We wish to thank Pr. Marcello Vichi for offering many insightful comments and helping us to improve the manuscript. Here we offer detailed responses to all questions. Reviewer's comments are in black, our replies are in blue.

General Comments:

There is some need to further strengthen the concept of why the BGC-Argo data should be considered the most appropriate reference dataset for global model assessment, and how they relate to the other existing datasets (especially satellites, which are going to be superior for evaluating surface chlorophyll than BGC-Argo; see my comment 4 below). There is little doubt that the BGC-Argo program will become a reference climate data record in the longer term. Maybe the authors should provide some clearer recommendations to the readers in their final section. As it stands, the conclusion section appears truncated, with a series of comments that one would mostly expect in a report rather than in a journal article (see in this regard my comment 3 below).

REPLY: We have strengthened the concept of why we use the BGC-Argo float as the unique reference dataset in our study, in the abstract, and introduction.

In the abstract:

"The use of BGC-Argo observations as the single evaluation data set ensure the accuracy of the data as it is an homogenous data set with strict sampling methodologies and data quality control procedures" .

In the introduction:

"The BGC-Argo data set represents a significant improvement for the assessment of models comparing to large databases such as the World Ocean Database (WOD) (Boyer et al., 2013) or the Copernicus Marine Service in situ dataset (European Union-Copernicus Marine Service, 2015). Large databases are composed of data collected from various instrument types with heterogenous data sampling methodologies. Therefore, for a given variable, the accuracy numbers are not the same and change depending on the instrument type (European Union-Copernicus Marine Service, 2019). Consequently, this affects the overall accuracy over time due to the changing proportion of instrument types over the years. On the other hand, the BGC-Argo data set is an homogenous data set with strict and uniform sampling methodologies and data Quality-Control (QC) procedures. As a result, the BGC-Argo data set have a satisfactory level of accuracy, which remains stable over time (Johnson et al., 2017; Mignot et al., 2019).

Moreover, the number of quality-controlled observations collected every year by the BGC-Argo fleet is now greater than any other data set (Claustre et al., 2020). Using the BGC-Argo dataset as the single evaluation data set is therefore a way to ensure consistent accuracy. "

We have also provided recommendations in the manuscript and in the conclusion how to relate the BGC-Argo and satellite Chl*a.*

"While the assimilation decreases the model-BGC-argo data misfit for Chl$_{mixed}$ comparing to a simulation without assimilation (not shown), the model errors for the three metrics associated with Chla remains systematically larger than the BGC-Argo variability. Yet, it has been shown that, when comparing to the satellite Chla product assimilated (European Union-Copernicus Marine Service, 2022), the model-satellite misfit was lower than the variability of the satellite data (European Union-Copernicus Marine Service, 2019). This suggest that the model-BGC-Argo data misfit could originate, in part, from discrepancies between the satellite Chla product assimilated and the BGC-Argo data. We propose that studies should check the consistency between ocean colour products and BGC-Argo Chla products at the global scale as these two products are expected to be assimilated together in future operational BGC systems (Ford, 2021)."

We have also rewritten the conclusion entirely.

Section 3 is still confusing. I apologise with the authors if this is due to my own limitation, but I feel there could be other readers raising questions like mine. Somehow, the previous version of the manuscript was clearer, although I realize that this may be a consequence of all the other changes in this revision. I would suggest the users be clear with their definitions. They now indicate that 22 metrics can be extracted from the BGC-Argo datasets, but they do not explain clearly that these metrics have been grouped according to key components/processes of marine ecosystem functioning (i.e. the 4 sub-sections presented in Sec. 3). This grouping is evident in Table 2, but the text is unsatisfactory. The confusion is further augmented by naming one of the key processes "Oceanic pH" (one of the metrics) instead of "Ocean acidification". The authors say: "The metrics are described below", but actually they first describe the processes, and then how the metrics derived from the BGC-Argo data can be used to quantify these processes.

REPLY: We agree with reviewer. We have rewritten the section that defines the assessment metrics and made the grouping more evident in the text.

They should also explain why certain metrics are included in one grouping rather than another. For instance, the surface partial pressure of CO2, which is essential for estimating the air-sea flux, can be computed from pH and DIC, which have been included in two different groups. It is true that inorganic carbon is linked to both the physical solubility pump and the biological carbon pump, and this ambiguity should be recognized.

REPLY: We agree with the reviewer that DIC is linked to both the physical solubility pump and the biological carbon pump. However, it is now included in the carbonate chemistry metric considering that the classical variables for the study of carbonate chemistry are DIC, Alk, pH and pCO_2 (Williams and Follows, 2011).

I am (now) aware of the main intent of this manuscript. However, more effort should be put into demonstrating that this exercise is a contribution to the literature on global biogeochemical models and their assessment, rather than a report that could have been produced by CMEMS as part of their operational endeavour. For this reason, I would recommend the authors to improve their description of results, which is often written as a dry reporting of the model discrepancies. This is instead well done in Sec. 6, which is now very clear and combines the demonstrative aims with the provision of some directions for future research and/or analyses. I have given some more specific comments in the next section.

REPLY: We thank the reviewer for bringing this to our attention. The main text of the manuscript has been largely revised, and we have improve the descriptions of the results. We now provide directions for future research and/or analyses that are summarized in the conclusion:

“

Overall, the model surpasses the BGC-Argo climatology in predicting pH, DIC, Alk and O_2 in the mesopelagic and the mixed layers, as well as NO_3, Si and PO_4 in the mesopelagic layer. Concerning the other metrics, whose model predictions are outperformed by the BGC-Argo climatology, we provide suggestions to reduce the model-data misfit and thus to increase the model efficiency. For, PO_4, Si, and NO_3, we propose to test if the uncertain model error covariances during the assimilation of satellite Chla could lead to a degradation in predicting nutrients in the mixed layer. For Chla-related metrics, we recommend to check the consistency between ocean colour products and BGC-Argo Chla products at the global scale as it may explain part of the *misfit between the model, that assimilates satellite Chla, and BGC-Argo observations. The discrepancies between modelled and observed POC and OMZs have been already investigated in previous studies. It has been suggested that improving the BGC-Argo POC-b_{bp} conversion factor, tuning the model parameters and implementing missing processes in the model structure could decrease the model-data inconsistencies associated with POC dynamics. Similarly, the improvement of the ocean circulation in physical models should improve the accuracy of OMZs model predictions. Finally, pH_{mixed} and $pCO_{2\ mixed}$ should be better modelled if the uncertainties associated with DIC, Alk, temperature and salinity in the mixed layer are reduced.*

The method proposed here is also beneficial to inform about the BGC-Argo network design. In particular, the regions where BGC-Argo observations should be enhanced to reduce the model-data misfit through the assimilation of BGC-Argo data or process-oriented assessment studies. We strongly recommend to enhance the Arctic region, which is critically under sampled and is constantly outperformed by the BGC-Argo climatology. Likewise, BGC-Argo observations should be enriched in the Equatorial region and in the Southern Oceans, two regions where the model error barely exceed the BGC-Argo observations variability."

The authors rightly claim the unicity of this data set as well as its multivariate nature. However, this is not always put into practice in a demonstrative sense. I am particularly critical with Section 5.c, in which surface Chl is presented as an example of the maps. Why using sChl as the demonstrative metrics? This field is far better represented in terms of temporal frequency and spatial coverage by the satellite record and I'm sure the authors would recommend modellers to use this product for their validation. I am also sure satellite Chl has been used thoroughly before making the CMEMS model publicly available as a shared product.

REPLY: We thank the reviewer for bringing this to our attention. We have revised the methodology of the study and we have followed the approach of Allen et al. (2007) to put into practice the multivariate nature f the BGC-Argo data. The new methodology is summarized in the abstract:

" Here, we propose a new method to inform about the model predictive skill in a concise way. The method is based on the conjoint use of a K-means clustering technique -- an unsupervised learning algorithm, assessment metrics and BGC-Argo observations. The K-mean algorithm and the assessment metrics reduce the number of model data points to be evaluated. The metrics evaluate either the model state accuracy or the skill of the model in capturing emergent properties, such as the Deep Chlorophyll Maximums or

Oxygen Minimum Zones. The use of BGC-Argo observations as the single evaluation data set ensure the accuracy of the data as it is an interoperable homogenous multivariate data set with strict data quality-controlled procedures. The method is applied to the Copernicus Marine Service global forecasting system. The model performance is evaluated using the model efficiency statistical score that compare the model-observations misfit with the variability of the observations, and thus objectively quantifies whether the model outperforms the BGC-Argo climatology."

We cannot use the satellite Chl*a* for the validation of the model, as this product is already assimilated in the model.

I question the decision to not include in the main text one of the other variables that would not be available without the BGC-Argo dataset and CANYON-B. They are in the Appendix, and to me far more informative than sChl. As a modeller, my main question when reading this section is not how relevant the BGC-Argo dataset is to assess model performances, but rather how surface Chl from that dataset compares with the satellite record. This issue also applies to the results presented for the Atlantic time series in Sec 4.d. Why choose variables that have previously been used to assess models (nutrients and chlorophyll), instead of selecting new variables such as pH, DIC and POC, which would definitely give information on the processes of interest. In this case, these figures are not provided in the Appendix, which is a missed opportunity to demonstrate one of the main aims of this paper. If this is done because these results are not very good, then it is even more worrying.

REPLY: We agree with the reviewer. We have now included in the main text all variables derived from the BGC-Argo dataset and CANYON-B.

Specific comments

P1 L20-22 This is a generic statement for an abstract. The same can be said of BGC-Argo data, since rates are also not directly measured

REPLY: We have removed this sentence.

P2 L7 Has taken

REPLY: Thank you, we have made the correction.

P3 L7 All datasets are incomplete and have limitations, including the BGC-Argo

REPLY: We agree but we did not imply that the BGC-Argo dataset has no limitations.

P3 L30 Please explain why these AI methods cannot be applied to the other datasets

REPLY: These AI methods can also be applied to datasets that include temperature, salinity and oxygen measurements.

P4 L4 The dataset represents

REPLY: Thank you, we have made the correction.

P5 L2-L5 This sentence is unclear. What does it mean to be arduous? Is it a problem with the data set? Should the readers abstain from attempting it because it would not be possible? This sentence would be understandable if further discussed in the conclusions. As it stands, it seems the authors are justifying themselves for not having done it.

REPLY: We have removed this sentence.

P5 L19-20 There is a need to clarify from the beginning which are the variables directly measured with the on-board sensors of the BGC-Argo devices (primary variables?) and which ones are further derived (secondary or derived variables?). This would help in understanding the author's definition of metrics. This is further complicated in the reminder because some variables are a combination of derived and measured products (pH, NO3), and it is not always clear what is the percentage (for instance, in Sec. 5.d).

REPLY: We agree with the reviewer that some variables are primary while others are secondary. However, for simplicity, the variables derived directly from BGC-Argo or CANYON-B are mixed together. This is justified in the data section: "

"Finally, we complemented the existing BGC-Argo dataset with pseudo-observations of NO_3, PO4 , Si, and DIC concentrations as well as pH and pCO2 using the CANYON-B neural network (Bittig et al., 2018). CANYON-B estimates vertical profiles of nutrients as well as the carbonate system variables from concomitant measurements of floats pressure, temperature, salinity and O2 qualified in "Delayed" mode together with the associated geolocalization and date of sampling. The CANYON-B estimates of NO_3 and pH were merged with measured values on the rationale that CANYON-B estimates have RMS errors (NO_3 = 0.7 µmol kg^{-1} , pH = 0.013) (Bittig et al., 2018) that are of the same order of magnitude as those of the BGC-Argo observations errors (NO_3 = 0.5 µmol kg^{-1}, pH = 0.07) (Mignot et al., 2019; Johnson et al., 2017) .

Finally, we verified that the RMS errors of BGC-Argo data (both measured and from CANYON-B estimates) are lower than the RMS difference between the model and BGC-Argo data, so that the comparison of simulated properties with the BGC-Argo data leads to a meaningful evaluation of the model performance. We believe it is reasonable to draw conclusions on the model uncertainty from BGC-Argo data as long as the BGC-Argo errors are much lower than the model-observations RMS difference."

P7 L1-2 This is a very relevant addition. However, I do not see this concept further used in the presentation of the results. It is for instance not discussed when showing the RMSD in Fig. 5 and 6.

REPLY: We now use the model efficiency statistical score to assess the model performance.

"The model efficiency tests whether the model outperforms the BGC-Argo climatology ($0 < m_e < 1$,Fennel et al., 2022), or stated differently, if the model-data mean square difference is lower than the observation variance, i.e., $\sum_{i=1}^{N}(m_i - o_i)^2 < \sum_{i=1}^{N}(o_i - \bar{o})^2$."

P7 L20-22 This sentence is not connected with the following. It is customary to use the lower time frequency or coarser spatial resolution when comparing data and models (as done with the spatial maps in the results section). Why did the authors decide not to use weekly averages of the Argo data?

REPLY: We have connected the sentence with the following. We did not use the weekly averages because all model variables except POC are available as daily values.

P7 L25 to match

REPLY: Thank you, we have made the correction.

P7 L30 Was this done using the daily interpolation?

REPLY: Yes, it was done using the daily interpolation.

P12 L13 I would suggest using sparseness rather than scarcity. Argo data are still scarce.

REPLY: This sentence was removed during the rewriting of the main text.

P12 L21-22 Unclear sentence. Does it mean that showing this would confound the reader?

REPLY: This sentence was removed during the rewriting of the main text.

P13 L14 This section is presented as a technical report. I would recommend adding a few more sentences that point at the relationship between the metrics and the processes in section 3. For instance, when referring to the oxygen levels, make an explicit connection with sec. 3.a, and the same with the other variables. I think the value of the message would be further enhanced if there is a more direct connection between Sec. 3 and Sec. 5. The demonstrative aim of the manuscript is clear, but because there is no discussion section it would help to have some additional comments. Many questions arise, for instance, why Chl performs badly while nutrients don’t, while DIC is also good and spCO2 and spH are similarly worse? I am not asking the authors to offer full explanations since this would be beyond the scope of the work, but the indication that the BGC-Argo data help to highlight these discrepancies, which would not be possible with other datasets.

REPLY: This section was removed during the rewriting of the main text. In the new version of the manuscript we have made a direct connection between the variables when assessing the model.

P13 L20 close to the

REPLY: This sentence was removed during the rewriting of the main text.

P14 L2-4 This is another sentence that would be improved through references and linkages to Sec 3.

REPLY: This sentence was removed during the rewriting of the main text.

P14 L12 as well as

REPLY: This sentence was removed during the rewriting of the main text.

P14 L16 There is also a lack of sensitivity in the model for very low oxygen regions close to 0 umol kg-1. The model can have any number between 0 and 30 umol kg-1 when observed values are close to 0 umol kg-1. The feature reported in the text is relevant but the number of data is not very high. While the discrepancy around zero has a higher data density.

REPLY: This section was removed during the rewriting of the main text.

P14 L17 Cape Verde (https://en.wikipedia.org/wiki/Cape_Verde)

REPLY: This section was removed during the rewriting of the main text.

P14 L29 Figure 1 shows data counts, not Chl patterns. Please clarify.
REPLY: This sentence was removed during the rewriting of the main text.

P14 L31 Please explain the meaning of coherent. This should not be the first time this model is assessed against surface chl from satellites.

REPLY: This section was removed during the rewriting of the main text.

P14 L34 This is another comment that I would expect in a report. My understanding is that the aim of the work is to highlight what can be learned from the use of BGC-Argo data that is not possible with other datasets (e.g. satellite data).

REPLY: We have clarified the aim of the work in the introduction:

“The objectives of the present study are twofold. Our first aim is to propose a methodology that uses the BGC-Argo data set, an unsupervised learning algorithm and assessment metrics to simplify marine BGC model-data comparisons, and thus inform, in a concise way, about model performance. The second objective is to use this methodology to also identify ocean regions where the model-observations misfit is larger than the variability of the BGC-Argo data and thus inform the BGC-Argo observing system of regions that should be better sampled.”

P15 L6 Is there a reason for using quotation marks for the spring bloom?

REPLY: This sentence was removed during the rewriting of the main text.

P15 L18-22 This is another dry sentence used for a major misestimation, which would require some more context or a brief discussion. The percentages are extremely high. I am not questioning the model quality, rather the value of offering interpretations based on the assessment exercise.

REPLY: This sentence was removed during the rewriting of the main text.

P15 L28 Please indicate if these percentages are satisfactory with respect to the reference uncertainties indicated in the methods (P7 L1-2 and previous lines). This comment also applies to the previous point.

REPLY: This sentence was removed during the rewriting of the main text.

P15 L30-32 It would be helpful if the authors could add some comments on how the multivariate data from BGC-Argo allow to constrain models in a way that was sparse and more difficult 15 years ago. Consider for instance Vichi, Masina and Navarra (2007), in which all possible existing data were used to assess a global ocean BGC model. There is no need to add this reference, it's just one of the examples of how model assessment has been done in the literature.

REPLY: We have commented on the multivariate aspect of the BGC-Argo data in the introduction:

"The BGC-Argo floats also provide multivariate observations at high vertical and temporal resolutions and for long periods of time providing nearly continuous time series of the vertical distribution of several biogeochemical variables. This is not possible with discrete, univariate vertical samplings provided by cruise cast in situ measurements or from climatological values derived from the WOA. All these specificities overcome the limitations of the previous datasets, especially with respect to their univariate nature, as well as their limited vertical and temporal resolution. This opens new perspectives for the evaluation of BGC models(Gutknecht et al., 2019; Salon et al., 2019; Terzić et al., 2019)."

P17 L10 I suggest to use "limited" instead of lack

REPLY: This sentence was removed during the rewriting of the main text.

P17 L11 Increased number is not the only advantage. They are coherent, consolidated and sustainable. They could become equivalent to the concept of climate data records used for satellite data.

REPLY: This sentence was removed during the rewriting of the main text.

P17 L20-21 I would suggest to refer to the processes presented in Sec. 3

REPLY: This sentence was removed during the rewriting of the main text.

P21 Table 2 Please clarify in the section text if the definition used here is the same for both the model and the data

REPLY: We have clarified in the text that the definition of the metrics is the same for the model and the BGC-Argo data.

Fig. 4 and all the maps. It would be very helpful to add the maximum and minimum values of the range in the colorbar, to better understand the spread of data values

REPLY: This maps were removed during the rewriting of the main text.

##
##

We thank the reviewer #3 for their thoughtful comments. Here we offer detailed responses to all questions. Reviewer's comments are in black, our replies are in blue.

The manuscript "Using BGC-Argo floats for the assessment of marine biogeochemical models: a case study with CMEMS global forecast system" by Mignot et al. proposes 22 metrics for the assessment of biogeochemical models and applies them to a single model. As such, the analysis is a very welcomely comprehensive application of ocean BGC Argo observations, but is done in a vacuum without reference to previous or alternative modeling efforts. While this approach is fine from a technical report documentation perspective, it does not fit the standard of a scientific research paper. As such, it would seem more appropriate for "Geoscientific Model Development" than "Biogeosciences" in its present form.

REPLY: We have rewritten the main text of the manuscript in order to transform the manuscript into a scientific research article. In the introduction, the problematic is stated more clearly:

"The development of BGC models as well as the continuous increase in spatial and vertical resolutions has reached the point where the volume of model outputs has

dramatically increase. Simplification techniques are therefore required to provide decipherable information on model predictive skill. Allen et al. (2007) proposed a methodology for reducing the spatial dimensions in model assessment exercises, thereby providing concise information about the model performance. They use an unsupervised learning algorithm to classify the Southern North Sea into 5 coherent BGC regions based on modelled time series of temperature, NO_3, NO_3, and Si concentrations. They then evaluated the predictive capabilities of the model in each BGC region (instead of at each grid point), thus greatly reducing the number of points to be validated. An additional method for reducing the dimensions of model-data comparison is the use of assessment metrics (Hipsey et al., 2020; Russell et al., 2018). In particular, metrics such as depth-averaged state variables (e.g., mixed layer averaged Chla, NO_3, O_2, etc…), mass fluxes and process rates validation (e.g., primary production or division rates), or emergent properties validation [e.g., Deep Chlorophyll Maximum (DCM), or Oxygen Minimum Zone [OMZ]) are particularly useful to reduce the number of model's vertical layers to be compared with the observations. "

The objectives of the paper are also clearly explained:

"The objectives of the present study are twofold. Our first aim is to propose a methodology that uses the BGC-Argo data set, an unsupervised learning algorithm and assessment metrics to simplify marine BGC model-data comparisons, and thus inform, in a concise way, about model performance. The second objective is to use this methodology to also identify ocean regions where the model-observations misfit is larger than the variability of the BGC-Argo data and thus inform the BGC-Argo observing system of regions that should be better sampled. The first step of the method consists in defining 23 assessment metrics that are used both to construct the BGC regions and then to compare the model outputs with the BGC-Argo data. Second, following the approach of Allen et al. (Allen et al., 2007), we use an unsupervised learning algorithm, here a K-means clustering technique, to classify the global ocean into 8 coherent BGC regions based on the climatological modelled time series of the 23 assessments metrics. In the last step, the skill of the model in predicting the assessment metrics is evaluated in each BGC-region, using the model efficiency statistical score. Unlike other statistical metrics such the correlation coefficient, the bias or the root mean square difference, that does not quantifies objectively whether the model performance is acceptable or not; the model efficiency calculates whether the model outperforms an observational climatology (Fennel et al., 2022). Finally, the method is implemented using the Copernicus Marine Service global BGC forecasting system (European Union-Copernicus Marine Service, 2019)."

The conclusion also provides recommendation for future research/analysis:

“*Overall, the model surpasses the BGC-Argo climatology in predicting pH, DIC, Alk and O_2 in the mesopelagic and the mixed layers, as well as NO_3, Si and PO_4 in the mesopelagic layer. Concerning the other metrics, whose model predictions are outperformed by the BGC-Argo climatology, we provide suggestions to reduce the model-data misfit and thus to increase the model efficiency. For, PO_4, Si, and NO_3, we propose to test if the uncertain model error covariances during the assimilation of satellite Chla could lead to a degradation in predicting nutrients in the mixed layer. For Chla-related metrics, we recommend to check the consistency between ocean colour products and BGC-Argo Chla products at the global scale as it may explain part of the misfit between the model, that assimilates satellite Chla, and BGC-Argo observations. The discrepancies between modelled and observed POC and OMZs have been already investigated in previous studies. It has been suggested that improving the BGC-Argo POC-b_{bp} conversion factor, tuning the model parameters and implementing missing processes in the model structure could decrease the model-data inconsistencies associated with POC dynamics. Similarly, the improvement of the ocean circulation in physical models should improve the accuracy of OMZs model predictions. Finally, pH_{mixed} and $pCO_{2\ mixed}$ should be better modelled if the uncertainties associated with DIC, Alk, temperature and salinity in the mixed layer are reduced.*

The null hypothesis for establishing that the model is “good” should be defined.

REPLY: We now use the model efficiency statistical score to assess the performance of the model. The null hypothesis for establishing that the model is “good” is tested against the BGC-Argo climatology:

“The model efficiency tests whether the model outperforms the BGC-Argo climatology ($0 < m_e < 1$,Fennel et al., 2022), or stated differently, if the model-data mean square difference is lower than the observation variance, i.e., $\sum_{i=1}^{N}(m_i - o_i)^2 < \sum_{i=1}^{N}(o_i - \bar{o})^2$. “

Also, there are some really interesting of the value and needs for BGC Argo observations in the conclusions that are completely unsupported by the body of the manuscript… if the authors want to bring some of this Appendix material into the manuscript body so as to support these conclusions, (and leave the focus on just the current model) that would also be an appropriate means of turning the paper from a technical report on diagnostics into a scientific research paper.

REPLY: We thank the reviewer for this suggestion. All the Appendix material are now integrated into the manuscript body. We have made the design of the BGC-Argo observing system one the main objective of the study as summarized in the conclusion:

"

The method proposed here is also beneficial to inform about the BGC-Argo network design. In particular, the regions where BGC-Argo observations should be enhanced to reduce the model-data misfit through the assimilation of BGC-Argo data or process-oriented assessment studies. We strongly recommend to enhance the Arctic region, which is critically under sampled and is constantly outperformed by the BGC-Argo climatology. Likewise, BGC-Argo observations should be enriched in the Equatorial region and in the Southern Oceans, two regions where the model error barely exceed the BGC-Argo observations variability. "

Finally, the paper includes a multitude of language mistakes which I have tried to rectify in my technical comments.

REPLY: Thank you very much.

Technical comments:
1-16 – "a major tool" should be "major tools"

REPLY: Thank you, we have made the correction.

2-1 – "or" should be "and". Also, is there a difference between "These metrics" and "The metrics in the sentence before? If not, ". These metrics" should be "and"

REPLY: Thank you, we have made the correction.

2-3 – "suggest" seems an odd word here given that nearly all scientific papers display plots. Perhaps instead of "suggest" should be "recommend as a community standard"

REPLY: This sentence was removed during the rewriting of the main text.

2-7 – “had” should be “has”

REPLY: Thank you, we have made the correction.

2-14 – No, numerical simulations are not necessary “to monitor these ongoing changes”. Instead, the authors could say, “to contextualize monitoring of ongoing changes”

REPLY: Thank you, we have made the correction.

2-23 – remove “being”. Also, the attribution here with “mostly” is overconfidently placed on lack of BGC understanding. In many instances it is lack of understanding of the physics and lack of characterization of the forcing that are the bigger issues than the BGC parameterization.

REPLY: Thank you, we have made the correction.

2-25 – add comma before “and” .

REPLY: Thank you, we have made the correction.

2-30 – add “a” before “few”

REPLY: Thank you, we have made the correction.

3-3 – The list should reflect back to the same part of the sentence, not three different parts. If reflecting back to “to test their”, then it should be “to test their predictive skills, ability to reproduce BGC processes, and confidence intervals on model predictions” or if these are separate statements reflecting back to “their” and “to”, then “to test their predictive skills and ability to reproduce BGC processes and estimate confidence intervals on model predictions”

REPLY: Thank you, we have made the correction.

3-11 – “All these datasets neither have a” Should be “These datasets have neither”

REPLY: Thank you, we have made the correction.

3-12 – remove “can”

REPLY: Thank you, we have made the correction.

3-23 – “so far essentially sampled” should be “well sampled only”

REPLY: Thank you, we have made the correction.

3-24 – Add “the” before “regional”

REPLY: Thank you, we have made the correction.

3-24 – remove comma before “large”

REPLY: Thank you, we have made the correction.

3-25 – remove comma before “like”, and replace “or” with ‘and”

REPLY: Thank you, we have made the correction.

4-4 – “represent” should be “represents”.

REPLY: Thank you, we have made the correction.

Also, while this statement may be true in terms of quantity of data for a few parameters, it is not true in terms of either accuracy or comprehensiveness. Just because you can derive an estimate of SiO_4 from an O_2 sensor does not mean the dataset is better than actually measuring O_2 from a Winkler titration, much less using that value to extrapolate SiO_4.

4-6 – I do not know what “interoperability” means in this context. Is it something about the inherent environmental variability, or the measurement uncertainty?

4-8 – I don’t know what “separately” is being used for here. Are the authors saying that they are initializing the model with “WOA/WOD” and then evaluating performance separately with BGC-Argo? Or that the initialization of the model is done independently from WOA/WOD and then both WOA/WOD and BGC-Argo are used for independent evaluation?

REPLY: We agree that this paragraph was not clear. We have revised it. It now reads:

“

The BGC-Argo data set represents a significant improvement for the assessment of models comparing to large databases such as the World Ocean Database (WOD) (Boyer et al., 2013) or the Copernicus Marine Service in situ dataset (European Union-Copernicus Marine Service, 2015). Large databases are composed of data collected from various instrument types with heterogenous data sampling methodologies. Therefore, for a given variable, the accuracy numbers are not the same and change depending on the instrument type (European Union-Copernicus Marine Service, 2019). Consequently, this affects the overall accuracy over time due to the changing proportion of instrument types over the years. On the other hand, the BGC-Argo data set is an homogenous data set with strict and uniform sampling methodologies and data Quality-Control (QC) procedures. As a result, the BGC-Argo data set have a satisfactory level of accuracy, which remains stable over time (Johnson et al., 2017; Mignot et al., 2019). Moreover, the number of quality-controlled observations collected every year by the BGC-Argo fleet is now greater than any other data set (Claustre et al., 2020). Using the BGC-Argo dataset as the single evaluation data set is therefore a way to ensure consistent accuracy. “

4-18 – The sentence “We expect that the methodology employed here (from the data handling to the use of assessment metrics) would be useful and informative for other research teams interested in model evaluation with BGC-Argo floats.” Belongs in the discussion/conclusions, not in the introduction.

REPLY: This sentence was removed during the rewriting of the main text.

4-27 – “them” should be “these metrics”

REPLY: Thank you, we have made the correction.

4-29 – “. These metrics” should be “and”

REPLY: This sentence was removed during the rewriting of the main text.

4-31 to 5-5 – Again, the sentences beginning “Further, our validation framework could...” to “… is not addressed in this study” Belongs in the discussion/conclusions, not in the introduction.

REPLY: This sentence was removed during the rewriting of the main text.

4-33 – This sentence needs a lot of work. The authors could try adding “have” before “demonstrated”, “flux” before “calculation” and “the” before “basin”, remove the commas and remove “of mass fluxes and process rates” and see if it makes sense.

REPLY: This sentence was removed during the rewriting of the main text.

5-3 “use of the word “arduous” seems odd here. Whether something is hard to do is not necessarily relevant. More relevant is whether the effort is warranted… would it be too uncertain so as not to be robust?

REPLY: This sentence was removed during the rewriting of the main text.

5-7 – “follow: s” Should be “follows. S”

REPLY: Thank you, we have made the correction.

5-26 – “variable” should be “variables,”

REPLY: Thank you, we have made the correction.

5-32 – It would be helpful to site the WCRP standard here for essential climate variables, e.g Bojinski et al, 2014, “The concept of essential climate variables in support of climate research, applications, and policy”, BAMS

REPLY: We have cited the WCRP standard, as proposed by the reviewer.

6-1 – Unclear what is intended for “highest” here. Is it “highest quality” or “highest density” or something else?

REPLY: We agree it was not clear. “highest” is intended for “highest level of data modes”. We have changed it in the manuscript.

6-11,12 – “points” should be “point”

REPLY: Thank you, we have made the correction.

6-18 – So this means that the low values are biased high as the chance of a low positive value includes the possibility of the value being zero. How big is this problem? What fraction of the data had to be adjusted to zero?

REPLY: We now use an improved version of POC/bbp relationship. Consequently, there are no longer negative values.

6-23 – “floats” should be “float”

REPLY: Thank you, we have made the correction.

6-24 – add comma after “salinity”

REPLY: Thank you, we have made the correction.

6-25 – there should be a statement here on the carbon system data source that is used for the training of the algorithm… eventually the skill has to be traced back to the GLODAP or other data source.

REPLY: Ok, we have added a statement on the carbonate system data source.

6-34:7-2 – The authors should note that whether or not it is “reasonable” to draw these conclusions is also entirely reliant on both the BGC Argo data and the model capturing the underlying environmental variability.

REPLY: Ok, we agree.

7-8 remove comma

REPLY: Thank you, we have made the correction.

7-10 – remove “it” after “and”

REPLY: Thank you, we have made the correction.

7-19 – what is the advantage, if there is one, of saving only weekly and then recreating the daily values with interpolation? Is this to speed the model or otherwise reduce data size?

REPLY: The advantage is to reduce the data size.

7-26 – remove “values”. Again, is there an advantage of calculating output offline? Are CO2 fluxes calculated online and saved out? Perhaps it would be better to move this to the next section where the CO2 flux calculation is discussed.

REPLY: This sentence was removed during the rewriting of the main text.

7-32 – and “space to” between “and” and “the”

REPLY: Thank you, we have made the correction.

8-3 – The bias in MLD is provided, but what is the average MLD that would allow me to know the % bias?

REPLY: The BIAS is now indicated in %.

“*The overall mean square difference between the model and the data is equal to ~30% of the overall variance of the observations*”.

8-28 – This is a strange phrasing. It sounds from this that acidification does not impact the subsurface down to 200 m, on the “surface” and the 200-400 m range… Why not just say that acidification is expected to have its largest impact in the upper 400 m and then separately that the present analysis chooses the 200-400 m range of Kwiatkowski? Presumably the surface and 200-400 m ranges are shown to highlight different signals rather than to suggest the area in between is unimportant. This should b clarified.

REPLY: This sentence was removed during the rewriting of the main text.

9-12 – I would replace “first level” with “most simple but indirect level”, 9-13 – replace “of” with “associated with” and 9-17 – replace “second level” with “more process level” since the “second level” isn’t being pursued.

REPLY: Thank you, we have made the corrections.

9-22:9-26 – A brief statement and reference on the motivation for providing these mesopelagic estimates is warranted. Also, is there a reference or other rationale for this choice of varying depth range? This MLD-1000 m variable depth definition would seem to include the part of the euphotic zone below the mixed layer as “mesopelagic”, at least during the growing season. I would have thought the area below the mixed layer within the euphotic zone to look more like the surface than the mesopelagic, or “twilight zone”, a constant 200-1000 m range would have been easier to interpret, particularly against the 200-400 definition for pH.

REPLY: Ok , we have added a statement and reference on the motivation for providing these mesopelagic estimates:

“This two-layer comparison between model and BGC-Argo data provides an indirect evaluation of the key mesopelagic processes and fluxes associated with the carbonate chemistry, biological carbon pump and oxygen levels in the mixed, and mesopelagic layers.”

We have also added a rationale for the choice of this varying depth range:
“

The mesopelagic layer is defined as the layer between the MLD and 1000m. For simplicity, we use a simplified definition of the mesopelagic layer proposed by Dall' Olmo and Mork (2014). In their study, this layer is comprised between the deepest of the euphotic layer depth and the MLD, and 1000 m"

9-33 – remove second "processes" and end sentence after "production"

REPLY: Thank you, we have made the corrections.

10-10 – This sentence is very misleading. The vertical supply of NO3 to the surface is accompanied with remineralized DIC which is the reverse of the biological carbon pump. This sentence should be reworded.

REPLY: OK, we have reworded the sentence.

20-32 - Why define a biased average for O2 300? Shouldn't the average oxygen between 250-300 be referred to O2 275? Why not use the same 200-400 definition as pH? Or 250-350?

11-2 – Similarly, why define O2 1000 as O2 950-1000? Should this be o2 975, or alternatively, defined as 950-1050… do the floats only go down to 1000m? This would seem a reasonable justification if it were the case since gradients at this depth tend to be weak, but still wouldn't explain the odd 250-300 definition.

REPLY: This section was removed during the rewriting of the main text.

12-12 – "on a climatological level" should be "as a climatology"

REPLY: This section was removed during the rewriting of the main text.

12-13 – what is the purpose of "etc.."? "imposes" should be "requires"

12-14 – "in a climatological way" should be "as a climatology"
12-19 – why is "Biogeochemical-Argo Planning Group, 2016" in parenthesis here. Was this means of gridding a recommendation from this group? If so, please be explicit.
12-21 – "clarity" should be "clarity in visualization" or "simplicity in visualization"

12-26 – Add “While” before Taylor” and replace “but” with a comma in the next sentence. That would make It more clear that you are introducing a new topic rather than simply revisiting how great are the first three presentation methods.
12-33 – “for” should be “in”

13-7 – The sentence “Examples of the diagnostic plots described in section 4 in combination with the metrics defined in Section 3 are shown.” Seems redundant with the orientation statement in the introduction section and should be removed.

REPLY: These sections were removed during the rewriting of the main text.

13-16:13-25 – The null hypothesis that the reader should use to define “well represented” are not clear. Isn’t much or all of this fidelity due to the initial condition derived through the assimilation? I am not sure what to take from this. Is there an “unassimilated” version of the model with which the assimilation should be compared? Or a previous generation model? Or other unassimilative models such as CMIP6? Or is the objective just to show the broad contrast in pattern agreement between model and observations across variables? Why is pH so poorly predicted?

REPLY: As explained before, we now use the model efficiency statistical score to assess the performance of the model. The null hypothesis for establishing that the model is “good” is tested against the BGC-Argo climatology:

“The model efficiency tests whether the model outperforms the BGC-Argo climatology ($0 < m_e < 1$,Fennel et al., 2022), or stated differently, if the model-data mean square difference is lower than the observation variance, i.e., $\sum_{i=1}^{N}(m_i - o_i)^2 < \sum_{i=1}^{N}(o_i - \bar{o})^2$. “

We also provide some explanation as to why pH is so poorly predicted.

14-1:14-4 – This discussion of the value of Taylor diagrams is very superficial and somewhat misleading. The presentation here certainly shows what patterns and variability in different variables are relatively well reproduced, but whether this should inform future model development priorities entirely depends on the intended use of the model and associated requirements. Further, the most common scientific use of Taylor diagrams is the comparison of the same metric across models so that one can quantify the improvements.

REPLY: We agree. This section was removed during the rewriting of the main text.

14-25 – Without a frame of reference, it is not at all clear whether the model is good or bad. Like in the case of the Taylor diagrams, it seems like the analysis is being done in a vacuum without any awareness of other modeling efforts. There is also the lack of appreciation of the satellite derived estimate for this metric.

REPLY: We agree. This section was removed during the rewriting of the main text.

18-18 – The conclusions "Here, we showed that the spatial maps of model-observations comparison are also informative a posteriori, with respect to the network design, as they highlight sensitive areas where BGC-Argo observations are critical and where sustained BGC-Argo observations are required to better constrain the model. These maps correspond to the regions where the model uncertainty (see RMSD spatial maps in Figs. A22-A44) is the highest, i.e., the Equatorial belt with respect to the carbonate system variables, the Southern Ocean with respect to the nutrients and the DCM variables, and the western boundary currents and OMZs with respect to oxygen." Are very interesting scientific research conclusions but are not at all discussed in the body of the manuscript. This is totally unacceptable. The paper cannot bring in unsupported information at the conclusion stage referencing Appendix material. The authors need to show this or restate these conclusions as hypotheses for future work.

REPLY: We agree, as explained above. We have made the design of the BGC-Argo observing system one the main objective of the study as summarized in the conclusion:

"*The method proposed here is also beneficial to inform about the BGC-Argo network design. In particular, the regions where BGC-Argo observations should be enhanced to reduce the model-data misfit through the assimilation of BGC-Argo data or process-oriented assessment studies. We strongly recommend to enhance the Arctic region, which is critically under sampled and is constantly outperformed by the BGC-Argo climatology. Likewise, BGC-Argo observations should be enriched in the Equatorial region and in the Southern Oceans, two regions where the model error barely exceed the BGC-Argo observations variability.* "

References

Allen, J. I., Somerfield, P. J., and Gilbert, F. J.: Quantifying uncertainty in high-resolution coupled hydrodynamic-ecosystem models, J. Mar. Syst., 64, 3–14, https://doi.org/10.1016/j.jmarsys.2006.02.010, 2007.

Bittig, H. C., Steinhoff, T., Claustre, H., Fiedler, B., Williams, N. L., Sauzède, R., Körtzinger, A., and Gattuso, J.-P.: An alternative to static climatologies: robust estimation of open ocean CO2 variables and nutrient concentrations from T, S, and O2 data using Bayesian neural networks, Front. Mar. Sci., 5, 328, 2018.

Boyer, T. P., Antonov, J. I., Baranova, O. K., Garcia, H. E., Johnson, D. R., Mishonov, A. V., O'Brien, T. D., Seidov, D., Smolyar, I., and Zweng, M. M.: World ocean database 2013, 2013.

Claustre, H., Johnson, K. S., and Takeshita, Y.: Observing the Global Ocean with Biogeochemical-Argo, Annu. Rev. Mar. Sci., 12, annurev-marine-010419-010956, https://doi.org/10.1146/annurev-marine-010419-010956, 2020.

Dall'Olmo, G. and Mork, K. A.: Carbon export by small particles in the Norwegian Sea, Geophys. Res. Lett., 41, 2921–2927, https://doi.org/10.1002/2014GL059244, 2014.

European Union-Copernicus Marine Service: Global Ocean- In-Situ Near-Real-Time Observations, https://doi.org/10.48670/MOI-00036, 2015.

European Union-Copernicus Marine Service: Global Ocean Biogeochemistry Analysis and Forecast, https://doi.org/10.48670/MOI-00015, 2019.

European Union-Copernicus Marine Service: Global Ocean Colour (Copernicus-GlobColour), Bio-Geo-Chemical, L4 (monthly and interpolated) from Satellite Observations (Near Real Time), https://doi.org/10.48670/MOI-00279, 2022.

Fennel, K., Mattern, J. P., Doney, S. C., Bopp, L., Moore, A. M., Wang, B., and Yu, L.: Ocean biogeochemical modelling, Nat. Rev. Methods Primer, 2, 1–21, https://doi.org/10.1038/s43586-022-00154-2, 2022.

Ford, D.: Assimilating synthetic Biogeochemical-Argo and ocean colour observations into a global ocean model to inform observing system design, Biogeosciences, 18, 509–534, https://doi.org/10.5194/bg-18-509-2021, 2021.

Gutknecht, E., Reffray, G., Mignot, A., Dabrowski, T., and Sotillo, M. G.: Modelling the marine ecosystem of Iberia-Biscay-Ireland (IBI) European waters for CMEMS operational applications, Ocean Sci., 15, 1489–1516, https://doi.org/10.5194/os-15-1489-2019, 2019.

Hipsey, M. R., Gal, G., Arhonditsis, G. B., Carey, C. C., Elliott, J. A., Frassl, M. A., Janse, J. H., de Mora, L., and Robson, B. J.: A system of metrics for the assessment and improvement of aquatic ecosystem models, Environ. Model. Softw., 128, 104697, https://doi.org/10.1016/j.envsoft.2020.104697, 2020.

Johnson, Plant, J. N., Coletti, L. J., Jannasch, H. W., Sakamoto, C. M., Riser, S. C., Swift, D. D., Williams, N. L., Boss, E., Haëntjens, N., Talley, L. D., and Sarmiento, J. L.: Biogeochemical sensor performance in the SOCCOM profiling float array: SOCCOM BIOGEOCHEMICAL SENSOR PERFORMANCE, J. Geophys. Res. Oceans, 122, 6416–6436, https://doi.org/10.1002/2017JC012838, 2017.

Mignot, A., D'Ortenzio, F., Taillandier, V., Cossarini, G., and Salon, S.: Quantifying Observational Errors in Biogeochemical-Argo Oxygen, Nitrate, and Chlorophyll *a* Concentrations, Geophys. Res. Lett., 46, 4330–4337, https://doi.org/10.1029/2018GL080541, 2019.

Russell, J. L., Kamenkovich, I., Bitz, C., Ferrari, R., Gille, S. T., Goodman, P. J., Hallberg, R., Johnson, K., Khazmutdinova, K., and Marinov, I.: Metrics for the evaluation of the Southern Ocean in coupled climate models and earth system models, J. Geophys. Res. Oceans, 123, 3120–3143, 2018.

Salon, S., Cossarini, G., Bolzon, G., Feudale, L., Lazzari, P., Teruzzi, A., Solidoro, C., and Crise, A.: Novel metrics based on Biogeochemical Argo data to improve the model uncertainty evaluation of the CMEMS Mediterranean marine ecosystem forecasts, Ocean Sci., 15, 997–1022, https://doi.org/10.5194/os-15-997-2019, 2019.

Terzić, E., Lazzari, P., Organelli, E., Solidoro, C., Salon, S., D'Ortenzio, F., and Conan, P.: Merging bio-optical data from Biogeochemical-Argo floats and models in marine biogeochemistry, Biogeosciences, 16, 2527–2542, https://doi.org/10.5194/bg-16-2527-2019, 2019.

Williams, R. G. and Follows, M. J.: Ocean dynamics and the carbon cycle: Principles and mechanisms, Cambridge University Press, 2011.