This paper deals with closure of the global ocean mass budget for the period 2005-2020 using a combination of observation of model-based data products. This type of study is crucial to our understanding of observed climate change and identifying potential issues / limitations in observing capability and/or data processing. The manuscript is well-written and the figures are of high quality. Dealing with numerous observation and model-based datasets is always challenging in terms of understanding all the potential issues. My main comments below encourage the authors to offer a bit more discussion of the non-budget-closure and to include more quantitative information about how this could be accommodated by the various hypotheses they put forward. They could also consider the relative sizes of estimated uncertainties and/or instances where the uncertainty estimation may be limited by ensemble characteristics, or for other reasons. I find the manuscript to be suitable for publication subject to addressing my comments below.

I'm unsure about the term Global Mean Ocean Mass used throughout the manuscript. I tend to think of this as having units of mass per unit area, but I think we are talking about changes in the total ocean mass? Perhaps there is some explanation or convention that could be mentioned and this point clarified at the start of the manuscript.

I would like to see the authors spend a bit more discussion on the non-closure of the budgets shown in Figures 6, 7 and 8. Fundamentally, when a budget does not close, it suggests that the uncertainties in one or more components have been underestimated. This point is worthy of some discussion and perhaps some speculation on where limited ensembles or diversity across the ensemble may be playing a role in the uncertainty estimation. I led a recent paper where we presented a generic framework for using ensembles to characterise uncertainty, which may be of interest to the authors, Palmer et al [2021]: https://iopscience.iop.org/article/10.1088/1748-9326/abdaec

In the closing sentences of section 4.3, the authors cite "deep ocean below 2000 m depth, the atmospheric water vapour variations and the permafrost thawing” as potential
explanations for non-budget closure. I wonder if the authors could offer some more quantitative information in this regard. How large would the temperature variations below 2000 m depth have to account for the residuals, and so on for atmospheric water vapour, permafrost thawing. Would it be helpful if the Argo-based estimates of thermosteric sea-level change could include some estimate of the additional uncertainty below 2000 m? However, I suspect that the horizontal sampling uncertainty may still dominate - and perhaps this is underestimated, as mentioned in my comment on Figure 7. The paper by Allison et al [2019] https://iopscience.iop.org/article/10.1088/1748-9326/ab2b0b neatly illustrates the potential for mesoscale ocean “noise” to introduce spurious signals on a range of timescales, which may be inherent to the observational sampling and common to several (all?) data products? The authors may wish to comment in this regard.

Figure 6 seems to show a strong correlation between the WGHM TWS time series and the GRACE ensemble mean. Could these signals have been under- or over- estimated in one of the products? I think this point is worth some discussion. Can the authors comment on the different temporal resolution of the underlying datasets and ability to resolve the signals? This may also contribute to non-closure. This is mentioned briefly in section 4.3, but discussion of specific timeseries characteristics in section 4.1 and 4.2 could aid the reader.

Figure 7(a): The uncertainty on the in-situ thermosteric ensemble mean is much smaller than I would have expected. Was this informed simply from ensemble spread, as shown in Figure 5? Palmer et al [2021] argues that “structural uncertainty” from ensemble spread needs to be combined with some estimate of “internal/parametric uncertainty” in order to fully characterise the total uncertainty. I would encourage the authors to give this standpoint some consideration and update the uncertainty estimate if appropriate.

Figure 8: I don’t see panels (c) and (d) in this figure, as implied by the figure caption. I think perhaps the caption descriptions for (c) and (d) are intended to apply to panels (a) and (b)? Similar to Figure 6, there are some apparent correlations between the residuals and the TWS timeseries in particular. One thing to be cautious of is that fact that delayed-mode quality control of Argo floats typically takes 1-2 years to complete. Therefore, the last 1-2 years of data can be considered “provisional” and may be subject to revision, although I think this is generally considered more of an issue for salinity data, as noted here https://floats.pmel.noaa.gov/float-data-delayed-mode-quality-control

Figure 9: I’m not sure I understand the plot titles. The summation symbol would tend to suggest to me that the quantity subtracted is always (GIS + AIS + GIC + TWS), but this is not this is not the case for panel (c)? I think the similarity in the timeseries shown in Figure 9(b) strongly implies that GRACE and (GIS + AIS + GIC + TWS) must have similar timeseries, as shown in Figure 9(c), so I’m not sure how much additional information this really offers the reader. In addition to the trends, are there physical insights we can draw on from the variations/similarities in the residuals? In the figure caption, please clarify the precise period that trends are calculated over - e.g. 1st Jan 2015 to 31 Dec 2018 or similar.
On the residual trend, it's helpful to be explicit on whether the GRACE-determined mass trend is larger or smaller than the sum of individual components.

Stylistic choice, but I would recommend replacing “Besides” with “In addition”. Same sentence, suggest replacing “water vapour” with “water content”.

Please cite the latest IPCC AR6 report and specify a period for which the two-thirds statement applies (this has changed over time), as noted in the Working Group I summary for policymakers and Chapter 9 (Fox-Kemper et al, 2021).

Typo? Replace “float” with “floats”.

Replace “by the Argo float” with “by Argo floats”.

Could you briefly comment on the choice of GIA dataset and what effect a different dataset might have on your analysis? Some idea of the importance of this for the reader would be helpful.

Can you comment on the physical plausibility of some of the very sharp drops in the datasets seen in 2017? E.g. what would this imply for rainfall over land and subsequent river flows? How do these timeseries compare with timeseries of terrestrial land water storage shown later in the manuscript? I suspect that this cursory analysis would support “noise” as the main candidate explanation.

I would suggest a different notation for Epsilon between these equations, perhaps Epsilon1 and Epsilon2. This would make clear that these quantities are fundamentally different (a dash is often used when one quantity is a proxy for the other) - they are related to completely independent datasets.

I don’t understand why the standard uncertainties are raised to the power 3 before summing them. Is there a reference you can cite that explains this approach? Or offer some additional explanation.

Please include an explanation of the units of mass in global mean sea level equivalent. A second y-axis in units of Gt or similar could usefully be included.
Figure 2: Same comment as for Figure 1 applies here. Please consider this point for all subsequent figures (may be more appropriate to some figures than others).