We are grateful for your insightful comments, criticism, and time invested in our manuscript. Considering your comments and the questions raised by the other reviewer, we have introduced profound changes in the manuscript, addressing the critical points and aiming to make our results more understandable for the readers. Please note that the line numbers provided in our responses refer to the revised manuscript with tracked changes highlighted in yellow shading. In this response letter, the comments are **in black**, our replies are **in blue**, and the modified pieces of text in the main manuscript are **in orange**.

Response to Reviewer #2:

The authors present an automated quality control procedure for oxygen profiles. The authors then compare oxygen profiles from optodes and electrodes to discrete bottle samples to assess any bias. The authors state that they have taken on this important work in order to provide high-quality bias free data for scientific community to look at global deoxygenation over time on bias-free oxygen data. However, they do not demonstrate that their data product improves our ability to look at global deoxygenation. This could be demonstrated by looking at oxygen data before and after applying their QA procedure and bias adjustment.

Re: Thanks for the concern. Following your suggestion, we introduced changes in the manuscript to try to better demonstrate our ability to use our quality control procedure to identify erroneous observations. We use the high-quality WOCE dataset to demonstrate the ability of the QC procedure not only to identify outliers but also to retain the overwhelming majority of good observations – a pre-requisite feature of any QC procedure. We added Fig. 15, which shows the trajectories of Argo profiles exhibiting quality issues. In the revised version, we also added supplementary material that provides a detailed outlook of the distinct quality of the test performance.

The refined oxygen profile dataset linked to the submitted manuscript aims to facilitate climate change-related studies, including global ocean deoxygenation. We recognize that global oxygen time series based on original and validated data could better demonstrate the importance and impact of the quality control procedure. However, the estimation of the deoxygenation trends is related to the development of proper gap-filling methods, which remains beyond our work scope. Similarly, an in-depth analysis of global oxygen variability and the explanation of oxygen anomalies – as suggested in your comment – remains beyond the scope of the manuscript.

It is unclear how the presented QA procedure differs from the QA procedures implemented by each Argo DACs. Also, how do these 9 QA procedures differ from those outlined in QARTOD (https://repository.library.noaa.gov/view/noaa/18659)? I suspect that the first half of the paper could be significantly cut down by referencing these very similar methods and summarizing the results like in Table 2. From Table 2, I was surprised to find that ~65% of all CTD oxygen observations failed the "stuck value test" and that 60% of all CTD oxygen observations failed the "Local climatological range". This requires further discussion.

Re: These are great points.

(1) Regarding the difference between our QC system and the QARTOD system, we added the comparison of our QC procedure with that from the QARTOD manual. The manual is included now in the reference list. We find that several checks recommended by QARTOD for sensor data have also been implemented in our QC procedure. We note that the QARTOD manual points to the necessity of a *justified choice of the thresholds* for the QC tests (see page 3 of the manual), and this is exactly the point which is the focus of our QC procedure, where the choice of the thresholds is made not as "*ad hoc*" decision, but is based on the underlying statistical structure of the data. Several checks outlined and recommended by QARTOD are tailored for the real-time data flow and are less suitable for static archives.

As also noted in QARTOD, a manual spike test is highly recommended for oxygen sensor data. We improved the description of this check in the edited version and explained how spike thresholds were set.

Besides, we believe that our QC system is a useful addition to the existing Argo-QC because we are working on the QC-ed Argo data. Our QC system was able to identified oxygen profiles with quality issues retained in the DAC-controlled data.

In the meantime, we introduced changes to the oxygen vertical gradient test, which is not included in QARTOD. In the edited version, the local threshold limits are implemented using the box-plot method modified for skewed distributions, which is similar to the approach for oxygen concentration values.

(2) For the issue of the high rejection rate of CTD data, after double-checking the data and QC procedure, we still found it is true that CTD data in the WOD archive suffers from major quality issues. Additionally, we noted in the revised manuscript that CTD oxygen data failing our quality tests were also identified as outliers during the quality assessment at the NOAA National Center for Environmental Information for the World Ocean Database. Therefore, the CTD quality issues mentioned in our manuscript do not affect the results presented in the NOAA World Ocean Atlases for oxygen: "*The significant part of CTD oxygen outliers is attributed to the profile*

standard deviation check, which searches for profiles with identical or very similar oxygen values at all observed (reported) levels (Fig, 11a, check-5). Most of these profiles also fail the local climatological range check. We note that these profiles have also been identified as outliers during the compilation of the WOA18 (Garcia et al., 2018) and WOA23 (Reference) atlases of dissolved oxygen and have not impacted climatological oxygen distributions presented in these atlases.".

This work builds off the author's previous work creating temperature climatology from Argo floats. In this oxygen work, there is also an emphasis on Argo data but with comparison to CTD and Winkler data. The Winkler data is assumed to be accurate and it is unclear to me how the CTD (electrode sensor) comparison adds to the assessment of the Argo (optode sensor) oxygen data when both are compared to discrete oxygen samples (Winklers). And ultimately, since the authors are presenting quality controlled oxygen profiles from Argo assets I would remove the CTD profiles comparison from this paper, since the inclusion/discussion of the CTD oxygen data often seems like an afterthought.

Re: The author group has worked extensively on the historical temperature and salinity data (V. Gouretski and L. Cheng) and oxygen and other biogeochemical data processing (X. Xing and F. Chai), so we feel this paper is a good opportunity to link physical and biogeochemical data processing practices and to improve the data quality, taking advantage of the developments of the QC procedures tailored for the ocean physical variables.

Although the emphasis is on the Argo data (mainly because of the importance of the detected residual bias), CTD data is also a valuable source of oxygen data archive. We decided not to remove CTD data from this paper because: 1) the QC system is applicable for all instruments, including Argo, CTD and Bottle, so working on three different instruments demonstrates the capability of the proposed QC system; 2) documenting the quality issues of CTD data in WOD archive is also a valuable contribution to the community, which has not been done, to our knowledge. Thus, the community should be more careful about the CTD data. ; 3) Our results show that after the proper flagging of CTD oxygen profiles these are in a very good agreement with the reference Winkler data with only very small residual bias. We think the results of the quality control of CTD data represent a useful information for the oceanographic community and can be helpful for identifying the sources of corrupted data.

What do the authors mean by not-dummy oxygen values? Lines 279-280

Re: We changed the formulation about the non-dummy values: "*In contrast, for the CTD group, the histogram (Fig. 10c) exhibits a thick and long tail with a significant fraction of profiles having a high percentage of flagged levels.*".

Why would such a high level of CTD profiles fail quality checks? This is confusing, especially with the abstract stating the residual bias is negligible for CTD oxygen casts.

Re: Thanks. Please see our reply to your previous comment.

Lines 341-343: I'm not sure how Winkler measurements can be considered bias free when the authors just talked about inter-cruise offsets in Section 5.

Re: Yes, we agree that, strictly speaking, the Winkler data are not completely bias-free, and the inter-cruise differences have been established, for instance, in the early paper of one of the authors from the year 2000. The assumption is rather valid for the entire archive of Winkler profiles. We inserted the respective change in the text.

Nevertheless, we have more explicitly discussed this assumption in the limitations/caveats part of the manuscript, "*This study also has some limitations and caveats: (1) Although systematical errors have been identified for Argo oxygen data, the cause of the biases is still poorly known and requires further work. The differences between the DAC centers are also mysterious, and we suspect that the non-standard adjustment procedure developed by different National Argo Data Centers and the difference in sensors on Argo floats used in different countries might be responsible for the differences in diagnosed biases, which needs further confirmation. (2) Because the sources of biases are poorly known, the correction proposed in our study is largely empirical and only applies to the Argo data used in this study. If the Global Argo Data Center updates quality control and adjustment procedures, our bias corrections also require an update. (3) The QC procedure is designed to detect and flag the outliers. However, there are also risks of removing the "real extremes" in the ocean, especially under rapid climate change, as ocean extreme events are expected to become more frequent. One possible way to partly resolve this problem is imposing a trend in the local climatological range, accounting for the time-variation of the local oxygen distributions with climate change, which would help to reduce the false flag percentage of the real extreme data in the ocean. This requires further work when the local oxygen trends become clearer. (4) The Winkler data are used in this study as a reference. However, it is also possible that the Winkler data are not taken to the same standard, thus posing inconsistency within the Winkler dataset, especially for the data*

taken by different countries and time periods. Investigating the offsets on a cruise-by-cruise basis is also recommended in the future, as for CTD data.".

Comments on Figures

I generally found the figures hard to follow. Panels were often mislettered and colorbars were inconsistent and hard to follow throughout the manuscript. See Thyng et al 2015 (https://doi.org/10.5670/oceanog.2016.66) on improving color use in oceanographic sciences.

Re: Following your suggestion, we changed color palettes for all relevant figures

Figure 4: Not needed.

Re: We deleted Fig.4 following your advice.

Figure 6: Mislettered panels.

Re: Fig.6 letters have been corrected

Figure 12: Where these plots created from OSD data? Why were these depth horizons highlighted? G and K are labeled incorrectly according to the caption.

Re: Fig. 12 is now Fig.9. The underlying gridded fields are based on OSD and Argo data.
Please note that in the revised version, we do not assess the quality of the Argo profiling floats within the World Ocean Database (e.g. PFL instrumentation type). Instead, the quality assessment is made for the DAC-quality-controlled and adjusted oxygen profiles. The main reason for the change is that the WOD archive of PFL profiles contains both adjusted and unadjusted data, with raw, unadjusted profiles being substituted with adjusted profiles on a regular basis.

The choice of depth levels is justified in the text (below and above the main thermocline)

Figure 20: The panels are not lettered in alphabetical order

Throughout the manuscript there are a number of copy edits. The worst of which was the caption for Figure 19. Yearly number of BGC Argo profiles equipped with dfifreent tyees of opotde oxgyen sesnors (colored lines). Lightbulee shading corresponds to the total number of proifles: a) AOLM, b) Coriolis, c) JMA, d) CSIRO

Re: Fig.19 is now Fig.18. Lettering in the figure caption corresponds to that in the figure now.

Lastly, the data product is not available at the linked repository, could not be assessed as required for review for ESSD, and therefore results in my recommended rejection of this article.

Re: Unfortunately, we can not explain why the data product was not available at the linked repository, as we checked the availability of the data asking several colleagues to test the data access. We think it should be ok now. We regret that you could not access it while assessing our manuscript.