Ocean Sci., 6, 247–262, 2010
www.ocean-sci.net/6/247/2010/

Ocean Science

Characterization of mixing errors in a coupled physical biogeochemical model of the North Atlantic: implications for nonlinear estimation using Gaussian anamorphosis

D. Béal[1], **P. Brasseur**[1], **J.-M. Brankart**[1], **Y. Ourmières**[2], **and J. Verron**[1]

[1]LEGI/CNRS, Université de Grenoble, CNRS, BP 53X, 38041 Grenoble, France
[2]LSEET, Université du Sud Toulon Var, 83957 La Garde Cedex, France

Received: 4 June 2009 – Published in Ocean Sci. Discuss.: 30 June 2009
Revised: 20 January 2010 – Accepted: 21 January 2010 – Published: 17 February 2010

Abstract. In biogeochemical models coupled to ocean circulation models, vertical mixing is an important physical process which governs the nutrient supply and the plankton residence in the euphotic layer. However, vertical mixing is often poorly represented in numerical simulations because of approximate parameterizations of sub-grid scale turbulence, wind forcing errors and other mis-represented processes such as restratification by mesoscale eddies. Getting a sufficient knowledge of the nature and structure of these errors is necessary to implement appropriate data assimilation methods and to evaluate if they can be controlled by a given observation system.

In this paper, Monte Carlo simulations are conducted to study mixing errors induced by approximate wind forcings in a three-dimensional coupled physical-biogeochemical model of the North Atlantic with a 1/4° horizontal resolution. An ensemble forecast involving 200 members is performed during the 1998 spring bloom, by prescribing perturbations of the wind forcing to generate mixing errors. The biogeochemical response is shown to be rather complex because of nonlinearities and threshold effects in the coupled model. The response of the surface phytoplankton depends on the region of interest and is particularly sensitive to the local stratification. In addition, the statistical relationships computed between the various physical and biogeochemical variables reflect the signature of the non-Gaussian behaviour of the system. It is shown that significant information on the ecosystem can be retrieved from observations of chlorophyll concentration or sea surface temperature if a simple nonlinear change of variables (anamorphosis) is performed by mapping separately and locally the ensemble percentiles of the distributions of each state variable on the Gaussian percentiles. The results of idealized observational updates (performed with perfect observations and neglecting horizontal correlations) indicate that the implementation of this anamorphosis method into sequential assimilation schemes can substantially improve the accuracy of the estimation with respect to classical computations based on the Gaussian assumption.

Correspondence to: P. Brasseur
(pierre.brasseur@hmg.inpg.fr)

1 Introduction

Our understanding of the ocean biogeochemistry and marine ecosystems has made significant progress during the past decade. Coupled physical-biogeochemical models (CPBM) are becoming a useful source of information for many practical applications of societal and environmental importance, such as the monitoring and forecasting of marine resources, water quality and the ocean carbon cycle. Biogeochemical models are bound to be an essential component of the operational oceanographic systems that are being implemented, for instance, in the frame of the MERSEA and MyOcean European projects (Brasseur et al., 2009). In order to provide an accurate depiction of the essential biological variables, these models should be used in conjunction with global scale observation systems involving ocean colour satellites and profiling floats that, in the near future, will measure the subsurface concentration of oxygen, chlorophyll and nutrients (e.g., Gruber et al., 2006). The optimal merging of these multiple types of information requires the development of purpose-built assimilation methods, taking into account the specificities of the coupled physical-biogeochemical models, and of the data available for assimilation.

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

In order to design appropriate assimilation methods and to evaluate the level of control that can be expected from a given observation system, it is necessary to explore the structure of the errors that affect the model and the observations. A standard way to explore model errors is to perform Monte Carlo simulations (e.g., Evensen, 1994). This requires making prior assumptions about the possible sources of errors, originating for instance in a set of model parameters or in forcing functions. One then postulates a prior probability distribution for these errors, from which a sample is drawn. Model integrations are then performed for each element of the sample, and the resulting ensemble simulation provides an image of the model error structure (a sample of its probability distribution). From this image, it is then possible to diagnose how the original errors cascade on the various model state variables, if the errors are correlated in space and time, if robust relationships exist between observed and unobserved variables, if these relationships are close to linearity, how a given observing system can be used to control these errors, etc. Ensemble statistics can also be used to determine to which extent the probability distribution functions (pdfs) are Gaussian, and from this, the theoretical properties of the assimilation methods required to control the errors. In the context of marine ecosystem modelling, it is useful for instance to understand the level of control that can be expected from ocean colour data.

A key objective of the present study is to provide a characterization of mixing errors and their impact in coupled physical biogeochemical simulations. Another objective is to study the implications of the observed statistical behaviour for estimation and data assimilation methods. In this paper, a Monte Carlo method is applied to the study of mixing errors in a coupled physical-biogeochemical model of the North Atlantic ocean (described in Sect. 2.1), with a specific focus on the analysis of the ecosystem response to these errors. It is indeed well known that a cautious control of the ocean stratification and vertical mixing is crucial for consistent data assimilation in such coupled models (Berline et al., 2006), because it directly affects the nutrient supply and plankton residence time in the euphotic layer. Erroneous vertical mixing can be triggered by imperfections at different modelling stages, such as the wind forcing, the turbulent closure scheme or even the representation of mesoscale eddies through the restratification of the upper ocean (Oschlies, 2002).

To perform the Monte Carlo experiments, perturbations are applied to the wind forcing, which is the physical mechanism chosen here to trigger mixing errors in the coupled model. Common knowledge suggests that these errors propagate into the system according to the scheme of Fig. 1. Wind perturbations first induce perturbations of the mixed layer dynamics which translate into modifications of the mixed layer depth (MLD) and sea-surface temperature (SST). Deepening or shallowing of the mixed layer then modifies the nutrient supply in the euphotic layer, and subsequently the phytoplankton production in the euphotic layer. The impact on the biogeochemical state can be measured by the surface nitrate (NO_3) and phytoplankton (PHY) concentration. The latter is directly related to surface chlorophyll concentration (CHL), a quantity that is well observed through ocean colour satellites. By following this conceptual causal chain in the ensemble, it is possible to characterize the statistical dependence between the successive model variables and the observed quantities, their variations in space and time, and eventually the possibility to inverse the observed information back to the model space and forcing functions. These questions are examined in Sect. 3.

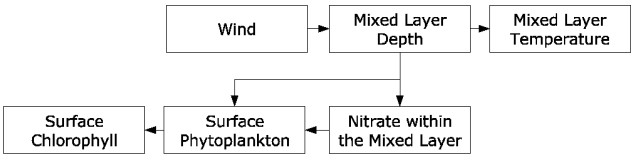

Fig. 1. Illustration of the conceptual transfer function between wind errors and the variables of a coupled physical-biogeochemical model. The arrows show the dominant effect that can be intuitively expected from ocean mixed layer and ecosystem dynamics.

One of the results of the ensemble simulations is that even for short-term forecasts (1 day), the relationships between ecosystem variables and observations are not close to linear, so that they cannot be fully exploited by a linear estimation method. For such a system, nonlinear methods are useful to improve the quality of the estimates. However, general nonlinear assimilation methods (e.g., particle filters as in Losa et al., 2003) which make no specific assumption about the shape of the prior pdf are too expensive for application to large size CPBM (16×10^6 state variables in our model), mainly because the identification of a general multivariate pdf with so many state variables would require too many ensemble members. Therefore, simplified solutions are needed to cope with real size problems.

A possible approach to non-Gaussian estimation problems is the use of anamorphosis transformations (i.e., Bertino et al., 2003; Lenartz et al., 2007), making nonlinear changes of variables to transform the forecast pdf (of arbitrary shape) into a Gaussian pdf. At first glance, this does not necessarily simplify the problem because identifying the change of variables requires a perfect knowledge of the original multivariate pdf, i.e. an ensemble as large as previously mentioned for particle filters. The simplified solution that we investigate in this paper is to perform the change of variable separately and locally for each state variable. In this way, a moderate size ensemble is usually sufficient to identify the change of variable and transform each marginal pdf to a nearly Gaussian pdf (see discussion in Sect. 4). This is obviously not sufficient to guarantee that the joint distribution becomes Gaussian. However, it is usually possible to detect from the ensemble the situations for which the approximation is accurate and the situations for which it is not. Both occur in our

case study, so that we will be able to evaluate the relevance of the scheme in any situation. A quantitative evaluation of the expected improvement with respect to linear estimates is also attempted to conclude the study.

2 Ocean model and wind forcing perturbations

2.1 The coupled physical-biogeochemical model

The CPBM used for the ensemble simulation was originally developed by Ourmières et al. (2009) for investigating the relative importance of nutrient vs. physical data to constrain the seasonal development of the phytoplankton bloom in the North Atlantic. The components of the coupled model include a NEMO/OPA9 circulation model of the North Atlantic basin at a 1/4° horizontal resolution (see Sect. 2.1.1), and a biogeochemical model derived from the 6-compartment LOBSTER formulation (see Sect. 2.1.2). The reference simulation (without wind perturbations), that is used as a reference for the Monte Carlo simulations, is described in Sect. 2.1.3.

2.1.1 The North Atlantic Ocean circulation model

The circulation model is a DRAKKAR configuration (The DRAKKAR Group, 2007) of the free surface primitive equation model NEMO/OPA (Madec et al., 1998). The domain covered is the North Atlantic basin from 20° S to 80° N and from 98° W to 23° E, with 1/4° resolution horizontal grid (Barnier et al., 2006). The vertical discretization is done using 45 geopotential levels, with a grid spacing increasing from 6 m at the surface to 250 m at the bottom. Vertical mixing of momentum and tracers is modelled by the TKE turbulence closure scheme (Blanke and Delecluse, 1993), and convection is parameterized with enhanced diffusivity and viscosity. Buffer zones are defined at the southern, northern and eastern (Mediterranean) boundaries with relaxation of temperature (TEM) and salinity (SAL) to Levitus climatology (Levitus et al., 2001). The forcing fluxes are calculated using bulk formulations and the ERA40 atmospheric forcing fields (Uppala et al., 2005). The prognostic variables include the zonal and meridional velocity components (U and V), temperature, salinity and sea surface height (SSH).

2.1.2 The LOBSTER biogeochemical model

LOBSTER (LOcean Biogeochemical Simulation Tools for Ecosystem and Resources) is a nitrogen-based ecosystem model with 6 prognostic variables in the euphotic layer: nitrate (NO_3), ammonium (NH_4), phytoplankton (PHY), zooplankton (ZOO), detritus and semi-labile dissolved organic nitrogen (Levy et al., 2005a). The bottom of the euphotic layer is prescribed at a constant depth of 191 m. Below the euphotic layer, the model considers very simple parameterizations of decay to nitrate, detritus sedimentation and remineralization of zooplankton mortality. LOBSTER is coupled on-line to the circulation model without feedback of the biogeochemical variables on the physics. The coupling frequency is equal to the circulation model time-step (40 min). The on-line coupling as well as the maximum frequency is thought to allow accurate diagnostics of the ecosystem evolution without possible problems brought by the use of averaged physical fields as an off-line configuration would need. More detail about the model equations is available in Levy et al. (2005a and 2005b) and about the North Atlantic implementation in Ourmières et al. (2009).

2.1.3 Reference simulation of the coupled model

The reference simulation of the coupled model used in this study corresponds to year 1998 of the FREE simulation described in Ourmières et al. (2009) and performed without data assimilation. In this simulation, the U, V and SSH fields are initialized to zero, while the TEM and SAL fields are interpolated from the December Levitus climatology (Levitus et al., 1998). Then, the physical model is run for 12 years from 1 January, 1984 to 1 January, 1996, providing a balanced physical ocean state to start the biogeochemical model spin-up. At that time, the nitrate field is initialized with the December climatology 2001 (Conkright et al., 2002) interpolated on the model grid. The other biogeochemical fields are set to constant values in the euphotic zone and to zero below: zooplankton is set to 0.01 mmol N/m^3, phytoplankton to 0.1 mmol N/m^3 and ammonium, dissolved organic matter and detritus to 0.001 mmol N/m^3. The coupled model is then run for 2 years starting 1 January, 1996 and using the physical ocean state obtained after 12 years of spin-up. Ourmières et al. (2009) analysed the convergence of the run and showed that the model is able to reproduce satisfying seasonal cycles of the biogeochemical variables. In this study, we will analyse the 1-month period between 15 April and 15 May, 1998, i.e. when the bloom event occurs.

2.2 Perturbed simulations

In order to generate an ensemble of model runs impacted by mixing errors in the upper ocean, Monte Carlo simulations are performed using perturbations of the surface forcings. Perturbations of the wind stress are considered here as the only source of mixing errors, while in reality these errors originate from a variety of approximations in the parameterization of sub-grid scale turbulence, in the specification of the surface boundary conditions for momentum, heat and salinity, and from other mis-represented dynamical processes such as restratification by mesoscale eddies. We proceed in two steps, assuming that the uncertainty in the wind can be estimated from the variability of ERA40 winds of March, April and May during 1985–2000: (i) the covariance of the wind variability is calculated using the ERA40 database, and (ii) the wind perturbations are randomly sampled from

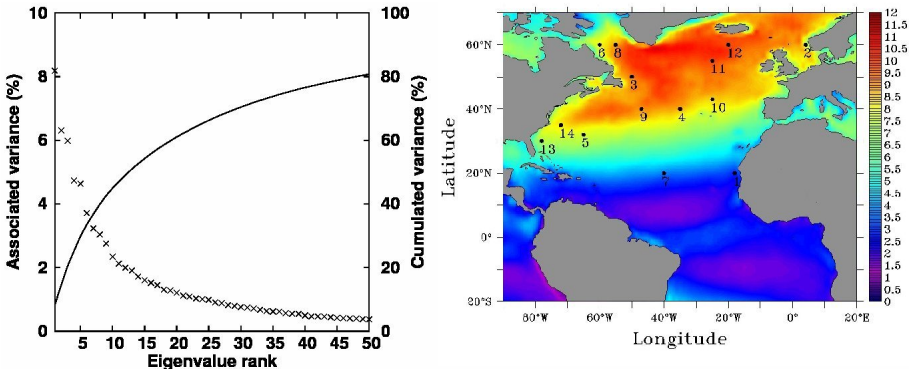

Fig. 2. (Left) Percentage of explained variance (left axis) and cumulated variance (right axis) for the first 50 EOFs computed from the variability of the ERA40 1985–2000 wind archives. (Right) Wind stress standard deviation (in N/m^2) calculated over the used archives.

a Gaussian probability distribution function with zero mean and this pre-calculated covariance.

In practice, an ensemble composed of one wind field every 4 days is extracted from the 1985–2000 ERA40 winds during the 3 months period centred on 15 April. This ensemble contains 368 members representative of the season during which the Monte Carlo simulations are performed. A multivariate EOF (Empirical Orthogonal Function) analysis of this ensemble is performed combining the u and v components of the wind, and the first 50 dominant EOFs (representing 80% of the wind variance) are used to generate the perturbations. Figure 2 (left panel) illustrates the first 50 eigenvalues in decreasing order, their corresponding percentage of explained variance and the cumulated percentage of explained variance. Figure 2 (right panel) also shows the standard deviation of the resulting wind stress variability which is also the expected standard deviation of the wind stress perturbations. It is especially large over the subpolar gyre and over the Gulf Stream region. As mentioned above, wind perturbations generate anomalies of the biogeochemical model variables. As a result, a more intense ecosystem response is expected in the subpolar and Gulf Stream regions. These regions are also where the intensity of the spring bloom is maximum in the reference simulation.

The Monte Carlo simulations are then performed using an ensemble of 200 time-varying perturbations of the wind forcing. Assuming that the typical decorrelation time scale of wind errors is about 4 days, independent samples of 200 members are drawn every 4 days with the covariance defined above. These are then interpolated linearly in time to obtain perturbations every 6 h, which is the input frequency of forcing fields in the ocean model. In practice, this corresponds to sample independent coefficients for each EOF from $\mathcal{N}(0,1)$ every 4 days, interpolate them in time to obtain the perturbation amplitude $\alpha_i(t)$ for every EOF$_i$, $i = 1...50$, and them compute the perturbed wind using Eq. (1). It is worth noting that in Eq. (1), the normalized EOFs are multiplied by the squared root of the corresponding eigenvalue, so that each EOF is a column of the squared root of the perturbation covariance matrix.

$$\begin{pmatrix} u(t) \\ v(t) \end{pmatrix}_{\text{pert}} = \begin{pmatrix} u(t) \\ v(t) \end{pmatrix}_{\text{reference}} + \alpha_1(t)\,\text{EOF}_1 + \ldots + \alpha_{50}(t)\,\text{EOF}_{50} \quad (1)$$

3 Study of the ensemble forecast

The objective of this section is to describe the ensemble response of the model to the wind perturbations described in Sect. 2. This response is analysed by studying the ensemble forecast at 14 stations in the North Atlantic (see their location in Fig. 2), especially at BATS (Bermuda Atlantic Time Series, station 5), INDIA (Ocean Weather Station India, station 11) and NABE (North Atlantic Bloom Experiment, station 12) biogeochemical stations, and in the Gulf Stream (station 14, noted GS). For three of these stations (BATS, INDIA and GS), ensemble scatterplots are presented to characterize the relationships that can be deduced from the transfer function in Fig. 1, i.e. between WND and MLD, MLD and TEM, MLD and NO_3, MLD and PHY, or NO_3 and PHY. (WND is the wind stress modulus expressed in N/m^2). To interpret the mechanisms behind these relationships, we also analyse the ensemble of TEM, NO_3 and PHY vertical profiles at these stations.

In addition, the information extracted from the ensemble are synthesized using two statistics (presented for all 14 stations in Table 1):

– the linear correlation coefficient (Pearson):

$$r = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2}\sqrt{\sum_i (y_i - \overline{y})^2}} \quad (2)$$

where $x=(x_i)_{i=1}^n$ and $y=(y_i)_{i=1}^n$ are n-size samples of 2 random discrete variables and $\overline{x}$ and $\overline{y}$ are the respective means of these samples;

Table 1. Linear vs rank correlation coefficient between variables at 14 stations of the North Atlantic domain, as obtained from the 1-day ensemble forecast. A significantly higher rank correlation (in bold) means that anamorphosis is likely to be useful.

Stations	WND/MLD	MLD/TEM	MLD/NO_3	TEM/PHY	SAL/N03	NO_3/PHY
Mauritania (1)	**0.83/0.94**	**−0.88/−0.97**	**0.87/0.97**	**0.85/0.96**	−0.93/−0.93	**−0.81/−0.94**
Norway (2)	0.85/0.06	0.98/0.91	0.75/0.37	−0.48/−0.01	0.97/0.93	−0.95/−0.67
New Foundland (3)	0.91/0.63	0.95/0.93	0.80/0.75	−0.79/−0.59	0.99/1.00	−0.96/−0.91
Acores (4)	0.87/0.87	**−0.95/−1.00**	**0.95/0.99**	0.99/0.98	−0.97/−0.98	−0.98/−0.99
BATS (5)	0.88/0.91	**−0.78/−1.00**	**0.85/0.97**	0.99/0.98	−0.99/−0.99	−0.99/−0.95
Labrador 1 (6)	0.89/0.81	**0.79/0.94**	**0.88/0.99**	−0.87/−0.84	0.98/0.98	−0.99/−0.97
Subtropical Gyre (7)	**0.72/0.92**	**−0.83/−0.93**	0.32/0.31	0.78/0.62	**−0.45/−0.69**	−0.84/−0.61
Labrador (8)	0.76/0.69	0.29/0.37	0.89/0.94	−0.31/−0.33	0.98/0.84	−1.00/−0.99
Gulf Stream (9)	0.90/0.91	**−0.92/−0.98**	0.89/0.4	**0.91/0.99**	−0.32/−0.17	0.85/0.39
Pomme (10)	**0.87/0.93**	−0.99/−1.00	**0.96/0.99**	0.99/1.00	0.05/0.41	**−0.93/−0.98**
INDIA (11)	**0.31/0.48**	−1.00/−0.98	0.45/0.41	0.53/0.48	**0.93/0.97**	−0.99/−0.98
NABE (12)	**0.09/0.22**	−0.97/−0.94	0.51/0.46	0.34/0.29	0.80/0.83	−0.90/−0.86
Gulf Stream 1 (13)	0.37/0.07	−0.67/−0.64	0.70/0.32	−0.10/−0.20	**0.72/0.82**	−0.73/−0.65
Gulf Stream 2 (14)	0.22/0.13	−0.98/−1.00	0.97/0.95	0.93/0.96	−0.73/−.72	−0.97/−0.99

– the rank correlation (Spearman) that is identical to the linear correlation except that each value x_i (respectively y_i) is replaced by the value of its rank R_i (respectively S_i) in the sample (e.g. R_i is the index of x_i in the sorted sample). The sequence R_i (or S_i) thus contains all integers between 1 and n:

$$r_s = \frac{\sum_i (R_i - \overline{R})(S_i - \overline{S})}{\sqrt{\sum_i (R_i - \overline{R})^2}\sqrt{\sum_i (S_i - \overline{S})^2}} \quad (3)$$

where $\overline{R}$ and $\overline{S}$ are respectively the mean of R and S. The rank correlation is useful to detect nonlinear relationships between variables (see for instance Press et al., 1992, chapter 14).

We also study how these correlations between model variables evolve with time, and the time scales over which the correlations with observed quantities can be considered robust enough to be exploited by a data assimilation system.

3.1 The ensemble response at three locations

By looking at the ensemble forecast after only one day of run, we will see that mixing is the dominant mechanism responsible for the propagation of wind forcing errors to the other state variables, in most locations. This is because the daily time scale is too short to trigger intense dynamical interactions between the biogeochemical variables of the LOBSTER model.

The corresponding correlation statistics are given in Table 1 for all 14 stations shown in Fig. 2. The ensemble response is analysed in details at three specific locations: at the BATS station (Fig. 3), the GS station (Fig. 4) and the INDIA station (Fig. 5). The ensemble statistics obtained at INDIA, BATS and GS provide good illustrations of statistical behaviours that are representative of very different stratification conditions. INDIA is located in a high-latitude, North Atlantic region dominated by strong wind variability (Fig. 2) and subject to strong convective events in winter. By contrast, BATS is representative of the subtropical gyre, with rather stable winds and well stratified upper ocean throughout the year. The GS station is located in the inter-gyre region, with intermediate wind variability and moderate stratification conditions. The figures show five scatterplots describing the transfer function in Fig. 1, as well as ensemble vertical profiles of temperature, nitrate and phytoplankton. We will discuss in sequence the propagation of uncertainties from the wind forcing to the physical properties, and then to the biogeochemical properties of the mixed layer.

3.1.1 Relationships between wind forcing and physical properties of the mixed layer

As a first step, we analyze the cascade of errors from the wind forcing to the physical variables (first line in Fig. 1).

WND/MLD. Wind errors generate different types of response on the mixed layer depth (see WND/MLD scatterplots in Figs. 3, 4 and 5). As a general rule, the larger the wind, the deeper the mixed layer; however, there are significant differences between the 3 situations. The scale of the plots shows that the amplitude of the MLD and TEM perturbations observed at INDIA are significantly smaller that the corresponding perturbations at BATS and GS, in spite of similar perturbations of the wind. Further, the spread around the linear regression line is larger at the INDIA station, while

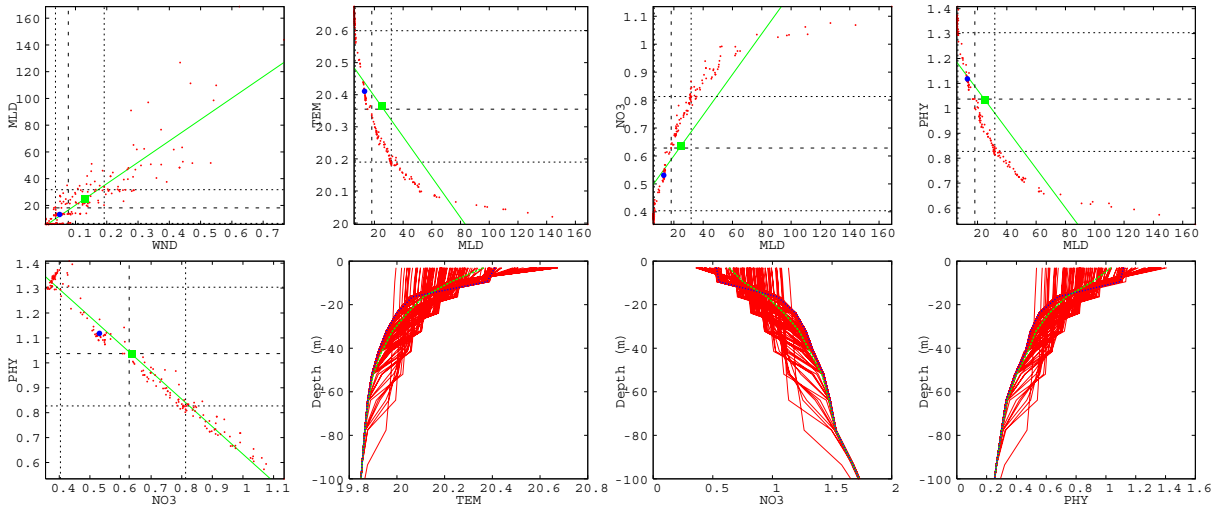

Fig. 3. Scatterplots of 1-day ensemble forecasts at BATS (65° W/32° N) station: the red points correspond to the 200 ensemble members; the blue point corresponds to the reference (unperturbed) run; the green square is the ensemble mean; the green line represents the linear regression of the ensemble; the black doted lines indicate the quartiles of the distribution. Vertical profiles: the red lines correspond to the 200 ensemble members, the blue line is the profile of the reference run, the green line is the mean profile.

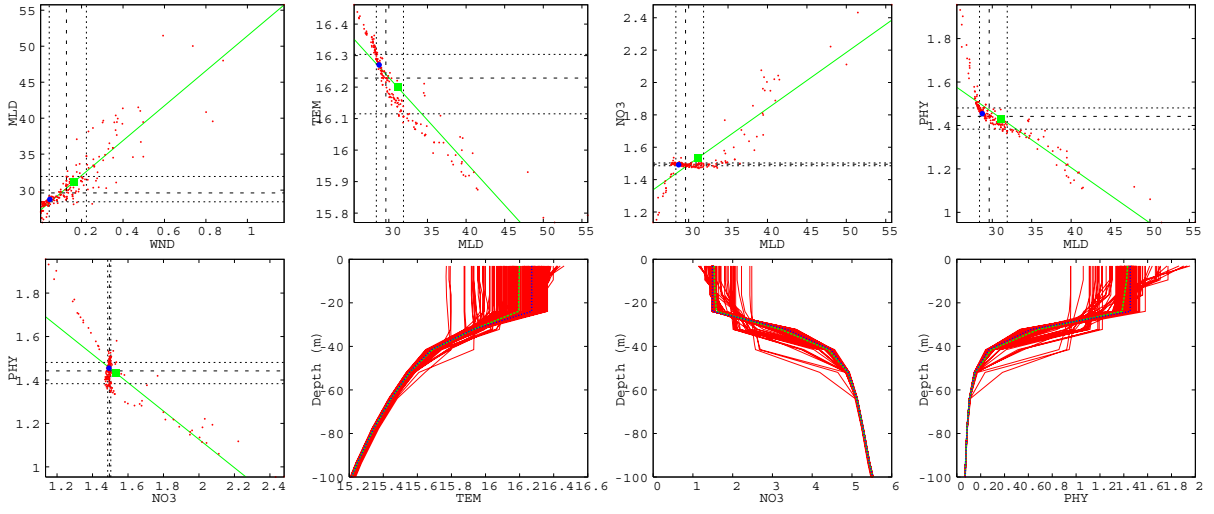

Fig. 4. Same as Fig. 3 but for the Gulf Stream (47° W/40° N) station.

such spread does not occur in the same way at the other stations. The relationship between WND and MLD is obviously nonlinear at INDIA station. For large wind anomalies, one can observe a sort of saturation of mixed layer depth perturbations. This can be explained by the very different mixed layer structures of the 3 reference states: at BATS, the mixed layer is very shallow and the turbulent energy brought by the wind immediately propagates down to the thermocline. The exactly opposite situation occurs at INDIA, where the water column of the reference run is well mixed down to around 400 m. As a result, the mixed layer depth is relatively insensitive to wind anomalies.

MLD/TEM. In general, the consequence of the mixed layer deepening when wind forcing increases is a cooling of the sea surface (see TEM/MLD plots in Figs. 3, 4 and 5). The mixing of warm surface water with cold water at depth results in a cooling of the mixed layer. The TEM/MLD relationships decrease monotonously, but not necessarily in a linear way. The shape of this relationship obviously depends on the shape of the vertical TEM profile. Moreover, the statistics of Table 1 show very high rank correlations, meaning that a quite robust relationship exist for this combination of variables (except at the Labrador station).

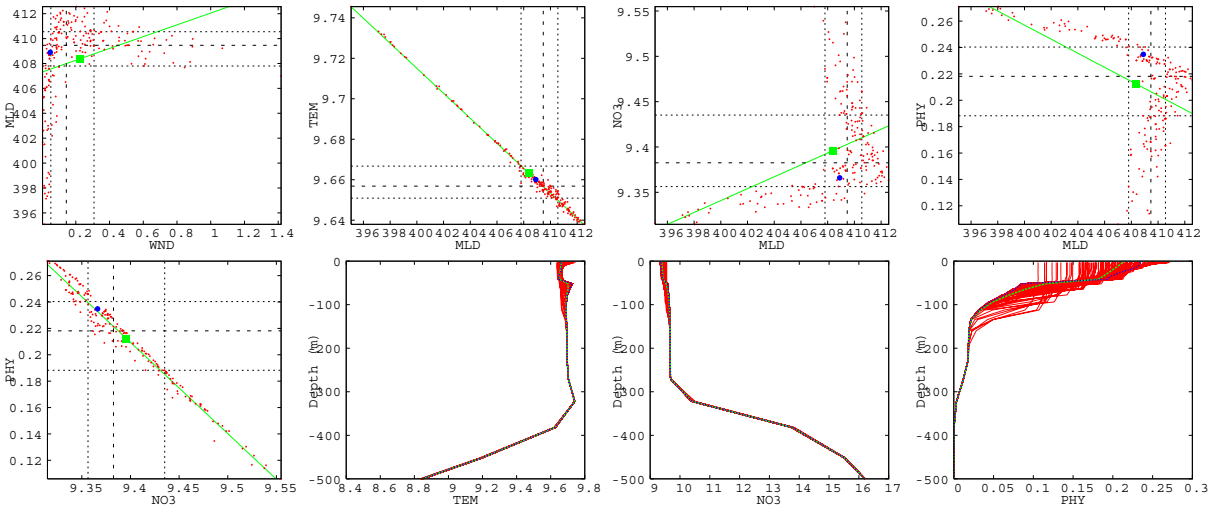

Fig. 5. Same as Fig. 3 but for the INDIA (25° W/55° N) station.

3.1.2 Relationships between mixed layer and biogeochemical properties

As a second step, we analyze the cascade of errors from the mixed layer to biogeochemical variables (second line in Fig. 1).

MLD/NO$_3$. Deepening of the mixed layer is expected to bring nitrate to the surface by mixing nutrient-rich deep water with nutrient-depleted surface water. This is exactly what happens at BATS and GS stations, where a nonlinear increase of NO_3 concentration is observed when the mixed layer deepens. From the scatterplot of the Gulf Stream station, one can however notice the existence of a plateau around the reference NO_3 concentration of 1.5 mmol m^{-3}: perturbations of the wind below some threshold are unable to propagate anomalies down to the nutricline depth. By contrast, the wind reduction yields restratification of the water column, which favours the consumption of NO_3 by phytoplankton. At INDIA station, we observe the same phenomenology as for MLD: the wind perturbations are not strong enough to significantly modify the NO_3 concentration over the whole 400 m mixed layer.

MLD/PHY. A nonlinear decrease of PHY concentration is observed when the mixed layer deepens. Since phytoplankton concentration typically dominates in the euphotic zone and weakens at depth, phytoplankton is expulsed from surface layers by mixing, and the MLD and PHY variables are negatively correlated. This is an exactly opposite behaviour compared to nitrate at BATS and GS stations, where mixing seems to be the dominant effect. It is interesting to note that such negative correlation could also be interpreted as the combined effect of shallowing MLD and increasing irradiance, as it typically occurs during bloom events. The INDIA station still shows a complex response which is difficult to interpret by simple mechanisms.

Finally, we analyze the scatterplots between the NO_3 and PHY biogeochemical variables.

NO$_3$/PHY. The scatterplots are characterized by well-defined relationships with pretty high correlations, sometimes altered by threshold effects as illustrated for the GS station. The statistics show that surface phytoplankton generally decreases when nitrate concentration increases. On the vertical, inverse distributions of phytoplankton and nutrient are observed over the water column. One can note interestingly that this general trend is consistent with the basic mechanism of phytoplankton growth which requires nutrient consumption in the euphotic layer. In the LOBSTER model, the phytoplankton growth is made possible by 2 different pathways: the new production sustained by nitrate, and the regenerated production sustained by ammonium. A cluster of high phytoplankton concentrations can be observed at BATS station for poor nitrate values, which might be explained by the regenerated phytoplankton production associated to very thin MLD. This is an example where a biogeochemical mechanism, different than mixing, transforms the error propagation in the coupled model.

In summary, the results discussed here above indicate that the propagation of wind errors after a one-day forecast is strongly dependent on the local stratification of the ocean, and that mixing is the dominant mechanism explaining the behaviour of the ensemble. In a first approximation, the state variables (TEM, NO_3, PHY) can be considered as passive tracers as long as the lead time remains small (one day). Further, the relationships between variables are generally loosing their robustness when the mixed layer deepens. The response of the CPBM after one day can be very complex, demonstrating nonlinear relationships between state variables with sometimes threshold effects. In the following section, we will focus on the evolution of the ensemble spread and the corresponding correlations with time.

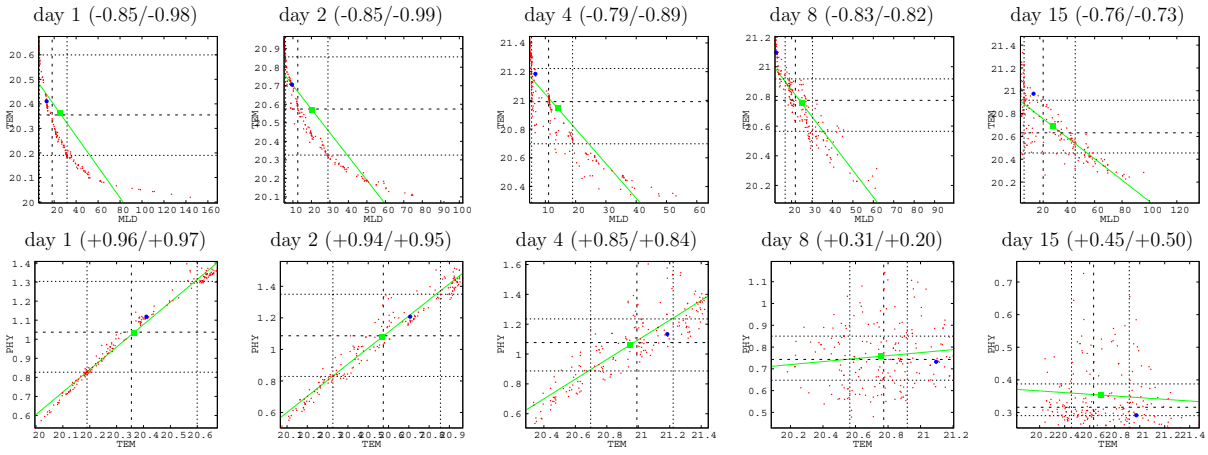

Fig. 6. Scatterplots of ensemble forecasts at BATS station (65° W/32° N) after 1, 2, 4, 8 and 15 days (from left to right): MLD/TEM (top line) and TEM/PHY (bottom line) relationships. Similar colour code as in Fig. 3.

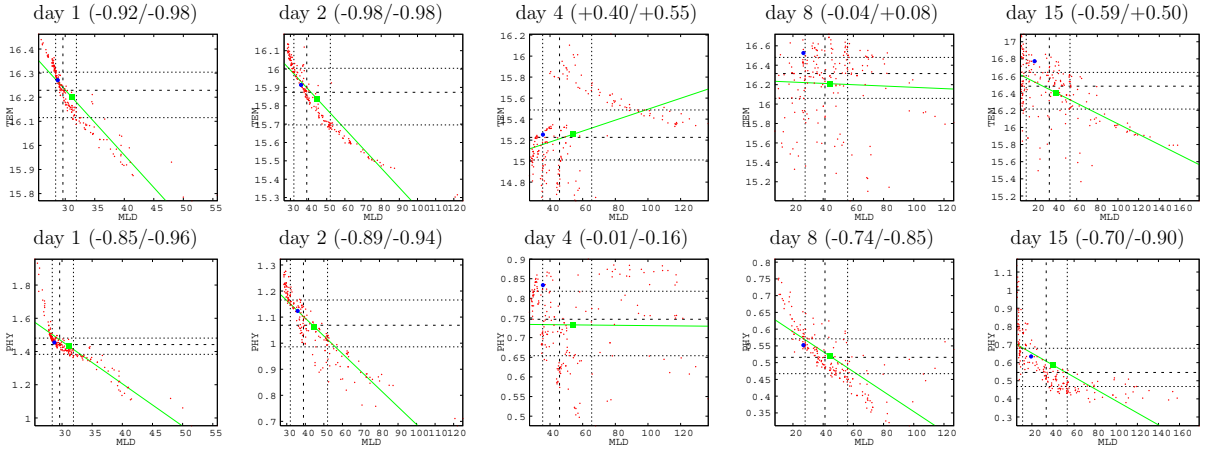

Fig. 7. Scatterplots of ensemble forecasts at Gulf Stream station (47° W/40° N) after 1, 2, 4, 8 and 15 days (from left to right): MLD/TEM (top line) and MLD/PHY (bottom line) relationships. Similar colour code as in Fig. 3.

3.2 Temporal evolution of the ensemble response

The objective of this section is to analyse the stability of these statistical relationships over a 2 week period after the application of wind perturbations. Figures 6 (BATS station) and 7 (Gulf Stream) show the scatterplots after 1, 2, 4, 8 and 15 days of run, illustrating the temporal evolution of relationships between variables. The discussion of the temporal evolution of the ensemble response at INDIA station has not been addressed in detail because it leads to conclusions that are very similar to the GS case and does not bring novel information about the stability of the ensemble covariance.

The spread of the ensemble with time is the first general trend clearly illustrated by these 2 figures. The more the experiment lasts, the larger the dispersion (following each line from left to right), and the variables tend to decorrelate with time. This is particularly visible for MLD/TEM, PHY/TEM and MLD/PHY relationships, leading for instance to an almost complete decorrelation after 8 or 15 days between PHY and TEM at BATS station. Note that sometimes a decorrelation during the first days of run can be followed by the recorrelation of the variables, as for example for the MLD/PHY at the GS station before and after the 4th day of run.

The shape of the relationships may also change with time. For instance, the nonlinear TEM/MLD relationship at BATS station is getting almost linear after the 8th day of run (except for small MLD values). More than that, initially well-defined relationships such as TEM/MLD and PHY/MLD at the GS station are becoming fuzzy after 4 days of run, and recover some structure after 8 or 15 days, but with a different shape. Finally, scatterplots could also disperse in such a way that no relationship exists anymore (e.g., PHY/TEM scatterplots on Fig. 6 after 8 days).

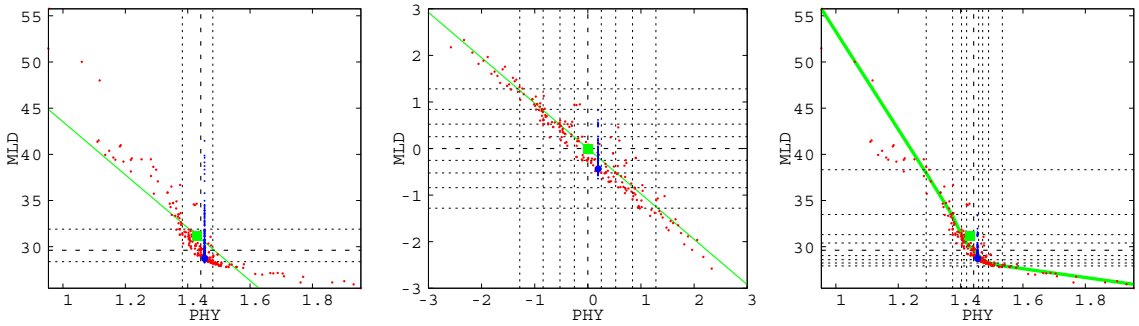

Fig. 8. Observational update at BATS station (65° W/32° N) using a perfect phytoplankton observation. The figure shows the 1-day ensemble forecast (red dots), with mean (green square) and linear regression line (thin green line), the reference simulation (large blue dot) that gives the PHY observation and the update ensemble (blue dots). The left panel illustrates a linear observational update performed in the original state space. In the middle panel the linear observation update is performed in a transformed state space (by anamorphosis). In the right panel the solution showed in the middle panel is transformed back into the original state space. The linear regression line of the middle panel (thin green line) transforms into the thick green line of the right panel. Dashed lines are medianes, and dotted lines are percentiles (quartiles in the left panel and deciles in the other panels).

As a conclusion, the ensemble response of the CPBM at lead times greater than one day is quite complex, with often enhanced dispersion and structural modification of the relationships. The temporal evolution of the scatterplots shows that reasonable relationships are sometimes preserved after 4 days of wind perturbation (e.g., at BATS station), and sometimes not (e.g., at the GS Station). In particular, relationships at BATS station obtained after a 4-day forecast could be used to determine the cascade of errors from WND to MLD, from MLD to TEM, and finally from MLD and TEM to PHY. From that kind of information, it is in principle possible to evaluate the potential utilization of observed chlorophyll data to control the state variables of the CPBM. In order to assess which state variable of the CPBM can be estimated using surface chlorophyll measurements over typical data assimilation time scales of 4 to 6 days, we use the examples of BATS or GS stations after 4 days of wind perturbations. These examples illustrate how the chain of errors in Fig. 1 can be used as a conceptual mechanism to quantify the potential performance of a linear observational update (even if, in practice, the observational update of sequential assimilation schemes should not be segmented into substeps according to this chain of errors because it would make the estimation process suboptimal and increase the complexity of the analysis step). A well-known limitation of the linear methods is indeed that the quality of the observational update requires linear relationships with sufficiently low dispersion to compute accurate inverse estimates of unobserved variables. The analysis of our results (Figs. 6 and 7) indicates that, even if a linear update might be somewhat beneficial at these stations, the clear non-Gaussian behaviour of the ensemble ideally requires more advanced methods. In the next section, we will demonstrate how linear updating methods can be upgraded to take into account such non-Gaussian behaviours.

4 Toward data assimilation: inference method using anamorphosis

The diagnostics of the ensemble forecasts presented in the previous section show the omnipresence of non-Gaussian behaviours as well as nonlinear relationships between state variables, which should be taken into account to produce an optimal update of the state of the system using the available observations. In the first subsection (Sect. 4.1), we first illustrate the problems that occur if a linear (Gaussian) observational update is used. This is done at the surface of the ocean using the reference phytoplankton as observation, and each member of the ensemble as background state. In a second stage (Sect. 4.2), a simple nonlinear transformation of the variables (anamorphosis) is proposed to improve the observational update. And finally, in Sect. 4.3, we discuss the impact of this anamorphic transformation for the whole North Atlantic domain.

4.1 Problems with linear observational update

In conventional Kalman filters, the linear observational update is computed using the formula:

$$\mathbf{x}^a = \mathbf{x}^f + \mathbf{K}\left(\mathbf{y} - \mathbf{H}\mathbf{x}^f\right) \tag{4}$$

where $\mathbf{x}^f$ is the forecast (or background) state, $\mathbf{y}$, the observation vector, $\mathbf{H}$, the observation operator and $\mathbf{K}$, the Kalman gain. It minimizes the estimation error variance (and thus corresponds to the best linear unbiased estimate) if the gain is computed by:

$$\mathbf{K} = \mathbf{P}^f\mathbf{H}^T\left(\mathbf{H}\mathbf{P}^f\mathbf{H}^T + \mathbf{R}\right)^{-1} \tag{5}$$

where $\mathbf{P}^f$ is the forecast (or background) error covariance matrix and $\mathbf{R}$, the observation error covariance matrix. This

solution also provides the absolute minimum error variance estimate (not only the best linear one) providing that the probability distributions are Gaussian. In this case, it also corresponds to the maximum likelihood estimate. Conversely, if the pdf are not Gaussian, better estimates exist in general.

In this paper, we restrict ourselves to the problem of estimating one state variable from the perfect observation of another state variable. For instance, in Fig. 8, we estimate the mixed layer depth from one phytoplankton observation. We use the reference simulation (large blue dot) as observation, and in order to get a solution that is statistically valid, we use sucessively each member of the ensemble as background (red dots). The solution will be deduced from the distribution of the updated values (small blue dots). We first focus on the left panel of Fig. 8 which illustrates the linear observational update. For that specific example, formula (6) can be rewritten

$$\mathrm{MLD}^a = \mathrm{MLD}^f + \gamma \frac{\sigma_{\mathrm{MLD}}}{\sigma_{\mathrm{PHY}}} \left(\mathrm{PHY}^o - \mathrm{PHY}^f \right) \quad (6)$$

where $(\mathrm{PHY}^f, \mathrm{MLD}^f)$ are the background values (red points), PHY^o is the observed value (abscissa of the large blue dot), $(\sigma_{\mathrm{PHY}}, \sigma_{\mathrm{MLD}})$ are the ensemble standard deviation for PHY and MLD, and γ is the linear correlation coefficient between PHY and MLD. Since the observation is perfect, all updated values $(\mathrm{PHY}^o, \mathrm{MLD}^a)$ (blue dots) are aligned vertically on the PHY^o value.

From the previous equation, it is apparent that the observational update (from the red point to the blue point) is done along a straight line with the given slope $\gamma \frac{\sigma_{\mathrm{MLD}}}{\sigma_{\mathrm{PHY}}}$, which is the slope of the linear regression line (in green on the figure) passing through the ensemble mean (green square). Hence, in this simple example, the ensemble observational update can be viewed as drawing from each red point a parallel to the green line and find the updated value at the intersection of this line with the vertical $\mathrm{PHY}=\mathrm{PHY}^o$.

But, from the ensemble displayed in Fig. 8 (red points), it is quite clear that the pdf is far from being Gaussian. For example, the quartiles of the marginal distributions (thin dashed lines) are not symmetric around the median (thick dashed line). On the other hand, in a general two-dimensional pdf, the regression curve (for instance for MLD) is defined (e.g. Von Mises, 1964) as the line with maximum MLD probability density for each value of the other variable (PHY). If a pdf is Gaussian, the regression curve is a straight line, which corresponds to the linear regression line defined above (drawn in green in the figure). Obviously, in our example, the maximum MLD probability for each PHY value is usually well above or well below the linear regression line, indicating again a non-Gaussian behaviour. Hence performing the observational update by following the linear regression line without exploiting the real shape of the distribution always produces suboptimal estimates, with significantly larger estimation errors. Moreover, we observe in Fig. 8 that the true regression line has a general positive curvature, so that the linear estimate is almost systematically above the true MLD value.

4.2 Nonlinear observational update using anamorphosis

4.2.1 Description of the anamorphosis transformation

In order to improve the observational update, we apply here a simplified method (similar to the one proposed by Bertino et al., 2003) with the general idea of transforming each marginal pdf into a pdf that is close to Gaussian. This is achieved by performing a change of variables (anamorphosis) separately for each single variable of the state vector (every physical/biogeochemical component at every horizontal/vertical location). For instance, Fig. 9 (left panel) shows the ensemble distribution of surface nitrate at the BATS station. Again, the pdf is obviously far from Gaussian. Let us denote by x the original random variable, and by $y=f(x)$, the transformed random variable. The objective is to find the function f defining a change of variables (anamorphosis) such that the random variable y is as close as possible to the Gaussian pdf $\mathcal{N}(0,1)$. Moreover, we want to infer f from the current ensemble description of the pdf of x. (This last point is the main difference with respect to the work of Bertino et al., 2003).

In order to reach this objective, the idea is to use the piecewise linear change of variable f remapping a set of percentiles of the pdf of x to the same percentiles of $\mathcal{N}(0,1)$. For instance, if x_k, $k=1,\ldots,p$ are the p percentiles of x (such that $p(x<x_k)=r_k$), for a given set of values $r_k, k=1,\ldots,p, 0<r_k<1, r_k<r_{k+1}$), and y_k are the corresponding percentiles of $\mathcal{N}(0,1)$, the function $f(x)$ writes:

$$f(x) = \begin{cases} y_1 & \text{for } x < x_1 \\ y_k + \frac{y_{k+1}-y_k}{x_{k+1}-x_k}(x-x_k) & \text{for } x \in [x_k, x_{k+1}] \\ y_p & \text{for } x > x_p \end{cases} \quad (7)$$

This change of variables is only uniequivocal on the range $[x_1, x_p]$ so that the reciprocal function is only defined on the range $[y_1, y_p]$. To go back to the original space, we use the transformation $x=g(y)$ defined by

$$g(y) = \begin{cases} x_1 & \text{for } y < y_1 \\ x_k + \frac{x_{k+1}-x_k}{y_{k+1}-y_k}(y-y_k) & \text{for } y \in [y_k, y_{k+1}] \\ x_p & \text{for } y > y_p \end{cases} \quad (8)$$

To reduce as much as possible the region of the state space out of the interval $[x_1, x_p]$, a possible solution is to include in the list of percentiles, the minimum of the ensemble as x_1 (as percentile $r_1=1/2n$ if n is the size of the ensemble) and the maximum of the ensemble as x_p (as percentile $r_p=\frac{2n-1}{2n}$). This definition of the anamorphosis functions corresponds to the most simple parameterization of the tails of the distribution: zero probability is assumed outside the range of the

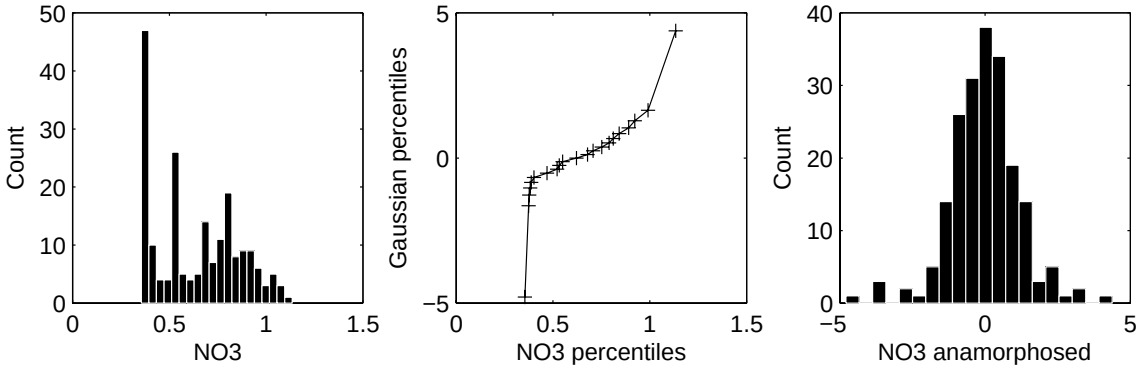

Fig. 9. Illustration of the anamorphosis transformation for nitrate at BATS station. The left panel shows the histogram of nitrate values in the 200-members 1-day ensemble forecast; the middle panel shows the piecewise linear change of variable mapping the nitrate percentiles to the $\mathcal{N}(0,1)$ percentiles and the right panel shows the histogram of the transformed variable.

ensemble forecast. This rely on the assumption that the ensemble forecast is a consistent (and thus unbiased) sample of the prior probability distribution, so that these tails correspond to a very small cumulated probability: if all statistics are correctly parameterized, only 0.5% of the updated values should fall outside the range $[-2.807, 2.807]$. If little is known about the extreme behaviour of the system, this may be a useful way of avoiding any kind of "extrapolation" outside the range of values explored by the ensemble forecast. More sophisticated options are nevertheless possible by introducing a prior assumption about the tails of the probability distribution (for instance a Gaussian assumption, as in Simon and Bertino, 2009). Whatever the parameterization of the tails, it is certainly important to check that they are not used more often that statistically acceptable (i.e. more than 0.5% in our case), which would indicate inappropriate ensemble statistics (for instance because of systematic errors), and that something should be done to improve the error parameterizations.

Figure 9 (middle panel) shows the transformation that is obtained for the surface nitrate concentration at BATS station, using p=20 equidistant percentiles (dividing the pdf into 20 equiprobable intervals), and Fig. 9 (right panel) shows the resulting distribution in the transformed space. By construction, this distribution has the same 20 percentiles as $\mathcal{N}(0,1)$ and is thus close to Gaussian. The quality of the transformation relies on one subjective choice, which is the set of percentiles r_k, k=1, …, p. The larger p, the more complex is the change of variables that it is possible to represent. But a complex transformation needs a large ensemble to be properly identified. It is certainly a good policy to keep p small with respect to the size of the ensemble ($p \ll n$), and to distribute the percentiles as regularly as possible, for instance (with p odd): $r_1 = \frac{1}{2n}$, $r_k = \frac{k-1}{p-1}$, $2 \leq k \leq p-1$, $r_p = \frac{2n-1}{2n}$. However, even a limited number of percentiles computed locally using 200 ensemble members can still be somewhat different from the asymptotic solution for $n \to \infty$. The inaccuracy that is introduced by not using perfectly stabilized percentiles is similar in nature to the inaccuracy that results from non-stabilized ensemble mean and covariance, and their effect on the accuracy of the optimal estimates should be checked with the same care. The various scatterplots presented in the paper clearly suggest that the general shape of the local anamorphosis functions would not be significantly modified by the addition of new particles.

Note that our approach is quite different from the Gaussian anamorphosis algorithm proposed by Simon and Bertino (2009) to assimilate ocean colour data in a North Atlantic model using the EnKF. In their study indeed, each model variables is transformed using the same monovariate anamorphosis function at all grid points of the model. Each function is thus computed with a much larger ensemble so that stabilized anamorphosis functions are more easily obtained. However, in view of the high inhomogeneity of the statistics over the North Atlantic, this solution would not have been applicable to our problem. The inaccuracy of the anamorphosis function resulting from an assumption of homogeneous statistics would have been far larger than the inaccuracy that results from the imperfect convergence of the percentiles of the distribution. In the present implementation, the transformation is thus computed locally using the ensemble statistics obtained at each particular grid point.

It is important to remark that with the definition (7) of the anamorphosis functions, this new solution does not introduce any spurious discontinuity in the estimation problem. If all ensemble members are spatially smooth, their percentiles and thus the anamorphosis functions are spatially smooth as well, and the spatial correlations among transformed variables can still be exploited by the observational update. However, even if no discontinuity is introduced, the anamorphosis transformations (whether f and g are global or local) are likely to modify the spatial correlation structure (i.e. the linear correlation coefficients, but not a non-linear measure like rank correlation, which is never altered

by anamorphosis transformations), and it would be interesting to investigate whether useful (linear) spatial correlations are introduced, amplified or destroyed by the transformation. This is a question that remains open (whether f and g are global or local) and it is out of the scope of this study, which concentrates on the modification of the correlations between variables at the same spatial location.

4.2.2 Observational update in the transformed space

We now apply this idea to the example presented in Sect. 4.1 (Fig. 8). We thus transform *separately* the MLD and PHY variables using their respective percentiles x_k (corresponding to: r_k=0.0025, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.9975, k=1,…,11). The transformed scatterplot is shown in Fig. 8 (middle panel). The dotted line corresponds to the Gaussian percentiles $y_k \sim -1.28$, -0.84, -0.52, -0.25, 0, 0.25, 0.52, 0.84, 1.28, k=2,…,10 (the two extreme ones are not drawn). By construction, each marginal pdf (for MLD and PHY) has got the same percentiles y_k as a Gaussian pdf. More remarkably, the mean of the transformed ensemble is close to the origin of the axes, and the regression line (green line) is close to the true regression curve (corresponding to maximum MLD probability for each PHY value): these are two features that are not guaranteed by the method and that depend on the shape of the initial ensemble distribution. Moreover, due to the transformation, the linear correlation coefficient between MLD and PHY has increased from 0.85 to 0.97. We thus observe that in this particular case, it is more appropriate to perform the ensemble observational update in this transformed space (blue dots) since moving parallely to the regression line (in green) is certainly here the right thing to do (even if there are still a few members that are significantly above the regression line).

After that, we transform the solution back into the original space using Eq. (8) (Fig. 8, right panel). As expected, the ensemble of updated values (blue dots) is closer to the true state (large blue dot). The updated ensemble error variance is thus much smaller than it was using directly the linear observational update (compare to Fig. 8, left panel). If we also transform back the linear regression line from the transformed space (the green straight line in the middle panel of Fig. 8), we obtain the thick green curve of the right panel. We observe that it is very close to the true nonlinear regression curve (maximum MLD probability for each PHY value). Performing the observational update in the transformed space is more or less like moving along this regression curve, which leads obviously to a smaller resulting error variance.

In order to analyse the situations in which the method is likely to work correctly, we now redo mentally the same exercise for some of the example scatterplots presented in Sect. 3. Four kinds of situations may be distinguished. (i) The data are well correlated and the regression line is linear (as for instance, in Fig. 3: WND/MLD, PHY/NO_3, in Fig. 4: WND/MLD or in Fig. 5: MLD/TEM, NO_3/PHY). In this situation (high rank correlation and high linear correlation), the linear observational update already exploits quite correctly the information contained in the observed variable, and only little improvement can be expected from the transformation. (ii) The data are well correlated, the regression curve is nonlinear and monotonous (as for instance MLD/TEM, MLD/NO_3, MLD/PHY in Fig. 3 and MLD/TEM, MLD/NO_3, MLD/PHY, NO_3/PHY in Fig. 4). In this situation (high rank correlation and low linear correlation, see Table 1), performing a linear observational update (following the linear regression line in green) is not a good solution, and making the simple anamorphosis described above always leads to a significant improvement. Exploiting adequately high rank correlations is the typical case for which the method is designed, and the solution is in this case closer to optimality. (iii) The data are well correlated (nonlinearly), the regression curve is nonlinear and non-monotonous (as for instance WND/MLD, MLD/NO_3 or MLD/PHY in Fig. 5). In this situation (low rank correlation and low linear correlation), our simplified method does not fully solve the problems of the linear observational update, and remains quite suboptimal. No separate transformation of the two variables can transform the non-monotonous regression curve into a straight line; a joint two-dimensional nonlinear transformation (or another method) would be needed here. However, the nonlinear method is not likely to be worse than the linear observational update. (iv) The data are poorly correlated (as can happen after a longer forecast in Figs. 6, 7 or 8). In this situation (no rank or linear correlation), transforming the variables does not help a lot: not much can anyway be expected from the multivariate observational update.

Up to now, the method has only been applied to a state vector made of 2 variables and with a perfect observation of one of the variable. However, the method is general and can be applied for any number of state variables and observations. One only needs to transform every state variables and observations separately (i.e. for every physical/biogeochemical component at every horizontal/vertical location) and perform the standard multivariate observational update in the transformed state space. If the observation operator is complex, transforming the corresponding observation requires computing the model equivalent to that observation for each member of the ensemble and find the function f given by Eq. (7) from this ensemble of value. If the observations are not perfect, a special care must also be taken to obtain a relevant Gaussian parameterization of the observation errors in the transformed space. The additional cost of these operations with respect to the linear observational update is very small so that the method can easily be applied to large size systems.

It is also worth noting that the method also solves the problem of inequality constraints that can exist on the value of some state variables, for instance $a \leq x \leq b$. The linear observational update (assuming Gaussianity) can indeed often

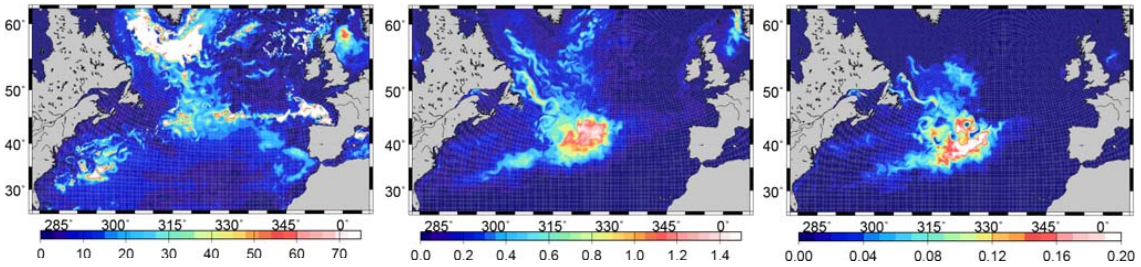

Fig. 10. Standard deviation of the 1-day ensemble forecast for the mixed layer depth (left panel), nitrate (middle panel) and zooplankton (right panel) concentrations.

violate such constraints, thus leading to inappropriate estimates. With anamorphosis 7 and 8, it is sufficient to choose $x_1 \geq a$ and $x_p \leq b$ for the final estimate to statisfy the inequality constraints. This can be compared to the truncated Gaussian filter proposed by Lauvernet et al. (2009) to solve the problem. By contrast to their approach, the method described here can only deal with inequality constraints that apply separately on each state variable. Moreover, a larger size ensemble is required to identify the anamorphosis than to identify a truncated Gaussian pdf. The truncated Gaussian filter is thus cheaper, it can deal with more general inequality constraints, but the shape of the prior pdfs is less general (truncated Gaussian pdfs are assumed).

4.3 Application of the non linear update over the North Atlantic

The operations performed in the previous section for the BATS station are here repeated at every model grid point, with the only purpose of generalizing the previous results over the whole Atlantic domain. This means that horizontal correlations are not taken into account here, and that this experiment cannot be considered as an optimal global observational update. The purpose of this simplification is still to concentrate on the improvement of the correlations between variables at the same spatial location. As in the previous section, the surface phytoplankton of the reference simulation is considered to be the observation (still assumed perfect), and the 1-day ensemble forecast at surface is used in the same way to compute the observational update (i) in the regular state space and (ii) in the anamorphosed state space. The effect of the transformation is characterized by the standard deviation of the updated ensemble.

Figure 10 shows the standard deviation of the 1-day ensemble forecast for the mixed layer depth, nitrate and zooplankton concentration before the observational update. It represents the standard deviation of the error that we want to reduce using the phytoplankton observations. The maps show that the largest MLD errors (left panel) are located in the Northern part of the domain that corresponds to large wind standard deviations (see Sect. 2). Large MLD errors usually yield large NO_3 errors (middle panel), as can be expected from the scheme in Fig. 1. In times, this leads to errors in the primary and secondary productions, that are nevertheless confined here to the Gulf Stream region (see ZOO errors standard deviations range, in the right panel of Fig. 10), because the spring bloom starts in that area at the time of this experiment (15 April).

Figure 11 shows the standard deviation reduction that is obtained with the linear observational update, i.e. the ratio of the updated ensemble standard deviation to the ensemble forecast standard deviation (that is shown in Fig. 10), and Fig. 12 shows the same result obtained using the anamorphosis scheme. These results can be analysed using the classification given in Sect. 4.2.2. (i) There are regions and variables for which the linear observational update is already very good and not much can be expected from anamorphosis to significantly improve the solution. In these regions, the variable (MLD, NO_3 or ZOO) is well correlated to PHY and the regression line is linear. (ii) There are also many regions where the error standard deviation can be substantially reduced by anamorphosis. In these regions, the variables are well correlated to PHY (high rank correlation) but along a nonlinear regression curve, so that they can be controlled through PHY observations but not with a linear analysis scheme. (iii) Finally, there are regions where nor the linear observational update, nor anamorphosis can reduce the forecast error that was induced by the wind perturbations. These errors cannot be controllable by PHY observations only. Direct observations would be necessary. This mostly corresponds to regions where the forecast error is small (see Fig. 10). Here, the wind is thus not likely to be one of the dominant sources of errors, so that no conclusion of practical consequence can be derived from this simplified study involving only wind errors.

Finally, in order to investigate the performance of the method for longer lead times, the same experiment has been repeated for the ensemble forecast at days 2, 4, 8 and 15. In order to summarize the results, Fig. 13 shows for each case study, the fraction of the domain (X-axis) for which the error reduction factor by the ensemble observational update (fully illustrated at day 1 by the maps in Figs. 11 and 12) is lower

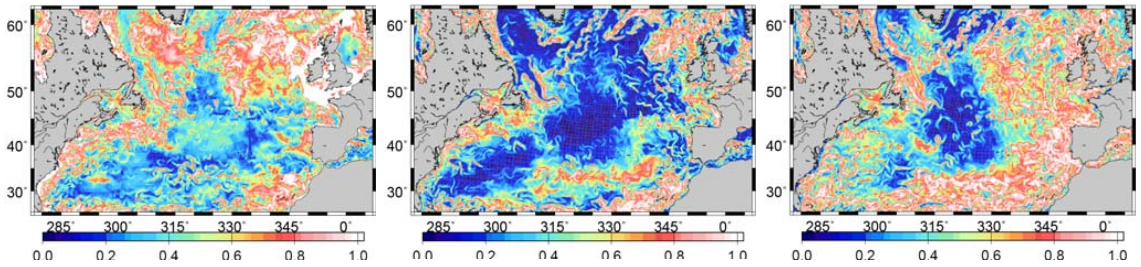

Fig. 11. Ratio of the updated ensemble standard deviation to the forecast ensemble standard deviation (shown in Fig. 10), as obtained using the linear observational update.

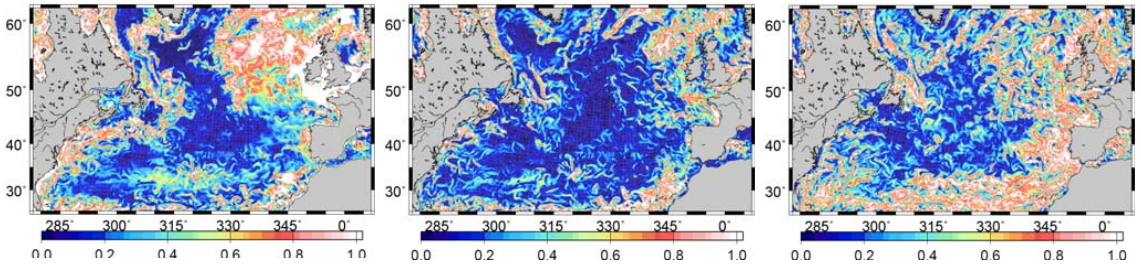

Fig. 12. Ratio of the updated ensemble standard deviation to the forecast ensemble standard deviation (shown in Fig. 10), as obtained using the nonlinear observational update (linear observational update in the transformed space).

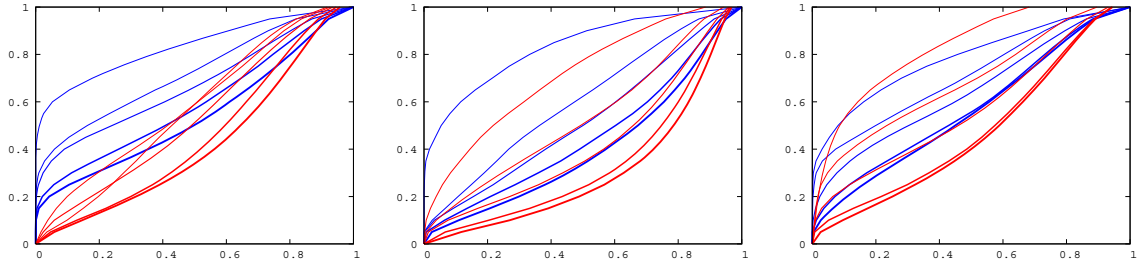

Fig. 13. Fraction of the domain (X-axis) for which the error reduction factor by the ensemble observational update (as illustrated for instance in Figs. 12 and 13 for day 1) is lower than a given value (Y-axis). The result is shown for the linear observational update (blue curves) and for the anamorphosis nonlinear observational update (red curves), at day 1, 2, 4, 8 and 15 (from thick curves to thin curves). The estimated variable is mixed layer depth (left panel), nitrate concentration (middle panel) or zooplankton concentration (right panel).

than a given value (Y-axis). Thus, the lower the curve, the largest fraction of the domain below a given reduction factor. For instance, at day 1 (thickest curves), the nonlinear observational update (with anamorphosis) is always better than the linear observational update (as already diagnosed from Figs. 11 and 12). As the lead time increases (from day 1 to day 15, from thick curves to thin curves), all three variables tend to decorrelate from phytoplankton observations (see Sect. 3.2), so that the accuracy of the estimation is deteriorating with time whatever the analysis scheme. We observe however that the nonlinear scheme remains most often significantly better from day 1 to day 15 (except for zooplankton at day 15), which means that there are many regions where nonlinear correlations can be exploited to improve the observational update.

5 Conclusions and perspectives

The Monte Carlo experiments that were designed to study mixing errors in a coupled physical biogeochemical model of the North Atlantic yield a number of conclusions in the perspective of ocean colour data assimilation. As a general rule, the results of the ensemble forecasts validate the conceptual transfer function proposed in the introduction (Fig. 1): the first order causal relationships summarized in the figure lead to tight correlations. However, the response is rather complex, depending in particular on the local stratification, in such a way that even the general features of the probability distributions can change radically in space and time (e.g. sign and strength of the correlation, shape of the regression curves, asymmetry between positive and negative

anomalies, presence of thresholds, ...). More embarassing, the tight correlations (in a nonlinear sense) observed for short term forecasts (1 day) decrease quickly with time, and thereby reduce the level of control that can be expected from a partial observing system like surface temperature and surface chlorophyll. Despite of this, our results suggest that, in many regions, a significant error variance reduction (on all variables shown in Fig. 1) can be obtained from these observations if the forecast does not exceed a few days (2 to 7 days as a function of the region), providing that the nonlinear relationships between the variables are appropriately exploited. For longer time-scales, the decorrelation observed in the ensemble runs could be the consequence of the short decorrelation time scale (4 days) adopted for the wind forcing perturbations, and it would be interesting to investigate the robustness of the results by using more persistent wind anomalies.

Nonlinearities in the model lead to many kinds of non-Gaussian behaviours, that cannot be properly handled by classical linear assimilation methods. In order to tackle this problem at moderate cost (i.e. in a way that is compatible with large size data assimilation problems), a simplified approximate nonlinear scheme has been studied in this paper. The idea is to perform a nonlinear change of variables (anamorphosis) separately for each state variable (locally in space and time), by remapping the ensemble percentiles of their marginal distribution to Gaussian percentiles. In that way, the additional cost of the observational update to make it nonlinear is negligible; the main cost is in the computation of an ensemble forecast that is sufficient in size to identify properly the transformation functions. The method has been evaluated using idealized inference experiments, in which several control variables (MLD, NO_3, ZOO) are estimated from a perfect and local chlorophyll observation. The results show that our simplified scheme is often sufficient to detect and to exploit the nonlinear relationships between observations and estimated variables, thus restoring the control of the system in situations for which linear estimation fails. In many regions of the North Atlantic, non-Gaussian behaviours are observed, which require a nonlinear estimation algorithm.

However, these experiments are still far from simulating a realistic analysis step of a true assimilation sequence, which would require at least to take explictly into account the horizontal correlations, and to introduce adequate parameterizations of the observation error statistics. Furthermore, these results have been produced for wind errors only, while many other error sources exist in basin scale CPBMs. Before general conclusions can be reached about the controllability of the system or about the least cost effective algorithm, it is necessary that similar studies be attempted for other important sources of errors, like the parameters governing the ecosystem processes, the light forcing, the vertical advection or the horizontal advection and diffusion. In following this research scenario, one should also be aware of possible nonlinear interactions between the error sources: conclusions obtained by considering them separately may no more be valid if they are present altogether.

Acknowledgements. This work is a contribution to the GMMC Mercator Vert project, and to the MERSEA (contract SIP3-CT-2003-502885) and MyOcean (Grant FP7-SPACE-2007-1-CT-218812-MYOCEAN) projects supported by the European Commission. We also wish to thank the anonymous reviewers for their useful comments and suggestions. The calculations were performed using HPC resources from GENCI-IDRIS (Grant 2009-011279).

Edited by: L. Bertino

The publication of this article is financed by CNRS-INSU.

References

Barnier, B., Madec, G., Penduff, T., Molines, J. M., Tréguier, A. M., Beckmann, A., Biastoch, A., Boning, C., Dengg, J., Gulev, S., Le Sommer, J., Rémy, E., Talandier, C., Theetten, S., Maltrud, M., and Mc Lean, J.: Impact of partial steps and momentum advection schemes in a global ocean circulation model at eddy permitting resolution, Ocean Dynam., 56(5–6), 543–567, 2006.

Berline, L., Brankart, J.-M., Brasseur, P., Ourmières, Y., and Verron, J.: Improving the physics of a coupled physical-biogeochemical model of the North Atlantic through data assimilation: impact on the ecosystem, J. Mar. Syst., 64(1–4), 153–172, 2007.

Bertino, L., Evensen, G., and Wackernagel, H.: Sequential Data Assimilation Techniques in Oceanography, Int. Stat. Rev., 71, 223–241, 2003.

Blanke, B. and Delecluse, P.: Variability of the tropical Atlantic ocean simulated by a general circulation model with two different mixed layer physics, J. Phys. Oceanogr., 23, 1363–1388, 1993.

Brasseur, P., Gruber, N., Barciela, R., Brander, K., Doron, M., El Moussaoui, A., Hobday, A., Huret, M., Kremeur, A.-S., Lehodey, P., Moulin, C., Murtugudde, R., Senina, I., Svendsen, E., and Matear, R.: Integrating biogeochemistry and ecology into ocean data assimilation systems, Oceanography, 22(3), 206–215, 2009.

Conkright, M. E., Antonov, J. I., Baranova, O., Boyer, T. P., Garcia, H. E., Gelfeld, R., Johnson, D., Locarnini, R. A., Murphy, P. P., O'Brien, T. D., Smolyar, I., and Stephens, C.: NOAA Atlas NESDIS 42, WORLD OCEAN DATABASE 2001 Volume 1: Introduction, US Gov. Printing Office, Washington DC, 160 pp., 2002.

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.

Evensen, G.: The Ensemble Kalman Filter: Theoretical Formulation and Practical Implementation, Ocean Dynamics, 53, 343–367, 2003.

Gruber, N., Doney, S. C., Emerson, S. R., Gilbert, D., Kobayashi, T., Körtzinger, A., Johnson, G. C., Johnson, K. S., Riser, S. C., and Ulloa, O.: The Argo-Oxygen program: a white paper to promote the addition of oxygen sensors to the international Argo float program, online available at: http://www-argo.ucsd.edu/o2_white_paper_web.pdf, 2007.

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 hydrostatic stability in ocean models, Ocean Modell., 27, 1–17, doi:10.1016/j.ocemod.2008.10.007, 2009.

Lenartz, F., Raick, C., Soetaert, K. and Grégoire, M.: Application of an Ensemble Kalman filter to a 1-D coupled hydrodynamic-ecosystem model of the Ligurian Sea, J. Mar. Syst., 68, 327–348, doi:10.1016/j.jmarsys.2006.12.001, 2007.

Levitus, S., Antonov, J. I., Boyer, T. P., and Stephens, C.: World Ocean Database 1998, National Oceanographic Data Center, Silver Spring, MD, 2001.

Lévy, M., Gavart, M., Mémery, L., Caniaux, G., and Paci, A.: A four-dimensional mesoscale map of the spring bloom in the northeast Atlantic (POMME experiment): Results of a prognostic model, J. Geophys. Res., 110, C07S21, doi:10.1029/2004JC002588, 2005a..

Lévy, M., Krémeur, A. S., and Mémery, L.: Description of the LOBSTER biogeochemical model implemented in the OPA system, Internal report IPSL/LOCEAN, 2005b.

Losa, S. N., Kivman, G. A., Schröter, J., and Wenzel, M.: Sequential weak constraint parameter estimation in an ecosystem model, J. Mar. Syst., 43, 31–49, 2003.

Madec, G., Delecluse, P., Imbard, M., and Lévy, C.: OPA 8.1 Ocean General Circulation Model reference manual, Note du Pole de modélisation, Institut Pierre-Simon Laplace (IPSL), France, No. 11, 91 pp., 1998.

Ourmières, Y., Brasseur, P., Lévy, M., Brankart, J.-M., and Verron, J.: On the key role of nutrient data to constrain a coupled physical-biogeochemical assimilative model of the North Atlantic Ocean, J. Mar. Syst., 75, 100–115, doi:10.1016/j.jmarsys.2008.08.003, 2009.

Oschlies, A.: Improved Representation of Upper-Ocean Dynamics and Mixed Layer Depths in a Model of the North Atlantic on Switching from Eddy-Permitting to Eddy-Resolving Grid Resolution, J. Phys. Oceanogr., 32, 2277–2298, 2002.

Press, W. H., Flannery, B. P., Teukolsky, S. A., and Vetterling, W. T.: Numerical Recipes in FORTRAN 77, Cambridge University Press, 992 pp., 1992.

Simon, E. and Bertino, L.: 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, Ocean Sci., 5, 495–510, 2009, http://www.ocean-sci.net/5/495/2009/.

The DRAKKAR Group: Eddy-permitting Ocean Circulation Hindcasts of past decades, CLIVAR Exchanges 42, 12, 8–10, 2007.

Uppala, S. M., Kallberg, P. W., Simmons, A. J., et al.: The ERA-40 reanalysis, Q. J. Roy. Meteor. Soc., 131, 2961–3012, 2005.

Von Mises, R.: Mathematical Theory of Probability and Statistics. Academic Press, New York, 694 pp., 1964.