Reply to Referee #1

We would like to thank the reviewer for his/her careful reading of the manuscript and for his/her appreciation of the work done in this paper. We did our best to take his/her remarks into account in a revised version of the manuscript (see explanation below).

General comments:

1. Section 2.2: EOFs are used to generate an ensemble of wind perturbations. It appears that the authors used the time-variance of the wind stress as the error variance of the wind field which seems excessive to me. This is equivalent to assume that the averaged wind is as accurate as instantaneous wind fields. Since the main objective of this paper is to assess the non-linear and the non-gaussian model response to wind errors, it is important to choose wind perturbations with a realistic amplitude. One could anticipate that non-linear saturation effects would be less of a problem if smaller and more realistic wind perturbations are used.

We agree with the referee that inferring the error variance of the wind from the time-variance of the wind stress is probably an excessive approximation. Perturbations of the wind forcing that are constructed using such assumption should not be considered as representative of the actual uncertainties in the wind field. However, the main objective of the paper is to investigate the overall impact of mixing errors on biogeochemical variables in the upper ocean, whatever their origin. The wind is one of the most obvious mechanisms to generate mixing errors, but these may also originate in other modelling options such as the formulation of surface heat or fresh water forcing, the turbulent closure scheme, the parameterization of deep convection etc. It is practically very difficult to quantify the relative importance of the different error sources. Given our motivation to analyse the whole cascade of mixing errors on biogeochemical variables, the generation of wind errors with realistic amplitudes is probably not an essential requirement for this study.
We recognize that some parts of the original text dealing with the actual motivation of the study were misleading, and a number of corrections have been made to improve this aspect. A new paragraph has been added in the Introduction section to clarify the objectives of the paper as follows: **"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 nonlinear estimation and data assimilation methods."** The title of the paper has been modified in order to avoid the misleading impression that controllability issues are a central aspect of this study.
We also tried to make more clear that the Monte Carlo experiments are performed by prescribing perturbations in the wind forcing which **is the physical mechanism chosen here to trigger mixing errors in the coupled model**. Further, the first paragraph of section 2.2 has been rewritten as follows to recall that the simulated wind-induced errors are in fact representative of different error sources responsible for spurious mixing in the ocean upper layer: **"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."**

2. Another issue with the perturbation scheme is the chosen time-scale of 4 days. The first EOFs represent generally slow and large-scale processes. The seasonal cycle is for *example often the most dominant EOF. However, the EOFs corresponding to the seasonal cycle is perturbed also with a time-scale of 4 days which effectively means that the model solution can switch from a summer to a wind regime over only this time scale. Later in the manuscript, the authors were able to provide a clear explanation of the model error propagation over 1 day, but noticed that over longer time scales (2 to 15) the variables tend to decorrelate. It seems to me that this negative resolution is simply a result of the short decorrelation time scale of the wind forcing and its large amplitude. Due to the highly variable error forcing, the ensemble members are quite far away from an equilibrium dynamics (and tend to move further away the equilibrium as the simulation goes on). I'm wondering how robust the presented results are relative to the time-scale of the wind perturbation (and their amplitude).*

In section 2.2, first paragraph, the statement that the uncertainty of the wind was estimated from the seasonal variability of ERA winds was indeed a mistake that has been corrected. The EOFs of the wind field are representative of the variability of the wind during a period of 3 months centered on April 15th (i.e. from March 1st to May 31st). Therefore, one can expect that the ensemble of wind stress fields will be significantly impacted by the variability of synoptic weather patterns, and this is basically what motivates the 4-day time scale adopted to sample the ERA40 archive and to build the time series of perturbed wind forcing. The 4-day time scale may not be the optimal choice, nevertheless it is clear that the computed EOFs do not represent the seasonal cycle and are not responsible for a winter/summer switch, as suggested by the referee.
It is true, however, that the decorrelation observed in the ensemble runs at longer time-scales (~ 2 weeks) could be a consequence of both the quite stong (and unrealistic) amplitude of the wind perturbations and the 4-day time scale chosen to build the perturbations. A word of caution has been added in the text to draw the attention of the reader on this point: **"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."**

3. The model state can be viewed as a time integrated response on the wind stress (and other fluxes). If the wind field suddenly changes, the instantaneous wind might indeed not be related to other instantaneous variables (which includes the effect of past winds). It might thus be interesting to look not only the relationship between instantaneous wind and the model variable but also to some integrated quantities such as the time integrated surface turbulent flux. This can still be useful in a data assimilation context because an error in the time integrated turbulent flux can be corrected in a approach similar to the incremental analysis update approach (Bloom et al. 1996, MWR).

This is a very good suggestion indeed. Computing the suggested integrated quantities such as surface turbulent fluxes, however, would necessitate a new ensemble integration with the perturbed forcing. This process would require new resources that are currently not available, but it might be envisaged as a follow-up to the present study.
Further, it should be reminded that there is no "sudden" change in the wind fields since linear interpolation is applied between independent perturbations every 4 days to obtain the 6-h forcing fields required by the model integration (see end of section 2.2). To some extent, the results illustrated by the scatterplots of figures 3-5 are precisely describing relationships between the wind perturbations at the initial time and the time integrated response of model state variables after one day. Some of the scatterplots of figures 6 and 7 are indeed more questionable, especially those showing WND/MLD relationships after 8 and 15 days. As a result, t**he first raw of Figures 6 and 7 have been removed** since the corresponding scatterplots are not essential for the understanding of the rest of the paper.

4. The authors test their scheme in a configuration which is similar to a twin experiment: they assimilate observations extracted from another model run to assess the impact on the model solution. However, here the observations are taken from the unperturbed reference simulation which is to me a questionable choice. The ensemble tends to be centered around the observations and per construction the ensemble average would be close to the "true" solution. It would be better to choose one ensemble member as the truth (and not using it to derive the statistics such as covariance and the anamorphosis transform).

The experiment discussed in the paper is indeed similar to a “twin” experiment in the sense that the data used in the observational update are taken from the unperturbed, reference simulation. We would not follow the suggestion of the referee in extracting the data from one single perturbed run, because the interpretation of the results could depend on the particular member chosen. A perfectly clean way to tackle the problem of using data from perturbed model runs would be to repeat the process for all members of the ensemble (not just for one member), and subsequently compute statistical averages of the different estimates. This would require much more computations, and make the interpretation of the results even more difficult.
From this comment and the following (as well as from other comments from Referee #2), we realized that there was a misunderstanding about the actual purpose of the “twin experiment” discussed in section 4. We actually performed a kind of “very simplified observational update" where spatial correlations are neglected, without explaining sufficiently the purpose of the experiment. This experiment was not meant to be an example of global and optimal analysis step, even less an assimilation experiment in which spatial correlations would have been an essential ingredient. In reality, Figures 11 to 14 were just added in the manuscript to provide at every possible location the kind of results presented in Figure 9, i.e. the reduction of the posterior ensemble variance if anamorphosis is applied. Our experiment is thus not really a twin experiment; it is even more idealized but it is certainly sufficient to show the effect that we want to illustrate. In the response to Referee #2, we describe how the text was modified to avoid this misunderstanding.

5. section 4.3: The authors assess the benefit of the transformation compared to the linear approach by calculating the standard deviation of the updated ensemble. The ensemble spread of the updated ensemble is indeed a measure of uncertainty but it is in this case too closely related to the update approach which the authors want to validate. A method which over-estimates (in a unrealistic way) the relationship between observed and non-observed variables would lead to a lower spread than a method which uses a more realistic assumption. An extreme case to clarify this point would be an "assimilation scheme" which updates the MLD by MLD = PHY/2. This scheme would have no spread at all since PHY observations are assumed to be perfect. But this is clearly a very bad assimilation scheme. Why the authors do not look to the ensemble RMS error between the MLD and the "true" MLD corresponding to the observed PHY? Idem for nitrate and zooplankton.

Yes, the reviewer is right. The comparison of the spread of updated ensembles is by no means a sufficient criterion to assess the quality of the observational update, because bad schemes are likely to produce ensemble spreads that have no connection with the real error. Even well-tuned assimilation systems can produce divergent error statistics, which progressively tend to grossly overestimate or underestimate the real error. Validating an assimilation scheme using the resulting ensemble spread is therefore insufficient in many cases.
But this is not what we do. We just use this simple diagnostic to illustrate the effect that non-Gaussian behaviours can have on the ensemble observational update, and to explain in particular what can happen if the nonlinearity of the regression curves is not properly taken into account (as in Fig. 9). This is sufficient to show how decisive anamorphosis can be to improve the accuracy of the observational update, but this is insufficient to validate the overall system as in twin assimilation experiments. Such a validation would have required a

better specification of other ingredients like simulating observation errors, parameterizing these errors consistently, accounting for horizontal correlations,... Doing this properly would be necessary to compute a real validation diagnostic (as suggested here by the reviewer), but it would also become more difficult to isolate the role of anamorphosis among many other effects. Our purpose here is only the demonstration of a new algorithmic feature rather than a full validation of an assimilation scheme.

Minor comments :

** section 2.1.2: what is the barotropic time step?*
The model time step is 2400 s. However there is no actual “barotropic” time step used in this model. The time-splitting approach is similar to the method introduced by Roullet and Madec (JGR, 2000) and applied in Barnier et al. (2006) and many similar configurations. We don’t feel it necessary to develop this aspect into more details here, as it is not a central aspect of the present paper.

** section 2.2: From the manuscript it is not clear if the authors combined the u- and v-components in the EOF calculation and perturbation scheme or if the components were used independently.*
The multivariate EOFs are computed by combining the u- and v- components of the wind at 10m height. As a result, the perturbations of the velocity components and not independent. **This aspect has been clarified in the text**.

** Please define acronyms such as BATS and INDIA*
We added definitions of the following acronyms in the first paragraph of section 3:
- BATS: Bermuda Atlantic Time Series
- INDIA : Ocean Weather Station India
- NABE : North Atlantic Bloom Experiment

** It would be useful to remind the reader from time to time that wind refers to wind stress and not wind speed.*
Done

** section 3.3: You describe how an update would affect successively PHY -> MLD -> Temperature. Is this chain of interaction only a conceptual view or do you propose actually to perform the assimilation in sub-steps? The standard scheme would correct directly temperature from PHY (using their covariance and possibly involving an anamorphosis). In the case of a gaussian distributed variable, an assimilation with substeps would lead to an suboptimal analysis since some processes link PHY directly to temperature without involving MLD (e.g. mortality).*
The chain of interactions illustrated by Figure 1 should be considered as a conceptual view only, because (i) in general it would correspond to a sub-optimal estimation process, and (ii) it would require additional (and in fact unnecessary) computing resources. A comment on this point has been added in the conclusions of the paper: **“In practice however, 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 sub-optimal and increase the computational complexity of the analysis step.”**

** section 4.2.1: the inverse anamorphosis transform is defined over [y_1, y_p]. However, the linear analysis update may produce a value outside of this range (in some extreme cases). What would be the strategy to transform this value back to a physical one?*
It is true that estimated values can fall outside the interval where the anamorphosis transform is defined. A detailed answer to this point is given in our response to the Referee #2, who raised a similar question (see major comment C: Truncation of the analysis).

** section 4.2.2: The back-transformed linear regression of the anamorphosed variable is referred to as the "true nonlinear regression". Unless this is a mathematical term already used previously, I would avoid calling it this way. There is not guarantee that the anamorphosed variables can be linked by a linear regression when the anamorphosis is performed on the variables independently (or if just one variable is transformed). And,*
** page 1309: " the regression line is nonlinear and non-monotonous (... Figure 5)". This is a bit confusing at first since figure 5 does not show a nonlinear and non-monotonous regression function (only a linear one). It would be clearer if the authors refer to the general tendency in those scatter plots which have those characteristics. Also "a *nonlinear* regression *line*" is a contradiction. Consider to change to wording here.*
We must admit the ambiguity of the terms used in the original version of the manuscript. We therefore changed "regression lines" by "regression curves" throughout the text, when relevant.

** page 1309: "... the proposed solution is in this case very close to optimality": As the authors know, different optimality criteria are used in data assimilation (leading to different solutions for non-gaussian distributed variables). To which definition of optimality the authors refer here?*
This sentence is somewhat redundant with the previous one in the paragraph, and is event incorrect to some extent. It has been removed as it doesn't bring anything new in the discussion.

** conclusions: "Before general conclusions can be reached about the controllability of the system or about the least cost effective algorithm, ...": should that read the "least costly algorithm" or the "most cost effective algorithm" ?*
Thank you to have pointed this error. The correct wording is indeed "the most cost effective algorithm".

Reply to Referee #2

We would like to thank the reviewer for his/her careful reading of the manuscript and for his/her appreciation of the work done in this paper. We did our best to take his/her remarks into account in a revised version of the manuscript (see explanation below).

General comments:

A. It seems to be difficult to get a general conclusion in term of controllability of mixing error from your study which appears very localized in time as well as in space. The focused period corresponds to one month of the spring bloom when I guess the nonlinearities of the biogeochemical model are the most important. Furthermore, the relations between variables are really investigated on only two stations, the third one (INDIA) being dropped without justification when looking at the temporal evolution of the ensemble response (§3.2). You should better justify this spatio-temporal localization in introduction and explain why you do not validate your conceptual transfer during the other periods of the year and why you are focusing only on the BATS and GS station.

We agree with the referee that the present study doesn't provide a comprehensive demonstration of the controllability of mixing errors based on surface chlorophyll data. This aspect is addressed at conceptual level only. In the framework of control theory, "controllability" is a very precise mathematical measure of the ability of external input data (e.g. surface chlorophyll) to move the state of a system from any (wrong) initial state to any other final (true) state, in a finite time interval. A full assessment of controllability conditions would have required a detailed analysis of the ensemble response in the different provinces of the North Atlantic, during different periods of the year, under different observation conditions etc. Such analysis was clearly out of the scope of the present study. From this point of view, the title of the original manuscript was somewhat misleading, and has been modified as follows **"Characterization of mixing errors in a coupled physical biogeochemical model of the North Atlantic: implications for nonlinear estimation using Gaussian anamorphosis"**. Further, we have removed the references to controllability and observability concepts that had been abusively introduced in section 3.3.
The approach followed in this study was rather to identify typical situations where modelling errors cannot be properly treated using conventional estimation methods based on the Gaussian assumption. The choice was made to identify a small number of situations localised in space and time that exemplify the need for estimation methods compliant with non-Gaussian error distributions. In order to better justify the selection made, the first paragraph of section 3.2 has been amended as follows: **"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 (Figure 2) and subject to strong convective events in winter. By contrast, BATS is representative of the subtropical gyre, with low 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."** These 3 contrasting situations motivate the analysis discussed in the paper and provide justification for developing the anamorphosis transformation approach presented in section 4. Regarding the temporal evolution of the ensemble response, the following sentence has been added: **"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 doesn't bring novel information about the stability of the ensemble covariance."**

Concerning the assimilation experiments, it would have been interesting to see how the introduction of anamorphosis functions impacts the behavior of the ensemble after several cycles of forecasts and analysis. I am wondering if the "substantial reduction of error variance" that you noted in your unique analysis is a result that can be repeatably observed? Furthermore did you try to restart your systems from the analyzed ensembles? What was the behavior of the error during this forecast step for both ensembles? It would have been also interesting to show results of assimilation experiments performed in winter when the behavior of the biogeochemical model is strongly different of the one during the bloom period. Did you realize experiments at that time?

We are well aware that this work is only a first and preliminary step toward the long-term objective to perfom full-fledged assimilation runs for correcting the biogeochemical state of a coupled model in presence of mixing errors. As explained in the reponse to comment C here below (see in particular the item on spatial correlations), the experiment discussed in section 4 should not be considered as a true analysis step of an assimilation sequence. As a result, the computed updates of the biogeochemical variables are not appropriate candidates to re-initialize a coupled ocean model run.

B I'm wondering why did you work with perfect observations. Real observations of surface chlorophyll present large errors. So the conclusion of your study may drastically change with the introduction of the observation error compulsory when dealing with real observations. Could you better justify this choice?

Surface chlorophyll data products derived from ocean colour satellite missions are often characterized by observation errors that are parameterized as a fraction of the remotely-sensed signal (rather than using an absolute error level). This raises both theoretical and practical issues regarding the formalism required to rigorously account for such error representations when computing observational updates. Simon and Bertino (2009) introduce an *ad hoc* strategy to specify the observation error variance as a given percentage of the observation itself which is a good idea, but we believe that a thorough analysis of this new formulation should be conducted before any numerical implementation. As the purpose of the numerical tests described in Section 4 was primarily to investigate the relevance of the anamorphosis scheme introduced in the second part of the paper, we prefered using perfect observations in order to isolate the impact of the anamorphosis transformations from other side effects.

C Anamorphosis functions.
Your construction of the anamorphosis function raises important issues that may be problematic in a realistic framework (assimilation of real observations).

We agree with the reviewer that the purpose of any such algorithm must be to deal with realistic applications, and it is true that the original manuscript was not always clear about the important statistical issues that are raised by the reviewer.

Spatial independence of the variables. *As written p.1307, l.11-13, one anamorphosis function per variable is built at each grid point. It implicitly assumes that the variables are spatially independent. It results to a spatially monovariate observational update in the transformed space (as done in §4.3). I am wondering how does this loss of spatial correlation in the update affect the performances of the assimilation? As you do not assimilate observations localized in the vicinity of the grid point, the update may be very sensitive to the error (and the relevance of its estimation specified in the filter) of the unique observation that is used. Furthermore, what is your strategy with grid points that are not cover by an observation? We can also imagine that it may lead to spatial discontinuities for the transformed variables y, the anamorphosis functions being different from a grid point to* *another one, and then to spatial discontinuities in the analyzed original variables. Did you note such phenomenon in your experiments?*

First, when constructing the anamorphosis functions, we look for a separate transformation for every variable of the state vector (**every physical/biogeochemical component at every horizontal/vertical location**). Then, when all variables are transformed, the anaysis step can be performed as usual by exploiting the (hopefully linear) correlations between variables of different nature and/or at different locations. There is no reason why spatial correlations should not be used in the same way as correlations between variables (which have also been transformed separately). The misunderstanding came from the fact that we performed a kind of "very simplified observational update" where spatial correlations are neglected, without explaining sufficiently the purpose of the experiment. This experiment was not meant to be an example of global and optimal analysis step, in which spatial correlations would have been an essential ingredient. In reality, Figs 11 to 14 were just added to the manuscript to provide at every possible location the result presented in Fig. 9, i.e. the reduction of the posterior ensemble variance (using just one perfect local observation) when anamorphosis is applied.

On the other hand, it would have been important to better explain in the paper that, even if the anamorphosis functions are different from one grid point to the next, the method 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. In the results of Figs 11 to 14, there was thus no spurious discontinuity in the estimated map. (This requires that the observations are themselves continuous, but this condition will drop as soon as horizontal correlations will be taken into account.) However, even if no discontinuity is introduced, it is true that the anamorphosis transformations (whether f and g are global or local) modify the spatial correlation structure (i.e. the linear correlation coefficients, but not a nonlinear 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.

In order to clarify the paper on this respect, the following text has been added at the end of section 4.2.1: **"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) is likely to modify the spatial correlation structure (i.e. the linear correlation coefficients, but not a nonlinear 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."** A sentence has also been added at the beginning of section 4.3 to clarify the purpose of the experiments: **"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,…"**

Stability of the anamorphosis functions.

The anamorphosis functions are built on the dynamical ensemble statistics (p.1307, l.11-13). It means that their shape may be greatly affected by the random draws generating the perturbations if the size of the ensemble is too small. Did you check the stability of your anamorphosis functions to the random process present in your Monte Carlo approach? Have you an estimation of the minimum size of the ensemble required to avoid such phenomenon?

It is indeed very likely that anamorphosis functions (i.e. percentiles of the distributions) computed locally using only n=200 particles are still somewhat different from the asymptotic solution for $n \to \infty$. To some extent, this convergence can also be questioned (and seldom verified) for the ensemble mean and covariance, which are other key ingredients of the observational update. However, 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. Moreover, no incoherence (only inaccuracy) is introduced by not using stabilized functions, and we show that, with our particular choice (even if not yet perfectly converged, and thus slightly approximate and suboptimal), we can significantly improve the linear correlation coefficients between observed and unobserved biogeochemical variables. It is also worth noting that assuming that the problem becomes Gaussian after univariate anamorphosis transformation is already a suboptimal approximation, which (in our problem) is much more penalizing that using imperfectly stabilized anamorphosis functions.

In their study, Simon and Bertino (2009) solve this stability problem by constructing one single anamorphosis function for every model variable, whatever their spatial location. In that way, by mixing all data from all model gridpoints, a much larger ensemble is available and stabilized anamorphosis functions can be 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 order to mitigate the two risks (inhomogeneity of the statistics or non-convergence of the anamorphosis function) simultaneously with a limited number of particles, an intermediate solution would be to compute the anamorphosis functions using all data within a given distance (with a weight decreasing to zero to avoid any discontinuity), but this technical solution would introduce additonal unknown parameters which could only be tuned by sensitivity experiments or quite arbitrary assumptions. In order to better explain this point, the following text has been added in section 4.2.1: (i) at the end of the 3rd paragraph: **"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."** and (ii) in the 4th (and last) paragraph: **"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."**

Truncation of the analysis. *It seems that you do not define tails to the anamorphosis functions. The minimum and maximum values of the transformed random variable y are* *defined by the percentile $r_1 = 1/2n$ and $r_p = (2n\text{-}1)/2n$. With n=200, it leads to $y_1 = -2.807$ and $y_p = 2.807$ (p. 1307, l.5 - 18). How do you process values out of $[-2.807, 2.807]$ that may appear during the analysis steps in the transformed space? Do you truncate them to x_1 and x_p when pulling back in the original space (as written in the manuscript)? In that case the functions would not be bijective. Furthermore you may favor arbitrary values (x_1 and x_p) due to the truncation, that are strongly influenced by the bias model. Would not it be better to consider to extend towards infinity the range of the transformed values to prevent such truncation?*

About the tails of the prior probability distributions (which would define the tails of the anamorphosis functions), there is no information in the ensemble forecast. The parameterization of these tails thus rely on prior assumptions, which can depend on every particular application. A classical choice would be to go back to a Gaussian assumption for the tails of the distribution (as in Simon and Bertino, 2009), but any other parameterization is possible according to the expected behaviour of the system. In our application, we decided to parameterize these tails by giving zero probability to any value outside the range of the forecast ensemble. This relies 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]. (By our function g, updated values outside this range are reset to x_1 and x_p.) 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 that have been explored by the ensemble forecast. Because extrapolation with a bad assumption about the shape of the tails can easily lead to unrealistic estimates (especially for unobserved quantities), such as with the linear observational update if the Gaussian assumption is not valid.

On the other hand, it is true that with this parameterization of the tails (zero probability outside the ensemble range of values), the functions are not bijective (g is the inverse of f in the ensemble range only), so that the terminology “anamorphosis” is somewhat improper. We can argue that our solution is the limit of exponential or Gaussian tails with an exponential or Gaussian decrease rate tending to infinity. But to be more rigourous, we should say that we perform the anamorphosis transformation between intervals $[x_1,x_p]$ and [-2.807,2.807] and assume a truncated Gaussian distribution on this transformed interval.

In order to clarify this point, the following text has been added in section 4.2.1 (at the end of the second paragraph): **“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 forecast ensemble. 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).”**

Model bias*. At each horizontal grid point, the data set used to build the anamorphosis function is made of the values of the n members of your ensemble. It means that the transformation will be strongly affected by the model bias. The way to define dynamically the interval of the original variable x (p. 1307, l.5 - 18, the bounds being defined by the forecast ensemble) may lead to an exclusion of relevant values for the observational update. For example, how do you deal with observations that will not be in the range [x1; xp]? This local adjustment of the anamorphosis functions using the ensemble statistics (p.1307, l.11-13) requires a model without or with a low bias. Well the coupled physical biogeochemical* *models, particularly the biogeochemical component, can present strong biases, that may damage the observational update. How will you take into account the problem of bias in future realistic configuration?*

The existence of model or observation biases is a general difficulty in data assimilation problems, which can only be solved by assuming that one of the input information is unbiased (e.g. one of the observation dataset) and by trying to shift the other input information (i.e. their probability distribution) using this reference. If the biased input information is an ensemble model forecast, solutions must be found to "shift" the ensemble towards the right region of the state space. If everything can be assumed linear, this problem reduces to shift all forecast particles by the same quantity, which must be computed using the reference dataset. If the model is nonlinear, the structure of the ensemble can depend on the existence of the shift, so that the model operator must be modified (e.g. by adding process noise) until unbiased forecast are produced. In this case, it is essential that the ensemble forecast explore the right region of the state space. This is the only way to obtain relevant statistics and thus optimal estimates.
There is no bias in the present study since, by construction, the experiments are idealized and the ensemble forecasts are consistent samples of the forecast error probability distributions. Such consistency, however, is difficult to obtain in realistic applications. The ensemble forecasts are sometimes incompatible with the observations (i.e. the forecast and observation prior probability distributions only overlap in regions with negligible prior probability). This is can be symptomatic of inappropriate parameterization of model or observations uncertainties, which it is dangerous to compensate by relying on poor parameterization of the tails of the distributions. This would certainly not be the best way of exploiting the rich information that can be obtained by an expensive ensemble forecast. If tails with specified Gaussian shape should be used systematically to compensate for a model bias, it is certainly more efficient not to perform the ensemble forecast and use directly a specified shape of the distribution (with ensemble OI for instance). On the contrary, if the estimated fields are found too often outside of the ensemble bounds (with a frequence larger than 0.5% in our case), this should be diagnosed, and something must be done to modify the error parameterization.
A word of caution has been added in section 4.2.1 (at the end of the second paragraph) to insist on the necessity of diagnosing inconsistent statistics, especially if the observations are systematically found outside of the ensemble range: **"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".**

Observation error. *I am not sure to understand your suggestion to deal with the observation error in the transformed space. If I am not mistaken, the observation error standard deviation in the transformed space is obtained by transforming the one from the original space with the "local slope" of the anamorphosis function. When you say "local slope", does it correspond to the slope associated with the percentile of the observation error standard deviation or the one associated with the value of the observation? Furthermore, I am wondering if this sort of "linearized" approximation can lead to the introduction of bias in the estimation of the observation error in the transformed space (overestimation or underestimation)? For the case of chlorophyll observation for which the error can be assumed to be lognormal, would it be easier to directly define the transformed observation error standard deviation by the percentage of error assumed for the original observation?*

The explanation about observation errors (only one sentence) was not very clear in the paper. We wrote it as a side remark, because we assumed perfect observations (and we only used the regression formula (6) locally in space). But the subject would indeed require more detailed explanations.

Since anamorphosis is providing a transformed control vector, the direct way to deal with the observations is to modify the observation operator by adding the function g on the right of the original operator H, i.e. g is applied to the transformed control vector to go back to the original control space and then H is applied to go to observation space. In addition, in the Kalman filter formula, we will need the linear tangent of this modified observation operator. If the original H is linear, this only requires multiplying H by the local slope of the g functions. In this case, we keep the original observations and we also keep the original parameterization of the observation error covariance matrix. However, with this method, we would usually end up with a nonlinear observation operator, even if the original H is linear. This is thus not a very good solution and it is certainly better to also transform the observations so that the observation operator remains close to linear.
With our transformation method, which determines the anamorphosis functions f and g from the percentiles of the original ensemble, a possible solution is to obtain similar anamorphosis functions in observation space (from the percentiles of the same ensemble transformed in observation space by the observation operator H), and use these functions to also transform the observations. With this solution, if the original H is linear, it is likely to remain close to linear with the transformed variables. (If the original H is just a linear interpolation, it is even possible to keep the same linear interpolation in the transformed space, knowing that this will correspond to another kind of interpolation for the original variables.) Nevertheless, in order to apply the linear observational update formulas with these transformed vectors, we need to postulate a Gaussian observation error probability distribution for the transformed observations. And since observation error statistics are usually given for the original observations, they need to be approximately converted into a Gaussian assumption for the transformed observations.
If the observation error standard deviation is small enough (i.e. if the slope of f does not change much over a few observation error standard deviations), then a Gaussian probability distribution for the original observations approximately transforms into a Gaussian probability distribution for the transformed observations. And the error standard deviation for the transformed observation can easily be obtained by multiplying the error standard deviation for the original observations by the local slope of the anamorphosis function (i.e. the slope of the anamorphosis function computed in observation space, around the value of each observation). If the observation error standard deviation is not small, then a Gaussian probability distribution for the original observations does not transform into a Gaussian probability distribution for the transformed observations. And assuming a Gaussian error probability distribution for the transformed observations (as we must do), means assuming a non-Gaussian distribution for the original observations. Hence, it is true that, in this case, unbiased original observations can transform into biased observations. Specific care should thus be taken to adjust the standard deviation and the mean of the transformed (Gaussian) distribution in such a way that, when transformed back into the original space, it becomes as consistent as possible (no bias and similar dispersion) with the postulated observation probability distribution. (Incidentally, for chlorophyll observations, it would not be correct to define the transformed observation error standard deviation by a percentage of error as for the original observations, because our anamorphosis transforms the median of the original ensemble into zero. Thus, if the original observation error is proportional to the observation value, the observations smaller than the median [negative transformed values] must have a relatively smaller error, and the observation larger than the median [positive transformed values] must have a relatively larger error. This cannot be a percentage of the transformed value, which would only results in giving low confidence to observations that are far from the ensemble median.)
Overall, this discussion shows that this subject is well beyond the scope of our paper (in which observations are error free), and that it is impossible to provide a correct presentation of the problem in a few sentences. Hence, since this is not needed to understand the method that we propose and the results that we obtain, we prefered to remove from the paper any reference to the parameterization of observation error in the transformed space. We have thus replaced the inaccurate sentence of the original paper by: **"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."

D §4.3 Application of the non linear update over the North Atlantic
In your discussion, the impact of the introduction of anamorphosis function is only evaluated by the reduction of the spread of the ensemble. The framework of twin experiments allows you to access to the true state, that's why I am wondering why are you not looking at the RMS error of the solution? This diagnostic could be useful to check the relevance of the estimation of the error evaluated from your ensemble (P_f and P_a). So p.1312, l.12, you say: "Fortunately, they mostly corresponds to regions where the forecast ensemble error is small" when talking about areas where both assimilation methods are not efficient. First, it seems to be reasonable to get low corrections in areas where the filter diagnoses low errors (low forecast ensemble error). Secondly, it does not mean that the RMS error is low as the ensemble may underestimate the error. Unfortunately, these areas of low forecast ensemble error may also be areas of significant error (areas of strong model bias for example). You should add this diagnostic and also remove "Fortunately".

As indicated in our response to Reviewer #1 who raised a similar question, this simple diagnostic is just used to illustrate the effect that non-Gaussian behaviours can have on the ensemble observational update, and to explain in particular what can happen if the nonlinearity of the regression curves is not properly taken into account (as in Fig. 9). This is sufficient to show how decisive anamorphosis can be to improve the accuracy of the observational update, but this is insufficient to validate the overall system as in twin assimilation experiments. Such a validation would have required a better specification of other ingredients like defining properly a reference simulation, simulating observation errors, parameterizing these errors consistently, accounting for horizontal correlations,...
Regarding the sentence p. 1312, l.12: we agree to remove the word "Fortunately" which is indeed not necessary, but we would disagree with the statement that the ensemble may underestimate the error, because the idealized experimental conditions precisely yield an ensemble that is perfectly consistent with the actual (unbiased) error in the system.

E I am wondering what are the differences between the Monte Carlo method used in your experiments and an Ensemble Kalman filter? It seems to be the same methods but I was not able to find the expression EnKF in the manuscript:
– p.1291, l.20: reference to Evensen (1994). You could add also Evensen (2003) and/or Evensen (2006).
– p.1292, l.7: "the Monte Carlo method".
– p.1304, l.10-15: it seems that you use an EnKF with perfect observations (Evensen, 1994). If the method used to assimilate data is an EnKF, please write it explicitly in the manuscript.

As far as we know, a Monte Carlo method is the generic designation of a method based on random perturbation experiments (e.g. to forecast error statistics as in Evensen 1994), which do not necessarily involve a filtering procedure. We have checked the correct wording throughout the text and **added citations to Evensen (2003) and Evensen (2006) whenever appropriate**. We did not mention EnKF initially because we do not perform assimilation experiments in this paper, and because what we say is not particular to EnKF. But since it is true that the method is well-suited to work with EnKF, it is now **mentioned in the paper with appropriate references**.

Minor comments

• *p.1290, l.22: "a simple nonlinear change of variables". You could introduce the word "anamorphosis". It will be easier to understand the title.*
We believe that the original wording is preferential in the abstract, whereas in the core text we first introduce the term "anamorphosis" and then explain the change of variables.

• *p.1291, l.23: "One then postulates a prior probability distribution for these errors". You could be more explicit and talk about Gaussian distribution as you perform Kalman filter analysis in the manuscript.*
Not at this point of the manuscript, since the goal is precisely to explore the transfer function of errors which are not Gaussian in general.

• *p.1291, l.28-p.1292.l1: ", if these relationships are linear". No sense: the theory of Kalman filtering assumes that the model and observation errors are independant.*
Thanks, the sentence has been corrected as follows: "if robust relationships exist between observed and unobserved variables".

• *p.1293, l.14: "A possible approach to nonlinear estimation problems is the use of anamorphosis transformations (Bertino et al, 2003)". The introduction of anamorphosis function is more dedicated to deal with the non-Gaussianity of the variables rather than performing nonlinear estimations. The EnKF with Gaussian anamorphosis suggested by Bertino et al can be interpreted as a linear estimation method (the analysis scheme being linear in the transformed space). Please rewrite the sentence.*
The original sentence was indeed a bit vague; it has been rewritten more accurately by replacing "nonlinear" by "non-Gaussian". In our view however, the EnKF complemented by the type of anamorphosis transformation which is presented in this paper could not be considered as a linear estimation method because (i) the ensemble forecast step in non-linear in essence, (ii) the anamorphosis functions are determined interactively using the ensemble statistics information. This is significantly different from a method where the anamorphosis function would be specified a priori (e.g. as a log function) without consideration of the ensemble statistics.

• *p.1293, l.19-20: " The simplified solution that we propose in this paper is to perform the change of variable separately for each state variable". The use of monovariate anamorphosis functions corresponds to what has been suggested by Bertino et al (2003). Please rewrite the sentence.*
The sentence has been corrected: "The simplified solution that we investigate in this paper is to perform the change of variable separately for each state variable, as initially proposed by Bertino (2003)."

p.1293, l.24-25: "However, it is usually possible to diagnose [...] is not". If you are talking about the multi-Gaussianity of the state vector, you will have to perform qui^2 tests. Unfortunately, it is very expensive for large systems (as coupled physical biogeochemical model).
We are not really talking about such things. The word "diagnose" is perhaps misleading and has been replaced by "detect".

• *p.1294, l.9: "central model simulation" → "control simulation" (?).*

"reference simulation" is more appropriate here.

• *p.1295, l.20 and l.24: "initialised" and "initialized". Please unify.*
Done

• *p.1296, l.12: "interanual" → "interannual" (?).*

This word has been removed in response to a comment made by Refereee #1.

• *p.1296, l.19: "EOF". Please define the acronyms.*
Done

• *§2.2 Wind ensemble perturbations*
The ensemble is generated by the introduction of wind perturbations. These perturbations are built from an EOF analysis of ERA40 winds. It raises the problem of the transfer of biases present in the ERA40 database to the perturbations. We can imagine that such process may favor particular structures of perturbations, leading to a less relevant modelling of the ocean error subspace. Did you note such problem? Furthermore, what are the expected benefits for this EOF approach comparing to a spectral method (Evensen, 2003) for example?
We haven't noticed such problems but we haven't either analysed in detail the structure of the wind perturbations.

• *p.1297, l.10: "perturbated" replaced by "perturbed".*
Done

• *p.1298, l.10-l.28: you should consider to remove the paragraph dealing with the rank correlation. This diagnostic appears only in the table 1 and the results are not qualitatively exploited in the discussion.*
The rank correlation is an interesting statistics which is not altered by anamorphosis transformation. A comment has been added in the discussion to emphasize this nice conservative property, and this is also why we suggest to keep this paragraph.

• *p.1300, l.1-6: I am not sure to understand the explanations given for the scatterplot at INDIA station. How can you explain this shape of the scatterplots? Is it due to the layers under 400m that are too thick to "feel" and transfer the wind perturbations? Furthermore, why does the increase of the wind stress lead to a slight decrease of the mixed layer depth?*
We tried to clarify the explanations as follows: **"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."** As explained later in this section, this is probably due to the absolute value of the mixed layer thickness itself, but also to the difficulty to diagnose precisely the MLD from a numerical model with finite discretization levels on the vertical. The slight decrease of the mixed layer depth (~ 2 m) is probably not very significant.

• *p.1301, l.6-7: "The INDIA station still shows a complex response which is difficult to interpret by simple mechanisms". Do you have a precise idea of the processes involved in this area, even if they are not simple? Did you note equivalent scatterplots in other stations (not shown and discussed in the manuscript) or are they only localized in this area?*
We haven't noticed similar responses in other stations. Anyway, the MLD values below average correspond to very low winds and high phytoplankton concentration near the surface. This probably corresponds to weak turbulence level, leading to sufficient residence time of phytoplankton cells in the euphotic layer to sustain primary production. When the wind increases, the lower phytoplankton concentration could result from the combined effect of increased mixing and smaller residence time of phytoplankton in the euphotic layer. But this is only a speculative interpretation which needs to be consolidated.

• *p.1302, l.27 -p.1303, l.1: "In particular, relationships [...] PHY". Is it not too ambitious to try to generalize the forecast length of your data assimilation system from a study realized at a given datum on an unique point? For example the 4-day length seems to be relevant in april at BATS station but not at GS station. How can we deal with such localized (in time and space) information?*
We fully agree with this comment, and by the way we didn't pretend that the 4-day length scale should be used to generalize the time window of an assimilation scheme at other stations or situations. The BATS test case is developed without any other ambition than the simple illustration of the conceptual mechanism further described in section 3.3.

• *§3.2 Temporal evolution of the ensemble response It would have been interesting to look at the evolution of the ensemble over 2 weeks at INDIA station. Your conclusions for short term forecast (1 day) in §3.1 is that the mixing being low (due to the stratification of the water column) in this area, other processes are enough important to significantly influence the dynamics of the biogeochemical variables. We can imagine that the increase of the spread of the ensemble, resulting from long forecasts, may lead to a destabilization of the structure of the water column (at least for several outliers), and then to an increase of the dominance of the mixing in the behavior of the system.*
The interpretation of the Referee makes sense indeed, but the temporal evolution of the ensemble at INDIA over two weeks doesn't yield new conclusions with respect to those already obtained from the analysis of BATS and GS evolutions, i.e. (i) increase of the ensemble spread with time and (ii) decorrelation of the model variables. This is why the discussion of the ensemble evolution in this section has not been further developed.

• *§3.3 Observability of physical and biogeochemical variables using chlorophyll data.*
The benefits for the manuscript of including this discussion are not obvious. First, the way to get the update ensemble (blue points on figure 8) is not clear. For example, we have to read §4 to understand why there is only one abscissa. Furthermore it does not bring more information than what was written previously in §3 and what is written in §4. You should consider to remove the paragraph.
We agree with the referee that this section is not essential to discuss the key points of the paper. Section 3.3. and Figure 8 are therefore removed in the revised version of the paper.

• *p.1304, l.7: "non-Gaussian behaviours". It is more or less the first occurrence of the non-Gaussianity of the variables in the manuscript. It would have been more relevant to diagnose the non-Gaussianity of the variables in the previous discussions.*
We have answered to that comment by correcting the way the concept of anamorphosis is introduced in the paper, including non-Gaussianity issues etc. (see in particular the response to related comment p 1293 here above).

• *p.1305, l.16: "formula (6) rewrites" replaced by "formula (4) reads" (?).*
Corrected.

• *p.1306, l.17: "the true regression line has a general positive curvature". The expression "true regression line" seems not to be suitable, even if the mean is understandable. Maybe "segmented regression" would be more relevant. Idem p.1308, l.24. • p.1309, l.15-29: The expressions "regression line is linear" and " regression line is nonlinear" seem to be not suitable. Maybe you should talk about linear/nonlinear relations between the variables.*
The wording has been changed: whenever relevant, "regression line" has been replaced by "regression curve" throughout the text.

• *p.1309, l.26: "always leads to a significant improvement". Is it a result from additional experiments or a conclusion extrapolated from the figures shown in the manuscript? You state that the linear relationship obtain between the transformed variables on figure 9 can not be guaranteed. Can you justify this generalization?*
"always" is not appropriate in this sentence and has been removed.

• *p.1310, l.6-9 : case (iv). Even if the anamorphosis functions do not significantly improve the correlations between the variables, you may get benefits from the improvement of the distribution of the transformed variables. It seems to be worth to transform the variables in that case also.*
We agree (as indicated in the text), but we just say that "not much can be expected from such a transformation".

• *p.1310, l.15-17: You should remove the brackets.* Done.

Reply to Referee #3

We would like to thank the reviewer for his/her careful reading of the manuscript and for his/her appreciation of the work done in this paper. We did our best to take his/her remarks into account in a revised version of the manuscript (see explanation below).

General comments:

Major comments:

Part. 3. Study of the ensemble forecast (p. 1297) ". . . at a dozen of locations in the North Atlantic. . ." The paper could be improved if the authors have discussed in some more details Table 1 (the spatial variability of the considered relationships between the physical/biogeochemical components) and then justified their choice of the 3 locations of interest. For instance, when discussing TEM/MLD relationships, correlation at the Labrador Sea, Norway, Newfoundland stations could be mentioned (with some explanation) as an exception in generally obtained negative correlation. The authors should try to interpret the WND/MLD (and MLD/NO3 and MLD/PHY) relations at the weather station INDIA. Convection is the main mixing mechanism at the latitudes. It could be that changes in wind stress may intensify or contrary make weaker the mixing (if less dense surface water is coming to the location due to wind circulation).

We repeat here below part of the response given to Referee #2 regarding the justification of BATS, GS and INDIA stations. The approach followed in this study was to identify typical situations where modelling errors cannot be properly treated using conventional estimation methods based on the Gaussian assumption. The choice was made to identify a small number of situations localised in space and time that exemplify the need for estimation methods compliant with non-Gaussian error distributions. In order to better justify the selection made, the first paragraph of section 3.2 has been amended as follows: **"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 dominate by strong wind variability (Figure 2) and subject to strong convective events in winter. By contrast, BATS is representative of the subtropical gyre, with low winds and well stratified upper layer throughout the year. The GS station is located in the inter-gyre region, with intermediate wind variability and moderate stratification conditions."**
Further, **a new paragraph has been added to describe the main trends identified from the statistics of Table 1, and a number of possible interpretations as suggested by the referee**.

Here the question about WND perturbations arises. How did the authors perturb the wind stress: whether they considered u and v component independently or already combined amplitude?
The perturbations of the wind stress which is applied to the ocean model result from perturbations of the wind velocity at anemometric height. In order to build the ensemble of forcings described by equation (1), multivariate EOFs of the ERA winds are computed by combining the u- and v- components of the wind. As a result, the perturbations of the velocity components and not independent. **This aspect has been clarified in the text**.

p. 1301, when making a discussion on NO3/PHY relationship, it would be better to mention about surface nutrient (NO3) consumption while phytoplankton growing, than explaining the obtained negative correlation by ". . .inverse distribution of those two quantities over the water column." Such a distribution could be the consequence but not the reason.

The proposed correction should certainly clarify the discussion. The first part of the discussion have been rewritten as follows: “**Surface phytoplankton generally decreases when nitrate concentration increases. This general trend is consistent with the basic mechanism of phytoplankton growth, which requires nutrient consumption in the euphotic layer. This process generally results in the observation of inverse distributions of phytoplankton and nutrient over the water column. The scatterplots can be characterized by well-defined relationships with pretty high correlations, sometimes altered by threshold effects such as illustrated for the Gulf Stream Station.”**

Assessing the filter performance with and without the anamorphosis transformation, would it be possible to present RMS error relative to reference solution (which used to generate chlorophyll data for assimilation in the twin experiment). It would be also of interest to see results of a next data assimilation step (since reduced ensemble spread does not always guarantee better forecast and state update at the next analysis steps).
A response to this comment has already been made to Referee #1 and Referee #2 regarding the same sort of issue. If a true filter had been implemented, it would certainly be possible to compute RMS errors relative to pseudo-observations generated in the framework of twin experiments. But this has not been done yet. We are aware that this work is only a first and preliminary step toward the long-term objective to perfom full-fledged assimilation runs for correcting the biogeochemical state of a coupled model. As explained in the reponse to comment C from Referee #2, the experiment discussed in section 4 should not be considered as a true analysis step of an assimilation sequence. As a result, the computed updates of the biogeochemical variables are not appropriate candidates to re-initialize a coupled ocean model run.

Minor comments:

p.1291, “Coupled physical–biogeochemical models” it would be a better place/time to introduce your acronym CPBM;
Done.

p. 1292, “a Monte Carlo method” instead of “the Monte Carlo method”;
Done.

p. 1295, “the December climatology (Conkright et al., 2002)” instead of “the December Levitus climatology 2001 (Conkright et al., 2002)” ?
Of course, corrected.

p. 1297, WND, what kind of wind stress characteristic (u, v components or. . .) is considered?
Done.

p. 1297,1298, acronyms WND, MLD, TEM, would be better to define them, as well as BATS, INDIA NABE
Done.

p. 1303, “ the cascade of errors from WND to MLD, from MLD to TEM, and finally from TEM to PHY ” What about direct “from MLD to PHY”? The authors consider the period of phytoplankton bloom, and the main mechanism of this event is shallowing MLD + increasing irradiance. Analyzing MLD/PHY correlation (p. 1301) the authors could also make an accent on this event.

Thank you for this remark. The direct cascade of errors from MLD to PHY has to be mentioned too. Regarding the negative correlation between MLD and PHY, a further comment is made: **"This negative correlation could be interpreted as either the result of passive mixing of phytoplankton in the mixed layer, or the combined effect of shallowing MLD and increasing irradiance, which typically occurs during bloom events."**

p. 1304, Why is the results for INDIA station not presented? "
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 doesn't bring novel information about the stability of the ensemble covariance.

p. 1309, when distinguishing the "four kind of situations" of where and how well the state update with the anamorphosis transformation performs, would it be worth referring to the locations (BATS, GS, . . . INDIA stations) where the certain situation exists but not only to the figures (Fig 3, 4, 5).
We have made a number of corrections along these suggestions.

p. 1310, " . . .the method has been only applied to a state vector made of 2 variables. . ." please repeat which variables.
Done.

Fig. 7, 8, Would not it be possible to use the same scale for similar scatterplots.
Figure 8 has been removed in response to a comment from Revewer #2.

Finally, all typos or misprints pointed out by the reviewer have been corrected.