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, your comments are **in black**, our replies are **in blue**, and the texts in the main manuscript are **in orange**.

Response to Reviewer #1:

General comments:

The manuscript is a valuable contribution to oceanographic research, especially in the context of understanding and monitoring ocean oxygen levels. It is critical to provide high-quality, bias-free ocean oxygen level data. This paper introduces a novel automated quality control procedure. The novel quality control procedure and bias assessment methodology have the potential to significantly enhance the reliability of ocean oxygen datasets. However, to fully realize its potential and solidify its standing as be a substantial contribution to the field, the manuscript would benefit from more rigorous validation, a detailed discussion of its broader implications, and a transparent discussion regarding the potential limitations.

Re: We want to express our gratitude for your evaluation of our work and for your insightful comments, criticism, and time invested in the revision process. Here, we introduce our changes in the revised manuscript to address your comments, which we believe have substantially improved our paper.

(1) "more rigorous validation". Following your suggestion, we introduced changes in the manuscript to try to 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 outlier statistics for each quality check along with the examples of profiles impacted by the respective check.

(2) "a detailed discussion of its broader implications": see our response below and the examples of the added discussion.

(3) "a transparent discussion regarding the potential limitations". This is a great point, and we have listed the limitations/caveats in the final section: "*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."*.

It is good to know the data quality of these commonly used published datasets. Due to the large volume of oxygen profile data, the authors' work is a lot and appreciated. However, as a data paper of ESSD, data quality control processes are not enough, it is also needed for the cost of ignoring these data bias. A discussion of the implications of this new dataset and quality control procedure on the broader field of oceanography would enrich the manuscript. Moreover, the methodology for handling anomalies in oxygen measurements, or 'spikes', needs clarity. The ocean's dynamic nature and the rapid measurement of oxygen profiles mean that spikes in oxygen levels due to abrupt changes in factors like nutrients, currents, or water masses are plausible. A detailed explanation of how these anomalies are approached and analyzed would provide valuable context and strengthen the trust in the methodology employed.

Re: Thanks for raising several important issues, here we addressed these points one-by-one. And We hope we have addressed all these concerns.

(1) For the comment of "However, as a data paper of ESSD, data quality control processes are not enough, it is also needed for the cost of ignoring these data bias.". This study aims to provide a new QC procedure to oxygen observations and also an assessment of the bias in Argo and CTD data, which yields a new in situ oxygen dataset. This database, like other efforts, such as GLODAP and WOD, will support further use of oxygen data in estimating and understanding oxygen changes at different temporospatial scales. We think it is within the scope of ESSD. The "cost of ignoring these data bias" is a good point, we have addressed this issue with several revisions:

1) many examples of the QCs identified by each QC check are provided in the Supplementary Material, so it can be readily seen that these are different kinds of bad data that should be flagged in the database, which can definitely impact the follow-on use of data;

2) More discussions of oxygen trends are provided in the revised manuscripts, to better link the magnitude of bias to the state of the estimate of ocean deoxygenation, see our next reply. An example is, in the discussion section, we add: "*Our calculations find a negative residual oxygen bias in the range -0.66 to -3.72 μmol kg^{-1} for all individual DAC datasets except CSIRO and MEDS. The residual positive bias for CSIRO and BODC profiles is within the range of 0.40-0.76 μmol kg^{-1}. This bias is crucial to accurately identify the deoxygenation trend, as current assessments suggest an upper 1000 m O_2 content decrease of 0.2–1.2 μmol kg^{-1} dec^{-1} during 1970–2010 (Gulev et al. 2023).*".

3) The "costs" can be partly seen by comparing the difference of oxygen climatology with or without bias correction. See Fig.X1, a significant difference is apparent, and this is compariable to the difference between IAP (the authors' group data) and GOBAI (which did not adjust Argo data).

However, we are hesitating to put these figures in this study, because one needs a spatial interpolation derive get gridded fields, which is beyond the scope of this study (the interpolation technique itself needs a rigorous evaluation). Nevertheless, we are open for more thoughts about this.

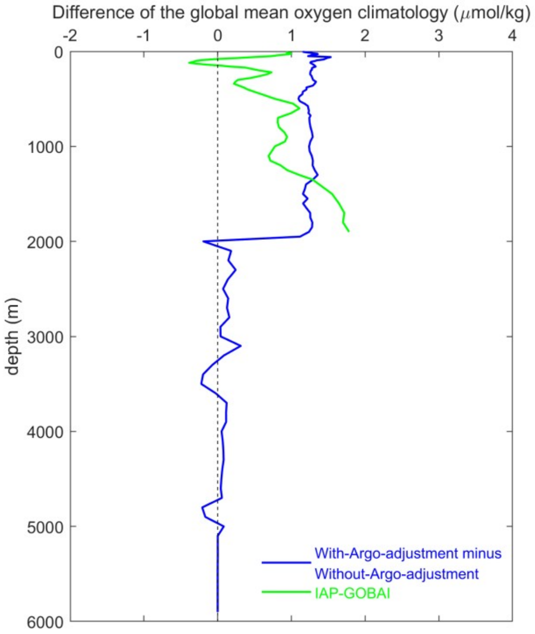

Figure X1. Difference of the Global mean oxygen climatology derived by using data with and without bias adjustment of Argo data (blue), from 1m to 6000m. For comparison, global oxygen difference between IAP and GOBAI data are shown (green). The spatial interpolation of oxygen data used in IAP group is similar to the temperature and salinity reconstruction introduced in Cheng et al. 2017, 2020. GOBAI uses Argo data without additional adjustment.

(2) For the suggestion of "*A discussion of the implications of this new dataset and quality control procedure on the broader field of oceanography would enrich the manuscript.*". We have included a number of discussions on the implications. For instance, in the introduction, a paragraph is added "*These quality issues impede the various applications of oxygen data, for instance, investigating how much oxygen the ocean has lost in the past decades (Levin et al., 2018; Gregoire et al., 2021). Previous assessments indicate the decline of open ocean full-depth O_2 content of 0.3%~2% since the 1960s, with an upper 1000 m O_2 content decrease of 0.5–3.3% (0.2–1.2 μmol kg^{-1} dec^{-1}) during 1970–2010 (Gulev et al. 2023). The maximum estimate is at least 6 times larger than the minimum one, suggesting substantial uncertainty in quantifying the open ocean oxygen changes, which is a grand challenge for the accurate assessment of deoxygenation (Helm et al. 2011; Long et al. 2016; Ito et al. 2017; Schmidtko et al. 2017; Breitburg et al. 2018). Furthermore, there is a mismatch between observed and modelled trends in dissolved upper-ocean oxygen over the last 50 years (Stramma et al. 2012). Uncertainties and differences between estimates are at least partly attributed to the oxygen data quality issues and inconsistency introduced by different instrument types (e.g. different precision, instrument-specific errors/biases) (Gregoire et al., 2021). For example, some BGC-Argo data conduct in-air oxygen measurements which can be used to correct potential systematic errors, while in other cases a climatology isd used (i.e. World Ocean Atlas) as a reference (Bittig and Körtzinger, 2015; Gregoire et al., 2021). Therefore, a consistent and thorough assessment of oxygen data quality, including a uniform data quality control for all instruments and instrumental bias assessments/corrections, is critical to*

providing a homogeneous ocean oxygen database for various follow-on applications, including quantification of the trend of ocean deoxygenation".

And a paragraph in the final Discussion section:" *In summary, this study proposed a new quality control approach and bias assessment for the CTD, bottle, and Argo oxygen data and investigated the consistency between these three primary instrumentation types. Our investigations ensured the consistency between the three datatypes and provided a solid basis for merging them into a single, integrated, and homogeneous oxygen database. Therefore, the database obtained in this study supports the next-step assessment and understanding of the change in ocean oxygen levels.*".

(3) For the last concern about better describing the methodologies, we have improved our description of methods in the revised manuscript (see the track-changed manuscript), including the spike test. We compared our QC procedure with that from the QARTOD manual (the Argo community). The manual is included now in the reference list. Several checks QARTOD recommended for sensor data have been implemented in our QC procedure. We note that 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.

Specific comments:

Line 36-38: This is only true for coastal regions.

Added "in the coastal regions"

Line 153: So you used the same method to the oceanic oxygen distribution? Make it clear here.

Yes it is, a sentence added here "*In the current study we use the Hubert and Vandervieren (2008) adjusted boxplot method as modified by Adil and Irshad (2015).*"

Line 222-223: Why is multiple extrema unrealistic? Any mechanisms behind?

Thanks. The description of the test has been changed. This is a statistical view of the oxygen changes with depth; by definition, we introduced in the manuscript, "*Multiple extrema check aims to identify profiles whose shape significantly deviates from the majority of profiles.*". And we added: "*The larger the extremum magnitude, the less frequent the corresponding profiles. Physically, an oxygen profile at a location is not likely to exhibit too large and too frequent oscillations of oxygen concentrations. Thus, the profiles with many/big extrema are likely erroneous. The histogram for*

Argo profiles differs from those for OSD and CTD because it is based on profiles already validated by the respective DACs.”

Further, we show through many examples that the profiles with multiple extrema are likely unrealistic (see Supplement 6).