Review of "The Mixed Layer Depth in the Ocean Model Intercomparison Project (OMIP): Impact of Resolving Mesoscale Eddies" by Treguier et al.

Reply to reviewer #1 (Steve Griffies)

This is a well-written summary of the comparison of mixed layers in a suite of OMIP2 simulations, including both coarse (1degree) and fine grids. The results are provocative and provide a benchmark and motivation for future work. I support publication and only have a few comments.

Many thanks for your review of our manuscript and your encouraging comments.

-line 26: We never really "validate" climate models. Instead we "evaluate" models.
We have replaced "validation" by "evaluation".

-line 74: "period period"
Typo corrected.

-line 188: "of course" is subjective; suggest removing
We agree, "of course" has been removed.

-In many many places in the manuscript, the word "resolution" is used when you really mean "grid spacing". Resolution is a non-dimensional number whereas grid spacing measures the distance between grid points, typically in degrees or km. In most places this quibble is not so important since we "know what is meant" even if we do not say it clearly. But on line 425 one reads the rather confusing sort of statement that can result when "resolution" and "grid spacing" are interchanged: "with a horizontal resolution of less than 1km". Does that phrase mean the grid spacing is coarser than 1km or finer than 1km?? This is the sort of confusion that a novice (and experienced) reader can come across when "resolution" is used when "grid spacing" is meant. I suggest clarifying all uses of "resolution".

Thanks for this suggestion, we have modified the manuscript accordingly (line numbers refer to the first version of the manuscript):
- abstract, line 24, "1° grid spacing".
- line 95, replace "1°" by "1° grid spacing".
- line 99, replace "up to 1/16°" by "up to 1/16° grid spacing".
- line 142, replace "horizontal resolution" by "horizontal grid spacing"
- line 147, replace "0.1° resolution" by "0.1° configuration"
- line 148, replace "1° resolution" by "1° configuration"
- line 149 and line 150, replace "Both resolutions" by "Both configurations"
- line 151, replace "at 0.1° resolution" by "at 0.1° grid spacing"
- line 152, replace "at 1° resolution" by "at 1° grid spacing"
- line 158, replace "nominal resolution" by "nominal grid spacing"
- line 166, replace "1/16° horizontal resolution" by "1/16° horizontal grid spacing"
- line 180, replace "horizontal resolution" by "horizontal grid spacing".
- line 377, replace "a low resolution (1°) reanalysis" by "a reanalysis with coarse (1°) grid spacing"
- line 391, remove "high resolution" before "1/16° CMCC model"
- line 394, replace "to reach a 1° resolution" by "to a nominal 1° grid"
- line 425, replace "horizontal resolution of less than 1km" to "grid spacing of less than 1 km".

-line 575: The authors observe that the MLD is more widely spread among the OMIP simulations than SST. This is a very important statement that perhaps has been noted before but is worth emphasizing more. It offers an important counter-point to those who discount OMIP simulations for using a prescribed atmosphere and so "have the answer given to them". There is a lot more physics going into the MLD than just that provided by the SST and SSS.

Thanks for this interesting comment. We have added two sentences to emphasize our statement more, line 578 in section 6:
" It is important to note that the mixed layer heat content, an important variable for climate, is not well constrained by the SST alone and that the mixed layer depth depends on internal ocean dynamics. As noted by Griffies et al., (2009) and Danabasoglu et al. (2014), OMIP experiments reveal the key influence of these ocean dynamics (and their representation in models) on major climate-relevant processes such as mixed layer properties, water mass subduction and meridional overturning."

-Now for my slightly nontrivial and somewhat self-serving comment. Namely, the authors point to the sensitivity of the MLD to the density threshold (Figure 2), the upper layer depth, and time sampling. In the end, they propose a 10m upper layer starting point rather than the "top grid cell" advised by Griffies et al (2016). I think this is a good suggestion. Yet they also support the density threshold approach, even though it suffers from the problems they note in their Figure 2, as well as those problems noted by BM04 whereby the density threshold should be a function of the SST/SSS given the nonlinear equation of state. These limitations and hyper-sensititivities motivated Reichl et al (2022) to propose a potential energy-based threshold. That approach also can use monthly mean T/S to directly compare model to obs, and it can be implemented online. So it is a practical approach and simpler than some methods like Holte and Talley, though more complex than the density threshold method.

I do not ask the authors to add the energy approach to their analysis, as that would be more work than I can reasonably request. But I do suggest they be somewhat more circumspect about the density threshold for future MIPs. I might be wrong, but at this point I think a potential energy approach ala Reichl et al is a physically compelling and practical approach that avoids many of the problems with the density approach.

We agree that we could mention in more detail the new approach presented by Reichl et al. As you point out, more work is necessary to fully implement the potential energy method at the global scale, and this is beyond the scope of our paper. We have added the following sentence line 219:
"**The potential energy-based method proposed by Reichl et al. (2022) may have the advantage of being less sensitive to the model's vertical resolution than other methods, but more research is needed to understand how to choose a potential energy threshold, and whether it is possible to define one that would be relevant globally and for all seasons**".

Reply to reviewer #2 (George Nurser)

Major comments

This is an important and timely piece of work on a key topic that is only going to become more important for climate science. The work presented here is rigorous and careful, and the authors have taken care to point out the inconsistencies in the model datasets employed here —there are not many eddy resolving runs available forced by the same atmospheric forcing datasets, and different vertical resolutions and parameterisations are used by different models. Hopefully this MS will stimulate further work on this subject.

The suggested definition of the model MLD as the shallowest depth where as $\rho - \rho 10m \geq 0.03$ kg m-3 (at least when calculating MLD online) should be followed in future work.

It is an important finding that finding that (generally) the MLD thus defined is indeed shallower at eddy-resolving resolution, presumably because the eddy parameterisations at coarse resolutions are not as effective at fluxing buoyancy upwards as are the resolved eddies at coarse resolution.

It is also interesting to see how sensitive the MLD are to the actual ML model and to model vertical resolution, especially in summer.

We thank you very much for these positive comments on our manuscript.

I was not fully convinced, however, of the argument that the MLD calculated from the monthly-average temperature and salinity, MLD($\langle T \rangle$, $\langle S \rangle$), was a totally adequate reflection of the monthly-mean of the MLD calculated online at each time step $\langle$MLD(T, S)$\rangle$, and thus a fully comparable field to the MLD calculated from individual observational profiles. For the zonally averaged $\langle$MLD(T,S)$\rangle$ − MLD($\langle T \rangle$,$\langle S \rangle$) in Fig. 6 seem to reach ~30–40 m in late spring. Moreover of the two models plotted in Fig. 6, IAP-LICOM is very much an outlier, and FSU-HYCOM has notably coarse vertical resolution (~ 36 layers) and so would not expect to show much of an effect, especially in summer.

We did not want to claim that "the mixed layer depth computed from the monthly averaged T and S is a totally adequate reflection of the monthly-mean of the MLD calculated from observational profiles". Rather, we have found that it was the best possible choice for our model intercomparison, because the MLDs calculated online in the different models are impossible to compare with each other, due to the different methods (as shown in table 1). We have added a sentence to better explain our strategy, line 387:
"We acknowledge that computing MLDs from monthly averaged T and S is not satisfying, and that the only motivation to do so is to allow the intercomparison of models for which the MLDs computed online are not comparable with each other".

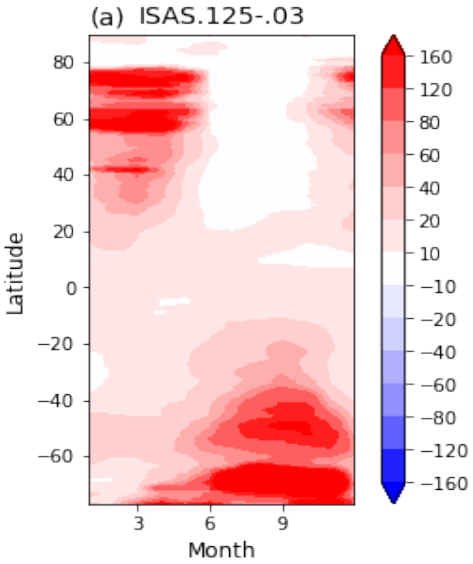
Fig. 6 would be more convincing if it included output from a higher resolution, "good" model, like CMCC-NEMO which has 98 levels. Why not use NCAR-POP and CMCC-NEMO as were used in Fig. 3 to show the impact of using different reference levels?

Your suggestion to use other models for Figure 6 is of course relevant, but we used the two models for which the datasets were accessible. Daily 3D fields of temperature and salinity have been published on ESGF for the IAP-LICOM model, although the retrieval of these

fields is a challenge (which is the reason why only one year, 1998, was used). Daily 3D fields are available for FSU-HYCOM, but not published. For Fig.6, E. Chassignet and A. Bozec have recomputed the MLD from one year of daily output, using the method we have chosen in our paper, to enable the comparison.

I agree that for purposes of comparability it was necessary to use MLD(⟨T⟩,⟨S⟩), but suggest that it would have been better to have tried a larger density difference criterion when using monthly-mean averages. For it is conventional when calculating MLD using monthly-mean observational climatologies to use a larger density difference, e.g. WOA 2018, which uses ρ − ρ10m ≥ 0.125 kg m-3 s. It would be relatively straightforward to calculate MLD using monthly-mean T and S, with a range of density differences starting up from 0.03 kg m-3 and ranging up to 0.125 kg m-3 using e.g. CMCC-NEMO, and choose the density difference that minimised deviations from the monthly-average of the MLD calculated online with the 0.03 kg m-3 density criterion (i.e minimise the difference field plotted out in the panels in Fig. 6).

We agree that the use of a larger density jump when working with monthly model output could make differences such as shown in Fig.6 smaller for any given month, but it is unlikely that a single value of the density jump would effectively reduce the difference at all latitudes and for all months. As an example, the difference in MLD computed with the two density thresholds (0.03 and 0.125) is shown below for the ISAS climatology. Its seasonal cycle differs from the seasonal cycle of the daily-monthly MLD differences shown in our Fig.6. For this reason we do not wish to advocate this approach in our manuscript.



Recommendation

This manuscript will be very useful for the field and should be published subject to minor corrections.

We thank the reviewer for this positive recommendation.

Detailed comments

P6, l138. Does IAP-LICOM include the wave generated mixing of Qiao, et al., GRL, 2004?
No, it does not.

P6, l164 "NEMO version 3.64" Is this what you mean?
Thanks for catching this error, we meant NEMO version 3.6. We have corrected the number.

P14, Fig. 4. Caption. Observation datasets ⇒ observational datasets.
Corrected.

P17, Fig. 4. Why not use NCAR-POP and CMCC-NEMO for consistency?

I suppose you mean Fig 6. Unfortunately, the daily 3D fields have not been saved for NCAR-POP, and we did not have the resources necessary for the re-computation of the MLD from the 3D daily fields stored at CMCC (see our answer to your comment on Fig 6 above).