Evaluation. The paper is not acceptable in its present form. It may become acceptable after substantial improvements, describing more precisely what the authors have done, and allowing better understanding of the results they have obtained and of the significance of these results.

The paper describes the implementation of the assimilation technique of Back and Forth Nudging (BFN) on the Nucleus for European Modeling of the Ocean (NEMO) primitive equation model. The results of a number of 'identical twin' experiments are presented, and a comparison is made, both in terms of intrinsic quality of the results and of computational cost, with results obtained through the well-established technique of 4D Variational Assimilation (4D-Var). I think the results obtained by the authors are fundamentally worth publishing, but they are very difficult to understand and evaluate. Additional explanations and clarifications are necessary. Here are my main comments and suggestions for improvement. They are limited at this stage to what I consider as the most important aspects of the paper. I defer other comments to a possible future version of the paper.

1. The purpose of assimilation of observations is to estimate the state of the observed system, using all available relevant information. In the present case (as is usually the case in assimilation of atmospheric or oceanic observations), the relevant information consists of the observations proper, and of a numerical model (here NEMO) describing the dynamics of the flow. The purpose of the authors is to estimate through BFN the state of the flow at the initial time of the assimilation window, expressed in the format of the model, and denoted $x(0)$.

A basic question is the degree of under/overdeterminacy of the corresponding estimation problem, *i.e.*, how many scalar parameters are to be determined from how many scalar pieces of information ? If there are more parameters to be determined than available pieces of information, the estimation problem is underdetermined, and assimilation will, at best, reconstruct only part of the flow.

From what I understand (top of p. 1087), and in agreement with the general formulation of Primitive Equation models, the initial condition $x(0)$ is specified by the values, over the three-dimensional grid of the model, of the temperature and the two velocity components, and by the values of the sea-surface height SSH. Denoting by H and L the number of gridpoints in the horizontal and the number of levels in the vertical respectively, this makes $M = (3L + 1)*H$ parameters to be determined.

Observations are defined as daily SSH observations (p. 1089, l. -4), *i.e.* H values per day. In the case of a 2-day window, that makes 3H values, much less than the required M. The authors nevertheless present their results as being successful. They write (p. 1090, ll. 23-24) that a linear interpolation is made between observations to produce values at every timestep. The value of the model timestep does not seem to be given in the paper, but is the interpolation sufficient to overdetermine the problem ? Even if it mathematically does, the additional 'observations' thus obtained will be largely redundant, resulting in poor numerical conditioning and very slow convergence. Have you done experiments without interpolation ?

It might be that as many as M parameters are not actually necessary. There must exist an approximate geostrophic balance in the flow, thus reducing the number of effective parameters to $(L + 1)*H$. But then the condition for balance must be introduced somewhere in the assimilation process (otherwise experience shows that the assimilation tends to use the non-geostrophic degrees of freedom to fit the noise in the observations). It may also be that some components of the flow are in effect 'slaved' to other components, to which they readjust so rapidly that their initial values are in practice irrelevant. But, if that is so, say it clearly.

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

2. The performance of the method is generally assessed in the paper through the relative error $||x\text{-}x^{\text{true}}||/||x^{\text{true}}||$ (p. 1091, l.1). 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 $||x^{\text{true}}||$ 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.

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. 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.

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 x^b as in Eq. (11) ? If yes, how was it defined, and how was the associated matrix covariance matrix **B** defined ? 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.

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.

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.