We sincerely thank the Anonymous Referee #1 for the thorough and constructive comments on our manuscript. We appreciate the effort taken to highlight both the strengths and the shortcomings of our study, particularly regarding the use of observational data sets and the robustness of the model ranking exercise. Based on the instructions of the editorial support team of Copernicus Publications, we will provide a point-by-point response and outline the revisions we will undertake below before we make revision on the manuscript. Italic font will be used to distinguish our replies from the reviewer' comments.

Review of "Evaluating the performance of CMIP6 models in simulating

Southern Ocean biogeochemistry" by Ming Cheng et al.

Scope of the manuscript, general comments and recommendation

The manuscript by Cheng et al. evaluates the performance of the biogeochemical part of CMIP6 models in reproducing Southern Ocean biogeochemical observations. As the Southern Ocean is one of the regions where biogeochemical models diverge most strongly, this is an important subject for a study, especially since biogeochemical models have become quite a bit more complex on average in the transition from CMIP5 to CMIP6 (Seferian et al, .

The evaluation in the manuscript is performed using the typical tools used in that type of study, namely looking at biases, correlation etc. between model output and climatologies of observations, in the end combining the different metrics into an overall ranking of the models. The study is, however, untypical, in that it attempts to judge the models not only against the 'classical' observations, for which good climatologies are avaialable, namely the macronutrients and chlorophyll, but also against observations of the micronutrient iron estimated depths and chlorophyll levels of deep chlorophyll maxima, where those are present, and finally the concentration of POC and even separately the biomasses of zooplankton, detritus and bacteria. Other 'standard' observations, like satellite-based net primary production, dissolved inorganic carbon and total alkalinity are not taken into account.

While I think that the attempt to include new variables into the assessment of biogeochemical models is a progress, the manuscript does not take into account the uncertain state of our knowledge in many of the variables that the authors use. In my view the mauscript is too uncritical of the observational database that they use to compare the models against, and consequently too confident in the ability to judge model outcomes.

Here are my main criticisms concerning this point:

- Firstly, for their iron validation, the authors use the combination of observed bottle data from Tagliabue et al. (2012). This data is (unlike the attribution of this dataset to GEOTRACES, made in the manuscript, which is simply wrong) mostly a compilation of pre-GEOTRACES data of high quality. Since the publication of this data set, a large number of additional data has become available through the GEOTRACES intermediate data products, especially for the Southern Ocean. Why has this data not been taken into account?

We acknowledge the error in attributing the Tagliabue et al. (2012) compilation to GEOTRACES. This will be corrected with the addition of the GEOTRACES IDP2021 *reference. In addition, we will extend the iron evaluation to include the most recent GEOTRACES Intermediate Data Product (IDP2021v2) for the Southern Ocean and repeat the comparison. This will improve the robustness of our iron assessment.*

- For the evaluation of the depth of the deep chlorophyll maximum and chlorophyll concentration at the maximum, the authors have chosen the product from Copernicus, which is based on the works of Sauzede et al. (2016). The authors mention that this dataset estimates POC and chlorophyll using a neural network method, but do not give any further details. Here is therefore my summary of the method: The data set estimates the vertical distribution of particle backscatter (which can be used as a measure of POC) from the large data base of ARGO vertical profiles of temperature and salinity, and co-located surface satellite estimates of particle backscatter and chlorophyll a from MODIS. Actually, contrary to the statement made in the manuscript, the method presented in Sauzede et al (2016) only describes the estimation of POC profiles, NOT of chlorophyll. For the chlorophyll estimation one should probably cite the data manual (https://documentation.marine.copernicus.eu/QUID/CMEMS-MOB-QUID-015-010.pdf). While this data set is unique in that it for the first time allows a look at the vertical distribution of biological activity in the ocean, it is not 'observations' (which is how it is repeatedly referenced to in the manuscript), but a fairly indirect estimate. The limits of this data set and its possible errors are not discussed at all in this manuscript, and neither are the error estimates, which are present in the data themselves, taken into account in the model assessment. Instead, the data set is uncritically taken as 'truth'.

We agree that the Copernicus product for chlorophyll and POC is an indirect reconstruction based on neural network methods, not direct observations. As mentioned in the user manual of the Copernicus product, the current vertical chlorophyll profile product is generated by the latest neural network-based method called SOCA2024, upgraded from SOCA2016 described in Sauzède et al. (2016). That means, the chlorophyll product is uses the same method as with POC product, even though Sauzède et al. (2016) only mentioned the neural method to produce vertical POC profiles. We will explicitly state that these are observation-based estimates, not direct Chl a measurements. We will also discuss the uncertainties and potential errors in the data set. And we will adjust the language throughout to avoid overstating confidence.

- Why is the same data set also taken for the evaluation of surface chlorophyll and POC? As the processing of the data in the copernicus product involves chlorophyll and backscatter estimates from MODIS, it would remove one possible source of error to directly use the satellite data here. Actually at this point it should be discussed that the standard algorithm used in satellite estimates of chlorophyll has been questioned in the Southern Ocean by Johnson et al. 2009 (which is cited in the manuscript); the algorithm proposed in Johnson et al. 2009 gives on average higher values of chlorophyll in the Southern Ocean than the standard algorithm used at that time for SeaWIFS. I think this also hold for the GlobColor product used in the copernicus data set, but I have to admit that this is getting beyod my expertise. But I think it illustrates yet another source of uncertainty in the 'observations' that should be discussed.

We acknowledge that using the analysed Copernicus data set for surface chlorophyll and POC is not ideal. In the revised manuscript, we will include direct satellite-based product (MODIS) as an independent comparison for surface chlorophyll and POC. We will also *discuss the uncertainty of Southern Ocean chlorophyll satellite-based product as highlighted by Johnson et al. (2013), and how this may affect inter-model comparison and model-observation comparison.*

- Just out of curiosity: Many model assessments also use satellite-based estimates of net primary production. Is there a specific reason why this was not done here?

We initially did the evaluation of NPP performance. When considering simulation on phytoplankton may be overweighted in model ranking and length of manuscript, we decided not to put NPP evaluation in the manuscript. Of course, we will consider putting the NPP data back to the manuscript.

- And finally, the authors use ONE number of how POC is distributed over phytoplankton, zooplankton, dead organic matter and bacteria that has been estimated for the Southern Ocean to convert the copernicus estimate of POC into one of phytoplankton, zooplanton, detritus, and bacterial carbon biomass. In their tables 6 and 7 they then judge whether models 'underestimate zooplankton' etc. But when you actually read the paper by Yang et al. 2022, one immediately realizes the limits of that comparison. Firstly, the paper does not describe microzooplankton, but only the biomass of zooplankton that can be caught in plankton nets. Secondly, the biomasses of the three zooplankton groups studied in that paper (mesozooplankton, krill and salps) has a large regional variability, as for example shown in their figure 2. While the Yang paper indeed demonstrates that there is an inverted trophic pyramid in the Southern Ocean, the actual biomass numbers probably have a large uncertainty from sampling bias. Taking the one biomass number for the whole Southern Ocean obtained here then for conversion of a totally different POC estimate into zooplankton biomass further leads to errors. To add to that, the authors do not describe how they have combined the estimates from the three different papers cited into one. In my view it makes sense to investigate whether models obtain a similar inverted trophic pyramid as described in Yang et al, but not to write sentences like 'Most models describe integrated phytoplankton carbon reasonably well with values comparable to observations' when the observations are just indirect estimates of POC from copernicus, multiplied by one Southern Ocean estimate of the phytoplankton carbon:POC ratio, and then not taking possible erors into account. The whole section starting line 412 to line 445 in my view should be scrapped.

We accept that our approach of applying a single partitioning ratio from Yang et al. (2022) is oversimplified and neglects large regional variability and sampling uncertainties. In our original research, we used annual POC data to avoid the effect of missing monthly data of some models. In this case, most models underestimated surface POC concentration according to that the effect of low data availability in winter months on calculation of annual mean. We did some work on classifying carbon type to address the potential points that the types of carbon in the models may have biases. In the revised manuscript, we will redo the surface POC comparison, by changing the annual comparison to summer comparison, to reduce the effect of errors on annual mean calculation. Also, we will delete the section between line 412 and line 445 about discussing POC classification.

Given these criticisms I don't think the paper can be published without quite major revisions. To make it publishable, I think the following needs to be done:

- Extend the data set used for the comparison of modeled iron by the data from the lates GEOTRACES intermediate data product and repeat the comparison.

Dissolved iron comparison will be repeated using GEOTRACES IDP2021v2.

- Redo the comparison of deep chlorophyll maximum frequency and chlorophyll levels taking into account the uncertainty of the copernicus data set.

The uncertainty of the Copernicus data set will be discussed in comparison of DCM.

- use (at least in addition to the copernicus data set) the direct satellite-based estimated of chlorophyl and POC from MODIS for the surface comparison; possibly also discuss the issue of the chlorophyll algorithm uncertainty raised by Johnson et al, 2009.

Surface chlorophyll and POC comparison will be repeated using MODIS data set. And the uncertainty of chlorophyll algorithm will be discussed.

- either remove the comparison with the different components of POC completely or do it properly by accounting for the error margins

The comparison with the different components of POC will be completely removed.

I think all these changes would probably be incompatible with the strong focus of the paper on 'ranking' of the different models, i.e. saying which one is 'the best', which comes second etc. Given the uncertainty of the data sets used, which is completely neglected in the present manuscript, I don't really think this can be done with any confidence.

As this will require more or less a complete rewrite of the manuscript

I limit my further specific comments to the most important ones.

Specific comments

Line 135-136: '.. we use yearly data instead, as carbon export predominantly occurs during summer months': I don't understand the reasoning here. If carbon export predomnantly occurs in summer, does not using annual POC values make the connection of export less reliable?

As we mentioned above, we will redo the surface POC comparison by changing the annual comparison to a summer comparison. The new POC comparison will include all 14 models.

Formula 4: The formula for root-mean-square difference is given here corectly; but in the Taylor diagnam one should use the RSMD after correction for the mean model-data bias, otherwise the connection between CC, SSD and RSMD that is used to construct the diagram does not hold (Taylor 2001). Was this done here?

We ensure that the RMSD values were bias-corrected, so the Taylor diagrams were correctly plotted.

line 153: 'the number of grid points..' Does that depend on the grid resolution? Is that a problem?

Actually, the DCM frequency is calculated based on the area of the grid points, not simply the number of grid points. We will change "the number of grid points" to "the area proportion".

Table S1: Were the calculations of CC and other statistical quantities for chlorophyll done using log-transformed data, as is done most of the times?

We did not apply log transformation when calculating CC and other statistical metrics for chlorophyll. For visualisation, we used a moving scale.

Comparison of surface nitrate and silicate: Given that the Southern Ocean is an upwelling region, would it make sense to also check the concentration of these nutrients in Circumpolar Depp water with data when tryng to explain the model-data difference at the surface?

This is a sensible suggestion. In our manuscript, we mainly focus on biogeochemical performance and the effect of biogeochemical processes on biogeochemical performance. We acknowledge that CDW nutrient concentrations influence surface fields in the Southern Ocean, we will consider comparing CDW nutrient concentrations although this analysis is beyond our current scope. We will need to explore this.

When comparing dissolved iron with the Tagliabue et al. 2012 data set, mean bias estimates are given. Does a mean make sense in such a sparse data set? Should one perhaps at least also have a look at the median?

The dissolved iron data from Tagliabue et al. (2012) are distributed to 1°×1° grids by calculate their median of closest samples to plot the surface dissolved iron map. In this case, we compared the dissolved iron difference by calculating the mean. We will provide more details on how we used the iron dataset and how we have utilised the GEOTRACES IDP2021v2 data product, noting that most of the Tagliabue et al. (2012) included data from the IPY 2007-2008.

In the iron comparison, repeatedly the 'limited availability of observational data' is referred to, which is correct. But the data is not that limited, given the GEOTRACES data that is ignored here.

We will redo the dissolved iron comparison by using the latest GEOTRACES product (IDP2021v2).

Model ranking: it is unclear to me how the different statistical quantities to judge model-'data' agreement are converted into one ranking. Is the lowest RSMD the criterium, the highest CC?

The overall ranking of each model is based on its ranking of the different variables. The ranking of a variable for a model is based on rankings of four statistics: MBE (the lowest |MBE| have highest ranking), SSD (the closest to 1 have highest ranking), RMSD (the lowest have highest ranking), and CC (the highest have highest ranking). We will more detailly describe the criterium of ranking in the method section.

Line 383: "DCMs are primarili driven by photoacclimation". No, not all of them, see Cornec et al. 2021. The whole discussion of DCMs and the factors driving them is a bit superficial.

We agree with that not all of DCMs are driven by photoacclimation. Cornec et al. (2021) indicated that around half of DCMs are driven by photoacclimation and another half are DBMs. This situation in models is different to conditions within the "real" water column. In *models, chlorophyll only represents live phytoplankton, while it will be excluded from the count after phytoplankton dies and is transfer to the detritus pool. However, in the real water column, chlorophyll can also be detected in died phytoplankton. In addition, Boyd et al. (2024) suggested DCM and DBM formation and persistence can be a result from recycled iron within the subsurface associated with the maximum in ammonium and upward silicate transport from depth which support diatom production. The challenge is most models do not simulate this well. These structural difference between the real water column and models makes simulating DCMs in models challenging. In this case, the modelled DCMs are not as strong as them discovered in the water column. We will add related content to the manuscript to interpret the bias on DCMs between observation and simulation.*

References

Cornec, M., Claustre, H., Mignot, A., Guidi, L., Lacour, L., Poteau, A., et al. (2021). Deep chlorophyll maxima in the global ocean: Occurrences, drivers and characteristics. Global Biogeochemical Cycles, 35, e2020GB006759. https://doi. org/10.1029/2020GB006759

Additional References

Boyd, P. W., Antoine, D., Baldry, K., Cornec, M., Ellwood, M., Halfter, S., Lacour, L., Latour, P., Strzepek, R. F., Trull, T. W., & Rohr, T. (2024). Controls on Polar Southern Ocean Deep Chlorophyll Maxima: Viewpoints From Multiple Observational Platforms. Global Biogeochemical Cycles, 38(3). https://doi.org/10.1029/2023gb008033

Cornec, M., Claustre, H., Mignot, A., Guidi, L., Lacour, L., Poteau, A., D'Ortenzio, F., Gentili, B., & Schmechtig, C. (2021). Deep Chlorophyll Maxima in the Global Ocean: Occurrences, Drivers and Characteristics. Global Biogeochemical Cycles, 35(4). https://doi.org/10.1029/2020gb006759

Johnson, R., Strutton, P. G., Wright, S. W., Mcminn, A., & Meiners, K. M. (2013). Three improved satellite chlorophyll algorithms for the Southern Ocean. Journal of Geophysical Research: Oceans, 118(7), 3694-3703. https://doi.org/10.1002/jgrc.20270

Sauzѐde, R., Claustre, H., Uitz, J., Jamet, C., Dall'Olmo, G., D'Ortenzio, F., Gentili, B., Poteau, A., & Schmechtig, C. (2016). A neural network -based method for merging ocean color and Argo data to extend surface bio -optical properties to depth: Retrieval of the particulate backscattering coefficient. Journal of Geophysical Research: Oceans, 121(4), 2552-2571. https://doi.org/10.1002/2015jc011408

Tagliabue, A., Mtshali, T., Aumont, O., Bowie, A. R., Klunder, M. B., Roychoudhury, A. N., & Swart, S. (2012). A global compilation of dissolved iron measurements: focus on distributions and processes in the Southern Ocean. Biogeosciences, 9(6), 2333-2349. https://doi.org/10.5194/bg-9-2333-2012

Yang, G., Atkinson, A., Pakhomov, E. A., Hill, S. L., & Racault, M. F. (2022). Massive circumpolar biomass of Southern Ocean zooplankton: Implications for food web *structure, carbon export, and marine spatial planning. Limnology and Oceanography, 67(11), 2516-2530. https://doi.org/10.1002/lno.12219*