Reply to reviewer #1

We wish to thank the reviewer for her/his constructive comments. We reply to each comment below (original comments *in italics* and our response in regular font).

1 General comments

This paper presents a bred vector method for ocean model initialization relevant for seasonal-to-interannual coupled model predictions. The main novelty of the method compared to previous work is to extend the breeding to the entire water column of the ocean model, and focus on the transition time scale around one year lead time. The authors start by describing the model framework and discuss the choices made in their implementation of bred vectors in the MPIOM oceanic component. They then present an analysis in a perfect model framework to compare bred vector perturbations to a more traditional lagged initialization method. This manuscript is well-written, clearly organized and concise. I found the bred vector method description particularly clear. The authors made the effort of showing results for different lead times and different ocean variables. The Talagrand diagram figures over different ocean regions are a nice and synthetic way to present results.

The authors are well aware of the main limitations to their current study and discuss these in the paper. In the bred vector implementation section, the validation of the choices for the implementation is rather short, in the sense that some sensitivity tests were run, results are discussed but not necessarily shown. This could be done for instance in supplementary material for readers interested in specifics of the method. Another thing I would have expected to see in the article was an evaluation of the significance of the results presented. The scientific conclusions drawn from the study are valid in the framework described in the paper, and this study will greatly benefit from being adapted to initialized interannual integrations to see if conclusions hold. Followup papers on this work will be very interesting for the seasonal-to-decadal climate modelling community.

More specific comments/questions for the authors follow.

All general comments are brought up within the list of specific comments, and we reply within the list.

2 Specific comments

2.1 Introduction

p5191, l. 10: "Different techniques have been tested to initialize and perturb the ocean's surface": Instead of citing a bulk of papers, could you maybe list a few of these techniques and the relevant papers for each? Is there a review paper you could cite on these initialization techniques?

We have expanded the sentence into a short paragraph, where techniques and

respective papers are mentioned. We are not aware of a review paper for (ocean) initialization and/or ensemble generation.

p5192 l. 19: "At the expense of being limited to perfect model measures to quantify the ensemble spread": I am actually more concerned about the quantification of the model error than the ensemble spread when dealing with perfect model simulations. What exactly do you mean by this?

This was indeed ambiguously worded. We meant to express that the evaluation of spread is limited to an evaluation against the model's own spread instead of observations. We have now re-worded this: "At the expense of being limited to using the model itself as a reference when quantifying the ensemble spread,…"

2.2 Bred vector implementation
I really liked the description section which describes thoroughly the bred vectors method.

p5195, l. 3: "The breeding cycle itself consists in the last four steps" : For the sake of clarity it could be a good idea to use numbers instead of hyphens, and then say "steps n to n+3"...

Changed as suggested.

In the implementation section (3.2), the choices for several settings of the method are discussed. However, even if an assessment of the method as described here is made in the beginning of the results section 3.3 to justify that these choices are valid, it would be most interesting for readers who intend to reproduce these results in other ocean models to have more details on the tests that were made, by means of supplementary material for instance.

We carefully re-read the section, and attempted to provide additional information. However, as most tests yield very similar results (also to what is presented in figure 1), we could not identify many additional information that would warrant supplementary material. We did include some additional information in a few places in section 3.2.

p5196, l. 26: Did you try other normalizations for the bred vectors? Did you test other values for the normalization instead of 10 percent before coming to the conclusion that this was the most appropriate choice?

The main sensitivity that we have found was on the length of the breeding cycle and on the extent of the vertical profile for the norm. Other parameters such as the choice of the breeding variables (temperature, salinity, and/or velocity), different geographical regions for the computations of the (Pacific or Atlantic), and also stronger normalization factors all resulted in comparatively small differences.

In the results section, figure 1 is difficult to interpret given that there is no color bar. Do the colors correspond to the same values in all four figures shown?

Yes, normalized values are shown. This information is now added to the figure caption.

p5198, paragraph lines 16-22: "Neither choice fundamentally changes the regression maps (Fig. 1) [: : :] are analyzed." From the formulation of the sentence, I would expect to see these results in the figure somewhere. I would remove the "(Fig. 1)" if you don't intend to show this, or change the sentence.

As the results are indeed very similar to what is already shown in the manuscript, we do not add a figure. As suggested, we have removed the "Fig. 1" reference, and also added "(not shown)" at the end of the paragraph.

2.3 Ensemble generation
p 5199, l.13: "All quantities presented in this section are averaged over the ten ensemble members": this statement is confusing, since you are mostly looking at ensemble spread (as in the Talagrand diagrams).

Indeed, this statement was inappropriate and we removed it.

l.22: The description of how the Talagrand diagram is calculated could be more clear; what do you mean by "the ensemble is sorted according to a predicted value"? Furthermore, in the figures, you present Talagrand diagrams for rather large regions/boxes. It could be useful to mention in the text how you pool the data together in your analysis (described briefly in the caption for figure 4).

We have reworded and extended the paragraph:
For a Talagrand diagram, the ensemble is sorted, and the range of the sorted ensemble verified against a control value, verifying whether the control is within or outside the ensemble distribution. A flat Talagrand diagram indicates that the ensemble has the same probability distribution as the control. A U-shaped Talagrand diagram indicates too little spread in the ensemble, and a Gaussian shaped rank histogram indicates too much spread. We show Talagrand diagrams for different regions, as distributions vary considerably between different regions. For a Talagrand diagram over a specific region, all grid points and all ensemble members are individually included in the computation of the histogram without any prior spatial averaging. Note that Talagrand diagrams for very small regions might be too noisy to be interpreted for their ensemble spread, while a global average might not representative for the spread in a certain region.

l.27: "The spread-error ratio is then the ratio of this spread and the difference": do you mean root mean square difference?

Yes, corrected.

p5200, l.1: "A perfect ensemble would result in a spread-error ratio of 1." Could you be a little more precise about what you mean by "perfect" in this sentence?

Replaced by: 'An ensemble with the same spread as the reference simulation would result in a spread-error ratio of 1.'

In the results section, some figures are discussed quite briefly, leaving the reader to make his own interpretations as to whether the bred vectors initial perturbation method is indeed better than the lagged initialization. As mentioned earlier, and although this is sometimes a complicated task, results would highly benefit from significance assessments.

Please see below for our modifications as suggested for the individual figures (especially figures 5 and 6).

I find figure 4 and the other Talagrand diagram box figures very interesting, both in terms of content and in terms of ways of presenting synthetic experiment results. However, more information should be given on the shading used to indicate which ensemble is closer to a flat distribution. How was this calculated? Which threshold was chosen to discriminate between categories?

We have included the following two sentences in the paragraph that describes the Talagrand diagrams:
To compare two Talagrand diagrams we compute the deviation of the respective histogram from a flat distribution. We assume variations between two ensembles to be significant if the difference between the two deviations from a flat distribution is larger than 10 percent.

The figures showing the spread-error ratios (fig. 5, 6c and d) would benefit from a clearer separation in the colors used. The changes between both methods are some- times quite small and they are difficult to see when values remain in the blue-green color range.

We changed the color scale as suggested.

More specific comments on this section follow:
• p5200, l.8: the sentence must have been cut, and makes no sense as it is now.

Corrected.

• p5200, l.15: Is the 0.05 spread-error difference significant?

The spread-error figures now include not only a clearer color scale, but as well a white area around 1 (±0.05). Of course, this does not imply that differences

between the spread-error ratios of 0.05 are significant. However, a formal significance estimate of the limited ensemble against the model itself would suggest higher robustness that the present results can deliver.

• *p5201, l.2: The description of the differences between figures 5a and 5b lacks more detail, more quantitative information. Furthermore, although Talagrand diagram results shown previously give some idea that the improvements noted in 5b are mainly due to increasing the ensemble spread, it could be worth discussing the impact of error alongside the spread-error analysis shown here.*

We have included in the description of figure 5 a brief individual description of spread and skill:

While figure 4 indicates an increased spread in the bred initialized ensemble in many regions, the separate inspection of the error indicates smaller errors in the bred initialized ensemble in the tropics and in dynamically active regions like the western boundary currents.

[…] At one year lead time, the improvements in the spread-error ratio in the bred initialized ensemble stem almost exclusively from the improvement in the spread, while the errors are of comparable structure and magnitude both ensembles.

• *I am curious about what is happening in the bred vector initialization experiments around Japan. The spread-error ratio in figures 5b and 5d over this area is much higher than in the lagged initialization case, although judging from the Talagrand diagrams in figure 4 there is no clear difference for this region when compared to other boxes. Did you have a further look at this? How does the error over the region grow? Can this be related directly to the bred vector perturbations?*

The region with the larger spread-error ratio is smaller than the region that is shown in figure 4, and it is hence hard to clearly relate these two figures for such a region. Specifically for the Kuroshio rehion, we find at both 2-4 months and one year lead time a slightly smaller spread in the bred initialized ensemble than in the lagged initialized ensemble. The error, however, is considerably smaller in the bred initialized ensemble than in the lagged initialized ensemble. In turn the spread-error ratio is larger for the bred initialized ensemble than the lagged initialized ensemble.

2.4 Discussion
I very much appreciated this section as it underlines the fact that the present study is preliminary and details how it can be extended and applied to a more operational framework. In the case of initializing coupled ocean-atmosphere bred vectors, would you expect the optimal breeding cycle time to change much with respect to an ocean only breeding method?

As atmosphere-only breeding cycles are typically in the order of hours, ocean and atmosphere breeding cycles would likely have to be different.

3 Technical comments

You make an extensive use of "e.g.", which sometimes is not adapted to the construction of the sentence (see for example p 5193, l. 19).

We have now restricted the use of "e.g." to citations.

p 5193, l. 21: "uninitialized freely model": a word must be missing here...

Corrected.

Figure 1: the color bar is missing; there seems to be a white dot on figure 1c, is this due to highly negative values?

Yes. As for the colorbar: All values are normalized, this is now mentioned in the caption.

Figure 2: Since the values are quite similar in both experiments, I would suggest using the same color bar and scale so that the unperturbed and bred experiments can be compared more easily.

Changed as suggested.

p 5199, l. 21: (ii) is missing in the enumeration of the measures.

Corrected.

p 5203, l. 18: used IN/BY Yang et al. (missing word)

Corrected.

In the references section, some capital letters are missing (for instance, "north atlantic" p5206 l.10) in several article references.

We checked all references and capitalized letters where needed.