We thank the editor Dr. Farneti in handling our manuscript, and Dr. Griffies, Dr. Bell, Dr. Hogg and Dr. Hirschi for their positive and constructive comments. We have acknowledged their work in the Acknowledgements section. Please find our point-by-point reply below in red text.

# Referee #4 (Joel Hirschi)

I think this manuscript is a most interesting and timely illustration of how storage and analysis of large model datasets may evolve in the coming years. The latest generation of ocean (and also of atmosphere) models now routinely produce datasets of O(Tera-Petabytes). The storage and analysis of these datasets is a major challenge – which often results in cutting edge simulations being underexploited. Here the authors use the cloud-based framework proposed by the Pangeo project to produce an intercomparison of a set of 8 submesoscale-permitting ocean models. The manuscript shows the potential of the cloud framework and provides an assessment of the mixed-layer instability (MLI) parameterisation of Fox-Kemper (2011) across a set of submesoscale-permitting models. The manuscript is clear and well-written and will be an excellent contribution to GMD. I only have a few minor points listed below that might benefit from some clarification.
We thank the reviewer for his supportive comments.

## Comments:

1. The main motivation of the study is to demonstrate a framework for the intercomparison and analysis of datasets of O(Tera- Petabytes). However, the region of focus around the Gulf Stream separation is actually quite small (~1000 km x 1000 km) and the size of the datasets will be Gigabytes rather than Terabytes or more. The choice to focus on the Gulf Stream separation region is well motivated as this is a region where SWOT tracks will cross. Nevertheless, I wonder if something can be said about how easily the system would scale if comparison and analysis were extended to e.g. the largest domain (North Atlantic) that all 8 models have in common or to global analyses (ie. when the amount of data indeed gets in the order of Petabytes…). Could OSN handle this amount of data? Could it be uploaded onto the cloud within a reasonable amount of time?
In terms of technology, OSN and the Pangeo pipeline are capable of handling petabytes of data. We have uploaded the 4 regions shown in Figure 1 to OSN and the process scaled well. We have added this in lines 88-89 as: "The entire process of zarrifying the data, fluxing them to OSN and cataloging scaled well for the four regions shown in Figure 1." The primary limitation comes from the acquisition of funding to support and maintain such storage on the cloud.

2. I found the MLI assessment most interesting as it adds an interesting piece of science and the agreement seen in Figures 6 and D1 is surprisingly good. However, I am not sure that the explanation given in Appendix D as to why the histogram values shown in Figure D1 are falling under the one-to-one line is correct. Isn't this rather the consequence of taking the spatial median? If the local (i.e. for each grid cell) values are taken for Ce, there is by construction a perfect alignment of the histograms with the one-to-one lines. Any departure from the one to one lines has

therefore to result from summarising the spatial variability with one value (i.e. Ce(t, x,y) --> Ce(t)). I also note here that across the models the slope for the histograms is steeper than the one-to-one lines. As before, I feel that the slope will be affected depending on which value you chose for Ce (e.g. median, mean, mode, 1st, 3rd quartile…etc). Depending on which value you pick and on the distribution of the values Ce(t,x,y), I expect that the histogram values can move above, onto, or below the one-to-one line and that the slope can increase or decrease. What do the distributions of values Ce(t,x,y) actually look like? It might be nice to see an example. This distribution may be a useful guide for deciding on the value of the efficiency coefficient Ce.

If we permit Ce to take spatially negative values, then the reviewer is correct in that “there is by construction a perfect alignment of the histograms with the one-to-one lines” as it is the ratio between the submesoscale buoyancy flux ($w'b'$) and its value predicted from the parametrization. Namely, Ce(t,x,y) is indeed locally negative where frontolysis dominates as $w'b' < 0$ whereas the parametrization by construct can only take positive values (Fig. E1). While Ce is a tuning parameter, energetic consistency of the parametrization requires it to take only positive values (Fox-Kemper et al., 2011). Although the spatial median was chosen so that Ce(t) is less sensitive to spatial extrema, and in order to have the time series of $\overline{w^s b^s}^z$ and the parametrization to agree in the spatially averaged sense, we agree that the histogram would depend on the values chosen for Ce. We have added the histogram for ‘Ce’s diagnosed as the spatial mean and mode in Appendix E. While the joint histogram seems to align closer to the one-to-one line when the spatial mode is used than the median (Fig. E1), the parametrization tends to overestimate the submesoscale buoyancy flux (Fig. E2). We, therefore, recommend the usage of spatial medians in estimating Ce.

## Details:

- Figure 5: I suggest to label the panels with w, wm, ws and b, bm, bs.
  Done.

- Figures 6, D1: Use "Ce"  rather than "C".
  Done.

# Reference


- Fox-Kemper, B., Danabasoglu, G., Ferrari, R., Griffies, S.M., Hallberg, R.W., Holland, M.M., Maltrud, M.E., Peacock, S. & Samuels, B.L. (2011). Parameterization of mixed layer eddies. III: Implementation and impact in global ocean climate simulations. *Ocean Modelling*. doi:10.1016/j.ocemod.2010.09.002;