First reviewer:

We would like to thank the reviewer for her/his careful reading of the article and for the useful comments which helped us to improve and clarify the manuscript. We have addressed all the comments as explained below. Also, we have proposed major changes in the article to put into light the comparison DBFN-4Dvar and to improve its readability.

1) (...) In any case, specify the values of H, L (and M), and discuss the question of the under/overdeterminacy of the estimation problem, especially for short assimilation windows.

Thank you for raising this point. In the proposed experiments we have $H = 9720$, $L = 11$ and accordingly $M = 116640$. Thinking the problem without a background term and with daily SSH observations, i.e. daily H observations, it would be necessary to have at least a Data Assimilation window (DAw) of 12 days to determine the problem. Using time interpolation, in a two days DAw we have 192 time steps, which is by far enough to determine the problem from a mathematical point of view.
However, writing the equation (1) in a discrete form we have:

$$\vec{x}^{t+1} = \vec{x}^t + \delta t \left(F(\vec{x}^t + K(\vec{x}^t_{obs} - H(\vec{x}^t))\right)$$

which shows that the nudging algorithm provides an estimate of the full system state at time $t+1$ independently of the size of the observation space, i.e $\vec{x}^t$ is indeed a background term.
We added to the article a paragraph highlighting the importance of the background for the DBFN and 4Dvar when the number of observations are smaller than the size of the state space to be identified.

And also, look at what happens in the absence of linear interpolation between observations.

The reviewer's question is good, indeed we think we have partially adressed this question in the article Sect. 5.2.2 where we discuss the effect of data sparsity on the solution. Our conclusion is that in the absence of time interpolation and considering a scalar gain, the estimates are poorer than with interpolation (this is true at least for the assimilation of SSH with a scalar gain). This is due to nature of the SSH variable in our model. Looking to the momentum equation:

$$\frac{\partial u_h}{\partial t} = -u_h \nabla_h u_h - f.\hat{k} \times u_h - \frac{1}{\rho_0} \nabla_h (p_h + p_s) + S,$$

where $p_h(x,y,z,t) = \int_{\zeta=z}^{\zeta=0} g\rho(T,S,\zeta)d\zeta$ is the hydrostatic pressure and $p_s = g\rho\eta$ is the free surface pressure, and to the equation governing the free surface (η) evolution:

$$\frac{\partial \eta}{\partial t} = -\nabla_h (H\bar{U}_h) + K(\eta_{obs} - \eta),$$
$$\bar{U}_h = \frac{1}{H} \int_{-H}^{0} u_h dz$$

we see that nudging the free surface implies a correction of the barotropic velocity. Then, since the barotropic component of the velocity field propagates much faster than the baroclinic component, the corrections rapidly readjust to their initial value, as you have pointed out in your comments. When using time interpolation we see that the baroclinic components are adjusted by the model itself, probably through interactions between the barotropic and the baroclinic modes.
When the matrix K is constructed to spread the observations to the non-observed variables, e.g 3D-velocities and temperature, Fig.13 shows that we do not need to time interpolate the observations.
In the new version of the article we took out the scalar gain experiments and the experiments using daily observation of the full SSH field because it is an extremely non realistic situation. Then, we added results about the comparison DBFN vs 4Dvar using a more realistic observation network. For these experiments, the observations are available every four days and no time interpolation is done.

2) The performance of the method is generally assessed in the paper through the relative error. There are however exceptions : errors are evaluated in Fig. 7 through rms values (with unspecified units ...), which does not make comparison with other results very easy. But the relative error is not anyway a good measure,to the extent that the value of is not specified. In the case of temperature in particular, where (I think) x true is expressed in K, the value of the relative error is very small, and it is not possible to have an obvious understanding of the significance of that error. A much better measure would be the error relative the intrinsic variability of the variable under consideration. When you describe the model, give the intrinsic variability of each variable (or a typical range of variation over say, one month), and evaluate the estimation errors with respect to that variability. Physical units might also be used for evaluating the errors, but will not be very significant for non- oceanographer readers.

We have changed the figures to consider the RMS error, which answers the first issue raised by the reviewer, and we have also specified the units for the figures.

3) Figure 5, and all similar figures that follow, show that the estimation error increases almost systematically over successive assimilation windows. That is obviously due to some form of cycling from one window to the next. The authors do not say how that cycling is done, but there is simply no point in doing it if it leads to an increase in the estimation error. It is preferable to restart the assimilation from scratch at each new window.

This is a good point raised by the reviewer. We have added one section (5.3) explaining that one assimilation cycle is defined as the process of identification of an initial condition through the iterative process followed by a forecast spanning the assimilation window, which provides a background or first guess to the next assimilation cycle.

In theory, under strong observability conditions the BFN estimation is independent of the background field (see for instance Auroux and Blum,2005). It means that the system forgets the background information and the solution is a function only of the observations, which means that there is no reason for taking the latest forecast as the new background. However in reality, only a small portion of the state space is observed in a DAw, and thus the D/BFN may take some assimilation cycles to forget the background. Indeed, the numerical experiments we have been conducting have shown that starting the iterations with a background closer to the true state speeds up the convergence, and that is why we have considered cycling of the DBFN algorithm.
However, as the reviewer has mentioned, under large DAw the system diverges after some cycles. Divergence may depend on the K matrix, DAw length and the observation network. Thus, depending on the algorithm configuration, it would be worth resetting the assimilation system after some cycles.
We made changes in the article in order to consider only a gain matrix K which updates the full state vector, thus avoiding the divergence features presented in the Sect. 5.1.
Inconclusive considerations as to the origin or the effect of the increase, like the ones relative to Figures 9 and 10, are simply irrelevant if they do not tell how to avoid the increase in the first place. Just mention that cycling, as you have attempted to do it, has the effect of increasing the estimation error.

We agree with the reviewer suggestion and took out figures 9 and 10 from the manuscript as well as most assertions about the role of diffusion in the divergence of the DBFN.

4) From what I understand, the most significant result of the paper is that BFN is numerically more efficient than 4D-Var (subsection 5.2.3). The description of the 4D-Var experiment is however much too succinct. Was there a background term xb as in Eq. (11) ? If yes, how was it defined, and how was the associated matrix covariance matrix B defined ?

Yes, the reviewer is right about the importance of the comparison DBFN vs 4Dvar. We have improved the description of the 4Dvar configuration as suggested. The 4Dvar that we have used considers a background term of the form:

$$\frac{1}{2}(\delta x_0)^T B^{-1}(\delta x_0)$$

where B is the background error covariance matrix. This matrix is supposed to model the spatial covariances of the background errors of a given variable as well as the cross-covariance between the errors of different variables. Since the state space is too big, it is impossible to store the entire B matrix. Therefore, Derber and Bouttier (1999) have proposed the decomposition of the multivariate problem into a sequence of several univariate problems. This is accomplished by decomposing the variables into a balanced component and an unbalanced component. This is done for all variables except one which should be kept without decomposition in order to define the balanced and unbalanced components of the other variables. We used the decomposition proposed by Weaver (2005)

for which the temperature is the "seed" variable and then thanks to some physical constraints such as the geostrophic balance, the hydrostatic balance and the principle of water mass conservation all other state variables may be decomposed into a balanced (B) component and an unbalanced (U) component. Thus, each model variable may be written as:

$$\begin{aligned} T &= T \\ S &= S_B + S_U = G_{ST}(T) + S_U \\ \eta &= \eta_B + \eta_U = G_{\eta\rho}(\rho) + \eta_U \\ u &= u_B + u_U = G_{u\rho}(\rho) + u_U \\ v &= v_B + v_U = G_{v\rho}(\rho) + v_U \end{aligned} \tag{1}$$

where

$$\begin{aligned} \rho &= G_{\rho T}(T) + G_{\rho S}(S) \\ p &= G_{pp}(\rho) + G_{p\eta}(\eta) \end{aligned} \tag{2}$$

Then we can define the linearized balance operator G such as:

$$B = GB_UG^T$$

Since a covariance matrix may be written as the product of variances and correlations, B may be expressed as:

$$B = G\Lambda^T C\Lambda G^T$$

where Λ is a diagonal matrix of error standard deviation, for which the climatological standard deviation are the entries, and C is an univariate correlation matrix modeled using the generalized diffusion equation discussed extensively in Weaver and Courtier (2001), Weaver et al. (2005).
We have changed the article accordingly. Also, in many aspects we have brought modifications to the paper in order to treat more deeply this point.

I mention that, if a background term is present, the comparison with BFN is not clean, since the information contained in the background is given to 4D-Var, but not to BFN. I also mention that, if a background is present, the corresponding estimation problem is automatically overdetermined, since an estimate, however inaccurate, will be available for each model state variable.

As we said before, we have made a lot of modifications on this point. Now we believe the paper adresses this question.
5) The authors write on several occasions (e.g., p. 1103,ll. 21-23) that the nudging term in eqs(3) and (4) is small in comparison with other terms. But no evidence is given to that effect.

Thank you for raising this point. This effect can be seen in the Fig.(15) of the article, for example, by the absence of shocks after each combination of data and observation. When great increments are added to the model, i.e. when $K(x^{obs} - H(x))$ is greater than the dominant terms in the equations, it generally leads to the generation of inertia-gravity waves that tends to dominate the energy spectra and disrupt the forecasts.

We agree with the reviewer that, rather than making statements about it, it would be more convenient to present the momentum balance budget including the nudging term. However, since we believe that it is not necessary here to enter in such details, we prefer just to remove these assertions from the article.

6) Figure 4 shows spectra of the model fields (in backward-forward integrations without nudging if I understand correctly). Similar spectra for the estimation error would be very useful, by showing which spatial scales are, or are not, reconstructed by BFN.

The reviewer raised a good point. We added to the article a figure (Fig.(1) in this text) similar to the Fig.(4) calculated for a two years data assimilation experiments. The figure shows that the effective resolution of the model is not affected by the diffusive character of the DBFN algorithm. It is clear that there is a reduction of the energy for the scales close to the grid scale, but the energy for scales greater than $7 \times \Delta x$ is not affected. It means that the diffusion-induced errors presented in the section 4 are "controlled" by the assimilation of sea surface height observations. Comparing with the 4Dvar, there is no great differences in the surface spectrum for the 5 days and 10 days windows, which once more proves the reliability of the DBFN for assimilation of oceanic observations. The deep ocean kinetic energy spectrum shows a relatively high energy for scales larger than $100km$. This is due to the high variance of the PLS estimator at deep layers as shown in the Fig (2) of this letter. This means that this behavior is not intrinsic to the D/BFN methodology but to our implementation scheme.
Accordingly, we have added to the article a discussion about the Fig.1 of this letter and about the gain K used in our DBFN experiments.

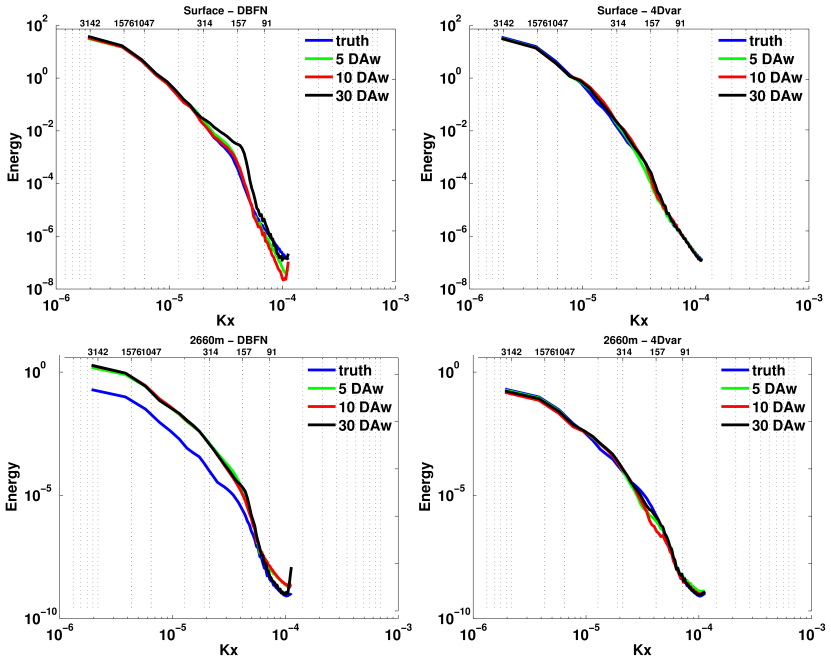

Figure 1: Kinetic energy mean power spectra calculated using the first layer (top) and a layer at 2660m (bottom) and using the 650 days of the assimilation experiments using the DBFN (left) and the 4Dvar (right). Blue curves represent the "true" power spectra; Green curves represent the power spectra calculated for the 5 days DAw; Red curves represent the power spectra calculated for the 10 days DAw and Black curves represent the power spectra calculated for the 30 days DAw. In the bottom abscissa the tick-labels stand for longitudinal wave-number (rad/m) while in the top abscissa the tick-labels stand for the corresponding wavelengths in km units.

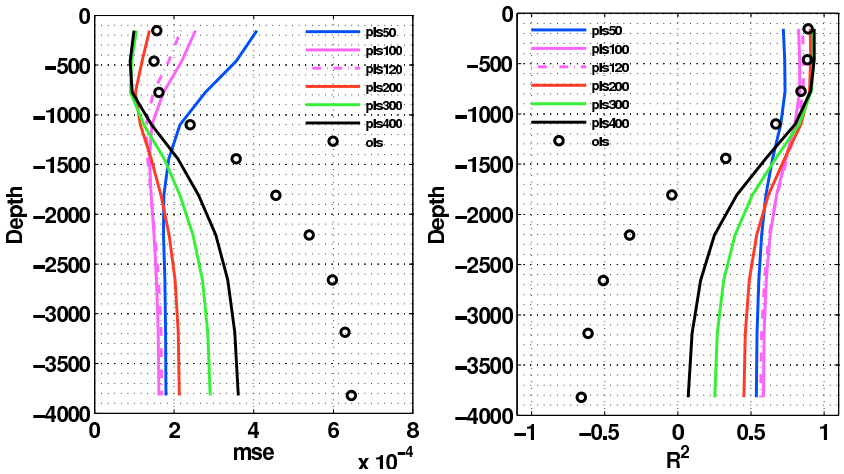

Figure 2: Mean Squared error of the residuals (left panel) and R^2 score (right panel) for the PLS algorithm using different number of modes, indicated in the legend, and for Ordinary Least Square regression. Results of prediction, i.e. the statistics are calculated using objects not used in the construction of the regression model.

Remarks

As stated before, we have made major changes in the article. Here we describe how the structure of the article changes:

- Subsection 3.1 "Ocean model configuration" is abbreviated and included into Section 3;
- Subsection 3.2 "Data Assimilation experiments" is transformed into Section 5;
- Section 5 contains four subsections:
  1. "Prescription of the DBFN gain": explains how the matrix K is calculated;
  2. "The 4Dvar background term configuration": details the background term used in the 4Dvar;
  3. "Assimilation cycle": explains how and why the methods are cycled;
  4. "Observation network": describes the assimilated observations and discusses the undetermination of the assimilation problem;
- The Section 5 is shortened and becomes Section 6. Subsection 5.1 "Experiments with scalar nudging coefficients", 5.2.1 "Daily gridded SSH observations" and 5.2.2 "Temporal data sparsity" are removed from the article;
- Subsection 5.2.3 "Intercomparisons" is transformed into subsection 6.1 "Reference experiment";
- Subsection 6.2 "Sensitivity experiments" is created to describe sensitivity tests with respect to the length of the Data Assimilation window. The DBFN and 4Dvar are compared.

Second reviewer:
We would like to thank the reviewer for her/his careful reading of the article and for the useful comments which helped us to improve and clarify the manuscript. We have addressed all the comments as explained below. Also, we have proposed major changes in the article to put into light the comparison DBFN-4Dvar and to improve its readability.

a) the main point regards Eq.(6) and (7). The authors only cite a personal communication that should explain why after an infinite number of iterations their algorithm should converge to a trajectory calculated without the diffusive and the nudging term. We think that this point is important in driving the reader in the comprehension of the results presented. So the authors should give some theoretical justifications and verify it in their results.

Thanks for raising this question. The personal communication concerns the statements that under convergence conditions and under the hypothesis that at convergence both forward and backward trajectories are equal, then Eqs. (6) and (7) hold. To see this we write the DBFN system as:

$$\begin{aligned} \frac{\partial \vec{x}_k}{\partial t} &= \mathcal{F}(\vec{x}_k) + \nu \Delta \vec{x}_k + \vec{K}(\vec{x}_{obs} - \mathcal{H}(\vec{x}_k)) \\ \vec{x}_k(0) &= \tilde{\vec{x}}_{k-1}(0), \qquad\qquad 0 < t < T, \end{aligned} \tag{3}$$

$$\begin{aligned} \frac{\partial \tilde{\vec{x}}_k}{\partial t} &= \mathcal{F}(\tilde{\vec{x}}_k) - \nu \Delta \tilde{\vec{x}}_k - \vec{K}'(\vec{x}_{obs} - \mathcal{H}(\tilde{\vec{x}}_k)) \\ \tilde{\vec{x}}_k(T) &= \vec{x}_k(T), \qquad\qquad T > t > 0. \end{aligned} \tag{4}$$

where $k \in N_{\geq 1}$ stands for iterations.
We see that if $\vec{K}' = \vec{K}$ and the forward and backward limit trajectory are equal, i.e $\tilde{\vec{x}}_\infty = \vec{x}_\infty$, then taking the sum between Eqs.(3) and (4) shows that $\vec{x}_\infty$ satisfies the model equations without the Nudging and diffusion:

$$\frac{\partial \vec{x}_\infty}{\partial t} = \mathcal{F}(\vec{x}_\infty) \tag{5}$$

while taking the difference between Eqs.(3) and (4) shows that $\vec{x}_\infty$ satisfies the Poisson equation:

$$\Delta \vec{x}_\infty = -\frac{\vec{K}}{\nu}(\vec{x}_{obs} - \mathcal{H}(\vec{x}_\infty)) \tag{6}$$

Concerning the BFN and DBFN convergence, we prefer to make references to past works in the introduction to avoid increasing the length of the article: Auroux and Blum (2005) for a ODE linear system ; Ramdani et al. (2010) for reversible linear PDE equations (Wave and Schrödinger equations); Auroux and Nodet (2012) for linear and non-linear transport equation under viscous and non viscous conditions.
Then, we added a paragraph to the article explaining how we obtain Eqs. (6) and (7) and the issues related to convergence.

b) it is not clear the behavior of the diffusive term in the backward

integration. We understood that this term eliminates the small scale structures both in forward and in backward integration. The sign indicated in (4) suggests this interpretation but some sentences at pag. 1080, line 15 and following let the reader quite confused.

We agree with the reviewer, the diffusion term as it is written in Eqs.(3) and (4) eliminates the small scale structure both in forward and backward mode. However, the point discussed in pag. 1080 is to clarify that ideally the true inverse model should not dissipate energy both in forward and backward integration. If analytically this makes the backward integration ill-posed, numerically and for finite Data Assimilation window it is the very small scales (high wavenumber) that pose the problem. That is why we suggest the use of the BFN (not the DBFN) followed by a digital filter which eliminates the necessary energy to keep the numerical solution stable.
Since this comment may be a source of confusion, we decided to take it out of the article and just say that for sake of stability we used the DBFN.

c) It is completely unclear what are the different kinds of K's used. At Pag. 1080 it seems (we use latex notation) that K =k H^T R^-1, then the authors speak about a "K based on the PLS regression model", somewhere else (e.g. Pag. 1083) it seems that after the DBFN the PLS regression is used. We strongly suggest the authors to make the technical details of the different experiments of their method clear.

We thank the reviewer for this useful remark. We have used two versions of K. One is a scalar, and in this case we can interpret $K = kH^T R^{-1}$ with $H^T = Id$ and R the observation error covariance, which in our case is diagonal with equal entries. The other one relies in the covariance (correlations) calculated thanks to the PLS regression. In this case, the updating scheme can be seen as a rough approximation of the two steps update for EnKF. As we have already said, we made several changes in the article, thus in the new version only the Kalman-like gain is used. Accordingly, we have added the subsection 5.1 "Prescription of the DBFN gain" to clarify our choice of the gain matrix K.

d) We agree with the other referee that the relative error is not a good measure of the difference between two states. We suggest the use of the RMS or of the RMS normalized by the standard deviation.

We have changed the figures to consider the RMS error.

e) This point regards the DFBN technique: the authors state the in absence of observations the iterations converge to an homogeneous state. This means that after several iterations the analysis is completely independent of the dynamics equation (F(x) in Eq.(2)).

We think that with our explaination given to the remark a) the reviewer will certainly better understand this point. Indeed, we do not say that without observation the solution is totally independent of the model, since it is stated that the trajectory at convergence is a solution of the model F. Indeed, in the complete absence of observations it is not worth considering the iterations.

We included in the article a better explaination about this point, accordingly with our answer to the topic a).
We think that with no diffusive term, after several iterations, the model is in some sense "forced" to become equal to the observations in the observed points. Reading the manuscript we have understood that the authors think that with a balance of the diffusive and nudging terms, the trajectory should converge to an actual trajectory- of the model without diffusion. If this is correct the authors should better clarify and also prove that this behavior holds, at least in the model under examination.

What we mean by the paragraphs of lines 3-15 on page 1082 is that the trajectory at the convergence satisfies both the model equations without diffusion terms and the Poisson equation. It does not necessarily mean that it is the solution of the Back and Forth Nudging when considering the model without diffusion. Moreover, we agree with the reviewer when he says that "with no diffusive term, after several iterations, the model is in some sense "forced" to become equal to the observations in the observed points". To see this, we just need to write Eqs.(3) and (4) without the diffusion term and take the difference between them:

$$K(x^{obs} - H(x_\infty)) = 0$$

To show that at convergence the solution satisfies the model equation without the nudging and diffusion terms, we have configured an experiment for which the true state comes from a higher resolution model ($3km$). The projection of the higher resolution model onto our mesh is viewed as the model trajectory without diffusion. This trajectory is assimilated using the DBFN algorithm and then we compare the kinetic energy spectrum for the high resolution model, a typical spectrum for our configuration and the spectrum after the assimilation of the high resolution observation. Figure 3 presents these spectrums. We readily see that the reconstructed spectrum is much closer to the high resolution model than to the typical spectrum for our configuration.

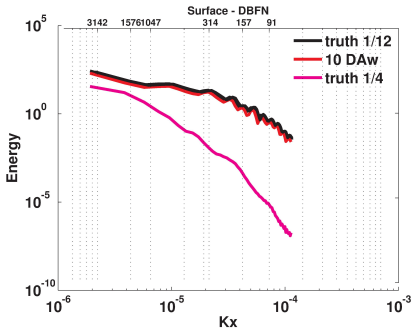

Figure 3: Kinetic energy mean power spectra calculated using the first layer (top) using the forecast of the assimilation experiments using the DBFN and assimilating high resolution observations. Black curve represents the "true" power spectra at high resolution; Red curve represents the power spectra calculated for the 10 days DAw and Magenta curve represents a typical spectrum for our configuration. In the bottom abscissa the tick-labels stand for longitudinal wave-number (rad/m) while in the top abscissa the tick-labels stand for the corresponding wavelengths in km units.

Remarks

As stated before, we have made major changes in the article. Here we describe how the structure of the article changes:

- Subsection 3.1 "Ocean model configuration" is abbreviated and included into Section 3;
- Subsection 3.2 "Data Assimilation experiments" is transformed into Section 5;
- Section 5 contains four subsections:
  1. "Prescription of the DBFN gain": explains how the matrix K is calculated;
  2. "The 4Dvar background term configuration": details the background term used in the 4Dvar;
  3. "Assimilation cycle": explains how and why the methods are cycled;
  4. "Observation network": describes the assimilated observations and discusses the undetermination of the assimilation problem;
- The Section 5 is shortened and becomes Section 6. Subsection 5.1 "Experiments with scalar nudging coefficients", 5.2.1 "Daily gridded SSH observations" and 5.2.2 "Temporal data sparsity" are removed from the article;
- Subsection 5.2.3 "Intercomparisons" is transformed into subsection 6.1 "Reference experiment";
- Subsection 6.2 "Sensitivity experiments" is created to describe sensitivity tests with respect to the length of the Data Assimilation window. The DBFN and 4Dvar are compared.