The authors thank Dr. Catia Domingues for the extensive suggestions on the paper, here the point-to-point replies are provided in blue, the comments are in black, the modified texts for the manuscript are shown in orange.

Major comments:

1) No dedicated section on caveats.

Why do caveats are not really discussed to inform users? For example, the gridding process (section 2.6) relies on CMIP model covariances for infilling, so this IAPv4 product (and earlier versions) are not purely based on observations. This observational-model mixed approach has circular implications for studies focused on comparison or evaluation of CMIP models, detection & attribution as well as in constraining CMIP model projections. In order words, the use of IAPv4 is not appropriate for these types of studies.

Re: We do have an extensive discussion on the remaining issues, see the last four paragraph of the final section, I copied here "**Despite several marked improvements, issues needing further investigation remain.** Although inter-annual and decadal-scale changes of satellite-based EEI and observational OHC are generally consistent, a mismatch remains between EEI and OHC for their month-to-month variation, as the monthly variation of OHC is still much larger than implied by EEI. There are several possibilities, in our opinion: first, there is substantial heat storage and release for land and ice monthly, which needs to be accurately quantified; second, the accuracy of OHC estimate on a monthly basis still needs to be improved for month-to-month variation because of the limited data coverage; third, the EEI observed by CERES also suffers from sampling biases on a monthly basis (Loeb et al., 2009). Thus, a better understanding of the monthly variation of OHC and EEI is still a research priority. Besides, the failure to close the 2015-2023 sea level budget indicates that the underlying data still has bias problems, which need to be explored and resolved.

Second, the application of CODC-QC in IAPv4 leads to a stronger ocean warming rate in the past decade than WOD-QC used in IAPv3 because WOD-QC removes more positive temperature anomalies than CODC-QC. This could imply that the rate of increase in OHC is still slightly underestimated and deserves an in-depth investigation. Several fundamental questions must be answered: first, are there still real temperature extremes being removed by CODC-QC, such as in small warm/cold eddies? Are the extremes well sampled by the current observation system? If not, what is the impact? Moreover, it is clear that the high latitudes where sea ice occurs are not well sampled and need more attention.

Third, during the development of the data product, we discovered that much metadata relating to the profiles in the World Ocean Database is missing and that much existing metadata is incorrect, also giving rise to duplicate profiles, putting a strain on the overall quality of a database of oceanic observations. More than ever, long-term concerted efforts are needed to eliminate duplicate profiles and identify and correct missing metadata using statistical methods, expert control, or machine learning techniques. For example, the International Quality-Controlled Database (IQuOD) group is coordinating some activities related to data processing

techniques, uncertainty quantification, and improving the overall quality of ocean data (Cowley et al., 2021).

Furthermore, the quantification of uncertainty for *in situ* measurements, gridded T/OHC values, and the global OHC estimates need to be improved. IAPv4 only accounts for the instrumental error and sampling/mapping error. In the future, comprehensive quantification of other uncertainty sources will be made, including the choice of climatology, vertical interpolation, XBT/MBT/APB/Bottle corrections, etc. It is also necessary to analyze the correlation between these error sources. This also helps to understand regions with larger uncertainty for OHC estimates, which supports the design of the global ocean observing system."

For the specific point of using CMIP5 models, we do not regard this technique as a caveat; instead, based on our previous studies (Cheng&Zhu, 2016, Cheng et al. 2017, 2020) and independent evaluations by many publications, this approach provides a good estimate of ocean temperature and OHC change and the estimate uncertainty is also provided. It is up to users to use this dataset or not, if the methods are transparently described, as in Section 2.6 "For each month, IAPv3 used 40 model simulations (historical runs) from the Coupled Model Intercomparison Project phase 5 (CMIP5) to provide a flow-dependent ensemble, which is then constrained by observations to provide optimized spatial covariance. IAP mapping uses model-based covariance because we argue that spatial covariance can never be satisfactorily parametrized by some simple basic functions (such as Gaussian) given its complexity. With model-based, flow-dependent, and dynamically-consistent covariance, the IAP mapping provides a more realistic reconstruction than other approaches based on Gaussian-based parameterized covariance, as evaluated by many studies (Cheng et al., 2017; Cheng et al., 2020; Dangendorf et al., 2021; Nerem et al., 2018)."

2) Although sea surface temperature was evaluated, no evaluation for the abyssal ocean was done (below 2000 m), relatively important given that this is a new aspect from previous versions and which also differ from published analyses. This could be done by subsampling the gridded data where profiles exist and compare differences.

Re: For the upper 2000 m estimate, we have subsampled the Argo-period observations which has near-global-ocean-coverage (Cheng et al. 2017) to evaluate the estimate. However, the deep ocean changes are different as we don't have data with satisfied global coverage to do this test. Therefore, this study simply applied the IAP mapping method to the deep ocean observations and gave an estimate, similar to previous studies (Purkey and Johnson, 2010, and EN4 data), just different mapping techniques are applied. There is no published method to evaluate the available deep ocean OHC estimates (i.e., subsampling the gridded data). We have mentioned this aspect in the Discussion section as a caveat to caution the users: "Fourth, the deep ocean changes below 2000 m are estimated based on the currently available data, including data from hydrological sections and Deep-Argo. IAP mapping technique is applied. Because of the lack of independent observations with global ocean coverage, evaluating the deep ocean change estimate is still an open issue. Thus, the below-2000 m estimate should be used with caution, as also indicated in previous estimates (Purkey and Johnson, 2010; Desbruyères et al. 2017; Good et

al. 2013). A community-agreed evaluation approach for the deep ocean changes is critically needed."

Other comments:

Line 51: Gridding methods are also the main source of spread among observational estimates, as found in Boyer et al. (2016) and Savita et al. (2022). Please inform the reader.
Re: Gridded methods are to resolve the "irregular and incomplete data coverage" issue, so this aspect has been mentioned here.

Line 69: Missing Domingues et al. and Johnson et al. in the list.
Re: Good references, but as this is not a review paper, and these are just examples of a large set of literature. Please refer to the recent review paper (Cheng et al. 2022) for more comprehensive information. Domingues et al. are not referred to here because there is no T/OHC gridded fields available for this dataset, which is not comparable to other data products. The papers by the G.Johnson group are cited in some places and their data have been added in the revised manuscript for comparison; for example,

Johnson, G. C., Purkey, S. G., Zilberman, N. V., and Roemmich, D.: Deep Argo Quantifies Bottom Water Warming Rates in the Southwest Pacific Basin. Geophys. Res. Lett., 46, 2662-2669, https://doi.org/10.1098/rsta.2022.0188, 2019.

Lyman, J. M., and Johnson, G. C.: Estimating Global Ocean Heat Content Changes in the Upper 1800 m since 1950 and the Influence of Climatology Choice. J. Climate, 27, 1945-1957, https://doi.org/10.1175/JCLI-D-12-00752.1, 2014.

Lyman, J. M., Good, S. A., Gouretski, V. V., Ishii, M., Johnson, G. C., Palmer, M. D., Smith, D. M., and Willis, J. K.: Robust warming of the global upper ocean. Nature, 465, 334-337, https://doi.org/10.1038/nature09043, 2010.

Lyman, J. M., and G. C. Johnson, 2023: Global High-Resolution Random Forest Regression Maps of Ocean Heat Content Anomalies Using In Situ and Satellite Data. J. Atmos. Oceanic Technol., 40, 575–586, https://doi.org/10.1175/JTECH-D-22-0058.1.

Line135: My understanding is that the grey list is for operational centres. Profiles on that list should not be removed in your case. Please check with Argo data management team.
Re: As described by the Argo community, the "grey list" contains a list of active Argo floats that are suspected of malfunctioning (Wong et al. 2023). For example, in Argo webpage, it states "*In early 2007, it was discovered that Argo profiles from SOLO floats with FSI CTD (Argo Program WHOI) may have incorrect pressure values. The problem did not affect any other combination of instrument and sensor. In GTS TESAC messages, potentially affected instruments can be identified by instrument type 852 (SOLO FSI, see WMO Code Table 1770). Some profiles can be corrected automatically and some need additional study. The automatic fix for these profiles was instituted on 10 October, 2007. For profiles need additional attention, the float will stay on the greylist until the profiles have been fixed. While studying the pressure offset errors, a related problem was discovered in a group of WHOI/SBE profiles. For the affected WHOI/SBE instruments, all profiles*

have been corrected and are available on the GDACS as of 14 September 2007." Described in (https://argo.ucsd.edu/data/data-faq/)."

Thus, although it is definitely upon the users to choose to use or not use the grey list, it is highly likely that floats on the grey lists are problematic data identified by the Argo group. Therefore, the IAP group decided to use the grey list to reduce the risk of including data from malfunctioning floats.

Wong, A. P. S., Gilson, J., and Cabanes, C.: Argo salinity: bias and uncertainty evaluation, Earth Syst. Sci. Data, 15, 383–393, https://doi.org/10.5194/essd-15-383-2023, 2023.

Line 136: Why those data are not directly available via WOD?
Re: They can be available via WOD (if the WOD group agrees) after the paper's publication (as the data will be publicly accessible, we have another publication on the way to describe these data).

Line 177: There are several definitions for extreme events. Which one are you using? Please include reference and rationale for selecting one of the various definitions.
Re: The sentence has been modified to "Local climatological ranges change with time to account for the long-term trends of ocean temperature accompanied by more frequent extreme events (e.g., Oliver et al., 2018). Previously, the use of the static local ranges tended to remove too many "extreme values" (at the tails of the temperature distributions) associated with climate change in recent years that were actually real, leading to a QC-procedure-related bias in the gridded dataset and OHC estimate (Tan et al., 2023);". The definition of the extremes is not relevant here. We did not define any extremes. It is a general description of high/low-temperature values at the tails of the temperature distribution. Please refer to Tan et al. (2023) for more information on how QC works.

Line 179: What is the reference for the real events?
Re: We added Oliver et al.2018 as an example for sea surface temperature and Sun et al. (2023) for subsurface changes.

Line 191: What is the reference for the manually QC-ed datasets?
Re: Tan et al. (2023) used Quata and WOCE datasets, and we added references here (Gouretski and Koltermann, 2004; Thresher et al., 2008).

Line 206: Has this been observed before in other publications, for example, Roquet et al?
Re: After double-checking the profiles before and after QC in 2008 below 4000 meters, and we find that this high rejection rate (~81.9%) is because of the gross errors in CTD data, not in Glider data (see Fig. R1 below). We also noted that this has not been reported before, to our knowledge, but we think it is easy to be identified by any QC system because they are apparent outliers.

Therefore, the sentence has been modified to "For example, the higher rejection rate within 2008-2009 below 4000 meters is because of the gross errors in the CTD data."

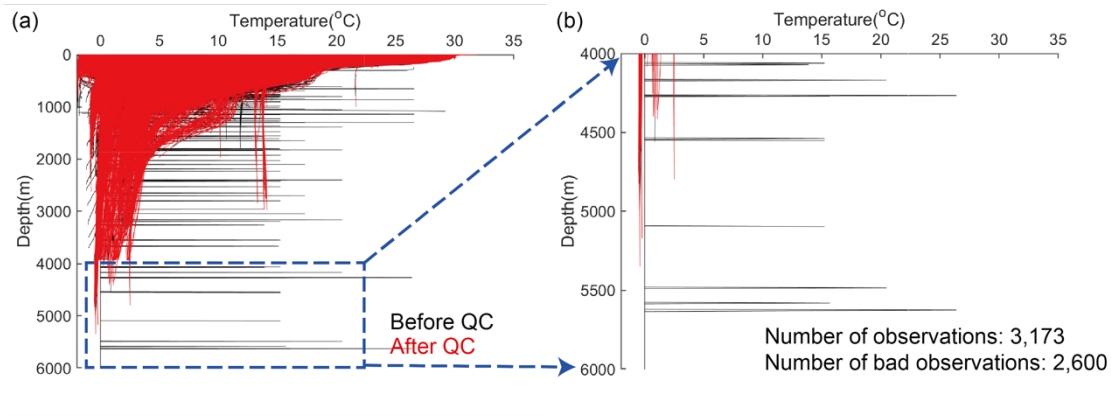

Fig. R1. (a) All CTD profiles in 2008. Here, the black color denotes profiles without performing any QC. The red color denotes the good profiles identified by CODC-QC. (b) is the zoom-in of (a) between 4,000 and 6,000 meters.

Line 215: It is true that Gouretski and Koltermann (2007) were the first to report on the XBT biases. However, Domingues et al. (2008) were the first to demonstrate the significant impact on the magnitude and variability of the global upper-ocean warming over multiple decades (see also AR5, ocean observations chapter).

Re: We respectfully disagree with this statement. First, Gouretski and Koltermann (2007) study is not the first to report on the XBT biases. The XBT bias has been found way back to old times around 1970 (see Cheng et al. 2016 BAMS for an overview). Second, Gouretski and Koltermann (2007) is the first to show that XBT biases have a global impact on OHC, and their abstract says "*We use a global hydrographic dataset to study the effect of instrument related biases on the estimates of long-term temperature changes in the global ocean since the 1950s. The largest discrepancies are found between the expendable bathythermographs (XBT) and bottle and CTD data, with XBT temperatures being positively biased by 0.2–0.4°C on average. Since the XBT data are the largest proportion of the dataset, this bias results in a significant World Ocean warming artefact when time periods before and after introduction of XBT are compared. Using bias-corrected XBT data we argue reduces the ocean heat content change since the 1950s by a factor of 0.62. Our estimate of the ocean heat content increase (0–3000 m) between 1957–66 and 1987–96 is 12.8·10^{22} J. Because of imperfect sampling this estimate has an uncertainty of at least 8·10^{22} J*". It is clear that it shows the global impact of XBT biases on OHC.Gouretski and Koltermann (2007) also demonstrated that un-corrected XBT data were responsible for an artificial decadal-scale variability of the global OHC time series.

As our paper is not a review paper and it is not a good place to discuss the history of XBT bias, we suggest to refer to the community papers Abraham et al. (2013), Cheng et al. (2016) and Goni et al. (2019) for more information.

Line 228: Please include comparison plots for the older and newer coefficients in the Suppl. Material, so readers can compare the differences arising from the update during the overlapping periods.

Re: The latest XBT bias correction scheme has been provided in www.ocean.iap.ac.cn/ under the label of Data Service -> New techniques. The code for implementing CH14 is also available (http://www.ocean.iap.ac.cn/ftp/images_files/XBT_cor_Matlab_code2023.zip).

Line 268: See also Boyer et al. (2016) for the impact of climatological choices.

Re: Boyer et al. (2016) added here.

Line 279: Refer to relevant figures in Rhein et al. 2013 (Suppl. Material).

Re: We are confused by this comment. This sentence in our paper stresses the inconsistency of different baselines at different locations could violate the spatial structure of the anomaly field. This aspect is not assessed in Rhein et al. (2013) (IPCC-AR5 reference). The most relevant plot in Rhein et al. (2013) is the Number of temperature profiles extending to 700 m depth in each 1° × 1° square, by decade, between 65°N and 65°S from the 1950s to 2000s. But this is only marginally relevant to our statement. Indeed, IPCC-AR5 (2013) only mentioned that the choice of climatology is one uncertainty source of OHC estimate, I quote: "*but other sources of uncertainty include the different assumptions regarding mapping and integrating UOHCs in sparsely sampled regions, differences in quality control of temperature data, and differences among baseline climatologies used for estimating changes in heat content (Lyman et al., 2010).*". On this basis, we'd like to avoid confusing readers and cite only the most relevant references here.

Line 309: How do these thresholds compare with the choices in Willis et al. 2007?

Re: I guess you are referring to Willis et al. 2008. This study used Argo data and did not provide thresholds for vertical intervals. Thus, if you can be more specific and point us to the resources, we would appreciate it.

Willis, J. K., D. P. Chambers, and R. S. Nerem, 2008: Assessing the globally averaged sea level budget on seasonal to interannual timescales. J. Geophys. Res. Oceans, 113, C06015.

Line 316: How does the distribution of depth levels in IAPv4 compare with WOA?

Re: Here are the standard levels in WOA (102 levels from 0 m to 5500 m) and IAPv4 (119 levels from 1 m to 6000m).

WOA_levels=[0,5,10,15,20,25,30,35,40,45,50,55,60,65,70,75,80,85,90,95,100,125,150,175,200,225,250,275,300,325,350,375,400,425,450,475,500,550,600,650,700,750,800,850,900,950,1000,1050,1100,1150,1200,1250,1300,1350,1400,1450,1500,1550,1600,1650,1700,1750,1800,1850,1900,1950,2000,2100,2200,2300,2400,2500,2600,2700,2800,2900,3000,3100,3200,3300,3400,3500,3600,3700,3800,3900,4000,4100,4200,4300,4400,4500,4600,4700,4800,4900,5000,5100,5200,5300,5400,5500]

IAPv4_levels=[1,5,10,15,20,25,30,35,40,45,50,55,60,65,70,75,80,85,90,95,100,110,120,130,140,150,160,170,180,190,200,220,240,260,280,300,320,340,360,380,400,425,450,475,500,525,550,575,600,625,650,675,700,750,800,850,900,950,1000,1050,1100,1150,1200,1250,1300,1350,1400,1

450,1500,1550,1600,1650,1700,1750,1800,1850,1900,1950,2000,2100,2200,2300,2400,2500,2600,2700,2800,2900,3000,3100,3200,3300,3400,3500,3600,3700,3800,3900,4000,4100,4200,4300,4400,4500,4600,4700,4800,4900,5000,5100,5200,5300,5400,5500,5600,5700,5800,5900,6000]
;

93 levels of the WOA and IAPv4 standard levels are the same. Most differences come from 100~400 m, where IAPv4 levels are denser than WOA, that's because we find the denser levels could help to improve the calculation of the mixed layer depth in the middle latitude regions. The set of standard levels is mostly practical, taking account of the vertical scales of variability. The standard levels are different for different groups.

Line 322: monthly "mean" climatology?
Re: Yes, "mean" added here.

Line 323: Why not median? (instead of mean)
Re: Before, the use of "median" usually aimed to minimize the impacts of (erroneous) outliers. However, after our QC, the data quality was satisfied in our case, so we used the arithmetic mean here to better define the gridded averages. And we did not find using a median significantly impacts our reconstruction.
A more fundamental question behind this choice is: given the skewed distribution of the temperatures in each 1deg grid and 1 month (i.e. not Gaussian in most places, see a statistic on skewness in Tan et al. 2023), what is the target estimate? Gridded median or gridded mean? They are different quantities. In our case, we want to estimate the gridded mean for IAPv4.

Line 327: Please include reference or evidence which shows that is physically grounded.
Re: This is a basic physics (more of a common sense) I think it is well-accepted by physical oceanographers: generally, upper ocean variability (month-to-month and inter-annual) is higher than in the deep ocean, and the deep ocean changes are generally slower than the upper ocean (because of the interactions between sea water and the "fast" atmosphere). To be more specific, this sentence is modified to "This process takes advantage of the larger persistence of anomalies (generally smaller monthly and inter-annual variability) in the deep ocean than in the upper ocean and thus is physically grounded."

Line 334: Is this procedure originally based on this reference? Should it be included?
Smith, D.M. & Murphy, J.M. (2007) An objective ocean temperature and salinity analysis using covariances from a global climate model. *Journal of Geophysical Research: Oceans*, **122**, C02022.
Re: The sentence in our paper on line 334 says "IAPv4 adopted a similar mapping approach (Ensemble Optimal Interpolation with dynamic ensemble: EnOI-DE) as in IAPv3 introduced in Cheng and Zhu (2016)". This is a description of EnOI-DE approach used for both IAPv4 and IAPv3. Thus, this sentence is not relevant to Smith&Murphy (2007), as they used an optimal

interpolation method with spatial covariance provided by a single model, so the approach is different.

Line 340: Does it account for narrow high-latitude fronts (e.g. across ACC)? Does the approach have awareness or does it mix water from two sides of fronts?

Re: The covariance is defined by the model ensemble, so if the model's simulation has fronts across ACC, it can represent them in error covariances. Here we provide an example, showing the temperature anomaly reconstruction in December 2023 at 500m (http://www.ocean.iap.ac.cn/), it is clear that the temperature anomalies along ACC can be well represented and the reconstruction is physically meaningful (I mean, describe the physics that we know).

For your second question "Does the approach have awareness or does it mix water from two sides of fronts?". I'm confused, how does any approach mix water from two sides of fronts?? I would appreciate it if you could provide specific comments on how to look into this.

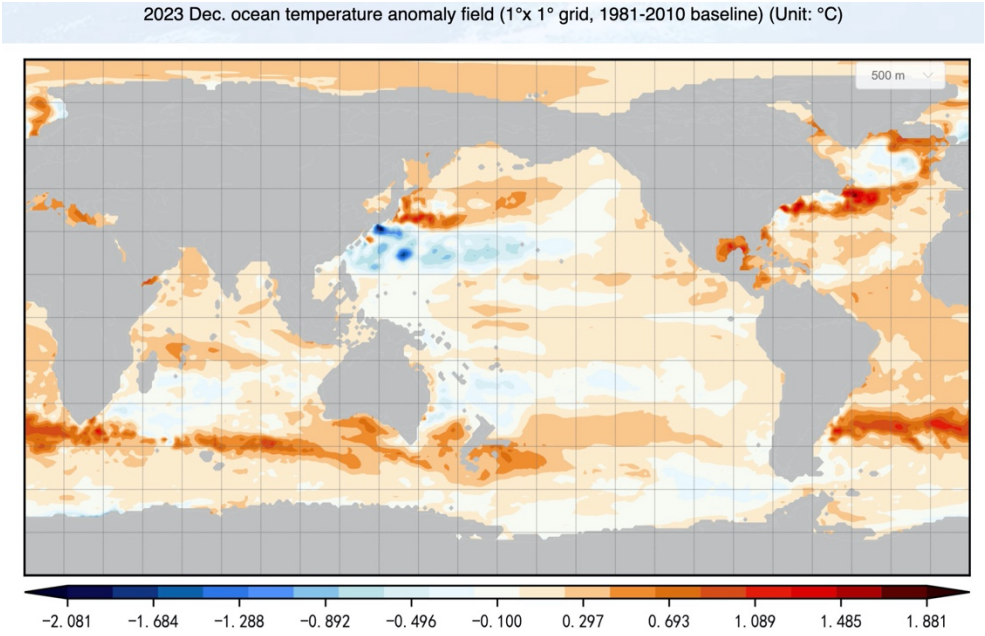

Figure R2. Temperature anomaly reconstruction field (IAPv4) at 500m at December 2023.

Line 348: What are the implications for certain studies when the gridded estimates are not purely observational?

Re: We would appreciate your being specific on this question. The method is transparent and peer-reviewed before. We refer to our replies to your major comments for the point of combing models with observations.

Line 352: Can this approach be benchmarked via the IAPSO's ME4OH working group best practices?

Re: ME4OH is an ongoing project which is in its initial phase, and there is no general agreement on how to benchmark the mapping approach; in other words, there is no published paper on this

issue. That's why we chose not to mention this activity before it can provide a community-accepted benchmark approach.

Line 370: Also compare to Savita et al. 2022.
Re: This sentence is about a large variance of temperature in some eddy-rich regions, which is not relevant to Savita et al. 2022, which compares OHC estimates based on different approaches.

Table 1: Does the radius of influence cross ocean basins or does it have awareness? What about frontal structures, particularly in the Southern Ocean?
Re: The radius of influence does not cross the land; this is a good point, and we mentioned it in the manuscript: "The radius of influence does not cross the land."

Line 396: See also Savita et al. 2022 and Meyssignac et al. 2019 studies on the impact of ocean masks.
Re: Both references added.

Lines 403-423: What is the difference it makes to the global values? 1%, 20%?
Re: Some quantification of the impacts on OHC is provided: " Since the open ocean accounts for the vast majority of the global ocean volume, the influence of the VC method on the global OHC trend is small. For example, the upper 2000 m OHC trend with VC is ~0.15% (~0.45%) smaller than without VC from 1958-2023 (2005-2023) for IAPv4. However, it can significantly affect regional OHC estimates, especially in regions with complex topography. For example, the Maritime Continent region's 0-2000 m OHC trend is reduced by 6.9% (4.2%) after applying VC from 1958-2023 (2005-2023) (Jin et al. 2024)."

Line 449: How can you say it is a superior dataset? Compared to what? What happens if other datasets have (compensating) issues? How do you know the other datasets used in the budget are perfect?
Re: The sentence in our manuscript is: "A superior dataset should be capable of closing the sea level and the Earth's energy budgets.". This is a general statement on one metric to show the performance of the dataset, not a comment on any specific dataset. Couldn't a good/reliable dataset close the sea level and energy budgets? This is a physical-based metric.

We also did not say that the other datasets used in the budget are perfect; all estimates have uncertainty attached to them, and the sea level budget closure can be assessed in the context of uncertainty (as we did, as all other published studies such as Church et al. 2011, Frederikse et al., 2020, IPCC). Even in your paper, Domingues et al. 2008, you used the budget closure to indicate the improvement of your reconstruction, as I quote: "***The improved closure of the sea-level budget over multi-decadal periods (Fig. 3b) and the better agreement in the magnitude of observed and simulated decadal variability (Fig. 2c, d) increase confidence in the present results and represent progress since the last two IPCC reports2,19.***". Thus, we are confused about your questions on this sentence.

Line 476: Please include figure in Suppl. Material to show this point.

Re: This sentence says: "This approach provides an effective method to quantify the local trend by minimizing the impact of year-to-year variability and start/end points.". We refer to the peer-reviewed study (Cheng et al. 2022b) for more information; this is a dedicated study to introduce and evaluate this LOWESS-based approach to quantifying the local trends. We did not repeat the analysis here.

Line 496: surface area?

Re: Yes, "surface" added.

Line 508: Which of the improvements is making the most difference?

Re: As said, the most important thing is the use of "degree distance" instead of "km distance". The sentence is "The improvement in IAPv4 is mainly because of the methodology improvements: IAPv3 used 1990–2005 data to construct climatology which suffered from errors related to sparse data coverage, use of "degree distance" instead of "km distance", and other error sources.".

Line 528: Please include figure in Suppl. Material to demonstrate this point. What about the added profiles?

Re: The difference is already shown in Fig.7; the upper panels are IAPv3, where the degree distance is used, and the lower panels are IAPv4, where the km distance is used. The features are very illustrating; for example, you can see in IAPv3 that the anomalies are "as rays emerging from the pole" (new texts added in the revised manuscript), which shows the impact of using "degree distance" because if the degree distance is used, the actual km distance goes smaller when going closer to the pole. Hence, the influencing distance of each observation is more and more limited in space.

Line 540: What is the definition of subtropics and midlatitudes? Please include the latitudinal range for each.

Re: There is no clear boundary for the MLD feature changes: they are just changing gradually with space. Nevertheless, in the revised manuscript, we roughly provided the latitudinal ranges: "The MLD shows a much stronger seasonal variation in the subtropics and midlatitudes (for example, 20°~70° in both hemispheres) than in other regions (including the tropics, for example, 20°S~20°N) ".

Line 566: What do you mean by "quantitatively consistent"?

Re: This sentence has been removed.

Line 574: Sparser observations in the ocean or satellite SST?

Re: Both. Sentence modified to "The largest difference between IAPv4 and other SST products comes mainly from the Pacific and the Southern Ocean before 1980, associated with sparser in situ observations for both SST and subsurface temperature data."

Line 608-610: Please include figure in Sup. Material to demonstrate these two points.

Re: The sentence you are referring to is "The updated MBT and XBT corrections are mainly responsible for the difference between 1980 and 2000. Data QC impacts the month-to-month variation of the OHC time series.". We kindly refer you to three previous studies for this statement. The references are now added in this sentence.

XBT: Cheng et al. (2014) Fig.17: Cheng, L., Zhu, J., Cowley, R., Boyer, T., and Wijffels, S.: Time, Probe Type, and Temperature Variable Bias Corrections to Historical Expendable Bathythermograph Observations. J. Atmos. Ocean. Technol., 31, 1793-1825,_ https://doi.org/10.1175/jtech-d-13-00197.1, 2014.

MBT: Gouretski and Cheng (2020), Fig.10: Gouretski, V., and Cheng, L.: Correction for Systematic Errors in the Global Dataset of Temperature Profiles from Mechanical Bathythermographs. J. Atmos. Ocean. Technol., 37, 841-855, https://doi.org/10.1175/jtech-d-19-0205.1, 2020.

QC: Tan et al. (2023), Fig.15: Tan, Z., Cheng, L., Gouretski, V., Zhang, B., Wang, Y., Li, F., Liu, Z., & Zhu, J.: A new automatic quality control system for ocean profile observations and impact on ocean warming estimate. Deep Sea Research Part I: Oceanographic Research Papers, 194, 103961, https://doi.org/10.1016/j.dsr.2022.103961, 2023.

Lines 620-629: How does the gridded data subsampled at the locations (x,y,z,t) of the actual profiles compare? Is there any significant difference between them? Does the gridded product also use Deep Argo floats? Do the wide known significant CMIP model drifts (particularly in the deep ocean) were removed before the calculation of the co-variances?

Re: Yes, the gridded product used Deep Argo floats in combined with other available data.

The drift is not removed before calculation because drift removal at a global scale only impacts the global T/OHC and does not impact spatial covariance. If one removes the drift at each grid box with local linear or quadratic regression, it generates problems for the model representation of covariance because the resultant fields are not dynamically consistent any more (as the errors in drift removal at local scales are not perfect).

For the comparison of gridded data with the in situ profiles, here we provide an example of the difference between the reconstruction and the observations at 3100m (Fig. R3, upper) compared with the temperature trend at 3100m from reconstructed data (Fig. R3, middle). It seems that the difference pattern is significantly different from the trend pattern. In general, there are no substantial systematical biases in the reconstruction (Fig. R3, lower). For instance, the major warming appears in the Southern Ocean with trends varying from 0.01~0.05 ^{o}C dec^{-1} from 1991 to 2023, so the total temperature changes are 0.33~1.65 ^{o}C. The reconstruction and observations differences are around zero. Using other dept levels yields similar results.

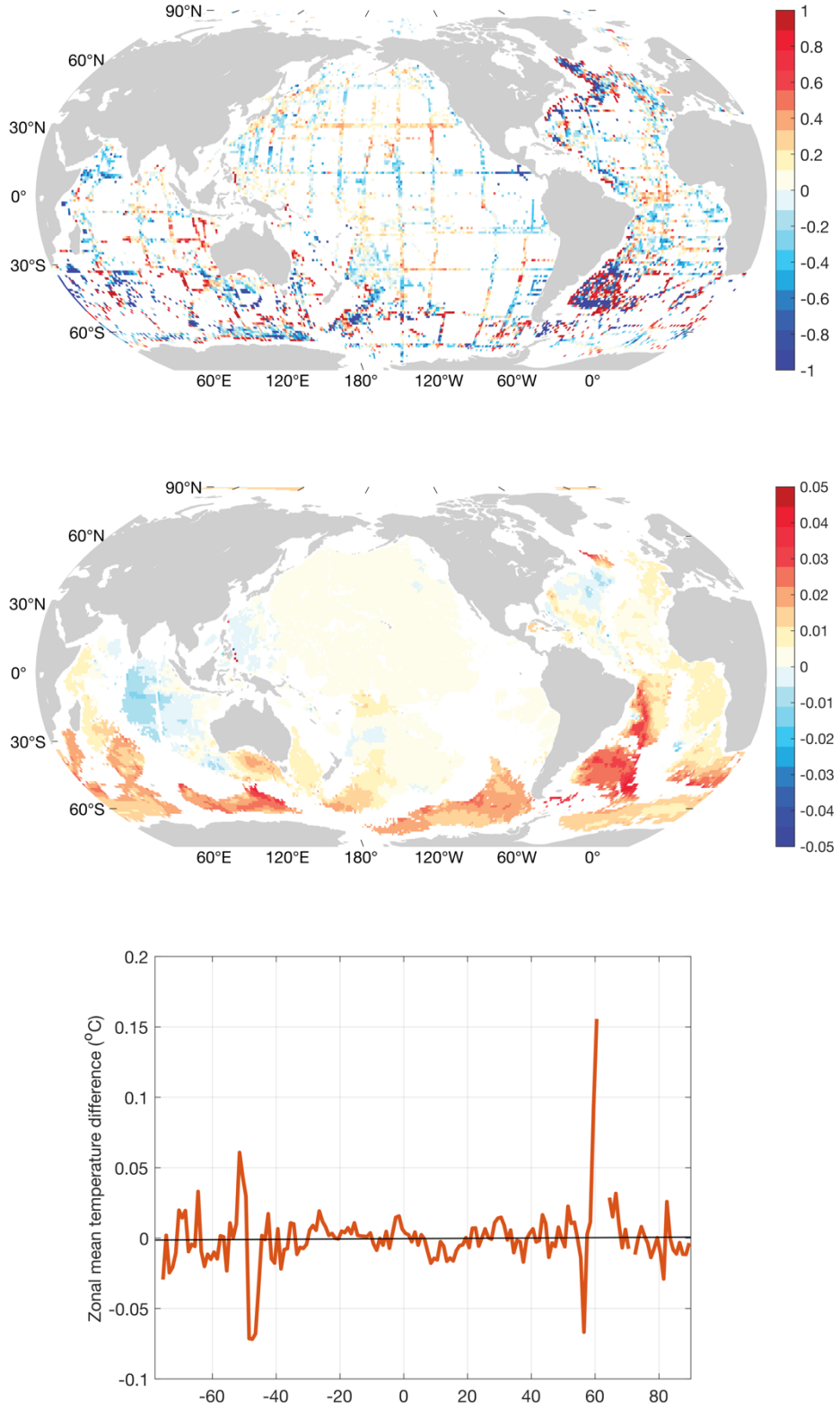

Figure R3. (upper) difference between the reconstruction and the in situ temperature observations at 3100m (upper), the unit is degree Ceilsus. (middle) the temperature trend at 3100m from reconstructed data in IAPv4, the unit of $^{o}C\ dec^{-1}$. The data are for 1991-2023 period.

(lower) the zonal mean difference between the reconstruction and the in situ temperature observations at 3100m from 1991-2023.

Figure 11, panel a: There is a large interannual variability around 2000-2005, just before the Argo array achieve its global float target. Could this unusual step change (compared to the other variability observed for the entire record) arise from the radical change in the observing system? Should a cautionary note be included in the text?
Re: There is no evidence that the variation between 2000-2005 is "unusual". Based on the subsample test by Cheng et al. (2017), the current mapping approach does not lead to significant biases associated with sampling changes, including the 2000-2005 period.

Line 640: Was this reported in a paper before Trenberth et al. 2016? Please cite reference if exists. Please consult with Loeb/Sato.
Re: We don't know a dedicated paper reporting this before Trenberth et al. 2016, please let us know if you know any studies.

Line 649: Did you apply the same smoother to the other timeseries? Is the comparison fair?
Re: Yes, all time series are smoothed in the same way, as stated in the table caption "12-month running mean (13-points are used, with start-point and endpoint 656 weighted by 0.5)". The same approach is also used in Trenberth et al. 2016. One can do a cross-check.

Line 666: How does your SST data compare with satellite SST along boundary current regions? Does it look realistic? Or is it too warm because the QC is not flagging data errors? Missing proper evaluation.
Re: The SST trend pattern has been compared with other data in Fig.10; IAPv4 is consistent with other data, and the pattern is definitely realistic, please check section 3.3 texts.
This section is for OHC, and we compare the OHC trend map between IAPv4 and IAPv3; this sentence is a state of fact shown in Figs. 12, 13 that "The IAPv4 shows stronger warming near the boundary currents regions". For the explanatory sentence of "mainly because of the improved QC that does not flag high-temperature anomalies.", this has been tested and published in Tan et al. (2023). But to better show you this issue, here we provide an example to show the QC's impact (Fig. R1, below). You can see that the impact of QC on OHC is mainly in the eddy-rich regions (Fig. R1 upper panel), and the WOD-QC that is used in IAPv3 is removing more positive anomalies than CODC-QC (Fig. R1 lower panel). A careful check of profiles indicates that most of the removed positive anomalies look realistic. Moreover, altimetry data show that these positive anomalies are located within warm eddies and are associated with thermocline changes (so the temperature anomalies are big at the sea subsurface, around the thermocline).

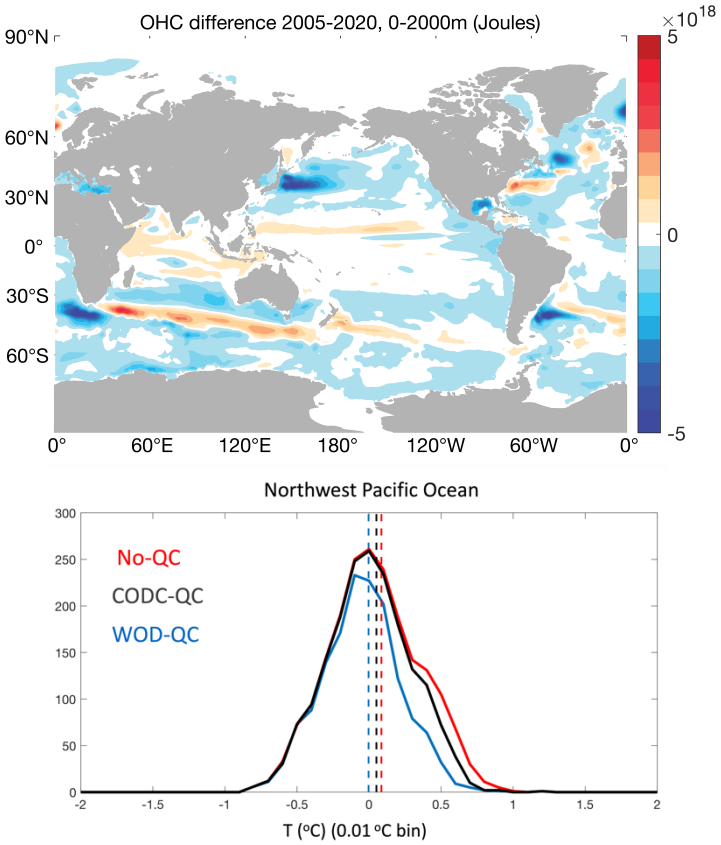

Fig. R1. (upper) Mean Upper 2000m OHC difference between 2005-2020 between the reconstruction results based on WOD-QC (used in IAPv3) and CODC-QC (used in IAPv4). In the two reconstructions, the other data processing procedures are the same. (lower) The distribution of 500 m temperature anomalies in each 0.01°C bin in the Northwest Pacific Ocean for observations without any QC (red), with WOD-QC (blue) and CODC-QC (black).

Figure 12: Where is there statistical difference between IAPv3 and v4?

Re: We decided not to provide any statistical check on the differences for IAPv4 minus IAPv3, because of the following reasons: 1) it is nearly impossible to have a correct statistical check, because the statistical significance check of the different field relies on the covariance of the two products, because they are not independent. This estimate is nearly impossible; 2) even if we can identify regions with insignificant differences between IAPv4 and IAPv3, they might be important because OHC is an integral quantity, summing up the small numbers might yield a big/non-negligible difference for OHC. In this regard, providing significance for regional differences is misleading and gives readers the wrong impression that the insignificant regions are not important.

For these reasons, we feel that giving the trend maps of IAPv3 and IAPv4 and then showing their differences are sufficient just to illustrate the impact on the trend and trend patterns. The users can then assess if the differences are important or not based on their needs.

Line 686: Why does deep ocean warming occur after 1990 and not before? What is the physical explanation and evidence? Do we have enough deep ocean observations before 1990s?

Re: We here refer to the peer-reviewed paper Cheng et al. 2017 again that using a subsample test indicates that the deep ocean changes within 700-2000 m can be reliably reconstructed since 1960. The uncertainty is relatively large though, see Fig.11 error range.
It is beyond the scope of this study to explain why deep ocean warming occurred after 1990 and not before. This study aims to describe the data/methodologies/results.

Line 712-713: Is the interannual variability statistically significant? Where are the error envelope for the other datasets?
Re: There is substantial/significant inter-annual variability in the derived MHT time series, which are also physically meaningful. We refer to Trenberth et al. (2016, 2019a, 2019b) for more details on the analyses of MHT variability. Also, similar approaches are used to derive MHT in other groups, such as Mayer et al. (2022) and Liu et al. (2020); there are dedicated studies on the inter-annual variability of MHT with similar approach.
For the other datasets, the error range cannot be given here because the uncertainty estimates require the complete estimates of local OHC uncertainty and the spatial error covariances, which are not available for other datasets. This does not impact the conclusion of this study because we are here to compare how better different estimates based on different OHC data fall into the RAPID envelope and their correlation with the RAPID time series.

Trenberth, K. E., Y. Zhang, J. T. Fasullo, and L. Cheng, 2019a: Observation-Based Estimates of Global and Basin Ocean Meridional Heat Transport Time Series. J. Climate, 32, 4567–4583, https://doi.org/10.1175/JCLI-D-18-0872.1.

Trenberth, K. E., and J. Fasullo, 2017: Atlantic meridional heat transports computed from balancing Earth's energy locally. Geophys. Res. Lett., 44, 1919–1927, https://doi.org/10.1002/2016GL072475.

Trenberth, Kevin E. and Zhang, Yongxin, 2019b, Observed Interhemispheric Meridional Heat Transports and the Role of the Indonesian Throughflow in the Pacific Ocean, Journal of Climate Vol. 32, No. 24, pp 8523, 1520-0442

Mayer, Johannes, Mayer, Michael, Haimberger, Leopold, and Liu, Chunlei, 2022, Comparison of Surface Energy Fluxes from Global to Local Scale. Journal of Climate Vol. 35, No. 14, pp 4551, 1520-0442

Liu, Chunlei, Allan, Richard P., Mayer, Michael, Hyder, Patrick, Desbruyères, Damien, Cheng, Lijing, Xu, Jianjun, Xu, Feng, and Zhang, Yu, 2020, "Variability in the global energy budget and transports 1985–2017" Climate Dynamics Vol. 55, No. 11-12, pp 3381, 1432-0894

Figure 16: Please add uncertainty timeseries to demonstrate how an improved ocean observing system is making a difference in reducing uncertainty.
Re: Thanks. Good point. We added the uncertainty range (95% CI) in the shading of our energy budget estimate.

Figure 17: Please include other estimates (IAPv3, EN4, ISH, Johnson et al) for comparison as, for instance, done in Figure 14.

Re: A comparison of different products is currently under development in the other papers in a more comprehensive way (GEWEX GDAP group study led by Dr. Maria Hakuba) with a standardised approach to derive dOHC/dt, common mask and other issues, so we don't want to replicate the efforts add more details here. This paper mainly aims to describe IAPv4 compared with IAPv3.

Line 837 and Table 3: Incorrect AR6-related statements and values. AR6 trends were not based on least-squares fit nor on Frederikse et al. 2020. "Based on the ensemble approach of Palmer et al. (2021) and an updated WCRP Global Sea Level Budget Group (2018) assessment (Figure 2.28) GMSL rose at a rate of 1.32 [0.58 to 2.06] mm yr^{-1} for the period 1901–1971, increasing to 1.87 [0.82 to 2.92] mm yr^{-1}between 1971 and 2006, and further increasing to 3.69 [3.21 to 4.17] mm yr^{-1} for 2006–2018 (*high confidence*). The average rate for 1901–2018 was 1.73 [1.28 to 2.17] mm yr^{-1} with a total rise of 0.20 [0.15 to 0.25] m (Table 9.5)."

Table 9.5 | Observed contributions to global mean sea level (GMSL) change for five different periods. Values are expressed as the total change (Δ) in the annual mean or year mid-point value over each period (mm) along with the equivalent rate (mm yr–1). The very likely ranges appear in brackets based on the various section assessments as indicated. Uncertainties for the sum of contributions are added in quadrature, assuming independence. Percentages are based on central estimate contributions compared to the central estimate of the sum of contributions.

Re: Thanks for this information. Table 3 has been removed.

In the revised manuscript, what we were doing is simpler: download AR6 data and replace the thermosteric sea level time series with IAPv4 thermosteric sea level to check the impact.

Please check the report and/or with authors. Chapters 2 and 9.

https://iopscience.iop.org/article/10.1088/1748-9326/abdaec/meta

Re: Thanks for this information. Please check the supplementary material of Cheng et al. 2022 NREE review paper for our comments on the caveats of this uncertainty estimate approach.

Table 3: Does IAPv4 have GMSL estimates as this table implies?

Re: Table 3 has been removed.

Line 850: If salinity change is irrelevant for global mean sea level, should it be wise to compute thermosteric rather than steric sea level for budget purposes? As salinity data tend to be less reliable than temperature data?

Re: In the revised manuscript, to compare with IPCC-AR6 results, only thermosteric sea level is considered.

Lines 901-903: What is the relative importance of each factor? Which factor(s) is(are) making the most difference?

Re: This is a good topic for a separate paper (actually we are already working on that) to understand the relative importance of each factor. It is not a simple answer because the different factors are not independent of each other: changing one technique can impact the contribution of the other factors. Thus, this deserves a dedicated study.

Lines 908-919: Please include this in a caveat section, along with other caveats (e.g. use of model covariance and applications not recommended).
Re: See our reply to your first major comment.

Line 922: Does any of the products have enough spatio-temporal resolution to resolve mesoscale variability? Is this being aliased or properly accounted for as error?
Re: None of the 1-degree resolution data products can resolve meso-scale variability (with typical spatial scales of 20~300 km). The key point here is to properly represent the statistical feature/characteristics of the mesoscale eddies (i.e. the averaged temperature changes). The errors related to this issue are "representative error" as quantified and represented in IAPv4 mapping approach (see Method section for details).

Line 928-936: IQuOD is doing much more than just uncertainty. Please represent the comprehensive IQuOD activities properly, and how that might support other activities, such as yours, reanalyses, etc.
Re: Yes, we agree that IQuOD plans to do more, here we just want to describe the most relevant activities, which are already quite broad "For example, the International Quality-Controlled Database (IQuOD) group is coordinating some activities related to data processing techniques, uncertainty quantification, and improving the overall quality of ocean data (Cowley et al., 2021).". We believe it is not a place to completely introduce IQuOD, even though I'm the current co-chair of IQuOD. Please note, similar to your comments on ME4OH, a very extensive discussion on the "wish list" of these projects is dangerous, instead, what has actually been done and the suggestions grounded on the present study is more scientifically meaningful.

Line 937-944: Please cite Boyer et al. 2016, Savita et al. 2022, and IAPSO's ME4OH best practice working group:
https://iapso-ocean.org/images/stories/_working_groups/Best_practice_study_groups/mapeval4oceanheat__2021-proposal.pdf
Re: I'm not sure it is proper to attribute these recommendations only to Boyer et al. 2016, Savita et al. 2022, and IAPSO's ME4OH. First, these comments in the present study are more specific than those presented in Boyer et al. 2016, Savita et al. 2022, for example, we suggest "In the future, comprehensive quantification of other uncertainty sources will be made, including the choice of climatology, vertical interpolation, XBT/MBT/APB/Bottle corrections, etc. It is also necessary to analyze the correlation between these error sources.". This is very specific and has not been fully discussed before. Second, there are more studies recommending various aspects of issues related to the OHC estimates; only attributing these points to your mentioned

references is not proper. Third, our recommendations are grounded on the results of the present study.

Boyer et al. 2016 and Savita et al. 2022 are cited in our manuscript in the proper places. We did not cite the ME4OH document because it is not peer-reviewed literature, and no publication from the ME4OH group has been available until now (to our best knowledge).