Ocean Sci. Discuss., 6, 617–652, 2009
www.ocean-sci-discuss.net/6/617/2009/

Ocean Science
Discussions

Application of the Gaussian anamorphosis to assimilation in a 3-D coupled physical-ecosystem model of the North Atlantic with the EnKF: a twin experiment

E. Simon and L. Bertino

Nansen Environmental and Remote Sensing Center, Norway

Received: 19 February 2009 – Accepted: 18 March 2009 – Published: 23 March 2009

Correspondence to: E. Simon (ehouarn.simon@nersc.no)

Published by Copernicus Publications on behalf of the European Geosciences Union.

Abstract

We consider the application of the Ensemble Kalman Filter (EnKF) to a coupled ocean ecosystem model (HYCOM-NORWECOM). Such models, especially the ecosystem models, are characterized by strongly non-linear interactions active in ocean blooms and present important limitations for the use of data assimilation methods based on linear statistical analysis. Besides the non-linearity of the model, one is confronted with physical/biological limitations, the analysis state having to be consistent with the model, especially with the constraints of positiveness of some variables. Furthermore the non-Gaussian distributions of the biogeochemical variables break an important assumption of the linear analysis, leading to a loss of optimality of the filter. We present an extension of the EnKF dealing with these limitations by introducing a non-linear change of variables (anamorphosis function) in order to execute the analysis step in a Gaussian space. We present also the initial results of the application of this non-Gaussian extension of the EnKF to the assimilation of simulated chlorophyll surface concentration data in a North Atlantic configuration of the HYCOM NORWECOM coupled model.

1 Introduction

The context of this work lies in the study and the forecast of the dynamics of the ocean and the evolution of its biology. Important economical stakes involve a better optimization of the management of the natural environment, especially by fisheries. So analysis and short term forecasts of the primary production will be more and more useful to environmental agencies for monitoring algal blooms and possible movement of the fish populations. For the particular case of Norway, an important issue is the possible movement of fish populations following the sea-ice retreat from the Norwegian Arctic to the Russian Arctic. Such perspectives have led to the developments of numerical ecosystem models during the last decades, as well as their coupling with existing physical ocean models. These couplings are made either on- or off-line, bring vertical 1-D as well as 3-D physical models in and express the trade-off between our need in term of modelling and forecast and the available computing resources.

Nevertheless these models present numerous uncertainties linked to the complexity of the processes that they try to represent and the parameterization that they introduce. Numerical ocean models are still imperfect and present many errors due to some theoretical approximations, the numerical schemes as well as the resolution that are used. Even though many improvements have been made in the modelling of ocean ecosystem, the models are still too simple in comparison of the complexity of the ocean biology. Finally, the multi-scale interactions between the physics and the biology of the oceans are still little known, leading also to errors and uncertainties in the coupling of both numerical models. So numerical ocean ecosystem models alone are not sufficient for understanding and forecasting the real ocean.

An other source of information lies in the observations of the ocean biology. The use of satellites allowed the community to obtain important informations on the surface biology. So the observed surface ocean color gives informations on the distribution of the surface chlorophyll for a large area of the oceans, and thus the distribution of the phytoplankton. Satellite observations are also dependent of the atmospheric conditions (clouds, etc.), leading to more or less important lack of data in the surface of the ocean. Finally, the observations can present important errors, especially for satellite data near the coast. So errors on surface chlorophyll provided from observations of the surface ocean color is of the order of 30% of the value. In the same way, in situ measurements led to a better understanding of the vertical components of the biological systems in the interior of the ocean. Nevertheless these informations have heterogeneous spatial and temporal distributions. The in situ data networks are still quite poor, mainly localized near the coast, and finally are not able to provide information covering the 3-D global ocean.

Hence the interest of combining in an optimal way the information providing by the models and the observations. This is the objective of data assimilation methods. These methods can be classified in two categories: the stochastic approach based on the theory of the statistical estimation – the Kalman filter (Kalman, 1960) and its extensions – and the variational approach based on the theory of the optimal control (Sasaki, 1955; Lions, 1968; Le Dimet, 1982; Le Dimet and Talagrand, 1986; Courtier et al., 1994). These methods can be applied on important classes of problems: the optimization of parameters of the model conditionally to the observations, the sensitivity analysis of the model (to parameters, observations, etc.) and the state estimation. Both are equivalent for linear systems. Data assimilation methods have been successfully applied in the fields of meteorology and physical oceanography and some of them are now used for operational forecast. Nevertheless their application in ecosystem forecasting is quite recent: they have started to be applied on ecosystem model mainly this last decade. Furthermore, the use of biological observations could be relevant to improve the forecast of the physical model, leading to a real interest for coupled ocean-biogeochemical models.

Data assimilation methods based on the Kalman filter have been successfully applied in numerous cases. In 1-D vertical ocean ecosystem models, real biological in situ data have been assimilated with an Ensemble Kalman Filter (EnKF) by Allen et al. (2003), Torres et al. (2006). In 3-D ocean ecosystem models, twin experiments of assimilation of simulated satellite surface chlorophyll data with a SEEK filter (Pham et al., 1998) in a North Atlantic configuration have been done by Carmillet et al. (2001). Twin experiments of assimilation of simulated in situ data with a SEIK filter (Pham, 2001) in the Cretan Sea have been realized by Triantafyllou et al. (2003). Finally for realistic experiments in 3-D ocean ecosystem models, we can note the works of Natvik and Evensen (2003a,b), who have assimilated successfully real surface ocean color data with an EnKF over a short period (2 months) in a North Atlantic configuration, the works of Nerger and Gregg (2007) who assimilated SeaWiFS data with a monovariate SEIK filter in a global ocean configuration, and those of Gregg (2008) who demonstrated the capabilities of a monovariate assimilation of SeaWiFS data with a simple method (Conditional Relaxation Scheme Method) over long periods. For a more important overview of works dealing with the problem of data assimilation in ocean ecosystem model, we refer to Gregg et al. (2009).

We focus in this present paper on the application of the EnKF (Evensen, 1994, 2003, 2006) for state estimation in coupled ocean ecosystem models. Such models present important practical and theoretical limitations for the application of data assimilation methods based on linear statistical analysis.

On the one hand, the strongly nonlinear behavior of ecosystem models (especially during the period of the Spring bloom) raises the question of the stochastic model to be used (Bertino et al., 2003).

On the other hand one is also confronted with physical/biological limitations: the analysis state has to be consistent with the model, especially under the constraints of positiveness of some variables. Most variables of ecosystem models are concentrations of a given tracer, and so cannot be negative. Nevertheless this problem is also known for the assimilation in physical ocean models. One thinks for example to the correction of layer thickness while assimilating data in hybrid coordinates model (HYCOM). Several solutions have been suggested to deal with such problems. The one of Thacker (2007) introduces inequality constraints via Lagrange multipliers, leading to a 2-passes 3D-Var. Such approach can also be applied to a Kalman filter. Into the framework of stochastic methods, Lauvernet et al. (2009) develops a truncated Gaussian filter with inequality constraints. But positiveness is only one example of non-Gaussianity among many others. We focus here a more general approach to non-Gaussianity.

Finally the non-Gaussian distributions of most biogeochemical variables break an important assumption of the linear analysis, leading to a loss of optimality of the EnKF (and other filters). The optimality of the linear statistical analysis is proved under some assumptions, notably an assumption of Gaussianity made on the distribution of the variables (of the model and the observations) and the errors.

A way to deal with these two last limitations is the introduction of anamorphosis functions in the filter, as suggested by Bertino et al. (2003). They presented an EnKF in which they introduce non-linear changes of variables (anamorphosis function) in order to realize the analysis step in a Gaussian space. Numerical experiments with a

1-D ocean ecosystem model led to promising results. The present paper comes within the continuity of these works and deals with the application of this extension of the EnKF in a more realistic 3-D ocean ecosystem models.

The outline of the paper is as follows. We present the EnKF with Gaussian anamorphosis and a way to build a monovariate anamorphosis function in Sect. 2. We describe our experimental framework in Sect. 3. Results of the methods are discussed in Sect. 4, and we present our conclusion in Sect. 5.

2 The Ensemble Kalman filter with Gaussian anamorphosis

We describe in this section the algorithm of the EnKF with Gaussian anamorphosis suggested by Bertino et al. (2003). The principle is simple and consists in introducing non-linear changes of variables in order to realize the analysis step in a "Gaussian" space, while the forecast step is realized in the physical space.

The main benefit of such algorithm is to remove in one pass two important limitations of the application of linear statistical analysis scheme in ecosystem models (described in introduction). The assumption of Gaussian distribution of the variables appears now to be relevant for the transformed variables during the analysis step. Furthermore there is no "physical" limitation (constraint of positiveness, etc.) on the transformed variables during the analysis, removing post-processing steps that are compulsory when the analysis state vector is not consistent with the physical model.

2.1 Algorithm

The algorithm is based of the skeleton of the EnKF and divides into two steps:

Forecast: the forecast step corresponds to the one of the EnKF. It is a propagation step that uses a Monte Carlo sampling to approximate the forecast density by N realizations:

$$\forall i \in \mathbb{N}_N, \quad \mathbf{x}_n^{f,i} = f_{n-1}(\mathbf{x}_{n-1}^{a,i}, \epsilon_n^{m,i}) \qquad (1)$$

with $\mathbf{x}_n$ the state vector at time t_n, f_{n-1} the nonlinear model and ϵ_n^m the model error.

Analysis: the analysis step conditions each forecast member to the new observation $\mathbf{y}_n$ by a linear update. The anamorphosis functions are introduced in this step.

For each variable of the model, at time t_n, we apply a function ψ_n which is a nonlinear bijective function from the physical space to a Gaussian space. We treat each variable separately. In order to simplify the notations, we assume that we have one variable in our model (so one function ψ_n). It reads:

$$\forall i \in \mathbb{N}_N, \quad \tilde{\mathbf{x}}_n^{f,i} = \psi_n(\mathbf{x}_n^{f,i}) \tag{2}$$

In practice, it means that we apply the changes of variables for each variables in every point of the discretized domain.

In the same way, we introduce an anamorphosis function χ_n for the observations $\mathbf{y}_n$ at time t_n:

$$\tilde{\mathbf{y}}_n = \chi_n(\mathbf{y}_n). \tag{3}$$

Given the observation operator $\mathbf{H}$ linking the physical variables and the observations. We define the observation operator $\tilde{\mathbf{H}}_n$ linking the transformed variables and observations by the formula

$$\tilde{\mathbf{H}}_n = \chi_n \circ \mathbf{H} \circ \psi_n^{-1} \tag{4}$$

By assuming that $\tilde{\mathbf{H}}_n$ is linear, the linear analysis equation in the Gaussian space reads formally as the classical linear analysis equation:

$$\forall i \in \mathbb{N}_N, \quad \tilde{\mathbf{x}}_n^{a,i} = \tilde{\mathbf{x}}_n^{f,i} + \tilde{\mathbf{K}}_n(\tilde{\mathbf{y}}_n - \tilde{\mathbf{H}}_n\tilde{\mathbf{x}}_n^{f,i} + \epsilon_n^{o,i}) \tag{5}$$

with $\tilde{\mathbf{K}}_n$ the classical Kalman gain matrix in the Gaussian space and $\epsilon_n^{o,i}$ the observation errors in the Gaussian space which follow a Normal law $(\epsilon_n^{o,i} \sim \mathcal{N}(0, \tilde{\Sigma}^o))$.

The return to the physical space is realized by using the inverse of the anamorphosis function:

$$\forall i \in \mathbb{N}_N, \quad \mathbf{x}_n^{a,i} = \psi_n^{-1}(\tilde{\mathbf{x}}_n^{a,i}) \tag{6}$$

The analyzed mean $\mathbf{x}_n^a$ and the covariance matrix $\mathbf{C}_n^a$ are approximated by the ensemble average and covariance of $(\mathbf{x}_n^{a,i})_{i\in\mathbb{N}_N}$.

Remarks

1. The construction of relevant anamorphosis functions χ_n and ψ_n is not straightforward. Analytic functions as log or Cox-Box can be used for variables which have initially a "good" distribution, but are not guaranteed to improve the distribution in general case. A more general way to build relevant anamorphosis function can be obtained from the empirical marginal distribution. More details about their constructions are given afterwards.
2. The use of nonlinear functions may introduce non linearities on the transformed observation operator $\tilde{\mathbf{H}}$. In some practical cases, a "good" choice of $\mathbf{H}_n$ and χ_n leads to a linear operator. In the case when $\mathbf{H}_n$ is extracting measurements from the state vector, this is not an issue. So it can not be guaranteed for general cases. For a nonlinear $\tilde{\mathbf{H}}$, we suggest to use the EnKF analysis scheme for nonlinear measurements suggested by Evensen (2003, 2006).
3. This algorithm based of the use of monovariate anamorphosis functions does not handle multivariate non-Gaussianity of the state vector. Even if each transformed variables follows a Gaussian distribution, their bivariate (and more generally their multivariate) distributions will not be necessary bi-Gaussian (resp. multi-Gaussian). In practice this property is really difficult to check due to the large size of the vectors. We assume that the improvements of the monovariate distributions will improve the multivariate distribution. More sophisticated transformations should be investigated in the future (see Schölzel and Friedrichs, 2008).

2.2 Construction of a monovariate anamorphosis function

The performances of the extended EnKF described above are strongly dependent of the choice of the anamorphosis functions ψ_n and χ_n. Several strategies can be applied to the construction of functions that improve the Gaussianity of the distribution of the variables. A first idea is to use "classical" analytic function as the logarithm function or the Cox-box functions.

Rather than using analytic functions that need to have a prior knowledge of the distribution of variables, we construct the anamorphosis functions directly from a sample of variables.The idea is to build the anamorphosis functions from the empirical marginal distributions of the variables. For that we assume that the variables at different locations and on a limited time period are identically distributed conditionally to the past observations and the physics. The algorithm of the construction of a monovariate anamorphosis function (one function per variable) divides into three parts:

1. *Construction of the experimental anamorphosis function based on the empirical marginal distribution*. Such function and the way to build it are well known in the geostatistical community. More details can be found in Chilès and Delfiner (1999). The computational costs of this step are negligible in comparison with the costs of forecast steps in the EnKF.

2. *Interpolation of the experimental anamorphosis function*. Classical polynomial interpolations can be used. Nevertheless high order polynomial interpolations generate oscillations (close to the extrema of the empirical anamorphosis) that need a particular treatment when defining the tails of the monotonic function. We choose linear interpolation instead.

3. *Definition of the tails of the function*. It is an important step due to the fact that one defines the bounds of the physical variables. The definition of the physical bounds is the way to introduce the physical constraints of the model (for example a minimum value equal to zero will correspond to a constraint of positiveness). For the bounds of the Gaussian space, one has to take unlikely high values of the analysis into account which leads to extend the tails towards infinity.

These three steps of the construction of the anamorphosis function for the chlorophyll-*a* variable are summarized Fig. 1.

Remarks

1. The anamorphosis function of a Gaussian variable is linear.
2. The anamorphosis functions as constructed here are designed for continuous distributions and may not improve "pathological" distributions such as Dirac or bimodal.
3. Without Monte-Carlo sampling the introduction of nonlinear functions in order to realize the linear analysis estimation in an other space can lead to an assimilation bias as follows.

   $$\mathrm{E}[\psi_n^{-1}(\tilde{\mathbf{x}}_n^a)] \neq \psi_n^{-1}(\mathrm{E}[\tilde{\mathbf{x}}_n^a]) \qquad (7)$$

   The bias only has an explicit expression in a few particular cases, like the exponential. One general way to avoid the bias is to sample randomly the forecast distribution. In the EnKF, this sampling is realized by using an ensemble during the forecast step. Nevertheless for the other methods as Ensemble Optimal Interpolation (EnOI) or Extended Kalman Filter (EKF), samplings are compulsory.
4. We assume that the variables at different locations in space are identically distributed. In practice, this assumption can been not checked for localized events, leading to a loss of relevance of anamorphosis functions. The spatial refinements of these functions is still an open issue and has to be investigated.

3 Description of the experimental framework

3.1 The coupled ocean ecosystem model

The experiments were performed in a North Atlantic and Arctic configuration of the HYCOM-NORWECOM coupled model. We describe briefly this configuration, which corresponds to the coarse resolution one in Hansen and Samuelsen (2009).

The domain of the model cover the North Atlantic and the Arctic oceans from 30° S. The grid was created using the conformal mapping algorithm outlined in Bentsen et al. (1999).

The physical model used is the HYbrid Coordinate Ocean Model, HYCOM, (Bleck, 2002). The vertical coordinates are isopycnal in the open, stratified ocean, and change to z-level coordinates in the mixed layer and/or unstratified seas. The model uses 23 layers with the minimum thickness of the top layer of 3 m. The model presents 216×144 horizontal grid points which leads to an horizontal resolution of 50 km. This is sufficient to resolve broadly the large-scale circulation.

The evolutions of the ice cover in the North part of the domain (mainly in the Arctic Ocean) is taken into account by an on-line coupling between the physical ocean model and an ice module including a thermodynamic model (Drange and Simonsen, 1996) and a dynamic model (using the elastic-viscous-plastic rheology of Hunke and Dukowicz, 1999). Finally the ERA40 synoptic fields and climatological river runoff (without nutrient) are used to force the model.

The ecosystem model is the NORWegian ECOlogical Model system, NORWECOM (Skogen and Søiland, 1998; Aksnes et al., 1995). This model includes two classes of phytoplanktons (diatom and flagellates), several classes of nutrients, and the classes of oxygen, detritus, inorganic suspended particulate matter (ISPM) and yellow substances. Nevertheless in our experiments ISPM and yellow substances were not activated. So the ecosystem state vector is made of 7 variables.

This configuration is illustrated Fig. 2 by a snapshot of surface chlorophyll-*a* on 22 October 1997.

3.2 Data assimilation experiments

We focus on data assimilation in the ecosystem model. The multivariate assimilation of both physical and biological states is a challenging work and is still an open issue. The state vector corresponds to the ecosystem state vector only, namely seven 3-D variables. Due to the lack of feedback in the coupling from the ecosystem model to the physical one, it means that the assimilation does not correct the ocean physical state.

Our aim is to compare the performances of the extended EnKF with Gaussian anamorphosis to those of a "classical" EnKF. In that way twin experiments have been realized: the true state and the observations are issued from a simulation of the coupled model. The benefits of such framework is the knowledge of all the component of the solution which leads us to check the impact of the assimilation, in space as well as in time, over all the variables of the model.

Two assimilation systems have been implemented in the same configuration described bellow. The first one called ECO corresponds to the direct application of the EnKF. A post-processing step is added to remove negative values as well as too important values: negative values are increased to zero while unlikely high values are replaced by an arbitrary upper bound (this value corresponds to the physical maximum bound introduced in the construction of the anamorphosis functions, cf. Table 1). The second one called ANA corresponds to the application of the EnKF with Gaussian anamorphosis. No post-processing step is included, as the method does not require any.

The temporal linking of the experiments is as follows. Started from an already spun-up simulation at the date of 10 July 1997, the true state is generated by running the model without perturbation, while the ensemble is generated by running the same model with perturbations (more details about the generation of the ensemble come below). This simulation is issued from works of Hansen and Samuelsen (2009) and corresponds to the results of a spun-up started in 1958. At this date the spring bloom is at a late stage and the concentration of phytoplankton starts to decrease. Then data assimilation is included as from 24 September 1997. At this date the spring bloom is over and the global concentration of phytoplankton is low and decreases. Assimilation cycles are then performed over one year with a frequency of one analysis step per week.

The synthetic observations are the surface chlorophyll-*a* obtained by a spatial sampling of the noised true state (Eq. 8) every third grid index. Furthermore the observations under ice or too close to coasts (the depth of the water column must be greater than 300 m) are not assimilated in order to take into account several constraints of the assimilation of realistic satellite data. Finally the observations present in the southern boundary area (last 15 grid points in the y-direction) are not assimilated either, nor are the observations present in the Arctic ocean (first 50 grid points in the y-direction). It leads to a time evolutive network of observations illustrated Fig. 3 on 31 December 1997.

The observations are defined as follows

$$\mathbf{y}_n = \mathbf{H}_n\mathbf{x}_n^t \times e^{(Z_n-\sigma^2/2)} \qquad (8)$$

with $Z_n \sim \mathcal{N}(0, \sigma{=}0.2)$. It means that we construct the observations by adding to the true surface chlorophyll-*a*, which is assumed having a lognormal distribution, an observation error around 6–7%, which is lower than the "usual" error of real satellite data (around 30%). Nevertheless such value increases the difficulties for both systems to assimilate the observations. $\frac{\sigma^2}{2}$ is a bias reduction term.

The initial ensemble as from 24 September 1997 is the same for both systems (ECO and ANA). It is made up of 100 members obtained by running the model from 10 July 1997 with perturbations of the atmospheric fields in the physical model only (as done in Natvik and Evensen, 2003a). These random perturbations are generated by a spectral method (Evensen, 2003) in which the residual error is simulated using a spatial decorrelation radius of 250 km. The decorrelation time-scale is of five days. The standard deviations of the fields perturbed are: 0.03 $\mathrm{N.m^{-2}}$ for the eastward and northward drag coefficient, $\sqrt{2.5}\,\mathrm{m\,s^{-1}}$ for the wind speed, $\sqrt{0.005}\,\mathrm{W\,m^{-2}}$ for the radiative fluxes and 3° Celsius for the air temperature. These values correspond to the ones use in the TOPAZ operational forecast and monitoring system (Bertino and Lisæter, 2008).

Finally both systems use localization as suggested by Evensen (2003). The radius is constant and equal to 500 km (10 cell-grids in the two horizontal directions) which leads to assimilate in each point between 2 and 10 observations depending on the area. The aim of this work being the comparison of the intrinsic behavior of the two assimilation systems, we have not introduced advanced operational processes as the decrease of the radius close to the coast for example, in order to have a better understanding of the benefits of anamorphosis functions.

3.3 Construction of the monovariate anamorphosis functions

We assume that each variable and the chlorophyll-*a* at different locations in space are identically distributed in a time period of three months centered on the datum of the analysis step. In that way we obtain time evolutive anamorphosis functions. The choice of three months is motivated by the time scale of bloom phenomenon which is about 4 months. Such moving windows allow to catch the differences of distribution at the beginning and the end of the bloom in the construction of the anamorphosis functions.

The experimental anamorphosis functions are computed from weekly outputs from a four years integration of the model. The anamorphosis function is piecewise linear, using linear interpolation of the experimental anamorphosis function. The tails of the anamorphosis are defined as follows:

- Physical bounds: the minimum values are equal to zero (constraint of positiveness) and the maximum values are unlikely high values summarized in Table 1.
- Gaussian bounds: the minimum values are equal to -9 (value with a probability around 1×10^{-19}). We do not define maximum values, the right tails extending towards infinity.

Remark

In case of model bias (which would occur with assimilation of real data), the model-based anamorphosis functions may be impaired by the bias, especially when using a short moving window. For example, the main bloom could be modeled too early or too late by a couple of weeks, which would make high concentrations of plankton too likely or too unlikely at different stages of the bloom. Thus the moving time window should be shorter than the bloom, but not too short by comparison to usual ecosystem model delays. We consider three months as a reasonable compromise.

The results of the application of anamorphosis functions on the distribution of the diatoms (phytoplankton) and the silicate (nutrient) are shown in Fig. 4 during three periods of the year: in winter (31 December 1997) when the primary production is low, during the Spring bloom (14 May 1998) and in fall (3 September 1998) when the concentration of phytoplankton decreases slowly. In this present study, we focus particularly on diatoms which are linked to the chlorophyll-*a* (observation) by a linear relation and on the silicate which is rate limiting for the production of diatoms but not for flagellates (second phytoplankton class of the model).

First we note that the time evolutive anamorphosis functions provide more Gaussian distributed variables as expected. This is globally true for the other variables of the ecosystem model (not shown). Nevertheless the histogram of the transformed diatoms during the Spring bloom let appear the superimposition of two gaussians. It can be explained by the bloom in the Eastern part of the North Atlantic (mainly off Spain) in the ensemble which is earlier than the blooms present in the data set used for building the anamorphosis functions. So it means that we reach the problem of the bias of anamorphosis functions based on moving windows. A way to deal with this problem would be to include more extreme events in the data set used for the construction of the anamorphosis functions.

4 Data assimilation results

4.1 Overall error evolution

At first we are interested in the evolution in time of the true Root Mean Square error (RMS) and the ensemble standard deviations (STD) of the solution of the two systems.

The expression at time t_n of these two quantities is as follows:

$$\begin{aligned}\mathrm{RMS}(t_n) &= \sqrt{\frac{1}{\#\Omega}\sum_{\mathbf{k}\in\Omega}(\mathbf{x}^t(t_n,\mathbf{k})-\bar{\mathbf{x}}(t_n,\mathbf{k}))^2}\\ \mathrm{STD}(t_n) &= \sqrt{\frac{1}{N-1}\frac{1}{\#\Omega}\sum_{\mathbf{k}\in\Omega}\sum_{m=1}^{N}(\mathbf{x}^m(t_n,\mathbf{k})-\bar{\mathbf{x}}(t_n,\mathbf{k}))^2}\end{aligned} \qquad (9)$$

with Ω the domain of computation, $\#\Omega$ the number of grid points of the domain used for the computation of the RMS and STD, N the number of members, $\mathbf{x}^t$ the true state, and $\bar{\mathbf{x}}$ the mean of the ensemble.

Figure 5 represents the evolution of the RMS error and the standard deviations over one year for the surface chlorophyll-*a* (what we observe). In that case Ω is the top layer of the model. We note that the RMS error of the ECO configuration (plain EnKF with a simple post-processing) is clearly more important than the RMS error of the solution issued from the ANA configuration, especially during the period of the Spring bloom. For that period, the ECO error on the surface chlorophyll-*a* reaches a maximum of $7\,\mathrm{mg\,m^{-3}}$ (same order of the solution) when the ANA error reaches $0.4\,\mathrm{mg\,m^{-3}}$.

Furthermore we observe three phases in the evolution of the curves. The first one corresponds to the end of the bloom and the winter (October 1997–March 1998). During that phase, the analysis steps mainly damage the solution without the use of anamorphosis functions, as indicated by the positive increments of RMS at each assimilation step. We note also that the standard deviation is longer than the RMS error for both systems, expressing an over-estimation of the error by the filters. It remains also the case all over the year for the ANA system. The second phase corresponds to the Spring bloom. Without anamorphosis function, the RMS error and the standard deviation strongly increase. The bloom is too important as well as too early. The RMS error and the standard deviation increase also with the use of anamorphosis function. Nevertheless their values are much lower than the ones of the ECO configuration. Furthermore the lack of observations in shallow waters leads to some difficulties to correct the solution in several areas (cf. Sect. 4.3). Finally the third phase corresponds to the end of the bloom. The RMS error and the standard deviation decrease slowly to only reach their initial values in the ANA experiment, but the final errors are twice larger at the end of the ECO experiment.

4.2 Errors in the sub-surface

In order to explore the multivariate aspect of the data assimilation, we focus on the evolution of the RMS error and the standard deviation, computed on one only point $(-58.8^\circ, 38.7^\circ)$ in the area of the Gulf Stream, for the diatoms and the silicate. This point, called P_8 and localized by a black cross on Fig. 6, is in the 8th layer (waters between 30 m and 38 m) of the model, the deepest one locally before vanishing of the diatoms. As the concentrations of diatoms at this point can change quickly with time, it is a good indicator of the front of structures.

Once again we note that the introduction of anamorphosis functions allows an important reduction of the RMS error in comparison with the assimilation using a plain EnKF both for the diatoms (Fig. 7) and the silicate (Fig. 8).

The three previous phases appear clearly on the ECO configuration but the error is 10 to 100 times lower with the use of Gaussian anamorphosis. We observe again that the analysis steps of the plain EnKF often also occur damages to the solution, but it is also the case with anamorphosis functions. Although the use of the anamorphosis improves spectacularly the solution in sub-surface, it does not guarantee the update will always reduce the error.

4.3 Regional distribution of the errors

We examine the spatial localization of the error on the surface chlorophyll-*a*. Figures 9, 10 and 11 represent the maps of the surface chlorophyll-*a* component of $\bar{\mathbf{x}}^a - \mathbf{x}^t$ on 31 December 1997, 14 May 1998 and 3 September 1998.

On 31 December, we first note an important error along the North American coast in the ECO configuration. Analysis steps produce too important quantities of phytoplankton which are then advected on the shelf. Furthermore we observe a slight positive error due to the post-processing steps inside the domain. By removing negative values, these steps increase artificially the mean of the ensemble for each variable. This increases the primary production in area where the concentration of phytoplankton should be null. The combination of these two biases during the winter time is responsible of the strong and too early bloom observed Fig. 5. With the use of anamorphosis functions, we note a slight error in the Gulf of Mexico but the assimilation does not damage the solution along the North American coast. Finally as said previously, the observations present in the southern boundary area are not assimilated, that's why important errors remain in this part of the domain.

On 14 May, during the Spring bloom, we note a very important error (in the order of the mean value) from the North American coast to Europa in the solution issued from the ECO configuration. The bloom is too important in this part of the domain and the data assimilation can not remove completely the residual errors. This is not the case for the solution issued from the ANA configuration. The error is more important than in winter, but remains smaller than the error of the ECO solution. Furthermore the lack of observations on the European North West Shelf leads to important persistent errors in the North Sea (between UK and Norway) for both configurations.

After the Spring bloom, on 3 September, we observe errors in a chlorophyll-*a* structure localized in the Sub-Arctic Gyre for both configurations, these errors being less important for the ANA experiment. We note also an important error for the ECO configuration in the North Sea and the Barents Sea where no observations are present. These important errors are the direct consequences of the too strong bloom and the advection of important quantities of phytoplankton and nutrients that have not been totally corrected. Finally we observe the remains of important errors in shallow waters (close to the North American coasts and in the Channel and Irish Seas) in the experiment without the introduction of anamorphosis functions.

4.4 Occurrence of outliers

In the time period of fall/winter 1997–1998, the ECO analysis steps damage the surface solution. On 22 October 1997, datum of the first wrong analysis step, the mean of the ECO forecast ensemble shows important abnormal surface chlorophyll-*a* structures in shallow waters along Florida which are not present in the ANA mean forecast ensemble. Scatter plots (not shown) of surface model diatoms against surface chlorophyll-*a* observations in these areas reveal the presence of outliers in the ECO ensemble, to which the ensemble correlations are sensitive, leading to unphysical updates and the deterioration of the solution.The same scatter plots (not shown) with the transformed ANA forecast ensemble and the transformed observations do not reveal the presence of outliers, and rather indicate the variables are uncorrelated at this location and time. So the use of anamorphosis functions allows to reduce the probability of the presence of outliers by reshaping the ensemble clouds following a normal law when a long-tailed distribution (such as that of diatoms or chlorophyll-*a*) is reshaped into a Gaussian.

Nevertheless outliers may appear locally in areas where the limitations of the algorithm are reached. So that the scatter plots (not shown) of transformed model diatoms in the point P_8 (vertical limit of the vanishing of phytoplankton) against transformed surface observations reveal the presence of outliers. Refinements in space of the anamorphosis functions as well as more sophisticated transformations (in order to handle multivariate non-Gaussianity) should be investigated.

5 Conclusions

A twin experiment has been conducted with a realistic coupled physical-ecosystem model of the North Atlantic and Arctic Oceans, assimilating simulated surface chlorophyll-*a* with an EnKF, with and without Gaussian anamorphosis.

The study reveals that applying the plain EnKF with a simple post-processing of negative values leads to several assimilation biases. First we note an underestimation of the error by the dynamic filter leading to an incomplete control and a filter divergence. We observe also too high residual errors, about 90% of the error being directly or indirectly caused by assimilation biases. Finally we note a too early start of the Spring bloom, especially in shallow waters and a too late persistence of the bloom in fall.

These assimilation biases are not obvious effects of the post-processing of negative values and could have been erroneously blamed on the model or on observations in operational settings. The Gaussian anamorphosis can efficiently remedy for these biases and can be applied operationally for negligible additional costs.

The Gaussian anamorphosis is by no means reserved to the EnKF but is naturally applied there because of Monte-Carlo formalism. It could be applied in a non-Monte-Carlo method provided that a random sampling is performed before the analysis step.

The assimilation of real satellite data with the EnKF with Gaussian anamorphosis has now to be investigated. It raises the challenging problem of model bias, well known in the data assimilation community, and particularly crucial for the use of anamorphosis functions built on the empirical marginal distributions of model variables. Furthermore two limitations of the algorithm have been reached during these experiments: the first one concerns the assumption on an identical spatial distribution of the variables in the construction of the anamorphosis functions and the second one concerns the monovariate aspect of the algorithm. Works on the refinements in space of the anamorphosis functions or on multivariate transformations would allow a practical improvement of the algorithm.

Acknowledgements. This study has been funded by the eVITA-EnKF project from the Research Council of Norway and the MERSEA IP from the European Commission's 7th FP. A grant of CPU time from the Norwegian Supercomputing Project (NOTUR) has been used. The authors are thankful to C. Hansen and K. A. Lisæter for providing the HYCOM-NORWECOM system.

References

Aksnes, D., Ulvestad, K., Baliño, B., Berntsen, J., and Svendsen, E.: Ecological modelling in coastal waters: towards predictive physical-chemical-biological simulation models, Ophelia, 41, 5–36, 1995. 627

Allen, J. I., Eknes, M., and Evensen, G.: An Ensemble Kalman Filter with a complex marine ecosystem model: hindcasting phytoplankton in the Cretan Sea, Ann. Geophys., 21, 399–411, 2003, http://www.ann-geophys.net/21/399/2003/. 620

Bentsen, M., Evensen, G., Drange, H., and Jenkins, A. D.: Coordinate transformation on a sphere using conformal mapping, Mon. Weather Rev., 127, 2733–2740, 1999. 627

Bertino, L., Evensen, G., and Wackernagel, H.: Sequential Data Assimilation Techniques in Oceanography, International Statistical Reviews, 71, 223–241, 2003. 621, 622

Bertino, L. and Lisæter, K. A.: The TOPAZ monitoring and prediction system for the Atlantic and Arctic Oceans, J. Operational Oceanogr., 1, 2, 15–19, 2008. 629

Bleck, R.: An oceanic general circulation model framed in isopycnic-cartesian coordinates, Ocean Model., 4, 55–88, 2002. 627

Carmillet, V., Brankart, J.-M., Brasseur, P., Drange, H., Evensen, G., and Verron, J.: A singular evolutive extended Kalman filter to assimilate ocean color data in a coupled physical-biogeochemical model of the North Atlantic ocean, Ocean Model., 3, 167–192, 2001. 620

Chilès, J.-P. and Delfiner, P.: Geostatistics: Modeling Spatial Uncertainty, Wiley, New York, 1999. 625

Courtier, P., Thépaut, J. N., and Hollingsworth, A.: A strategy for operational implementation of 4D-Var, using an incremental approach, Q. J. Roy. Meteor. Soc., 120, 1367–1387, 1994. 620

Drange, H. and Simonsen, K.: Formulation of air-sea fluxes in the ESOP2 version of MICOM, NERSC Report 125, Nansen Environmental and Remote Sensing Center, Norway, 1996. 627

Evensen, G.: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics, J. Geophys. Res., 99(C5), 10143–10182, 1994. 621

Evensen, G.: The Ensemble Kalman filter: theorotical formulation and practical implementation, Ocean Dynam., 53, 343–367, 2003. 621, 624, 629

Evensen, G.: Data Assimilation, The Ensemble Kalman Filter, Springer, 2006. 621, 624

Gregg, W. W.: Assimilation of SeaWiFS ocean chlorophyll data into a three-dimensional global ocean model, J. Mar. Syst., 69, 205–225, 2008. 620

Gregg, W. W., Friedrichs, M. A. M., Robinson, A. R., Rose, K. A., Schlitzer, R., Thompson, K. R., and Doney, S. C.: Skill assessment in ocean biological data assimilation, J. Mar. Syst., 76, 1–2, 16–33, 2009. 621

Hansen, C. and Samuelsen, A.: Influence of horizontal model grid resolution on the simulated primary production in an embedded primary production model in the Norwegian Sea, J. Mar. Syst., 75, 1–2, 236–244, 2009. 626, 628

Hunke, E. and Dukowicz, J.: An elastic-viscous-plastic model for sea-ice dynamics, J. Geophys. Res., 27, 1849–1867, 1999. 627

Kalman, R. E.: A new approach to linear filtering and prediction problems, Trans. ASME Ser. D, J. Basic Eng., 82D, 35–45,1960. 620

Lauvernet, C., Brankart, J.-M., Castruccio, F., Broquet, G., Brasseur, P., and Verron, J.: A truncated Gaussian filter for data assimilation with inequality constraints: application to the hydro-static stability condition in ocean models, Ocean Model., 27, 1–17, 2009. 621

Le Dimet, F.-X.: A general formalism of variational analysis, CIMMS Report, 1982. 620

Le Dimet, F.-X. and Talagrand, O.: Variational algorithms for analysis and assimilation of meteorological observations: Theorotical apects, Tellus, 38A, 97–110, 1986. 620

Lions, J. L.: Contrôle optimal des systèmes gouvernés par des équations aux dérivées partielles, Dunod, Paris, 1968. 620

Natvik, L. J. and Evensen, G.: Assimilation of ocean colour data into a biochemical model of the North Atlantic. Part 1. Data assimilation experiments, J. Mar. Syst., 40–41, 127–153, 2003. 620, 629

Natvik, L. J. and Evensen, G.: Assimilation of ocean colour data into a biochemical model of the North Atlantic. Part 2. Statistical analysis, J. Mar. Syst., 40–41, 155–169, 2003. 620

Nerger, L. and Gregg, W. W.: Assimilation of SeaWiFS ocean chlorophyll data into a global ocean model using a local SEIK filter, J. Mar. Syst., 68, 237–254, 2007. 620

Pham, D. T., Verron, J., and Roubaud, M.-C.: A singular evolutive extended Kalman filter for data assimilation in oceanography, J. Mar. Syst., 16, 3–4, 323–340, 1998. 620

Pham, D. T.: Stochastic methods for sequential data assimilation in strongly nonlinear systems, Mon. Weather Rev., 129(5), 1194–1207, 2001. 620

Sasaki, Y.: A fundamental study of the numerical prediction based on the variational principle,

J. Meteor. Soc. Japan, 33, 262–265, 1955. 620

Schölzel, C. and Friederichs, P.: Multivariate non-normally distributed random variables in climate research – introduction to the copula approach, Nonlin. Processes Geophys., 15, 761–772, 2008, http://www.nonlin-processes-geophys.net/15/761/2008/. 624

Skogen, M. and Søiland, H.: A user's guide to NORWECOM v2.0. The NORWegian Ecological Model system, Technical Report Fisken og Havet 18, Institute of Marine Research, Norway,1998. 627

Thacker, W. C.: Data assimilation with inequality constraints, Ocean Model., 16, 264–276, 2007. 621

Torres, R., Allen, J. I., and Figueiras, F. G.: Sequential data assimilation in an upwelling influenced estuary, J. Mar. Syst., 60, 317–329, 2006. 620

Triantafyllou, G., Hoteit, I., and Petihakis, G.: A singular evolutive interpolated Kalman filter for efficient data assimilation in a 3-D complex physical-biogeochemical model of the Cretan Sea, J. Mar. Syst., 40–41, 213–231, 2003. 620

Table 1. Anamorphosis functions: maximal physical bounds.

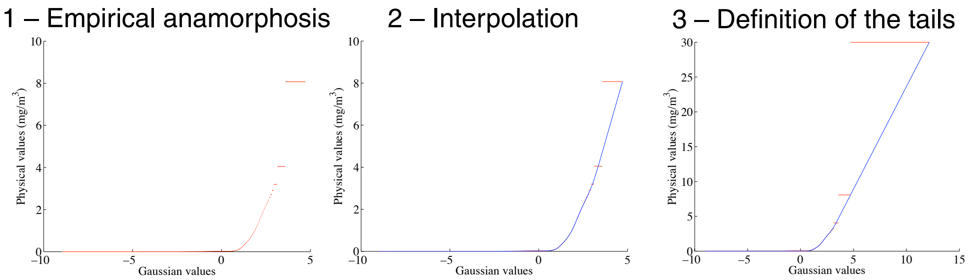
Variables	NIT	PHO	SIL	DET	SIS	FLA	DIA	CHLA
$\mathrm{mg\,m^{-3}}$	1000.	210.	4000.	100.	200.	150.	150.	30.

OSD

6, 617–652, 2009

Gaussian anamorphosis in a 3-D ecosystem model

E. Simon and L. Bertino



1 – Empirical anamorphosis

2 – Interpolation

3 – Definition of the tails

Fig. 1. Surface chlorophyll-*a* observations: the steps of the construction of a monovariate anamorphosis function.

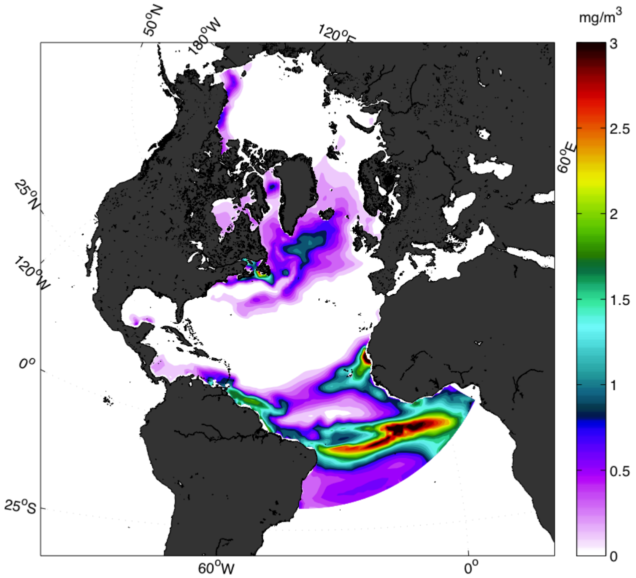


Fig. 2. Arctic and North Atlantic configuration: surface chlorophyll-*a* concentration (mg/m^3) on 22 October 1997.

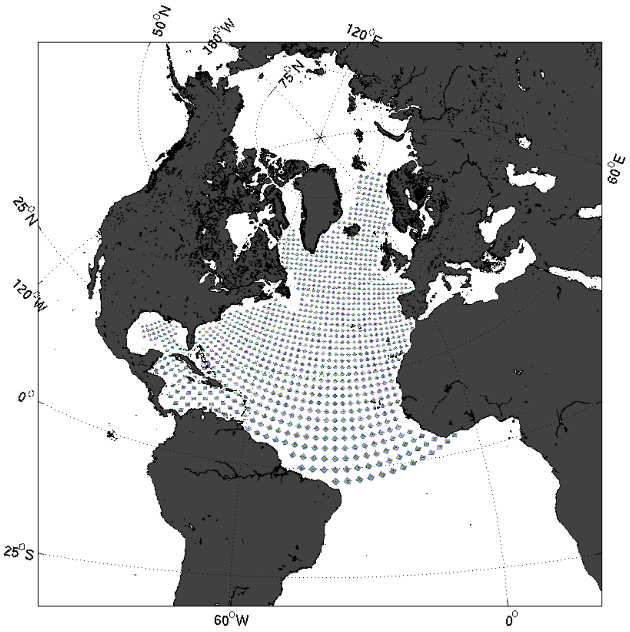


Fig. 3. Surface chlorophyll observations: network of available observations on 31 December 1997.

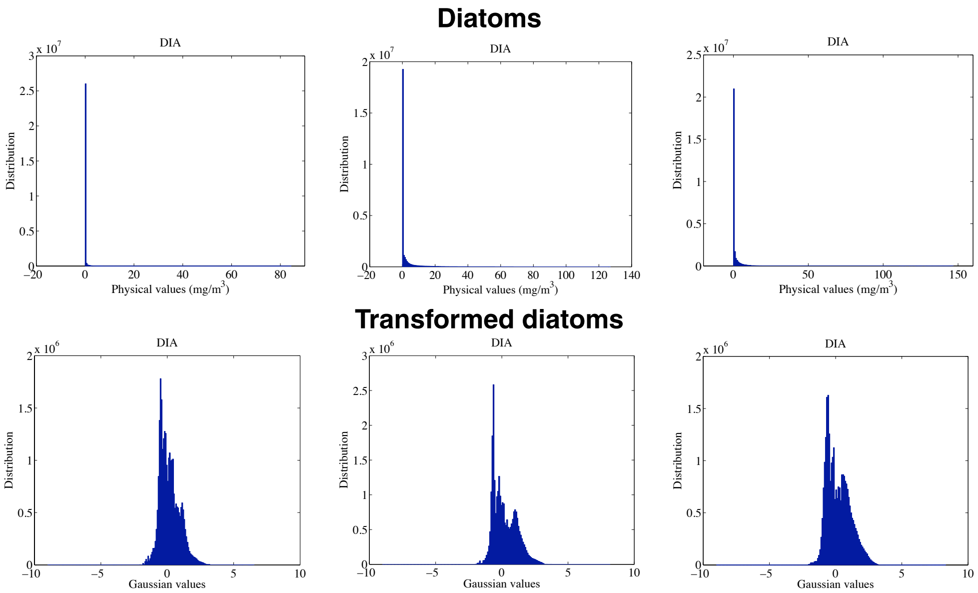
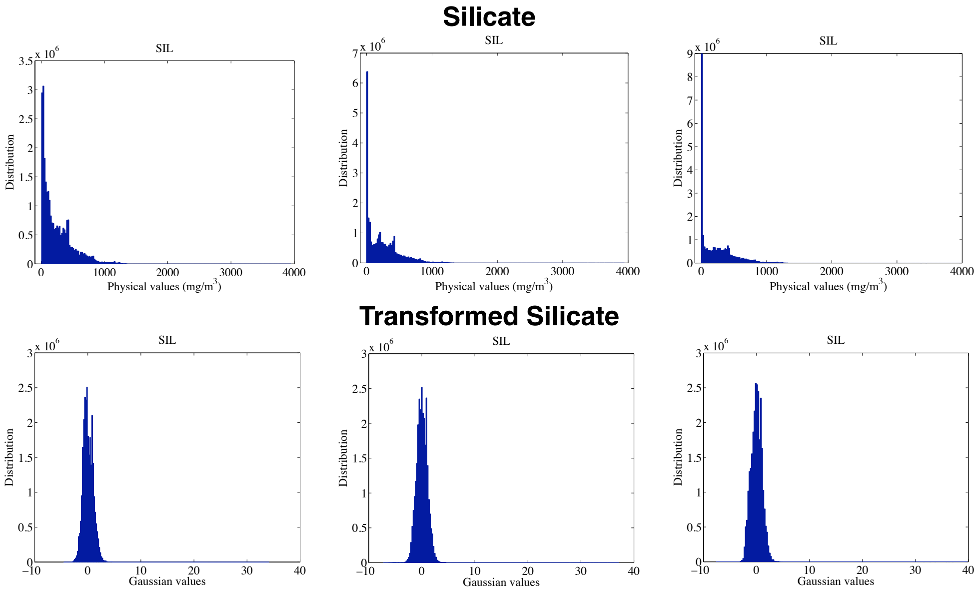


Fig. 4. Distributions of 3-D physical and transformed variables. Left: 31 December 1997; center: 14 May 1998; right: 3 September 1998.

Title Page
Abstract
Introduction
Conclusions
References
Tables
Figures
Back
Close
Full Screen / Esc
Printer-friendly Version
Interactive Discussion



Silicate

Transformed Silicate

Fig. 4. Continued.

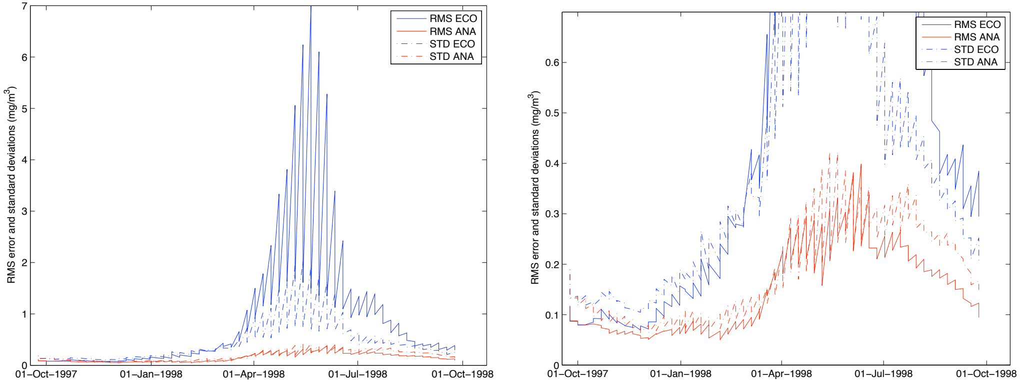


Fig. 5. Surface chlorophyll-*a*: one year evolution of the RMS error and the standard deviations (mg/m^3). On the right: zoom on the low values.

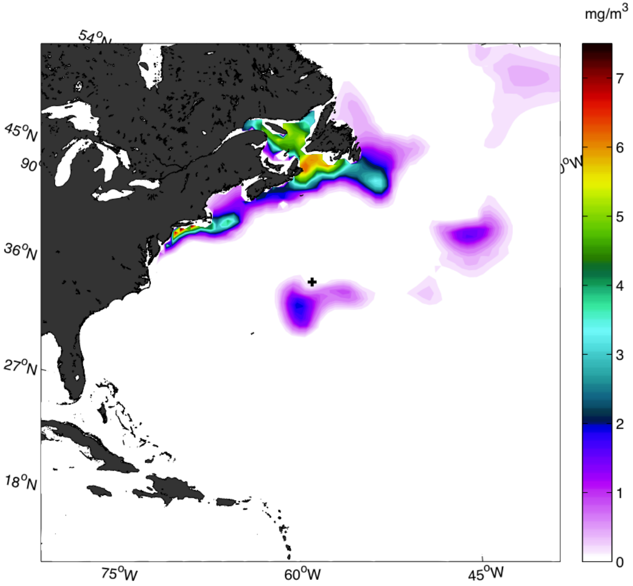


Fig. 6. Diatoms concentration (mg/m^3): the 8th layer on 31 December 1997. The point P_8 is localized by a cross.

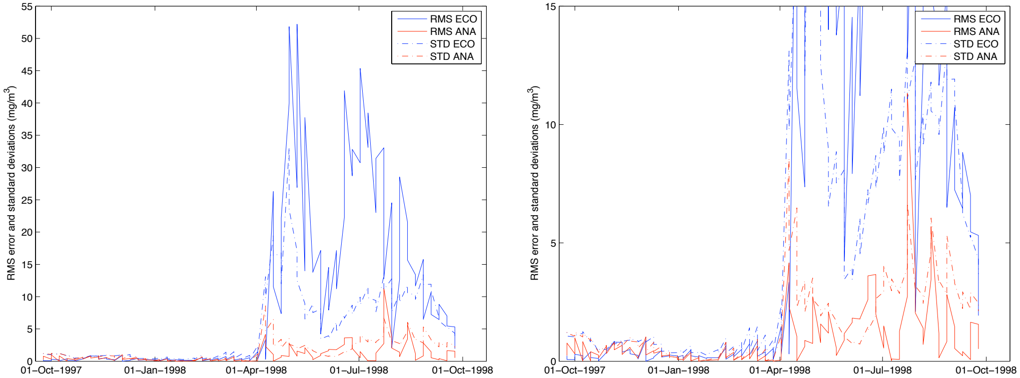


Fig. 7. Diatoms at point P_8: one year evolution of the RMS error and the standard deviations (mg/m^3). On the right: zoom on the low values.

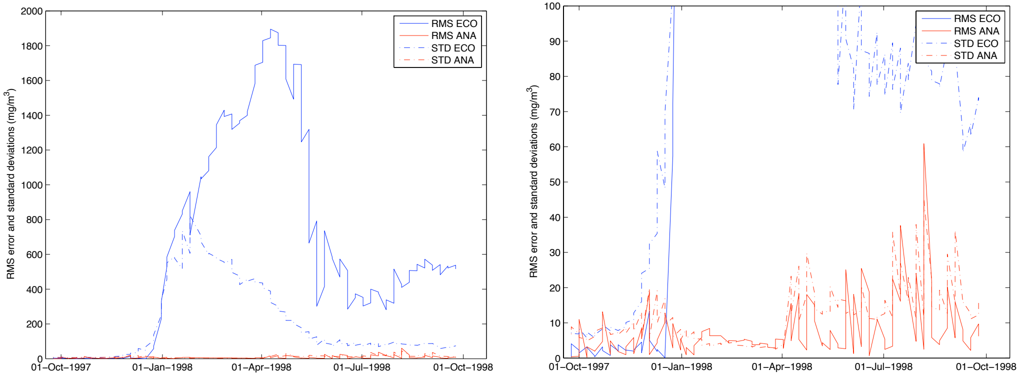
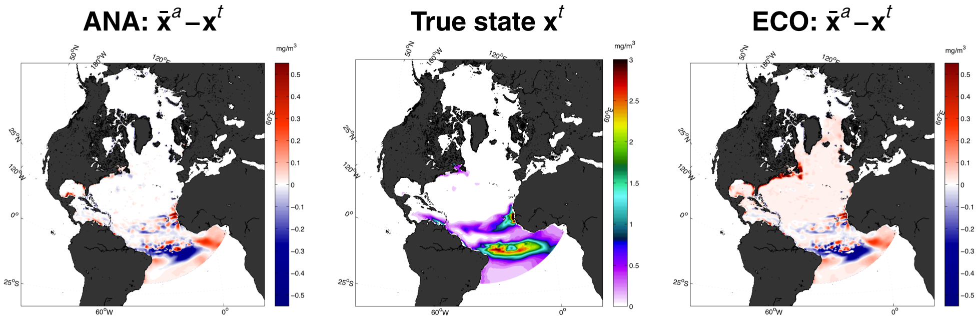


Fig. 8. Silicate at point P_8: one year evolution of the RMS error and the standard deviations (mg/m^3). On the right: zoom on the low values.

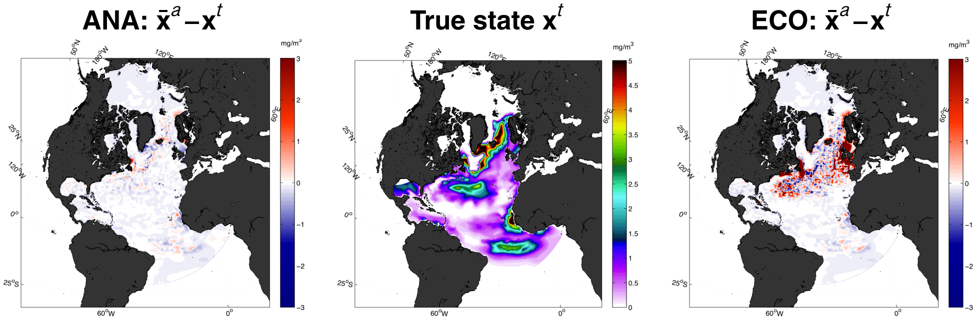
Title Page
Abstract
Introduction
Conclusions
References
Tables
Figures
◀◀
▶▶
◀
▶
Back
Close
Full Screen / Esc
Printer-friendly Version
Interactive Discussion


ANA: $\bar{\mathbf{x}}^a - \mathbf{x}^t$

True state $\mathbf{x}^t$

ECO: $\bar{\mathbf{x}}^a - \mathbf{x}^t$

Fig. 9. $\bar{\mathbf{x}}^a - \mathbf{x}^t$: surface chlorophyll-*a* component (mg/m^3) on 31 December 1997.



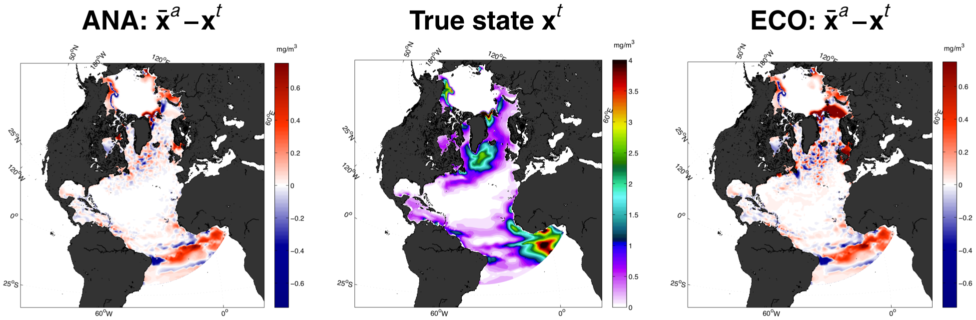
ANA: $\bar{\mathbf{x}}^a - \mathbf{x}^t$

True state $\mathbf{x}^t$

ECO: $\bar{\mathbf{x}}^a - \mathbf{x}^t$

Fig. 10. $\bar{\mathbf{x}}^a - \mathbf{x}^t$: surface chlorophyll-*a* component (mg/m^3) on 14 May 1998.

Title Page
Abstract
Introduction
Conclusions
References
Tables
Figures
◀|
|▶
◀
▶
Back
Close
Full Screen / Esc
Printer-friendly Version
Interactive Discussion



ANA: $\bar{\mathbf{x}}^a - \mathbf{x}^t$

True state $\mathbf{x}^t$

ECO: $\bar{\mathbf{x}}^a - \mathbf{x}^t$

Fig. 11. $\bar{\mathbf{x}}^a - \mathbf{x}^t$: surface chlorophyll-*a* component (mg/m^3) on 3 September 1998.

Title Page
Abstract
Introduction
Conclusions
References
Tables
Figures
◀◀
▶▶
◀
▶
Back
Close
Full Screen / Esc
Printer-friendly Version
Interactive Discussion