Ocean Sci. Discuss., 9, C618–C628, 2012
www.ocean-sci-discuss.net/9/C618/2012/

Ocean Science Discussions

Interactive comment on "Toward a multivariate reanalysis of the North Atlantic ocean biogeochemistry during 1998–2006 based on the assimilation of SeaWiFS chlorophyll data" *by* C. Fontana et al.

Anonymous Referee #1

Received and published: 27 June 2012

In this paper, the authors investigate the ability to provide multivariate reanalysis of the ocean biogeochemistry by assimilating surface chlorophyll concentrations with variants of the Kalman filter. For that purpose, two systems based on a SEEK filter with static error sub-space assimilate SeaWiFS chlorophyll data every 8 days in a North Atlantic configuration of the coupled model NEMO-LOBSTER during 9 years (1998-2006). Observations of the physical ocean are not assimilated. The first system applies the traditional linear scheme of the SEEK filter while the second system includes anamorphic transformations to perform the analysis step with Gaussian distributed transformed

variables and observations as suggested by Bertino et al. (2003). Both systems approximate the forecast error covariance matrix from historical ensembles based on the same deterministic simulation which is also used as a control run for the validation. The performances of the data assimilation systems are assessed with respect to the assimilated SeaWiFS data and nitrate concentrations from the World Ocean Atlas (in-situ data and climatology). The results shown suggests that the system including the anamorphic transformations provides better estimates of the ocean biogeochemistry than the traditional linear update.

I find the manuscript globally well written. However, I am not the most competent to assess the level of English. I think that several points in the manuscript require clarifications. It concerns the diagnostics that are shown and more importantly, the description of the anamorphic transformations which is too succinct. The strategy chosen to build and apply the anamorphic transformation raises significant issues that are not addressed in the manuscript. However, I can I have misunderstood what is really done. In that case, it strengthens my opinion that the authors should provide technical details on their anamorphic transformations.

For that reasons, I would recommend revisions before considering the publication of this manuscript.

1 Major comments

A Time-averaging of the simulations

Most of the figures shown in the manuscript correspond to time-averaged diagnostics over N-day periods, with N being successively equal to 60, 16 and 30. However, the variables used to perform the averaging are not clearly specified. Does it only include the outputs of the analysis steps or does it include daily outputs obtained during the forecast steps? For example, considering the 60-day window used in Figure 2, do you average 7-8 analysis outputs or 60 model outputs (including the analysis)? A clear definition of the time-averaging is required to better assess the results.

B Description of the Gaussian anamorphosis

I think that the description of the anamorphic transformations in §3.2 is not specific enough. As written, the strategy to build the anamorphosis functions is equivalent to the one used in Doron et al. (2011). If I am not mistaken, it means that for each grid cell (i, j, k), the monovariate transformations $\phi_{i,j,k,n}$ (one per variable) are built from the historical temporal ensemble that is used to compute the EOF at time n. Each sample is made of 210 values and the tails of the distribution are defined by assuming zero probability to values out of the range of the sample. Am I correct? I think that this information should be specified in §3.2 "Specific setup of the assimilation system" in order to better understand what is done in this study.

Furthermore, the strategy to specify the observation error for the transformed observations is not described. The reference to Doron et al. (2011) is not really helpful, because Doron and coauthors performed twin experiments with perfect observations leading to the specification of a very low value for the observation error. However, the experiments presented here assimilate real observations with large errors. It is mentioned that the error "associated with each distinct observation pixel is set to 30% of the considered data". I presume that this sentence stands for the system with the linear update. How does it work in practice for the system with the Gaussian anamorphosis? This must be specified in the manuscript.

C Tails of the anamorphosis functions

Following the approach suggested in Doron et al. (equations (9) and (10)), the tails of the anamorphic functions are defined by truncating the values out of the

range of the ensemble locally built from the historical simulation. The transformations are not bijective out of the range of the ensemble (no extrapolation): the values larger than the maximum (resp. lower than the minimum) of the ensemble are assigned to the same "Gaussian value" than this maximum (resp. minimum). In practice, it means that the values larger (lower) than the maximum (minimum) of the ensemble are considered as equal to this maximum (minimum).

On the one hand, the transformations are locally built - one ensemble per grid cell - from a deterministic simulation with a model which can be biased (poorly known parameters, erroneous parameterizations,..) and can present large errors (complex coupled system). On the other hand, observations present large errors and can also be biased. Consequently, the number of occurrences of observations out of the range of the local ensembles can be large in such a framework. What is the strategy adopted to transform these observations? Are they truncated to the bounds of the ensemble? If yes, I am wondering what is the impact of these truncations on the results shown in the manuscript.

Because the transformations are not bijective it means that the system with the linear update and the system including the Gaussian anamorphosis do not assimilate the same set of observations. Let y be a phytoplankton observation built from the SeaWiFS data and x_1 and x_p be the minimum and maximum of the local grid cell ensemble built from the historical simulation. The observation y is assimilated in the system with the linear update. However, the system with the nonlinear update assimilates the observation $\tilde{y} = \begin{cases} x_1 & \text{for} \quad y \leq x_1 \\ y & \text{for} \quad x_1 \leq y \leq x_p \\ x_p & \text{for} \quad x_p \leq y \end{cases}$.

In practice, the system with anamorphic transformations might assimilate data defined from the free run (one bound of the sample) depending on the season and the area (for example region #2 during spring and summer?). The frequency of occurrences of these "synthetic" data must be quantified because assimilating data from the free run simulation (potentially erroneous and biased) is highly questionable. It is necessary to evaluate the time evolution of the differences between the data sets assimilated in the two different systems. It would not be surprising to note large differences between the solutions obtained with the different systems if there are significant differences between the assimilated observations.

For the same reason, the transformation back to the original space leads to analyzed state variables in the ranges defined by the local historical ensembles. In practice, the solution is constrained every 8 days to be in a range of values defined from the historical run for each grid cell. These constraints specifying maximum values for the solution are not present in the system assimilating observations with the linear update. This can explain why we note a drift towards large values for the nitrate in the solution obtained with the linear update and not in the solution obtained with the nonlinear update. Again, it would be worthy to investigate how systematic are the truncations due to the transformation of the variables.

I am wondering why the authors have not considered to include tails in their transformations. It could have been simply done by extending to infinity (or unlikely values) the first and last segments of their anamorphic transformations. The estimation could be less constrained by the uncertain deterministic simulation. It would be better to rerun the experiment with such an approach in order to assess the impact of the truncations on the results shown in the manuscript.

Specific comments

- p. 1891, l. 12-14 "this issue partly explains why most of the pioneer studies dealing with ocean color data assimilation were first carried out using pseudo-data rather than with real data": Twin experiments are not necessarily performed due to the poor quality of the data. They offer a perfect framework for the assessment of methodological developments: the true state and most of the error sources are known which is not the case when assimilating real observations. The sentence could be rewritten.
- p. 1891, l. 19 "making the traditional assimilation framework inappropriate to develop this applications (Doron et al., 2011)": A reference to Bertino et al. (2003) might be relevant as well.
- p. 1892, l. 6-8 "(i) to identify the best possible implementation of a multivariate ocean color assimilative system based on state-of-the-art methods": I am not really sure that this point is addressed in the manuscript. For practical reasons, the authors use a simplified version of the SEEK filter and assess the performances of two different analysis schemes. But, they do not consider ensemble methods that could constitute "better implementations of a multivariate ocean color data assimilative system".
- p. 1895, l. 4-5 "and composite data would make the diagnostics of the assimilation experiments rather complex": I am not rally sure to understand what are the issues arising from the use of merged data. Composite data sets are expected to lead to a better spatio-temporal coverage of the area, so why is it important to use one sensor only?
- p. 1894-1895 §2.2 "The ocean color data set and associated errors": The binning period of the observations in not specified. As the assimilation is performed every 8 days, I presume the authors use a time-averaged 8-days product. Is it correct?

If yes, are these observations assimilated at a date corresponding to the center of the 8-days time window? This information should be specified in the manuscript.

- p. 1895 l.25 - p. 1896, l. 3 "The sequential approach is consistent [..] rather than assuming that one particular aspect of the model [..] is responsible for the model/data misfits": I am not sure to understand the meaning of this sentence. The authors might consider to clarify this point.

- p. 1896, l. 13-14 " the upgrade of the assimilation scheme toward a fully explicit ensemble scheme will be straightforward in forthcoming applications": I presume that the upgrade of the assimilation system towards an ensemble scheme might require to work on the anamorphic transformations as well. A first issue concerns the choice of the samples used to build the empirical anamorphosis functions (from the forecast ensemble, from previous simulations as done in this study). A second issue is related to the tails of the transformations as defined in this study (see major comment [C]). Because the transformations truncate values out of range of the local samples, we might observe a depletion of the ensemble for some variables in different grid cells - either when transforming the forecast ensemble (not in agreement with the historical simulation) or transforming back the analysis (not in agreement with the forecast ensemble or the historical simulation) - and a divergence of the filter. A discussion on the strategy to define and apply the anamorphic transformation in the framework of explicit ensemble-based Kalman filtering could be included.

- p. 1896, l.18-19 "in the first version, the analysis is performed using the original state variables": I presume that negative values produced by the analysis steps are processed before the propagation steps. What is the strategy adopted? A simple increase of these values to zero?

- p. 1896, l. 29 - p. 1897, l.1 "The parametrization of the anamorphic transformation is equivalent to the one in Doron et al. (2011)": More details would be helpful

(see major comment [B]).

- p. 1897, l. 7-9 "it is possible to parameterize the error statistics [..] any extrapolation outside the range of values described by the ensemble is avoided": The issues arising from this parameterization should be discussed (see major comment [C]).

- Introduction and §2.3 "Assimilation method": According to the bibliography and the way the references are included in both sections, it seems that only people from LEGI have worked on Gaussian anamorphosis extension of Kalman filters since the original works of Bertino et al. (2003). It could have been mentioned that anamorphosis functions were also used in Simon and Bertino (2009). In the framework of the ocean biogeochemistry, the works of Simon and Bertino (2012) investigating the strategy to build the anamorphic transformations, or Ciavatta et al. (2011) using logarithmic transformations to handle the positiveness of the variables could have been cited as well.

  Ciavatta S., Torres R., Saux-Picart S. Allen J.I.: Can ocean color assimilation improve biogeochemical hindcasts in shelf seas?, *Journal of Geophysical Research*, 116, C12043, 2011.

  Simon E. and Bertino L.: Gaussian anamorphosis extension of the DEnKF for combined state parameter estimation: application to a 1D ocean ecosystem model, *Journal of Marine Systems*, 89, 1-18, 2012.

- p. 1898, l. 26-27 "The error associated with each distinct observation pixel is set to 30% of the considered data": I am wondering what is the strategy to specify the error of the transformed observations (see major comment [B]). What does "30% of the considered data" mean for the transformed observations?

- p. 1899, l. 10-12 "It is noteworthy that Hu et al. (2011) recently proposed equivalent parametrization [..] experiment": The reference is 2012.

- p. 1899, l. 25 "Figure 6 provides": it seems to be more natural to label this figure "Figure 2" . Furthermore, few words on the temporal coverage of the data set would be helpful (basically, how many data are available every year).
- p. 1900, l. 9 "over successive 60-day periods": Does it correspond to an average of 60 daily output or 7-8 analysis outputs (see major comment [A])?
- p. 1901, l. 4-5 "this is a crucial point since the free run is actually sampled to compute EOF basis used in the assimilation scheme": It could be specified that this is also crucial for the nonlinear analysis scheme since the local transformations are built from this simulation, aren't they?
- p. 1901, l. 7-13 "Considering the runs with data assimilation, the bloom starts almost in phase with the observations [..]. During the bloom, the model values are in good agreement with the observations [..] the model values are still underestimated on average by the end of the year": I agree that both data assimilation systems improve the solution compared to the free run. However, these results could be nuanced. In March-April (second row), large concentrations of chlorophyll are present in the observations along the North American coasts and the Subpolar Gyre, in the Bay of Biscay and in the North Sea while they are not present in the simulation assimilating these data. It gives the feeling that the bloom has already started in the observations and not in the model simulations. In the same way, we note large chlorophyll concentrations in the Subpolar Gyre in the observations during the period July-August (fourth row) that are not present in the model simulations.
- p. 1901, l. 18-20 "This issue is related to the fixed-based variant of the SEEK filter chosen to assimilate data [..] increasing the temporal window during which the EOF are computed": It could also suggest to propagate the error covariance matrix during the estimation process.

- p. 1902, l. 9-11 "These patterns do not appear in the anamorphosis run either, suggesting that the corrections applied to the non-observed variables are more realistic when using a non-linear assimilation scheme": Could it be related to the fact that the anamorphic transformations do not allow values outside the range of values of the local sample defined from the free run simulation? The transformation back and forth of one variable is equivalent to apply a truncation of values below and above given thresholds (the bounds of the sample). It means that the nitrate concentrations are constrained to belong to a range of values defined by the free run every 8 days. It can prevent a drift of the simulation towards large values, however it might not be the most elegant way to do it.. I am wondering why would be the impact of specifying minimum and maximum values (based on the free run) for the variables in the model on the simulation with the linear update. Would it prevent the occurrence of these localized spots in the phytoplankton concentration?
- p. 1902, l. 16-17 "these specific processes are most relevant in the context of coastal ocean color data assimilation experiments (Fontana et al., 2009, 2010; Hu et al., 2011)": The works of Ciavatta et al. (2011) assimilating ocean color data in the English Channel could be relevant as well.

  Ciavatta S., Torres R., Saux-Picart S. Allen J.I.: Can ocean color assimilation improve biogeochemical hindcasts in shelf seas?, *Journal of Geophysical Research*, 116, C12043, 2011.
- p. 1902, l. 24 "time-averaging over 16-day periods": Does it correspond to an average of 16 daily output or 2 analysis outputs (see major comment [A])?
- p. 1903, l. 17-19 " the difference between the linear and non-linear runs which is generally small, suggesting that the multivariate corrections have similar effects in both experiments" : I do not fully agree concerning region #2. The dates of the bloom peaks are in agreement for both linear and nonlinear runs (slightly later

than the free run). However, the peaks in chlorophyll concentration are lower in the solution obtained with the nonlinear run during the second half of the experiment (since 2001). These concentrations are close to the values present in the free run while the concentrations in the linear run are close to the observations. These differences seem to be present in the North Sea and the English Channel during the period Mai-August as highlighted in figure 2 (second and third rows). We note low concentrations in the free run and nonlinear run while we note large concentrations in the observations and the linear run. Again, could it be due to the fact that the anamorphic transformations constrain the observations and the solution to be inside the range of values of the local samples defined from the free run simulation (see major comment [C])? What would have been the result if "extrapolations" were added?

- §4.3 "Surface chlorophyll concentration forecast": The value of the mode of the three distributions could be specified.
- p. 1909, l. 18 "between the each experiment": between each experiment?
- p. 1912, l. 24 "sub-tropcial"
- p. 1928 Figure 9: It would be helpful to specify what are the experiments (a-d) in the legend.

Interactive comment on Ocean Sci. Discuss., 9, 1887, 2012.