Review paper : Earth system models overestimate the sensitivity of apparent oxygen utilisation to age change in the deep ocean

This study investigates the relationship between apparent oxygen utilisation (AOU) trends and water mass age trends in the deep ocean using Earth system models and observational data from GLODAPv2. The authors establish linear relationships between these variables to assess the biological carbon pump's sensitivity to circulation changes under contemporary and future climate scenarios. While this work addresses an important aspect of ocean biogeochemistry relevant to climate projections and would be of interest to the journal, several aspects require further clarification or investigation to enhance the paper's overall strength.

Major Comments

- **Water Mass Definition Methodology**: The water mass classification using PO* thresholds appears somewhat arbitrary and model-dependent. The authors mention "iterative testing" to refine thresholds but provide insufficient detail about this process. Since most of the results are based on this classification, I belive a more detailed section about the derivation of these thresholds is necessary. I also dont understand why there is a difference between the number of water mass definitions for ESMs and observations (four vs two water masses).

- **Linear Relationship Assumption Validity:** While the authors acknowledge in the introduction that linear relationships between AOU and age can break down in certain regions and under specific conditions of the BCP. It would be interesting to further discuss under which conditions this linearity assumption fails in a transient climate scenario. For example, it is unclear to me under which conditions this linear relation still applies (regionally or globally) on SSP5-8.5.

- **Drift Correction Methodology:** The drift correction procedure using a pi-Control simulation makes sens, however, it is not specified how the actual years of the pi-control are adjusted to fit 1850-2099.

- **Observational Data Limitations and T-S Binning**: The T-S space binning approach for observational data analysis is not well justified. The choice of bin resolution (0.027 and 0.0023) seems arbitrary, and the methodology results in very small sample sizes (only 23 and 118 significant joint trends for Southern and Atlantic water masses). The authods should better emphasise the reason behind using this TS binning.

- **On the significance of the observed ovserestimation in ESMs:** The authods show clearly that observational values are "on the low side of the ESMs range" which indicates overestimation of the sensitivity of AOU to age change. While they briefly acknowledge limitations in computing estimates for the contemporary period based on observations (lines around 279), I'm not totaly convinced that the overestimation in ESMs is not due to having more data. For instance, if the ESMs were sampled at the observational locations and the analysis repeated with the same methodological choices as for observations data (e.g. same watter mass definition, same trend calculation approach with the same TS bining), are we going to still going to observe an ovserestimation ?

I hope that the authors will understand my comments in a constructive way, and that I value their work and the time they invested in the preparation of the manuscript. It might be that I have misunderstood something, in this case, if something wasn't clear for me as a reviewer, it is possible that it wouldn't be clear also for the readers.