Review of gmd_2022_27: **Cloud-based framework for inter-comparing submesoscale permitting realistic ocean models** by Takaya Uchida and others

Review by Mike Bell

Summary of paper: This paper outlines a new approach to generating analysis-ready cloud-optimised (ARCO) data whose purpose is to enable open-source scientific analysis of large ocean / atmosphere data sets. It uses this approach to inter-compare @8 high resolution ocean model simulations in the Gulf Stream separation region with a focus on the sea surface height fields and the vertical buoyancy fluxes by sub-mesoscale mixed-layer eddies (SMLE).

This paper is generally well-written and presents some interesting results. It is also part of an important pioneering effort to improve the intercomparison of large data sets. So I expect to recommend that it is accepted for publication after some revision. In intercomparisons of this sort it is important to help the reader to work out what are the most significant scientific results. I make a number of suggestions below that I hope will help the authors to improve their presentation in this regard. I also ask some questions about the ARCO approach which relate to its sustainability. There is finally a list of minor points. In these “grammar” indicates something is not quite right in the sentence; the problem hinges on the word I’ve picked out.

Presentation of results (assisting reader to more easily grasp the main scientific points)

1. L103 and L108-109: Some more comments on Appendix A at this stage would assist the reader. For example

- Say that most models are spun up for only 12-18 months. Would it be possible to show HYCOM50 data after a 12-18 month spin-up? (were the data required to do that archived?) This would reduce the inhomogeneity of the data.
- Could Table A1 include the date of the start of the spin-up and the start and end of the analysis period (or at its least its length)? This is needed for the SSH analysis
- Table A2: were some of the bathymetries smoothed or edited more than others?
- Table A3: note that vertical resolution as well as horizontal resolution varies significantly between the models
- Table A4: is the flux form (rather than vector form) for momentum the default? If the vector form is used is a Hollingsworth correction the default? Please be explicit about this (this is relevant to the vorticity plot)
- Table A4: note that some models have biharmonic viscosities and others do not (relevant to vorticity plot)
- Table A5: note that FESOM-GS and ORCA36 do not have tidal forcing whilst the others have at least the leading 5 tidal forcings.

2. L104: Is it not possible to discuss the differences that are all too visible in Figure 2 in more detail? Couldn’t you include an analysis of the time-scale of viscous damping in the models near the grid-scale to see to what extent that could account for the results?
3. Colour scales in figures 2, 3 and 5: Is it possible to use a colour scale which has a somewhat better dynamic range. I can only really make out 5 colours; dark red/blue, light red/blue, white.
4. L128-129: It seems to me that Figure C1 is more scientifically interesting than Figure 3 as the differences due to tidal forcing differences are reduced and there is a comparison with AVISO data. Useful intercomparisons of this sort can be a lot of work. More careful removal of the tides might give a more interesting comparison.

5. L133-134: “interesting”: this is a necessary preliminary step in assessing the std deviation. Can you group the models into tidally forced and unforced or indicate “no tides” on the title of the panel for ORCA36 and FESOM-GS?
6. L139: ”applying tidal forcing”. Another possibility is to remove tides from the SWOT data (as is routinely done for altimeter data). Atmospheric surface pressure forcing is relevant as well as tidal forcing. It's only when the tidal and mesoscale interact that both need to be treated within one model.
7. L140-154: I haven't managed to work out what scientifically valuable points can be extracted from Figure 4. Most of the points made relate to differences in the forcing. There are probably too many lines on the plots. Perhaps you could separate the winter plot into two plots each with 4 lines. Do you compare the winter and summer plots in the text?
8. Section 3.2 seems to be carefully done and a good analysis though I have a question about equation (2) – see minor points.
9. L209-212: The sub-mesoscale buoyancy flux diagnosed from the models is relatively large in FIO-COM32. It's quite hard for the reader to compare the models because the scales on the ordinate in Figure 6 are different in each of the panels.
10. L259: See comment on L139.
11. Figure D1: negative values of C(t) puzzled me. This implies that just occasionally the best fit is obtained using a negative C for the whole domain. I think you calculate C by minimising the square differences of the parametrised and actual fluxes. One would usually plot C(t)*MLI on the x-axis. The slope of the scatter fit would then be shallower than the 1:1 line. It would be informative to give some correlation coefficients. Figure D1 map plots: The scales on some map plots are twice those on others. This makes intercomparison difficult.

Sustainability of ARCO methodology

In order to understand the ARCO methodology, I read the Stern et al. 2022 (S22) paper. That paper is well organised and very helpful. The summary of it in section 2 complements it well, being significantly shorter, and comprehensible on its own.

S22 section 4.1 mentions that one of the lessons (re-)learnt from your SWOT exercise was that data transfers between sites is slow and hence geographical proximity is important. Does the S22 approach need to be adjusted in recognition of this point?

I have no expertise in cloud computing. But there are some basic principles of data access that seem to me to be quite generic. The largest data sets have to be stored on cheaper forms of data storage like cartridges that are slow to load up on a system. If a data set is spread across 100s of cartridges access to the data will be slow/expensive. So to be analysis-ready, data has to be sub-setted in a way that suits the type of analysis that will be performed. The data might need to be laid down in perhaps 2 or 3 different ways to enable a wide range of analyses. It's not clear to me what sub-setting approaches you have taken. Is the restriction of the data to the GS region and the time period you have chosen sufficient? For example, for the 3D data required for the SMLE analysis, was the data laid down in a way that was tailored to your analysis – or was it small enough to be stored on disk? It seems to me that the sustainability of the ARCO approach may depend on astute or pragmatic solutions to these issues.

Many readers will have similar questions in their mind when they read the paper. If the previous paragraph contains invalid assumptions it would be useful to point them out so that the proposed approach is better understood. The co-authors include the main authors on S22, and the paper is not particularly long, so discussion of such points might be within the scope of your paper. The discussion of sustainability otherwise seems lacking in depth.

L236: 100 Euros per month: is this the total cost (that seems unlikely!) or the cost per user given the quoted 64 cores and 256 GB of memory?

Minor points

L18: “each party”: sometimes one group (often an independent group) is elected to do the analysis

L22: grammar: “needed”

L28: grammar: “by”

L42-43: As I understand it from Stern et al 2022, the data were collected in one place and this was a slow step. In principle they could have been stored close to their origin - but in ARCO formats. I suspect this would create its own difficulties.

L84: ”which unifies the API” could you say “which unifies the API to read and load the data” ?

L103: The tables are in appendix A not appendix B.

L112-115: You might consider re-writing this to avoid saying SSH is also known as ADT. The SWOT ADT (not the SSH) needs to be compared with the model’s SSH.

L119-120: You can calculate standard deviation of a from $\overline{a^2} - (\bar{a})^2$ so don’t really need access to the temporal dimension for this diagnostic.

L129: grammar: tend -> tends

L180: grammar: were

L199-203: I’m not sure I follow this sentence. I believe that the square of the horizontal buoyancy gradient is expected to scale with $(\Delta s)^{-1}$ and there should be a factor $\Delta s/L_f$ in (2) where L_f is a “modified mixed-layer Rossby radius”. I think the analysis that has been performed must take this into account.

L269: grammar; “were”