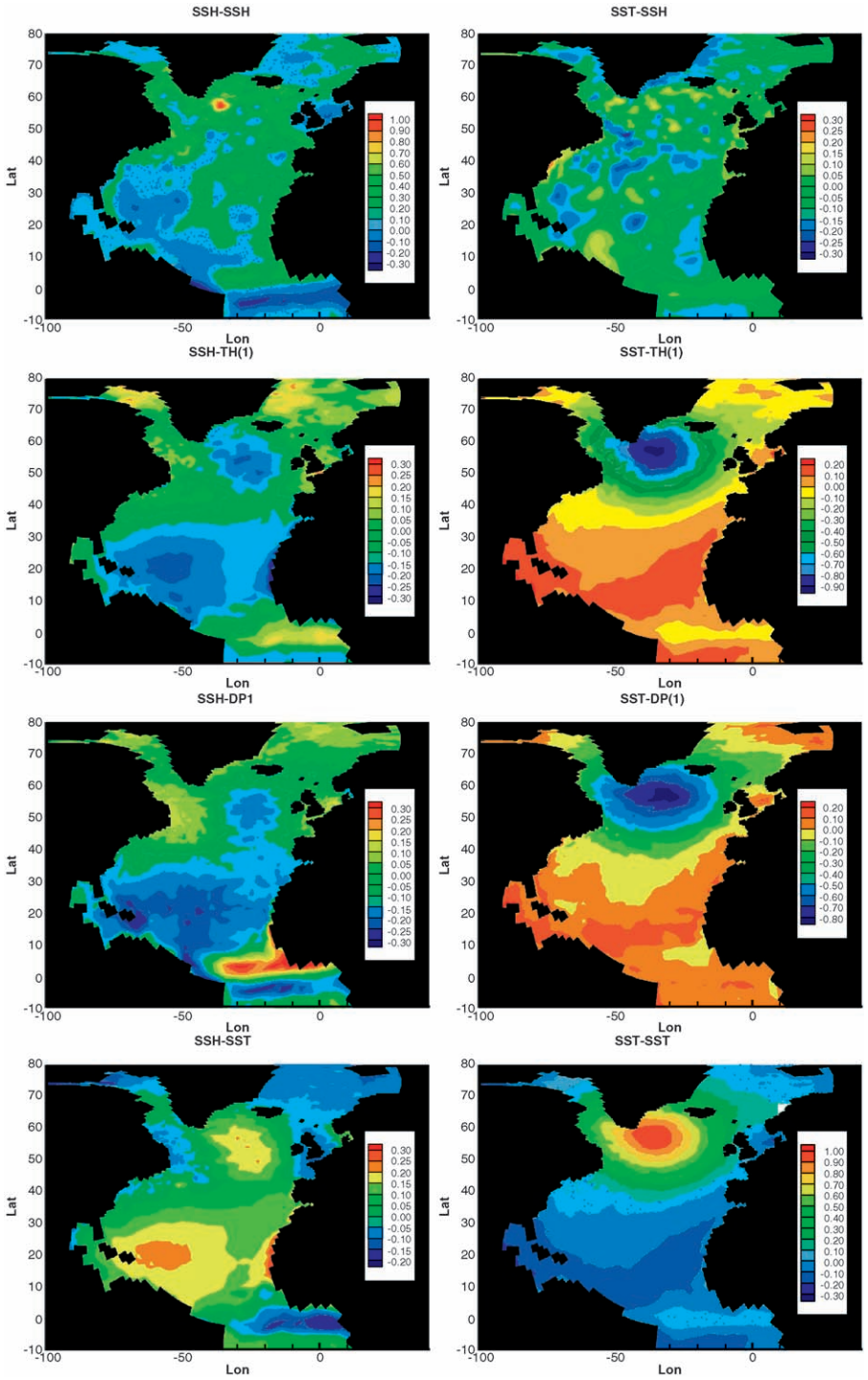

Fig. 5. Correlations between MICOM variables corresponding to location 37°W and 57.5°N, calculated from EnKF forecast ensemble day 184.

forecast ensemble on day 184, the specific location chosen is 37°W and 57.5°N. At this point, the SSH–SSH and SST–SST correlations are identical to one and decrease with increasing distance from the chosen point. The horizontal decorrelation length scales in the SSH–SSH and SST–SST correlation plots are not equal. This is due to the fact that SSH is more influenced by dynamical processes in the ocean, while SST is more dominated by smooth atmospheric conditions. The other plots show weak correlations between the variables involved, except from the SST–TH(1) and SST–DP(1) correlation plots that show a rather strong negative correlation between the variables. This confirms the expectation that a warming of the mixed layer should lead to a reduction of the mixed layer thickness and density.

6.2. Sea-level anomalies

The upper-left plot in Fig. 6 shows the observed SLA on day 184, which is the first day of the assimilation experiment. The middle-left plot shows the forecast of SLA and the lower-left plot shows the innovation between the observed SLA and the SLA forecast on day 184. Clearly, there are large discrepancies in the model produced and observed SLA, especially in the central Gulf Stream region. The large discrepancies are expected since the Gulf Stream extension is a highly energetic area dominated by strong meso-scale activity which is not resolved by the coarse resolution in the model. Another model problem is also related to the well-known northward shift of the Gulf Stream separation, which means that the model will have the wrong placement of the Gulf Stream front.

The right plots in Fig. 6 show the innovation between the observed SLA and the EnKS analysis (upper), the EnKF analysis (middle) and the SEEK analysis (lower). The best result is obtained with the EnKS scheme which updates the SLA field using future measurement of SLA and SST up till day 244.

The SLA innovation plots for day 244 are given in Fig. 7. The upper-left plot shows the difference between the free-run and observed SLA, while the difference between the observed and EnKF/EnKS forecast of SLA and the observed and SEEK SLA forecast are given in the middle-left and lower-left plots, respectively. In general, the differences between the observed SLA and EnKS/EnKF forecast of SLA have decreased from day 184 to day 244 (compare lower-left plot in Fig. 6 and middle-left plot in Fig. 7). The discrepancies have especially decreased outside the coast of South America, in the central Gulf Stream region and in the Norwegian Sea. This is because the assimilation of SST and SLA keeps the model state close to the observed data. A similar comparison for the SEEK forecast of SLA shows a strong reduction in innovation values along the South America coast lines and in the Gulf Stream.

The right plots in Fig. 7 show the SLA innovation fields for the EnKS analysis (upper), the EnKF analysis (middle) and the SEEK analysis (lower). By comparing the innovation fields corresponding to the forecast and analysed SLA (respectively left and right plots in Fig. 7), it is clear that both the EnKF, EnKS and SEEK assimilation schemes "force" the SLA field towards the observed SLA data. Moreover, in the central Gulf Stream region, the difference between the observed and analysed SLA data is smaller in the SEEK run than in the EnKF/EnKS run. This can be explained by the fact that the Gulf Stream region is high variability area and, consequently, the subspace span by the dominant error modes in the SEEK filter is larger in this region than in regions with less variability. In the northern Norwegian Sea, the SEEK assimilation (as opposed to the EnKF and EnKS runs) results in relatively large discrepancies between the model data and the observations.

The EnKF and EnKS analysis are very similar on day 244. They are, however, not identical because the observation field has been smoothed according to the model grid resolution in the EnKS assimilation. This smoothing procedure has been necessary in order to reduce the total number of observations in the EnKS analysis.

To summarize, in general, a time reduction of the discrepancies between the SLA forecast and the SLA observations can be found for all assimilation schemes. Moreover, in general the quality of the free-run SLA (as compared to the observed data) is worse than the quality of forecasts produced in the assimilation runs. In addition, the EnKS analyses are generally closer to the observed data than the SEEK and EnKF analyses. This could be due to the smoother properties of the EnKS.

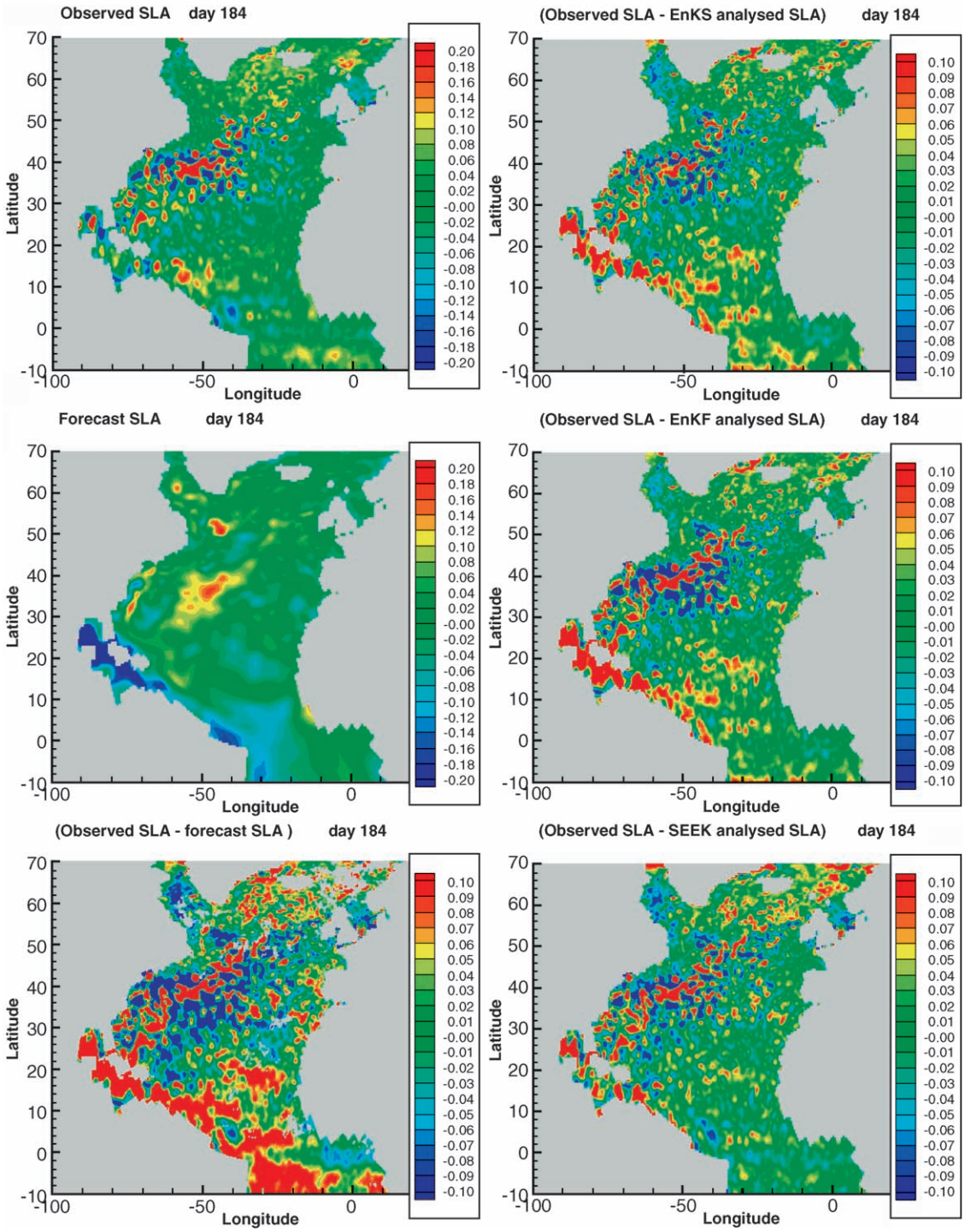

Fig. 6. Numerical results for SLA field, day 184 1996.

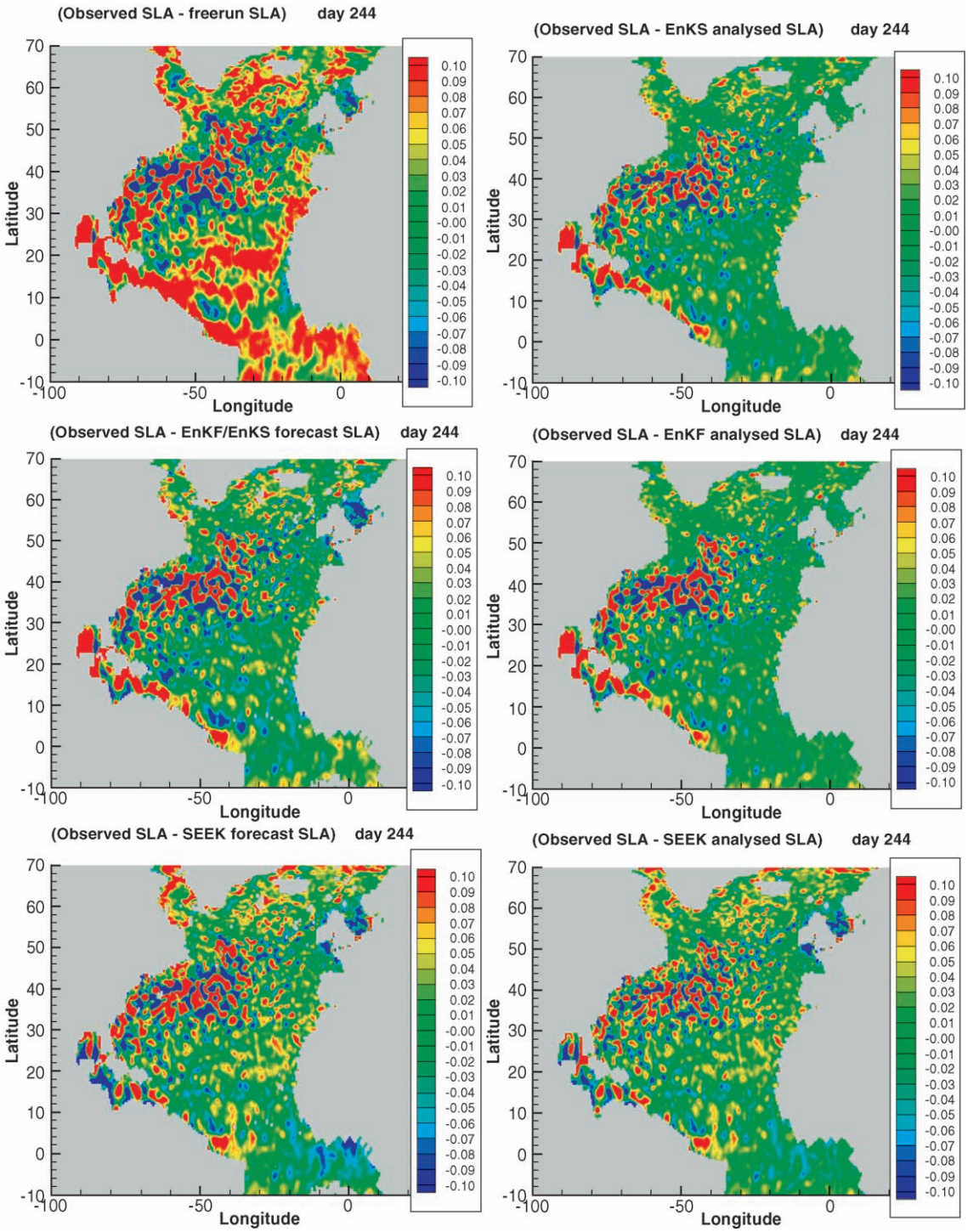

Fig. 7. Numerical results for SLA field, day 244 1996.

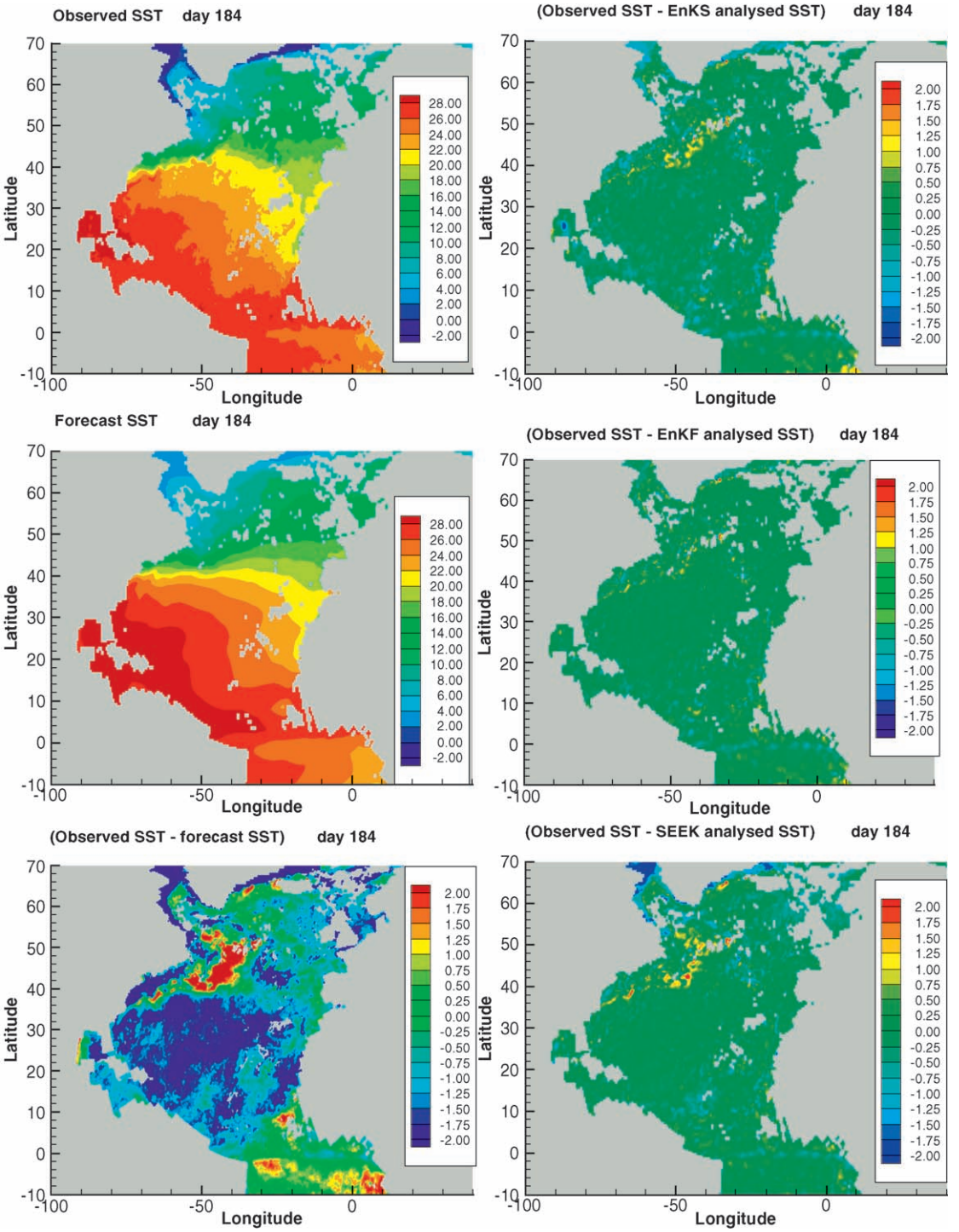

Fig. 8. Numerical results for SST field, day 184 1996.

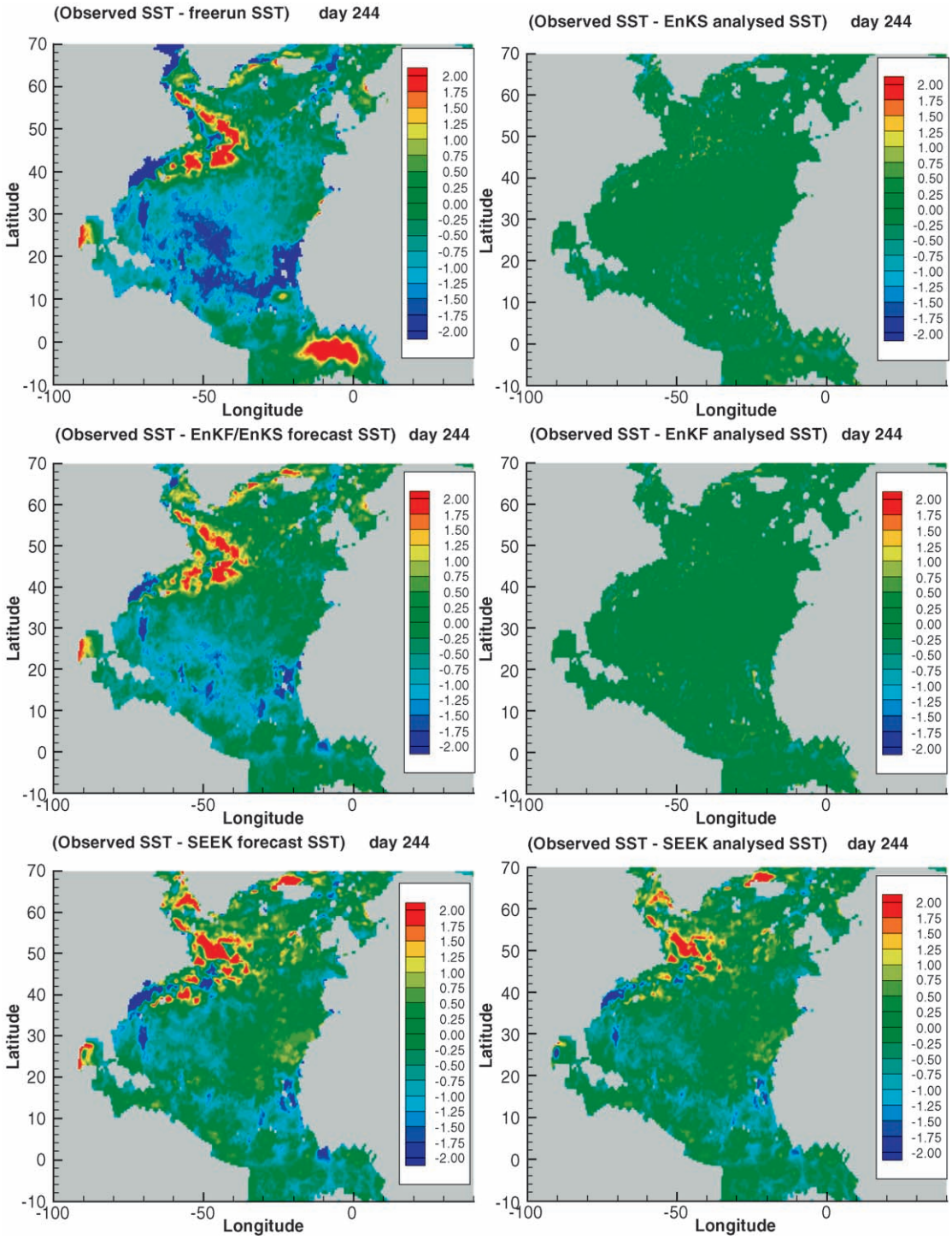

Fig. 9. Numerical results for SST field, day 244 1996.

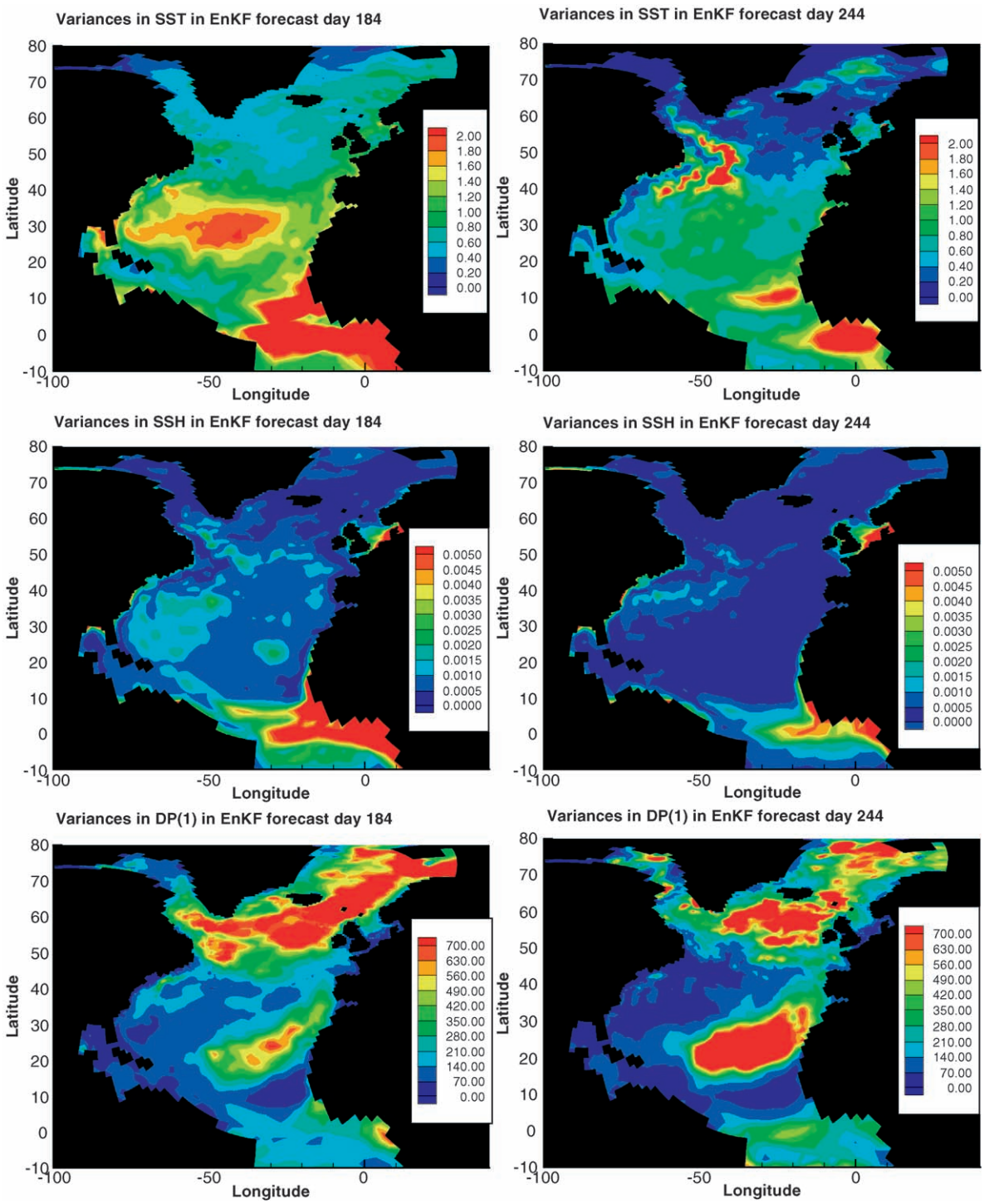

Fig. 10. Variances in EnKF forecast ensemble, on day 184 (left) and 244 (right). SST (upper), SSH (middle) and mixed layer thickness (lower).

6.3. Sea-surface temperature

Plots of observed and forecasted SST on day 184 are given in Fig. 8, upper left and middle left, respectively. The difference between the two data fields is given in the lower left plot. Large discrepancies, due to the coarse model resolution and the unrealistic Gulf Stream separation at Cape Hatteras, are found in the central Gulf Stream region. The model SST is generally higher than the observed SST between latitudes 10°N and 40°N. The difference between the observed and the updated SST fields resulting from the EnKS, EnKF and SEEK analyses are seen in the right plots in Fig. 8. It is clear that the analyzed fields are more similar to the observed SST than the forecast fields are indicating that the assimilation schemes give model updates that are close to the observations. Moreover, the EnKS analysis is somewhat further away from the observations than the EnKF analysis, especially in the Gulf Stream region. This can be explained by the fact that the model temperature of the mixed layer is in direct contact with the atmosphere and, therefore, changes on the timescale of the atmospheric changes. Hence, the SST has a rather short decorrelation scale and the correlations are negligible after 10 days. Consequently, the EnKS assimilation of SST data backward in time seems to be unnecessary and even seems to bring noise into the system (due to artificial time correlations in the ensembles corresponding to different data assimilation times). Since the SLA field has a much longer time decorrelation scale than the SST field, the EnKS behaves differently within the SLA assimilation context.

The numerical results for the SST field on day 244 are given in Fig. 9. By comparing the plots of the difference between observed SST and EnKF forecast of SST on day 184 (lower left in Fig. 8) and day 244 (middle left in Fig. 9), we observe that the discrepancies between the model and observations have decreased almost everywhere in the model domain. The same tendency is seen for the SEEK forecasts of SST indicating that the data assimilation efficiently forces the model SST to follow the evolution in time of the observed SST. This can also be confirmed by

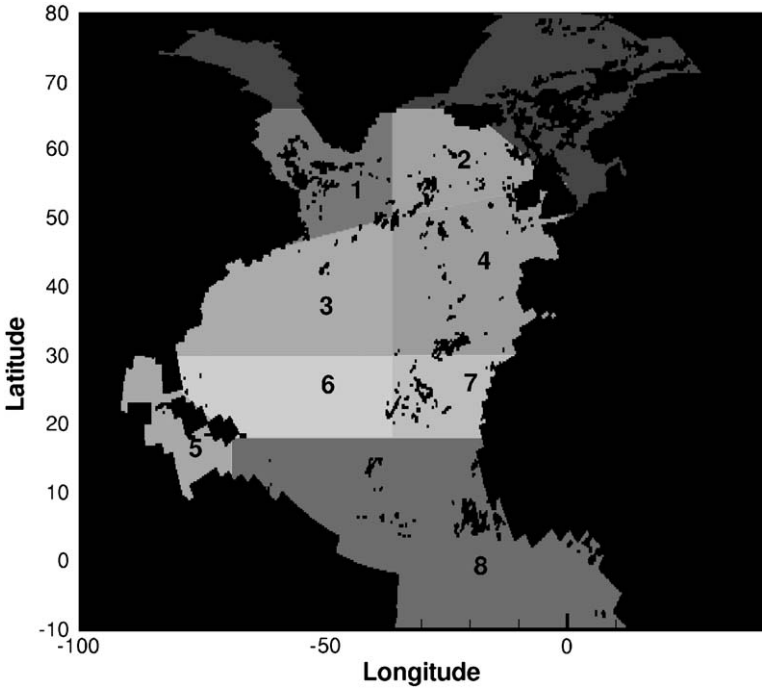

Fig. 11. Subdomains used in the RMSE calculations.

comparing plots of the difference between free-run and observed SST, with the corresponding plots of SEEK and EnKF forecasts (see left plots in Fig. 9). The right plots in Fig. 9 show the difference between the observed and model SST after assimilation of SLA and SST. Again, the plots clearly show that all assimilation schemes efficiently update the model SST towards the observed SST.

Note that the EnKF forecast of SST is typically closer to the observations than the SEEK forecasts. This can partly be due to the SST relaxation that has only been applied in the EnKF and EnKS runs. In general, the updates in SST are larger in the EnKF and EnKS runs than in the SEEK filter run. On the other hand, strong SLA updates can be found both in the SEEK and EnKS assimilation runs. This is perhaps due to the different ways the model error covariances are calculated in the EnKF and SEEK filter. Possibly, the variances (errors) in model SLA are higher in the SEEK run than in the EnKF run, and vice versa for the SST field. In all cases, the analysis corrects the model field towards the observed data field and typically the innovation between observations and model forecast data decreases as a function of time (not shown). The innovation plots for the free-run, SEEK and EnKF assimilation runs clearly indicate that the quality of

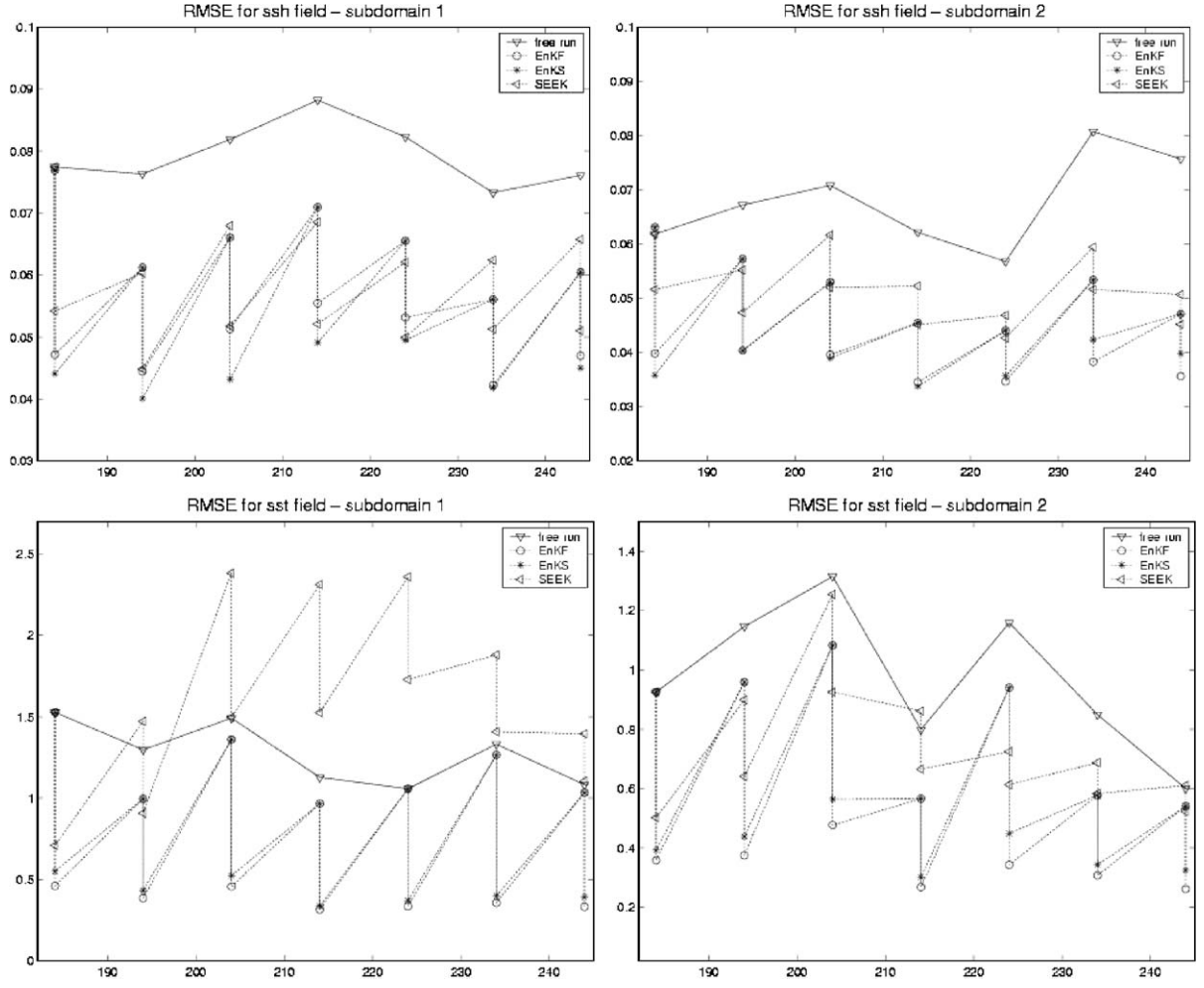

Fig. 12. Time series of RMSE values for subdomains 1 (left) and 2 (right).

the SST and SLA forecasts typically improves when assimilating SLA and SST. However, the quality of the forecast from the assimilation runs is slightly worse in some specific regions (compared to the observations). This can be due to the relatively coarse resolution in the model or the assumption on Gaussian statistics that is made in the analysis schemes.

6.4. Time evolution of modelling errors

The modelling errors for the different variables in the model are approximated from the ensemble of model states in the EnKF and EnKS and from the dominant error modes in the SEEK filter. For the EnKF and EnKS, the size of the modelling errors is determined by the spreading of the model states in the ensemble. Using the ensemble of model states to calculate the variances for the different model variables, the modelling errors can be investigated. Fig. 10 shows the variances in the SST, SSH and mixed layer thickness on day 184 (left) and day 244 (right). By comparing the leftmost and rightmost plots, it can be seen that the modelling error for these variables typically decreases as a function of time. This is an effect of the EnKF assimilation scheme that efficiently decreases the distance between the ensemble members. Hence, even though the spreading of the ensemble members increases during the forecasting of the

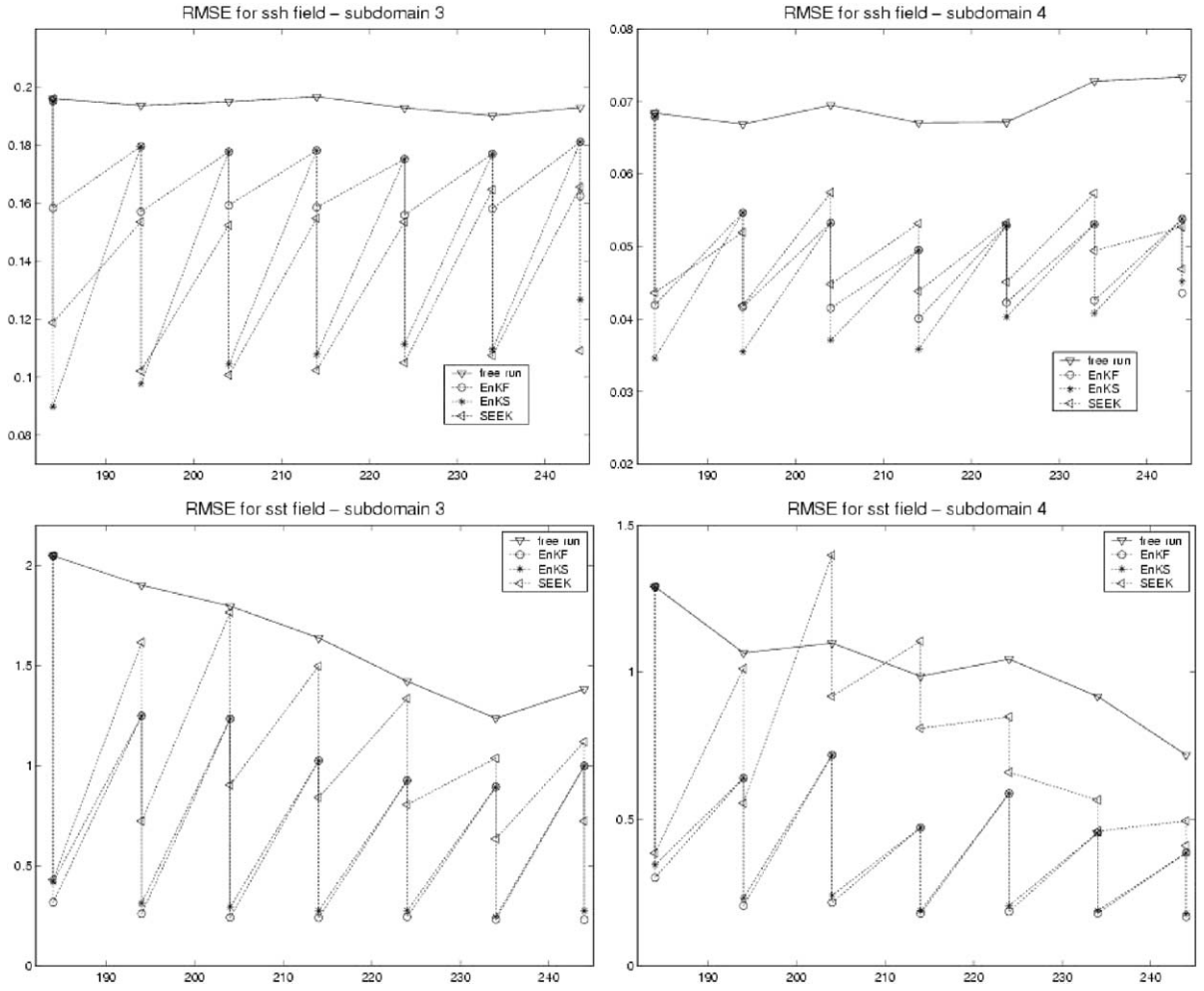

Fig. 13. Time series of RMSE values for subdomains 3 (left) and 4 (right).

ensemble, the effect of the assimilation over time is a general reduction in ensemble variance.

By comparing the patterns in the residual plot for the SLA field in Fig. 7 (middle left) and the SLA variances plot for day 244 in Fig. 10 (middle right), we can observe that the modelling of the SLA errors is to some extent consistent with the observed residuals. It is consistent in the sense that low and high residual values in Fig. 7 generally correspond to high variance values in Fig. 10. The exception is outside the West coast of Africa. Similar observations are made when comparing the residual and variances plots for the SST fields in Fig. 9 (middle left) and Fig. 10 (upper right). The magnitude of the SLA and SST residual errors on the other hand are not necessarily consistent with the magnitude of the square root of the corresponding variances. This can be due to the way the modelling errors are incorporated through the random atmospheric forcing, e.g., the specific standard deviation values chosen for the different atmospheric forcing fields might be appropriate in some regions of the ocean and not appropriate in others.

6.5. Root mean square error of innovation fields

In order to measure the typical "distance" in phase space between the predicted or analysed model state

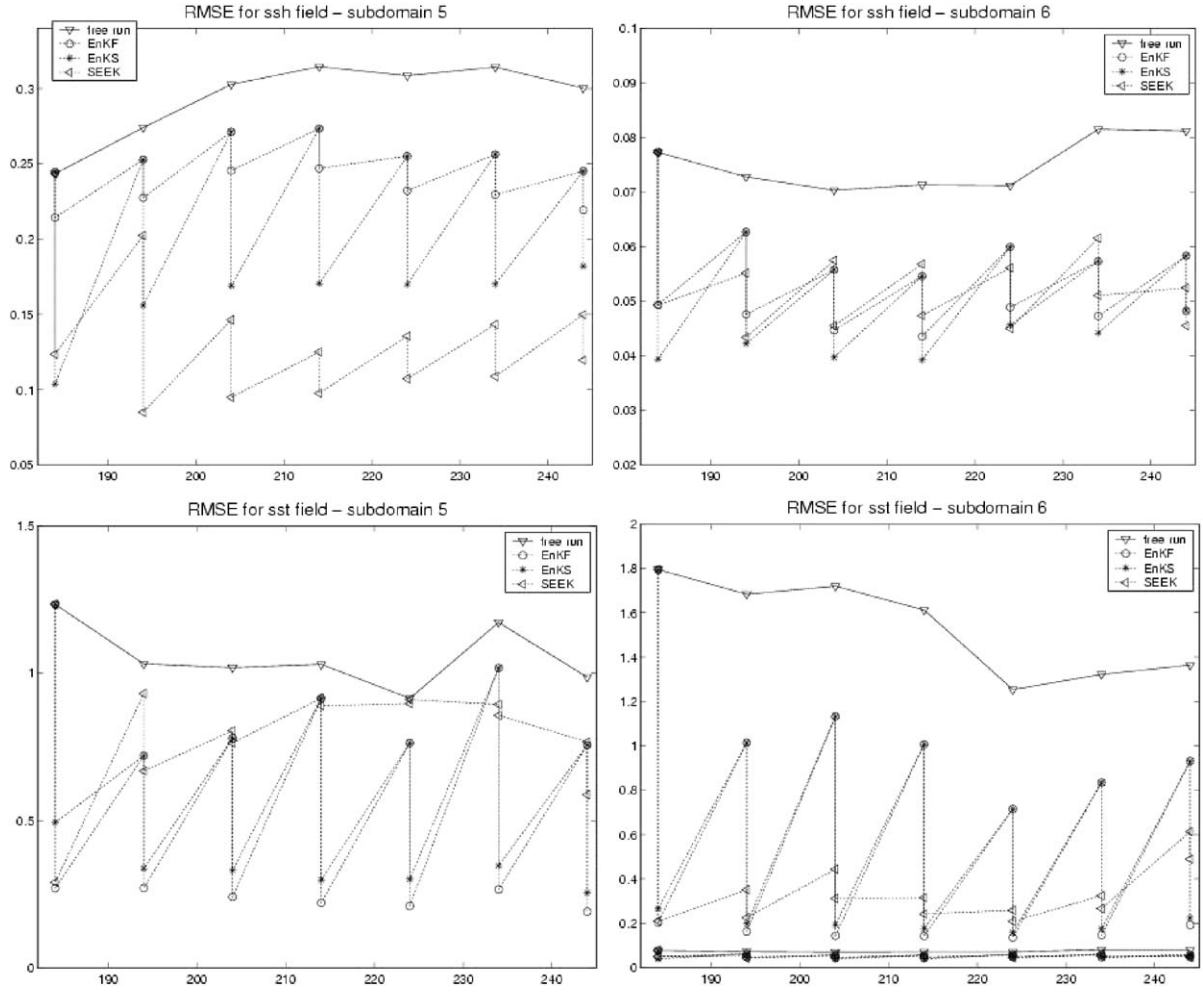

Fig. 14. Time series of RMSE values for subdomains 5 (left) and 6 (right).

and the corresponding observed state, we consider the RMSE norm given by

$$\mathrm{RMSE}(x) = \mathrm{sqrt}\left(\frac{1}{p}\sum_{j=1}^{p} x_j^2\right), \tag{47}$$

where $x_j = d_j - (H\psi)_j$ and p is the number of grid-points where observations are defined. The RMSE value is calculated in specific predefined subdomains of our model domain (see Fig. 11). Since the SEEK filter has not assimilated data in coastal zones, only data that are measured at locations where ocean depth exceeds 500 m contribute in the calculations. The RMSE is also calculated using data from a free-run simulation in order to study the impact of the data assimilation on the model state. The effect of the data assimilation can be examined by intercomparing the RMSE values corresponding to the free-run simulation and the model runs with data assimilation, respectively.

The RMSE values that result from the free-run, EnKF, EnKS and SEEK filter assimilation cycles can be found in Figs. 12–15. The upper plots in the figures show the RMSE values for the SLA fields and the lower plots show the corresponding RMSE values for the SST fields. The SLA–RMSE values in the different subdomains indicate that the typical distance between model and observed SLA is reduced when data assimilation is applied in the model run (compare free-run

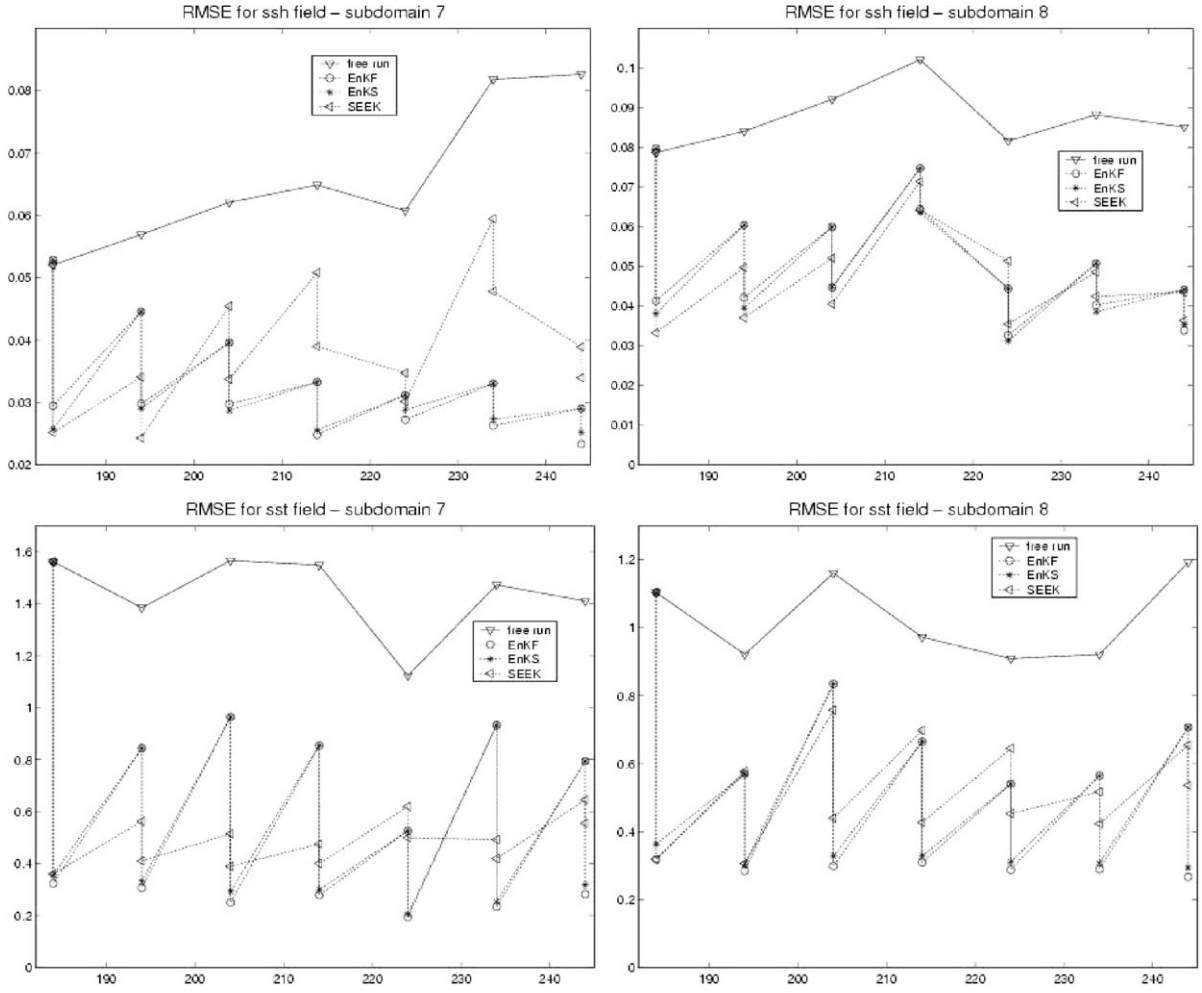

Fig. 15. Time series of RMSE values for subdomains 7 (left) and 8 (right).

RMSE values and RMSE values resulting from the EnKF, EnKS and SEEK filter assimilation cycles). Moreover, in most subdomains, e.g., subdomains 1, 4 and 6, the EnKS analysed SLA fields typically give the smallest RMSE values. Again, this is due to the smoothing properties of the EnKS. In subdomain 3, however, the SEEK filter performs better than the EnKS. Moreover, the SLA–RMSE values calculated

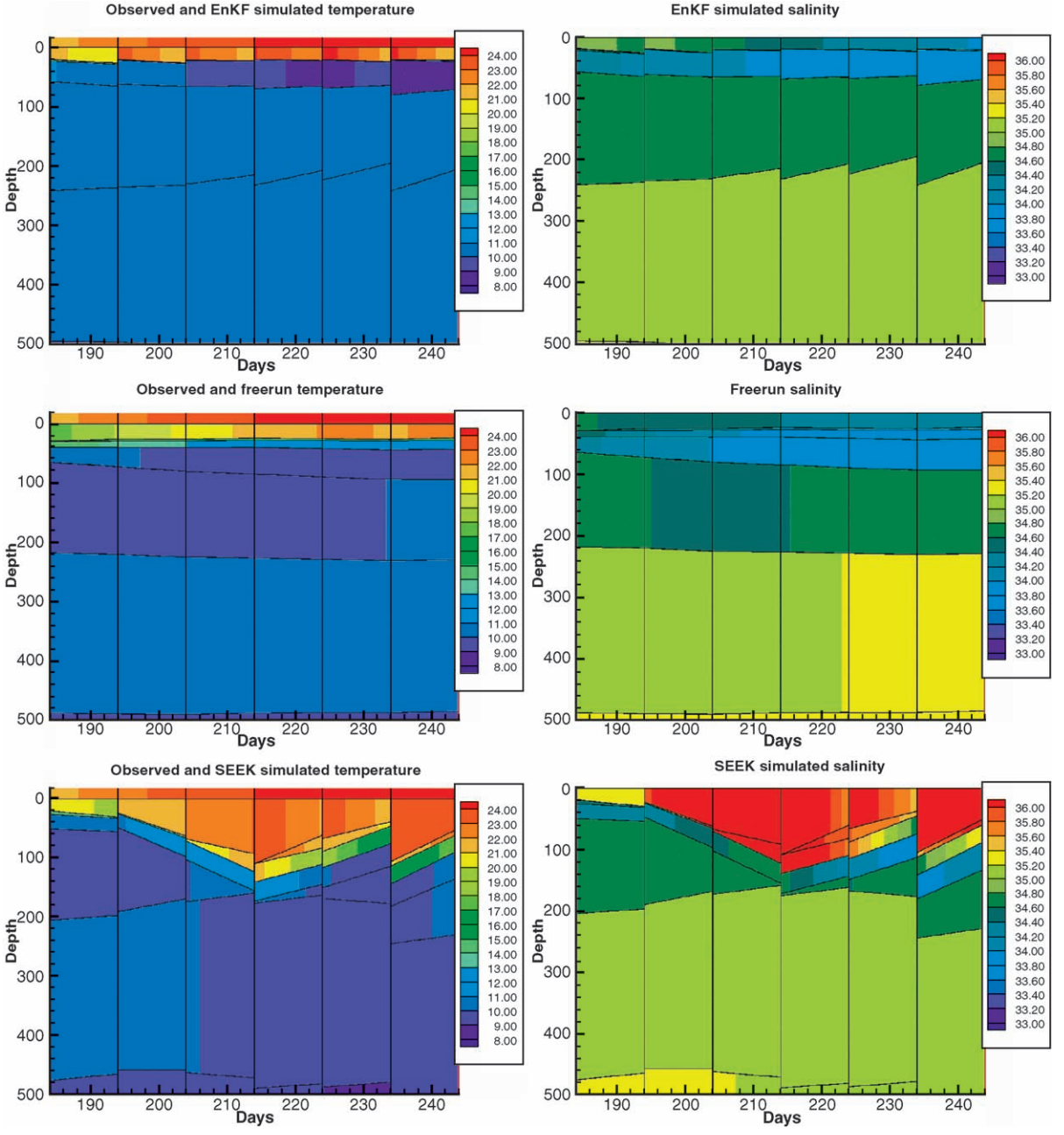

Fig. 16. Time–depth sections of observed and simulated temperature (left) and salinity (right). EnKF data (upper), free-run data (middle) and SEEK data (lower).

from the SEEK forecasts are even smaller than the SLA–RMSE values calculated from the EnKF analysis in this region. Also in subdomain 5, the SEEK performance is the best. Both subdomains 3 and 5 are high variability areas, i.e., regions where the SEEK filter has shown to provide strong updates. In other subdomains,

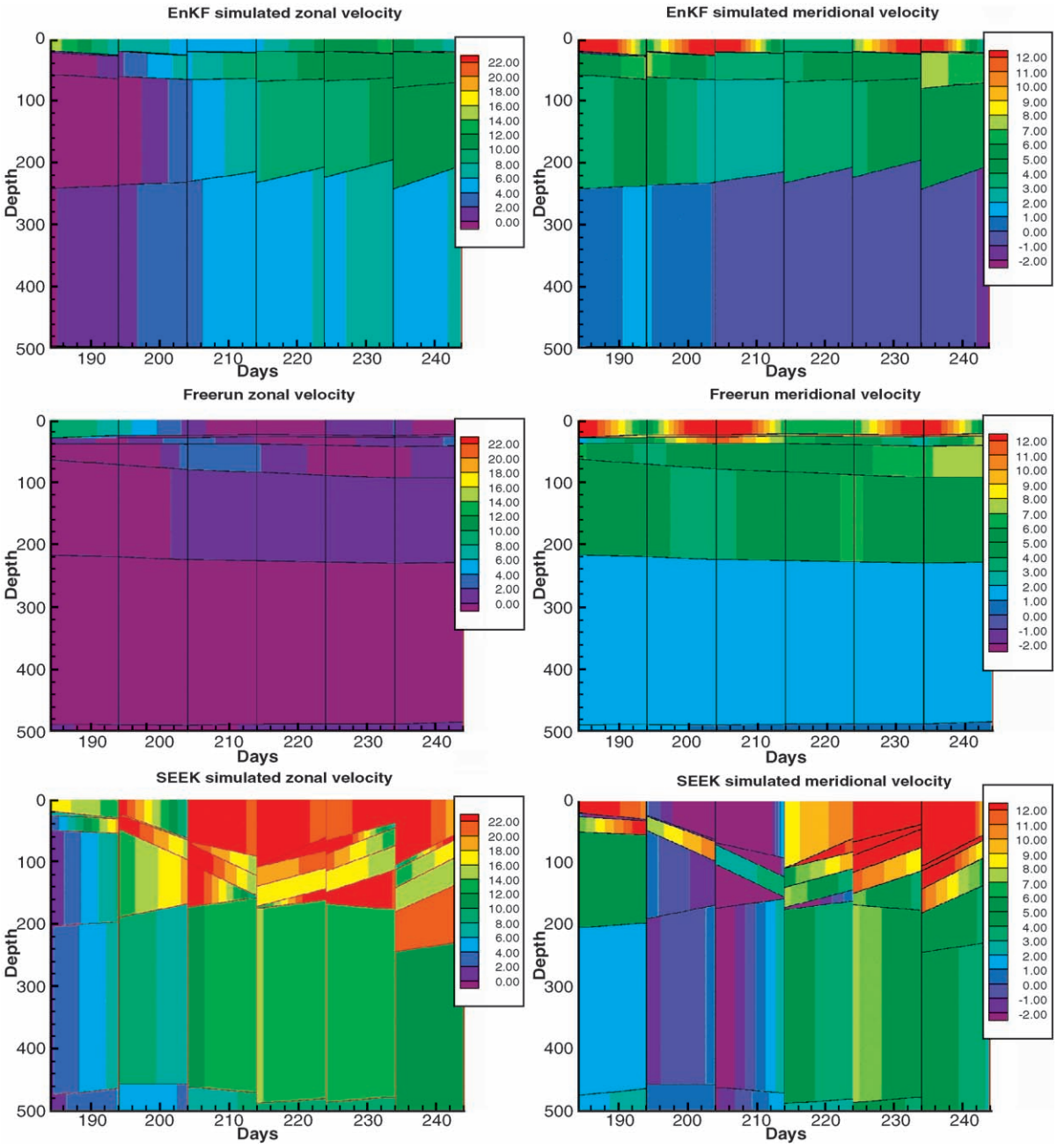

Fig. 17. Time–depth sections of observed and simulated zonal velocity (left) and meridional velocity (right). EnKF data (upper), free-run data (middle) and SEEK data (lower).

typically with less variability, the performance of the various assimilation schemes is generally in the favor of the EnKS.

The SST–RMSE values calculated from the SEEK forecast are generally larger than the corresponding RMSE values from the EnKF and EnKS forecasts. This can be due to the SST relaxation that has been considered in the EnKF and EnKS runs. In some cases, the SST–RMSE values from the SEEK run are even larger than the SST–RMSE values from the free-run, see, e.g., subdomain 1 in Fig. 12. The exceptions can be found in subdomains 6, 7 and 8, i.e., in the area close to the equator. From the lower plots in Figs. 12–15, it is seen that the EnKF-analysed SST fields typically result in smaller RMSE values than the EnKS-analysed SST fields. This is due to the relatively short decorrelation timescale for the SST field.

6.6. Time–depth sections

The assimilation of SST and SLA impacts the model state not only in the surface layer, but also in the isopycnic layers below. In order to examine this effect on the temperature, salinity, zonal and meridional velocities, we have plotted the time evolution of these variables at a selected location in the Gulf Stream region (− 45°W, 42.2°N) in Figs. 16 and 17.

The temperature sections for the EnKF, free-run and SEEK simulations are given in Fig. 16, upper left, middle left and lower left, respectively. The plots include the observed SST, which is given in the fictive top layer with negative depth. The effect of the assimilation is clearly seen for all runs. The isopycnic layers are adjusted at each assimilation step, the adjustment in the SEEK filter is generally stronger than in the EnKF run. In both the SEEK filter and EnKF, the simulated temperatures in the mixed layer adjust towards the observed SST every assimilation update. However, since there is a negative correlation between SST and the mixed layer depth, a decrease in mixed layer thickness is expected when the model SST is increased towards the observed SST. This is only found in the EnKF temperature section plot. A stronger impact of the observed SLA in the SEEK

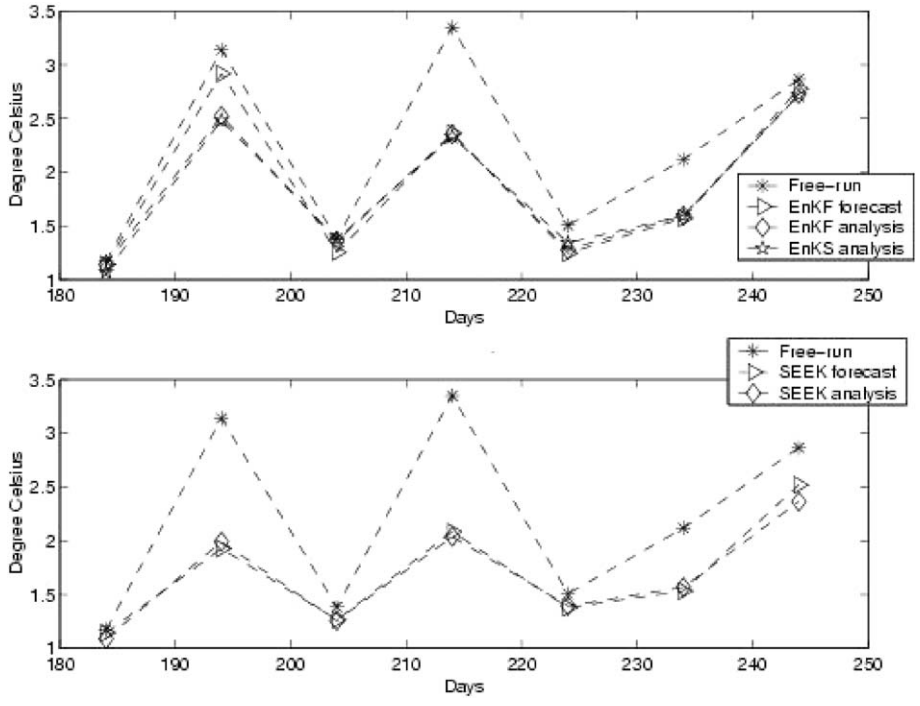

Fig. 18. XBT–RMSE plots for the EnKF and EnKS runs (upper) and the SEEK run (lower).

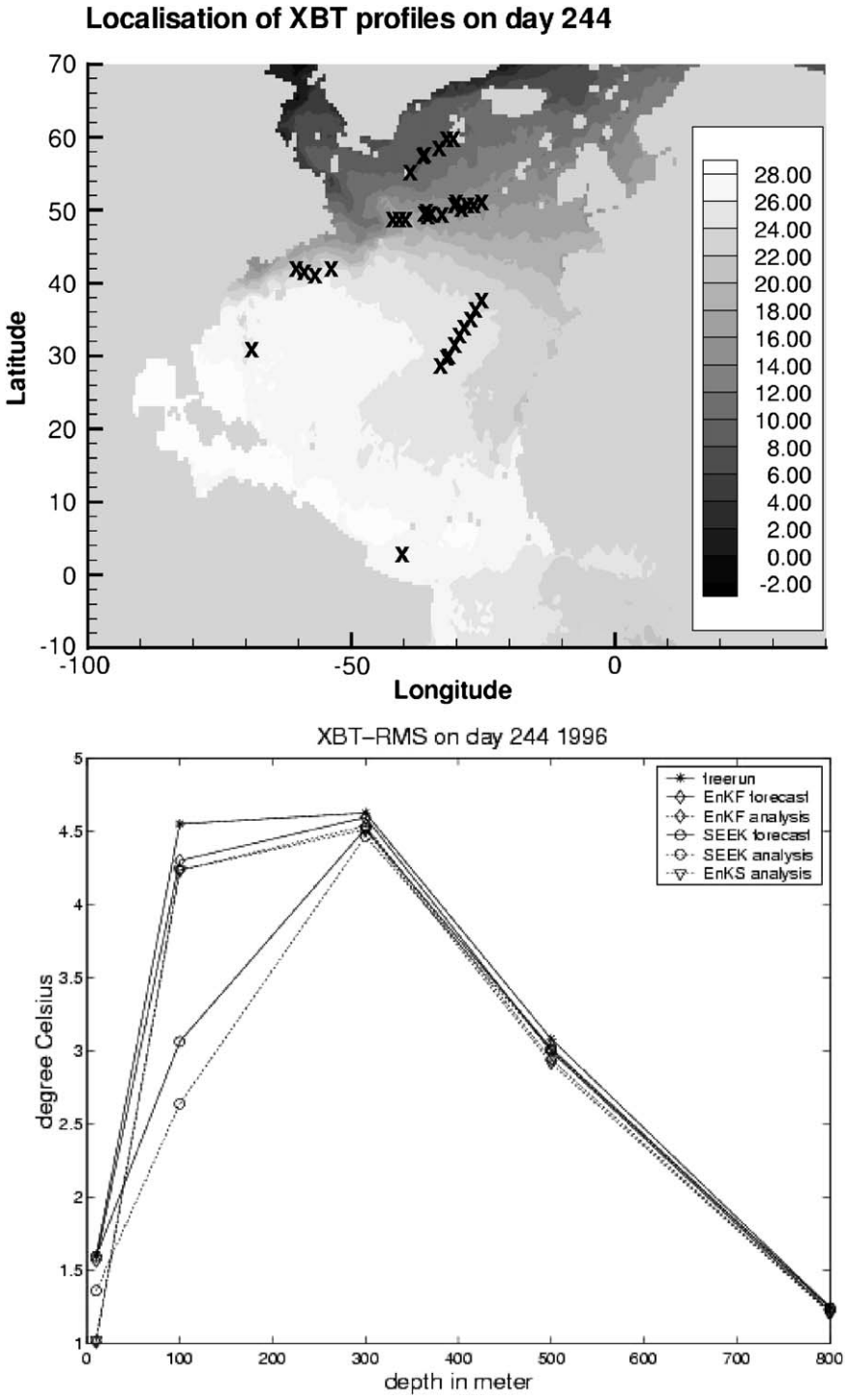

Fig. 19. The localisation of the XBT data on day 244 (upper) and the corresponding RMSE versus depth plot for the free-run, EnKF, EnKS and SEEK assimilation runs (lower).

filter than in the EnKF may explain this result. Only the upper 500 m of the ocean is plotted, since the updates below are rather weak.

The salinity section is given in rightmost plots in Fig. 16. Clearly, the data assimilation does not impact the salinity values in the isopycnic layers much, i.e., the salinity update is rather small within each isopycnic layer. However, by comparing the free-run and assimilation salinity fields it is seen that the time evolution of the salinity is changed due to the assimilation of SLA and SST. Hence, the assimilation (especially the SEEK assimilation) has an indirect effect on the salinity conditions.

The updates in zonal and meridional velocity are rather weak in the EnKF sections and strong in the SEEK sections, see Fig. 17. Hence, the two assimilation schemes approximate the covariances between the velocity components and the observed data types in a completely different manner. The SEEK filter approximates strong correlations while the EnKF approximates weak. This is probably due to the different ways the error space (covariances) is represented and propagated forward in time within the SEEK and EnKF.

6.7. Validation using independent XBT data

XBT data has been extracted from the TOGA/WOCE/CLIVAR data bases on day 184 and every 10 days throughout the assimilation experiment. The number of profiles extracted is 21 on day 184, 25 on day 194, 13 on day 204, 9 on day 214, 26 on day 224, 16 on day 234 and 35 on day 244. Only data measured on the following depths has been considered: 10, 100, 300, 500, 800 and 1000 m. The corresponding model data from the free-run, EnKF, EnKS and SEEK filter assimilation runs has been calculated, by measuring the data in the specific isopycnic layers that are located at each depth and location. XBT–RMSE values were calculated using all the XBT data collected throughout the assimilation period. Fig. 18 (upper) shows the XBT–RMSE values as a function of time for the free-run, EnKF and EnKS simulations. In general, the RMSE values resulting from the assimilation runs are lower than the corresponding values from the free run. The same tendency can clearly be seen in Fig. 18 (lower), which shows XBT–RMSE values for the free-run and SEEK filter simulations. Hence, these results show that the assimilation schemes generally improve the ocean circulation. On day 194, the XBT–RMSE value for the EnKF forecast is noticeably higher than the XBT–RMSE values for the SEEK forecast (but still lower than the free-run XBT–RMSE value). This can be explained by the fact that the random atmospheric forcing in the EnKF forecasting of the ensemble of model states introduces a relatively large modelling error in the temperature field. Hence, the effect of tuning the standard deviation values in the random atmospheric forcing should be further investigated. It should be noted that the number of measurements is limited and in addition the model and XBT data are measured on the same day but not necessarily on the same point of time.

Fig. 19 (upper) shows the location of the XBT data profiles on day 244, while Fig. 19 (lower) shows the RMSE values as a function of depth on this day. The RMSE values, which are plotted for both the free run, the EnKF, EnKS, SEEK runs (both before and after assimilation) indicate that the effect of assimilating SLA and SST is most enhanced in the upper 200–300 m of the ocean. Moreover, the lowest RMSE values and strongest corrections after assimilation are found in the SEEK run on this specific day. However all the assimilation runs typically give smaller RMSE values than the free run on day 244, indicating a positive effect from the assimilation on the ocean circulation.

7. Conclusions

In this work, the SEEK, EnKF and EnKS assimilation schemes have been applied with the ocean general circulation model, MICOM, in the North Atlantic. The multivariate schemes have shown to efficiently update the model state using real satellite measurements of SLA and SST.

The impact of assimilating the ocean surface data is clearly seen, especially on the SLA and SST fields, but also on the salinity fields, subsurface temperature, the thickness of the isopycnic layers and velocity fields. In general, the assimilation of SLA and SST improves the model fields with respect to real observations. Validation against independent XBT in situ data indicates that the model temperature resembles the real data more closely in the assimilation runs than

in the free run. The assimilation schemes are however not able to fully compensate for the model error in high variability areas such as the central Gulf Stream region. One reason can be the relatively coarse model resolution in this area, which makes the model unable to reproduce fine-scale variations in the observations. In order to get more realistic results, it will be needed to primarily improve the model itself by refining the resolution and using more accurate atmospheric forcings.

The general conclusion of this work is, however, that all methods performed well enough to participate to the development of preoperational prototypes and real-time demonstrations. The results of the assimilation experiments support the statement that the three systems have similar performances. However, because the implementations of the EnKF, the EnKS and the SEEK filter were a bit different, and because a good independent data set with global coverage was lacking, a thorough investigation of the relative performances of the methods was not quite possible.

From the technological point of view, we have shown that both methods are feasible for this kind of problem, and both have their merits and disadvantages. From the hindcast experiments, we learned that an ensemble size of about 150 is sufficient to obtain an EnKF that works satisfactorily. The same number is needed for the EnKS. The EnKS is slightly more expensive than the EnKF because some extra matrix–vector operations have to be performed. In contrast, the SEEK filter could be run with a number of error modes of order 10 only. Given the tremendous growth of computer power expected in the coming years, ensemble methods are, thus, perfectly suitable to develop prototypes of ocean monitoring and forecasting systems.

Last but not least, it must be mentioned that ensemble-based data assimilation methods do not require much recoding given the original model code. This means that model updates are easily handled in such systems, showing another strong point in favour of these methods.

In the EU-funded TOPAZ project, the preoperational ocean monitoring and forecasting system established in the DIADEM project will be further developed. Encouraged by the promising data assimilation results in DIADEM, the assimilation system will be subject to further developments with, e.g., assimilation of in situ and buoy data, sea-surface salinity, ice concentration and thickness. Especially, we will in the future focus on the multivariate impact of assimilating these data types using independent data resources to validate the assimilation results.

Acknowledgements

The study was supported by the European Commission through the MAST-III project DIADEM under contract MAS3-CT98-0167. In addition, it has received support from a grant of CPU time from the Norwegian super computing committee (TRU).

References

Anderson, J.L., Anderson, S.L., 1999. A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. Mon. Weather Rev. 127, 2741–2758.

Bleck, R., Boudra, D., 1986. Wind-driven spin-up in eddy-resolving ocean models formulated in isopycnic and isobaric coordinates. J. Geophys. Res. 91, 7611–7621.

Bleck, R., Hanson, H.P., Hu, D., Kraus, R.B., 1989. Mixed layer-thermocline interaction in a three-dimensional isopycnic coordinate model. J. Phys. Oceanogr. 19, 1417–1439.

Bleck, R., Rooth, C., Hu, D., Smith, L.T., 1992. Salinity-driven thermohaline transients in a wind- and thermohaline-forced isopycnic coordinate model of the North Atlantic. J. Phys. Oceanogr. 22, 1486–1515.

Brankart, J.M., Testut, C.E., Brasseur, P., Verron, J., 2003. Implementation of a multivariate data assimilation scheme for isopycnic coordinate ocean models: application to a 1993–96 hindcast of the North Atlantic Ocean circulation. J. Geophys. Res. 108, 3074.

Brasseur, P., Ballabrera, J., Verron, J., 1999. Assimilation of altimetric data in the mid-latitude oceans using the SEEK filter with an eddy-resolving primitive equation model. J. Mar. Syst. 4, 269–294.

Eanes, R.J., Bettadpur, S.V., 1996. The CSR 3.0 Global Ocean Tide Model, CSR-TM-95-06, Center for Space Research, Univ. of Texas, Austin, USA.

Evensen, G., 1992. Using the Extended Kalman Filter with a multilayer quasi-geostrophic ocean model. J. Geophys. Res. 97, 17905–17924.

Evensen, G., 1994. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. J. Geophys. Res. 99, 10143–10162.

Evensen, G., van Leeuwen, P.J., 1996. Assimilation of Geosat altimeter data for the Agulhas current using the ensemble Kalman filter with a quasi-geostrophic model. Mon. Weather Rev. 124, 85–96.

Evensen, G., van Leeuwen, P.J., 2000. An ensemble Kalman smoother for nonlinear dynamics. Mon. Weather Rev. 128, 1852–1867.

Gaspar, P., Grégoris, Y., Lefevre, J.-M., 1990. A simple eddy kinetic model for simulations of the oceanic vertical mixing: tests at Station Papa and Long-Term Upper Ocean Study Site. J. Geophys. Res. 95, 16179–16193.

Gaspar, P., Ogor, F., Traon, P.Y.L., Zanife, O.Z., 1994. An ensemble Kalman smoother for nonlinear dynamics. J. Geophys. Res. 99, 24981–24994.

Higdon, R.L., Bennett, A.F., 1996. Stability analysis of operator splitting for large-scale ocean modeling. J. Comput. Phys. 123, 311–329.

Houtekamer, P., Mitchell, H., 1998. Data assimilation using an ensemble Kalman filter technique. Mon. Weather Rev. 126, 796–811.

Jazwinski, A.H., 1970. Stochastic Processes and Filtering Theory. Academic Press, San Diego, CA.

Miller, R.N., Ghil, M., Gauthiez, F., 1994. Advanced data assimilation in strongly nonlinear dynamical systems. J. Atmos. Sci. 51, 1037–1056.

Natvik, L.-J., Evensen, G., 2003a. Assimilation of ocean colour data into a biochemical model of the North Atlantic: Part 1. Data assimilation experiments. J. Mar. Syst. (this issue).

Natvik, L.-J., Evensen, G., 2003b. Assimilation of ocean colour data into a biochemical model of the North Atlantic: Part 2. Statistical analysis. J. Mar. Syst. (this issue).

Pham, D.T., 2001. Stochastic methods for sequential data assimilation in strongly nonlinear systems. Mon. Weather Rev. 29, 1194–1207.

Pham, D.T., Verron, J., Roubaud, M.C., 1998. Singular evolutive extended Kalman filter with EOF initialization for data assimilation in oceanography. J. Mar. Syst. 16, 323–340.

Smith, L.T., Boudra, D.B., Bleck, R., 1990. A wind-driven isopycnic coordinate model of the North and Equatorial Atlantic Ocean: 2. The Atlantic basin experiment. J. Geophys. Res. 95, 13105–13128.

Traon, P.-Y.L., Nadal, F., Ducet, N., 1998. An improved mapping method of multi-satellite altimeter data. J. Atmos. Ocean. Technol. 15, 522–534.

van Leeuwen, P.J., 2001. An ensemble smoother with error estimates. Mon. Weather Rev. 129, 709–728.

van Leeuwen, P.J., Evensen, G., 1996. Data assimilation and inverse methods in terms of a probabilistic formulation. Mon. Weather Rev. 124, 2898–2913.

Verron, J., Gourdeau, D.T.P.L., Murtugudde, R., Busalacchi, A.J., 1999. An extended Kalman filter to assimilate satellite altimeter data into a nonlinear numerical model of the tropical Pacific Ocean: method and validation. J. Geophys. Res. 104, 5441–5458.