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 #2 (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.
We thank the referee for his thorough review of our manuscript.

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
    a. 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.
    Regarding HYCOM50, it was spun up from rest and integrated for 20 years. Sensitivity experiments were performed starting from year 15. It took a minimum of 5 years to reach mechanical equilibrium (Chassignet and Xu, 2017).
    For the first 5 years of the HYCOM50 integration, only saved monthly fields were saved. Daily afterwards. We have added this in the table caption.
    We note that Fig. 2 is for illustration purposes and not an in-depth model comparison since the models all have differences in their setup. The integration time of 12-18 months would suggest that the large-scale features from the other simulations are still sensitive to their respective initial conditions.

    b. 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
For the sake of storage, only three months for summer (Aug., Sep., Oct.) and winter (Feb., Mar., Apr.) are saved from an arbitrary year per simulation, which are then used for the SSH analysis. We have added the year of outputs used to Table B1.

c. Table A2: were some of the bathymetries smoothed or edited more than others?
We have added this information to Table B4.

d. Table A3: note that vertical resolution as well as horizontal resolution varies significantly between the models
Added in the table caption.

e. 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)
We have added this information to the table.

f. Table A4: note that some models have biharmonic viscosities and others do not (relevant to vorticity plot)
Added in the table caption.

g. Table A5: note that FESOM-GS and ORCA36 do not have tidal forcing whilst the others have at least the leading 5 tidal forcings.
Added in the table caption.

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?
This is indeed an important point and a likely culprit for the difference we see in Fig. 2. We would like to leave a more in-depth examination for a subsequent paper where we discuss the effect of numerics on the resolved submesoscale flow. Here, we have kept the focus of the manuscript on the implementation and application of the Pangeo Forge framework and showcased a few example diagnostics.

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.
We have changed the colormap of the standard deviation in Fig. 3 to purple/orange to differentiate from the blue/red colormap used for demonstrating the mean.

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.
While we agree that the comparison with AVISO is interesting (i.e. Fig. C1), in light of

SWOT, we would like to keep Fig. 3 in the main text to highlight the importance of including tidal forcing in numerical simulations.

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?
   We have added “no tides” in the titles.

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.
   To our knowledge, the accurate removal of tidal forcing is an area of on-going research (particularly for the non-phase-locked (incoherent) part of the internal tide signals; Zaron and Ray, 2018; Carrere et al., 2021). The benefit of having tidally forced simulations is that we can develop and test such methods of removing tides. We have noted this in line 149.

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?
   We have split the spectra plots so as to enlarge them.
   We have also added in the text in lines 165-171: “It is interesting to note that at time scales of O(1-10 days), most runs show higher variability during winter than summer (Figure 4a,c), while the tidally forced runs show higher variability at time scales shorter than O(1 day) during summer (Figure 4b,d). The seasonality at time scales shorter than O(1 day) is reversed for ORCA36, a run with no tidal forcing. It is possible that the increase in forward cascade of energy stimulated by the tides are the culprit for higher SSH variability at time scales shorter than the inertial frequency during summer than winter for the tidally forced runs and vice versa for the non-tidally forced runs (Barkan et al., 2021). The overall higher SSH variability at time scales longer than the inertial frequency during winter than summer, on the other hand, is likely due to wind-driven inertial waves (Flexas et al., 2019).”

8. Section 3.2 seems to be carefully done and a good analysis though I have a question about equation (2) – see minor points.
   Please see our reply to the referee’s minor point below.

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.
   We agree that the flux is relatively large in FIO-COM32. Unifying the y-axes would mean that we would have the same axes for all simulations as the axis used for FIO-COM32. This would unfortunately make the temporal fluctuations of HYCOM50 difficult to observe. We have added in the caption: “Note that the y axes vary depending on the magnitude diagnosed from each simulation in order to highlight its

temporal variability."

10. L259: See comment on L139.
    Please see our reply corresponding to 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.
    C(t) is never plotted in Fig. D1 (now Fig. 7) and is always positive (cf. blue lines in Fig. 6). It is calculated by taking the spatial median of C(t,x,y) diagnosed at each grid point by taking the ratio of MLI and the actual flux. This is described in lines 225-226 as: "We diagnosed Ce by taking the ratio between the right-hand and left-hand side of equation (2) at each grid point and time step, and then the horizontal spatial median of it."
    The aim of the map plots is to exhibit the actual buoyancy flux and its equivalent predicted from the MLI parametrization within each model. If we were to unify the scaling, this would make some panels saturate while making it difficult to see the signal for others. We have increased the size of the figure and added in its caption: "Note that the range of colorbar differs depending on the magnitude diagnosed from each model to highlight their spatial features and comparison between the submesoscale buoyancy flux and its equivalent predicted from the parametrization per simulation."

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?
We agree with the referee that in light of limited storage (which is always a constraint in storing and distributing model outputs), some astute planning is needed beforehand to decide what variables and which regions are to be stored.

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.
As the referee correctly points out, the amount of data to be stored on the cloud is limited by the funding allocated to the cloud storage. The daily averaging of 3D data (as opposed to hourly outputs) was indeed done due to storage constraints. While we have focused the sub-setting to regional data that correspond to a few SWOT Crossover regions, we believe a 10x10 degree subset over the upper 1000m allows for various analyses to be conducted. For example, the dataset was not tailored for the SMLE analysis where storage of hourly outputs would have allowed for further analysis on how the interaction between inertia- and internal-gravity waves and submesocale flows would have affected the SMLE parametrization. We have noted in the manuscript in line 82: "... (due to cloud storage constraints)."

L236: 1000 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?
We have added more details on the total cost in lines 266-271 as: "Currently as of writing, the cloud storage provided by OSN is funded by an NSF grant acquired by the Climate Data Science Laboratory at Columbia University, and the JupyterHub on Google Cloud Platform by Centre National d'Études Spatiales (CNES) funding. The cost of cloud resources for the JupyterHub with parallelized computation adds up to roughly 1000 € per month with the maximum computational resources of 64 cores and 256 gigabytes of memory per user; the resources scale on-demand, while the cost of operating the scalable Kubernetes infrastructure is managed by a vendor (2i2c) for a few thousand dollars a month. Although this may seem expensive…"

Minor points

- L18: "each party": sometimes one group (often an independent group) is elected to do the analysis
  We have added: "(often an independent group)".

- L22: grammar: "needed"
  Corrected.

- L28: grammar: "by"
  Corrected.

- 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.
We believe that having the ARCO data storage centralized facilitates the managing of data.

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

- L103: The tables are in appendix A not appendix B.
We have corrected the table referencings to Appendix B. Appendix A describes the Pangeo Forge recipes.

- 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.
In response to referee #1’s comments, we have rephrased this as: “In light of the SWOT mission, the primary variable of interest is the ocean dynamic sea level. The AVISO estimate of this quantity is called the Absolute Dynamic Topography (ADT), while the closely related model diagnostic is their Sea Surface Height (SSH) after correcting for the inverse barometer effect if atmospheric pressure variability was simulated. Technically, SSH is defined as the geodetic height of the sea surface above the reference ellipsoid, while ocean dynamic sea level (or ADT) is defined relative to the geoid, but in models where the geoid and reference ellipsoid coincide these two definitions are in practice the same (Gregory et al., 2019). Furthermore, in the specific comparisons made here, a regional average of the ocean dynamic sea level estimates is removed first, so that large-scale, slow changes (e.g., ice sheet contributions) are excluded from the comparison.”

- L119-120: You can calculate standard deviation of a from $\overline{a^2} - (\overline{a})^2$ so don’t really need access to the temporal dimension for this diagnostic.
While we agree with the referee, $\overline{a^2}$ is not always a saved variable as model output.

- L129: grammar: tend -> tends
Corrected.

- L180: grammar: “were”
Corrected.

- 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 Δs/Lf in (2) where Lf is a “modified mixed-layer Rossby radius”. I think the analysis that has been performed must take this into account.
Our understanding is that the Δs factor is there to account for weaker resolved buoyancy gradients when the model resolution is coarse. In other words, the resolved buoyancy gradient at $O(0.1^{\circ})$ is different/weaker than the coarse-grained fields of buoyancy resolved at $O(1/50^{\circ})$.
Quoting from Fox-Kemper et al. (2011): “*The MLE parameterization (5) is*

proportional to the horizontal density gradient, a quantity that depends strongly on horizontal resolution. Coarser models have weaker gradients than finer, and sparser observations have weaker gradients than denser. Additionally, the MLE parameterization in (5) is based on one resolved front, rather than a sea of statistically-distributed fronts of varying strength and orientation. Fortunately, one can scale for these effects based on an analysis of the horizontal wavenumber spectrum of near-surface density variance. The Δs/Lf factor in (6) is the result of this analysis (Section 2.1.3). This rescaling can be done with some confidence, as the same near-surface density variance spectrum is found in observations (Section 2.1.1) and in model hierarchies designed to study the effects of differing resolution (Section 2.1.2).”

As the model outputs we use are all submesoscale permitting, we have left the scaling factor (Δs) out from our analyses. We have noted this as: “Δs was omitted due to all our model outputs partially resolving the submesoscale buoyancy flux. Furthermore, as Δs doesn't vary much among the models, this factor would not contribute much to the overall differences between models, in comparison to the greater variability due to numerics, etc., this manuscript is meant to introduce.” in lines 222-224.

- L269: grammar; “were”
  Corrected.

Reference

- Barkan, R., Srinivasan, K., Yang, L., McWilliams, J.C., Gula, J. & Vic, C. (2021). Oceanic mesoscale eddy depletion catalyzed by internal waves. *Geophysical Research Letters*. doi:10.1029/2021GL094376;
- Carrere, L., Arbic, B.K., Dushaw, B., Egbert, G., Erofeeva, S., Lyard, F., Ray, R.D., Ubelmann, C., Zaron, E. & Zhao, Z., et al. (2021). Accuracy assessment of global internal-tide models using satellite altimetry. *Ocean Science*. doi:10.5194/os-17-147-2021;
- Chassignet, E. & Xiaobiao, X. (2017) Impact of Horizontal Resolution (1/12 to 1/50) on Gulf Stream Separation, Penetration, and Variability. *Journal of Physical Oceanography*. doi:10.1175/jpo-d-17-0031.1;
- Flexas, M.M., Thompson, A.F., Torres, H.S., Klein, P., Farrar, J.T., Zhang, H. & Menemenlis, D. (2019). Global estimates of the energy transfer from the wind to the ocean, with emphasis on near-inertial oscillations. *Journal of Geophysical Research: Oceans*. doi:10.1029/2018JC014453;
- Gregory, J. M., Griffies, S. M., Hughes, C. W., Lowe, J. A., Church, J. A., Fukimori, I., Gomez, N., Kopp, R. E., Landerer, F., Le Cozannet, G. and others. (2019) Concepts and terminology for sea level: Mean, variability and change, both local and global. *Surveys in Geophysics*. doi:10.1007/s10712-019-09525-z;
- 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;

- Zaron, E.D. & Ray, R.D. (2018). Aliased tidal variability in mesoscale sea level anomaly maps. *Journal of atmospheric and oceanic technology*. doi:10.1175/JTECH-D-18-0089.1;