Reply to Anonymous Reviewer #1

Dear Reviewer,

Thank you very much for your detailed and thorough review of the submitted manuscript. Please, find below our response to the issues raised in your review. All pieces of the text changed according to your comments are written in blue.

1. Reconcile internal numerical inconsistencies in the OHC results. The OHC increase between the 1920–1944 and 1945–1969 periods is given as 149 ZJ in Section 8.5 (line 424) but as 139 ZJ in the Conclusions (line 484). Similarly, the total warming since 1932 is 456 ZJ at line 424 but 457 ZJ at line 444. These are presumably the corrected vs. uncorrected values or simple typographical slips, but as written the paper contradicts itself on its headline numbers. Please reconcile these figures throughout the Abstract, Section 8.5, and Section 9, and state explicitly whether quoted values are for original or depth-corrected profiles.

We have reconciled the internal consistency of the OHC results. It is now stated where quoted values are for original or for the depth-corrected profiles. Please note that we recalculated all OHC estimates and error bars using a specific heat capacity of 3991.87 J/kg °C for seawater, as the original estimates used the value for freshwater.

2.Correct the standard-depth list in Section 8.5 (lines 436–438). The list of 67 standard levels contains apparent duplications and omissions: "30" appears twice ("15, 30, 25, 30, 35…"), "2000" appears twice ("2100, 2000, 2300"-presumably 2200 was intended), and "3200" appears twice (presumably 3100). As listed, the levels do not sum to 67 unique depths, and the claim that they represent 1.2% of all 1-m levels cannot be verified. Please correct the list and re-verify the 49.5% statistic if affected.

The standard depth list was corrected (duplicate levels deleted) and moved into the supporting information section.

3.Resolve citation-year and author-name mismatches between text and reference list. (a) Reiniger and Ross is cited as 1966 throughout the text (lines 20, 163, 471) but dated 1968 in the reference list; (b) Chapman is cited as 1955 (line 100) but listed as 1959; (c) Zanna et al. is cited as 2018 throughout but listed as 2019; (d) "Spillhouse, 1938" (line 50) should be Spilhaus, consistent

with the reference list; (e) “Gourteski” at lines 30 and 85 should read Gouretski. These are easily fixed but essential for a data paper that readers will mine for sources.

Thanks for the comments. We have inserted respective corrections in the reference list and in the text.

4.Make the Challenger expedition dates consistent. The expedition is variously dated 1872–1876 (lines 35, 74), 1873–1875 (line 320), and 1873–1876 (lines 413, 578). The voyage departed in December 1872 and returned in May 1876; please adopt one convention (or clarify that the dates at line 413 refer to the analysis period rather than the voyage) and apply it throughout.

The Challenger expedition dates were made consistent using the period 1873-1876. This agrees with the period given by Thomson et al. (1885).

5. Check Figure 15 legend against the stated analysis periods. The legends in Fig. 15 appear to read “2010–2017” for the most recent period, whereas the text and Section 8.1 define the period as 2010–2025. If the figure was carried over from an earlier analysis, please regenerate the legend; if the plotted period genuinely differs, the discrepancy must be explained. Relatedly, in Section 8.4 the error-estimation procedure is clearly described, but the resulting uncertainty magnitudes are never stated numerically in the text—adding one or two sentences quantifying the basin- and global-scale standard errors (or pointing to the error bars in Figs. 15–16 with representative values) would materially strengthen the section at negligible cost.

Fig. 15 was corrected. We rewrote Section 8.4, also taking into account the second reviewer's comments. We added characteristic temperature uncertainties in Section 8.4, as you suggested.

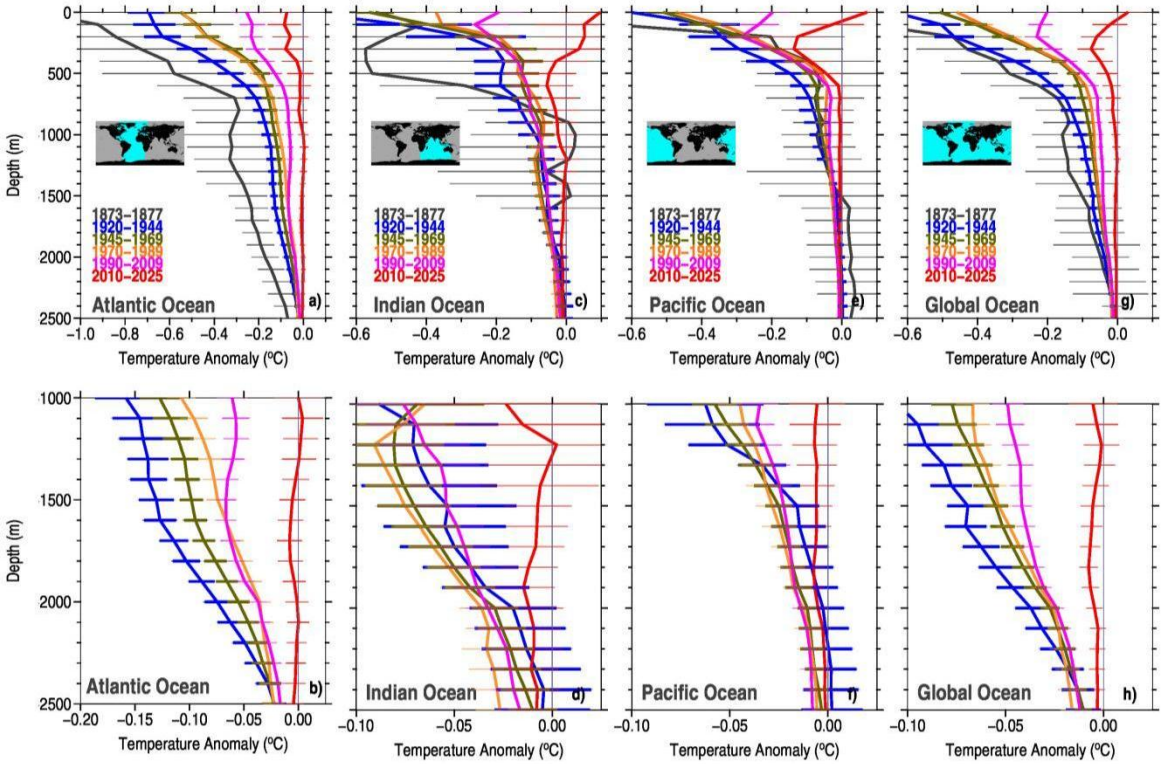

Figure 15. a, c, e, g) Temperature anomaly vertical profiles for the three main oceans and for the entire global ocean and for selected six time periods with standard error bars; b, d, f, h) same but for the depth range below 1000 m. Note the different temperature scale for the Atlantic Ocean.(revised)

"8.4 Error estimation

The box-averaged temperature anomalies in each 100 m layer are calculated using all observed anomaly profiles within the respective 5º × 5º latitude/longitude box. Instrumental errors and mesoscale variability introduce noise, which biases the estimated box-averaged anomalies. Boyce (1966) estimates the instrumental error for reversing thermometers as ± 0.04 ºC. Errors associated with the mesoscale variability are much larger. According to **Fig. S7**, the standard deviation of temperature anomalies within a 2° × 4° latitude/longitude box exceeds 3 ºC in the western boundary currents and in the equatorial belt in the upper 200-meter layer. We assume temperature anomalies within a 5 ° × 5 ° latitude/longitude box are independent because the observations span 20 years or longer, and each mean box anomaly is associated with its respective standard error. For the basin- and global mean anomalies, the errors are obtained by the summation in quadrature of within-box errors. In order to estimate the number of degrees of freedom, we calculated spatial autocovariances in east-west and north-south directions using box-averaged anomalies for the best-sampled period 1990-2009 (**Fig. S10**). The meridional correlation radius is determined by the zero-crossing and is taken to be 1 meridional distance (555 km) between the boxes. The correlation along longitude has a long tail, so we define the correlation radius as the distance at which the correlation function falls below 0.1, which averages two box separations in the zonal direction. We

note that for the earliest periods (1873-1877 and 1920-1944) global and basin means are estimated from observations (box-averaged values) that cover only a small fraction of the respective area (see Fig. 14), leading to a mapping error. This kind of error was not estimated in the current study and is not considered in **Figs. 15-16**."

6. Clarify data access in Section 10. The dataset DOI requires email registration and login. ESSD policy expects data to be publicly and freely accessible without barriers wherever possible. Please either remove the registration requirement, or briefly justify it and confirm that access is granted automatically and without restriction, so the editor can verify compliance with the journal's data policy.

We have also made the data available on the Zenodo platform (https://zenodo.org/records/22764102), in addition to http://doi.org/10.12157/IOCAS.20260520.001, which requires an extra login to access the data. The Data Availability section in the manuscript has been modified accordingly.

7. The manuscript requires a careful language edit. The science is clear, but grammatical errors and typographical slips occur throughout and occasionally rise to the level of technical error. The most important instance: "Jung's modulus" at line 459 should be Young's modulus (and note the same sentence refers to "hemp wire" where "hemp rope" is meant, and "four time higher" should read "four times higher"). A non-exhaustive list of additional corrections follows; I would encourage the authors to have the full manuscript professionally proofread rather than relying solely on this list.

- Line 11–12: "observations which abundance, quality and accuracy changed" → "observations whose abundance, quality, and accuracy have changed."
  Done.
- Line 35 and line 483: "marks the begin of" / "the begin of the ocean warming" → "beginning."
  Done.
- Line 45: "Fritjoff Nansen" → "Fridtjof Nansen."
  Done.
- Line 49: "hydrographic hydrographic profiles" — duplicated word.
  Done.

- Line 62: “we access for the first time the change” → “we assess.”

  Done.

- Line 126: “An disproportionately large number” → “A disproportionately large number.”

  Lines 137, 421, 472: “undergone” used as a simple past (“the data … undergone three steps”; “the global ocean undergone”; “profiles undergone”) → “underwent” (or “have undergone”).

  Done.

- Line 197: “performed for a the resampled profile” — delete “a.”

  Done.

- Line 260: “Here we imply this method” → “apply.”

  Done.

- Line 292: “In a deviation to the widely used World Ocean Atlas climatologies” → “In contrast to.”

  Done.

- Line 309: “mimimum/maximum” → “minimum/maximum.”

  Done.

- Line 415: “with wights equal to” → “weights.”

  Done.

- Line 417: “There several significant differences” → “There are several…”

  Done.

- Reference list: “littele ice age” (Gebbie and Huybers), “Warming of heworld ocean” (Levitus et al.), “Pysics and Chemistry” (Thomson and Murray 1889), “Washigton D.C.” (U.S. Naval Oceanographic Office), “Manual fort hydrological works” (State Oceanographic Institute), and “Her Majesty Stationary Office” → “Stationery Office.” The Frankignoul (1981) entry should also be reformatted to match the journal’s reference style (year at the end).

  Done.

We conducted a careful language edit and made corrections according to your detailed list above – thank you so much for your careful reading!

Please note that we substantially reworked Section 4.3, because the old comparison of RR and FR interpolation methods was not adequate. In the old comparison method, the interpolation by the two

methods was applied to the same climatological (WAGHC) temperature profiles, so that the RR method already "knows" the shape of the climatological profile. For this reason, we used a different approach to compare the RR and FR methods. Now we use the original observed level quality-controlled profiles, which are interpolated vertically using both methods. Before interpolation, artificial data gaps are created by removing data levels. The error is defined as the difference between the original (omitted) temperature and the interpolated value. The subsequent comparison of the interpolated values with the observed (artificially removed) values shows that 1) the FR method has a smaller global root-mean-square interpolation error and 2) the RR interpolation introduces a positive temperature bias (global mean temperature error). For this reason, only the FR method is used. We updated Fig. 9 and rewrote Section 4.3.

The new text of Section 4.3 is (Lines 225-259 in the revised manuscript):

"As shown in **Fig. 7**, the global root-mean-square interpolation error in the RR method increases with gap size. To reduce the error for large vertical gaps, the climate anomaly method suggested by Frankignoul (1981) is used (FR method). In this method, deviations of individual temperature profiles from the mean climatology are linearly interpolated over depth, and the interpolated value at a level is the sum of the mean climatological temperature and the interpolated anomaly. **Fig. 8** shows several selected observed profiles with big vertical gaps along with the respective interpolated profiles. For large gaps, temperature differences between the two methods can exceed several degrees. For large gaps, the RR method tends to follow simple linear interpolation across the gap, resulting in a significant deviation from the climatological profile.

To further illustrate the differences between the two methods, we used the complete validated observed level profile dataset. To estimate interpolation errors, artificial data gaps are created. Error estimation was performed for two choices of the vertical gap size. In a 'small gap' case, for each temperature profile with N observed levels, a total of N-2 reduced level profiles can be created, with one omitted observed level k between the uppermost (k=1) and the lowest level (k=N) of the full original profile. In a 'big gap' case, in addition to level k, levels k-1 and k+1 are omitted as well. Vertical interpolation of reduced profiles is performed at the depth of the removed level k, and the difference between the observed and interpolated temperature estimates the interpolation error. The vertical distance between the levels k+1 and k-1 ('small gap' case) and k+2 and k-2 ('big gap' case) gives the size of the artificial vertical gap. We performed interpolation using the RR and FR methods, respectively. Finally, we calculated the overall RMS (**Fig. 9a**) and mean (**Fig. 9b**) interpolation errors for the WAGHC standard depth layers. **Fig. 9c** shows the mean vertical gap size for the two cases. Our calculations demonstrate that the FR method results in a smaller interpolation error compared to the RR method. The layer-mean errors (**Fig. 9b**) show that the RR method

produces a systematic positive bias, which is especially strong in the 500-1000 m layer, with maximum values up to 0.05/0.2 °C for the small- and big-gap cases, respectively. In contrast, for the same layer, the FR method shows a maximum negative bias of only - 0.01/- 0.04 °C for the small- and big-gap cases, respectively. The spatial distribution of the mean interpolation error is shown in **Fig. S4**. For all three selected layers, the FR interpolation method results in smaller interpolation errors compared to the RR method. The difference between the two methods is more pronounced above and within the main thermocline (layers 182-218 m and 798-878 m, respectively). For the layer 182-218 m, the largest errors are found within the strongly stratified equatorial belt (**Fig. S4a, b**) for both the FR and RR methods. Within the main thermocline (798-878 m), the RR method shows the largest positive errors, up to 0.4 °C, across the entire tropical and subtropical belt.

The overall performance of the interpolation method described in the above sections is further illustrated by the temperature/depth histograms based on the observed and interpolated data, respectively (**Fig. 10**). To obtain these histograms, the number for each temperature-depth bin is normalized by the number of observations (observed or interpolated, respectively) in the most data-abundant temperature bin for each depth. A good agreement between the two histograms suggests the absence of spurious features in interpolated profiles."

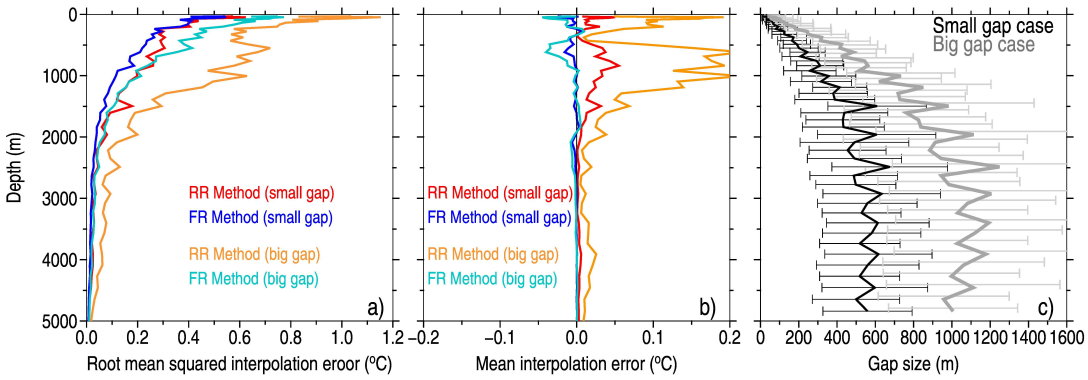

Figure 9. Global root mean squared (a) and global mean (b) error for interpolated temperature values obtained by means of RR method (red and orange) and FR method (blue and cyan). Red and blue profiles correspond to the small gap case, orange and cyan correspond to the big gap case. Explanation Is given in the text. Mean gap size versus depth for the small (black) and big (grey) gap case respectively. Horizontal bars correspond to one standard deviation. (revised)

Reply to the Anonymous Reviewer #2

Dear Reviewer,

Thank you very much for your detailed and thoughtful review of the submitted manuscript. Please find below our response to the issues raised in your review. All pieces of the text changed according to your comments are written in blue.

* SPECIFIC COMMENTS *

* Sec 3.1: This section corresponds to Figs 5a-c. However, the text only discusses Fig 5a, not the other two. Are Figs 5b and c not relevant here? If so, consider removing them to simplify the figure.

The reference was changed to the Fig.5a-c in Line 142 in the revised manuscript.

* Sec 3.2: It is unclear how the manual quality control is done. What are the criteria for determining if a profile is an outlier? Please explain the procedure and cite relevant references.

We rewrote and expanded the section on manual quality control to explain how it was done. Also, an additional reference to Swift (2010) was added (Lines 147-160 in the revised manuscript):

"**3.2 Manual quality control**

Secondly, temperature profiles were subjected to manual quality control. The data have been screened for remaining outliers in each 5° × 20° Latitude/Longitude box (**Fig. S1**). The manual control was based on the visual inspection of temperature profiles and temperature anomaly profiles within each box. The visual outlier check is focused mostly on the deep parts of profiles (typically below 500-1000 m) where seasonal changes attenuate, and the temperature/depth relation is tight. Outlier diagnostics for the manual quality control are shown in **Fig. 5d-f.** Outliers are identified visually using observations that passed the automated quality control as a background reference. A higher percentage of manually identified outliers is clearly seen below 1000-1500 m, that is, in the depth range below the main thermocline. Examples of manually rejected observations are shown in **Fig. S2** for six randomly selected boxes. Temperature profiles for observed level data that passed both the automated and manual quality control are shown in **Fig. S3** for selected boxes. Overall, the manual quality control identified 0.17 % of all temperature observations as outliers, with 0.85 % of the profiles affected by the manual check. We note that, despite improved accuracy and reliability of modern oceanographic instrumentation, manual (expert) control and data evaluation remain necessary steps in the preliminary analysis of hydrographic data (Swift, 2010)."

* Sec 4.1

Please add a short description of the RR and FR methods in the preamble of section 4. In addition, please explain why the two methods are chosen here instead of a simple linear interpolation. The authors show that the polynomial interpolation (the RR method) causes unrealistic excursions in temperature profiles in Fig 6 (lines 182-186). For these cases, it seems that a simple linear interpolation would yield a similar result, but without the unrealistic excursions. Is this correct?

The beginning of Section 4 was rewritten and provides a description of both interpolation methods. We also included the reference to Li et al. (2022), where they show that a warm bias is introduced when linear interpolation is used (Lines 183-197 in the revised manuscript):

"**4.1 Case of uneven distribution of observed levels over depth**

The straightforward application of the RR method to observed level profiles was found to result in unrealistically large excursions on the interpolated temperature profile for a not-negligible number of profiles. Such cases typically occur when the observed levels are distributed so unevenly that adjacent vertical gaps differ by more than an order of magnitude. Such profiles occur, for instance, when the entire temperature profile at a hydrographic station is composed of two or more hydrographic casts, each covering a part of the entire water column. When such profiles are merged together in the final station table, the deepest observed depth of the shallower cast might be very close to the shallowest observed level of the deeper cast. In other cases, small gaps may result from including interpolated levels along with observed levels in the resulting station table, or from large gaps in profiles caused by outlier rejection. Since the interpolating polynomial must pass through each point of the original temperature profile, unrealistic overshoots appear on the interpolated profile. **Fig. 6** shows several examples of such profiles. In some cases, the overshooting amounts to several degrees. To avoid the overshooting, the original profiles are checked for pairs of levels which are very close to each other before interpolation. Such pairs are then replaced by one level referred to as the middle point, and temperature is taken as the average of the temperatures at the two levels. Another option is to use a simple linear interpolation in such cases."

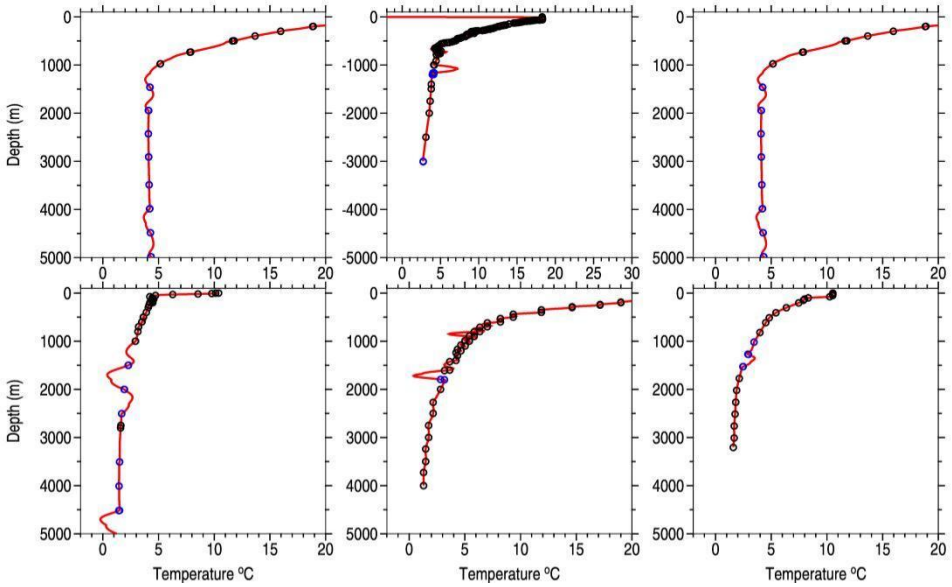

Figure 6. Examples of temperature profiles with uneven distribution of observed levels resulting in unrealistic overshooting of interpolation polynomials in RR method: observed levels (black circles), observed levels with a small distance to a neighbor level (blue circles), interpolated profile (red).

* Sec 4.2

Lines 202-203 "centre of the gap": But in Fig 7, the x-axis label is "depth of the upper gap boundary", not the depth of the centre of the gap.

Thank you for noting that! We have made the figure and the text consistent. The center of the depth is meant.

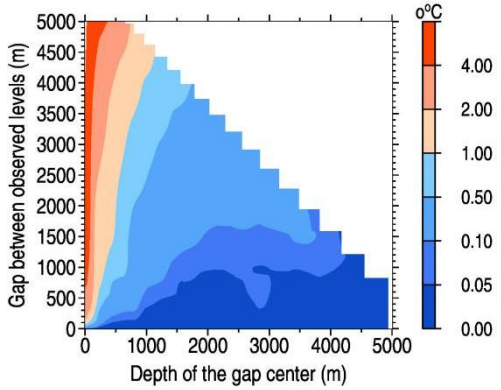

Figure 7. Global root mean squared interpolation error versus depth and the vertical gap size. Error refers to the center of the gap. (revised)

Is the interpolation error a function of geographical locations? Ocean temperature profiles have different characteristics in different regions, e.g. weak stratification at high latitudes vs strong

stratification in the tropics. Hence, I expect that you may get different interpolation errors from the WAGHC profiles at different locations.

Yes, your comment is quite true: the error varies geographically, and Fig. 7 provides only an overall guideline. This is now explicitly noted in the revised text (Lines 215-217): "Fig. 7 serves as a guideline and characterizes the global mean interpolation error. Regional interpolation errors may be different due to variations in the vertical temperature stratification over the global ocean.".

* Sec 4.3 and Fig 9

Several key details are missing from this section.

What is the definition of the mean error and the mean absolute error? What is the dimension on which the "mean" operation is applied?

Now we provide the details in the revised text. We provided the mean error and the root mean squared error; the mean absolute error is no longer used in the revised manuscript. The root mean squared error is calculated over all WAGHC grid nodes between 70° S and 70° N, which is described in Section 4.2 (Lines 213-214 in the revised manuscript).

Similarly, in the revised section 4.3, where we compare the FR and RR methods, we provide the overall root mean squared and mean interpolation errors based on all profiles.

Why is the interpolation error always positive in Fig 7, but has both positive and negative values in Fig 9?

In Fig 9, which panels are the mean error and which are the mean absolute error? (I can't find this information in the caption)

In Fig 9, does the interpolation error includes errors for both climatology and anomaly? If so, it is no surprise that the FR method has a smaller error. This is because: a) the climatology signal is much larger than the anomaly signal and b) the FR method already knows what the climatology signal is. However, we are more interested in temperature change than the climatology. Given this, it may be better to focus on the anomaly signal when evaluating the interpolation error.

It would be more consistent to use the same type of diagram for sections 4.2 and 4.3, since they both examine the interpolation error. Fig 9 shows the error in different depth ranges using different panels. Could those panels of different depth ranges be put into a single panel like in Fig 7?

Thank you for highlighting the fact that the comparison of FR and RR methods as conducted in the original version of the manuscript and illustrated by the old Fig. 9 is not adequate. In the old comparison method the interpolation by the two methods was applied to the same climatological (WAGHC) temperature profiles, so that the RR method already "knows" the shape of the climatological profile. For this reason, we selected a different approach in comparing the RR and the FR methods. Now we use the original observed level quality-controlled profiles which are interpolated vertically using the both methods. Before interpolation artificial data gaps are created by removing data levels. The error is defined as the difference between the original (omitted) temperature and the interpolated value. The subsequent comparison of the interpolated values with the observed (artificially removed) values shows that 1) the FR method has a smaller global root-mean square interpolation error and (2) that the RR interpolation Introduces a positive temperature bias (global mean temperature error). For this reason, only FR method is used. Fig.9 was updated respectively and Section 4.3 was re-written.

The new text of the Section 4.3 is (Lines 221-255 in the revised manuscript):

"As shown in Fig. 7, the global root-mean-square interpolation error in the RR method increases with gap size. To reduce the error for large vertical gaps, the climate anomaly method suggested by Frankignoul (1981) is used (FR method). In this method, deviations of individual temperature profiles from the mean climatology are linearly interpolated over depth, and the interpolated value at a level is the sum of the mean climatological temperature and the interpolated anomaly. **Fig. 8** shows several selected observed profiles with big vertical gaps along with the respective interpolated profiles. For large gaps, temperature differences between the two methods can exceed several degrees. For large gaps, the RR method tends to follow simple linear interpolation across the gap, resulting in a significant deviation from the climatological profile.

To further illustrate the differences between the two methods, we used the complete validated observed level profile dataset. To estimate interpolation errors, artificial data gaps are created. Error estimation was performed for two choices of the vertical gap size. In a 'small gap' case, for each temperature profile with N observed levels, a total of N-2 reduced level profiles can be created, with one omitted observed level k between the uppermost (k=1) and the lowest level (k=N) of the full original profile. In a 'big gap' case, in addition to level k, levels $k-1$ and $k+1$ are omitted as well. Vertical interpolation of reduced profiles is performed at the depth of the removed level k, and the difference between the observed and interpolated temperature estimates the interpolation error. The vertical distance between the levels k+1 and k-1 ('small gap' case) and k+2 and k-2 ('big gap' case) gives the size of the artificial vertical gap. We performed interpolation using the RR and FR methods, respectively. Finally, we calculated the overall RMS (Fig. 9a) and mean (Fig. 9b)

interpolation errors for the WAGHC standard depth layers. **Fig. 9c** shows the mean vertical gap size for the two cases. Our calculations demonstrate that the FR method results in a smaller interpolation error compared to the RR method. The layer-mean errors (Fig. 9b) show that the RR method produces a systematic positive bias, which is especially strong in the 500-1000 m layer, with maximum values up to 0.05/0.2 °C for the small- and big-gap cases, respectively. In contrast, for the same layer, the FR method shows a maximum negative bias of only -0.01/- 0.04 °C for the small- and big-gap cases, respectively. The spatial distribution of the mean interpolation error is shown in **Fig. S4**. For all three selected layers, the FR interpolation method results in smaller interpolation errors compared to the RR method. The difference between the two methods is more pronounced above and within the main thermocline (layers 182-218 m and 798-878 m, respectively). For the layer 182-218 m, the largest errors are found within the strongly stratified equatorial belt (Fig. S4a, b) for both the FR and RR methods. Within the main thermocline (798-878 m), the RR method shows the largest positive errors, up to 0.4 °C, across the entire tropical and subtropical belt.

The overall performance of the interpolation method described in the above sections is further illustrated by the temperature/depth histograms based on the observed and interpolated data, respectively (**Fig.10**). To obtain these histograms, the number for each temperature-depth bin is normalized by the number of observations (observed or interpolated, respectively) in the most data-abundant temperature bin for each depth. A good agreement between the two histograms suggests the absence of spurious features in interpolated profiles."

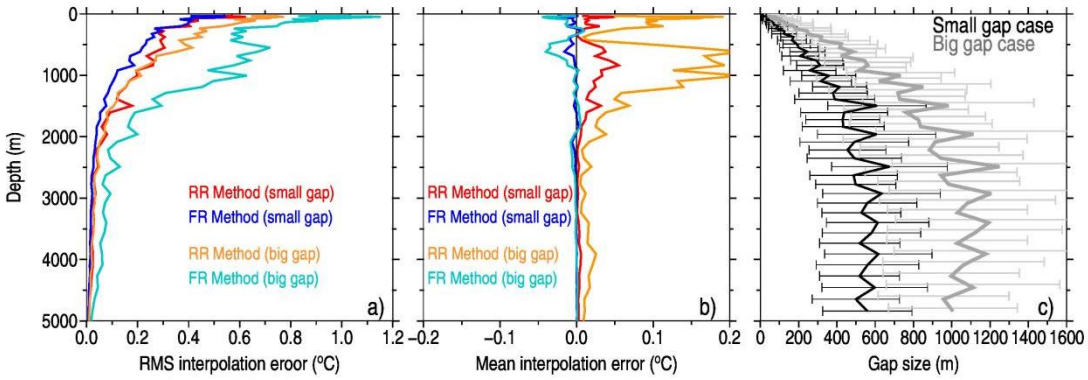

Figure 9. Global root mean squared (a) and global mean (b) error for interpolated temperature values obtained by means of RR method (red and orange) and FR method (blue and cyan). Red and blue profiles correspond to the small gap case, orange and cyan correspond to the big gap case. Explanation Is given in the text. Mean gap size versus depth for the small (black) and big (grey) gap case respectively. Horizontal bars correspond to one standard deviation. (revised)

Line 225: Is the histogram for the interpolated data also normalised by the number of observations? It seems a more sensible choice is to use the number of interpolated data points as the normalisation factor.

We changed that piece of text to make it understandable (Lines 252-254 in the revised manuscript): "To obtain these histograms, the number for each temperature-depth bin is normalized by the number of observations (observed or interpolated, respectively) in the most data-abundant temperature bin for each depth."

Line 226: Do you mean "the most data abundant *temperature* bin for each depth" here?

We changed that piece of text (Lines 252-254 in the revised manuscript): "To obtain these histograms, the number for each temperature-depth bin is normalized by the number of observations (observed or interpolated, respectively) in the most data-abundant temperature bin for each depth."

* Sec 5

Line 269: If the standard levels are removed, why the frequencies for the standard levels are not zero in Fig 11? Perhaps this is caused by the Gaussian smoothing? What is the width of the Gaussian kernel and how is it determined?

Yes, that happens because of smoothing. We added the explanation in Lines 294-295 in the revised manuscript: "We obtained these histograms after removing the standard levels and then smoothing with a Gaussian kernel (the frequencies at standard depths are not zero due to the Gaussian smoothing). After some experimenting, we applied a Gaussian kernel width of 20 m to prevent over-smoothing of the histograms."

* Sec 6

Lines 311-312: Do you mean "σ_MEAN(z) is the global mean of σ(x,y,z)"?

Thank you for noting that. We changed the text in Line 337 in the revised manuscript.

* Sec 7

Fig 13: The depth is denoted using positive values in (d), but negative values in (f).

Figure 13 was corrected.

Fig 13 caption: Lines 344-345 are inconsistent with the text at line 334. Also, "same as f)" should be "g) same as f)".

Corrected!

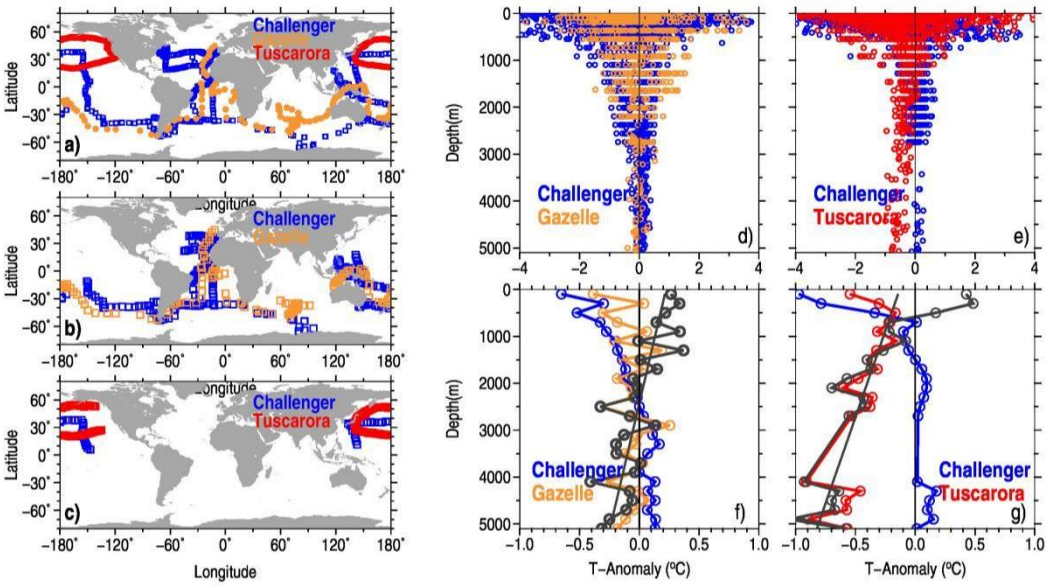

Figure 13. a) Station locations for "Challenger", "Gazelle", and "Tuscarora" expeditions; b) location of "Challenger" and "Gazelle" stations used for data inter-comparisons; c) same as b) but for "Challenger" and "Tuscarora" data; d) comparison of temperature anomalies for "Challenger" and "Gazelle" data; e) same as d) but for "Challenger" and "Tuscarora" data; f) mean temperature anomalies in 200-meter layers for "Challenger" (blue) and "Gazelle" (orange) data and the respective anomaly difference (black); g) same as f) but for "Challenger" (blue) and "Tuscarora" (red) data. (revised)

* Sec 8I suggest swapping "The earliest time period ..." with "The following time periods ..." at lines 351-353. When I read it, I thought "The following time periods" means that the time periods after the earliest time period, which is not what the authors meant here.

We introduced the change in Lines 378-379 in the revised manuscript: "The earliest time spans the quasi-global oceanographic survey conducted by "Challenger", "Gazelle", and "Tuscarora". The selected time periods are: 1873-1876, 1920-1944, 1945-1969, 1970-1989, 1990-2009, 2010-2025."

Line 365 "the northern part of the Atlantic ocean": The northern part of the Atlantic ocean shows a cooling in Fig 14l. Do you mean the northern part of the Pacific ocean?

Fig 14: This is a busy figure. To improve readability, I suggest labelling the time periods to the left of the figure and the depths at the top of the figure using bold font.

Yes, Fig. 14 is busy, and it was truncated, showing anomaly changes for only three depth layers. Please note that the choice of levels is different now. In the original version, the upper boundary of each layer was at the surface, so that for all choices the (much higher) upper layer anomalies dominated the anomaly pattern. Now we show anomaly difference fields for three layers: 0-500 m, 500-1000 m, and 1000-2000 m. Accordingly, we completely rewrote the description of the anomaly changes. To improve the readability of Figure 14, we added labels for the respective time periods.

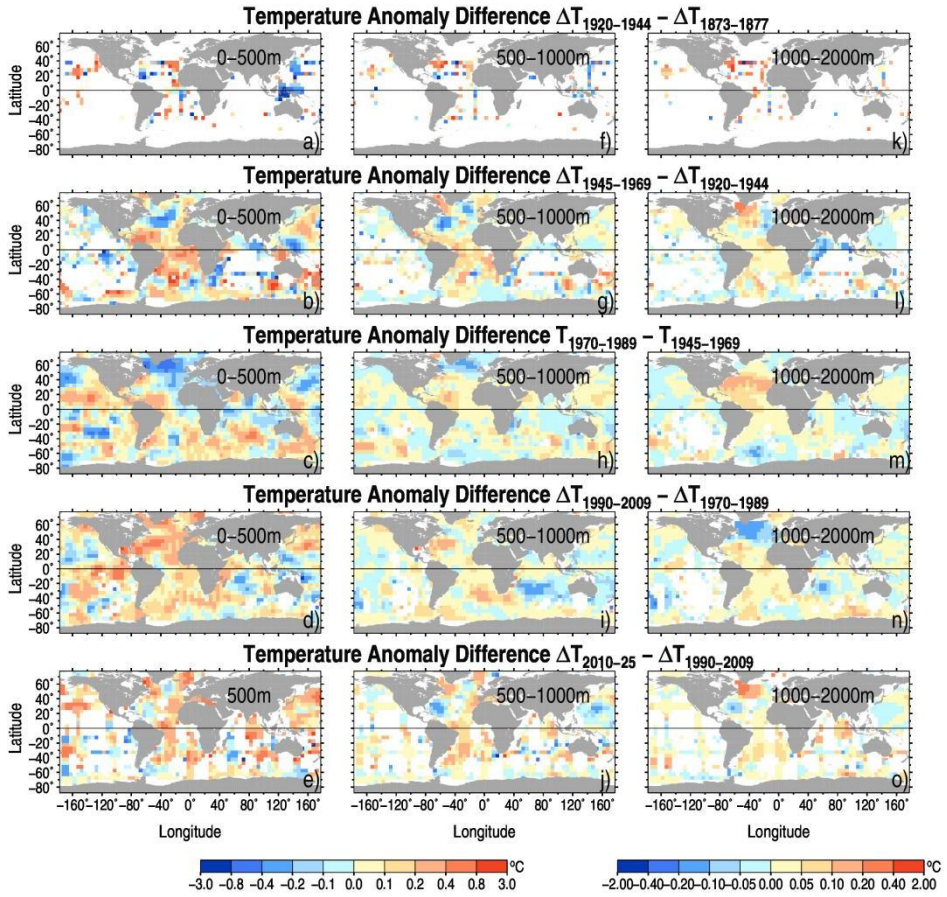

Figure 14. Layer-averaged temperature anomaly differences between selected time periods in 5° × 5° latitude/longitude boxes: 0-500 m (a-e), 500-1000 m (f-j), 1000-2000 m (k-o). Note a different color scale for the layer 1000-2000 m. (revised)

* Sec 8.4

This section would benefit from additional detail on several key points.

The text mentioned the instrumental errors and the mesoscale variability. How are they derived and what are their respective magnitudes?

Fig S8 shows how to estimate spatial covariances. There are several dots for a given lag in Fig S8. What are they?

How are the within-box error and the correlation radius combined to give the error for spatial means?

We have completely rewritten the beginning of Section 8.4 to meet your suggestions (Lines 425-442 in the revised manuscript), where the reference to Boyce (1966) on measurement uncertainty of reversing thermometers was added:

"The box-averaged temperature anomalies in each 100 m layer are calculated using all observed anomaly profiles within the respective 5º × 5º latitude/longitude box. Instrumental errors and mesoscale variability introduce noise, which biases the estimated box-averaged anomalies. Boyce (1966) estimates the instrumental error for reversing thermometers as ± 0.04 ºC. Errors associated with the mesoscale variability are much larger. According to **Fig. S7**, the standard deviation of temperature anomalies within a 2° × 4° latitude/longitude box exceeds 3 ºC in the western boundary currents and in the equatorial belt in the upper 200-meter layer. We assume temperature anomalies within a 5 ° × 5 ° latitude/longitude box are independent because the observations span 20 years or longer, and each mean box anomaly is associated with its respective standard error. For the basin- and global mean anomalies, the errors are obtained by the summation in quadrature of within-box errors. In order to estimate the number of degrees of freedom, we calculated spatial autocovariances in east-west and north-south directions using box-averaged anomalies for the best-sampled period 1990-2009 (**Fig. S10**). The meridional correlation radius is determined by the zero-crossing and is taken to be 1 meridional distance (555 km) between the boxes. The correlation along longitude has a long tail, so we define the correlation radius as the distance at which the correlation function falls below 0.1, which averages two box separations in the zonal direction. We note that for the earliest periods (1873-1877 and 1920-1944) global and basin means are estimated from observations (box-averaged values) that cover only a small fraction of the respective area (see Fig. 14), leading to a mapping error. This kind of error was not estimated in the current study and is not considered in **Figs. 15-16**."

For the earliest periods, basin means are estimated from observations that cover only a small fraction of the basin (see Fig 14 top row). This will probably introduce another error (a mapping error?). Is this considered in Figs 15 and 16?

Yes, that is true. The mapping error (error due to insufficient geographical sampling) is not investigated in our study. This is explicitly mentioned now at the end of Section 8.4 (Lines 439-442 in the revised manuscript):

"We note that for the earliest periods (1873-1877 and 1920-1944) global and basin means are estimated from observations (box-averaged values) that cover only a small fraction of the respective area (see Fig. 14), leading to a mapping error. This kind of error was not estimated in the current study and is not considered in Figs. 15-16."

* Sec 8.5

Lines 423-426 "149 ZJ" "456 ZJ" "30 %" "67 ZJ": Please include uncertainty ranges when discussing these results. The same comment applies to the "30 %" reported in the abstract at line 25.

Please note that we re-calculated our OHC estimates using the specific heat capacity for seawater (3991.87 J/kg°C). In the original version, we used the heat capacity for fresh water. We apologize for this error. Throughout the manuscript, we introduced the updated estimates along with their respective uncertainty ranges.

In the revised version of the abstract, the 30 % increase for the layer 0-2000 m of the OHC between 1920-1944 and 1945-1969 amounts to 32.7 ± 6.8 % of the total OHC increase between 1920-1944 and 2010-2025 (Lines 25-27 in the revised manuscript). The error estimation is explained in Section 8.4 of the revised manuscript.

Lines 442-443: Intuitively, why does the depth corrections tend to increase the warming rate estimate?

The explanation was added – an analogy with the impact of biases in the data from expandable bathythermographs. The reference to Gouretski and Reseghetti (2010) on biases in XBT data was added (Lines 483-485 in the revised manuscript).

Corrections are essentially zero for the modern period (say, from 1990), but corrections for old profiles become more important because the thermometric method for depth estimation was rarely used. The corrections typically make the older temperature profiles colder and so increase the heating rate. We have added corresponding discussions in Section 8.5 (Lines 482-484 in the revised manuscript): "Overall, depth corrections increase the warming rate because they shift observed depths upward, cooling temperature profiles under the prevailing negative temperature gradient."

Line 444: The depth correction factor is close to 0.99 at most depths in Table A1. That is, the correction would only change the depth of a profile by about 1%. However, the impact of the

correction on the ocean heat content change estimate is about 10%, as reported at line 444. This is interesting. Could you please comment on why a small depth correction could lead to a large increase in the estimate of ocean warming?

We completely rewrote the second part of Section 8.5 and added an explanation of the impact of these corrections (Lines 480-485 in the revised manuscript): "The applied depth correction factor (**Fig. 11c**) varies from 0.94 at 50 m to 0.99 at 1500 m, resulting in different regional impacts as the vertical temperature gradient changes geographically. The impact of corrections is most important in the upper part of the water column with typically stronger vertical temperature gradients. Accounting for similar systematic depth errors in temperature profiles from the expendable and mechanical bathythermographs was found to be important for the OHC calculation (Gouretski and Koltermann, 2007; Gouretski and Reseghetti, 2010)."

We note that the depth correction is not linearly related to the change in OHC because the temperature varies nonlinearly with depth.

Line 446: How is the heating rate estimated in Fig 16b? Is it simply the difference between two adjacent time-mean anomalies in Fig 16a divided by the time span between them?

Yes, the explanation was added (Lines 486-487 in the revised manuscript): "The heating rate is estimated as the difference between two adjacent time-mean global OHC anomalies divided by the time span between them."

It would be useful to also show the ocean heating rate from Pan et al., Zanna et al and Wu et al in Fig 16b. In addition, please consider showing heating rate in W per m^2, so that it is easy to compare it with the TOA energy imbalance.

What is the uncertainty range in Fig 16?

We have re-plotted Fig. 16. Now the OHC time series for all studies mentioned in the text with the respective uncertainties, are shown (Fig.16a). Taking into account your suggestion, we show the heating rate in W/m2 (Fig. 16b). Here, the uncertainties are available for our results and for the results by Pan et al. (2026). Uncertainties for Wu et al. and Zanna et al. were unavailable to us.

We have also added some discussion in Section 8.6 (Lines 475-480 in the revised manuscript): "We found that the difference between temperature anomalies for original and depth-corrected profiles can amount to 5-10 % of the OHC change depending on the time period. Overall, depth corrections increase the warming rate because they shift observed depths upward, cooling temperature profiles

under the prevailing negative temperature gradient. Applying depth corrections increases the total energy gain between 1930 and 2015 by 16.0 ± 7.2 %, from 435.7 ± 22.1 ZJ to 505.2 ± 22.2 ZJ. The respective OHC change between 1962 and 2015 increases by 8.5 ± 7.8 % from 293.4 ± 16.5 ZJ to 317.7 ± 16.5 ZJ."

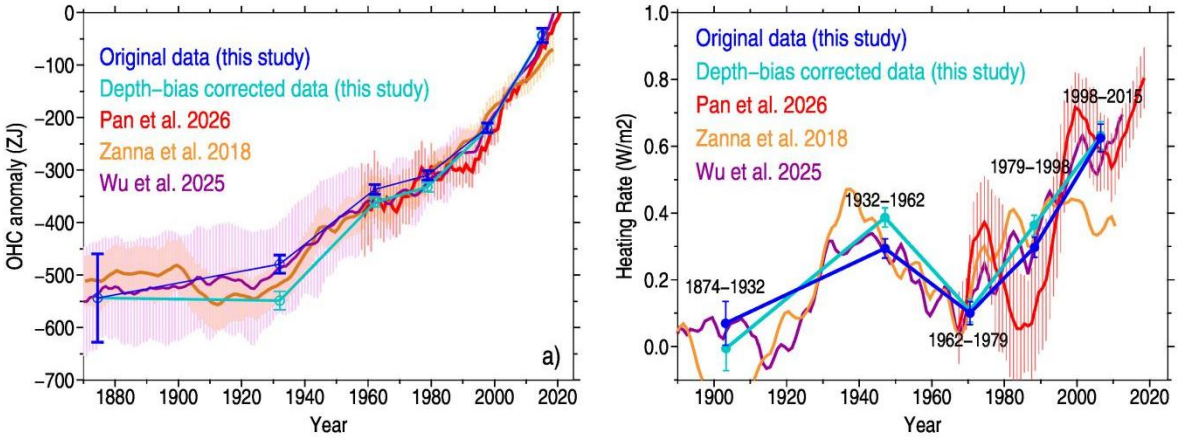

Figure 16. a) Global OHC anomalies for selected time periods for original data (blue) and depth-bias corrected data (cyan), OHC anomaly time series from Pan et al. (2026) (red), Zanna et al. (2018) orange), and Wu et al. (2025) (magenta) are shown for comparison; b) heating rate for five time periods: original profiles (black), depth-bias corrected profiles (blue), Pan et al.2026 (red). Heating rate time series for Zanna et al. (2018) and Wu et al. (2025) are based on the smoothed OHC time series using a 15-year running window. The following error bars are shown: standard error (this study); 95 % confidence interval (Pan et al., 2025); one standard deviation (Zanna et al, 2018); one standard deviation from 73 ensemble members (Wu et al., 2025). (revised)

Lines 460-461: The text says that applying the depth correction would result in a positive heating rate between 1870s and 1920s. However, Fig. 16b shows that doing so reduces the 1874-1932 heating rate from a positive value to near zero, which seems to contradict the text.

Thank you for noting that. There is no contradiction because, for the earliest time, we have not applied depth corrections. We believe the earliest profiles need corrections as well, but we do not provide them in the current study. We added the corresponding explanation in the text (Lines 505-513 in the revised manuscript): "We did not apply depth corrections for the earliest data from "Challenger", "Gazelle", and "Tuscarora" because a different instrumentation was used on these cruises (a hemp rope instead of a steel one, no bottles attached to the wire). For this reason, the heating rate between the first two time periods is close to zero in the case of the depth-corrected profiles (Fig. 16b, blue line). However, two arguments are in favor of a positive depth bias also in the oldest hydrographic profiles: a) unavoidable deviation of the sounding line from the vertical

under real meteorological conditions; b) a significantly larger elasticity of the hemp rope compared to the steel wire, with the Young's modulus for steel being four times higher than that for hemp. Allowing for possible positive depth bias in the earliest "Challenger", "Gazelle", and "Tuscarora" data would likely result in a positive heating rate between the 1870s and 1920s."

* Other items

Line 21 "a smaller ... error": The comparison is unclear. Do you mean a smaller error compared with linear interpolation?

The text was changed to make it more understandable as (Lines 21-23 in the revised manuscript): "This method is shown to result in a smaller interpolation error compared with the linear interpolation and the Reiniger and Ross (1968) parabolic method."

The y-axis label is "percent profiles" in Figs 3a-b and "Frequency %" in Fig 4. They seems to denote the same quantity? If so, please make them consistent.

Thank you, we made the figures consistent.

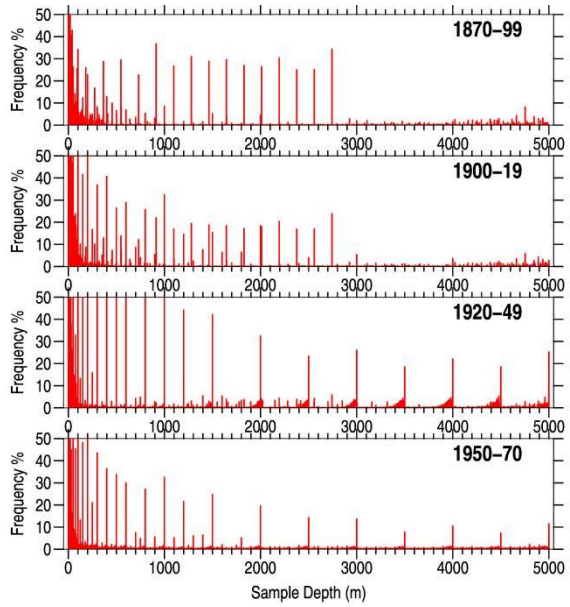

Figure 4. Sample depth frequency histograms for OSD temperature profiles for selected time periods. (revised)

Fig 5: Some of the panel labels and text are difficult to read. Please consider moving them outside the panels to improve readability.

We improved readability and re-plotted the figure.

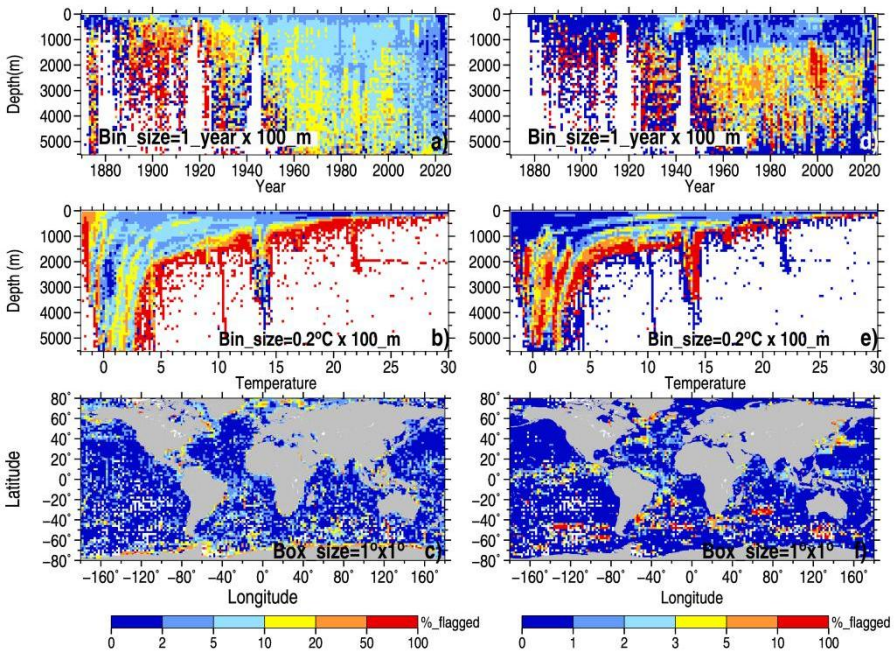

Figure 5. Outlier percentage for OSD temperature profiles: a-c) automated statistical quality control; d-f) manual quality control. Outlier percentage is shown for depth/year bins (a, d), depth/temperature bins (b, e), 1° × 1° latitude/longitude boxes (c, f). (revised).

Line 149: quality controlled -> quality control

Changed. Line 147 in the revised manuscript: 3.2 Manual quality control.

Lines 164-166: Please check if there is a missing comma between:

1) "observed temperature" and "and the ..."

2) "(FR method)" and "which interpolates …"

We have resolved these issues and revised this paragraph as (Lines 170-177 in the revised manuscript): “Two methods were compared to vertically interpolate temperature profiles: the Reiniger and Ross (1968) interpolation method (RR method) and the climate anomaly method by Frankignoul (1981) (FR method). The RR method constructs two three-point parabolas, giving two interpolated values. The final interpolated value is a distance-weighted average of the two estimates, aiming to prevent artificial extrema. The FR method uses the local climatological temperature profile in addition to the observed temperatures and interpolates temperature anomalies (deviations of individual temperature values from the respective climatological values). We do not use a simple linear interpolation here and opt for the RR and FR methods instead. As shown by Li et al. (2022),

using linear interpolation leads to a positive temperature bias for sparsely vertically sampled historical temperature profiles."

Line 197 "a the resampled profile" -> "a resampled profile"

The suggested change has been introduced (Line 208 in the revised manuscript).